ISIS technical report series, Vol. 11, November 2000

Intelligent Sensory Information Systems
University of Amsterdam
The Netherlands

Face Detection Methods
A Critical Evaluation

Thang V. Pham and Marcel Worring
Intelligent Sensory Information Systems
Department of Computer Science
University of Amsterdam
The Netherlands

We give an overview of face detection methods in literature. We also present our experiments to evaluate some of these methods. Existing method can be broadly divided into two categories: template-based and feature-based methods. Template-based methods, making use of statistical pattern recognition techniques, are robust in detecting frontal faces. Feature-based methods are better suited for detecting non-frontal faces. Such methods are, however, dependent on the feature detection results which are often not reliable. The use of color information has been examined also. Experimental results show that simple use of color information leads to a low degree of detection accuracy.

Contents

Intelligent Sensory Information Systems
Department of Computer Science
University of Amsterdam
Kruislaan 403
1098 SJ Amsterdam
The Netherlands

tel: +31 20 525 7463
fax: +31 20 525 7490
http://www.science.uva.nl/research/isis

Corresponding author:
Thang V. Pham
tel: +31(20)525 7528
vietp@science.uva.nl
http://www.science.uva.nl/~vietp

1 Introduction

Face detection is an important step in any automatic face recognition system. Given an image of arbitrary size, the task is to detect the presence of any human face appearing in the image. Detection is a challenging task since human faces may appear in different scales, orientations, and with different head poses. The imaging condition, such as illumination, also affects the appearance of human faces considerably. Moreover, human faces are nonrigid objects. There are a lot of variations due to varying facial expressions. In addition, the presence of other objects or facial features such as glasses, make-up or beards contribute substantially to the variation of facial appearance in an image.

A large number of face detection methods have been proposed in literature. Yang et al. [38] give a comprehensive survey on existing methods. In this report, we discuss methods in a greater details, especially the merits of each method. In addition, experiments have been carried out to evaluate the applicability of some methods for our news video dataset.

Face detection methods can be broadly divided into two classes: template-based and feature-based detection. In the first class, the pixel set in an image window is classified to see if there is a human face at that location. In the second approach, faces are detected by grouping facial features according to their geometric configuration in a model of a face, or by segmenting candidate facial regions based on color information and further verifying these regions based on their shapes and pixel values.

The next section of the report gives a brief overview of test sets which are used to evaluate face detection methods in literature. The template-based approach to the face detection problem is discussed in section 3 and the feature-based approach in section 4. Section 5 presents our experiments and a summary of the report is given in section 6.

2 Benchmark test sets

In this section, we give a brief description of test sets for face detection in literature.

The FERET (FacE REcognition Technology) testset [18] was created to assess the performance of face recognition algorithm. Each image in this dataset consists of a face on a uniform background. Although it is not suitable for evaluating methods for face detection in a general scene, many authors report their detection results on this dataset.

Sung and Poggio [30] compiled two test sets. The first one consists of 301 frontal and near-frontal faces of 71 different people. All images are high quality digitized images with fair amount of lighting variation. The second test set consists of 23 images with 149 face patterns. Most of the images have complex background patterns. There is also a wide range of variation in image quality. The second test set is often referred to as the MIT test set in literature.

Rowley at el. [23, 22] created the CMU test set. It consist of 130 images with a total of 507 frontal faces, including the MIT test set described above. The images are collected from the World Wide Web, scanned from photographs and newspaper pictures, and digitized from broadcast television[1]. Note that some authors exclude the line draw faces of the dataset in their evaluation.

3 Template-based detection

In this class of methods, a window of fixed size is scanned through the image. At each location the corresponding part of the image is classified as a face or non-face pattern. To detect faces of different scales, either the input image is scaled down or the size of the template is adjusted appropriately. Many techniques from pattern recognition have been applied to distinguish between face and non-face patterns. In the next section, statistical pattern recognition based methods are presented. Section 2.2 discusses rule-based methods for classification.

3.1 Statistical pattern recognition based methods

In a general object detection setting, let $x = \{x_1, x_2, \ldots, x_n\}$ be the pixel values of an image window where n is the number of pixels. Furthermore, let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_M\}$ be the set of object classes. There are several methods for estimating the probability $P(x|\Omega = \omega_c)$ for x to belong to the object class ω_c. Typically, a training set of N patterns $\{x^t\}_{t=1}^{N}$,each pattern belonging to one of the classes in Ω, is used to estimate $P(x|\Omega = \omega_c)$, $c = 1, \ldots, M$. To classify a new pattern z, the maximum likelihood principle is often used:

$$\Omega^* = \arg\max_{\omega_c} P(z|\Omega = \omega_c) \tag{1}$$

In face detection $\Omega = \{f, nf\}$, that is the object classes are restricted to: face and non-face. In the following, we describe methods that explicitly estimate the probabilistic models of these two classes.

Moghaddam and Pentland [15] use Principle Component Analysis (PCA) to estimate $P(x|\Omega)$. Two probabilistic models are considered: a single Gaussian distribution and a Mixture-of-Gaussians. For computational feasibility, each mean normalized image pattern is decomposed into two mutually exclusive and complementary subspaces: the principle subspace F spanned by the M eigenvectors corresponding to the M largest eigenvalues and its complement $\bar{F}$ composed of the remaining $N - M$ eigenvectors. The component in the subspace $\bar{F}$ is called "distance-from-feature-space" (DFFS) which is an Euclidean distance. This value is also the residual reconstruction error which can be computed as:

$$\begin{aligned} \epsilon^2(X) &= \sum_{i=M+1}^{N} y_i^2 \\ &= ||x - \bar{x}||^2 - \sum_{i=1}^{M} y_i^2 \end{aligned} \tag{2}$$

[1] This data set is available at:
http://muc.ius.cs.cmu.edu/IUS/eyes_usr17/har/har1/usr0/har/faces/test/

where $y = \{y_i\}_{i=1}^{M}$ are the projected components in F.

The component in the subspace F is called "distance-in-feature-space" (DIFS) which can be interpreted in terms of the probability distribution of y in F. The authors show that in case of a Gaussian distribution, the probability density $P(x|\Omega)$ can be approximated using M components in the principle subspace F only. Hence, the approximated probability $\hat{P}(x|\Omega)$ can be computed as:

$$\hat{P}(x, \Omega) = \left[\frac{exp(-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i})}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}}\right]\left[\frac{exp(-\frac{\epsilon^2(X)}{2\rho})}{(2\pi\rho)^{(N-M)/2}}\right] \quad (3)$$

where $\{\lambda_i\}_{i=1}^{M}$ are the M largest eigenvalues and ρ is the arithmetic average of the eigenvalues in the orthogonal subspace $\bar{F}$.

In case $P_F(x|\Omega)$ cannot be adequately modeled using a single Gaussian, a Mixture-of-Gaussians model can be used.

To detect faces, eq. (1) is used. The density estimation $\hat{P}(x|\Omega)$ is computed for each image vector x at location (i, j) and the class Ω is determined. A system based on unimodal Gaussian with $M = 10$ dimensional subspace has been tested on the FERET data set and a correct detection of 97% is reported. However, as noted by many researchers, this data set is quite simple for the task of face detection. The advantage of this method is that it can be applied to detect other objects such as eyes, nose and mouth. A disadvantage of this system is that each window has to be projected onto a subspace before classification. This involves a matrix multiplication for every image window and the time spent is considerable. In addition, since non-face patterns are not used in the training process, the system is not robust in detecting faces in general scenes.

Schneiderman and Kanade [26] estimate the probability $P(x|\Omega)$ based on local appearance. Each image window x is decomposed into overlapping subregions of a fixed size.

$$x = \{r_i, p_i\}_{i=1}^{N_r}$$

where r_i is the subpattern of x at location p_i.

The authors note that alignment error for subregions is less than that of the whole template when some degree of geometric distortion occurs. More importantly, subregion decomposition has the power of emphasizing distinctive parts of the object. The statistical dependency is not modeled to simplify the conditional probability function. Hence, the probability $P(x|\Omega)$ can be computed as:

$$\begin{aligned} P(x|\Omega) &= P(\{r_i, p_i\}|\Omega) \\ &= \prod_{i=1}^{n_r} P(r_i, p_i|\Omega) \quad (4) \end{aligned}$$

In order to estimate the probability $P(r_i, p_i|\Omega)$, several preprocessing steps are carried out including linear projection, sparse coding, discretization. Both face (f) and non-face (nf) classes are modeled. The non-face patterns are collected in a "bootstrap" fashion as in [30]. A pattern x is declared a face pattern if the likelihood ratio is above a threshold λ.

$$\frac{P(x|\Omega = f)}{P(x|\Omega = nf)} > \lambda$$

This method appears superior to the others. The correct detection rate is greater than 99.6% for the FERET test set. It is also among the best performers when evaluated against the CMU test set. The results show it could handle well slightly rotated faces. A drawback of this method is the size of the window (64×64). The window must be large enough to capture local details of human faces.

In [5], Colmenarez and Huang use a family of Markov processes to model face and non-face distributions. Given a distribution function P, assume a first order Markov process, and let S be a list of indices that is some permutation of $(1, ..., N)$. We have:

$$P(x|S) = P(x_{S_1}) \prod_{i=2}^{N} P(x_{S_i}|x_{S_{i-1}}) \tag{5}$$

The goal of the learning process is to find probability models maximizing the Kullback divergence for the training set. The Kullback divergence, giving a non-negative measure for the difference between the two distributions, can be computed as:

$$H(S) = H(x_{S_1}) + \sum_{2}^{N} H(x_{S_i}||x_{S_{i-1}}) \tag{6}$$

where

$$H(x_i) = \sum_{x_i} P(x_i|\Omega = f) log \frac{P(x_i|\Omega = f)}{P(x_i|\Omega = nf)} \tag{7}$$

$$H(x_i||x_j) = \sum_{x_i} \sum_{x_j} P(x_i, x_j|\Omega = f) log \frac{P(x_i|x_j, \Omega = f)}{P(x_i|x_j, \Omega = nf)} \tag{8}$$

From a training set, the probability measure is computed for both face and non-face classes. Then for each pair of pixels x_i and x_j the divergence is computed. The problem now reduces to finding a set of indices S maximizing (6). This is computationally equivalent to the Traveling Salesman Problem. A greedy algorithm is applied to find an approximate solution. A table of likelihood ratios is calculated that allows the likelihood ratio of a new pattern to be computed with $N-1$ additions only. The training phase of the algorithm is quite fast since it requires only one pass through the training set to collect data statistics. The classification of a new pattern is also fast because only $N - 1$ additions are required. However, the fact that to calculate the divergence, the probability of all possible values of one pixel x_i must be estimated puts a constraint on the dimensionality of x_i. The system requantizes each window into 4 gray levels. This provides a partial explanation for the large number of false alarms returned by this method.

Yang et al. [37] present two methods using a mixture of linear subspaces. The first method uses (common) factor analysis which is a statistical method to model the structure of high dimensional data using only a small number of latent variables. Factor analysis assumes that the variance of a single variable can be decomposed into common variance and unique variance. Unlike principle component analysis (PCA) which considers the total variance of all variables, factor analysis analyzes only the common variance of the observed variables. An EM algorithm is used to estimate the model's parameters.

The second method uses Fisher Linear Discriminant. First, the data set is clustered into 25 face and 25 non-face classes using Self Organizing Map (SOM). It is interesting to see that nearby prototypes in the two dimensional feature map have similar intensity and pose (Figure 1). A projection matrix is then determined by Fisher Linear Discriminant which maximizes the ratio between the between-class variance and within-class variance. The whole training set is projected onto this subspace and Gaussian distributions are used to model each class conditional density. Parameters of the model are estimated with maximum likelihood. New patterns z are also classified by using the maximum likelihood principle as in (1). Good results have been reported for both methods. Nevertheless, important parameters such as the number of clusters for the face and non-face class in the second method are much dependent on the size of the training set. Generic rules for selecting those parameters are not known.

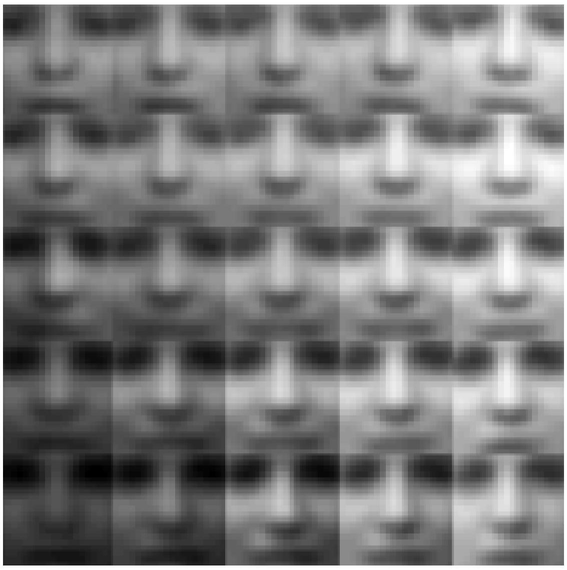

Figure 1: 25 face prototypes

Sung and Poggio [30] present a face detection system which models the distribution of face and non-face patterns. First, a modified version of the classical k-means clustering algorithms is applied to the training set of face patterns and non-face patterns to divide each set into 6 clusters. The result of this process is 6 face centroids and 6 non-face centroids together with their covariance matrices. The non-face patterns are generated in a "bootstrap" fashion by iteratively building a system using existing non-face patterns and collecting false positives of this system on a set of random images, then rebuild the classifier with the newly collected non-face patterns added to the current set. The advantage of this approach is that only those non-face patterns close to face clusters are collected and modeled. It is also possible that after each iteration, the face and non face patterns in the database are re-clustered to approximate the face distribution more accurately. A vector of 12 distances between the normalized window and the model's 12 clusters is computed. A two-valued distance metric is defined. The first value is the normalized Mahalanobis distance and the second value is the Euclidean distance between the test pattern and its projection to a lower dimensional subspace. These 12 two-valued distance measures are fed into a standard multilayer perceptron network classifier to identify face from non-face patterns. The system performs quite well on the CMU test set. However, the number of face and non-face clusters are chosen quite arbitrary. It is not clear how the system would perform for a different number of face and non-face clusters. As in [37], there is no rule for selecting these parameters.

Rowley et al. [21, 23] present a neural network based face detection system. A retinally connected neural network is used to discriminate face and non-face patterns. The non-face patterns are also generated by a "bootstrap" method as in [30]. The authors observe that with the same training process but with random initial weights and random initial non-face images, the detection result of faces are identical. Different errors or false alarms are made by different network. By using multiple networks and some heuristics, the number of false alarms produced by the system is reduced substantially. This system is robust in detecting frontal faces. The detection result is among the best reported in literature. This work is also extended to detect faces with in-plane rotation in [21, 24]. A derotation network is placed before the detector network to detect the orientation of each pattern. The detection rate for frontal faces is degraded, but a good detection result is reported for detecting faces of different orientations. Nevertheless, the pose variation problem is not considered.

The SNoW (Sparse Network of Winnows) learning architecture is used in [39]. This generic learning technique is tailored for learning in domains in which the number of features is very large. This method gives the best frontal face detection result in literature. An detection rate of 94.8% is reported for the CMU test set.

A method that makes use of information theory is described in [13]. From a set of face and non-face patterns, Most Informative Pixels (MIP) are defined as those maximizing the class separation between face and non-face classes. The classification is carried out on the MIP. The "distance-from-feature-space" metric (DFFS) (see 3.1) is used for classification. A detection rate of 94.1% is reported for the MIT test set and 88.3% for an early version of the CMU test set[2].

Osuma et al. [16] apply a Support Vector Machine (SVM) to the face detection task. SVM is a new technique in pattern recognition which aims at minimizing the upper bound on the generalization error. In order to train a large data set with SVM

[2] There are 42 images with a total of 169 faces.

efficiently, a decomposition algorithm is proposed in which a subset of the original data set is used and iteratively updated to train the classifier. A optimal condition and a strategy for improvement is specified so that after each iteration the system can decide if it has reached the optimal solution and define a way to improve the cost function if it has not. The system is trained with a large data set. The detection accuracy is comparable to that of other best methods.

In [34], Wu at el. use a fuzzy pattern matching method. First, both skin color and hair color distributions are estimated. Each pixel in the image is associated with a skin color likeness and hair color likeness value. These values range from 0 to 1. The skin color distribution model (SCDM) and the hair color distribution model (HCDM) are a fuzzy set of skin color and hair color respectively. Five head shape models are then built including frontal view, left side view, right side view, left diagonal view and right diagonal view from a training set of images. Each rectangular face region is selected and so are the skin part and the hair part in each region. The rectangle of the face region is divided into m × n square cells and each cell is assigned two values: the proportion of the skin area and the hair area in it. The values in each cell of a model is the average value of the corresponding cells in the training faces. The frontal view model is built based on the frontal view, 15 degree rotated to the left and 15 degree rotated to the right faces. Faces rotated 30 and 45 degree to the left (or right) are used to create left (or right) diagonal view models. Faces rotated 60, 75 and 90 degree to the left (or right) are used to create left (or right) side view models. To detect faces in an image, all rectangular regions of a given size are compared with the head shape models. Each rectangle is divided into m × n square subregions and each of them corresponds to a cell in a head shape model. The matching degree between a rectangle and a head shape model is done using fuzzy theory ([34] eq. (12)(11))

$$Match(rect, model) = \frac{\sum\limits_{square \in rect} match(square, cell)}{m \times n} \tag{9}$$

A drawback of this method is that it fails when the face regions are merged with other regions which also have a skin color.

3.2 Rule-based classification

One of the earliest face detection systems is proposed in [35]. A hierarchical knowledge based pattern recognition method is applied. Faces are searched for at three levels in which a template candidate is verified by a set of rules. First, an image is divided into equal cells of size $n \times n$. At level one, all template 4 × 4 cells are searched for face candidates. An example of a rule at this level is that 4 cells have a uniform gray level. Face candidates from level one are passed onto level two in which the resolution of each face candidate is doubled, that is the size of each cell is reduced by half and the template now consists of 8 × 8 square cells. Another set of rules are used to further verify face candidates at this level and remaining candidates are passed to level three. At level three, more discriminative rules are applied. Features for decision include the length, slope, curvature, concavity, center of gravity, and the coordinate of the ends of an edge. Only eyes and mouth edges are studied. This method is simple and it provides an efficient way to locate faces initially. However, the alignment problem due to the use of coarse resolution at level one and two needs more attention.

Dai and Nakano [6] propose an approach based on face texture and color information. A face texture model is constructed based on the space grey-level dependency matrix (SGLD). A SGLD matrix consists of elements $P_{ab}(m,n)$ where $P_{ab}(m,n)$ is the number of occurrences in which two pixels displaced by a vector (m,n) in the image have the grey level a and b, respectively. From a SGLD matrix, a set of textural features is derived. Accordingly, a set of feature matrices whose elements are textural features is formed. Based on these feature matrices and observations of the intensity level of facial images a set of inequalities is derived. Parameters of the inequalities are determined by experiments. This set of inequalities forms the face texture models. In [6] the I component of the YIQ color space is used for discriminating between skin and non-skin region. A window of size 16×20 is scanned through the image. At each image location, if the average intensity of the window does not fall in a certain range, the window is rejected. Otherwise, the feature matrices are computed and the window is classified as a face if all equalities of the face texture model are satisfied. One disadvantage of this method is that the SGLD matrix has to be computed for every window which is quite computationally expensive.

4 Feature-based detection

One of the major shortcomings of template-based methods is that they are very sensitive to both rotation and scaling. To detect faces of different scales, the input image or the template can be resized to an appropriate value so that classification or matching can be carried out. The problem of in-plane and out-of-plane rotation (or pose variation), however, is more serious. In order to detect rotated faces, another training set of faces of different views and rotations may be required, or a completely new set of rules must be devised.

Feature-based methods offer a sensible solution to both rotation and scaling problems. In this approach, various types of facial features are detected and, based on the human face model, faces are declared using these features. Human skin color is an important feature. It is a rather surprising finding that the human skin color falls into a small range in different color spaces regardless of races [36]. Many researchers have taken advantage of this in their approach to the problem.

Feature-based methods can be divided into two groups. The first is a bottom-up approach in which facial features are detected and grouped according to their geometric relationship. Bottom-up methods are summarized in the next section. The other approach is top-down where facial regions are first located using global color information. Each region is then further analyzed and non-face regions are rejected. Methods in this class are discussed in section 3.2

4.1 Bottom-up methods

One of the earliest methods is proposed by Govindaraju et al. in [8]. In this method, a face model consists of features which are described in terms of curves and structural relationships between them. Different sets of features are used in two stages. First, curves of the face are extracted to generate face candidates. Features detected in the first stage are grouped by matching against the face model using a cost function. A group of features with a cost less than a predefined threshold forms a hypothesized face candidate. Eyes and symmetry about the vertical axis are used in the second stage to test the hypothesized face candidates. This method is quite limited since several assumptions have been made. For example a face is bounded by edges, faces are frontal and the size of the face is within a fixed range. Because of the variability of shape between different people, it is hard to devise a reliable cost function.

In [12], Leung et al. proposed the use of random label graph matching to group features. First, features such as eyes, nose and nostrils are identified by convolving the input image with a set of Gaussian derivative filters at different scales and orientations. A vector of filter responses at a particular spatial location is then matched against a template vector response. A feature is detected at a location if the degree of matching is above a threshold. From a pool of facial features, constellations are found and ranked based on the probability density of the normalized distances between features. A drawback of this method is that since the distances between facial features are used in the matching, it can accommodate only faces with little rotation in depth. Also, the computation time for grouping features is substantial. A controlled search in which features are searched only in their expected region has been used to speedup the search process.

Another feature-based system is presented by Yow and Cipolla [42] [41]. Features are detected using elongated Gaussian derivative filters. This filter is able to detect features like corner of the eyes, mouth and nose, even in the case when faces are not frontal. In order to reduce the number of false alarms, other filtering techniques are used. Also, an upper limit of the edge intensity at each feature location is imposed to eliminate features which are falsely detected at strong edge locations. After a large number of feature points are detected, edges around each point are examined and similar edges are linked. Those feature points that have roughly parallel edges on both sizes are selected. They are then classified into different feature classes based on measurements such as edge length and edge strength. The facial feature classes are eyebrow, eye, nose and mouth. The features can now be grouped based on the face model. An example of the measurement of the face model is the aspect ratio of a facial region. False positives are eliminated further by the use of a belief network. Using the Gaussian derivative filters, corners of facial features are reliably detected. It is also robust to rotation and change of viewpoint. However, many background image locations are also detected as features and it makes the task of feature grouping considerably hard. Also, the use of many hard-coded face model rules may lead to serious model error or may not be sufficient to discriminate face from non-face arrangement. In addition, although most feature corners are detected, they are not well localized due to the variability of poses and facial expressions. The performance of this method on the CMU test set is not very encouraging. This could be explained by the low image quality of the CMU test set A.

Cai and Goshtasby [4] proposed another method using color information. First, the skin color is classified adaptively. Facial features that do not have skin color such as eyes and mouth appear as a small blob inside the face region. By detecting small non-skin color regions within a skin region, a rough estimation of the facial features are located. According to different possible arrangements of facial features in the face region, the face model is overlaid in the face region and the cross correlation coefficient is computed. The face model is constructed from a set of normalized faces. A threshold is selected and regions whose matching rate higher than this threshold are declared as faces. A pitfall of this method is that the mechanism of overlaying the model on the face region is rather ad hoc. It is based on a set of rules that could have model error. Also, it is expected that the system would not cope well with pose variation because only a limited number of templates are used in the matching process. It is not clear whether an addition of templates of different views would increase the performance of the system. Another drawback of this method is that when facial features are small, they also appear having skin color and the algorithm will fail to detect faces.

Jebara et al. [10] use another method to detect facial features in the face region. Skin color regions are first segmented and feature detection is carried out by using a symmetry transformation [19, 11]. The detected features are tracked in a sequence and a 3D structure of the face can be constructed.

Sun et al. [29] also make use of the local symmetry information. A local symmetry map [14, 19] is obtained and fused with the color map which present the skin color likeness of each pixel. Facial features correspond with the local maxima of the map resulted. These features are then grouped according to a face geometry model.

Both methods described above are based on a assumption that facial regions can be segmented from the background reliably using color information only. This assumption limits the applicability of both methods.

4.2 Top-down methods

In this class of methods, face regions are segmented from the background using color information. Different color spaces have been used. Face candidate are formed from these regions and further verified. The advantage of these methods is that they are very fast. Faces can be detected in real time. However, no method seems able to resolve the problem when background objects also have skin color, especially when they are merged with the face region. Also, these methods are very sensitive to varying lighting conditions. In the following, we summarize some of the methods in this category.

In [27], skin-like regions are first segmented. A region growing algorithm is applied at a coarse resolution of the segmented image to determine connected components. Small size components are rejected and a best-fit ellipse is then computed for each connected component on the basis of moments. Those ellipses that are good approximations of the connected components are selected and considered as face candidates. The goodness measure is computed based on the ratio of the number of pixels of the region outside the ellipse and pixels inside the ellipse that do not belong to the region against the size of the region. These face candidates can be further verified by searching for facial features inside the ellipses. Facial feature detection is carried out by examining the grey scale x- and y-relief of the connected components or using a modified version of the watershed algorithm.

Saber and Tekalp [25] propose another skin color segmentation algorithm using Gibbs Random Field filtering. The result of this filtering process is a collection of segmented skin regions. Each region is then approximated by an ellipse. The distances between each region shape and its ellipse are computed using the Hausdorff distance measure. Region and ellipse pairs whose distance is greater than a predefined threshold are rejected. Feature detection is then carried out inside the ellipse for each of the remaining regions. First, the two eyes are detected by considering "holes" inside the ellipse mask. The centroid of each detected hole are candidates for eyes. Based on the observation on the relative positions of eyes in a face, five cost functions are derived and the two holes that have a minimum weighted sum of these cost functions are marked as eyes. The tip of the nose and the mouth are located based on the location of the eyes. The weight used in combining cost functions and parameters used in locating the nose and mouth are determined empirically.

In [40], a multiscale image segmentation method [1] is used to extract homogeneous regions at different scales. The segmentation result is determined by a homogeneity scale parameter δ_g and a spatial scale parameter δ_s. The CIE LUV color space without the luminance component is used to build the human skin color model. Parameters of the distribution are obtained by maximum likelihood estimation. A threshold is used to determine if a pixel has a skin color and a region is considered skin region if above 70% of its pixels have skin color. The segmentation is performed at different scales. At each scale, regions are merged until the shape of the merged region is approximately elliptic. The goodness of the approximation is based on the number of pixels of the region inside its ellipse shape. Regions that have a ratio between the major and minor axis less than 1.7 are considered face candidates. A candidate region with dark areas or holes inside are considered human faces based on the observation that facial features either do not have skin color or are darker than their surrounding areas.

In contrast to merging, Wei and Sethi [33] use an iterative partitioning of the human skin region to detect faces. First, the binary image of the segmented skin region is obtained by performing skin color classification at each pixel location. A morphological closing is then applied followed by a opening to remove small regions and detach regions which are connected by a thin strip. The structuring element size is adaptive to the size of the input image. A region that can be approximated well by an ellipse is considered a face candidate region. The ratio between the length of the major and minor axis together with the ratio between the area of the ellipse and the actual area of the region are used to assess how well a region is approximated by its ellipse. Other regions will be partitioned into smaller regions that may be approximated by ellipses. Subregions whose size are too small are rejected. Furthermore, subregions whose shapes are approximately elliptic are considered face candidate regions and the rest are further partitioned. The partitioning process stops when all subregions are too small and no face candidate is found. Faces are detected by verifying features in the face candidate regions. A histogram-based thresholding scheme is applied. The histogram of the Y component of each face region candidate is computed and a value is chosen such that 18% of the pixels have an intensity below it. After applying the threshold to each face candidate region with its computed threshold, facial features should correspond to the dark parts. Based on relative positions of these dark parts and their shapes within a face region, a face region is classified as face or not.

Wang and Sung [32] also use morphological operations and knowledge about the human face for shape analysis in detecting frontal view faces. Face regions are first extracted using color information. A morphological closing operation is performed to fill holes inside the face regions. The face contour is then extracted by analyzing the skeleton of the face region. The purpose is to separate ears and neck parts from the face region. The skeleton of the segmented region is obtained using the morphological skeleton operation. A line tracing and merging algorithm is developed and applied to extract lines from the skeleton image. All lines are examined by a set of rules and their endpoints for fitting the contour of the face are collected. Some points of the color edge image are also added to the contour fitting points set according to the same set of rules. The face contour is modeled as an ellipse using the least square fitting algorithm with the set of collected points to fit. Once the face contour has been extracted, facial features are detected using a algorithm that is similar to the one used in [28].

Terrillon et al. [31] use a neural network to discriminate between human face and other objects. Skin color regions are extracted using color information and 11 lowest-order moments are computed for each region. These moments are translation, scale and in-plane rotation invariant. A feed-forward multilayer perceptron is used to classify these regions using the 11 moments as inputs. The method works only when facial regions can be well segmented from the background.

5 Experiments

5.1 Test set and detection measures

One hundred frames from a news video are were for evaluation. The frames were taken at every 300 frames of the sequence. Although the number of frames are quite small, there are a lot of variations in size, rotation and pose of faces which appear in this test set. A groundtruth was prepared for this test set in which only faces with both eyes visible are included. Small faces were also excluded. There are 85 faces in total of which 37 faces are frontal and 48 faces are non-frontal.

Each implementation is run to detect faces in all 100 frames. The total number of detected frontal faces, detected non-frontal faces and false alarms are reported. In this experiment, the execution time is not considered, though clear differences among methods were observed.

In addition, some template-based methods have been evaluated against a test set of 1905 face and 2200 non-face patterns.

5.2 Face detection methods

A number of methods were evaluated. In this section, we will briefly describe how each method was implemented. Available libraries have also been used for evaluation. For methods that require learning, a training set of 3000 face and 9443 non-face patterns were used which were independent of our test. The data are provided by

Bunschoten [3] and extended with more non-face patterns by using a “bootstrap” method as described in [30].

Face detection methods:

1. **Rowley’s neural network** [21]. This face detector is obtained from CMU. This is one of the most significant works in face detection. It has been used in several related projects including Informedia and NameIt.

2. **Colmenares and Huang’s information theory based** [5]. This method is quite different from other classification-based methods in that the learning phase is very fast. The classification of a new pattern is also fast.

3. **Template matching.** The face template is the one used in [4]. A template of size 30x30 is used. A fast scanning mechanism is devised to reduce the number of tested windows.

4. **Simple color** [27]. The $c1c2c3$ color space [7] is used to segment the face regions.

5. **Wei’s partitioning** [33]. Each rejected skin color region is further partitioned into subregions of which some could be face regions.

6. **Principle Component Analysis** [15]. In contrast to [15], non-face patterns are also used.

7. **Fisher Linear Discriminant** [37].

8. **Texture based classification** [6]. The parameters of this method is chosen by using a much larger training set than in [6]. The parameters are chosen so that 100% of the face patterns satisfy all rules of the algorithm.

Feature detection:

1. **Yow’s method** [42]. An elongated Gaussian kernel of ratio 1:3 is applied to detect facial features.

2. **Reisfeld’s method**[19, 20]. Similar techniques have been used for feature detection in [10, 29].

5.3 Result

The following is the results of face and feature detection methods in this experiment. Parameters are selected manually. Methods in table 1 are evaluated against a set of 100 frames from the news videos. Methods in table 2 are evaluated against a set of 1905 face and 2200 non-face patterns.

Figure 2 and 3 are sample outputs of the two feature detection methods.

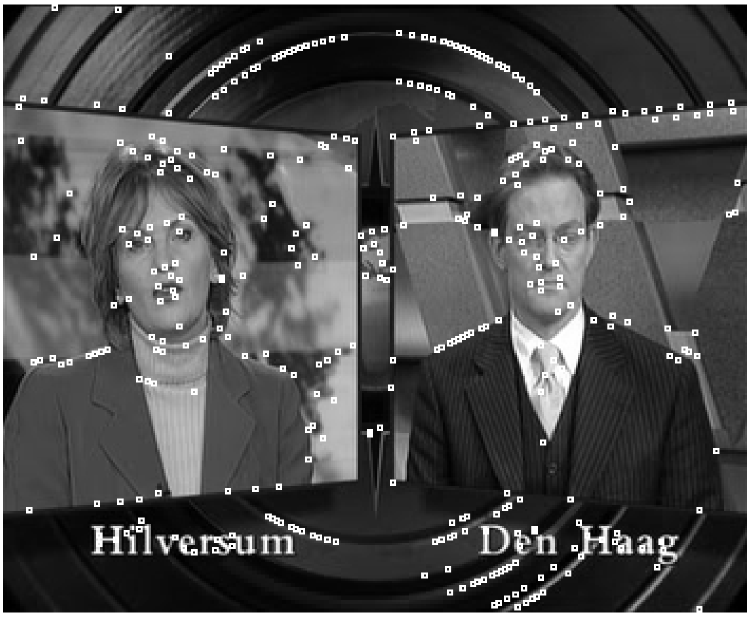

Figure 2: Result of applying an elongated Gaussian filter (threshold = 1)

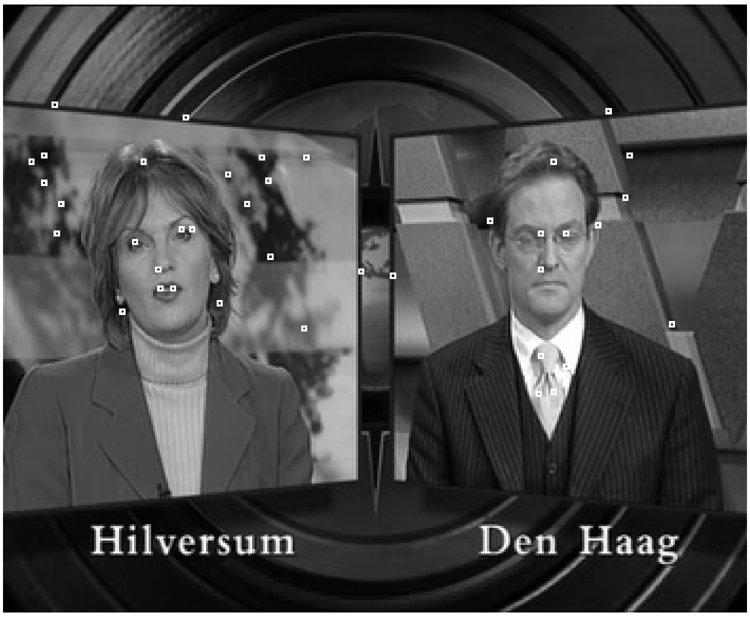

Figure 3: Reisfeld's radial dark transform (kernel size = 3, threshold = 30)

Methods	Parameters	Detection Rates(%)			False De.
		Fr	Non-fr	Total	Total
Rowley	w/o eye detection	91.89	58.53	72.94	7
	with eye detection	16.21	18.75	17.65	0
Colmenarez & Huang	threshold = 50	75.68	20.83	44.71	39
Template matching	threshold = 0.8	91.89	41.67	63.52	213
	threshold = 0.82	86.48	47.91	64.71	98
Simple color		18.92	35.42	28.24	29
Wei partitioning		18.92	35.42	28.24	61

Table 1: Performance of face detection methods against a test set of 37 frontal faces and 48 non-frontal faces. Images are taken from a news video.

Methods	Parameters	Error Rates
PCA - zero mean, unit norm	5 eigenvectors	22.66%
	10	4.51%
	20	1.95%
	45	1.71%
	60	1.85%
Fisher Linear Discriminant	8 classes	4.19%
	18	5.97%
	32	9.99%
	50	25.14%
Dai's texture-based	100% correct detection rate	22.43% false alarms

Table 2: Performance of face detection methods with a test set of 1905 face and 2200 non-face patterns.

5.4 Discussion

Experimental results show that Rowley's face detector has the best performance. The detection results of our implementations are not good. This could be explained by the fact that our training set is quite limited. A total of 4905 face patterns are used in comparison with 16,810 face patterns used in [37] or 118920 face patterns used in [26].

The classification method using PCA is simple yet very effective. With an error rate of 1.71%, a high number of of tested windows will result in many false alarms. The method using Fisher Linear Discriminant is among the best performers proposed in literature. However, the result shown here is very poor. It seems that the small number of face patterns does not represent each face class reliably. The classification time for both method is considerable since each new pattern has to be projected onto a lower dimensional space, which involves a matrix multiplication. The texture-based method Dai's could give a very high detection rate for face patterns, but it also give a lot of false alarms. This could be useful as a verification mechanism.

Methods that make use of plain color information are not reliable for the news video dataset. Most of color-based methods assume that the facial region can be segmented from background using only color information. This assumption does not usually hold for news video.

As for bottom-up feature-based methods, the feature detection step is very important. The method used by Yow in [42] does reliably detect corners of facial features. However, a lot of false alarms are also generated (Figure 2). The second method gives a better detection results. However, it is very sensitive to scale. More importantly, in both methods the detection result is not well localized. This make the relative distances between facial features unreliable for the feature grouping step.

Methods	Advantages	Disadvantages
Neural Network SVM PCA FLD Distribution based Local appearance	Robust for detecting frontal faces. Powerful when combined with other methods that reduces the search space.	Slow Hard to extend Require a large set of training examples.
Information based	Fast training and classification.	High number of false alarms.
Rule-based	Easy to incorporate prior knowledge	Model-error prone Not very discriminative
Feature based (bottom-up)	Good solution to the rotation problem	Hard to detect facial features
Color-based	Fast	Not reliable in a general scene.

Table 3: Face detection methods: advantages and disadvantages

6 Conclusion

Face detection is still a difficult problem in Computer Vision. In this report we have discussed several methods for face detection in literature. Experiments have been carried out to evaluate the effectiveness of some methods when applied to news video data.

Statistical pattern recognition -based methods give roughly the same detection rate. They are robust for detecting frontal faces. They also can detect slightly rotated faces. Nevertheless, for unrestricted imaging conditions it is necessary to have a face detector that could detect well faces which are rotated in depth.

Neural network-based as well as other template-based methods can detect frontal faces. However, it is very hard to extend these systems to detect faces with out-of-plane rotation. Another large training set of faces with different poses may be required and even if such a training set is available, it is not guaranteed that existing classification methods will still give a good performance. The feature-based approach promises an elegant solution to the problem of out-of-plane rotation. However, it is very dependent on the feature detection phase. With a large variation in facial expression, it is not surprising that the facial feature detection in a general scene is not reliable. This makes the feature-based methods not as robust as template-based methods.

There are some directions which may lead to a successful face detection system. It is reasonable to expect that other methods that classify the whole image windows will not improve the detection result significantly. A promising approach is being carried out at MIT [9]. Instead of classifying the whole image window, the system detects different components of a face such as the eyes, nose and mouth by using a learning method. The geometric relations among these components are grouped using another learning method. The system is expected to handle in-depth rotation much better than existing template-based methods. Also, it should be more robust than the existing feature-based methods because of the learning processes.

In detecting faces with different head poses, the discriminative powers of different parts of the face change. This observation was made in [17, 2]. An investigation into the discriminative powers of different parts of the face are useful for face detection as well as recognition.

Color information has been used with limited success. This is largely because of the fact that the background may also have human skin color. However, the face skin color and the background color are still different. A better color segmentation method may successfully segment the face from background. The difficulty is that the size of the face has to be large enough for the segmentation to be reliable. Also, facial features and lighting condition may lead to undesirable segmentation result.

Finally, other feature such as the face outline or motion information may be a strong cue for detecting human faces in a video.

References

[1] N. Ahuja. A transform for multiscale segmentation by integrated edge and region detection. *IEEE PAMI*, 18(12):1211–1235, 1996.

[2] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE PAMI*, 1993.

[3] R. Bunschoten. Face detection using neural networks. Master's thesis, University of Amsterdam, 1999.

[4] J. Cai and A. Goshtasby. Detecting human faces in color images. *Image and Vision Computing*, 18:63–74, 1999.

[5] A. Colmenarez and T. Huang. Face detection with information-based maximum discrimination. In *Proc. of CVPR'97*, 1997.

[6] Y. Dai and Y. Nakano. Face-texture model based on sgld and its application in face detection in a color scene. *Pattern Recognition*, 29(6):1007–1017, 1996.

[7] T. Gevers and A. Smeulders. Color based object recognition. *Pattern Recognition*, 32:453–464, March 1999.

[8] V. Govindaraju, S.N. Srihari, and D.B. Sher. A computational model for face location. In *Proc. of The Third International Conference on Computer Vision*, pages 718–721, 1990.

[9] B. Heisele and T. Poggo. Face detection. http :// www.ai.mit.edu/ lab/ abstracts/ pdf/ z-heisele.pdf.

[10] T. Jebara and A. Pentland. Parameterized structure from motion for 3d adaptive feedback tracking of faces. In *Proc. of CVPR'97*, 1997.

[11] M.F. Kelly and M.D. Levine. Annular symmetry operators: A method for locating and describing objects. In *Proc. of The Fifth International Conference on Computer Vision*, pages 1016–1021, 1995.

[12] T.K. Leung, M.C. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. In *Proc. of The Fifth International Conference on Computer Vision*, 1995.

[13] M. Lew and N. Huijsmans. Information theory and face detection. In *Proc. of ICPR'96*, 1996.

[14] C.-C. Lin and W.-C. Lin. Extracting facial features by an inhibitory mechanism based on gradient distributions. *Pattern Recognition*, 29(12):2079–2101, 1996.

[15] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE PAMI*, 19(7), 1997.

[16] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Proc. of CVPR'97*, Puerto Rico, 1997.

[17] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proc. of CVPR'94*, 1994.

[18] P. J. Phillips, P. J. Rauss, and S. Z. Der. FERET (face recognition technology) recognition algorithm development and test results. Technical Report ARL-TR-995, US Army Research Laboratory, 1996.

[19] D. Reisfeld and Y. Yeshurun. Robust detection of facial features by generalized symmetry. In *Proc. of 11th IAPR International Conference on Pattern Recognition*, volume 1, pages 117–120, 1992.

[20] D. Reisfeld and Y. Yeshurun. Preprocessing of face images: Detection of features and pose normalization. *Computer Vision and Image Understanding*, 71(3):413–430, 1998.

[21] H. A. Rowley. *Neural network-based face detection.* PhD thesis, Computer Science Department, Carnegie Mellon University, 1999.

[22] H. A. Rowley, S. Baluja, and T. Kanade. Test images for the face detection task. http://muc.ius.cs.cmu.edu/IUS/eyes_usr17/har/har1/usr0/har/faces/test/.

[23] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE PAMI*, 20(1):23–38, 1998.

[24] H. A. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. In *Proc. of CVPR'98*, pages 38–44, 1998.

[25] E. Saber and A. M. Tekalp. Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions. *Pattern Recognition Letters*, 19(8):669–680, 1998.

[26] H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *Proc. of CVPR'98*, 1998.

[27] K. Sobottka and I. Pitas. Looking for faces and facial features in color images. In *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Application*, volume 7, pages 124–137. Russian Academy of Sciences, 1997.

[28] R. Stiefelhagen and J. Yang. A model-based gaze tracking system. *International Journal of Artificial Intelligence Tools*, 6(2):193–209, 1997.

[29] Q. B. Sun, W. M. Huang, and J. K. Wu. Face detection based on color and local symmetry information. In *Proc. of the Third International Conference on Automatic Face and Gesture Recognition*, pages 130–135, Nara, Japan, 1998.

[30] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE PAMI*, 20(1):39–51, 1998.

[31] J.-C Terrillon, M. David, and S. Akamatsu. Automatic detection of human faces in natural scene images by use of a skin color model and of invariant moments. In *Proc. of the Third International Conference on Automatic Face and Gesture Recognition*, pages 112–117, Nara, Japan, 1998.

[32] J-G Wang and E. Sung. Frontal-view face detection and facial feature extraction using color and morphological operations. *Pattern Recognition Letters*, 20:1053–1068, 1999.

[33] G. Wei and I. K. Sethi. Face detection for image annotation. *Pattern Recognition Letters*, 20(11):1313–1321, 1999.

[34] H. Wu, Q. Chen, and M. Yachida. Face detection from color images using a fuzzy pattern matching method. *IEEE PAMI*, 21(6):557–563, 1996.

[35] G. Yang and T. Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53–63, 1994.

[36] J. Yang, W. Lu, and A. Waibel. Skin-color modeling and adaptation. Technical Report CMU-CS-97-146, School of Computer Science, Carnegie Mellon University, May 1997.

[37] M. H. Yang, N. Ahuja, and D.Kriegman. Face detection using mixtures of linear subspaces. In *Proc. of the Fourth International Conference on Automatic Face and Gesture Recognition*, Paris, France, March 2000.

[38] M. H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. Submitted.

[39] M. H. Yang, D. Roth, and N. Ahuja. A snow-based face detector. In *Advances in Neural Information Processing Systems 12*. MIT Press, 2000.

[40] M.H. Yang and N. Ahuja. Detecting human faces in color images. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 127–130, Chicago, 1998.

[41] K.C. Yow. *Automatic human face detection and localization.* PhD thesis, Downing College, the University of Cambridge, 1998.

[42] K.C. Yow and R. Cipolla. Scale and orientation invariance in human face detection. In *Proc. British Machine Vision Conference*, pages 745–754, 1996.

Acknowledgements

This research is sponsored by the ICES Multimedia Information Analysis project and TNO-TPD.
The face dataset is provided by Roland Bunschoten.

ISIS reports

This report is in the series of ISIS technical reports. The series editor is Rein van den Boomgaard (rein@science.uva.nl). Within this series the following titles are available:

References

[1] Hieu T. Nguyen, M. Worring, and A. Dev. Robust motion-based segmentation in video sequences. Technical Report 4, Intelligent Sensory Information Systems Group, University of Amsterdam, December 1998.

[2] J. Vendrig, M. Worring, and A.W.M. Smeulders. Filter image browsing: interactive image retrieval by using database overviews. Technical Report 5, Intelligent Sensory Information Systems Group, University of Amsterdam, December 1998.

[3] M. Worring and A.W.M. Smeulders. Content based internet access to paper documents. Technical Report 6, Intelligent Sensory Information Systems Group, University of Amsterdam, December 1998.

[4] S.D. Olabarriaga, P.R. Pfluger, and A.W.M. Smeulders. Piecewise dm: A locally controllable deformable model. Technical Report 7, Intelligent Sensory Information Systems Group, University of Amsterdam, 1999.

[5] S.D. Olabarriaga and A.W.M. Smeulders. Interaction in the segmentation of medical images, a survey. Technical Report 8, Intelligent Sensory Information Systems Group, University of Amsterdam, 1999.

[6] R. v.d. Boomgaard, E.A. Engbers, and A.W.M. Smeulders. Decomposition of separable concave structuring functions. Technical Report 9, Intelligent Sensory Information Systems Group, University of Amsterdam, 1999.

[7] G. Stijnman and R. v.d. Boomgaard. Background estimation in video sequences. Technical Report 10, Intelligent Sensory Information Systems Group, University of Amsterdam, 2000.

[8] T.V. Pham and M. Worring. Face detection methods: A critical evaluation. Technical Report 11, Intelligent Sensory Information Systems Group, University of Amsterdam, 2000.

You may order copies of the ISIS technical reports from the corresponding author or the series editor. Most of the reports can also be found on the web pages of the ISIS group (http://www.science.uva.nl/research/isis).

Intelligent Sensory Information Systems
University of Amsterdam
The Netherlands