A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry

Michael C. Burl[1], Markus Weber[2], and Pietro Perona[2,3]

[1] Jet Propulsion Laboratory
M/S 525-3660, 4800 Oak Grove Drive
Pasadena, CA 91109, U.S.A.
Michael.C.Burl@jpl.nasa.gov
[2] California Institute of Technology
MC 136-93
Pasadena, CA 91125, U.S.A.
{mweber,perona}@vision.caltech.edu
[3] Università di Padova, Italy

Abstract. Many object classes, including human faces, can be modeled as a set of characteristic parts arranged in a variable spatial configuration. We introduce a simplified model of a deformable object class and derive the optimal detector for this model. However, the optimal detector is not realizable except under special circumstances (independent part positions). A cousin of the optimal detector is developed which uses "soft" part detectors with a probabilistic description of the spatial arrangement of the parts. Spatial arrangements are modeled probabilistically using shape statistics to achieve invariance to translation, rotation, and scaling. Improved recognition performance over methods based on "hard" part detectors is demonstrated for the problem of face detection in cluttered scenes.

1 Introduction

Visual recognition of objects (chairs, sneakers, faces, cups, cars) is one of the most challenging problems in computer vision and artificial intelligence. Historically, there has been a progression in recognition research from the particular to the general. Researchers initially worked on the problem of recognizing individual objects; however, during the last five years the emphasis has shifted to recognizing classes of objects which are visually similar.

One line of research has concentrated on exploiting photometric aspects of objects. Matched filtering (template matching) was an initial attempt along these lines. More modern approaches use classification in subspaces of filter responses, where the set of filters is selected based on human receptive fields, principal components analysis [12, 23, 16, 2], linear discriminant analysis, or by training with perceptron-like architectures [22, 20]. These methods allow one to accomodate a broader range of variation in the appearance of the target object than is possible using a simple matched filter.

A second line of research has used geometric constraints between low level object features. Methods such as alignment [11], geometric invariants [15], combinations of views [24, 21], and geometric hashing [26, 19] fit within this category.

Further generalization has been obtained by allowing an object to be represented as a collection of more complex features (or texture patches) connected with a deformable geometrical model. The neocognitron architecture [10] may be seen as an early representative. More recently, Yuille [27] proposed to use deformable templates to be fit to contrast profiles by gradient descent of a suitable energy function. Lades, von der Malsburg and colleagues [13, 25] proposed to use jet-based detectors and deformable meshes for encoding shape. Their work opened a number of interesting questions: (a) how to derive the energy function that encodes shape from a given set of examples, (b) how to initialize automatically the model so that it converges to the desired object despite a cluttered background in the image, and (c) how to handle partial occlusion of the object. Lanitis, Cootes et al. [14, 6, 7] proposed to use principal components analysis (applied to the shape of an object rather than the photometric appearance) to address the first issue. Pope and Lowe [17, 18] used probability theory to model the variation in shape of triples of features. Brunelli and Poggio [1] showed that an ad hoc face detector consisting of individual features linked together with crude geometry constraints outperformed a rigid correlation-based "full-face" detector.

Burl, Leung, and Perona [3, 4] introduced a principled framework for representing object deformations using probabilistic shape models. Local part detectors were used to identify candidate locations for object parts. These candidates were then grouped into object hypotheses and scored based on the spatial arrangement of the parts. This approach was shown to work well for detecting human faces in cluttered backgrounds and with partial occlusion. There is no guarantee, however, that first "hard-detecting" the object parts and then looking for the proper configuration of parts is the best approach. (Under a "hard" detection strategy, if the response of a part detector is above threshold, only the position of the part is recorded; the actual response values are not retained for subsequent processing.)

In this paper, we reconsider from first principles the problem of detecting an object consisting of characteristic parts arranged in a deformable configuration. The key result is that we should employ a "soft-detection" strategy and seek the arrangement of part locations that maximizes the sum of the shape log-likelihood ratio *and* the responses to the part detectors. This criteria, which combines both the local photometry (part match) and the global geometry (shape likelihood) provides a significant improvement over the "hard-detection" strategy used previously.

In Sect. 2 we provide a mathematical model for deformable object classes. The optimal detector for this model is derived from first principles in Sect. 3. We then investigate, in Sect. 4, an approximation to the optimal detector which is invariant to translation, rotation and scaling. In Sect. 5 we present evidence which verifies the practical benefits of our theoretical findings.

2 Deformable Object Classes

We are interested in object classes in which instances from the class can be modeled as a set of characteristic *parts* in a deformable spatial configuration. As an example, consider human faces, which consist of two eyes, a nose, and mouth. These parts appear in an arrangement that depends on an individual's facial geometry, expression, and pose, as well as the viewpoint of the observer.

We do not offer a precise definition of what constitutes an object "part", but we are generally referring to any feature of the object that can be reliably detected and localized using only the local image information. Hence, a part may be defined through a variety of visual cues such as a distinctive photometric pattern, texture, color, motion, or symmetry. Parts may also be defined at multiple scales. A coarse resolution view of the head can be considered a "part" as can a fine resolution view of an eye corner. The parts may be object-specific (eyes, nose, mouth) or generic (blobs, corners, textures).

2.1 Simplified Model

Consider a 2-D object consisting of N photometric parts $P_i(x, y)$, each occuring in the image at a particular spatial location (x_i, y_i). The parts P_i can be thought of as small image patches that are placed down at the appropriate positions. Mathematically, the image T of an object is given by:

$$T(x, y) = \sum_{i=1}^{N} P_i(x - x_i,\ y - y_i) \tag{1}$$

For convenience, we will assume, that the $P_i(x, y)$ are defined for any pair (x, y), but are non-zero only inside a relatively small neighborhood around $(0, 0)$.

Let $\boldsymbol{X}$ be the vector describing the positions of the object parts, i.e.

$$\boldsymbol{X} = \begin{bmatrix} x_1 \; x_2 \ldots x_N \; y_1 \; y_2 \ldots y_N \end{bmatrix}^T \tag{2}$$

An *object class* can now be defined as the set of objects induced by a set of vectors $\{\boldsymbol{X}_k\}$. In particular, we assume that the part positions are distributed according to a joint probability density $p_{\boldsymbol{X}}(\boldsymbol{X})$. We will designate the resulting object class as $\mathcal{T}$. To generate an object from this class, we first generate a random vector $\boldsymbol{X}$ according to the density $p_{\mathbf{X}}(\boldsymbol{X})$. Since this vector determines the part positions, we simply place the corresponding pattern P_i at each of these positions.

Note that no assumption about $p_{\boldsymbol{X}}(\boldsymbol{X})$ is made at this time. It should be clear, however, that through $p_{\boldsymbol{X}}(\boldsymbol{X})$ we can control properties of the object class, such as the range of meaningful object *shapes*, as well as tolerable ranges of certain transformations, such as rotation, scaling and translation.

3 Derivation of the Optimal Detector

The basic problem can be stated as follows: given an image $\mathcal{I}$ determine whether the image contains an instance from $\mathcal{T}$ (hypothesis ω_1) or whether the image is background-only (hypothesis ω_2). In our previous work we proposed a two-step solution to this problem: (1) apply feature detectors to the image in order to identify candidate locations for each of the object parts and (2) given the candidate locations, find the set of candidates with the most object-like spatial configuration. However, there is nothing to say that first hard-detecting candidate object parts is the right strategy. In the following section, we will directly derive the optimal detector starting from the pixel image $\mathcal{I}$.

3.1 Optimal Detector

The optimal decision statistic is given by the likelihood ratio

$$\Lambda = \frac{p(\mathcal{I}|\omega_1)}{p(\mathcal{I}|\omega_2)} \tag{3}$$

We can rewrite the numerator by conditioning on the spatial positions $\boldsymbol{X}$ of the object parts. Hence,

$$\Lambda = \frac{\sum_{\boldsymbol{X}} p(\mathcal{I}|\boldsymbol{X},\omega_1)\cdot p(\boldsymbol{X}|\omega_1)}{p(\mathcal{I}|\omega_2)} \tag{4}$$

where the summation goes over all possible configurations of the object parts. Assuming that parts do not overlap, we can divide the image into $N+1$ regions, $\mathcal{I}^0, \mathcal{I}^1, \ldots, \mathcal{I}^N$, where $\mathcal{I}^i$ is an image which is equal to $\mathcal{I}$ in the area occupied by the non-zero portion of part P_i (positioned according to $\boldsymbol{X}$) and zero otherwise. $\mathcal{I}^0$ denotes the background. Assuming furthermore that the background is independent across regions, we obtain

$$\begin{aligned}
\Lambda &= \frac{\sum_{\boldsymbol{X}} \prod_{i=0}^{N} p(\mathcal{I}^i|\boldsymbol{X},\omega_1)\cdot p(\boldsymbol{X}|\omega_1)}{p(\mathcal{I}|\omega_2)} && (5)\\
&= \sum_{\boldsymbol{X}} \left[\prod_{i=1}^{N} \frac{p(\mathcal{I}^i|\boldsymbol{X},\omega_1)}{p(\mathcal{I}^i|\omega_2)}\right]\cdot p(\boldsymbol{X}|\omega_1) && (6)\\
&= \sum_{\boldsymbol{X}} \left[\prod_{i=1}^{N} \lambda_i(x_i,y_i)\right]\cdot p(\boldsymbol{X}|\omega_1) && (7)
\end{aligned}$$

Here, the $\lambda_i(x_i,y_i) = \frac{p(\mathcal{I}^i|\boldsymbol{X},\omega_1)}{p(\mathcal{I}^i|\omega_2)}$ can be interpreted as likelihood ratios expressing the likelihood of part P_i being present in the image at location (x_i, y_i). Note that $\lambda_0(x,y)$ is equal to one, under the hypothesis that the statistics of the background region do not depend on the presence or absence of the object.

We can specialize this derivation by introducing a particular part detection method. For example, assuming that the object is embedded in white Gaussian noise, we can substitute Gaussian class conditional densities and obtain

$$\lambda_i = \frac{\mathcal{N}\left(\mathcal{I}^i;\ \boldsymbol{\mu}_{\boldsymbol{X}}, \sigma^2 \boldsymbol{I}\right)}{\mathcal{N}\left(\mathcal{I}^i;\ \mathbf{0}, \sigma^2 \boldsymbol{I}\right)} \tag{8}$$

Here, $\boldsymbol{\mu}_{\boldsymbol{X}}$ is the object with parts positioned at $\mathbf{X}$, $\mathbf{0}$ shall denote a vector of zeros and $\boldsymbol{I}$ is the identity matrix. Expanding the Gaussian densities and combining terms yields:

$$\begin{aligned}
\lambda_i &= \exp\left(\frac{\boldsymbol{\mu}_{\boldsymbol{X}}^T \mathcal{I}^i}{\sigma^2} - \frac{\boldsymbol{\mu}_{\boldsymbol{X}}^T \boldsymbol{\mu}_{\boldsymbol{X}}}{2\sigma^2}\right) \\
&= \exp\left(-\frac{\boldsymbol{\mu}_{\boldsymbol{X}}^T \boldsymbol{\mu}_{\boldsymbol{X}}}{2\sigma^2}\right) \cdot \exp\left(\frac{\boldsymbol{\mu}_{\boldsymbol{X}}^T \mathcal{I}^i}{\sigma^2}\right) \\
&= c \cdot \exp\left(\frac{\mu_{\boldsymbol{X}}^T \mathcal{I}^i}{\sigma^2}\right) \qquad (9)
\end{aligned}$$

where σ^2 is the variance of the pixel noise and c depends only on the energy in the object image and is therefore constant independent of $\boldsymbol{X}$, provided the parts do not overlap. Equation (9) simply restates the well known fact that matched filtering is the optimal part detection strategy under this noise model. Writing A_i for the response image obtained by correlating part i with the image $\mathcal{I}$ and normalizing by σ^2, we finally obtain

$$\mathit{\Lambda} = c \cdot \sum_{\boldsymbol{X}} \left[\prod_{i=1}^{N} \exp\left(A_i(x_i, y_i)\right)\right] \cdot p(\boldsymbol{X}) \tag{10}$$

The constant c does not affect the form of the decision rule, so we will omit it from our subsequent equations.

3.2 Independent Part Positions

If the part positions are independent, $p(\boldsymbol{X})$ can also be expressed as a product

$$p(\boldsymbol{X}) = \prod_{i=1}^{N} p_i(x_i, y_i) \tag{11}$$

Thus, we have

$$\mathit{\Lambda} = \sum_{\boldsymbol{X}} \left[\prod_{i=1}^{N} \lambda_i(x_i, y_i) p_i(x_i, y_i)\right]$$

For the special case of additive white Gaussian noise, we obtain

$$\mathit{\Lambda} = \sum_{\boldsymbol{X}} \left[\prod_{i=1}^{N} \exp\left(A_i(x_i, y_i)\right) p_i(x_i, y_i)\right]$$

$$
\begin{aligned}
&= \sum_{\boldsymbol{X}} \left[\prod_{i=1}^{N} \exp\left(A_i(x_i, y_i) + \log p_i(x_i, y_i)\right) \right] \\
&= \prod_{i=1}^{N} \left[\sum_{(x_i, y_i)} \exp\left(A_i(x_i, y_i) + \log p_i(x_i, y_i)\right) \right] \qquad (12)
\end{aligned}
$$

Thus, we need to compute the correlation response image (normalized by σ^2) for each object part. To this image, we add the log probability that the part will occur at a given spatial position, take the exponential, and sum over the whole image. This process is repeated for each object part. Finally, the product of scores over all the object parts yields the likelihood ratio.

Note, that the detector is not invariant to translation, rotation, and scaling since the term $p_i(x_i, y_i)$ includes information about the absolute coordinates of the parts.

3.3 Jointly Distributed Part Positions

If the part positions are *not independent*, we must introduce an approximation since summing over all *combinations* of part positions as in (7) is infeasible. The basic idea—similar to a winner-take-all strategy—is to assume that the summation is dominated by one term corresponding to a specific combination $\boldsymbol{X}_0$ of the part positions. With this assumption, we have

$$
\begin{aligned}
\Lambda \;\approx\; \Lambda_0 &= \prod_{i=1}^{N} \lambda_i(x_i, y_i) \cdot p(\boldsymbol{X}_0) \\
\log \Lambda_0 &= \sum_{i=1}^{N} \log \lambda_i(x_i, y_i) \;+\; \log p(\boldsymbol{X}_0) \qquad (13)
\end{aligned}
$$

and in the case of additive white Gaussian noise

$$
\log \Lambda_0 = \left(\sum_{i=1}^{N} A_i(x_{0i}, y_{0i}) \right) \;+\; \log p(\boldsymbol{X}_0) \qquad (14)
$$

The strategy now is to find a set of part positions such that the matched filter responses are high and the overall configuration of the parts is consistent with $p(\boldsymbol{X}|\omega_1)$. Again, the resulting detector is not invariant to translation, rotation, and scaling.

4 TRS-invariant Approximation to the Optimal Detector

The approximate log-likelihood ratio given in (13) can readily be interpreted as a combination of two terms: the first term, $\sum A_i$, measures how well the hypothesized parts in the image match the actual model parts, while the second term, $p(\boldsymbol{X}_0)$, measures how well the hypothesized spatial arrangement matches the ideal model arrangement. The second term, the configuration match, is specified as a probability density over the absolute coordinates of the parts, which in practice is not useful since (a) there is no way to know or estimate this density and (b) this formulation does not provide TRS-invariance.

We can make use of the theory developed in our previous work (see [4] or [5]) to write down a TRS-invariant detector that closely follows the form of (13). In particular, we know how to factor the term $p(\boldsymbol{X}_0)$ into a part that depends purely on shape and a part that depends purely on pose:

$$p_{\boldsymbol{X}}(\boldsymbol{X}_0) = p_{\boldsymbol{U}}(\boldsymbol{U}_0(\boldsymbol{X}_0)) \cdot p_{\boldsymbol{\Theta}}(\boldsymbol{\Theta_0}(\boldsymbol{X_0})) \tag{15}$$

Here, $\boldsymbol{U}$ denotes the *shape* of the constellation and the vector $\boldsymbol{\Theta}$ captures the pose parameters. Computing $\mathbf{U}(\boldsymbol{X})$ corresponds to transforming a constellation $\boldsymbol{X}$ in the image to so-called *shape space* by mapping two part positions (the *base-line pair*) to fixed reference positions. In shape space, the positions of the remaining $N-2$ parts define the shape of the configuration, written as

$$\boldsymbol{U} = \begin{bmatrix} u_3 \; u_4 \; \ldots \; u_N \; v_3 \; v_4 \; \ldots \; v_N \end{bmatrix}^T \tag{16}$$

If $p_{\boldsymbol{X}}(\boldsymbol{X})$ is a joint Gaussian density, then the shape density, $p_{\boldsymbol{U}}(\boldsymbol{U})$, can be computed in closed form as shown by Dryden and Mardia [8]. This established, we can obtain the TRS-invariant detector by dropping the pose information completely and working with shape variables $\boldsymbol{U}_0$, instead of figure space variables $\boldsymbol{X}_0$. The resulting log-likelihood ratio is then

$$\log \Lambda_1 = \sum_{i=1}^{N} A_i(x_{0i}, y_{0i}) \; + \; K \cdot \log \frac{p_U(\boldsymbol{U}_0|\omega_1)}{p_U(\boldsymbol{U}_0|\omega_2)} \tag{17}$$

The shape likelihood ratio, rather than just $p_{\boldsymbol{U}}(\boldsymbol{U}_0)$, is used in place of $p_{\boldsymbol{X}}(\boldsymbol{X}_0)$ to provide invariance to the choice of baseline features. The likelihood ratio also assigns lower scores to configurations that have higher probabilities of accidental occurrence. The factor of K provides a weighted trade-off between the part match and shape match terms, since the units of measurement for the two terms will no longer agree. (The proper setting for this value can be estimated from training data).

An object hypothesis is now just a set of N coordinates specifying the (hypothesized) spatial positions of the object parts. Any hypothesis can be assigned a score based on (17). It is no longer the case that hypotheses must consist only of points corresponding to the best part matches. The trade-off between having the parts match well and having the shape match well may imply that it is better to accept a slightly worse part match in favor of a better shape match or vice versa.

We do not have a procedure for finding the hypothesis that optimizes $\log \Lambda_1$. One heuristic approach $\mathcal{A}_1$ is to identify candidate part locations at maxima of the part detector responses and combine these into hypotheses using the conditional search procedure described in [4]. However, instead of discarding the response values, these should be summed and combined with the shape likelihood. In this approach, the emphasis is on finding the best part matches and accepting whatever spatial configuration occurs. There is no guarantee that the procedure will maximize $\log \Lambda_1$.

Figure 1 illustrates the gain of approach $\mathcal{A}_1$ over hard detection. The two components of the goodness function (sum of responses and shape log-likelihood) can be seen as dimensions in a two dimensional space. Evaluating the goodness function is equivalent to projecting the data onto a particular direction, which is determined by the trade-off factor K. A technique known as "Fisher's Linear Discriminant" [9] provides us with the direction which maximizes the separability of the two classes. If the sum of the detector responses had no discriminative power, the value of K would tend toward infinity. This would correspond to a horizontal line in the figure. The advantage of soft detection is further illustrated in Fig. 2.

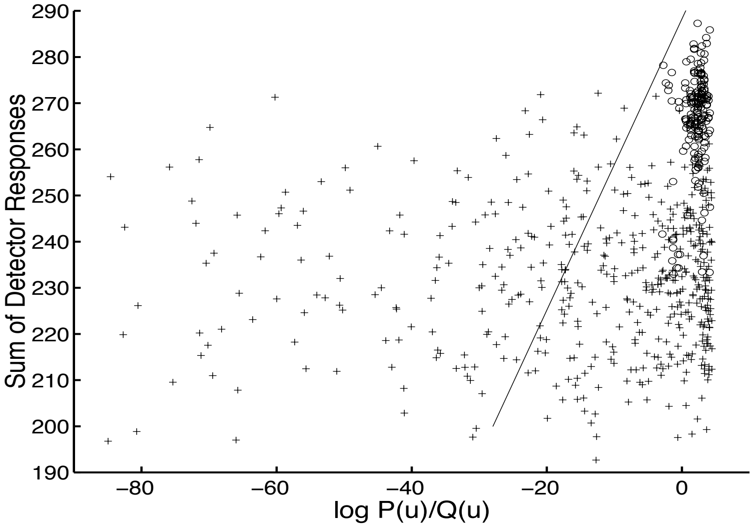


Fig. 1. Illustration of the advantage of *soft* detection. The sum of the detector outputs is plotted against the shape log-likelihood, for a set of face (o) and background (+) samples. Also shown is a line onto which the data should be projected (derived by Fisher's Linear Discriminant method).

A second approach, $\mathcal{A}_2$, is to insist on the best shape match and accept whatever part matches occur. This method is roughly equivalent to using a rigid matched filter for the entire object, but applying it at multiple orientations and scales.

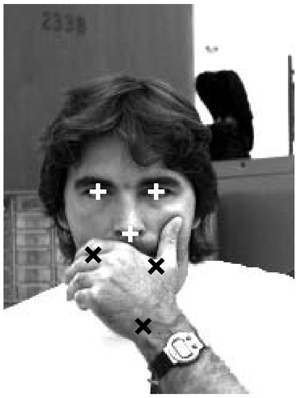


Fig. 2. Two constellations of candidates for part locations are shown. The background constellation (black 'x') yields a greater shape likelihood value than the correct hypothesis (white '+'). However, when the detector response values are taken into consideration, the correct hypothesis will score higher.

Finally, we tested a third approach, $\mathcal{A}_3$, that intuitively seems appealing. Candidate part locations are identified as before in $\mathcal{A}_1$ at local maxima in the part response image. From pairs of candidate parts, the locations of the other parts are estimated to provide an initial hypothesis. (So far, this is equivalent to using a fixed-shape template anchored at the two baseline points). From the initial hypothesis, however, a gradient-style search is employed to find a local maximum of $\log \Lambda_1$. Individual part positions are pulled by two forces. One force tries to maximize the response value while the other force tries to improve the shape of the configuration.

5 Experiments

We conducted a series of experiments aimed at evaluating the improvements over hard detection of object parts, brought about by the different approaches described in the previous section. To test our method, we chose the problem of detecting faces from frontal views. A grayscale image sequence of 400 frames was acquired from a person performing head movements and facial expressions in front of a cluttered background. The images were 320 × 240 pixels in size, while the face occupied a region of approximately 40 pixels in height. Our face model was comprised of five parts, namely eyes, nose tip and mouth corners.

For the part detectors we applied a correlation based method—similar to a matched filter—acting not on the grayscale image, but on a transformed version of the image that characterizes the dominant local orientation. We found this method, which we previously described in [5], to be more robust against variations in illumination than grayscale correlation. The part detectors were trained

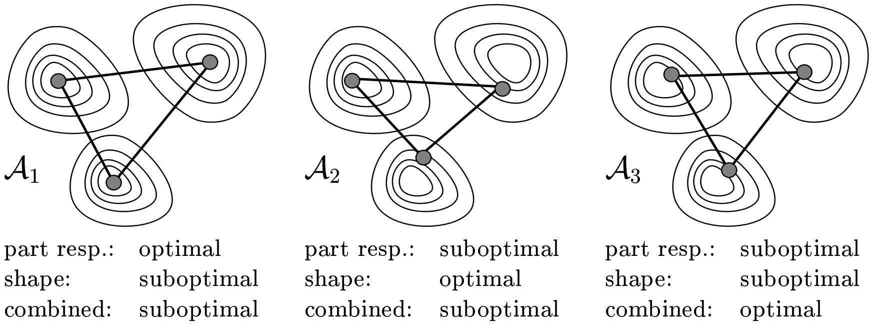
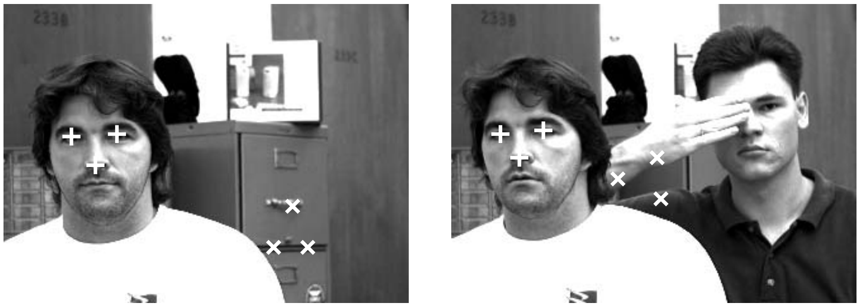


Fig. 3. Pictorial illustration of the three approaches $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$ discussed in the text. For each approach we show a set of three contours which represent the superposition of response functions from three part detectors. With approach $\mathcal{A}_1$ the detector responses are optimal, but the combination of responses and shape is suboptimal. With approach $\mathcal{A}_2$ the shape likelihood is optimal, but the combination is still suboptimal. Only under approach $\mathcal{A}_3$ is the combined likelihood function optimized by seeking a compromise between contributions from the detector responses and shape.



	Best Correct	Best False
Σ Responses	101.1	93.7
Shape Log-LH.	1.457	-0.096
Weighted Total	101.5	93.7

	Best Correct	Best False
Σ Responses	96.4	94.8
Shape Log-LH.	3.460	-3.530
Weighted Total	97.5	93.7

Fig. 4. Examples from the sequence of 400 frames used in the experiments. The highest scoring correct and incorrect constellations are shown for each frame. The tables give the values for shape log-likelihood, sum of detector responses as well as overall goodness function.

on images of a second person. In order to establish ground truth for the part locations, each frame of the sequence was hand-labeled.

Prior to the experiment, shape statistics had been collected from the face of a third person by fitting a joint Gaussian density with full covariance matrix to data extracted from a sequence of 150 images, taken under a semi-controlled pose as discussed in [4].

5.1 Soft Detection vs. Hard Detection

In a first experiment, we found that using the five features described above, recognition on our test sequence under the hard detection paradigm was almost perfect, making it difficult to demonstrate any further improvements. Therefore, in order to render the task more challenging, we based the following experiments on the upper three features (eyes and nose tip) only. In this setting, approach $\mathcal{A}_1$, i.e. combining the part responses with the shape likelihood without any further effort to maximize the overall goodness function (17), yields a significant increase in recognition performance. This result is illustrated in Fig. 5, where ROC (*Receiver Operating Characteristics*) curves are shown for the hard detection method as well as for approach $\mathcal{A}_1$.

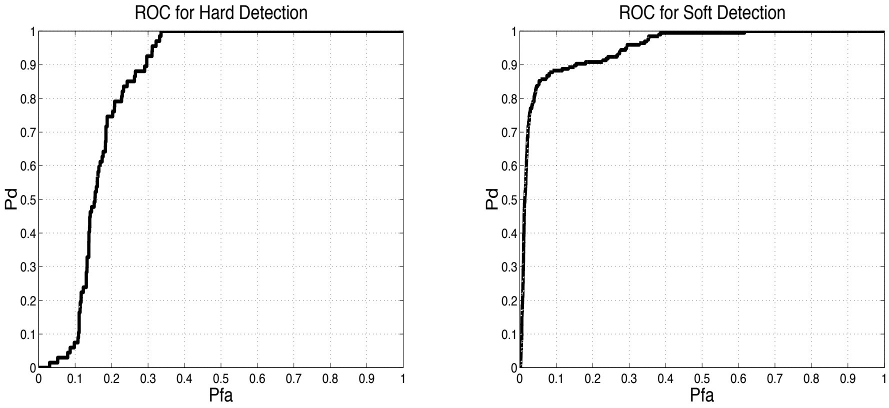


Fig. 5. The two ROC curves show the performance of hard vs. soft detection of features as a trade-off between detection probability, P_{d}, and probability of false alarm, P_{fa}. The soft detection method $\mathcal{A}_1$ clearly outperforms the hard detection strategy, especially in the low false alarm range.

5.2 Gradient Descent Optimization

Approach $\mathcal{A}_3$ was tested in a second experiment by performing a gradient descent maximization of the goodness criteria with respect to the hypothesized part positions in the image. There are two potential benefits from doing this: improved detection performance and improved localization accuracy of the part positions. A cubic spline interpolation of the detector response images was used in order to provide the minimization algorithm with a continuous and differentiable objective function. Local maxima of the detector response maps were used as initial estimates for the part positions. We found that, on average, optimal part positions were found within a distance of less than one pixel from the initial positions.

Fig. 6 shows the detection performance of the method before and after optimization of (17). There does not seem to be any noticeable improvement over approach $\mathcal{A}_1$. This result is somewhat surprising, but not entirely counterintuitive. This is because by optimizing the goodness criteria, we are improving the score of the constellations from both classes, ω_1 and ω_2. It is not clear that, on average, we are achieving a better separation of the classes in terms of their respective distribution of the goodness criteria. From a different perspective, this is a positive result, because the gradient descent optimization is computationally very expensive, whereas we have already been able to develop a 2 Hz real-time implementation of approach $\mathcal{A}_1$ on a PC with Pentium processor (233 MHz).

Since our part detectors did not exhibit a significant localization error for the test data at hand, we have not been able to determine whether approach $\mathcal{A}_3$ might provide improved localization accuracy.

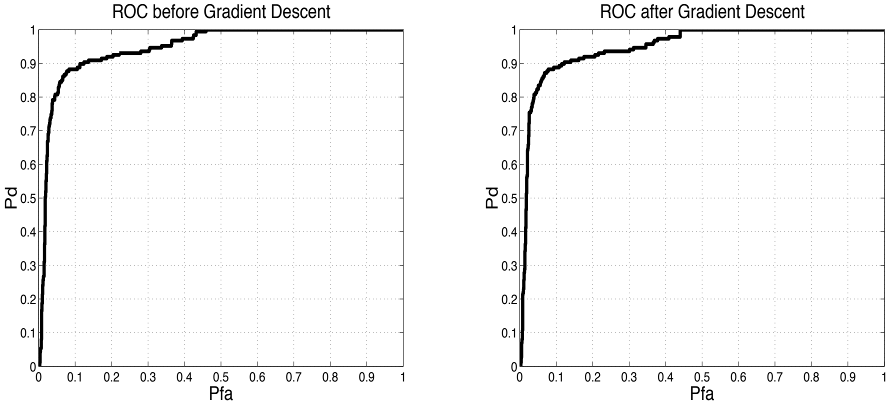


Fig. 6. The ROC performance does not significantly improve after Gradient Descent Optimization of the goodness criteria.

6 Conclusion

We have reconsidered from first principles the problem of detecting deformable object classes of which human faces are a special case. The optimal detector for object class $\mathcal{T}$ was derived for the case of independent part positions. When the part positions are jointly distributed the optimal detector is too complicated to evaluate, but it can be approximated using a winner-take-all simplification. In both cases, the detector is composed of two terms: the first term measures how well the hypothesized parts in the image match the actual model parts, while the second term measures how well the hypothesized spatial arrangement matches the ideal model arrangement.

The configuration match is specified in terms of the absolute positions of the object parts, therefore the optimal detector cannot be used in practice. However, using previous theoretical results, we were able to write an expression that closely follows the form of (13), but only exploits the *shape* of the configuration. The resulting criteria combines the part match with shape match and is invariant to translation, rotation, and scaling.

Although we do not have a procedure for finding the hypothesis that maximizes the overall goodness function, a heuristic approach $\mathcal{A}_1$ worked very well. In this approach, candidate parts are identified and grouped into hypotheses as in the shape-only method, but, in addition, the response values (part matches) are retained and combined with the shape likelihood. A second approach, including a gradient descent optimization of the goodness function with respect to the part position in the image, did not provide significant improvement in recognition performance.

Acknowledgement

The research described in this paper has been carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. It was funded in part by the NSF Center for Neuromorphic Systems Engineering at Caltech.

References

1. R. Brunelli and T. Poggio. "Face Recognition: Features versus Templates". *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(10):1042–1052, October 1993.
2. M.C. Burl, U.M. Fayyad, P. Perona, P. Smyth, and M.P. Burl. "Automating the hunt for volcanoes on Venus". In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 1994.
3. M.C. Burl, T.K. Leung, and P. Perona. "Face Localization via Shape Statistics". In *Intl. Workshop on Automatic Face and Gesture Recognition*, 1995.
4. M.C. Burl, T.K. Leung, and P. Perona. "Recognition of Planar Object Classes". In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 1996.
5. M.C. Burl, M. Weber, T.K. Leung, and P. Perona. *From Segmentation to Interpretation and Back: Mathematical Methods in Computer Vision*, chapter "Recognition of Visual Object Classes". Springer, in press.
6. T.F. Cootes and C.J. Taylor. "Combining Point Distribution Models with Shape Models Based on Finite Element Analysis". *Image and Vision Computing*, 13(5):403–409, 1995.

7. T.F. Cootes and C.J. Taylor. “Locating Objects of Varying Shape Using Statistical Feature Detectors”. In *European Conf. on Computer Vision*, pages 465–474, 1996.
8. I.L. Dryden and K.V. Mardia. “General Shape Distributions in a Plane”. *Adv. Appl. Prob.*, 23:259–276, 1991.
9. R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, Inc., 1973.
10. K. Fukushima. Neural networks for visual-pattern recognition. *IEICE Trans. Inf. & Syst.*, 74(1):179–190, 1991.
11. D.P. Huttenlocher and S. Ullman. “Object Recognition Using Alignment”. In *Proc. 1^{st} Int. Conf. Computer Vision*, pages 102–111, 1987.
12. M. Kirby and L. Sirovich. Applications of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(1):103–108, Jan 1990.
13. M. Lades, J.C. Vorbruggen, J. Buhmann, J. Lange, C. v.d. Malsburg, R.P. Wurtz, and W. Konen. “Distortion Invariant Object Recognition in the Dynamic Link Architecture”. *IEEE Trans. Comput.*, 42(3):300–311, Mar 1993.
14. A. Lanitis, C.J. Taylor, T.F. Cootes, and T. Ahmed. “Automatic Interpretation of Human Faces and Hand Gestures Using Flexible Models”. In *International Workshop on Automatic Face- and Gesture-Recognition*, pages 90–103, 1995.
15. J. Mundy and A. Zisserman, editors. *Geometric invariance in computer vision.* MIT Press, Cambridge, Mass., 1992.
16. Hiroshi Murase and Shree Nayar. “Visual Learning and Recognition of 3-D Objects from Appearance”. *Int J. of Comp. Vis.*, 14:5–24, 1995.
17. Arthur R. Pope and David G. Lowe. “Modeling Positional Uncertainty in Object Recognition”. Technical report, Department of Computer Science, University of British Columbia, 1994. Technical Report # 94-32.
18. Arthur R. Pope and David G. Lowe. “Learning Feature Uncertainty Models for Object Recognition”. In *IEEE International Symposium on Computer Vision*, 1995.
19. I. Rigoutsos and R. Hummel. A bayesian approach to model matching with geometric hashing. *Comp. Vis. and Img. Understanding*, 62:11–26, Jul. 1995.
20. H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(1):23–38, Jan 1998.
21. A. Shashua. “On Geometric and Algebraic Aspects of 3D Affine and Projective Structures from Perspective 2D Views”. Technical Report A.I. Memo # 1405, MIT, 1993.
22. K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(1):39–51, Jan 1998.
23. M. Turk and A. Pentland. “Eigenfaces for Recognition”. *J. of Cognitive Neurosci.*, 3(1), 1991.
24. S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(10), 1991.
25. L. Wiskott and C. von der Malsburg. “A Neural System for the Recognition of Partially Occluded Objects in Cluttered Scenes”. *Int. J. of Pattern Recognition and Artificial Intelligence*, 7(4):935–948, 1993.
26. H.J. Wolfson. “Model-Based Object Recognition by Geometric Hashing”. In *Proc. 1^{st} Europ. Conf. Comput. Vision, LNCS-Series Vol. 427, Springer-Verlag*, pages 526–536, 1990.
27. A.L. Yuille. “Deformable Templates for Face Recognition”. *J. of Cognitive Neurosci.*, 3(1):59–70, 1991.