Appears in European Conference in Computer Vision, Copenhagen, Denmark, 2002.

What is the Role of Independence for Visual Recognition?

Nuno Vasconcelos[1]* and Gustavo Carneiro[2]

[1] Compaq Computer Corporation,
Cambridge Research Laboratory,
nuno.vasconcelos@compaq.com
[2] Department of Computer Science,
University of Toronto,
carneiro@cs.toronto.edu

Abstract. Independent representations have recently attracted significant attention from the biological vision and cognitive science communities. It has been 1) argued that properties such as sparseness and independence play a major role in visual perception, and 2) shown that imposing such properties on visual representations originates receptive fields similar to those found in human vision. We present a study of the impact of feature independence in the performance of visual recognition architectures. The contributions of this study are of both theoretical and empirical natures, and support two main conclusions. The first is that the intrinsic complexity of the recognition problem (Bayes error) is higher for independent representations. The increase can be significant, close to 10% in the databases we considered. The second is that criteria commonly used in independent component analysis are not sufficient to eliminate all the dependencies that impact recognition. In fact, "independent components" can be less independent than previous representations, such as principal components or wavelet bases.

1 Introduction

After decades of work in the area of visual recognition (in the multiple guises of object recognition, texture classification, and image retrieval, among others) there are still several fundamental questions on the subject which, by large, remain unanswered. One of the core components of any recognition architecture is the feature transformation, a mapping from the space of image pixels to a feature space with better properties for recognition. While numerous features have been proposed over the years for various recognition tasks, there has been small progress towards either 1) a universally good feature set, or 2) a universal and computationally efficient algorithm for the design of optimal features for any particular task.

In the absence of indisputable universal guidelines for feature design, one good source of inspiration has always been the human visual system. Ever since

* A longer version of the paper containing all proofs is available from the author.

the work of Hubel and Wiesel [9], it has been established that 1) visual processing is local, and 2) different groups in primary visual cortex (i.e. area V1) are tuned for detecting different types of stimulus (e.g. bars, edges, and so on). This indicates that, at the lowest level, the architecture of the human visual system can be well approximated by a multi-resolution representation localized in space and frequency, and several "biologically plausible" models of early vision are based on this principle [2, 3, 8, 13, 18, 19]. All these models share a basic common structure consisting of three layers: a *space/space-frequency* decomposition at the bottom, a middle stage introducing a non-linearity, and a final stage pooling the responses from several non-linear units. They therefore suggest the adoption of a mapping from pixel-based to space/space-frequency representations as a suitable universal feature transformation for recognition.

A space/space-frequency representation is obtained by convolving the image with a collection of elementary filters of reduced spatial support and tuned to different spatial frequencies and orientations. Traditionally, the exact shape of the filters was not considered very important, as long as they were localized in both space and frequency, and several elementary filters have been proposed in the literature, including *differences of Gaussians* [13], *Gabor functions* [16, 8], and *differences of offset Gaussians* [13], among others. More recently, this presumption has been challenged by various authors on the basis that the shape of the filters determines fundamentally important properties of the representation, such as sparseness [7, 15] and independence [1].

These claims have been supported by (quite successful) showings that the enforcement of sparseness or independence constraints on the design of the feature transformation leads to representations which exhibit remarkable similarity to the receptive fields of cells found in V1 [1, 15]. However, while the arguments are appealing and the pictures compelling, there is, to the best of our knowledge, no proof that sparseness or independence are, indeed, fundamental requirements for visual recognition. On the contrary, not all evidence supports this conjecture. For example, detailed statistical analysis of the coefficients of wavelet transforms (an alternative class of sparse features which exhibit similar receptive fields) has revealed the existence of clear inter-dependencies [17].

In what concerns the design of practical recognition systems, properties such as sparseness or independence are important only insofar as they enable higher-level goals such as computational efficiency or small probability of error. Under a Bayesian view of perception [12], these two goals are, in fact, closely inter-related: implementation of minimum probability of error (MPE) decisions requires accurate density estimates, which are very difficult to obtain in high-dimensional feature spaces. The advantage of an independent representation is to decouple the various dimensions of the space, allowing high dimensional estimates to be computed by the simple multiplication of scalars. In this sense, independence can be a crucial enabler for accurate recognition with reduced complexity. On the other hand, it is known that any feature transformation has the potential to increase Bayes error, the ultimate lower-bound on the probability of error that any recognition architecture can achieve, for a given feature space. It is not clear that independent feature spaces are guaranteed to exhibit lower Bayes error than non-independent ones. In fact, since the independence constraint restricts the set of admissible transforms, it is natural to expect the opposite.

Due to all of this, while there seem to be good reasons for the use of independent or sparse representations, it is not clear that they will lead to optimal recognition. Furthermore, it is usually very difficult to determine, in practice, if goals such as independence are actually achieved. In fact, because guaranteeing independence is a terribly difficult endeavor in high-dimensions, independent component analysis techniques typically resort to weaker goals, such as minimizing certain cumulants, or searching for non-Gaussian solutions. While an independent representation will meet these weaker goals, the reverse does not usually hold. In practice, it is in general quite difficult to evaluate by how much the true goal of independence has been missed.

In this work we address two questions regarding the role of independence. The first is fundamental in nature: "how important is independence for visual recognition?". The second is relevant for the design of recognition systems: "how realistic is the expectation of actually enforcing independence constraints in real recognition scenarios?". To study these questions we built a complete recognition system and compared the performance of various feature transforms which claim different degrees of independence: from generic features that make no independence claims (but were known to have good recognition performance), to features (resulting from independent component analysis) which are supposed to be independent, passing through transforms that only impose very weak forms of independence, such as decorrelation.

It turns out that, with the help of some simple theoretical results, the analysis of the recognition accuracy achieved by the different transforms already provides significant support for the following qualitative answers to the questions above. First, it seems to be the case that imposing independence constraints increases the intrinsic complexity (Bayes error) of the recognition problem. In fact, our data supports the conjecture that this intrinsic complexity is monotonically increasing on the degree of independence. Second, it seems clear that great care needs to be exercised in the selection of the independence measures used to guide the design of independent component transformations. In particular, our results show that approaches such as minimizing cumulants or searching for non-Gaussian solutions are not guaranteed to achieve this goal. In fact, they can lead to "independent components" that are less independent than those achieved with "decorrelating" representations such as principal component analysis or wavelets.

2 Bounds on recognition accuracy

A significant challenge for empirical evaluation is to provide some sort of guarantees that the observed results are generalizable. This challenge is particularly relevant in the context of visual recognition, since it is impossible to implement all the recognition architectures that could ever be conceived. For example, the fact that we rely on a Bayesian classification paradigm should not compromise the applicability of the conclusions to recognition scenarios based on alternative classification frameworks (e.g. discriminant techniques such as neural networks or support vector machines). This goal can only be met with recourse to theoretical insights on the performance of recognition systems, which are typically available in the form of bounds on the probability of classification error.

The most relevant of these bounds is that provided by the Bayes error, which is the minimum error that any architecture can achieve in a given classification problem.

Theorem 1 *Given a feature space $\mathcal{X}$ and a query $\mathbf{x} \in \mathcal{X}$, the decision function which minimizes the probability of classification error is the Bayes or maximum a posteriori (MAP) classifier*

$$g^*(\mathbf{x}) = \arg\max_i P_{Y|\mathbf{X}}(i|\mathbf{x}), \tag{1}$$

where Y is a random variable that assigns $\mathbf{x}$ to one of M classes, and $i \in \{1, \ldots, M\}$. Furthermore, the probability of error is lower bounded by the Bayes error

$$L^* = 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})], \tag{2}$$

where $E_{\mathbf{x}}$ means expectation with respect to $P_{\mathbf{X}}(\mathbf{x})$.

The significance of this theorem is that any insights on the Bayes error that may be derived from observations obtained with a particular recognition architecture are valid for all architectures, as long as the feature space $\mathcal{X}$ is the same. The following theorem shows that a feature transformation can never lead to smaller error in the transformed space than that achievable in the domain space.

Theorem 2 *Given a classification problem with observation space $\mathcal{Z}$ and a feature transformation*

$$T : \mathcal{Z} \rightarrow \mathcal{X},$$

then

$$L^*_{\mathcal{X}} \geq L^*_{\mathcal{Z}} \tag{3}$$

where $L^*_{\mathcal{Z}}$ and $L^*_{\mathcal{X}}$ are, respectively, the Bayes errors on $\mathcal{Z}$ and $\mathcal{X}$. Furthermore, equality is achieved if and only if T is an invertible transformation.

The last statement of the theorem is a worst-case result. In fact, for a specific classification problem, it may be possible to find non-invertible feature transformations that do not increase Bayes error. What is not possible is to find 1) a feature transformation that will reduce the Bayes error, or 2) a universal non-invertible feature transformation guaranteed not to increase the Bayes error on all classification problems.

Since Bayes error is an intrinsic measure of the complexity of a classification problem, the theorems above are applicable to any classification architecture. The following upper bounds are specific to a family of architectures that we will consider throughout this work, and are usually referred to as plug-in decision rules [6]. The basic idea is to rely on Bayes rule to invert (1)

$$g^*(\mathbf{x}) = \arg\max_i P_{\mathbf{X}|Y}(\mathbf{x}|i)P_Y(i), \tag{4}$$

and then estimate the quantities $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ and $P_Y(i)$ from training images. This leads to the following upper bound on the probability of error.

Theorem 3 *Given a classification problem with a feature space $\mathcal{X}$, unknown class probabilities $P_Y(i)$ and class conditional likelihood functions $P_{\mathbf{X}|Y}(\mathbf{x}|i)$, and a decision function*

$$g(\mathbf{x}) = \arg\max_i \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i), \tag{5}$$

the difference between the actual and Bayes error, is upper bounded by

$$P(g(\mathbf{X}) \neq Y) - L^*_{\mathcal{X}} \leq \sum_i \int |P_{\mathbf{X}|Y}(\mathbf{x}|i)P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i)|d\mathbf{x}. \tag{6}$$

In the remainder of this work we assume that the classes are a-priori equiprobable, i.e. $P_Y(i) = 1/M, \forall i$. This leads to the following corollary.

Corollary 1 *Given a classification problem with equiprobable classes, a feature space $\mathcal{X}$, unknown class conditional likelihood functions $P_{\mathbf{X}|Y}(\mathbf{x}|i)$, and a decision function*

$$g(\mathbf{x}) = \arg\max_i \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i), \tag{7}$$

the difference between the actual and Bayes error is upper bounded by

$$P(g(\mathbf{X}) \neq Y) - L^*_{\mathcal{X}} \leq \Delta_{g,\mathcal{X}} \tag{8}$$

where

$$\Delta_{g,\mathcal{X}} = \sum_i KL[P_{\mathbf{X}|Y}(\mathbf{x}|i)||\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)], \tag{9}$$

is the estimation error and

$$KL[P_{\mathbf{X}}(\mathbf{x})||Q_{\mathbf{X}}(\mathbf{x})] = \int P_{\mathbf{X}}(\mathbf{x}) \log \frac{P_{\mathbf{X}}(\mathbf{x})}{Q_{\mathbf{X}}(\mathbf{x})} d\mathbf{x} \tag{10}$$

is the relative entropy, or Kullback-Leibler divergence, between $P_{\mathbf{X}}(\mathbf{x})$ and $Q_{\mathbf{X}}(\mathbf{x})$.

Bounds (3) and (8) reflect the impact of both feature selection and density estimation on recognition accuracy. While the feature transformation determines the best possible achievable performance, the quality of the density estimates determines how close the actual error is to this lower bound. Hence, for problems where density estimation is accurate one expects the actual error to be close to the Bayes error. On the other hand, when density estimates are poor, there are no guarantees that this will be the case.

The latter tends to be the case for visual recognition, where high-dimensional feature spaces usually make density estimation a difficult problem. It is, therefore, difficult to determine if the error is mostly due to the intrinsic complexity of the problem (Bayes error) or to poor quality of density estimates. One of the contributions of this work is a strategy to circumvent this problem, based on the notion of embedded feature spaces [21].

Definition 1 *Given two vector spaces $\mathcal{X}_m$ and $\mathcal{X}_n$, $m < n$, such that $dim(\mathcal{X}_m) = m$ and $dim(\mathcal{X}_n) = n$ an embedding is a mapping*

$$\epsilon : \mathcal{X}_m \to \mathcal{X}_n \tag{11}$$

which is one-to-one.

A canonical example of embedding is the zero padding operator for Euclidean spaces

$$\iota_m^n : \mathbb{R}^m \to \mathbb{R}^n \tag{12}$$

where $\iota_m^n(\mathbf{x}) = (\mathbf{x}, \mathbf{0})$, $\mathbf{x} \in \mathbb{R}^m$, and $\mathbf{0} \in \mathbb{R}^{n-m}$.

Definition 2 *A sequence of vector spaces $\{\mathcal{X}_1, \ldots, \mathcal{X}_d\}$, such that $dim(\mathcal{X}_i) < dim(\mathcal{X}_{i+1})$, is called embedded if there exists a sequence of embeddings*

$$\epsilon_i : \mathcal{X}_i \to \mathcal{X}'_{i+1}, \; i = 1, \ldots, d-1, \tag{13}$$

such that $\mathcal{X}'_{i+1} \subset \mathcal{X}_{i+1}$.

The inverse operation of an embedding is a submersion.

Definition 3 *Given two vector spaces $\mathcal{X}_m$ and $\mathcal{X}_n$, $m < n$, such that $dim(\mathcal{X}_m) = m$ and $dim(\mathcal{X}_n) = n$ a submersion is a mapping*

$$\gamma : \mathcal{X}_n \to \mathcal{X}_m \tag{14}$$

which is surjective.

A canonical example of submersion is the projection of Euclidean spaces along the coordinate axes

$$\pi_m^n : \mathbb{R}^n \to \mathbb{R}^m \tag{15}$$

where $\pi_m^n(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n) = (x_1, \ldots, x_m)$. The following theorem shows that any linear feature transformation originates a sequence of embedded vector spaces with monotonically decreasing Bayes error, and monotonically increasing estimation error.

Theorem 4 *Let*

$$T : \mathbb{R}^d \to \mathcal{X} \subset \mathbb{R}^d,$$

be a linear feature transformation. Then,

$$\mathcal{X}_i = \pi_i^d(\mathcal{X}), i = 1, \ldots, d-1 \tag{16}$$

is a sequence of embedded feature spaces such that

$$L^*_{\mathcal{X}_{i+1}} \leq L^*_{\mathcal{X}_i}. \tag{17}$$

Furthermore, if $\mathbf{X}_1^d = \{\mathbf{X}_1, \ldots, \mathbf{X}_d\}$ *is a sequence or random variables such that* $\mathbf{X}_i \in \mathcal{X}_i$,

$$\mathbf{X}_i = \pi_i^d(\mathbf{X}), i = 1, \ldots, d \tag{18}$$

and $\{g(\mathbf{x})\}_1^d$ *a sequence of decision functions*

$$g_i(\mathbf{x}) = \arg\max_k \hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}|k) \tag{19}$$

then

$$\Delta_{g_{i+1},\mathcal{X}_{i+1}} \geq \Delta_{g_i,\mathcal{X}_i}. \tag{20}$$

Figure 1 illustrates the evolution of the upper and lower bounds on the probability of error as one considers successively higher-dimensional subspaces of $\mathcal{X}$. Since accurate density estimates can usually be obtained in low-dimensions, the two bounds tend to be close when the subspace dimension is small. In this case, the actual probability of error is dominated by the Bayes error. For higher-dimensional subspaces two distinct scenarios are possible, depending on the independence of the individual random variables X_i. Whenever these variables are dependent, the decrease in Bayes error tends to be cancelled by an increase in estimation error and the actual probability of error increases. In this case, the actual probability of error exhibits the concave shape depicted in the left plot, where an inflection point marks the subspace dimension for which Bayes error ceases to be dominant.

The right plot depicts the situation where the variables X_i are independent. In this case, it can be shown that

$$\Delta_{g_{i+1},\mathcal{X}_{i+1}} - \Delta_{g_i,\mathcal{X}_i} = \sum_k KL[P_{X_{i+1}|Y}(x|k)||\hat{p}_{X_{i+1}|Y}(x|k)], \tag{21}$$

i.e. the increase in overall estimation error is simply the sum of the errors of the individual scalar estimates. Since these errors tend to be small, one expects the overall probability of error to remain approximately flat.

Hence, the shape of the curve of probability of error as a function of the subspace dimension carries significant information about 1) the Bayes error in the full space $\mathcal{X}$ and 2) the independence of the component random variables X_i. We will see in section 5 that this information is sufficient to draw, with reasonable certainty, conclusions such as "the Bayes error of transform T is greater than that of transform U". With regards to independence, the ultimate test is, of course, to implement a recognition system based on estimates of the joint density $P_{\mathbf{X}}(\mathbf{x})$ and compare with a recognition system based on the independence assumption, i.e. $P_{\mathbf{X}}(\mathbf{x}) = \prod_i P_{X_i}(x_i)$. When independence holds, the two systems will achieve the same recognition rates. From now on, we will refer to the former system as based on *joint modeling* and to the latter as based on *independent modeling* or on the *product of marginals*.

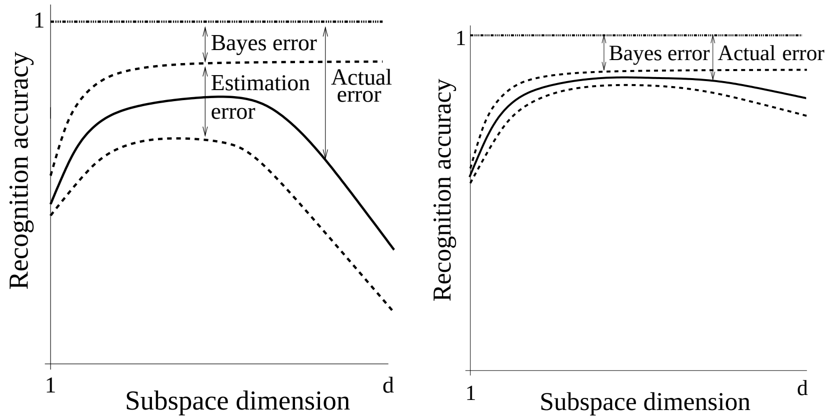

Fig. 1. Upper bound, lower bound, and actual probability of error as a function of subspace dimension. Left: dependent features. Right: independent features.

3 Feature transforms

Since the goal is to evaluate the impact of independence on visual recognition, it is natural to study transformations that lead to features with different degrees of independence. We restrict our attention to the set of transformations that perform some sort of space/space-frequency decomposition. In this context, the feature transformation is a mapping

$$\begin{aligned} T : \mathbb{R}^k &\to \mathbb{R}^d \\ \mathbf{z} &\to \mathbf{x} = \mathbf{W}\mathbf{z} \end{aligned}$$

where $\mathbf{z} \in \mathbb{R}^k$ is a $n \times n$ image patch with columns stacked into a k-dimensional vector ($k = n^2$) and $\mathbf{W}$ the transformation matrix. In general, $k \geq d$, and one can also define a reconstruction mapping

$$\begin{aligned} R : \mathbb{R}^d &\to \mathbb{R}^k \\ \mathbf{x} &\to \mathbf{z} = \mathbf{A}\mathbf{x} \end{aligned}$$

from features $\mathbf{x}$ to pixels $\mathbf{z}$. The columns of $\mathbf{A}$ are called basis functions of the transformation. When $d = k$ and $\mathbf{A} = \mathbf{W}^T$ the transformation is orthogonal. Various popular space/space-frequency representations are derived from orthogonal feature transforms.

Definition 4 *The Discrete Cosine Transform (DCT) [11] of size n is the orthogonal transform whose basis functions are defined by:*

$$A(i,j) = \alpha(i)\alpha(j)\cos\frac{(2x+1)i\pi}{2n}\cos\frac{(2y+1)j\pi}{2n},\ 0 \leq i,j,x,y < n \qquad (22)$$

where $\alpha = \sqrt{1/n}$ for $i = 0$, and $\alpha = \sqrt{2/n}$ otherwise.

The DCT has empirically been shown to have good decorrelation properties [11] and, in this sense, DCT features are at the bottom of the independence spectrum. Previous recognition results had shown, however, that it can lead to recognition rates comparable to or better than those of many features proposed in the recognition literature [20]. It is possible to show that, for certain classes of stochastic processes, the DCT converges asymptotically to the following transform [11].

Definition 5 *Principal Components Analysis (PCA) is the orthogonal transform defined by*

$$\mathbf{W} = \mathbf{D}^{-1/2}\mathbf{E}^T, \tag{23}$$

where $\mathbf{EDE}^T$ *is the eigenvector decomposition of the covariance matrix* $E[\mathbf{zz}^T]$.

It is well known (and straightforward to show) that PCA generates uncorrelated features, i.e. $E[\mathbf{xx}^T] = \mathbf{I}$. While they originate spatial/spatial-frequency representations, the major limitation of the above transforms as models for visual perception is the arbitrary nature of their spatial localization (enforced by arbitrarily segmenting images into blocks). This can result in severe scaling mismatches if the block size does not match that of the image detail. Such scaling problems are alleviated by the wavelet representation.

Definition 6 *A wavelet transform (WT) [14] is the orthogonal transform whose basis functions are defined by*

$$A(i,j) = \sqrt{2^{k+l}}\Psi\left(2^k x - i\right)\Psi\left(2^l y - j\right)_{(0,0)\leq(i,j)<(2^k,2^l)}^{0\leq k,l<\log_2 n} \tag{24}$$

where $\Psi(x)$ *is a function (wavelet) that integrates to zero.*

Like the DCT, wavelets have been shown empirically to achieve good decorrelation. While this is an important part of independence (all of it when the inputs are Gaussian) there is in general a significant amount of higher-order dependencies that cannot be captured by orthogonal components [15]. Eliminating such dependencies is the goal of independent component analysis.

Definition 7 *Independent Component Analysis (ICA) [4] is a feature transform such that*

$$P_{\mathbf{X}}(\mathbf{x}) = \prod_i P_{\mathbf{X}_i}(\mathbf{x}_i) \tag{25}$$

where $\mathbf{X} = \{X_1, \ldots, X_d\}$ *is the random process from which feature vectors are drawn.*

An equivalent definition is to require that the mutual information between features is zero (see [1] for details). The exact details of ICA depend on the particular algorithm used to learn the basis from a training sample. Since independence is usually difficult to measure and enforce if d is large, ICA techniques tend to settle for less ambitious goals. The most popular solution is to minimize a contrast function which is guaranteed to be zero if the inputs are independent. Examples of such contrast functions are higher order correlations and information-theoretic objective functions[4]. In this work, we consider representatives from the two types: the method developed by Comon [5], which uses a contrast function based on high-order cumulants, and the FastICA algorithm [10], that relies on the negative entropy of the features.

4 Experimental set-up

In order to evaluate the recognition accuracy achievable with the various feature transformations, we conducted experiments on two image databases: the Brodatz texture database, and the Corel database of stock photography. Brodatz is a standard benchmark for texture classification under controlled imaging conditions, and no distractors. Corel is a good testing ground for recognition in the context of natural scenes (e.g. no control over lighting or object pose, cluttered backgrounds).

Brodatz contains 112 gray-scale textures that were broken down into 9 128×128 patches, leading to a total of 1008 images. This set was split into two subgroups, a *query* database containing the first patch of each texture and a *retrieval* database containing the remaining 8. In the case of Corel, we selected 15 image classes[1] each containing 100 color images. We then created a query and retrieval database by assigning each image to the query set with a probability 0.2.

All color images were converted to the YBR color space. Where applicable, the feature transformations were applied to each channel separately and the resulting feature vectors combined by interleaving the color components according to the pattern $YBRYBR\ldots$. For each channel, the feature space was 64-dimensional (three layers of wavelet decomposition and 8×8 image blocks in the remaining cases) and consecutive observations were extracted with a step of 2 (Brodatz) or 4 (Corel) pixels in each of the x and y directions. Public domain software by the authors of the techniques was used for learning the feature transformations. All learning was based in two $100,000$-point samples extracted randomly from the retrieval databases. Figure 2 presents the basis functions learned from Brodatz for PCA, ICA with the method by P. Comon, and ICA with the FastICA algorithm, as well as the DCT basis (wavelet basis do not have block-based support and are not shown).

Once the different bases were computed, all image patches were projected into each of them leading to a sample of feature vectors per image. Maximum likelihood (ML) parameters of a Gaussian mixture model were then estimated using the EM algorithm. The number of Gaussian components was held constant (several values were tried with qualitatively similar results, here we report results with 8 components), and a joint density for each of the embedded subspaces $\mathcal{X}_i$ was obtained by downward-projection of the joint density in $\mathcal{X}$ [21]. A Gaussian

[1] Arabian horses, Auto racing, Owls, Roses, Ski scenes, religious stained glass, sunsets and sunrises, coasts, Divers and diving, Land of the pyramids (pictures of Egypt), English country gardens, fireworks, Glaciers and mountains, Mayan and Aztec ruins, and Oil Paintings.

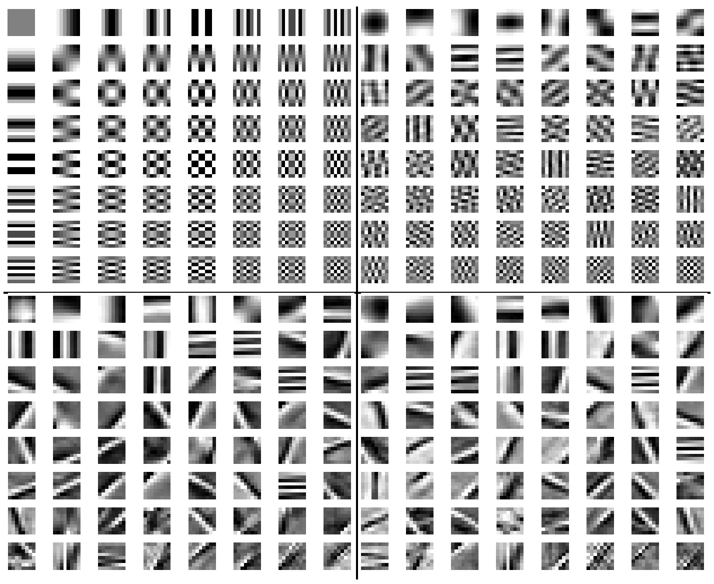

Fig. 2. Basis functions for DCT (top left), PCA (top right) ICA learned with Comon's method (bottom left) and ICA learned with the fastICA method (bottom right).

mixture with the same number of components was also fit to each of the scalar variables X_i to obtain the independent model.

To double-check the independence results we computed various statistical measures of independence. The first was the KL divergence between the joint and independent models $KL\left[P_{\mathbf{X}}(\mathbf{x})||\prod_i P_{X_i}(x_i)\right]$. Since we wanted an alternative measure of independence not affected by the quality of the mixture parameter estimates, we used histograms to compute this statistic. However, in order to avoid well known problems of histogram-based estimates in high dimensions, we only considered average pairwise divergences

$$\hat{KL}(X_i) = \frac{1}{d-1}\sum_{j\neq i} KL\left[P_{X_i,X_j}(x_i,x_j)||P_{X_i}(x_i)P_{X_j}(x_j)\right]. \qquad (26)$$

These divergence are measures of pairwise independence and should be zero whenever independence holds.

One popular way to measure dependencies of order larger than two is through high-order statistics, such as cross-cumulants. While the 2^{nd} order cross-cumulant

$$Cum[X_i, X_j] = E[X_i X_j], \forall i \neq j \qquad (27)$$

is a measure of the correlation between two variables, the 4^{th}-order cross-cumulant

$$\begin{aligned} Cum[X_i, X_j, X_k, X_l] &= E[X_i X_j X_k X_l] - E[X_i X_j]E[X_k X_l] \\ &- E[X_i X_k]E[X_j X_l] - E[X_i X_l]E[X_j X_k], \forall i \neq (j,k,l), \end{aligned}$$

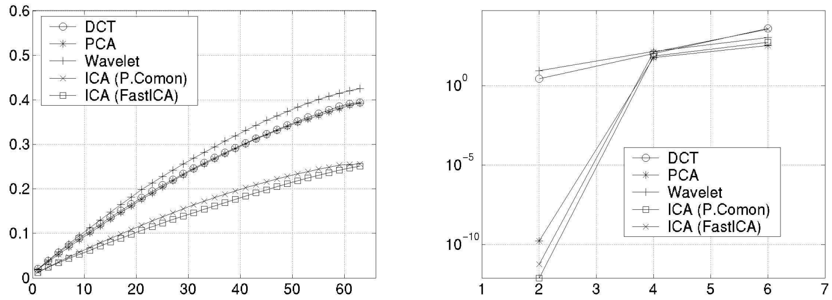

Fig. 3. Independence measures on Brodatz. Left: curve of cumulative average KL divergence $(i, \sum_{j=1}^{i} \hat{KL}(X_j))$. Right: cross-cumulant norm.

can serve both as a measure of 1) linear fourth-order dependence and 2) distance from Gaussianity [10], and higher-order cumulants capture dependencies of higher order. Unfortunately, the number of terms in a cumulant grows exponentially with its order and the computations involved rapidly become infeasible. We computed cumulants up to 6^{th}-order, but omit the formulas. All cumulant information was summarized by the norm of the off-diagonal terms (cross-cumulants), e.g.

$$||Cum_4|| = \sum_{i,j,k,l \neq (c,c,c,c)} Cum^2[X_i, X_j, X_k, X_l] \tag{28}$$

for the fourth-order cumulant. These statistics are zero when independence holds.

5 Results

Figure 3 presents the independence measures obtained on Brodatz. The curves on the left plot represent the cumulative average KL divergence (26) after reordering the X_j such that $\hat{KL}(X_{j+1}) < \hat{KL}(X_j)$. These curves suggest the existence of two groups: the first, consisting of the ICA techniques, achieves significantly better pairwise independence than the second, consisting of the decorrelating transforms. A somewhat different picture starts to emerge from the right plot which shows the evolution of the cumulant norm as a function of its order. While the ICA techniques (together with PCA) achieve the lowest 2^{nd} and 4^{th} order cumulant norms, the slope of the curve (between 4^{th} and 6^{th} order) is larger than that of the wavelet features. This indicates that, for higher orders, the curves are likely to cross, in which case the wavelet representation would be the most independent. This observation is supported by the results that follow and suggests that minimizing cumulants up to a certain order does not really provide any independence guarantees, since the dependencies can simply become of higher-order.

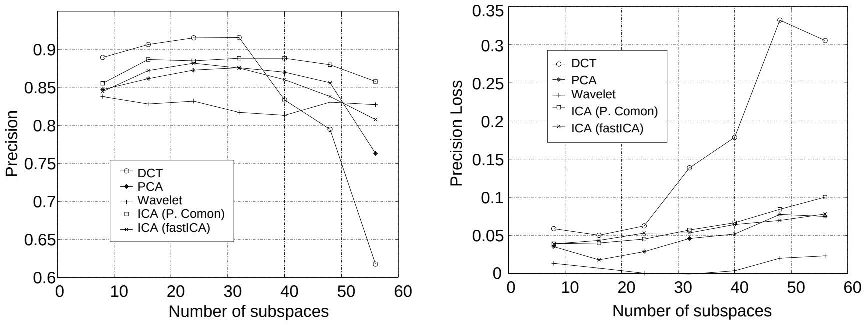

Fig. 4. Recognition results on Brodatz. Left: Precision, at 30% recall, achieved with joint modeling. Right: Precision loss inherent to the independence assumption.

In order to evaluate recognition accuracy we measured precision at various levels of recall[2]. Since the results were qualitatively similar for all levels, we only present curves of precision, as a function of subspace dimension, at 30% recall on Brodatz and 10% recall on Corel. The left plot of Figure 4 shows the precision achieved on Brodatz with joint modeling. The right plot presents the associated precision loss[3] when the joint model is replaced by the product of the marginals. This precision loss is a measure of the dependence between the features, since both models should lead to the same result when independence holds.

Two major conclusions can be taken from the figure. First, the ordering of transformations by degree of independence is quite surprising, with wavelets at the top, followed by PCA, the two ICA methods, and the DCT (as a distant last). While we want to avoid conclusions such as "feature transform X leads to weaker dependencies" that may not generalize to other databases, it is clear that this ordering is very different from that of Figure 3 (ICA techniques on top, then DCT and PCA, and finally wavelets). This can only mean that quantities such as pairwise KL divergence or a limited set of cross-cumulants do not really capture what is going on in terms of independence, at least the aspects that are important for recognition. While this is not completely surprising, since these measures only capture *pairwise* or *linear* dependencies, it clearly indicates that recognition is affected by much more sophisticated patterns of dependence. The logical conclusion is that ICA techniques designed to minimize measures such as those of Figure 3 may not always be of great use for recognition.

Second, the precision curves seem to comply very well with the theoretical arguments of section 2. In particular, they are concave (there is a large increase

---

[2] When the n most similar images to a query are retrieved, recall is the percentage of all relevant images that are contained in that set, and precision the percentage of the n which are relevant.

[3] By precision loss we mean the difference between the precision achieved with the joint and independent models.

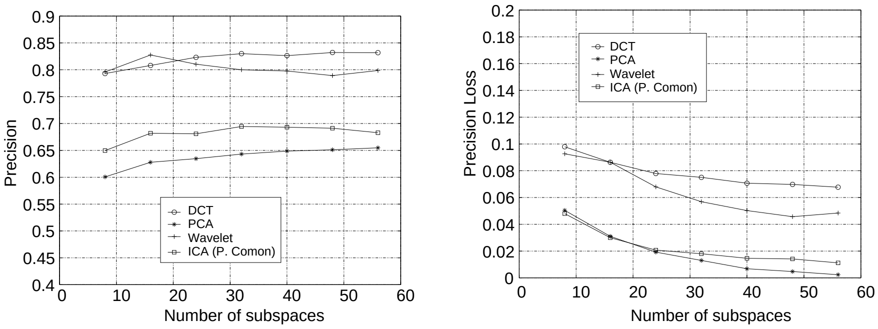

Fig. 5. Recognition results on Corel. Left: Precision, at 10% recall, achieved with joint modeling. Right: Precision loss inherent to the independence assumption.

in precision from 1 to 8 dimensions that we do not show for clarity of the graph), and tend to be flatter when the features are more independent. Remember that compliance with the theory implies that the curves are dominated by the Bayes error for all dimensions when the features are independent, and up to the inflection point when they are not. This is an important observation, since the more independent features (flatter curves) have smaller precision than that achieved at the inflection point of the less independent ones. In fact, a comparison of the two plots reveals significant evidence in support of the conjecture that precision at the inflection point is a monotonic function of the degree of dependence of the features! The natural conclusion is then that independence has a non-negligible cost in terms of Bayes error. In particular, the precision achieved with the most independent features (wavelet coefficients) is almost 10% bellow the peak precision achieved with the less independent ones (DCT).

This conclusion is also supported by Figure 5, which presents recognition results on Corel. Since this is a larger database and contains colored images, 192-dimensional feature space, the queries take significantly longer to compute. For this reason, we restricted the analysis to the first 64 dimensions (and only considered one of the ICA techniques) which are probably not enough to reach the inflection point in all cases. Nevertheless, one can still confidently say that the precision of the more independent feature transforms is roughly 10% lower than the peak precision of the less independent transforms. The only significant difference with respect to the results obtained on Brodatz is that ICA does appear to produce features which are very close to independent, while the wavelet coefficients are not independent.

References

1. A. Bell and T. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3328, December 1997.

2. J. Bergen and E. Adelson. Early Vision and Texture Perception. *Nature*, 333(6171):363–364, 1988.
3. J. Bergen and M. Landy. Computational Modeling of Visual Texture Segregation. In M. Landy and J. Movshon, editors, *Computational Models of Visual Processing*. MIT Press, 1991.
4. J. Cardoso. Blind Signal Separation: Statistical Principles. *Proceedings of the IEEE*, 90(8):2009–20026, October 1998.
5. P. Comon. Independent Component Analysis, A New concept? *Signal Processing*, 36:287–314, 1994.
6. L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.
7. D. Field. What is the goal of sensory coding? *Neural Computation*, 6(4):559–601, January 1989.
8. I. Fogel and D. Sagi. Gabor Filters as Texture Discriminators. *Biol. Cybern.*, 61:103–113, 1989.
9. D. Hubel and T. Wiesel. Brain Mechanisms of Vision. *Scientific American*, September 1979.
10. A. Hyvarinen and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13:411–430, 2000.
11. N. Jayant and P. Noll. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, 1984.
12. D. Knill and W. Richards. *Perception as Bayesian Inference*. Cambridge Univ. Press, 1996.
13. J. Malik and P. Perona. Preattentive Texture Discrimination with Early Vision Mechanisms. *Journal of the Optical Society of America*, 7(5):923–932, May 1990.
14. S. Mallat. A Theory for Multiresolution Signal Decomposition: the Wavelet Representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 11:674–693, July 1989.
15. B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.
16. M. Porat and Y. Zeevi. Localized Texture Processing in Vision: Analysis and Synthesis in the Gaborian Space. *IEEE Trans. on Biomedical Engineering*, 36(1):115–129, January 1989.
17. J. Portilla and E. Simoncelli. Texture Modeling and Synthesis using Joint Statistics of Complex Wavelet Coefficients. In *IEEE Workshop on Statistical and Computational Theories of Vision, Fort Collins, Colorado*, 1999.
18. D. Sagi. The Psychophysics of Texture Segmentation. In T. Papathomas, editor, *Early Vision and Beyond*, chapter 7. MIT Press, 1996.
19. A. Sutter, J. Beck, and N. Graham. Contrast and Spatial Variables in Texture Segregation: testing a simple spatial-frequency channels model. *Perceptual Psychophysics*, 46:312–332, 1989.
20. N. Vasconcelos. *Bayesian Models for Visual Information Retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.
21. N. Vasconcelos and A. Lippman. A Probabilistic Architecture for Content-based Image Retrieval. In *Proc. IEEE Computer Vision and Pattern Recognition Conf.*, Hilton Head, North Carolina, 2000.