Presented at CVPR98

Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition

Henry Schneiderman and Takeo Kanade
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

In this paper, we describe an algorithm for object recognition that explicitly models and estimates the posterior probability function, $P(\text{object}|image)$. We have chosen a functional form of the posterior probability function that captures the joint statistics of local appearance and position on the object as well as the statistics of local appearance in the visual world at large. We use a discrete representation of local appearance consisting of approximately 10^6 patterns. We compute an estimate of $P(\text{object}|image)$ in closed form by counting the frequency of occurrence of these patterns over various sets of training images. We have used this method for detecting human faces from frontal and profile views. The algorithm for frontal views has shown a detection rate of 93.0% with 88 false alarms on a set of 125 images containing 483 faces combining the MIT test set of Sung and Poggio with the CMU test sets of Rowley, Baluja, and Kanade. The algorithm for detection of profile views has also demonstrated promising results.

1. Introduction

In this paper we derive a probabilistic model for object recognition based primarily on local appearance. Local appearance is a strong constraint for object recognition when the object contains areas of distinctive detailing. For example, the human face consists of distinctive local regions such as the eyes, nose, and mouth. However, local appearance alone is usually not sufficient to recognize an object. For example, a human face becomes unintelligible to a human observer when the various features are not in the proper spatial arrangement. Therefore the joint probability of local appearance and position on the object must be modeled.

Nevertheless, representation of only the appearance of the object is still not sufficient for object recognition. Some local patterns on the object may be more unique than others. For example, the intensity patterns around the eyes of a human face are much more unique than the intensity patterns found on the cheeks. In order to represent the "uniqueness" of local appearance, the statistics of local appearance in the world at large must also be modeled.

The underlying representation we have chosen for local appearance is discrete. We have partitioned the space of local appearance into a finite number of patterns. The discrete nature of this representation allows us to estimate the overall statistical model, $P(\text{object}|image)$, in closed form by counting the frequency of occurrence of these patterns over various sets of "training" images.

In this paper we derive a functional form for the posterior probability function $P(\text{object}|image)$ that combines these representational elements. We then describe how we have applied this model to the detection of faces in frontal view and profile. We begin in section 2 with a review of Bayes decision rule. We then describe our strategy for deriving the functional form of the posterior probability function in section 3 and perform the actual derivation in section 4. In section 5, we describe how use training images to estimate a specific probability function within the framework of this functional form. In section 6 and 7 we give our results for frontal face detection and profile detection, respectively. In section 8 we compare our representation with other appearance-based recognition methods.

2. Review of Bayes decision rule

The posterior probability function gives the probability that the object is present given an input image. Knowledge of this function is all that is necessary to perform object recognition. For a given input image region, $x = image$, we decide whether the object is present or absent based on which probability is larger, $P(\text{object}|x)$ or $P(\overline{\text{object}}|x) = 1 - P(\text{object}|x)$, respectively. This choice is known as the maximum *a posteriori* (MAP) rule or the Bayes decision rule. Using this decision rule, we achieve optimal performance, in the sense of minimum rate of classification errors, if the posterior probability function is accurate.

3. Model derivation strategy

Unfortunately, it is not practically feasible to fully represent $P(\text{object}|image)$ and achieve optimal performance; it is too large and complex a function to represent. The best we can do is choose a simplified form of $P(\text{object}|image)$ that can be reliably estimated using the available training data.

Although a fully general form of $P(\text{object}|image)$ is intractable, it provides a useful starting point for derivation of a sim-

plified probabilistic model. In our derivation, we take this general form and apply successive simplifications to it until it is in a computationally feasible form. At each stage of this derivation we make our modeling decisions on the basis of domain knowledge and intuitive preferences.

This strategy of derivation provides an explicit record of all the representational simplifications made in deriving such a functional form. We then know not only those relationships we have modeled but those we have not modeled. For example, we make the implicit modeling decision not to represent the joint statistics of appearance across the full spatial extent of the object. This simplification along with various others become explicit through this derivation process.

4. Model derivation

In this section, we derive a functional form of the posterior probability function. This functional form was derived with the problem of frontal face detection in mind, but is generally applicable to a wider range of objects. For this reason, we describe the specific modeling choices we make for face detection after we have described the general nature of the simplification in each following section. Overall, we derive this functional form by applying approximately 13 simplifications and modifications to the general form of the posterior probability function. Variations on these modeling choices for face profile detection are described in section 7.

4.1. Notation

Throughout this document we make use of several notational conventions. Random variables are indicated in italics, e.g., *image*. When these random variables assume a specific value, the value is not italicized, e.g., $x = image$. Curly braces, {}, indicate aggregates. For example, $\{a_i\}$ represents all a_i: a_1, a_2, a_3, etc. We designate the class of all visual scenes that do not contain our object by the symbol $\overline{\text{object}}$.

4.2. General form of posterior probability function

The most general representation we consider is the posterior probability function of the object conditioned directly on the entire input image:

$$P(\text{object}|image) = P(\text{object}|pixel(1,1), pixel(1,2), \ldots, pixel(n,m)) \quad (1)$$

Where, $pixel(i,j)$ is the scalar intensity value (or color vector value for a color image) at location (i, j) in the *image*.

4.3. Size standardization

We first standardize the size of the object. Rather than model the object at all sizes simultaneously, we model the object at one standard size. This simplification allows us to express the posterior probability function conditioned an image *region* of fixed size, r_{reg} x c_{reg}:

$$P(\text{object}|region) = P(\text{object}|pixel(1,1), pixel(1,2), \ldots, pixel(r_{reg}, c_{reg})) \quad (2)$$

where, $pixel(i,j)$ is the scalar intensity value at pixel location (i, j) in the *region*.

In order to detect an object at any position in an image, we must then evaluate $P(\text{object}|region)$ for every overlapping *region* of this size within the image boundaries. Additionally, to detect the object at any size, we must repeat this process over a range of magnification scales of the original image.

We model faces that are normalized in size to 64x64. This size was chosen to be large enough to capture the detailed appearance of a human face.

4.4. Decomposition into class conditional probabilities

Using Bayes theorem, we can decompose the posterior probability function into the class conditional probabilities for the object, $P(region|\text{object})$, and non-object, $P(region|\overline{\text{object}})$, and the prior probabilities, $P(\text{object})$ and $P(\overline{\text{object}})$:

$$P(\text{object}|region) = \frac{P(region|\text{object})P(\text{object})}{P(region)} \quad (3)$$

where the unconditional probability of the image region, $P(region)$, is given by:

$$P(region|\text{object})P(\text{object}) + P(region|\overline{\text{object}})P(\overline{\text{object}}) \quad (4)$$

This decomposition allows us to separately estimate each of the class-conditional probability functions, $P(region|\text{object})$ and $P(region|\overline{\text{object}})$ from object and non-object training images, respectively. In the following sections we discuss how we simplify the functional forms for these probabilities.

Furthermore, using Bayes theorem, Bayes decision rule can re-written in an equivalent form as a likelihood ratio test:

$$\frac{P(region|\text{object})}{P(region|\overline{\text{object}})} \underset{\overline{\text{object}}}{\overset{\text{object}}{\gtrless}} \lambda = \frac{P(\overline{\text{object}})}{P(\text{object})} \quad (5)$$

Under this formulation we decide the object is present if the likelihood ratio (left side) is larger than the ratio of prior probabilities (right side). Otherwise we decide the object is not present.

Often we have little knowledge of the prior probabilities. By writing the decision rule this way all information concerning the priors is combined into one term, λ. This term can be viewed as a threshold controlling the sensitivity of the detector.

4.5. Decomposition into subregions

We decompose the input region into an aggregate of smaller *subregions* of fixed size, r_{sub} x c_{sub}:

$$region = \{subregion\} \tag{6}$$

where $subregion = (pattern, pos)$ contains two types of information: *pattern* - the array of pixel intensities over the subregion and *pos* - the subregion position with respect to the overall *region*. We consider all overlapping *subregions* within the larger *region*. For faces we use *subregions* of size 16x16.

With these modifications, the class conditional probability functions become:

$$\begin{aligned} P(region|\text{object}) &= P(\{subregion\}|\text{object}) \\ P(region|\overline{\text{object}}) &= P(\{subregion\}|\overline{\text{object}}) \end{aligned} \tag{7}$$

where there are n_{subs} *subregions* in a *region*.

We describe the advantages of this decomposition in sections 4.6.3 and 4.6.4.

4.6. No modeling of statistical dependency among subregions

We do not model statistical dependency among subregions. This simplification gives the following expression for the class conditional probability functions:

$$\begin{aligned} P(\{subregion\}|\text{object}) &\approx \prod_{k=1}^{n_{subs}} P(subregion_k|\text{object}) \\ P(\{subregion\}|\overline{\text{object}}) &\approx \prod_{k=1}^{n_{subs}} P(subregion_k|\overline{\text{object}}) \end{aligned} \tag{8}$$

Through these simplifications, our modeling requirements are reduced to representing $P(subregion_k|\text{object}) = P(pattern, pos|\text{object})$ and $P(subregion_k|\overline{\text{object}}) = P(pattern, pos|\overline{\text{object}})$ that describe the joint behavior of subregion appearance and position.

4.6.1. Model complexity reduction

The choice of not modeling the statistical dependency among subregions greatly reduces the complexity of the model. To illustrate the extent of this simplification, let us assume that a *region* is represented by an aggregate of *n subregions* and each subregion can take on *m* possible values describing its intensity *pattern*. The full statistical distribution for the object, $p(pattern_1, pattern_2, \ldots, pattern_n|\text{object})$, is then modeled over m^n discrete events. In contrast, the distribution, $p(pattern, pos|\text{object})$, is modeled over *mn* discrete events.

4.6.2. Loss of modeling power

Unfortunately, by not modeling this statistical dependency, there are many relationships we cannot represent. For example, we cannot represent attributes that are similar across the extent of the object, such as skin color on a human face. We cannot represent the structure in the brightness distribution across the object that is larger in extent than a subregion. For example, on human faces the forehead is usually brighter than the eye sockets [1]. We cannot represent any form of symmetry. We cannot represent if all parts of a geometric figure are connected [2].

However, this assumption does not impose a debilitating penalty for the problem of face detection because local features are salient and consistent among different faces, e.g., noses look relatively similar from individual to individual and appear in relatively the same position relative to the other facial attributes.

The application of this assumption to the recognition of other objects may not be as successful. In particular, it could be argued that many objects are more distinguished by overall structure rather than individual features. For example, on a modern building, windows are distributed in a regularly spaced arrangement against the uniform texture of the building material. The distinguishing characteristics are not the individual windows, nor the specific spacing of the window arrangement, but simply the presence of some form of regular window spacing.

4.6.3. Small alignment errors when matching

Using the subregion decomposition, we can accommodate some degree of geometric distortion in the appearance of the object. The alignment error between a full-size template and a rotated version of the template will be quite significant -- see figure 1. If we match individual subregions, the alignment error will be much less -- see figure 2. Similarly, subregion-based

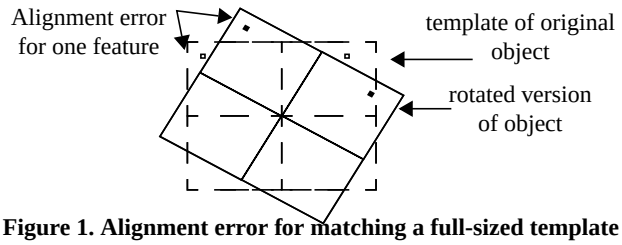


Figure 1. Alignment error for matching a full-sized template

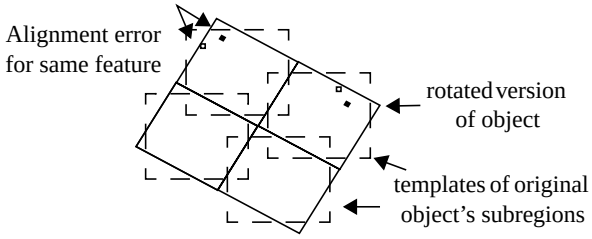


Figure 2. Alignment error for matching individual subregions

matching will reduce the alignment error for distortions in aspect ratio and magnification.

4.6.4. Emphasizing distinct parts of the object's appearance

The subregion decomposition provides a mechanism for emphasizing distinctive parts of the object's appearance over less distinctive parts. Let us consider the current expression for

the likelihood ratio:

$$\prod_{k=1}^{n_{\text{subs}}} \frac{P(subregion_{\text{k}}|\text{object})}{P(subregion_{\text{k}}|\overline{\text{object}})}$$

Distinctive areas on the object, *subregion*$_k$, will have a large value for $P(subregion_k|\text{object})/(P(subregion_k|\overline{\text{object}}))$ since the occurrence of these patterns is much more frequent on the object than in the world at large. Thus, such distinctive areas contribute more to the overall product given above.

4.7. Projection of the subregion intensity pattern

We linearly project the subregion intensity *pattern*, onto a lower dimensional space of dimension n_{pr}:

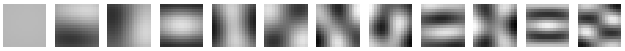
$$projection = A^T pattern \tag{9}$$

where *pattern* is rewritten as a column vector.

The columns of the projection operator, A, are chosen to be principal components computed from a sample of *subregions* collected from training images of the object.

A linear projection was chosen because of its computational efficiency using fast Fourier transforms. We chose principal components as basis functions because they minimize the mean square reconstruction error with respect the training set of object images.

For faces, we project the 16x16 subregion intensity *pattern* onto a 12 dimensional space. Below we show a set of principal components displayed as 16x16 arrays.



Overall, these principal components capture 98.4% of the total energy of their data set.

4.8. Sparse coding of projection

Typically, for any given pattern, its projection onto some of the eigenvectors will be negligibly small. Therefore, we apply sparse coding where for each pattern, we selectively represent only the n_{tr} largest responses among the n_{pr} eigenvectors.

For faces, we do not apply sparse coding to individual coordinates but instead to *groups* of coordinates. We first arrange the 12 coordinates into 9 different groups. Coordinates 1 through 6 are each assigned their own group. The remaining coordinates are grouped into pairs, 7 & 8, 9 & 10, and 11 & 12. This assignment of groups partially equalizes the amount of energy represented by each group. Then for each pattern, we select the response of the first group and the 5 groups that have the largest responses among the remaining 8 groups.

4.9. Discretization

We quantize the sparse coded representation into a finite number of patterns. Our expression for the class conditional probabilities is now given by:

$$\begin{aligned} P(pattern, pos|\text{object}) &\approx P(q1, pos|\text{object}) \\ P(pattern, pos|\overline{\text{object}}) &\approx P(q1, pos|\overline{\text{object}}) \end{aligned} \tag{10}$$

where $q1$ can take on n_{q1} discrete values and $q1 = Q_1(pattern)$ combines projection, sparse coding, and quantization.

For faces, each of the original coordinates in the projection are quantized to a finite number of levels between 3 and 8. Overall, $q1$ can take on n_{q1} = 3,854,120 different values. In figure 3 we illustrate how the successive operations of projection, sparse coding, and quantization affect the appearance of an image and the mean square pixel reconstruction error (MSE).

A. Original Image

B. Reconstruction from projection of A. MSE = 165.6

C. Reconstruction from sparse coded version of B. MSE = 179.2

D. Reconstruction from quantized version of C. MSE = 295.8.

Figure 3. Image reconstruction error.

4.10. Decomposition of appearance and position

We decompose the class conditional probabilities into the product of two distributions using the probability chain rule:

$$\begin{aligned} p(q1, pos|\text{object}) &= p(pos|q1, \text{object})p(q1|\text{object}) \\ p(q1, pos|\overline{\text{object}}) &= p(pos|q1, \overline{\text{object}})p(q1|\overline{\text{object}}) \end{aligned} \tag{11}$$

No further reduction is performed on $P(q1|\text{object})$ and $P(q1|\overline{\text{object}})$. In the following sections we describe the simplifications we use for representing $p(pos|q1, \text{object})$ and $p(pos|q1, \overline{\text{object}})$. Each of these distributions describes the positional distribution of each subregion intensity pattern, x = $q1$, within the overall *region*.

4.11. Positional representation

In images of non-objects, there are no stable landmarks from which we can define a *region*-based coordinate system. Therefore, we model the positional distribution as uniform:

$$p(pos|q1, \overline{\text{object}}) \approx \frac{1}{n_{\text{subs}}} \tag{12}$$

In representing the positional distribution for objects, $P(pos|q1, \text{object})$, we reduce the resolution of subregion position, by mapping *pos* to a new variable, *pos*', over a coarser resolution.

We also reduce the number of discrete patterns. In doing so, we first compute an estimate of $P(q1|\text{object})$ from the training data of face images (see section 5). We then select those n_{est} patterns that have the largest frequency of occurrence, where $n_{est} << n_{q1}$. For these values of $q1$ we explicitly

estimate and smooth the distribution $p(pos'|q1, \text{object})$. For the remaining n_{q1} - n_{est} values of $q1$, we model $p(pos'|q1, \text{object})$ as a uniform distribution.

To reduce the number of patterns further, we group together patterns whose smoothed distributions, $p(pos'|q1, \text{object})$, are similar. We use a simple clustering technique based on VQ [3] to form these groups of patterns that have similar positional distributions. The final form of this distribution becomes:

$$P(pos'|q1, \text{object}) \approx P(pos'|q2, \text{object}) \tag{13}$$

Where the reduction in the number of patterns is expressed by $q2 = Q_2(q1)$ which maps the set of n_{est} patterns to a smaller set of n_{q2} composite patterns, represented by the $q2$.

For faces, we reduce the positional resolution of subregions from 48x48 to 16x16. We estimate the spatial distribution for n_{est} = 300,000 of the original n_{q1} = 3.8M patterns. We then combine these patterns to form a smaller set of n_{q2} = 20,000 composite patterns. Below we show examples of $P(pos'|q2, \text{object})$, for four different patterns (values of $q2$):

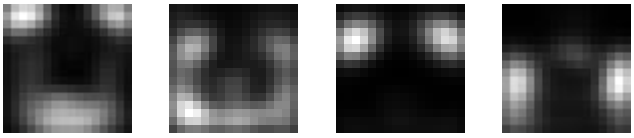


Figure 4. $P(pos'|q2, \text{object})$ **for several values of $q2$.**

4.12. Intensity normalization

We normalize the intensity over the entire *region* to have zero mean and unit variance. Since this normalization discards information about the mean and variance of original image region, it could be thought of as a small simplification to the posterior probability function.

Normalization reduces a known form of variation in the appearance of the object. By reducing this variation, we can obtain a better statistical estimate from a limited pool of training examples.

For faces, we compute the normalization coefficients only from the portion of the input region that contains the face. We perform this normalization separately on the left and right sides of the input *region*, to compensate for situations in which opposite sides of a face receive unequal amounts of illumination.

4.13. Multiresolution representation

We have only discussed the representation in the context of one level of resolution. This largely limits us to representing visual attributes that are the size of the subregion. To enhance our representation we consider multiple levels of resolution.

We form separate submodels of the class conditional probability functions, $P(region^j|\text{object})$ and $P(region^j|\overline{\text{object}})$, at several scales of resolution and we do not model the statistical dependencies among them. Thus, the expressions for the class condition probabilities become:

$$P(region|\text{object}) \approx \prod_{j=1}^{n_{magn}} P(region^j|\text{object})$$
$$P(region|\overline{\text{object}}) \approx \prod_{j=1}^{n_{magn}} P(region^j|\overline{\text{object}}) \tag{14}$$

where $region^j(u, v) = region(a_j u, a_j v)$ and a_j scales the region's resolution and there are n_{magn} scales of resolution.

For face detection, we use three levels of resolution given by, $a_1 = 1.0$, $a_2 = 0.577$, and $a_3 = 0.333$ as shown below in figure 5.

Figure 5. Scales of resolution used for face detection

4.14. Final form of Bayes decision rule

By fully substituting all simplifications of the class-conditional probability functions into equation (5), the overall expression for Bayes decision rule becomes:

$$\prod_{j=1}^{n_{magn}} \prod_{i=1}^{n_{subs}} \frac{P(q1_i^j|\text{object})P(pos_i^{j'}|q2_i, \text{object})}{P(q1_i^j|\overline{\text{object}})} \underset{\overline{\text{object}}}{\overset{\text{object}}{>\lessgtr}} \lambda = \frac{P(\overline{\text{object}})}{P(\text{object})} \tag{15}$$

5. Estimation

Equation (15) gives the final expression for the functional form of the likelihood ratio. We now use labelled training examples to estimate a specific likelihood ratio function within the structure of this functional form.

5.1. Training set for frontal face detection

We formed training sets from 991 faces images and 1,552 non-face images. We used the same set of images to train each of the level of resolution within the model. The magnification of these images is scaled appropriately for each resolution level.

To partially compensate for the limited number of face images, we expanded this training set by generating synthetic variations of these images. For each face image we generated *120* synthetic variations in orientation, size, aspect ratio, intensity, and background scenery.

5.2. Method of estimation

We break the estimation of the likelihood function into several components. For each scale of resolution j, we first estimate $P(q1^j|\text{object})$ and $P(q1^j|\overline{\text{object}})$ directly from face and

non-face training images, respectively. The estimates of these functions are then substituted directly into equation (15). We then estimate $P(pos^{j'}|q1^{j}, \text{object})$ from the face training images for n_{est} values of $q1$. We then derive $P(pos^{j'}|q2^{j}, \text{object})$ from $P(pos^{j'}|q1^{j}, \text{object})$ by the procedure outlined in section 4.11. $P(pos^{j'}|q2^{j}, \text{object})$ is then substituted into equation (15) giving us the complete estimate for the likelihood ratio function.

There are several principles that are common to the estimation of $P(q1^{j}|\overline{\text{object}})$, $P(q1^{j}|\text{object})$ and $P(pos^{'j}|q1^{j}, \text{object})$. Their estimates are computed in closed form. For $P(q1^{j}|\overline{\text{object}})$ and $P(q1^{j}|\text{object})$, we simply count how frequently each value of $q1$ occurs in the training data, using non-face images and face images respectively. Then for $P(pos^{j'}|q1^{j}, \text{object})$,we count how frequently each pattern, x = $q1$, occurs at each position *pos*‘ in the image region. These are maximum likelihood estimates and they are unbiased, consistent, and efficient (satisfy Cramer-Rao lower bound)[8].

6. Testing results for frontal faces

Each row in table 1 shows our performance for a different value of a detection threshold, λ' (closely related to the threshold λ given in equation (15)), on the test set of Sung and Poggio [4] (136 faces) excluding 3 images of line drawn faces. We searched for all faces between the sizes of 18x18 and 338x338 by evaluating each input image at 17 levels of magnification.

Table 1: Results on images from [4]

Schneiderman & Kanade (20 images)		Sung and Poggio [4](23 images)	
Detection rate	False alarms	Detection rate	False alarms
91.2%	12	84.6%	13
89.0%	3	79.7%	5

Similarly, table 2 shows our performance on the combined test sets of Sung and Poggio [4] and Rowley, Baluja, and Kanade [5](483 faces) excluding 5 images of line drawn faces. We searched each input image at 17 levels of magnification for faces from size 18x18 to 338x338.

Table 2: Results on images from [4] and [5].

Schneiderman & Kanade(125 images)		Rowley, Baluja, and Kanade [5] (130 images)	
Detection rate	False alarms	Detection rate	False alarms
93.0%	88	92.5%	862
90.5%	33	86.6%	79
77.0%	1	77.9%	2

Figure 6. Our results on a test image from [4]

Table 3 shows our performance on three portions of the FERET[7] face set consisting of subsets of 1000, 241, and 378 face images at profile angles of 0° (full frontal), 15°, and 22.5°, respectively. We searched each input image at 14 levels of magnification for faces from size 22x22 to 235x235.

Table 3: Results on FERET[7] images

	Schneiderman & Kanade		Rowley, Baluja, and Kanade[5]	
Data set	Detection rate	False alarms	Detection rate	False alarms
0° set	99.6%	1	98.7%	3
15° set	100.0%	0	99.6%	0
22.5° set	99.7%	2	95.5%	3

Moghaddam and Pentland [6] achieve a detection rate of 97%. on this test set. False alarm data was unreported.

7. Face profile detection

The same theory has been tested for face profile detection. There are several significant differences between our algorithm for profile detection and our algorithm for face detection. For profile detection we do not perform intensity normalization. Instead of measuring absolute intensity across the subregion (i.e. projection on to the first eigenvector for frontal faces), we measure the difference in intensity between a subregion and its neighboring subregions. This intensity information is quantized into 3 levels. We then combine this intensity information with result of sparse coding. In sparse coding we select the 4 largest projections among the 12 remaining eigenvectors. Instead of quantizing these selected responses, we simply indicate which group of 4 responses was selected, the sign of each

individual component, and which component is largest.

Below we show some preliminary results acquired for a fixed value of the detection threshold. The double bar indicates the front of the face:

8. Appearance-based methods for recognition

The representation described in this paper combines: joint statistics of local appearance and position on the object, statistics of local appearance in the world at large, a discrete non-parametric probability distribution, and estimation by counting the frequency of occurrence of a finite set of patterns in the training data. Many other methods share one or two of these concepts, but none, to our knowledge, have combined all of them.

In particular, our method differs significantly from appearance-based methods that emphasize global appearance over local appearance. For example, the methods [4], [9], [10], [11], model the full extent of the object at once. In particular the methods [4], [10], [11] implicitly give equal weighting to distinctive and non-distinctive areas on the object.

There are several methods [5], [6], [12],[13], [14], which capture the joint variation of local appearance and position on the object. These methods all differ from our approach in that they model the appearance of hand-selected features on the object rather than modeling local appearance across the full extent of the object. [5] captures local appearance through a multilayer perceptron architecture with hidden units that have localized support regions. However, this architecture rigidly fixes the spatial relationships of these localized receptive fields. [14] uses a Gaussian distribution to model the spatial variation in feature location. Their model of the non-face statistics is chosen completely by hand. [6] uses a mixture of Gaussians to model the statistics of the local features on a face and does not model the statistics of non-face appearance. The methods of [12] and [13] both reduce the dimensionality of the local regions by projection onto the principal components. Recognition is then performed by comparing a set templates representing the object to a set of image regions at the appropriate spacing as specified by the object model.

The method of [15] uses a discrete representation and estimation method similar to ours except they apply it to color rather than local appearance. We choose a discrete, non-parametric, representation of the probability distribution function because it greatly simplifies the estimation problem. Estimation of multimodal continuous parametric distributions (e.g. mixture models, multilayer perceptrons) is usually not possible in closed form and requires iterative estimation procedures which are not guaranteed to converge to a global optimum. Continuous valued non-parametric methods such as nearest neighbor and Parzen windows require storing all training examples and exhaustive comparison of training examples to each input. Because such methods require large training sets for even moderately high dimensional spaces they are prohibitive in storage and computational requirements.

References

[1]. M. Oren, et. al. “Pedestrian Detection Using Wavelet Templates.” CVPR, ‘97. pp. 193- 199.

[2]. M. Minsky and S. Papert. *Perceptrons: an Introduction to Computational Geometry* (expanded ed.) MIT Press. Cambridge, MA, 1988.

[3]. A. S. Pandya and R. B. Macy. *Pattern Recognition with Neural Networks in C++*. CRC Press. Boca Raton, FL. 1996.

[4]. K-K Sung, T. Poggio. “Example-based Learning of View-Based Human Face Detection.” ACCV ‘95 and AI Memo #1521, 1572, MIT.

[5]. H. Rowley, S. Baluja, T. Kanade. “Neural Network-Based Face Detection.” PAMI 20(1), January, 1998.

[6]. B. Moghaddam and A. Pentland. “Probabilistic Visual Learning for Object Representation.” PAMI, 19(7). pp. 696 - 710. July, 1997.

[7]. P. J. Phillips, et. al. “The FERET Evaluation Methodology for Face-Recognition Algorithms.” CVPR ‘97. pp. 137 - 143.

[8]. B. V. K. Vijaya Kumar. Lectures on Pattern Recognition. In publication.

[9]. D. Casasent and L. Neiberg. “Classifier and Shift-invariant Automatic Target Recognition Neural Networks.” Neural Networks. vol. 8, pp. 1117-1129, 1995.

[10]. H. Murase and S. Nayar. “Visual Learning and Recognition of 3D Objects from Appearance.” IJCV. 14(1), 1995, pp.5-24.

[11]. E. Osuna, R. Freund, F. Girosi. “Training Support Vector Machines: an Application to Face Detection.” CVPR ‘97. pp.130-136.

[12]. J. Krumm. “Eigenfeatures for Planar Pose Measurement of Partially Occluded Objects.” CVPR ‘96. pp. 55-60.

[13]. K. Ohta and K. Ikeuchi. “Recognition of Multi-Specularity Objects using Eigen-Window. Tech. Rep. CMU-CS-96-105. Feb. ‘96.

[14]. M. Burl and P. Perona. “Recognition of Planar Object Classes.” CVPR ‘96. pp. 223-230.

[15]. M. Swain, D. Ballard. “Color Indexing.” IJCV. 7(1):11-32. 1991.