# A Theoretically Optimal Probabilistic Classifier Using Class-Specific Features

Paul M. Baggenstoss (p.m.baggenstoss@ieee.org)
and Heinrich Niemann (niemann@immd5.informatik.uni-erlangen.de)
Lehrstuhl für Musterkennung (Informatik 5)
Universität Erlangen-Nürnberg,
Martensstrasse 3, 91058 Erlangen, Germany

This work was supported by Office of Naval Research (ONR-321US) and Naval Undersea Warfare Center, Newport RI, USA

January 8, 2001

## Abstract

In this paper, we present a new approach to the design of probabilistic classifiers. Rather than working with a common high-dimensional feature vector, the classifier is written in terms of separate feature vectors chosen specifically for each class and their low-dimensional PDFs. While sufficiency is not a requirement, if the feature vectors are sufficient to distinguish the corresponding class from a common (null) hypothesis, the method is equivalent to the maximum a posteriori probability (MAP) classifier. The method has applications to speech, image, and general pattern recognition problems.

## 1 Problem Statement

Consider the problem of classifying a data sample $\mathbf{x}$ into one of $M$ classes. The optimal maximum a posteriori (MAP) or Bayesian classifier is

$$\arg \max_{j=1}^{M} p(H_j|\mathbf{x}) = \arg \max_{j=1}^{M} p(\mathbf{x}|H_j)\, p(H_j). \quad (1)$$

If the LFs, $p(\mathbf{x}|H_j)$, are not known, it is necessary to estimate them from *training data*. To avoid dimensionality issues, it is often necessary to reduce $\mathbf{x}$ to a smaller vector of statistics or *features*, $\mathbf{z} = T(\mathbf{x})$. The traditional feature-based classifier is based on the PDF estimates of $\mathbf{z}$ under each hypothesis:

$$\arg \max_{j=1}^{M} \hat{p}(\mathbf{z}|H_j)\, p(H_j). \quad (2)$$

Because $\mathbf{z}$ needs to be sufficient for the entire problem, it often must contain a large number of features. The two fundamental problems in designing such a classifier are (1) to obtain a low-dimensional feature vector with sufficient information and (2) to obtain its joint PDF estimate under each class hypothesis. PDF estimation above a dimension of about 5 is problematic [**?**]. As a result, feature reduction is often needed [**?**]. High-dimensional PDF estimators can perform well as classifiers if there exists good separability among the classes in the high-dimensional space. Further performance improvements are difficult without addressing the dimensionality problem in a more direct manner.

Consider a set of *class-specific* feature vectors (FV):

$$\mathbf{z}_j \triangleq T_j(\mathbf{x}), \quad 1 \le j \le M,$$

which do not need to be unique (we could have $T_j = T_k$ for some $k \neq j$). Criterion for selecting features is discussed below. The dimension of the class-specific FVs will be equal to or lower than that of the common FV $\mathbf{z}$. Assume that we have available estimates of the PDFs of each FV under the corresponding hypothesis:

$$\hat{p}(\mathbf{z}_j|H_j) \quad 1 \le j \le M. \quad (3)$$

We seek a way to re-write the classifier (1) in terms of the PDFs of the class-specific features (3). To make fair likelihood function (LF) comparisons, it is necessary to "project" these PDFs back to the original data space.

## 2 Theoretical Results

We define the "projected" PDF as

$$\hat{p}(\mathbf{x}|H_j) \triangleq \left[\frac{p(\mathbf{x}|H_{0,j})}{p(T_j(\mathbf{x})|H_{0,j})}\right] \hat{p}(T_j(\mathbf{x})|H_j), \qquad (4)$$

where $H_{0,j}$ is class-dependent null hypothesis that can be a simplified case such as independent Gaussian or exponentially distributed noise. We will prove shortly that the functions $\{\hat{p}(\mathbf{x}|H_j)\}$ given in equation (4) are indeed PDFs and furthermore they induce the corresponding PDF $\hat{p}(\mathbf{z}_j|H_j)$ on $\mathbf{z}_j$. We assume that for each $j$, the PDFs $p(\mathbf{x}|H_{0,j})$ and $p(T_j(\mathbf{x})|H_{0,j})$ are known *exactly* and for all realizations of $\mathbf{x}$ we have $p(T_j(\mathbf{x})|H_{0,j}) > 0$.

**Theorem 1** *Let $\mathcal{X}$ be a range of possible realizations of $\mathbf{x}$. Let $p_x(\mathbf{x}|H_0)$ be a PDF defined on $\mathcal{X}$ and Let $p_x(\mathbf{x}|H_0) > 0$ for all $x \in \mathcal{X}$. Let $\mathcal{Z}$ be the image of $\mathcal{X}$ under the transformation $\mathbf{z} = T(\mathbf{x})$. Let $p_z(\mathbf{z}|H_0)$ be the PDF of $\mathbf{z}$ when $\mathbf{x}$ is drawn from the PDF $p_x(\mathbf{x}|H_0)$. Thus, $p_z(\mathbf{z}|H_0) > 0$ for all $z \in \mathcal{Z}$. Let $f_z(\mathbf{z})$ be any PDF defined on $\mathcal{Z}$. Then the function defined by*

$$f_x(\mathbf{x}) = \frac{p_x(\mathbf{x}|H_0)}{p_z(T(\mathbf{x})|H_0)}\, f_z(T(\mathbf{x})) \qquad (5)$$

*is a PDF defined on $\mathcal{X}$, thus it has unit area. Furthermore, if $\mathbf{x}$ is drawn from the distribution $f_x(\mathbf{x})$ as defined in (5), then the PDF of $\mathbf{z}$ will be $f_z(\mathbf{z})$.* **Proof:** *Let $M_z(\mathbf{y})$ be the joint moment generating function (MGF) of $\mathbf{z}$. By definition,*

$$\begin{aligned}
M_z(\mathbf{y}) &= \mathrm{E}_z\left\{e^{\mathbf{y}'\mathbf{z}}\right\} = \mathrm{E}_x\left\{e^{\mathbf{y}'T(\mathbf{x})}\right\} \\
&= \int_{\mathbf{x}\in\mathcal{X}} e^{\mathbf{y}'T(\mathbf{x})}\, \frac{p_x(\mathbf{x}|H_0)}{p_z(T(\mathbf{x})|H_0)}\, f_z(T(\mathbf{x}))\, \mathrm{d}\mathbf{x} \\
&= \int_{\mathbf{x}\in\mathcal{X}} e^{\mathbf{y}'T(\mathbf{x})}\, \frac{f_z(T(\mathbf{x}))}{p_z(T(\mathbf{x})|H_0)}\, p_x(\mathbf{x}|H_0)\, \mathrm{d}\mathbf{x} \\
&= \mathrm{E}_{x|H_0}\left\{e^{\mathbf{y}'T(\mathbf{x})}\, \frac{f_z(T(\mathbf{x}))}{p_z(T(\mathbf{x})|H_0)}\right\} \\
&= \mathrm{E}_{z|H_0}\left\{e^{\mathbf{y}'\mathbf{z}}\, \frac{f_z(\mathbf{z})}{p_z(\mathbf{z}|H_0)}\right\} \\
&= \int_{\mathbf{z}\in\mathcal{Z}} e^{\mathbf{y}'\mathbf{z}}\, \frac{f_z(\mathbf{z})}{p_z(\mathbf{z}|H_0)}\, p_z(\mathbf{z}|H_0)\, \mathrm{d}\mathbf{z} \\
&= \int_{\mathbf{z}\in\mathcal{Z}} e^{\mathbf{y}'\mathbf{z}}\, f_z(\mathbf{z})\, \mathrm{d}\mathbf{z},
\end{aligned}$$

*from which we conclude that the PDF of $\mathbf{z}$ is $f_z(\mathbf{z})$. The above proof may be modified to show that $f_x(\mathbf{x})$ has area 1.*

The PDF $f_x(\mathbf{x})$ may be thought of as a PDF constructed on $\mathcal{X}$ in such a way that $\mathbf{z} = T(\mathbf{x})$ is the sufficient statistic (SS) to distinguish $f_x(\mathbf{x})$ from $p_x(\mathbf{x}|H_0)$. By the invariant property of likelihood ratios for SSs, if $\mathbf{z}_j = T_j(\mathbf{x})$ is a SS for $H_j$ vs. $H_0$ and $f_z(\mathbf{z}) \to p(\mathbf{z}|H_j)$, then $f_x(\mathbf{x}) \to p(\mathbf{x}|H_j)$. Applying this to the problem at hand, we have the class-specific classifier

$$\arg\, \max_{j=1}^{M} \left[\frac{p(\mathbf{x}|H_{0,j})}{p(\mathbf{z}_j|H_{0,j})}\right] \hat{p}(\mathbf{z}_j|H_j)\; p(H_j). \qquad (6)$$

Furthermore, see that if for each $j$, $\mathbf{z}_j$ is a SS to distinguish $H_j$ from $H_{0,j}$, and $\hat{p}(\mathbf{z}_j|H_j) \to p(\mathbf{z}_j|H_j)$, (6) becomes the optimal MAP classier (1). This fact provides the theoretical guide for feature selection. That is, look for features which distinguish a given class from $H_{0,j}$. Or if $T_j(\,)$ is fixed, choose $H_{0,j}$ so that $T_j(\mathbf{x})$ is approximately sufficient for distinguishing $H_j$ from $H_{0,j}$. Advantages of the class-specific method include:

- Reduced feature PDF dimension.

- Modular architecture.
- The class-specific method relies partially on $p(\mathbf{z}_j|H_j)$ and partially on $p(\mathbf{z}_j|H_0)$, which is known a priori. The denominator has the effect of "assisting" that class for which the data appears least likely under the null hypothesis.

Note that in equation (6), the PDFs $p(\mathbf{x}|H_{0,j})$ and $p(\mathbf{z}_j|H_{0,j})$ must be accurately computed even if data samples are *significantly* different from the $H_{0,j}$ hypothesis. Therefore, accurate tail behavior is essential.

While $H_{0,j}$ may be dependent on $j$, having a common null hypothesis has some advantages. In this case, (6) becomes

$$\arg \max_{j=1}^{M} \frac{\hat{p}(\mathbf{z}_j|H_j)}{p(\mathbf{z}_j|H_0)}\, p(H_j), \tag{7}$$

which is in the form of a set of dedicated *detectors*. Having a detector-like structure has obvious advantages at low signal-to-noise ratio (SNR) because separate thresholds can be set on each detector to reject samples instead of forcing a decision.

## 2.1 Theory extension: Hidden Markov Modeling

An $M$-state HMM involves a set of $N$ state occurrences $\boldsymbol{\theta} \triangleq \{q[1] \dots q[N]\}$ where $1 \leq q[t] \leq M$. The sequence $\boldsymbol{\theta}$ is a realization of the Markov chain with state priors $\{\pi_j, j = 1, 2 \dots M\}$ and $M \times M$ state transition matrix $A = \{a_{ij}\}$. The observations $\mathbf{X} \triangleq \{\mathbf{x}[1], \mathbf{x}[2] \dots \mathbf{x}[N]\}$ are realizations from a set of state PDF's

$$p(\mathbf{x}|H_j), \quad j = 1, 2 \dots M,$$

where $H_j$ is the condition that state $j$ is true. We assume the observations are independent, thus

$$p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{t=1}^{N} p(\mathbf{x}[t]|H_{q[t]}).$$

The complete set of parameters defining the HMM are

$$\lambda \triangleq \left[\{\pi_j\}, \{a_{ij}\}, \{p(\cdot|H_{q[t]})\}\right],$$

where $\sum_{j=1}^{M} \pi_j = 1$, $\sum_{j=1}^{M} a_{ij} = 1$. The Baum-Welsh algorithm maximizes the LF over $\lambda$ [**?**]. The LF is written as [**?**]

$$p(\mathbf{X}; \hat{\lambda}) = \sum_{\boldsymbol{\theta}} \hat{\pi}_{q[1]}\, \hat{p}(\mathbf{x}[1]|H_{q[1]}) \cdot \prod_{n=2}^{N} a_{q[n-1]q[n]}\, \hat{p}(\mathbf{x}[n]|H_{q[n]}), \tag{8}$$

where the summation of $\boldsymbol{\theta}$ is over all possible state sequences of length $N$. To address the dimensionality issue, most implementations reduce the observations to a FV $\mathbf{Z} \triangleq \{\mathbf{z}[1], \mathbf{z}[2] \dots \mathbf{z}[N]\}$, where $\mathbf{z}[t] = T(\mathbf{x}[t])$. Instead, if we appy (4) with common $H_0$,

$$L(\mathbf{Z}; \hat{\lambda}) \triangleq \frac{p(\mathbf{X}; \hat{\lambda})}{p(\mathbf{X}|H_0)} = \sum_{\boldsymbol{\theta}} \hat{\pi}_{q[1]} \left[\frac{\hat{p}(\mathbf{z}_{q[1]}[1]|H_{q[1]})}{p(\mathbf{z}_{q[1]}[1]|H_0)}\right] \cdot \prod_{n=2}^{N} \left[a_{q[n-1]q[n]}\, \frac{\hat{p}(\mathbf{z}_{q[n]}[n]|H_{q[n]})}{p(\mathbf{z}_{q[n]}[n]|H_0)}\right], \tag{9}$$

A variation of the Baum-Welsh algorithm has been derived for estimation of the parameters of the feature PDFs $\{\hat{p}(\mathbf{z}_j|H_j)\}$ by modelling them as Gaussian mixtures [**?**]. The result is an HMM with state-dependent features. Because each state can have a different FV, it is possible to use special processing to take advantage of the special temporal or frequency-domain character of each state. This algorithm has been tested on simulated data and has show superior performance in comparison to the standard HMM. Part of this improvement is due a self-initialization effect. This effect is due to the dominant role of the denominator term $p(\mathbf{z}_j|H_0)$, which is known in advance.

## 2.2 Computer Simulation

A computer experiment designed using synthetic signals with known sufficient statistics was conducted to verify (7). The results are published in a recent paper [**?**]. The experiment featured three synthetic classes, each with a one-dimensional statistic. The performance of the class-specific classifier using one-dimensional PDF estimates was compared with the classifier constructed using the three-dimensional common FV composed of the three class-specific features. Since both classifiers used the same features and PDF estimation method (Gaussian mixtures), the performance comparison compared only the classifier architectures. The classification performance is

plotted in Figure 1 and shows that more than a factor of 10 fewer training samples are required by the class-specific classifier for the same level of performance.

Figure 1: Comparison of class-specific and traditional classifier showing probability of correct classification ($P_{cc}$) as a function of the number of training samples from each class. Each estimate of $P_{cc}$ is an average of 10 independent trials using 500 testing samples from each class in each trial.

# 3 Applications

## 3.1 Time-series (Speech) Analysis

The class-specific approach lends itself well to optimal time-series segmentation. Let the length-$T$ time-series be divided into $K$ segments with ending times $\boldsymbol{\tau} = [t_1, t_2 \ldots t_{K-1}]$. Within segment $k$, we assume the data is a realization of model $m_k$. Let $\boldsymbol{\mu} = [m_1, m_2 \ldots m_K]$, $1 \leq m_k \leq M$. Determination of $\boldsymbol{\tau}$ and $\boldsymbol{\mu}$ may be formulated as a maximum likelihood problem. If we assume that for a fixed $\boldsymbol{\mu}, \boldsymbol{\tau}$, the $K$ segments are independent,

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\tau}) = \prod_{k=1}^{K} p(x[t_{k-1}+1], \ldots, x[t_k]|H_{m_k}), \quad (10)$$

where $t_0 \triangleq 0$, and $t_K \triangleq T$. The maximization of this quantity may be written as

$$\max_{\boldsymbol{\mu}, \boldsymbol{\tau}} p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\tau}) = \max_{\boldsymbol{\tau}} \left\{ \max_{\boldsymbol{\mu}} p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\tau}) \right\}. \quad (11)$$

The inner maximization may be performed independently on each segment. Then, the problem may be solved without exhaustive search using dynamic programming. For each time $t$, the total log-likelihood of the best segmentation which ends at time $t$ may be calculated recursively. Aside from the computational aspects, the main difficulty of implementing (11) is the necessity to know the LFs $p(\mathbf{x}|H_j)$. Thus, they are often limited to one model whose parameters are allowed to change from segment to segment. An example is the segmentation of a DFT into constant-power segments [**?**]. In contrast, the class-specific method allows likelihood comparisons of competing models with different structure based only on their sufficient statistics via (4). Let

$$\mathbf{z}_j[t_1, t_2] \triangleq T_j(x[t_1], \ldots, x[t_2]).$$

Applying (4) with a common $H_0$,

$$\frac{p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\tau})}{p(\mathbf{x}|H_0)} = \prod_{k=1}^{K} \frac{p(\mathbf{z}_{m_k}[1+t_{k-1}, t_k]|H_{m_k})}{p(\mathbf{z}_{m_k}[1+t_{k-1}, t_k]|H_0)}. \quad (12)$$

Below, we explain the steps necessary for implementation of (12) and describe the specific choices we have made relative to the segmentation of speech data.

1. Selection of the $H_0$ hypothesis. For speech, it is useful to use $N(0,1)$ independent Gaussian noise.

2. Selection of sufficient statistics to differentiate each model from $H_0$. The two main models in speech are unvoiced ($H_1$) and voiced ($H_2$) processes. We selected as approximate sufficient statistics for $H_1$ the first seven autocorrelation function (ACF) lags $\mathbf{z}_1 = \mathbf{r}^6$, where $\mathbf{r}^6 = [r_0 \ldots r_6]$. For $H_2$, we used $\mathbf{z}_2 = [\mathbf{r}^8, r_\mathrm{p}, i_\mathrm{p}]$ where $r_\mathrm{p}, i_\mathrm{p}$ are the value and lag index if the highest ACF peak in the range of human pitch.

3. Determination of the denominator PDFs $p(\mathbf{z}_j|H_0)$. For a set of unweighted DFT-derived ACF estimates obtained from independent Gaussian noise, the exact joint moment generating function (MGF) may be determined, however a closed form expression for the joint PDF can not be found. Through application of the *saddlepoint approximation* or *tilted Edgeworth expansion* [**?**], accurate PDF approximations valid in the distant tails may be obtained [**?**]. For voiced speech, the introduction of the feature $i_\mathrm{p}$ requires using the PDF factorization

$$p(\mathbf{r}^8, r_\mathrm{p}, i_\mathrm{p}|H_0) = p(\mathbf{r}^8, r_\mathrm{p}|i_\mathrm{p}, H_0)\, p(i_\mathrm{p}|H_0), \quad (13)$$

where $p(i_\mathrm{p}|H_0)$ is approximated as a uniform distribution.

4. Determination of appropriate numerator PDFs $p(\mathbf{z}_j|H_j)$. These PDF may be obtained by PDF estimation using labeled training data, or may be constructed using prior knowledge. For voiced speech, the factorization (13) also applies under $H_2$ and we assume $p(i_\mathrm{p}|H_2)$ is uniform over the human pitch range. Often, it is useful to work with an alternative feature set with well-behaved statistics obtained by invertible transformation of $\mathbf{z}_j$. For $H_1$, we have found it useful to work with the alternative feature set $\mathbf{z}_1' = [\rho, \boldsymbol{\kappa}^6]$, where $\rho = \log(r_0)$, and $\boldsymbol{\kappa}^6 = [\kappa_1, \ldots, \kappa_6]$, and $\kappa_i = \log((1-K_i)/(1+K_i))$, where $K_i$, are the 6-th order reflection coefficients (RCs). The PDF $p(\mathbf{z}_1|H_1)$ may be obtained from $p(\mathbf{z}_1'|H_1)$ using a change of variables and the Jacobian of the transformation. We have found that the components of $\mathbf{z}_1'$ are approximately Gaussian and independent under $H_1$. Appropriate means and variances were obtained by observing typical data. For $H_2$, we have found it useful to work with $\mathbf{z}_2' = [\rho, \boldsymbol{\kappa}^8, \rho_\mathrm{p}, i_\mathrm{p}]$, where $\boldsymbol{\kappa}^8$ and $\rho$ are similarly defined and $\rho_\mathrm{p} = \log r_\mathrm{p}$.

An example of a segmented time-series is shown in Figure 2. This example was obtained by fitting voiced and un-voiced speech models to the segments. Good quality speech has been re-synthesized from the features from the segments obtained in this way.

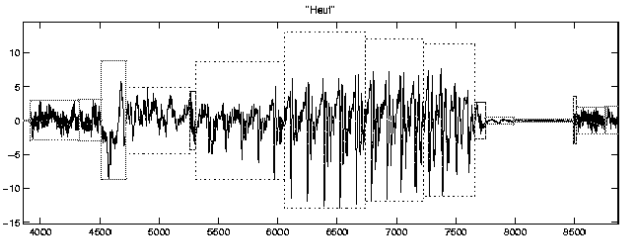


Figure 2: Example of optimal speech segmentation. Unvoiced segments are shown with finer dotted lines. Spoken word is german "heut", pronounced "hoit". The length of the voiced speech segments in the final solution are approximately equal to a multiple of the pitch period. This provides the best "fit" to the model which used unweighted DFT processing.

## 3.2 Image Recognition

The class-specific method may also be applied to image processing. To realize a benefit by direct application of equation (4), there should exist a low-dimensional FV for each object class which is nearly sufficient to distinguish the given object from $H_0$ (Gaussian or exponentially distributed independent noise on the image plane). Unfortunately, most object recognition problems are concerned with objects which are similar or share the same features. A better approach is to represent each object class as a collection of image primitives. These primitives can be then represented by different FVs. The recognition of the object classes can be accomplished by statistically modelling the spatial relationships among the image primitives. To test this concept, a simplified shape-recognition experiment was conducted. Figure 3 shows the original camera image which contains circles, squares, and a pentagon.

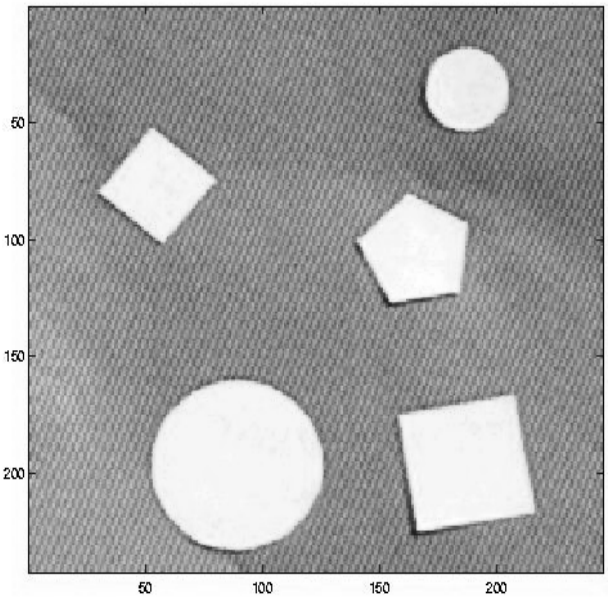


Figure 3: Original image (245 wide by 242 high).

The image of 245-by-242 pixels was pre-processed as follows. The image data is defined as $\{x_{ij},\ 1 \leq i \leq$

$N,\ 1 \leq j \leq M\}$. Let $\mathcal{N}(i,j)$ be circular neighborhood of 16 pixels radius around pixel (i,j):

$$\mathcal{N}(i,j) \triangleq (n,m) : \sqrt{(n-i)^2 + (m-j)} \leq 16.$$

We define $\mu(i,j)$ and $\sigma(i,j)$ as the sample mean and standard deviation of the data in neighborhood $\mathcal{N}(i,j)$. We then define the normalized neighborhood data at pixel $(i,j)$ as

$$\tilde{\mathcal{X}}(i,j) \triangleq \{\tilde{x}_{nm}^{ij}\} : (n,m) \in \mathcal{N}(i,j),$$

where

$$\tilde{x}_{nm}^{ij} \triangleq \frac{x_{nm} - \mu(i,j)}{\sigma(i,j)}.$$

The idea is then to test $\tilde{\mathcal{X}}(i,j)$ for each object primitive in each orientation. An example of a primitive for a 90-degree corner centered at pixel $(i,j)$ with an orientation of $\phi_l$=135 degrees is shown in Figure 4. Let there be $L$ orientations (we use 5-degree quantizations $L = 72$). Let

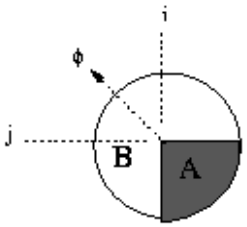


Figure 4: Image primitive for a 90-degree corner centered at pixel $(i,j)$ with an orientation of $\phi$=135 degrees.

$$Q(i,j,l,m) \triangleq \frac{p(\tilde{\mathcal{X}}(i,j)|H_m,\phi_l)}{p(\tilde{\mathcal{X}}(i,j)|H_0)},$$

$1 \leq l \leq L, \quad 1 \leq m \leq M$, where $M$ is the number of primitives. This quantity is the likelihood ratio test at pixel $(i,j)$ between the hypothesis that primitive $m$ is present in orientation $\phi_l$ and hypothesis $H_0$. To apply the class-specific method, we re-write this as

$$Q(i,j,l,m) \simeq \frac{p(\mathbf{z}_{i,j,l,m}|H_m,\theta_l)}{p(\mathbf{z}_{i,j,l,m}|H_0)},$$

where $\mathbf{z}_{i,j,l,m}$ is an approximate sufficient statistic for this binary test. Note that normalizing $\tilde{\mathcal{X}}(i,j)$ makes it easier to find approximate sufficient statistics to distinguish from zero-mean Gaussian noise of unit variance.

To illustrate the selection of approximate sufficient statistics, we consider a one-dimensional example. Consider a length-$N$ time-series $\mathbf{x} = [x_1, x_2 \ldots x_N]$. Let hypothesis $H_b$ be defined as

$$H_b : \begin{cases} x_t = a_1 + n_t, & 1 \leq t \leq b \\ \\ x_t = a_2 + n_t, & b > t \leq N, \end{cases}$$

where $a_1, a_2$ are two unknown constants. We define $H_0$ as independent $N(0,1)$ Gaussian noise. A good FV s the sums of the samples in the two regions:

$$\mathbf{z}_{1,b} = \sum_{t=1}^{b} x_t, \quad \mathbf{z}_{2,b} = \sum_{t=b+1}^{N} x_t$$

Thus, we have

$$\frac{p(\mathbf{x}|H_1,b)}{p(\mathbf{x}|H_0)} \simeq \frac{p(\mathbf{z}_{1,b},\mathbf{z}_{2,b}|H_1,b)}{p(\mathbf{z}_{1,b},\mathbf{z}_{2,b}|H_0)}. \tag{14}$$

Under $H_0$, $\mathbf{z}_{1,b}$ and $\mathbf{z}_{2,b}$ are mutually independent. Both are Gaussian with zero mean, while $\mathbf{z}_{1,b}$ has variance $b$, and $\mathbf{z}_{1,b}$ has variance $N-b$. Clearly ,

$$\begin{aligned} \log p(\mathbf{z}_{1,b},\mathbf{z}_{2,b}|H_0) = &-0.5\log(2\pi b) - \tfrac{1}{2b}\mathbf{z}_{1,b}^2 \\ &-0.5\log(2\pi(N-b)) \quad - \tfrac{1}{2(N-b)}\mathbf{z}_{2,b}^2. \end{aligned}$$

The numerator PDF in (14) needs to be defined. If we have no prior knowledge about $a_1, a_2$, we may assume uniform distributions or may ignore the prior distribution altogether. It has been experimentally verified that the numerator has minimal effect on the test. The classification decision may be made based only on the denominator, i.e. how *unlikely* the data is under $H_0$.

To extend the above example to the 90-degree corner primitive in Figure 4, we compute the sum of the samples in regions A and B. We then apply a similar argument to arrive at a formula for the joint distribution of the two region sums. We have created similar models for various corner widths, straight edges, and curved edges.

Depending on the application, the orientation information may or may not be important. If it is not important,

one may reduce the data by maximizing over the orientation $\phi_l$:

$$Q'(i,j,m) = \max_l \; Q(i,j,l,m).$$

An example of $Q'(i,j,m)$ for $m$ corresponding to the hypothesis of a 108-degree corner (as found in a pentagon) is shown in Figure 5. Note that the five 108-degree corners of the pentagon were detected, while none of the 90-degree corners in the image were detected. The image may be reduced in this way to small areas which are likely to be the locations of the desired primitives. Work is continuing to develop statistical models to describe the spatial relationship between these areas.

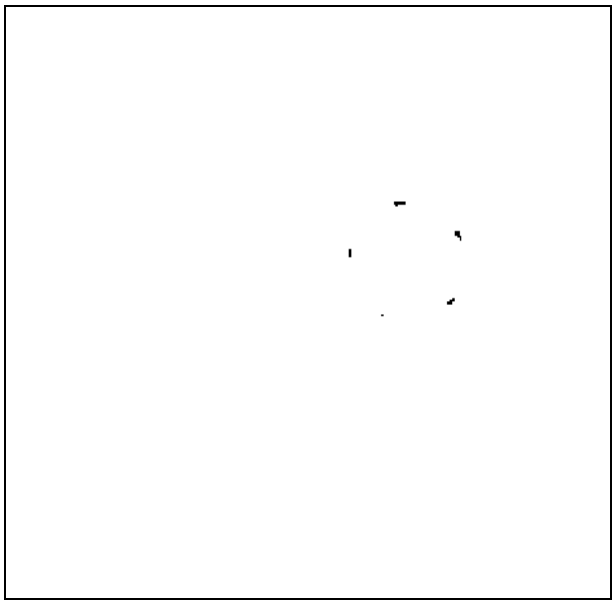

Figure 5: Image processed to detect 108-degree corners and thresholded for display. Numerator densities were ignored.

# 4 Conclusions

A new general method of pattern recognition is proposed. This method allows class-specific feature vectors to be used in an optimal classifier framework. The new method does not suffer from the same dimensionality issues as the traditional classifier because optimality of the method only requires feature vectors to be sufficient to distinguish the given class from a common (null) hypothesis. The method is based on a new theorem which permits "projecting" the feature vector PDFs back to the original data space. The method is showing promise in three applications: HMM modelling, time-series (speech) analysis, and image recognition.