Chapter 4

Feature transformation

In addition to making estimation easier, there are a few reasons why dimensionality reduction is a good idea in the context of evaluating image similarity. While, as seen in section 3.2.3, it is important to allow $\mathcal{Z}$ to be high-dimensional, it is also common for the interesting image structure to lie on a lower dimensional manifold [115, 179, 161]. The role of a feature transformation is to expose this manifold, allowing everything else to be discarded. If done right, this can be helpful in various ways.

First, if there is noise associated with the image capture process and the noise is uncorrelated with the image, the signal-to-noise ratio of the representation is usually improved. This happens because most of the noise energy tends to be in the dimensions that are eliminated, while most of the signal energy is in those that are retained. Second, the invariance of the representation to image transformations tends to improve since large regions of the original space are mapped into the same point in the manifold. Finally, because all points in the manifold provide a valid interpretation of the underlying scene, the low-dimensional projection can lead to judgments of similarity that are perceptually more relevant (i.e. intuitive for humans) than what is possible in the high-dimensional space.

For these reasons, most of the feature spaces used for image retrieval involve some form of dimensionality reduction. The interesting question is how to discard dimensions in a way that compromises as little as possible the achievable Bayes error. In this chapter, we argue that multi-resolution feature transformations that are prevalent in the compression

literature also have very good properties for retrieval.

4.1 Terms and notation

A *linear transform* is a map

$$A : R^n \to R^n$$

$$\mathbf{z} \mapsto \mathbf{A}\mathbf{z}$$

where $\mathbf{A} \in R^{n \times n}$. In the context of this thesis, $\mathbf{z} \in \mathcal{Z} \subseteq R^n$. The row vectors $\mathbf{a_i}$ of $\mathbf{A}$ are known as the *basis functions* of the transform. Since

$$x_i = \mathbf{a}_i \mathbf{z} = \sum_{j=1}^{n} a_{ij}\, z_j,\ i = 1, \ldots, n, \tag{4.1}$$

the components of the transformed vector $\mathbf{x}$ (known as the *transform coefficients*) are nothing more than the projections of $\mathbf{z}$ into these basis functions. When the basis vectors satisfy

$$\mathbf{a}_i^T \mathbf{a}_j = \delta_{i,j}, \tag{4.2}$$

where $\delta_{i,j}$ is the Kronecker delta function (2.3), the transform is said to be *orthonormal*. Orthonormality is a desirable property because the inverse of an orthonormal transformation is very simple to compute. This follows directly from (4.2), since

$$\mathbf{A}\mathbf{A}^T = \mathbf{I},$$

where $\mathbf{I}$ is the identity, and thus

$$\mathbf{A}^{-1} = \mathbf{A}^T,$$

i.e. a *unitary* matrix. The *linear projection* of R^n into R^k, for $k < n$, is the map

$$\pi_k : R^n \to R^k$$

$$\mathbf{x} \mapsto \mathbf{\Pi}_k \mathbf{x} \tag{4.3}$$

where $\mathbf{\Pi}_k = [\mathbf{I}_k\ \mathbf{0}_{n-k}]$, $\mathbf{I}_k$ is the $k \times k$ identity matrix and $\mathbf{0}_{n-k}$ a $k \times (n-k)$ matrix of zeros. The *embedding* of R^k into R^n is the map

$$\rho_k : R^k \to R^n$$

$$\mathbf{x} \mapsto \mathbf{\Pi}_k^T \mathbf{x}. \tag{4.4}$$

4.2 Previous approaches

A popular strategy in the retrieval literature for selecting a feature transformation is to simply decide what image properties are important (e.g. textureness, color, edginess, shape, etc.) and define an arbitrary set of features to capture these properties [118, 32, 125, 72, 164, 65, 153, 173, 195, 43, 80]. Other times, the features are selected in a more principled way but predominantly on the basis of a few of the above requirements, e.g. invariance [158, 156, 103, 49] or perceptual relevance [174, 94, 102, 175, 15, 57, 7]. Finally, many times the transformation is selected to provide good discrimination on a specific domain, e.g. texture, object, or faces databases [176, 19, 104, 179, 115, 129, 112].

All these strategies have flaws that are relevant in the context of CBIR. In the first case, it is difficult to know how important is the information that was left out and why other features would not perform better than the ones selected. The answer to this question can only be obtained through extensive experimental evaluation, but so far few exhaustive studies have been conducted [134, 96]. In the second case, it is usually unclear how the selected features perform under the requirements that were not considered for their selection. For example, a representation that has good invariance properties for the smooth surfaces that characterize most object databases may be discarding information that is crucial to characterize texture. Or a representation that captures perceptually relevant attributes for the characterization of texture may be discarding information that is crucial for the perception of faces. In the third case, the resulting features do not even make sense outside the domains for which they were developed.

One solution to these problems is to assemble different feature sets optimized for different criteria and different domains, and build a "society of models"[131, 110, 150]. The combination of multiple models has indeed become prevalent for the design of retrieval systems that can account for both color and texture [43, 118, 44, 164, 153, 175, 32, 101, 72]. However, it has serious drawbacks. First, it implies a significant increase in retrieval complexity since similarity has to be evaluated according to all the models. Second, it is usually not clear how to combine the different representations in order to achieve global inferences. In practice, it frequently requires users to specify weights for the different image attributes, a process that can be extremely non-intuitive. We take the alternate route of modeling

the joint density of the image observations over a spatial neighborhood exactly because it avoids these problems.

4.3 Minimizing the reconstruction error

We have already seen that such modeling requires a feature transformation that, for a given level of dimensionality reduction, is as close to invertible as possible. This, in turn, requires a precise definition of "as close to invertible as possible." Following a long tradition in image compression [74, 56, 29], we rely on *linear transformations* and use the minimization of the mean squared reconstruction error as a fidelity criterion.

Definition 2 *A feature transformation T_k provides dimensionality reduction of level $n-k$ if T_k is a map*

$$T_k : R^n \to R^k$$

defined by

$$T_k = \pi_k \circ A, \tag{4.5}$$

where A is an invertible linear transform.

Definition 3 *The mean squared reconstruction error for a feature transformation T_k defined by (4.5) is*

$$\mathcal{E}_k(\mathbf{z}) = E\left[||\mathbf{z} - (\mathbf{A}^{-1} \circ \rho_k \circ T_k(\mathbf{z}))||^2\right], \tag{4.6}$$

where $||\mathbf{z}||$ is the Euclidean norm of $\mathbf{z}$.

It is well known that this error is minimized by *principal component analysis* (PCA), also known as the *Karhunen-Loeve transform* (KLT), a dimensionality reduction technique that has found wide application in the vision and compression literatures [179, 115, 170, 129, 74, 29, 185]. It consists of finding the eigenvectors $\mathbf{e}_i$ and the eigenvalues λ_i, $i = 1, \ldots, n$, of the sample covariance of $\mathbf{z}$, i.e. the solution to

$$\hat{\mathbf{\Sigma}}_{\mathbf{z}}\mathbf{e}_i = \lambda_i \mathbf{e}_i,\ i = 1, \ldots, n \tag{4.7}$$

where

$$\hat{\Sigma}_{\mathbf{z}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \hat{\mu}_{\mathbf{z}})(\mathbf{z}_i - \hat{\mu}_{\mathbf{z}})^T,$$

$$\hat{\mu}_{\mathbf{z}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i,$$

and $\{\mathbf{z}_i\}_1^N$ a sample of observations from $\mathbf{z}$, projecting $\mathbf{z}$ onto the eigenvalue basis and discarding the dimensions associated with the smallest eigenvalues. If the eigenvalues are sorted in decreasing order

$$\lambda_1 > \ldots > \lambda_n,$$

this leads to

$$T_k^*(\mathbf{z}) = \pi_k(\mathbf{S}\mathbf{z}) \tag{4.8}$$

where $\mathbf{S} = [\mathbf{e}_1, \ldots, \mathbf{e}_n]^T$ and T_k^* is the optimal feature transformation for a dimensionality reduction of level $n - k$.

In addition to providing optimal dimensionality reduction, PCA has the advantage of generating *decorrelated* coefficients. Notice that, if $\mathbf{x}_i = \mathbf{S}\mathbf{z}_i$, then

$$\hat{\mu}_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i = \mathbf{S}\,\hat{\mu}_{\mathbf{z}}, \tag{4.9}$$

$$\hat{\Sigma}_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{S}(\mathbf{z}_i - \hat{\mu}_{\mathbf{z}})(\mathbf{z}_i - \hat{\mu}_{\mathbf{z}})^T\mathbf{S}^T = \mathbf{S}\hat{\Sigma}_{\mathbf{z}}\mathbf{S}^T, \tag{4.10}$$

and, from (4.7), $\hat{\Sigma}_{\mathbf{x}} = diag(\lambda_1, \ldots, \lambda_n)$. In practice, this means a reduction in the complexity of the subsequent density estimation that is larger than the simple dimensionality reduction from n to k. If, for example, a Gaussian model is used, decorrelation is equivalent to diagonal instead of full covariance matrices. This means that there will be k covariance parameters to estimate in $\mathcal{X}$, as opposed to n^2 parameters in $\mathcal{Z}$.

4.4 Discrete Cosine Transform

For simple statistical image models commonly used to evaluate the decorrelating abilities of a feature transformation, such as the first-order Gauss-Markov model, PCA is well approximated by an alternative transformation of lower implementation complexity: the *discrete*

cosine transform (DCT). This approximation is particularly good for signals that, like most of the images that we are interested in, exhibit strong pixel correlations [74, 73, 1, 71].

Several definitions of the DCT have been presented in the literature. In this thesis, we will use the one given in [74].

Definition 4 *The 1-D DCT is an orthonormal transform*

$$T : R^n \rightarrow R^n,$$

described by

$$[T(\mathbf{z})]_k = \sqrt{\frac{2}{n}}\alpha_k \sum_{i=0}^{n-1} z_i \cos \frac{(2i+1)k\pi}{2n}, \; k = 0, \ldots, n-1, \tag{4.11}$$

where

$$\alpha_k = \begin{cases} 1/\sqrt{2}, & \text{if } k = 0 \\ 1, & \text{if } k \neq 0. \end{cases} \tag{4.12}$$

By writing (4.11) in matrix form, it can easily be seen that the basis functions of the DCT are

$$\mathbf{t}_k = \sqrt{\frac{2}{N}}\alpha_k \cos \frac{(2i+1)k\pi}{2N}, \; i = 0, \ldots, N-1. \tag{4.13}$$

It is clear that the DCT performs a decomposition of the input signal into a sum of cosines of increasing frequency. In particular, the value of coefficient $T_0(\mathbf{z})$ is the mean or DC value of the input vector and commonly known as the *DC coefficient*. The remaining coefficients, associated with basis vectors of zero-mean, are known as *AC coefficients*. The extension of the 1-D DCT to the 2-D DCT is straightforward: the 2-D DCT is obtained by the separable application of the 1-D DCT to the rows and to the columns of the input image. In this case, the n basis vectors (4.13) are extended to a set of n^2 2-D basis functions that can be obtained by performing the outer products between all the 1-D basis vectors [29]. Figure 4.1 presents these basis functions for $n = 8$.

4.5 Perceptual relevance

Few universal results are known, at this point, about the mechanisms used by the human brain to evaluate image similarity. Ever since the seminal work of Hubel and Wiesel [67],

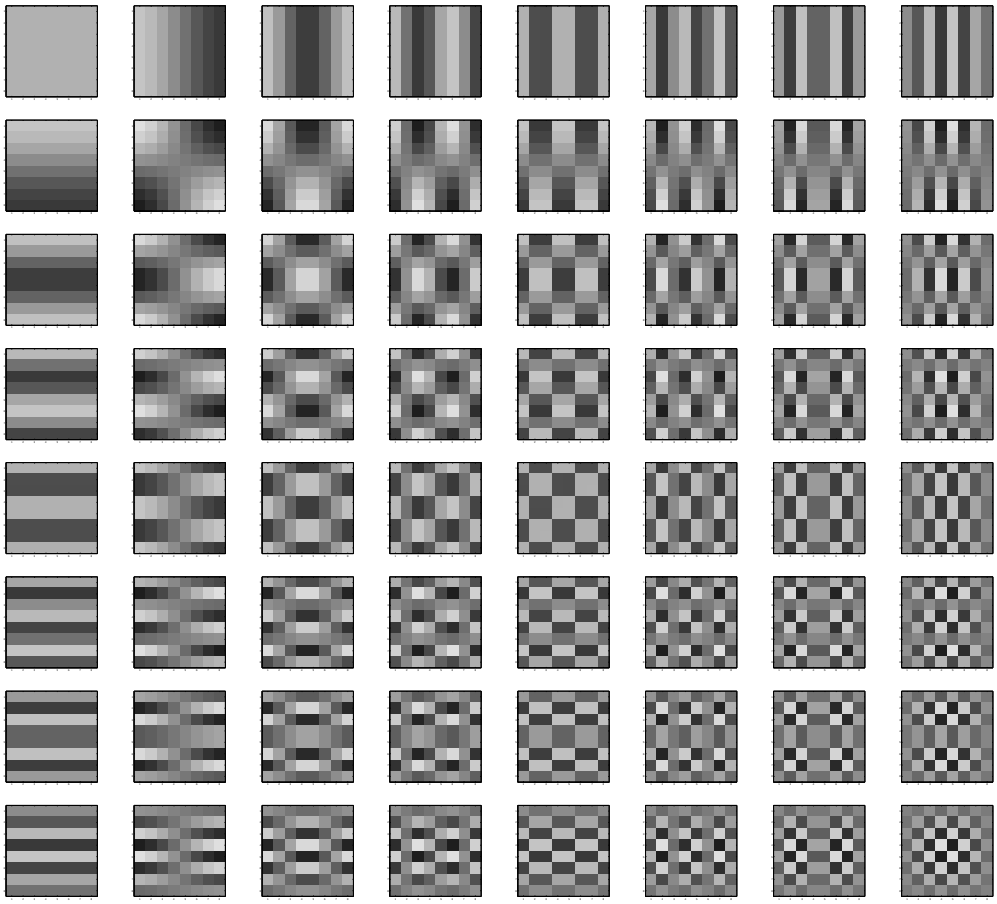

Figure 4.1: Basis Functions of the 2-D DCT of dimension 8.

it has been established that 1) 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 time, and several "biologically plausible" models of early vision are based on this principle [152, 98, 8, 46, 171, 9]. 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.

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. Several elementary filters have been proposed, including *differences of Gaussians* [98], *Gabor functions* [137, 46], and *differences of offset Gaussians* [98]. While Gabor functions seem to provide a better match to actual measurements from the visual cortex [35], the Gabor representation is also the most complex to design [34]. However, there seems to be some agreement in that the exact shape of the filters is not crucial, as long as the representation is localized in space and frequency.

With respect to the non-linearity, at least three different types have been proposed. Denoting by *channel* the output of each filter, these involve intra-channel processing only. If $g(x)$ is the nonlinearity, possible functions are *energy* ($g(x) = x^2$) [9, 46], *full-wave rectification* ($g(x) = |x|$) [8, 171, 46], and *half-wave rectification* ($g(x) = (\max(x, 0), -\min(x, 0))$) [98]. Finally, there is small agreement on the implementation of the final stage other than that it should involve pooling from the individual channel responses.

While biological plausibility is not a constraint for our representation, it is important that it can capture the fundamental characteristics of human visual processing since this is likely to lead to *perceptually* more relevant similarity judgments. The use of the DCT as feature transformation satisfies these requirements because, as illustrated by Figure 4.1, the DCT is localized in both space and frequency. The linear projection to achieve dimensionality reduction is even biologically plausible, since it simply consists of eliminating the filters associated with the frequency/orientation channels to be disregarded.

On the other hand, under the goal of preserving Bayes error, it makes little sense to

include a non-invertible linearity, like energy or full-wave rectification, in the model. Malik and Perona have shown that such non-linearities are also not likely to be implemented by the human visual system (by constructing examples of texture pairs where the sign of filter responses is the only property that allows their discrimination by humans) [98]. They suggest half-wave rectification which, being invertible, presents no evidence against our principles. In this case, the need for a non-linearity is simply a consequence of the implementation constraints of neural hardware. This point is important because several authors have argued for the use of the average channel energy as a feature for texture retrieval [102, 163, 40, 24, 137, 15]. Both the Bayesian principles and psychophysical evidence indicate that this is a bad idea. Finally, a third stage of pooling different channels may be consistent with a representation based on density estimates of the feature measurements. Not enough is known at this point to argue for or against this position.

A few perceptually based models of higher level have also been proposed in the literature [94, 174, 15, 155]. These, however, tend to be restricted to specific domains such as texture or color perception. Since there are no universal strategies for the experimental validation of the predictions made by these models, it is difficult to reach definitive conclusions about their strengths and weaknesses. In any case, these models tend to emphasize a decomposition into properties like randomness, periodicity, scale, and orientation that are all easily extracted from a representation localized in space and frequency.

More concrete evidence for the benefits of multi-resolution representations is that provided by decades of experience in image compression, where frequency decompositions are universally accepted as a good pre-processing stage to compression [74, 128, 63, 29]. The observation that discontinuities in the low-frequency components of an image (*blocking artifacts*) are much more noticeable than similar discontinuities in their high frequency counterparts indicates that low frequency information is perceptually more relevant than that in the high-frequencies. The facts that PCA is well approximated by the DCT for many natural images and the DCT is better matched to human perception are indeed the fundamental reasons for the widespread use of the DCT in image compression.

Finally, there is a long history of machine vision problems where multi-resolution representations are known to lead to the best solutions [20, 105, 22, 95, 4, 187] and some

striking recent advances in image synthesis, based on the statistical analysis of such representations, reinforce the idea that they are central to perceptually meaningful image modeling [64, 138, 135, 14, 151].

For all these reasons, multi-resolution spaces are natural candidates for CBIR from both the perceptual and Bayes error points of view. While we rely on the DCT, it should be pointed out that Bayesian retrieval is not tied to this particular transform. In fact, any other multi-resolution feature transformation could be employed, including wavelets [99, 100], Laplacian [20], or Gabor [137, 102] pyramids. Because most images are compressed using the DCT [128, 63], the DCT has the practical advantages of compatibility with a wide installed base of image processing hardware. This is the fundamental reason that motivated us to select it.

4.6 Experimental evaluation

In this section we present experimental results on the performance of the DCT features. Since the discussion on image representation will only be complete in the next chapter we postpone a detailed evaluation until then. Here, we simply want to dispel the common belief that the DCT coefficients are not a good feature set for texture recognition [96, 134, 163]. Figure 4.2 presents precision/recall curves, on the Brodatz database, for retrieval based on both the DCT and the MRSAR features. The DCT features were obtained by sliding an 8×8 window by increments of two pixels over each image to be processed. The implementation of MRSAR is that of [94]. Each feature transformation is combined with two similarity functions: the Mahalanobis distance, which is the standard in the texture literature, and the ML criteria, in a total of 4 image representations.

The figure confirms that, under MD, the performance of the DCT features is indeed terrible: precision is never better than 25%. However, a very different result is obtained when the similarity function is ML, in which case precision improves by up to 65% points! Hence, while the DCT is significantly worse than MRSAR under MD - a difference in precision that can be as high as 70% - it becomes competitive under ML - difference usually below 5%. Notice that, *when combined with ML, the DCT features even outperform the*

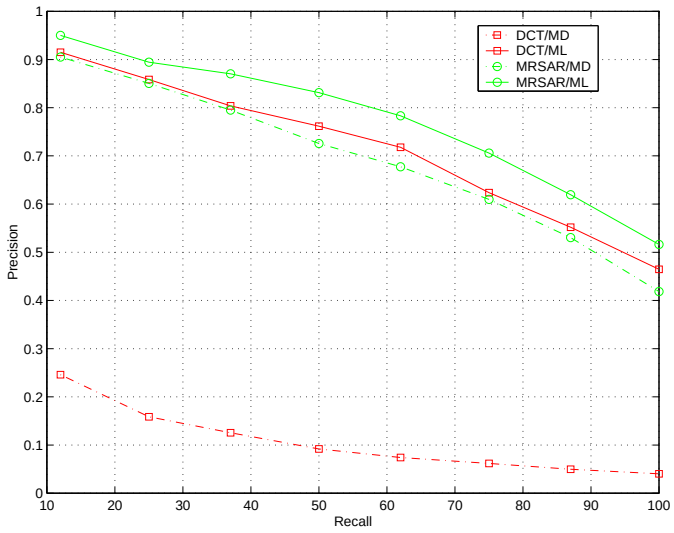

Figure 4.2: Precision/recall curves for Brodatz. In the legend, MRSAR means MRSAR features, DCT the DCT features, ML maximum likelihood, and MD the Mahalanobis distance.

standard MRSAR/MD combination.

Figure 4.3 provides an explanation for these observations. In the figure we present equi-probability contours for the best Gaussian fits to the features extracted from ten texture classes in the Brodatz database. In both plots, we show the joint density of the first two coefficients other than the pixel mean.

Since the DCT is an orthonormal transform, it preserves in $\mathcal{X}$ the shape of the densities in $\mathcal{Z}$. Hence, it is not surprising that there is a significant amount of overlap between the densities of different classes. On the other hand, following strategy S2 of Chapter 3, MRSAR tries to separate these densities as much as possible. Small overlap is a very important requirement under the MD since, as discussed in section 2.3.5, this metric does not consider the full query density, but only its mean. Consequently, because the DCT features have zero mean for all classes, retrieval is very error-prone when they are combined with MD. On the other hand, since ML can account for the entire query density, it has enough information to distinguish the different classes, even when the DCT features are used. It is therefore not surprising that the performance of the DCT improves so dramatically.

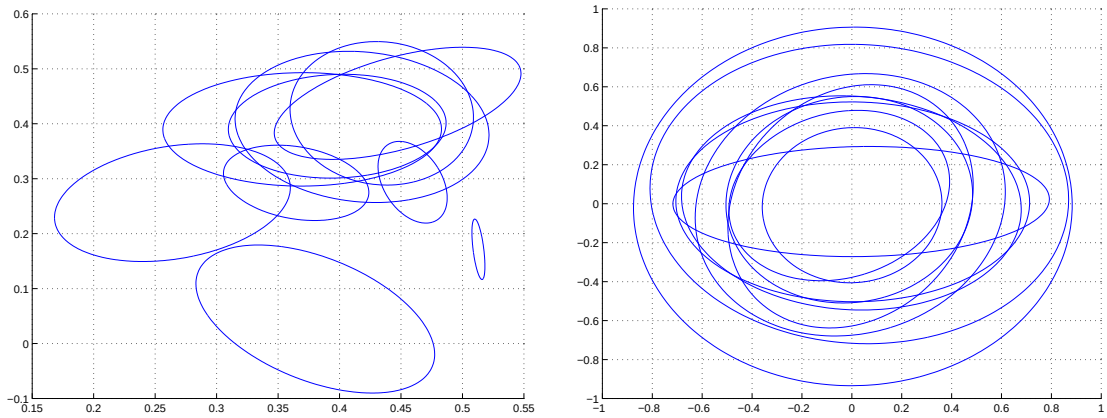

Figure 4.3: Gaussian fits (contours where probability drops to 65% of the maximum) to the features from ten texture classes from Brodatz. Only the first two components (other than the mean coefficient) of the feature space are shown. Left: MRSAR, right: DCT.

In summary, and contrary to prior beliefs, it is not the DCT coefficients that are a bad feature set for texture retrieval. Instead, the problem relies on the use of the Mahalanobis distance as a similarity criteria. The significance of this result is that, because the DCT features are generic, there is potential to design a unified representation for color and texture capable of doing well on generic databases. This is not possible with transformations that are highly specialized for texture, such as MRSAR.