Chapter 3

Image representations for retrieval

Numerous image representations have been proposed for image compression [74, 56, 128, 63, 73], object recognition [180, 62], texture analysis [144, 178, 152, 62, 61, 21] and, more recently, content-based retrieval [149, 2, 133, 132, 131]. Because we are interested in generic imagery (i.e. we want to make as few assumptions as possible regarding the content of the images under analysis) and it is still too difficult to segment such images in a semantically meaningful way, we will not consider here any representations that require segmentation either implicitly or explicitly. This includes many representations that are common in vision [115, 179, 77, 130, 194] and most of the ones used for shape-based retrieval [149].

In order to simplify the understanding of the remaining representations, it is useful to further decompose them into the two main components discussed in section 2.1: a feature transformation and a feature representation. In this chapter, we show that minimization of the probability of error, and the resulting Bayesian solution to the retrieval problem, provide us with concrete guidelines for the selection of feature spaces and representations. Interpretation of the strategies in current use according to these guidelines leads to insights about their major limitations and lays the ground for a better solution, that we will pursue in subsequent chapters.

3.1 Bayesian guidelines for image representation

In Chapter 2, we saw that one of the interesting properties of Bayesian retrieval is that it is optimal with respect to the minimization of error probability. In practice, however, good results can only be guaranteed if it is possible to achieve a probability of error close to the Bayes error. In this section, we look for theoretical guidelines that can help us achieve this goal.

3.1.1 Feature transformation

We start by analyzing the impact of a feature transformation on the overall probability of error.

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

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

the Bayes error on $\mathcal{X}$ can never be smaller than that on $\mathcal{Z}$. I.e.,

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

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.

Proof: The following proof is a straightforward extension to multiple classes of the one given in [38] for the two-class problem. From (2.10),

$$\begin{aligned}
L^*_{\mathcal{X}} &= 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})], \\
&= 1 - E_{T(\mathbf{z})}[\max_i P_{Y|\mathbf{X}}(i|T(\mathbf{z}))], \\
&= 1 - E_{T(\mathbf{z})}[\max_i \int P_{Y|\mathbf{Z},\mathbf{X}}(i|\mathbf{z}, T(\mathbf{z}))P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|T(\mathbf{z}))d\mathbf{z}], \\
&= 1 - E_{T(\mathbf{z})}[\max_i \int P_{Y|\mathbf{Z}}(i|\mathbf{z})P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|T(\mathbf{z}))d\mathbf{z}], \\
&= 1 - E_{T(\mathbf{z})}[\max_i E_{\mathbf{z}|\mathbf{X}}[P_{Y|\mathbf{Z}}(i|\mathbf{z})|\mathbf{X} = T(\mathbf{z})]], \\
&\geq 1 - E_{T(\mathbf{z})}[E_{\mathbf{z}|\mathbf{X}} \max_i [P_{Y|\mathbf{Z}}(i|\mathbf{z})|\mathbf{X} = T(\mathbf{z})]], \\
&= 1 - E_{\mathbf{z}}[\max_i P_{Y|\mathbf{Z}}(i|\mathbf{z})] = L^*_{\mathcal{Z}},
\end{aligned}$$

where we have used Jensen's inequality [31], and equality is achieved if and only if T is an invertible map.□

This theorem tells us that the choice of feature transformation is very relevant for the performance of a retrieval system. In particular, 1) any transformation can only increase or, at best, maintain the Bayes error achievable in the space of image observations, and 2) the only transformations that maintain the Bayes error are the invertible ones.

3.1.2 Feature representation

While a necessary condition, low Bayes error is not sufficient for accurate retrieval since the actual error may be much larger than the lower bound. The next theorem provides an upper bound for this difference.

Theorem 3 *Given a retrieval system 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*

$$g(\mathbf{x}) = \arg\max_i \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i), \tag{3.1}$$

the actual probability of error is upper bounded by

$$P(g(\mathbf{X}) \neq Y) \leq L^*_{\mathcal{X}} + \sum_i \int |P_{\mathbf{X}|Y}(\mathbf{x}|i)P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i)|d\mathbf{x}. \tag{3.2}$$

Proof: From (2.11),

$$P_{\mathbf{X},Y}(g(\mathbf{X}) \neq Y) - L^*_{\mathcal{X}} = \int [P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) - P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x})]P_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \tag{3.3}$$

and since, $\forall \mathbf{x} \in \mathcal{X}$ such that $g(\mathbf{x}) = g^*(\mathbf{x})$, we have

$$P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) = P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x}),$$

this is equivalent to

$$P_{\mathbf{X},Y}(g(\mathbf{X}) \neq Y) - L^*_{\mathcal{X}} = \int_E [P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) - P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x})]P_{\mathbf{X}}(\mathbf{x})d\mathbf{x}, \tag{3.4}$$

where

$$E = \{\mathbf{x}|\mathbf{x} \in \mathcal{X}, P_{\mathbf{X}}(\mathbf{x}) > 0, g(\mathbf{x}) \neq g^*(\mathbf{x})\}.$$

Letting

$$\Delta(\mathbf{x}) = P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) - P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x})$$

and defining the sets

$$E_i^* = \{\mathbf{x}|\mathbf{x} \in E, g^*(\mathbf{x}) = i\}$$

$$E_i = \{\mathbf{x}|\mathbf{x} \in E, g(\mathbf{x}) = i\},$$

it follows from (2.12) that, $\forall \mathbf{x} \in E_i^* \cap E_j$,

$$\Delta(\mathbf{x}) = P_{Y|\mathbf{X}}(i|\mathbf{x}) - P_{Y|\mathbf{X}}(j|\mathbf{x}).$$

Since, from (2.9),

$$P_{Y|\mathbf{X}}(i|\mathbf{x}) - P_{Y|\mathbf{X}}(j|\mathbf{x}) \geq 0 \ \forall \mathbf{x} \in E_i^*, \forall j \neq i$$

from (3.1) and the fact that $P_{\mathbf{X}}(\mathbf{x}) > 0 \, \forall \mathbf{x} \in E$,

$$\frac{\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|j)\hat{p}_Y(j)}{P_{\mathbf{X}}(\mathbf{x})} - \frac{\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i)}{P_{\mathbf{X}}(\mathbf{x})} \geq 0 \ \forall \mathbf{x} \in E_j, \forall i \neq j,$$

defining

$$\hat{p}_{Y|\mathbf{X}}(i|\mathbf{x}) = \frac{\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i)}{P_{\mathbf{X}}(\mathbf{x})},$$

we have, $\forall \mathbf{x} \in E_i^* \cap E_j$,

$$\begin{aligned}
\Delta(\mathbf{x}) &= P_{Y|\mathbf{X}}(i|\mathbf{x}) - P_{Y|\mathbf{X}}(j|\mathbf{x}) \\
&\leq P_{Y|\mathbf{X}}(i|\mathbf{x}) - P_{Y|\mathbf{X}}(j|\mathbf{x}) + \hat{p}_{Y|\mathbf{X}}(j|\mathbf{x}) - \hat{p}_{Y|\mathbf{X}}(i|\mathbf{x}) \\
&= |P_{Y|\mathbf{X}}(i|\mathbf{x}) - P_{Y|\mathbf{X}}(j|\mathbf{x}) + \hat{p}_{Y|\mathbf{X}}(j|\mathbf{x}) - \hat{p}_{Y|\mathbf{X}}(i|\mathbf{x})| \\
&\leq |P_{Y|\mathbf{X}}(i|\mathbf{x}) - \hat{p}_{Y|\mathbf{X}}(i|\mathbf{x})| + |P_{Y|\mathbf{X}}(j|\mathbf{x}) - \hat{p}_{Y|\mathbf{X}}(j|\mathbf{x})|
\end{aligned}$$

and

$$\begin{aligned}
\int_{E_i^* \cap E_j} \Delta(\mathbf{x}) P_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} &\leq \int_{E_i^* \cap E_j} |P_{\mathbf{X}|Y}(\mathbf{x}|i)P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i)| d\mathbf{x} \\
&+ \int_{E_i^* \cap E_j} |P_{\mathbf{X}|Y}(\mathbf{x}|j)P_Y(j) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|j)\hat{p}_Y(j)| d\mathbf{x}.
\end{aligned}$$

Using the fact that both collections of sets E_i^* and E_j partition E, we obtain

$$\begin{aligned}
\int_E \Delta(\mathbf{x}) P_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} &= \sum_{i,j} \int_{E_i^* \cap E_j} \Delta(\mathbf{x}) P_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&\leq \sum_i \int_{E_i^*} |P_{\mathbf{X}|Y}(\mathbf{x}|i) P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i) \hat{p}_Y(i)| d\mathbf{x} + \\
&\quad \sum_j \int_{E_j} |P_{\mathbf{X}|Y}(\mathbf{x}|j) P_Y(j) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|j) \hat{p}_Y(j)| d\mathbf{x} \\
&= \sum_i \left[\int_{E_i^*} |P_{\mathbf{X}|Y}(\mathbf{x}|i) P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i) \hat{p}_Y(i)| d\mathbf{x} \right. \\
&\quad + \left. \int_{E_i} |P_{\mathbf{X}|Y}(\mathbf{x}|i) P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i) \hat{p}_Y(i)| d\mathbf{x} \right] \\
&\leq \sum_i \int |P_{\mathbf{X}|Y}(\mathbf{x}|i) P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i) \hat{p}_Y(i)| d\mathbf{x}
\end{aligned}$$

where we have also used the fact that $E_i^* \cap E_i = \emptyset$. □

This theorem states that, if the Bayes error is small, accurate density estimation is a sufficient condition for high retrieval accuracy. In particular, good density estimation will suffice to guarantee optimal performance when the feature transformation is the identity.

3.2 Strategies for image representation

Together the two theorems are a convenient tool to analyze the balance between feature transformation and representation achieved by any retrieval strategy. We now proceed to do so for the two predominant strategies in the literature.

3.2.1 The color strategy

The theorems suggest that all that really matters for accurate retrieval is good density estimation. Since no feature transformation can reduce the Bayes error, there seems to be no advantage in using one. This is the rationale behind Strategy 1 (S1): *avoid feature transformations altogether and do all the estimation directly in $\mathcal{Z}$.* As Figure 3.1 illustrates, the main problem with this strategy is that density estimation can be difficult in $\mathcal{Z}$. Significant emphasis must therefore be given to the feature representation which is required to rely on a sophisticated density model. One possible solution, that has indeed become a de-facto standard for color-based retrieval [172, 139, 149, 96, 72, 150, 163, 164, 125, 68, 44, 167, 43, 168, 17], is the histogram. This solution is illustrated in Figure 3.1 b).

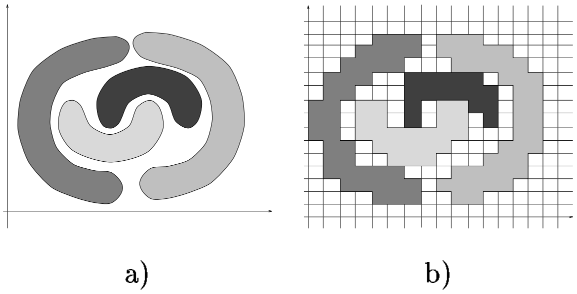


Figure 3.1: Example of a retrieval problem with four image classes. a) In the space of image observations, the class densities can have complicated shapes. b) Strategy 1 is to simply model the class densities as accurately as possible.

3.2.2 The texture strategy

Since accurate density estimation is usually a difficult problem [184, 162, 39], a feature transformation can be helpful if it makes estimation significantly easier in $\mathcal{X}$ than what it is in $\mathcal{Z}$. The rationale behind Strategy 2 (S2) is to exploit this as much as possible: *find a feature transformation that clearly separates the image classes in* $\mathcal{X}$*, rendering estimation trivial.* Ideally, in $\mathcal{X}$, each class should be characterized by a simple parametric density, such as the Gaussians in Figure 3.2, and a simple classifier should be able to guarantee performance close to the Bayes error.

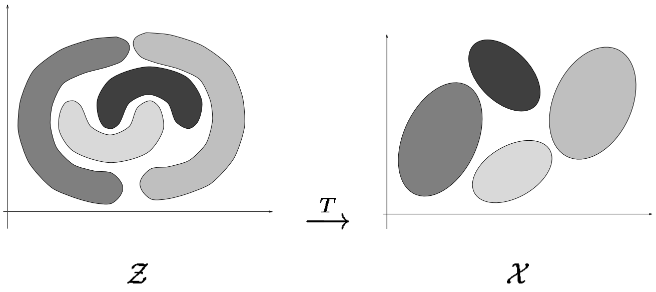


Figure 3.2: Example retrieval problem with four image classes. Strategy 2 is to find a feature transformation such that density estimation is much easier in $\mathcal{X}$ than in $\mathcal{Z}$.

Strategy S2 has become prevalent in the texture literature, where numerous feature transformations have been proposed to achieve *good discrimination* between different texture classes [163, 24, 96, 134, 102, 137, 40, 104, 178, 144, 174, 21, 176]. These transformations are then combined with simple similarity functions, like the Mahalanobis and Euclidean distances or variations of these, that assume Gaussianity in $\mathcal{X}$. More recently it has also been embraced by many retrieval systems [15, 118, 175, 150, 139, 153, 163, 96, 129, 7].

3.2.3 A critical analysis

Overall, none of the two strategies is consistently better than the other. While S1 has worked better for object recognition and color-based retrieval, S2 has proven more effective for the databases used by the texture community. Unfortunately, none of the two strategies is viable when the goal is to jointly model color and texture in the context of generic image databases.

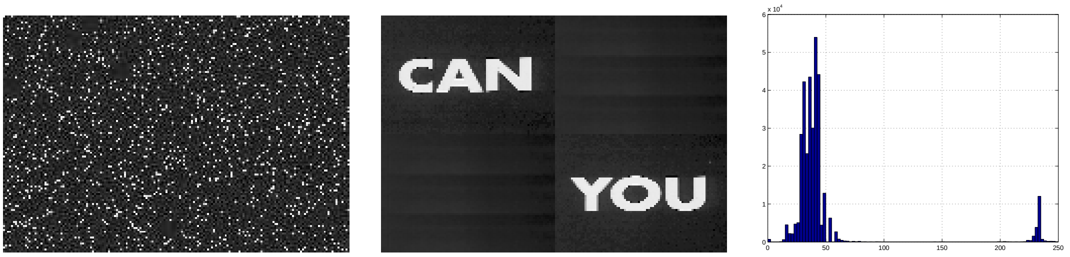
Limitations of strategy S1

While it works reasonably well when $\mathcal{Z}$ is a low-dimensional space, e.g. the 3-D space of pixel colors, S1 is of very limited use in high dimensions. This is a consequence of the well known curse of dimensionality: in higher dimensions, modeling requires more parameters and more data is required to achieve accurate estimation. Typically these relationships are non-linear. For example, the number of elements in the covariance matrix of a Gaussian is quadratic in the dimension of the space, and the number of cells in the histogram model increases exponentially with it[1].

In particular, for c color channels and observations with b pixels, the dimension of $\mathcal{Z}$ is $n = cb$. Hence, the complexity is at least linear and, in the case of the histogram exponential, in the size of the region of support of the observations. Consequently, accurate joint density estimates can only be obtained over very small spatial neighborhoods and the resulting representations cannot capture the spatial dependencies that are crucial for fine image discrimination. This is illustrated by Figure 3.3 where we present two images that,

[1] Assuming that the number of divisions in each coordinate axis is held constant.

although visually very dissimilar, are characterized by the same histogram [shown in c)]. In order to distinguish between these images, the representation must capture the fact that while on b) the white pixels cluster spatially, the same does not happen on a). This is an impossible task if the measurements do not have spatial support, e.g. the pixel colors commonly used under S1.



a) b) c)

Figure 3.3: a) An homogeneous and b) a non-homogeneous image that are visually dissimilar but have the same color histogram, shown in c).

Of course, there is no law stating that histograms cannot be computed in high dimensions, but in practice it is impossible to guarantee that the upper bound of Theorem 3 remains close to the Bayes error.

Limitations of strategy S2

For strategy S2, the main problem is the assumption that it is always possible to find a transformation that maps a collection of complicated densities in $\mathcal{Z}$ into a collection of simple densities in $\mathcal{X}$, without compromising Bayes error. The following theorem shows that, for multi-modal class-conditional densities, this is not possible with a generic feature transformation.

Theorem 4 *Consider a retrieval system with observation space $\mathcal{Z}$. If there exists a feature transformation T*

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

that preserves the Bayes error

$$L^*_{\mathcal{Z}} = L^*_{\mathcal{X}} \tag{3.5}$$

and maps a multi-modal density $P_{\mathbf{Z}|Y}(\mathbf{z}|i)$ on $\mathcal{Z}$ into a unimodal density $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ on $\mathcal{X}$ then 1) T is non-linear, and 2) T depends on $P_{\mathbf{Z}|Y}(\mathbf{z}|i)$.

Proof: From Theorem 2, (3.5) only holds if T is invertible, in which case [114]

$$det\left[J(\mathbf{z})\right] \neq 0 \ \forall \mathbf{z}$$

where $J(\mathbf{z})$ the Jacobian of T evaluated at $\mathbf{z}$

$$J_{i,j}(\mathbf{z}) = [D_{\mathbf{z}}T(\mathbf{z})]_{i,j} = \frac{\partial T_i}{\partial \mathbf{z}_j}(\mathbf{z}) \tag{3.6}$$

and $D_{\mathbf{z}}T(\mathbf{z})$ the vector derivative[2] of $T(\mathbf{z})$ with respect to $\mathbf{z}$. It follows, from the change of variables theorem [124], that the densities in $\mathcal{Z}$ and $\mathcal{X}$ are related by

$$P_{\mathbf{X}|Y}(T(\mathbf{z})|i) = det\left[J^{-1}(\mathbf{z})\right] P_{\mathbf{Z}|Y}(\mathbf{z}|i). \tag{3.7}$$

If T is linear $T(\mathbf{z}) = \mathbf{A}\mathbf{z}$ then $J(\mathbf{z}) = \mathbf{A}$ and, up to a scale factor, the two densities are equal

$$P_{\mathbf{Z}|Y}(T(\mathbf{z})|i) = det\left[\mathbf{A}^{-1}\right] P_{\mathbf{Z}|Y}(\mathbf{z}|i).$$

Hence, if $P_{\mathbf{Z}|Y}(\mathbf{z}|i)$ is multi-modal then so is $P_{\mathbf{X}|Y}(T(\mathbf{z})|i)$. This proves the first part of the theorem. If T is non-linear, by taking derivatives on both sides of (3.7)

$$\begin{aligned} D_{\mathbf{z}}\left[det[J^{-1}(\mathbf{z})]P_{\mathbf{Z}|Y}(\mathbf{z}|i)\right] &= D_{\mathbf{z}}P_{\mathbf{X}|Y}(T(\mathbf{z})|i) \\ &= D_{\mathbf{x}}P_{\mathbf{X}|Y}(\mathbf{x}|i)\Big|_{\mathbf{x}=T(\mathbf{z})} J(\mathbf{z}) \end{aligned}$$

and

$$D_{\mathbf{x}}P_{\mathbf{X}|Y}(\mathbf{x}|i)\Big|_{\mathbf{x}=T(\mathbf{z})} = D_{\mathbf{z}}\left[det[J^{-1}(\mathbf{z})]P_{\mathbf{Z}|Y}(\mathbf{z}|i)\right] J^{-1}(\mathbf{z}), \tag{3.8}$$

from which $D_{\mathbf{x}}P_{\mathbf{X}|Y}(\mathbf{x}|i)|_{\mathbf{x}=T(\mathbf{z})} = \mathbf{0}$ if and only if $D_{\mathbf{z}}\left[det[J^{-1}(\mathbf{z})]P_{\mathbf{Z}|Y}(\mathbf{z}|i)\right]$ is in the null space of $J^{-1}(\mathbf{z})$. Since $J(\mathbf{z})$ has full rank,

$$D_{\mathbf{x}}P_{\mathbf{X}|Y}(\mathbf{x}|i)\Big|_{\mathbf{x}=T(\mathbf{z})} = \mathbf{0} \Leftrightarrow D_{\mathbf{z}}\left[det[J^{-1}(\mathbf{z})]P_{\mathbf{Z}|Y}(\mathbf{z}|i)\right] = \mathbf{0}.$$

It follows that, if $\mathbf{x} = T(\mathbf{z})$ is the maximum of $P_{\mathbf{X}|Y}(\mathbf{x}|i)$, then

$$D_{\mathbf{z}}(det[J^{-1}(\mathbf{z})])P_{\mathbf{Z}|Y}(\mathbf{z}|i) + det[J^{-1}(\mathbf{z})]D_{\mathbf{z}}P_{\mathbf{Z}|Y}(\mathbf{z}|i) = 0,$$

[2] Several definitions have been proposed for the vector derivative. The one adopted here, equation (3.6), is that used in [114].

$$\frac{1}{det[J^{-1}(\mathbf{z})]}D_{\mathbf{z}}(det[J^{-1}(\mathbf{z})]) = -\frac{1}{P_{\mathbf{Z}|Y}(\mathbf{z}|i)}D_{\mathbf{z}}P_{Z|Y}(\mathbf{z}|i),$$

$$D_{\mathbf{z}}\left[\log\frac{det[J(\mathbf{z})]}{P_{\mathbf{Z}|Y}(\mathbf{z}|i)}\right] = \mathbf{0}.$$

Since the log is a monotonic function, this means that $\frac{det[J(\mathbf{z})]}{P_{\mathbf{Z}|Y}(\mathbf{z}|i)}$ has a critical point at $T^{-1}(\mathbf{x})$. For most parametric densities, $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ only has one critical point, implying that this will be the only critical point of $\frac{det[J(\mathbf{z})]}{P_{\mathbf{Z}|Y}(\mathbf{z}|i)}$. In any case, it follows that T depends on $P_{\mathbf{Z}|Y}(\mathbf{z}|i)$. □

The theorem explains why most texture retrieval approaches work well on databases of homogeneous images (like that of Figure 3.3 a)), but clearly fail when this is not the case. Since the pixel colors of non-homogeneous images (like that of Figure 3.3 b)) have different statistics according to their spatial location, the associated densities are inherently multi-modal. It is therefore impossible to find a generic transformation mapping them into a set of unimodal densities without compromising the Bayes error.

Yet, the vast majority of texture retrieval methods are based on a feature transformations that does not depend on the class conditional pdfs and the Gaussian representation (implicit in quadratic metrics like the Mahalanobis distance) [137, 21, 144, 178, 163, 96, 134, 102, 104, 174]. It is therefore not surprising that they cannot guarantee low Bayes error in $\mathcal{X}$. While data-dependent transformations have been proposed in the literature [40, 176], these usually imply finding a set of *discriminant* features that can only be computed by considering all the image classes simultaneously. This is impossible in the CBIR context since 1) there may be too many classes, and 2) the feature transformation has to be recomputed every time the database changes.

Putting it plainly, the theorem states that there is no such thing as a "free lunch". If we want to rely on simple models for density estimation, we will necessarily have to rely on a complicated feature transformation. And, in the end, the complexity of finding such a transformation may very well be orders of magnitude greater than that required by more sophisticated density estimation. Why then has the texture community been so focused on the question of finding good features for texture characterization? One possible explanation is that this is an historical consequence of the assumption that different textures can always be cleanly segmented and a texture classifier will operate on homogeneous texture patches[3].

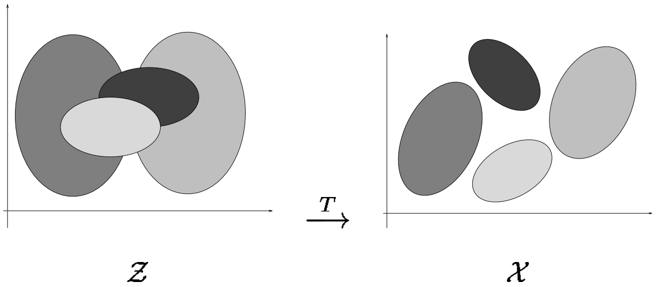


Figure 3.4: When the classes are Gaussian in $\mathcal{Z}$, a feature transformation can help by reducing their overlap in $\mathcal{X}$.

Since, by definition, homogeneous images have similar statistics everywhere, the densities of their observations are close to unimodal and any sensible feature transformation will generate unimodal densities in $\mathcal{X}$. For example, any linear transformation will generate a collection of Gaussians in $\mathcal{X}$ if the class-conditional pdfs are already Gaussian in $\mathcal{Z}$. In this case, as illustrated by Figure 3.4, a feature transformation can allow significant improvements in classification accuracy by making the classes in $\mathcal{X}$ more clearly separated than they are in $\mathcal{Z}$.

In practice, however, it is arguable that the segmentation problem can be cleanly solved before recognition. In fact, it it may never be possible to guarantee that the classifier will process samples from unimodal distributions. In this case, Theorem 4 shows that strategy S2 is hopeless as long as one insists on preserving the Bayes error. Unfortunately, unless this is the case, there is no guarantee that good performance in $\mathcal{X}$ will imply good accuracy in $\mathcal{Z}$, the ultimate goal of the retrieval system.

3.3 An alternative strategy

In the context of minimizing probability of error, the two standard strategies can be seen as two ends of a continuum: while strategy S1 is intransigent with respect to any loss in Bayes error and therefore asks too much from the feature representation; strategy S2

[3] Most of the databases used to evaluate texture recognition are indeed composed of homogeneous images.

constrains the representation to trivial models, expecting the feature transformation to do the impossible.

It seems that a wiser position would be to stand somewhere in between the two extrema. Since the overall probability of error is upper bounded by the sum of the Bayes and estimation errors, we need to consider the two *simultaneously*. While the crucial requirement for low Bayes error is *invertability* of the feature transformation, the crucial requirement for low estimation error is *low-dimensionality* in $\mathcal{X}$. Since we want $\mathcal{Z}$ to be high-dimensional, the two requirements are conflicting and a trade-off between invertability and dimensionality is required. This means that both the feature transformation and representation have an important role in the overall representation.

On one hand, the feature transformation should provide the *dimensionality reduction* necessary for density estimation to be feasible (but no more). On the other hand, the feature representation should be expressive enough to allow accurate estimates without requiring the dimension of $\mathcal{X}$ to be too low, therefore allowing the transformation to be close to invertible. This is the main idea behind our strategy.

Like strategy S2, we rely on a feature transformation. However, we limit its role to enabling dimensionality reduction; i.e. if we define a feature transformation to be of dimensionality reduction level $n-k$ when

$$T : R^n \to R^k,\ k \leq n,$$

then the *the optimal feature transformation is the one that, for a given level of dimensionality reduction, is as close to invertible as possible*. The idea of *close to invertible transformation* is intimately related to the idea of *semantics-preserving compression* advocated in the design of the Photobook system [129]. Here, we replace the idea of preserving semantics with the simpler and more generic goal of preserving information. It is very difficult to define semantics-preserving transformations without restricting databases to a specific domain or assuming the existence of a perfect segmentation algorithm.

Like strategy S1, we also place strong emphasis on the feature representation. Here, the goal is to guarantee that we will be operating as close to the Bayes error as possible for all levels of dimensionality reduction. In particular, as illustrated by Figure 3.5, we look for the representation that simultaneously satisfies the following requirements:

- like the Gaussian, is computationally tractable in high dimensions;
- like the histogram, can capture the details of multi-modal densities.

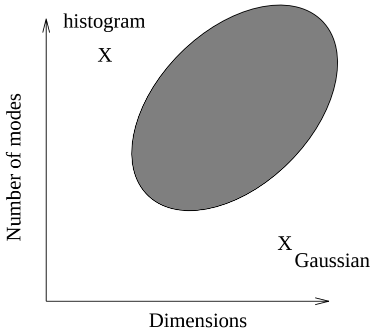


Figure 3.5: The space of feature representations. The histogram can account for multi-modal distributions, but is infeasible to compute in high dimensional feature spaces. The Gaussian is unimodal, but can be computed in high dimensions. The shaded area represents the region of the space where new feature representations are needed for the implementation of generic retrieval systems.

In the next chapter, we study the issue of dimensionality reduction. Feature representation is addressed in Chapter 5.