Graded Learning for Object Detection

Francois Fleuret\textsuperscript{1} and Donald Geman\textsuperscript{2}

Abstract

Our goal is to detect all instances of a generic object class, such as a face, in greyscale scenes. The design of the algorithm is motivated by computational efficiency. The search is coarse-to-fine in both the exploration of poses and the representation of the object class. Starting from training examples, we recursively learn a hierarchy of spatial arrangements of edge fragments, graded by their size (sparsity). The arrangements have no a priori semantic or geometric interpretation. Instead, they are selected to be "decomposable": Each can be split into two correlated subarrangements, each of which can be further divided, etc. As a result, the probability of an arrangement of size \( k \) appearing on an object instance decays slowly with \( k \). We demonstrate this both theoretically and in experiments in which detection means finding a sufficient number of arrangements of various sizes.

\textsuperscript{1} Avant-Projet IMEDIA, INRIA-Rocquencourt, Domaine de Voluceau, B.P.105, 78153 Le Chesnay. Supported in part by the CNET.

\textsuperscript{2} Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003; Supported in part by the ONR under contract N00014-97-1-0249 and the ARO under MURI grant DAAH04-96-1-0445.
1 Introduction

Starting with a training set of examples of a generic object class (e.g., “face”), our goal is to construct an algorithm to detect and localize all instances of this class in greyscale scenes. The examples are subimages containing a single instance of the object at various poses, for example frontal views of faces at a range of scales, tilts, etc. Whereas the “backgrounds” in the training samples might be very simple, the detection algorithm must function in natural, highly cluttered scenes. We measure performance by the false alarm rate and the size of the training set necessary to achieve a very small false negative rate.

This work belongs to an ongoing and broader project on visual recognition as a “twenty questions game” - a problem in efficient coding. We design features (image functionals) and search strategies at the same time and in a highly coarse-to-fine manner in both (i) the density of the object representations and (ii) the exploration of the space of poses. This paradigm was analyzed in the context of decision trees and stepwise uncertainty reduction in [2], [8], [9] and [16]. Although the approach to learning is different here, and no trees are induced, the final algorithm can be expressed as a huge, recursively coded tree of binary “queries” based on comparisons of intensity differences.

We learn a hierarchy of increasingly complex features. In order to declare detections, we successively check for a minimal number of each complexity. As in [3], the features are binary image functionals corresponding to the presence or absence of spatial arrangements of edge fragments. The fragments have an approximate location and an approximate orientation; the definition is purposely loose in order to accommodate geometric invariance. The arrangements have no a priori semantical or geometric interpretation. Instead, we simply want them to be as likely as possible on the object class. Specificity alone renders them rare in general backgrounds.

We introduce the notion of a “decomposable arrangement” and as well as an algorithm for discovering large numbers of these from training examples. This is the main contribution of the paper. “Decomposable” means that the arrangement can be divided into two highly correlated subarrangements, each of which further splits into two highly correlated smaller arrangements, and so forth all the way down the individual edge fragments. The learning algorithm is a procedure for building larger and larger arrangements in a recursive, bottom-up fashion. The motivation is that the probability of a decomposable arrangement of size $k$ appearing on an object instance decreases very gradually as $k$ increases, thus insuring features which are highly discriminating in separating objects from background. This statement will be justified theoretically and illustrated empirically.

In order to emphasize the new ideas and the statistical learning framework we shall only apply the learning algorithm to training examples in one particular subset of poses. The pose of a face means the shape and location of the triangle formed by the two eyes and the mouth. We will indicate how exactly the same routine is repeatedly applied to a series of successively more constrained learning sets (groups of triangles) in order to detect faces over a wide range of presentations; details and experiments are in [7]. In the end the processing is coarse-to-fine in both the subdivisions of pose space and the search for successively denser arrangements. The predecessor of this approach, demonstrated in [3], is to train at a single reference pose and then relax the geometrical constraints to accommodate other poses, using a generalized Hough transform to minimize computation.

2
Two experiments are described. One is synthetic - detecting “wings” in a world of lines. It is very challenging because, by design, the wings are locally indistinguishable from the “clutter” (diagonal lines). The other experiment is real - detecting frontal views of faces. We use the Olivetti training set of 300 faces: 10 pictures of each of 30 individuals. A relatively small training is sufficient since we only use it to estimate correlations, and training only requires a few minutes on a PC. In particular, we do not estimate a large system of coupled parameters as in other statistical learning methods. Also, invariance and other properties are largely “hard-wired” rather than learned.

2 Invariant Object Detection

The problem of detecting instances from a generic object class without information due to color, depth or motion has been widely studied in the computer vision literature. For example, in the case of faces, a variety of methods have been proposed, including artificial neural networks [12], [13], support vector machines [11], graph-matching [10], Bayesian inference [5], deformable templates [17], and the precursors of our methodology already cited.

One of the principal difficulties is the variation in the appearance of faces due to the vagaries of lighting; see for example the discussion in [14]. Our approach is to build a large degree of photometric invariance into the definition of the edge fragments by considering only comparisons of intensity differences; see Section 4. Similarly, our approach to geometric invariance is quite explicit. Each arrangement is a conjunction of elementary tests, which are themselves disjunctions of edge fragments; we then compare the number of arrangements present to a threshold. Detection is then finally based on a massive disjunction of conjunctions of highly localized edge fragments, a very literal treatment of geometric invariance and in sharp contrast to most other approaches; see [3] for a more complete discussion.

3 Statistical Framework

Let $I$ denote the set of (sub)images $I = \{I(x), x \in R\}$, where $R$ is a reference grid (e.g., $32 \times 32$) and $I(x)$ is quantized in a standard way, say to 256 grey levels. The subimages are divided into two categories - “object” and “background.” The object pictures contain a single instance of the object class at a reference pose, meaning that the object is roughly centered in the subimage and the scale of the object is roughly that of the grid $R$. The background pictures are everything else. Let $Y(I) \in \{0, 1\}$ stand for the class of $I$ with “0” corresponding to background.

Let $P$ denote a probability measure on $\mathcal{I}$. We can think of $P$ as the empirical measure on all subimages of all images in some large database. Then $P$ induces two conditional measures: $P_0(.) = P(. | Y = 0)$, the distribution on the background class and $P_1(.) = P(. | Y = 1)$, the distribution on the object class. Given a classifier $f : \mathcal{I} \rightarrow \{0, 1\}$, the false negative error rate is $\alpha(f) = P_1(f = 0)$ and the false positive error rate is $\beta(f) = P_0(f = 1)$. Ideally, we seek to minimize $\beta(f)$ subject to $\alpha(f) = 0$; this is part of the motivation for the classifier we define in Section 5.
The training set $\mathcal{L}$ is assumed to be a random sample from $\mathcal{I}$ under $P_1$. We use negative examples (i.e., samples from $P_0$) to estimate false alarm rates but not to construct the classifier. An important constraint is that the size of $\mathcal{L}$ is not sufficiently large to reliably estimate a number of inter-dependent parameters of the same order. Moreover, whereas the training set is a genuine random sample in the synthetic problem, this will rarely be the case in real problems.

Since our approach is inductive rather than deductive, we do not propose a model for either $P_0$ or $P_1$; instead, we rely on the empirical measure $\hat{P}_1$ induced from $\mathcal{L}$. Thus, we define a classifier directly from $\mathcal{L}$. By and large, training then amounts to estimating the probability distribution under $P_1$ of events in $\mathcal{I}$; these probabilities determine the components of the classifier and are derived by relative frequencies in $\mathcal{L}$. Another possibility would be to estimate both $P_0$ and $P_1$ and take corresponding the maximum likelihood classifier, but this appears to be impractical; the background distribution is extremely difficult to model and computing the maximum likelihood estimator is only feasible in simple cases.

4 A Hierarchy of Features

We seek a coarse-to-fine hierarchy of features, ranging from primitive (sparse and local) to complex (dense and global), whose statistics in the two populations become increasingly disparate. Computational efficiency is achieved by progressively checking for more and more complex attributes in order to separate instances of clutter from instances of object. Many subimages can be immediately dismissed as object candidates based on the primitive features, such as edge counts; more global confusions will require further examination involving higher-order correlations and longer-range dependencies.

4.1 Elementary Tests

All our features are conjunctions of binary, elementary tests. The elementary tests are local disjunctions of localized filters. In our experiments the local filters detect edge fragments; other, more sophisticated, filters might be more effective. Each edge filter is applied at each location in $R$, and has an orientation (horizontal or vertical) and a polarity (plus or minus). Consider a horizontal edge of positive polarity at location $x$. This means that $I(x) > I(x')$, where $x$ and $x'$ are vertical neighbors, and that $I(x) - I(x') > \max_y \{|I(x) - I(y)|, |I(x') - I(y)|\}$ for a certain number of neighboring pixels $y$. In Figure 6 we show a training face (left) together with the detected edge fragments (middle).

There is one elementary test $X_i = X_i(I)$ for each edge fragment, $i = 1, 2, ..., d$, where $d \approx 4|R|$. Test $X_i = 1$ if the corresponding edge is present in a small neighborhood of $x_i$ and $X_i = 0$ otherwise. The size of the neighborhood (the degree of "floating") depends on the subset in pose space that defines the learning problem; it is chosen to make the probability of the elementary tests of order $\frac{1}{y}$. 

4
4.2 Decomposable Arrangements

We refer to a subset $A \subset \{1, \ldots, d\}$ as an arrangement; the corresponding random variable

$$X_A(I) = \prod_{i \in A} X_i(I)$$

on $I$ is simply a spatial conjunction of elementary tests: $X_A = 1$ if and only if $X_i = 1$ for each $i \in A$. Let $\text{supp} X_i \subset R$ be the set of pixels which appear in the definition of $X_i$. In order to limit the family of arrangements we shall assume that $\text{supp} X_i \cap \text{supp} X_j = \emptyset$ whenever $i, j \in A$ and $i \neq j$. We write $|A|$ for the size of $A$. This is our pool of features; the classifier will be constructed from a subset of these.

We want to find arrangements $A$ for which the statistics of $X_A$ are as different as possible under $P_0$ and $P_1$. Since estimation under $P_0$ is problematic, we will attempt to obtain the desired disparity by constructing arrangements which are large but still likely under $P_1$. Size alone renders them rare under $P_0$. The construction is based on correlation. Let $\rho(U, V)$ denote the correlation coefficient of random variables $U$ and $V$ with respect to $P_1$. For binary variables with $0 < P_1(U = 1), P_1(V = 1) < 1$ we have

$$\rho(U, V) = \frac{P_1(U = 1, V = 1) - P_1(U = 1)P_1(V = 1)}{(P(U = 1)P(U = 0)P(V = 1)P(V = 0))^{1/2}}.$$

Consider arrangements $X_iX_j$ of size two. We could filter all such pairs by requiring that $\rho(X_i, X_j) \geq \rho$ for some threshold $\rho$, $0 < \rho < 1$. This yields pairs of elementary tests which tend to occur (or not occur) together on objects. Similarly, $X_iX_jX_k$ might be a good candidate for a discriminating arrangement of size three if, in addition, $\rho(X_iX_j, X_k) \geq \rho$. Continuing in this way, we can single out arrangements of size four by combining two “good” pairs $X_iX_j$ and $X_kX_l$ and further requiring that $\rho(X_iX_j, X_kX_l) \geq \rho$. And so forth.

Define a decomposition of $A$ to be any nested set of binary partitions (i.e., successive binary refinements) all the way down to individual elements of $\{1, 2, \ldots, d\}$. We shall also assume that a partition cell splits evenly if its size is even and splits into two child cells whose sizes differ by exactly one if its size is odd. Call it a $\rho$-decomposition if the correlation inequality holds at every split. In Figure 1 we show one decomposition of $A = \{1, 2, 4, 5, 9\}$. It is a $\rho$-decomposition if $\rho(X_1X_4, X_2X_5X_9) \geq \rho$, $\rho(X_1, X_4) \geq \rho$, $\rho(X_5X_9, X_2) \geq \rho$ and $\rho(X_5, X_9) \geq \rho$. Finally, an arrangement $A$, or the corresponding random variable $X_A$, will be called $\rho$-decomposable if there is at least one $\rho$-decomposition of $A$. Summarizing,

**Definition:** An arrangement $A$ is $\rho$-decomposable if it is an elementary feature or if exists two $\rho$-decomposable arrangements $B$ and $C$ with

- $A = B \cup C$, $B \cap C = \emptyset$
- $|B| - |C| \leq 1$
- $\rho(X_B, X_C) \geq \rho$

5
Figure 1: A \( \rho \)-decomposable arrangement can be broken down recursively into conjunctions of positively correlated smaller arrangements.

4.3 A Likelihood Bound

In general \( P_0(X_A = 1) \) and \( P_1(X_A = 1) \) depend on \( A \) and decrease as \(|A|\) increases. A reasonable assumption for \( P_0 \) is some type of exponential decrease, and indeed this is what we observe empirically. On the other hand, if \( A \) is a \( \rho \)-decomposable arrangement (\( \rho da \)), we might expect a slower rate of decrease under \( P_1 \). In fact, we can show that the rate of decrease is \( \rho^{\log_2 k} \). As a result, for “reasonable” values of \( \rho \), \( P_1(X_A = 1) \gg P_0(X_A = 1) \) for “large” \( A \). We cannot, however, say anything precise about the likelihood ratio since we do not propose a model for \( P_0 \). But we can give a precise lower bound on \( P(X_A = 1) \).

Let \( A(k, \rho) \) denote the set of all \( \rho da \)’s with \(|A| = k\).

**THEOREM:** For any \( k \geq 1 \), \( \rho > 0 \) and \( A \in A(k, \rho) \),

\[
P_1(X_A = 1) \geq \min_{1 \leq i \leq d} P_1(X_i = 1) \cdot \rho^{\log_2 k}.
\]

(1)

**Sketch of the Proof:** Let \( \xi = \min_{1 \leq i \leq d} P_1(X_i = 1) \) and write \( k = 2^l + j \) where \( l \geq 0 \) and \( 0 \leq j < 2^l \). It can be shown by an induction argument on \( k \) that

\[
P_1(X_A = 1) \geq \xi \cdot \rho^{l + j 2^{-l}}
\]

(2)

Then (2) implies (1) by the concavity of the function \( u \to \log_2 u \). The get the flavor of the induction step, suppose (2) is true for \( k = 1, 2 \) and consider the case \( k = 3 \). Then \( l = j = 1 \) and \( l + j 2^{-l} = \frac{3}{2} \). For binary \( U \), let \( v(U) = P_1(U = 1)P_1(U = 0) \), the variance.

Let \( A = \{a, b, c\} \) and suppose \( A \to \{a, b\} \cup \{c\}, \{a, b\} \to \{a\} \cup \{b\} \) is a \( \rho \)-decomposition. Then,

\[
P_1(X_a X_b X_c = 1) \geq P_1(X_a X_b = 1)P_1(X_c = 1) + \rho \left[ v(X_a X_b)v(X_c) \right]^{1/2}
\]

(3)

\[
\geq \rho \xi \cdot \xi + \rho \left[ \rho \xi (1 - \rho \xi) \xi (1 - \xi) \right]^{1/2}
\]

(4)

\[
\geq \xi^2 \rho + \rho \left[ \rho \xi^2 (1 - \xi)^2 \right]^{1/2}
\]

(5)

\[
= \xi \rho 3^2
\]

(6)

Inequality (3) is due to \( \rho(X_a X_b X_c) \geq \rho \), (4) holds because \( x \to x(1 - x) \) is increasing on \( 0 \leq x \leq 1/2 \) and (5) is implied by \( \rho \xi \leq \xi \).
The Model Classifier

Detection can be viewed as a sequence of hypothesis tests of increasing complexity relative to the null hypothesis of “background”; each test is applied only when all simpler ones have rejected the null. Since the overwhelming majority of subimages examined are in fact background, the result is a highly coarse-to-fine process in which very few subimages are investigated in detail.

Fix \( \rho \). Each hypothesis test is based on the number \( Z_{k, \rho} \) of \( \rho \mathrm{da} \)'s of size \( k \) present in the picture \( I \):

\[
Z_{k, \rho}(I) = \sum_{A \in A(k, \rho)} X_A(I).
\]

Let \( K \) be the largest size \( k \) which “covers” the object class in the sense that \( P_1(Z_{k, \rho} \geq 1) = 1 \). (In our experience it never happens that arrangements of size \( k \) cover but arrangements of size \( j < k \) do not.) Given thresholds \( \{c_1, \ldots, c_K\} \), we classify \( I \) as object if it contains more than \( c_k \) \( \rho \mathrm{da} \)'s for each \( k = 1, \ldots, K \). In other words, the classifier is

\[
f_\rho(I) = \prod_{k=1}^{K} 1\{Z_{k, \rho}(I) \geq c_k\}.
\]

The thresholds \( c_1, \ldots, c_K \) are defined by \( c_k = \max\{j : P_1(Z_{k, \rho} \geq j) = 1\} \). That is, they are the maximum values which yield \( \alpha(f_\rho) = 0 \). We implement \( f_\rho \) as a series of filters, as depicted in Figure 2.

Finally, the natural choice of \( \rho \) is the value minimizing the false positive error:

\[
\rho^* = \arg \min_\rho \beta(f_\rho).
\]

However, we have not yet performed any systematic exploration of the possible \( \rho \) values; in fact, from here on we assume \( \rho \) is fixed. In our experiments we take \( \rho = .1 \).

Learning

In practice we cannot construct \( f_\rho \) because we don’t have the sets \( A_{k, \rho}, k = 1, \ldots, K \). They require knowing \( P_1 \) and there might be a great many of them. Instead, our experiments
are based on an approximation. We cannot hope to estimate all the pdf's, so we attempt to find a fixed number \( n \) of these for each size \( k \leq K \). Thus, given \( \mathcal{L} \), one goal of learning is to estimate a subfamily of \( \mathcal{A}_C(k, \rho) \subseteq \mathcal{A}(k, \rho) \) of size \( n \) for each \( k \leq K \). The other learning task is to estimate the thresholds \( c_1, ..., c_K \).

Whereas the definition of a decomposable arrangement is top-down, the estimation procedure is bottom-up. Correlations are estimated under \( \hat{P}_1 \), the empirical measure derived from \( \mathcal{L} \). The construction is recursive: First build a family \( \{X_iX_j\} \), then a family \( \{X_iX_jX_k\} \), etc. In order to construct \( \mathcal{A}_C(2k, \rho) \) we only need \( \mathcal{A}_C(k, \rho) \); and for \( \mathcal{A}_C(2k+1, \rho) \) we only need \( \mathcal{A}_C(k, \rho) \) and \( \mathcal{A}_C(k+1, \rho) \).

Consider the even case. Let \( \mathcal{A}'_C(2k, \rho) \) be the set of all arrangements \( A_1 \cup A_2 \) where

- \( A_1, A_2 \in \mathcal{A}_C(k, \rho) \);
- \( \hat{\rho}(X_{A_1}, X_{A_2}) \geq \rho \);
- \( \text{support} X_{A_1} \cap \text{support} X_{A_2} = \emptyset \).

Here, \( \text{support} X_A = \cup_{i \in A} \text{support} X_i \).

We want to select a subset of \( \mathcal{A}'_C(2k, \rho) \) of size \( n \), if possible. Generally, \( n \ll |\mathcal{A}'_C(2k, \rho)| \ll n^2 \). If \( |\mathcal{A}'_C(2k, \rho)| \leq n \), put \( \mathcal{A}_C(2k, \rho) = \mathcal{A}'_C(2k, \rho) \). Otherwise, \( \mathcal{A}_C(2k, \rho) \) is randomly sampled from \( \mathcal{A}'_C(2k, \rho) \), one arrangement at a time, among all those which are present on those training example(s) which are covered by the least number of already chosen arrangements. In other words, the random sampling attempts to maximize the quantity \( \min_{\omega \in \mathcal{L}} \left\{ \sum_{A \in \mathcal{A}_C(2k, \rho)} X_A(\omega) \right\} \).

The recursive process is initialized with \( \mathcal{A}'_C(1, \rho) \), the \( n \) most common elementary tests, and terminates at the first \( k \) for which it is no longer possible to cover the object examples with \( n \) arrangements of size \( k \).

Finally, the natural estimators of the thresholds \( c_1, ..., c_K \) are

\[
\hat{c}_k = \max \left\{ c : \hat{P}_1 \left( \sum_{A \in \mathcal{A}_C(\rho, k)} X_A \geq c \right) = 1 \right\}, \quad k = 1, ..., K.
\]

Clearly, this over-estimates \( c_k \), in fact rather severely in many cases. Therefore, we take one-half this value in all our experiments; this yields \( \alpha(f) = 0 \), a fundamental constraint in our formulation.

7 Experiments with Wings

The experiment was designed to exaggerate problem of "clutter" - background structures which resemble the targets. We also wanted to evaluate the proposed detector and the learning algorithm in a controlled statistical setting. This means that \( \mathcal{L} \) is truly a random sample from \( P_1 \) and that the error rates \( \alpha(f) \) and \( \beta(f) \) of any proposed classifier, including maximum likelihood, can be estimated to high precision.
Figure 3: A “wing world” scenario.

Figure 4: “Wings” consist of two random lines of equal length starting from a common pixel.

7.1 Scenes

“Scenes” are binary and composed entirely of short, jagged lines; see Figure 3. The lines ("clutter") are realizations of random walks; the transition probabilities orient the segments either northwest or northeast. Scenes are then generated as follows: Randomly choose a set of locations $x_1, x_2, \ldots$ in the integer lattice, then randomly choose a direction, either NW or NE, for each location $x_i$ and generate a random walk of length $m$ starting at $x_i$. In the NW case, the walk proceeds either W, NW or N with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ respectively; in the NE case these directions are N, NE or E. The locations of the vertex of the “wings” are also randomly selected. Wings consist of two diagonal lines, one NW and one NE, starting from the same pixel and constrained by the symmetry condition illustrated in Figure 4.

The detection problem is difficult because every object instance is a distinguished configuration of two instances of clutter. Wings and background are inseparable based solely on local features; long-range correlations must be taken into account. On the other
hand, it is not necessary to consider very dense representations, such as a complete wing template, since a few black pixels oriented NW or NE already give ample evidence for a line segment. Finally, it is by no means obvious how a learning algorithm can “discover” the symmetry constraint or accommodate the large variation among wing instances.

### 7.2 Training and Testing

We generated a large scene and randomly extracted 1000 16 × 16 subimages centered at wing locations; thus |$\mathcal{L}$| = 1000. The optimal error rate seems to be around 2%, obtained by knowing the model geometry and checking for two diagonal lines which have the correct length and verify the symmetry constraint at the endpoints. This describes the maximum likelihood classifier

$$f_{\text{ML}}(I) = I_{\{P_1(I) > P_0(I)\}}$$

because $P_1(I) > P_0(I)$ for any wing subimage $I$ and because $P_1(I) = 0$ for any background subimage. Since every wing is found, the errors are all false positives.

In our classifier, the elementary tests are not the individual pixels, black or white. Instead, they are edge-based. “Edges” are two black pixels diagonally adjacent. Every elementary test $X_i$ has a location $x_i$ in the 16 × 16 grid and an orientation $\theta_i \in \{NW; NE\}$; $X_i = 1$ if there is an oriented edge of type $\theta_i$ anywhere in a $2 \times 2$ neighborhood of $x_i$. Such tests make the classifier stable with respect to small perturbations of the wing shape, but still precise enough to exploit the symmetry property.

The arrangements which are estimated from $\mathcal{L}$ accumulate on the wings and capture their shape. For small $A$ the locations are usually confined to one side, but for larger $A$ the longer range dependencies are exploited and $A$ contains pixels on both lines. And the probability distributions of the random variables $Z_k$ under $P_0$ and $P_1$ are indeed quite separated; the shapes of the cumulative distribution functions look rather similar to those for faces, shown in Figure 7. Finally, the best combined error rate we achieve is $\alpha + \beta \approx .07$; it is somewhat higher if we implement the constraint that $\alpha = 0$. In our view, separating these two populations is very difficult. For example, nearest-neighbors works poorly, even with “smart” (e.g., invariant) metrics.
8 Experiments with Faces

The pose is characterized by a triangle whose vertices are the positions of the two eyes and the mouth. Since our goal is to detect upright, frontal views of faces, the triangles point down and are basically isosceles. There are then five degrees of (linear) freedom: two for location, and one each for orientation, scale (e.g., the distance between the eyes) and elongation (e.g., the height-to-width ratio). We do not search for faces at all scales simultaneously. Rather, we first assume that all faces are within about ±25% of a minimal reference scale and then find the larger ones by re-running the algorithm on downsampled scenes.

The triangles are recursively partitioned. There is one learning problem per partition, but the algorithm is the same; only the training set changes. For example, we might first group together all poses whose location falls in a particular $4 \times 4$ window. The corresponding training set is then quite heterogeneous since all the scales, orientations, and elongations are mixed together and the location is only approximate. The first step in processing the scene would then be to evaluate the resulting classifier $f$ on the subimage $I$ around each non-overlapping window. Only the pixels in windows for which $f(I) = 1$ would be retained as possible face locations. Naturally, there are many false alarms because only simple, sparse arrangements can cover the training ensemble. However even a false positive rate of 50% means that half the scene is finished with very elementary processing, perhaps involving only arrangements of size one (i.e., edge statistics).

The next division might divide scales into “small” and “large.” This leads to two new, more restricted, learning problems. (Actually, it is enough to “expand” the arrangements learned for the small scale.) The training set is now more homogeneous and hence larger, more discriminating arrangements are found. We continue in this way, steadily restricting the amount of variation in the triangles. Eventually we might constrain the location to a $2 \times 2$ window and pin all other the parameters at one “reference pose” (triangle shape). In some sense, the limit of this process is a very global classifier, such as a deformable template [1], which might involve on-line optimization in order to refine nonlinear aspects of the pose and allow for face recognition.

8.1 Training Set

We will illustrate learning by fixing the shape of the triangle and locating it anywhere in a $4 \times 4$ window; as indicated above, this might be followed by further restricting the location.

The training set $\mathcal{L}$ is constructed from the Olivetti database of faces, which has ten frontal poses for each of forty individuals. After downsampling, each of these 400 pictures is $46 \times 56$ and we mark the locations of the eyes and mouth; there is relatively more variation in the scale and elongation of the triangles than in their orientation. We use 300 pictures (30 individuals) for training and the rest for estimating the false negative rate.

Recall that the direct input to the classifier is the feature vector of elementary tests $\{X_i\}$, not the original grey level image, and that each $X_i$ corresponds to a localized, oriented edge as described in Section 4.1. Hence, for each training image we compute an edge array which has four boolean features at each pixel corresponding to horizontal and
vertical edges, and two polarities. Each of these edge arrays is then transformed by an
affine mapping which carries the original triangle to the reference one. In other words,
the eyes and the mouth are sent to fixed locations. In Figure 6 we show one greyscale
picture (left) in the training set together with the original (middle) and transformed (right)
edge maps. The training database then has $300 \times 4^2 = 4800$ registered edge maps; the
elementary tests are local $(2 \times 2)$ disjunctions of these maps.

8.2 Learned Arrangements

Shown in Figure 7 are two arrangements learned from $L$, one of size six and one of size
eight. The orientations and polarities of the participating edges are indicated by the small
strips. Clearly the lefthand arrangement captures information about the face contour,
whereas the righthand one reflects typical correlations among edges above the eyes and in
the mouth. The lefthand arrangement splits into two subarrangements of size three, one
on each side of the face, and each of these divides into a single edge and a two paired edges.
These arrangements are quite “typical” of the thousands made. In fact, it turns out that
all the arrangements are composed of edges around the eyes, nose, mouth and forehead
outline. This is illustrated in Figure 8 in which the darkness of a pixel is proportional
to the number of times an edge whose support involves that pixel appears in any of the
learned arrangements.

One measure of the discriminating power of the arrangements is illustrated in Figure
9. The vertical axis indicates probability and the horizontal axis indicates arrangement
size. The curve is the theoretical lower bound in (1). We only use elementary tests with
$P_1$-values approximately one-half or higher, so that $\min_{1 \leq i \leq d} X_i \approx 0.5$ in (1). The value
taken for $\rho$ is .3; see Section 8.3. Thus the curve is $k \to .5(.3)^{\log_2 k}$. We randomly sampled
ten arrangements $A$ for each $k = 1,\ldots,7$ and computed the probabilities of the events
$\{X_A = 1\}$ under both $P_0$ and $P_1$; these are indicated, respectively, by the +’s and the o’s.
As can be seen, the actual likelihoods on faces lie considerably above the bound.

Finally, Figure 10 depicts the cumulative distribution functions, $P_0(Z_k \leq z)$ and
$P_1(Z_k \leq z)$, as estimated from data, for $k = 3$ and $k = 4$. As seen again, $\rho da$’s of
Figure 7: Examples of arrangements of sizes 6 and 8 on faces.

Figure 8: Areas of the face appearing in the arrangements; dark indicates high usage.
modest size are extremely rare under $P_0$ but not under $P_1$. Hence, for example, face and background subimages are separated rather well by $Z_4$.

### 8.3 Correlation Estimates

Are the $\rho$da's estimated from $\hat{P}_1$ actually $\rho$-decomposable with respect to $P_1$? Some are not and some are decomposable at even a higher level. Let $\rho_0 = .1$; this is the value used in our experiments. Recall that each constructed $A \in \mathcal{A}_L(k, \rho_0)$ has a proposed $\rho_0$-decomposition. One can then use additional data (e.g., the test set) to check this decomposition by re-estimating the correlations.

In fact, we can determine $\rho_{\text{max}}(A)$, the maximal value of $\rho$ for which the given decomposition of $A$ is a $\rho$-decomposition. This value may be smaller or larger than $\rho_0$. The results, shown in Figure 11, are conservative because checking whether $A \in \mathcal{A}(k, \rho_0)$, i.e., whether there is any $\rho_0$-decomposition of $A$, would lead to larger levels and shift the curve in Figure 11 to the right. We find, for example, that the decompositions for about 95% of the arrangements are valid at $\rho > 0$, 80% at $\rho \geq .1$ (the target value) and 45% at $\rho \geq .2$.

### 8.4 Estimated Error Rates

Error rates are difficult to estimate, especially without analyzing full scenes. To keep the paper self-contained and in the framework of statistical learning we simply classified the remaining 100 faces in the original database after registering them in the same fashion as the training images, and also a random sample of $46 \times 56$ subimages extracted from various greylevel scenes taken from the WWW. In [7] more realistic estimates are reported based on searching for faces in large scenes, with false positive rates expressed by the number of false alarms per pixel, as in [3].

Due to the conservative placement of the thresholds $c_1, ..., c_K$, the false negative rate is indeed zero. The false positive error as a function of the feature complexity, i.e., $\beta_k =$
Figure 10: Distribution functions of $Z_3$ and $Z_4$ on faces

Figure 11: Proportion of the learned decompositions which are actually decomposable at level $\rho$. 

15
\( \hat{p}_0(Z_1 \geq c_1, \ldots, Z_k \geq c_k) \) is given in Figure 12. The important property is the sharp decrease with \( k \); the final value (approximately 4\%) is relatively meaningless since this is but one step in a coarse-to-fine process. For example, if we restrict the location to a 2 \( \times \) 2, the rate drops to well below one percent.

9 Conclusion

We have formulated object detection as a learning problem constrained by invariance and efficiency. Our basic assumption is that it is ultimately possible to detect objects by discovering features which are considerably more likely under the “object distribution” than under the “background distribution.”

Although the context is statistical and inductive, we have not appealed to a general theory of inductive learning, for instance structural risk minimization [15], or to theoretical results about neural network classifiers [4],[6], or to generic representations, such as Bayes nets. Instead, the learning is highly dedicated to the vision problem. We capture the spatial dependency structure by estimating correlations among conjunctions of localized features. Each estimated parameter has an explicit interpretation, which makes training very fast. And we directly address the natural constraints - invariance to photometric and geometric distortions, very low false negative error and limited amounts of training data.

We have concentrated on a two-dimensional pattern and not a three-dimensional object. (We place the frontal views of faces in the former category.) It remains to be seen whether the methodology can be extended to handle all the degrees of freedom in imaging a 3D object. Obviously, simultaneously detecting many different object classes \( c = 1, \ldots, C \) is still more difficult. In our framework, the key question is “reusable parts” - how the size of the combined feature set \( \bigcup_{c=1,\ldots,C} \mathcal{A}_c(k) \) scales with \( C \), where \( \mathcal{A}_c(k) \) is the set of arrangements of size \( k \) dedicated to object \( c \). We want logarithmic scaling, resulting from representing different objects with the same arrangements whenever possible.

Finally, in defense of limited goals, nobody has yet demonstrated that objects from
even one generic class under constrained poses can be rapidly detected without errors in complex, natural scenes; visual selection by humans occurs within two hundred milliseconds and is virtually perfect.

References


