Abstract

We utilize the discriminant analysis to select wavelet features for efficient object detection. The analysis applies to the Bayesian classifier and is extended to the case of boosting. Based on the error analysis under the Bayesian decision rule, we reduce the number of coefficients involved in detection to lower the computational cost. Using a Hidden Markov Tree (HMT) model to describe the pattern distributions, we introduce the concept of error-bound-tree (EBT) to relate feature selection to error reduction. The scheme selects discriminative features that are adaptive to the pattern and allows the detector to reach a decision faster.

1. Introduction

Detecting real world objects, such as people, faces and cars, from clutter background images is a widely used technique. Many application systems demand accurate and yet quick detection. In other words, reducing detection error and reducing the computational complexity are two major issues. Most work on object detection focuses on performance improvement. The complexity issue has received less attention. In this work, we consider both issues. We perform error analysis to select the distinguishing features, which allows the reduction in computation without jeopardizing the performance.

In the high-dimensional space, the set of all possible images are distributed as a scattered cluster. In this cluster, the object type images occupy a small subset surrounded by the rest non-object images (Figure 1). Generally, the two classes are not linearly separable and linear discriminant analysis does not fulfill the task. Refined class descriptions such as Gaussian-mixtures and neural networks are more suitable, but they introduce high computational cost to the detector. Wavelet representation has been used in several object detection systems [1,2,3,4]. The transform has low complexity and enables a better description for the family of image signals.

In this paper, we analyze the error probability for the Bayesian classifier and evaluate the discriminability of the subband coefficients for feature selection. The analysis is extended to boosting by deriving the pseudo-distribution functions from a sequence of weak classifiers. In section 2, we represent class distributions by the Hidden Markov Tree (HMT) model that characterizes the inter-scale dependence between subband coefficients. In section 3 and 4, we propose an error-bound-tree to analyze the detection error for the Bayesian classifier and its boosting version. The adaptive feature selection scheme is hereby derived. Section 5 shows the test result on the set of frontal view faces. Section 6 summarizes the work.

2. Modeling class distributions

Previous studies on wavelet representation show that natural images can be sufficiently described by properly modeling their subband coefficients. We adopt the Hidden Markov Tree (HMT) model introduced by [5] to emphasize on the salient dependence among coefficients, i.e. the inter-scale dependence.

2.1. Hidden Markov Tree (HMT) model

The following notations are used for the HMT model:

- \( z_{l,x,y}^B \) : coefficient /state variable indexed by location \((x,y)\), subband \( B = \{LH, HL, LL\} \) and scale \( l \);
- \( parent(z_{l,x,y}^B) = z_{l-1,x,2y}^B \) : parent of \( z_{l,x,y}^B \);
- \( child(z_{l,x,y}^B) = \{z_{l,2x+m,2y+n}^B : m,n = 0,1\} \) : children of \( z_{l,x,y}^B \);

The overall set of subband coefficients is expressed as
It allows quick estimate of the state \( S_{l,x,y}^B \) from \( z_{l,x,y}^B \) by simple thresholding. Given enough training examples from both classes, the histogram distributions are easily estimated by counting the number of samples that fall in each cell.

3. Bayesian classifier and boosting

Given the class distributions, the Bayesian decision rule minimizes the classification error. Use \( O \) and \( \overline{O} \) to denote the object and non-object classes respectively, the Bayesian decision rule [7] is implemented by the log-likelihood ratio test,

\[
l(Z) = \log \frac{p_O(Z)}{p_{\overline{O}}(Z)} \geq \tau \quad Z \in O \quad (\tau = \log \frac{P(Z \in O)}{P(Z \in \overline{O})})
\]

Using (1), we adopt the sequential likelihood-ratio test.

3.1. Boosting

Boosting techniques lower classification error by combining the decisions from multiple instances of the classifier \( l'(Z) \). In Adaboost [8], the final hypothesis test is a weighted sum of these weak classifiers.

\[
l(Z) = \sum_{i=1}^{T} \alpha_i l'_i(Z) \quad \left( \sum_{i=1}^{T} \alpha_i = 1 \right)
\]

Denote \( p_{O/\overline{O}}(Z) \) as the distribution functions for the classifier \( l'(Z) \). The final hypothesis test is expressed as

\[
l(Z) = \log \frac{\overline{P}_O(Z)}{\overline{P}_{\overline{O}}(Z)} \quad \overline{P}_O/Z(Z) = \prod_{i=1}^{T} [P_{O/\overline{O}}(Z)]^{\alpha_i}
\]

\( \overline{P}_{O/\overline{O}}(Z) \) are non-negative and bounded functions. We defined them as the pseudo-density functions. The final hypothesis test is equivalent to the single Bayesian classifier derived from the pseudo-densities. Denote

\[
\overline{P}_{O/\overline{O}}(z_{l,x,y}^{LH}, z_{l,x,y}^{HL}, z_{l,x,y}^{HH}, z_{l,x,y}^{0,0,0}, z_{l,x,y}^{0,0,0}, z_{l,x,y}^{0,0,0}) = \prod_{i=1}^{T} [P_{O/\overline{O}}(z_{l,x,y}^{LH}, z_{l,x,y}^{HL}, z_{l,x,y}^{HH}, z_{l,x,y}^{0,0,0}, z_{l,x,y}^{0,0,0}, z_{l,x,y}^{0,0,0})]^{\alpha_i}
\]

(1) and (3) also hold for the pseudo-density functions. The following error analysis applies to both single Bayesian classifier as well as boosting with pseudo-densities.

3.2. Upper bound on Bayesian error

The probability of error is calculated from two types of error, the miss rate and false alarm rate.
\[ \varepsilon = P(Z \in O) \mid \Gamma_1, p_o(Z)dZ + P(Z \in \overline{O}) \mid \Gamma_2, p_\pi(Z)dZ \]  
(7)

where \( \Gamma_1 = \{ Z : l(Z) \geq \tau \} \), \( \Gamma_2 = \{ Z : l(Z) < \tau \} \).

Since the closed-form expression for the error bound is often intractable, we resort to an upper bound as an alternative. One useful lower/upper bound is defined by the Bhattacharyya distance [6], \( \| p_o(Z) p_\pi(Z) \|^{\frac{1}{2}} dZ \),

\[ \varepsilon_l \leq \varepsilon \leq \varepsilon_u \quad (\varepsilon_l \equiv \varepsilon_u^\perp) \]  
(8)

\[ \varepsilon_u = [P(Z \in O) P(Z \in \overline{O})]^{\frac{1}{2}} \| p_o(Z) p_\pi(Z) \|^{\frac{1}{2}} dZ \]  
(9)

Since \( P(Z \in O) P(Z \in \overline{O}) \) is an unknown constant, we only consider the term of Bhattacharyya distance as we continue the discussion.

\[ \varepsilon_B(Z) = \left[ \| p_o(Z) p_\pi(Z) \|^{\frac{1}{2}} dZ \right] \quad (0 \leq \varepsilon_B(Z) \leq 1) \]  
(10)

4. Error analysis and feature selection

By examining the error bound, we can adaptively select the coefficients that produce lower error rate. Such coefficients are put into test ahead of others to achieve a faster decision. In this way, we jointly optimize the detection performance and the computational efficiency.

4.1. Error-bound-tree (EBT)

If we use the same cells to define the histogram estimates for both class distributions, the upper bound for the error probability can be derived as

\[ \varepsilon_B(Z) = \sum_{(S_1,S_2,S_3)} \overline{w}_{S_1,S_2,S_3} \cdot \prod_{i=0}^{1} \prod_{j=0}^{1} \varepsilon_T^B(z_{i,j};S_1) \]  
\[ \cdot \prod_{i=0}^{1} \prod_{j=0}^{1} \varepsilon_T^B(z_{i,j};S_2) \cdot \prod_{i=0}^{1} \prod_{j=0}^{1} \varepsilon_T^B(z_{i,j};S_3) \]  
(11)

\[ \varepsilon_T^B(z_{i,j};S) = \left[ p_o\left( \text{tree}(z_{i,j}) \right) \right] \left[ S_{i,j} \right]^{\frac{1}{2}} \cdot d\left( \text{tree}(z_{i,j}) \right) \]  
(12)

\[ \varepsilon_B(z_{i,j};S) \geq \varepsilon_B(z_{i,j};S) \]  
(13)

This structure is defined as the Error-bound-tree (EBT). In EBT, each node represents the Bayesian error bound incurred by testing the subtree originating from the node given that its parent is in a certain state.

4.2. Adaptive feature selection (AFS)

To classify an image pattern, the wavelet coefficients are put into a sequential Bayesian test in an orderly fashion, beginning with the coarsest subbands. (11)-(12) shows that adding more coefficients to the test brings down the error bound. At each stage, the state information of the previously tested coefficients is available. The idea of feature selection is to use the EBT to predict which of the untested subtrees leads to the most error reduction and proceed along the corresponding branch. Statistically, the selected features are most discriminative in terms of error reduction and adaptive to the pattern under classification. These features can lead to a quicker decision and thus save the computation for processing less discriminative ones.

For jointly minimizing error probability and reducing computational cost, we define the normalized error bound as the criterion for feature selection,

\[ \varepsilon_B(z_{i,j};S) = \left[ \varepsilon_T^B(z_{i,j};S) \right] C\left( \text{tree}(z_{i,j}) \right) \]  
(14)

\( C(\text{tree}(z_{i,j})) \) denotes the cost to obtain the subtree \( \text{tree}(z_{i,j}) \) and its induced likelihood ratio. (11) defines the effective bounding provided by unit computation. In our experiment, uniform cost is assumed for computing each coefficient and its induced likelihood ratio. With adaptive feature selection, the detector reaches decisions faster by ignoring the features that are not discriminative for the pattern under detection. The sequential Bayesian
test with adaptive feature selection is then summarized as:
1. Compute the partial likelihood ratio (4) with coefficients from the coarsest scale.
2. Rank the remaining coefficients according to their effective bounding (14) defined by the EBT. The most discriminative ones are introduced to the sequential test. This step proceeds repeatedly until a decision is reached.

5. Experiments

To test the error analysis based feature selection scheme, we collected 32x32-pixel block images of 706 frontal-view faces and 45066 non-face pattern. Each face image produced 47 additional examples by slight rotation, scaling and shifting. Half of the samples from each class were used in training to obtain the class distributions of the HMT model, the EBT and the normalized bounded. The remaining samples were used for testing. 5-level decomposition was performed with Haar filters. Histograms with 9 cells were used for the distribution and 3 possible states were defined for each coefficient. The same partitions of states and histogram cells were shared by both classes.

Bayesian classifier with boosting and adaptive feature selection was applied to the test set. Figure 4 shows the ROC curves with different number of features involved. The performance almost reaches its optimum by using a small subset (about 160) of adaptive features, out of a total number of 1023 features. Figure 5 shows a few examples of the selected feature map.

Table 1 compares the AFS based sequential test with the standard sequential test from coarse to fine levels. With AFS, the test uses fewer features and yet achieves lower error rate.

<table>
<thead>
<tr>
<th>Sequential test + AFS</th>
<th>Standard sequential test</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M, F)</td>
<td>(N₁, N₂)</td>
</tr>
<tr>
<td>(2, 0)</td>
<td>(118, 126)</td>
</tr>
<tr>
<td>(5, 1)</td>
<td>(91, 91)</td>
</tr>
<tr>
<td>(33, 26)</td>
<td>(66, 59)</td>
</tr>
<tr>
<td>(633, 673)</td>
<td>(36, 28)</td>
</tr>
<tr>
<td>(1053, 600)</td>
<td>(54, 32)</td>
</tr>
</tbody>
</table>

Table 1. Comparison between AFS-based and standard sequential tests. M/F: number of misses/false positives; N₁/N₂: average number of features used in face/non-face class.

6. Conclusions

We have proposed a discriminant analysis method using error analysis. In our method, feature selection is directly guided by error reduction. In this way, we jointly optimize the performance and the computational complexity of the classifier. The analysis can be used to facilitate the tasks with computational constraints.

The overhead cost for feature selection is negligible when processing features is relatively expensive. The scheme can also be applied to other feature types used in classification, as well as temporally adaptive feature selection for tracking objects in image sequences.

References