ELSEVIER

Pattern Recognition Letters 23 (2002) 451–461

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Face detection by aggregated Bayesian network classifiers ☆

Thang V. Pham *, Marcel Worring, Arnold W.M. Smeulders

ISIS, Informatics Institute, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

Abstract

A face detection system is presented. A new classification method using forest-structured Bayesian networks is used. The method is used in an aggregated classifier to discriminate face from non-face patterns. The process of generating non-face patterns is integrated with the construction of the aggregated classifier. The face detection system performs well in comparison with other well-known methods. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Face detection; Aggregated classifiers; Bayesian network classifier; Optimum branching

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 (in-plane rotations), and with different head poses (out-of-plane rotations). The imaging conditions, including illumination direction and shadow, also affect the appearance of human faces. Moreover, human faces are non-rigid objects, as there are variations due to varying facial expressions. Presence of other devices such as glasses is another source of variation. Facial attributes such as make-up, wet skin, hairs and beards also contribute substantially to the variation of facial appearance. In addition, the appearance differences among races, and between male and female are considerable. A successful face detection system should be able to handle the multiple sources of variation.

A large number of face detection methods have been proposed in the literature. Face detection methods can be broadly divided into: model-based detection, feature-based detection and appearance-based detection.

In the model-based approach, various types of facial attributes such as the eyes, the nose and the corner of the mouth are detected by a deformable geometrical model. By grouping the facial attributes based on their known geometrical relationships, faces are detected (Leung et al., 1995; Yow and Cipolla, 1996). A drawback of this approach is that the detection of facial attributes is not reliable (Leung et al., 1995), which leads to systems that are not robust against varying facial expressions and presence of other devices. This approach is better suited

☆ This research is sponsored by the ICES Multimedia Information Analysis project and TNO-TPD.

* Corresponding author. Tel.: +31-20-525-7553.
E-mail address: vietp@science.uva.nl (T.V. Pham).

0167-8655/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-8655(01)00177-5

for facial expression recognition as opposed to face detection.

Among the feature-based approach, the most obvious feature is color. It is a rather surprising to find that the human skin color falls into a small range in different color spaces regardless of race. Many researchers have taken advantage of this fact in their approach to the problem (Yang and Waibel, 1996; Wei and Sethi, 1999). Typically, regions with skin color are segmented to form face candidates. Candidates are further verified on the bases of the geometric face model. We choose not to use color information in this paper. It is partly because of the lack of a common color test set to evaluate different methods.

In the appearance-based approach, human faces are treated as a pattern directly in terms of pixel intensities (Sung and Poggio, 1998; Rowley et al., 1998). A window of fixed size $N \times M$ is scanned over the image to find faces. The system may search for faces at multiple image scales by iteratively scaling down the image with some factor. At the core of the system is a classifier discriminating faces from non-face patterns. Each intensity in the window is one dimension in the $N \times M$ feature space. The appearance-based methods are often more robust than model-based or featured-based methods because various sources of variations can be handled by their presence in the training set.

This paper presents a face detection system in the appearance-based approach. The class/non-class classification problem needs to be addressed because it is not possible to obtain a representative set of non-face patterns for training. Furthermore, because of the manifold of sources of variation, a complex decision boundary is anticipated. In addition, the classification methods should have a very low false positive rate since the number of non-face patterns tested is normally much higher than that of face patterns. Also due to a large number of patterns which need to be tested, a fast classification step is desirable.

The paper is organized as follows. The following section gives an overview of appearance-based classification methods. The construction of an aggregated classifier is described in Section 3. Section 4 presents a new classification method using forest-structured Bayesian networks. The face detection system is described in Section 5. Experimental results are given in Section 6.

2. Literature on appearance-based face detection

It is the classification method and the type of features that characterize different appearance-based face detection systems. Many techniques from statistical pattern recognition have been applied to distinguish between faces and non-face patterns. The one-class classification problem can be solved by designing a mapping which concentrates the one class into a point while mapping the other, the non-class as widely spread over the feature space as possible as proposed by Gelsema and coworkers (cf. Landeweerd et al., 1983).

Let $X = (X_1, X_2, \ldots, X_n)$ be a random variable denoting patterns spanning the $n = N \times M$-dimensional vector space $\mathscr{R}$. Let $x = (x_1, x_2, \ldots, x_n)$ be an instantiation of X. In addition, let $Y = \{0, 1\}$ be the set of class labels, face and non-face, respectively. Furthermore, let the two-class conditional probability distribution be $P_0(X)$ and $P_1(X)$. Once both $P_0(X)$ and $P_1(X)$ are estimated, the Bayes decision rule (Duda and Hart, 1973) may be used to classify a new pattern:

$$\varphi(x) = \begin{cases} 0 & \text{if } \log \frac{P_0(x)}{P_1(x)} \geqslant \lambda, \\ 1 & \text{otherwise}, \end{cases} \tag{1}$$

where λ is an estimation of the log-ratio of the prior probability of the two classes. When it is not possible to obtain such approximation, one may assume equal class prior probabilities, that is $\lambda = 0$. This leads to the maximum likelihood decision rule. This leaves the question of how to learn $P_y(X)$ effectively. Generally, it is not possible to estimate $P_1(X)$. This problem is often ignored. For the current face detection problem and the current dataset, we still form both $P_0(X)$ and $P_1(X)$. As a consequence, the solution will be specific for this dataset only. When the dataset grows large (we have 160 000 random patches), it will in the end be representative for the class of non-faces.

Moghaddam and Pentland (1997) use principle component analysis to estimate the class conditional density. The vector space $\mathscr{R}$ is transformed into principle subspace E spanned by the V eigenvectors corresponding to the V largest eigenvalues and its complement $\bar{E}$ composed of the remaining eigenvectors. The authors show that in case of a Gaussian distribution, $P_y(X)$ can be approximated using the V components in the subspace E only. In case $P_y(X)$ cannot be adequately modeled using a single Gaussian, a mixture-of-Gaussians model can be used. A drawback of this method is that no guidelines are given to determine the number of dimension V. In addition, as each pattern is projected on to a subspace before classification, a matrix multiplication is involved. This is not desirable when the classification time is an important factor.

Sung and Poggio (1998) present a face detection system which models $P_0(X)$ and $P_1(X)$, each by six Gaussian clusters. To classify a new pattern, a vector of distances between the pattern and the model's 12 clusters is computed, then fed into a standard multilayer perceptron network classifier. A preprocessing step is applied before classification to compensate for sources of image variation. It includes illumination gradient correction and histogram equalization. A shortcoming of this method is that there is no rule for selecting the number of Gaussian clusters.

The paper by Rowley et al. (1998) is representative for a larger class of papers considering neural networks for face detection. A retinally connected neural network is used. There are three types of hidden units aiming at detecting different facial attributes that might be important for face detection. The network has a single, real-valued output. The preprocessing step in (Sung and Poggio, 1998) is adopted. The system performs well on the CMU test set (Rowley et al., 1998).

The naive Bayes classifier is used in (Schneiderman and Kanade, 2000). Each pattern window is decomposed into overlapping subregions. The subregions are assumed statistically independent. Hence, $P_y(X)$ can be computed as

$$P_y(X) = P_y(\{R_i, P_i\}_{i=1}^{N_r}) = \prod_{i=1}^{N_r} P_y(R_i, P_i) \tag{2}$$

for $y \in \{0, 1\}$. R_i is the subregion of X at location P_i and N_r is the number of subregions. The method has the power of emphasizing distinctive parts and encoding geometrical relations of a face, and hence contains elements of a model-based approach as well. A drawback of this method is the strong independence assumption. This might not lead to high classification accuracy because of the inherent dependency among overlapping subregions.

Colmenarez and Huang (1997) use first-order Markov processes to model the face and non-face distributions:

$$P_y(X \mid S) = P_y(X_{S_1}) \prod_{i=2}^{n} P_y(X_{S_i} \mid X_{S_{i-1}}) \tag{3}$$

for $y \in \{0, 1\}$. S is some permutation of $(1, \ldots, n)$ and used as a list of indices. The learning procedure searches for an S_{m} maximizing the Kullback–Leiber divergence between the two distributions $D(P_0(X) \,\|\, P_1(X))$:

$$S_{\mathrm{m}} = \arg\max_{S} D(P_0(X|S) \,\|\, P_1(X|S)) \tag{4}$$

where $D(P_0(X) \,\|\, P_1(X))$ is defined as

$$D(P_0(X) \,\|\, P_1(X)) = \sum_{x \in R} P_0(x) \log \frac{P_0(x)}{P_1(x)}. \tag{5}$$

The Kullback–Leiber divergence is a non-negative value and equals 0 only when the two distributions are identical. The Kullback–Leiber divergence is a measure of the discriminative power between the probability distributions of the two classes (Kullback, 1959). By maximizing this measure, it is expected that a high classification accuracy can be achieved. The maximization problem, in this case, is equivalent to the traveling salesman problem (Gondran and Minoux, 1984). An heuristic algorithm is applied to find an approximate solution. An advantage of this approach is that both training and classification steps are very fast.

Osuna et al. (1997) apply support vector machines (Vapnik, 1998) to the face detection problem, which aims at maximizing the margin between classes. In order to train a large data set with vector support, a decomposition algorithm is proposed, in which a subset of the original data set is used. It is then updated iteratively to train the classifier.

One common characteristic of all methods is that they try to capture the decision boundary by the model supported by their classifiers. However, for classes with multiple sources of variation such as human faces, the decision boundary can be very complex. This might lead to poor accuracy performance for methods that can model simple decision boundaries. It might also lead to complex classifiers with a slow classification step. Hence, there is a need for a method which can model a complex decision boundary while allowing fast classification.

3. Data space exploitation and aggregated classifiers

In this section, we present a method which handles a complex decision boundary by using multiple classifiers in aggregation. Aggregated classifiers allow a natural way for solving the class/non-class classification problem.

A general learning set $\mathscr{L}$ consists of data $\{(y^t, x^t),\ t = 1, \ldots, T\}$ where the x's are the patterns and the y's are their corresponding classes. The learning set is used to form a classifier $\varphi(x \mid \mathscr{L})$, that is the class of a new pattern x is determined by $\varphi(x \mid \mathscr{L})$. Let $\{\mathscr{L}_k;\ k = 1, \ldots, K\}$ denote the K data sets to be created in order. Let φ_i denote the aggregated classifier formed by using $\{\mathscr{L}_j;\ j = 1, \ldots, i\}$ for $i = 1, \ldots, K$. The procedure for creating the data set is as follows:

1. Consider a set of face patterns $\mathscr{L}^a$. In addition, initially a set of non-face patterns $\mathscr{L}_1^{\bar{a}}$ is created by selecting randomly from a set of images containing no human faces. $\mathscr{L}^a$ and $\mathscr{L}_1^{\bar{a}}$ together form $\mathscr{L}_1$:
$$\mathscr{L}_1 = \mathscr{L}^a \cup \mathscr{L}_1^{\bar{a}}. \tag{6}$$
2. For $i = 2, \ldots, K$, apply the face detection system using the aggregated classifier φ_{i-1} on a set of images containing no human faces. False positives returned form a set of non-face patterns $\mathscr{L}_i^{\bar{a}}$. Apparently, these cases are hard cases for classifier φ_{i-1}. This set $\mathscr{L}_i^{\bar{a}}$ and the training set of face patterns $\mathscr{L}^a$ form $\mathscr{L}_i$:
$$\mathscr{L}_i = \mathscr{L}^a \cup \mathscr{L}_i^{\bar{a}}. \tag{7}$$

The number of classifiers K may be selected according to the desired classification accuracy. Because of our selection of learning sets, if any component classifier returns a non-face decision, the pattern is classified as non-face.

We argue that this technique is suited for the face detection problem. A complex decision boundary caused by the manifold of variation is modeled by using multiple classifiers. Each has different level of difficulty of separating the two classes. Each component classifier need not be very complex, which could allow a fast classification step. In addition, the fact that a non-face pattern can be rejected at any level improves the classification time because of the normally large number of non-face patterns. Significantly, since the same face patterns, $\mathscr{L}^a$, are used for training, the true positive rate does not degrade multiplicatively as the number of component classifiers increases. Also, because the non-face patterns are generated in a bootstrap fashion, it is expected that the false positive rate decreases multiplicatively. This allows a very low false positive rate.

4. Forest-structured Bayesian network classifier

In this section a new classification method for the two-class problem is described. The method is in the same spirit as the Markov process-based method in (Colmenarez and Huang, 1997). However, forest-structured Bayesian networks are used to model the joint probability distribution of each class instead of Markov processes. We use this method in an aggregated classifier because it has a fast classification step.

Bayesian network is an efficient tool to model the joint distribution of variables (Pearl, 1988). The joint distribution $P_y(X_1, \ldots, X_n)$ can be expressed using a forest structured Bayesian network as follows:

$$P_y(x) = \prod_{i=1}^{n} P_y(X_i = x_i \mid \Pi_i = \pi_i) \tag{8}$$

for $y \in \{0, 1\}$. Π_i denote the parent of X_i in the network structure. $P_y(X_i = x_i \mid \Pi_i = \pi_i)$ are estimated from the training data $\mathscr{L}_i$ (Eqs. (6) or (7)). Fig. 1 illustrates a forest structured Bayesian network modeling the joint distribution of six random variables $\{X_1, \ldots, X_6\}$.

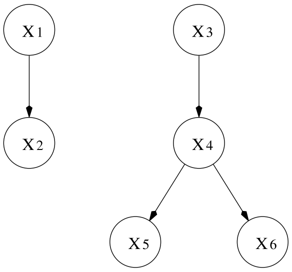

Fig. 1. A sample dependency model of six random variables with a forest structured Bayesian network: $P(X_1,\dots,X_6) = P(X_1)P(X_2\,|\,X_1)P(X_3)P(X_4\;|X_3)P(X_5\,|X_4)P(X_6\,|X_4)$.

We search for a network structure that maximizes the Kullback–Leiber divergence Eq. (5) between the two joint distributions.

The Kullback–Leiber divergence between two distributions in Eq. (8) can be obtained as

$$\begin{aligned} D(P_0(X)\,\|\,P_1(X)) &= \sum_x P_0(x)\log\prod_{i=1}^{n}\frac{P_0(x_i\,|\,\pi_i)}{P_1(x_i\,|\,\pi_i)} \\ &= \sum_{i=1}^{n}\sum_x P_0(x)\log\frac{P_0(x_i\,|\,\pi_i)}{P_1(x_i\,|\,\pi_i)} \\ &= \sum_{i=1}^{n}\sum_{x_i}\sum_{pa_i} P_0(x_i,\pi_i) \\ &\quad\times\log\frac{P_0(x_i\,|\,\pi_i)}{P_1(x_i\,|\,\pi_i)}. \end{aligned} \tag{9}$$

We show that the problem of maximizing Eq. (9) is equivalent to the maximum branching problem (Tarjan, 1977). In the maximum branching problem, a branching B of a directed graph G is a set of arcs such that:

1. if (x_1,y_1) and (x_2,y_2) are distinct arcs of B then $y_1 \neq y_2$.
2. B does not contain a cycle.

Given a real value $c(v,w)$ defined for each arc of G, a maximum branching of G is a branching such that $\sum_{(v,w)\in B} c(v,w)$ is maximum. It can be seen that maximizing $D(P_0(X)\,\|\,P_1(X))$ is equivalent to finding a maximum branching of a weighted directed graph constructed from the complete graph with node x_i's plus a node x_0 with an arc from x_0 to all other nodes. $W(i,j) = \sum_{x_i}\sum_{x_j} P_0(x_i,x_j)$ log $P_0(x_i\,|\,x_j)/P_1(x_i\,|\,x_j)$ is the weight associated with each arc in the graph. There are algorithms for solving the maximum branching problem in low order polynomial time (Tarjan, 1977).

To classify a pattern x, the Bayes decision rule Eq. (1) is used. Similar to the method in (Colmenarez and Huang, 1997), fast classification of a pattern can be achieved by constructing a table for all possible values of a variable and its parent. By using Eq. (8), the log likelihood value in Eq. (1) becomes

$$\begin{aligned} \log\frac{P_0(x)}{P_1(x)} &= \log\frac{\prod_{i=1}^{n} P_0(x_i\,|\,\pi_i)}{\prod_{i=1}^{n} P_1(x_i\,|\,\pi_i)} \\ &= \sum_{i=1}^{n}\log\frac{P_0(x_i\,|\,\pi_i)}{P_1(x_i\,|\,\pi_i)}. \end{aligned} \tag{10}$$

Once all possible values of $\log P_0(X_i\,|\,\Pi_i)/P_1(X_i\,|\,\Pi_i)$ for all i are computed, the classification of a new pattern can be carried out with only n additions. This allows a very fast classification step.

5. Face detection system

The architecture of the system is adopted from (Rowley et al., 1998). A window of size 20 × 20 is scanned over each image location to find face patterns. The size 20 × 20 is selected because it is large enough to capture details of human faces, while allowing a reasonable classification time. The system searches the input image at multiple scales by iteratively scaling down the image with a scale step of 20% until the image size is less than the window size.

Sources of variation are captured in the training set: illumination and shadows, facial expressions, glasses, make-up, hairs, beards, races and sexes. Limited orientation and head pose, namely frontal faces and near-frontal faces, are present.

We adopt two preprocessing operations from Sung and Poggio (1998): illumination gradient correction and histogram equalization. The former reduces the effect of heavy shadows and the latter normalizes the illumination contrast of the image. Finally, each pattern is quantized to six levels of gray values to enable the estimation of the discrete probabilities. Fig. 4 shows the quantized patterns from Fig. 3.

An aggregated classifier consisting of three Bayesian network classifiers, i.e. φ_3, is used to classify faces and non-face patterns. The number 3 was selected based on the tradeoff between the false positive rate and true positive rate (see Fig. 6). For $K > 3$, the true positive rate is low for the detection task.

A postprocessing step is carried out to eliminate overlapping detections. When overlapping occurs, a straightforward approach would be to select the window having the largest log likelihood value. This generates sparse maxima, of which most are false positives as is observed in (Rowley et al., 1998), that is most faces detected are detected at multiple positions nearby in place or in scale. We have repeated the experiment and arrived at the same conclusion. For each detected location, if the number of detections within a predefined neighborhood is less than a threshold, the location is rejected.

5.1. Data for training

For the purpose of this paper, a set of 1112 face examples was gathered from the Internet without selection. Color images were converted to grayscale images. Fig. 2 gives 30 randomly selected face examples. The dataset is split into two subsets at random: 1000 faces examples are used to create the training set and 112 used to create the test set. Thirty face patterns of size 20×20 are extracted from each original face examples by rotating the images about their center points by one random less than 10°, scaling by one random value selected from the interval 0.9 and 1.1, translating by one random value less than 0.5 pixel, and mirroring as in (Rowley et al., 1998). Fig. 3 illustrates 30 face patterns generated from one face example. In total, 33 360 face patterns were created (Fig. 4 shows the quantized patterns from Fig. 3).

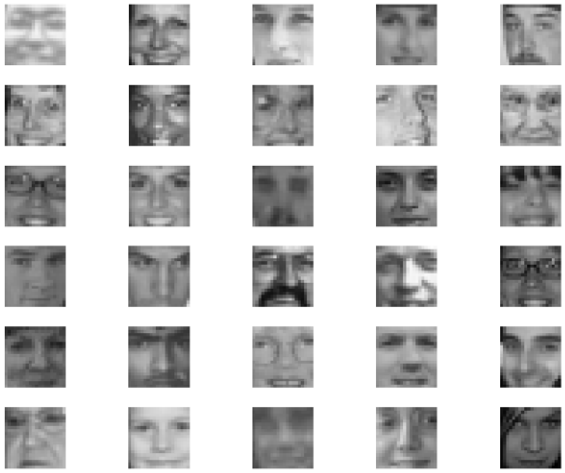

Fig. 2. Thirty of 1112 randomly selected face examples.

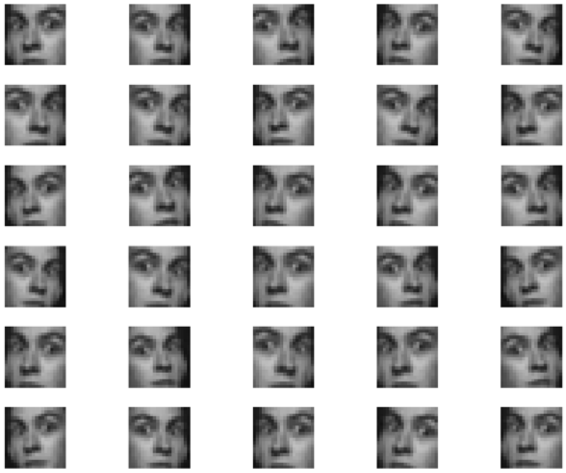

Fig. 3. An example of all 30 face patterns generated from each face example, yielding 30 000 patterns to train the system.

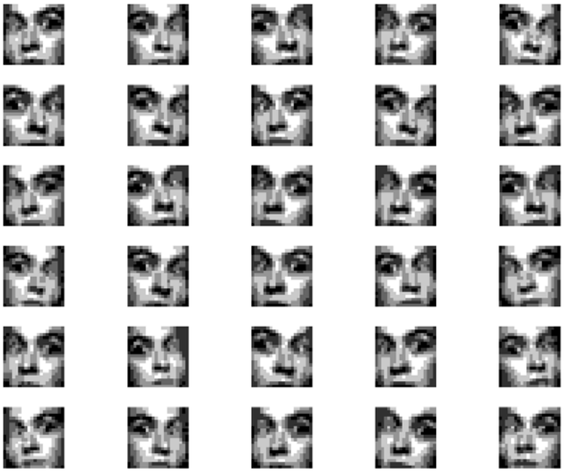

Fig. 4. Quantization to six levels of gray values of the patterns shown in Fig. 3. Note the preservation of geometrical layout after gray value normalization.

A set of 929 images containing no faces was also collected from the Internet. 360 000 non-face patterns are extracted from the images by randomly selecting a square from an image and subsampling it to patterns of size 20 × 20. Fig. 5 contains 30 non-face patterns. From the next level downwards, non-face patterns were generated as described in Section 3.

The dataset of 33 360 face and 360 000 non-face patterns is split into two subsets at random: the training set consists of 30 000 face and 160 000 non-face patterns, and the test set consists of 3360 face and 200 000 non-face patterns. This test set is referred to as the pattern test set, $\mathscr{L}^{\mathrm{T}}$. The face patterns of the two subsets were generated from two separate sets of face examples.

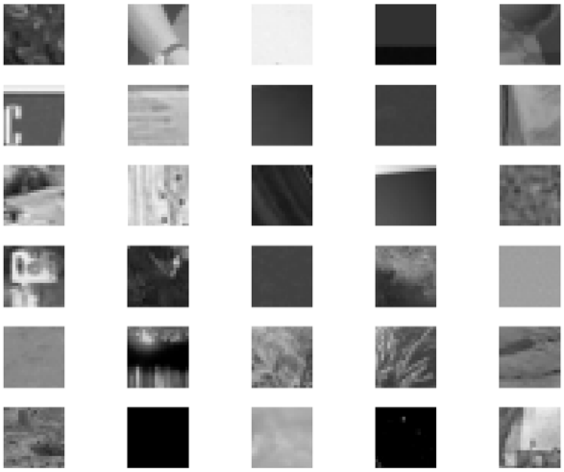

Fig. 5. Thirty non-face patterns randomly selected from the set of 160 000 non-face patterns.

6. Experimental results

6.1. Experiment with the number of component classifiers K

Fig. 6 shows the receiver operating characteristic curves for the four aggregated classifiers φ_1, φ_2, φ_3 and φ_4 on the pattern test set $\mathscr{L}^{\mathrm{T}}$. At a low false positive rate an aggregated classifier with higher value of K achieves higher true positive rate. However, saturation occurs with $K > 1$, i.e. it is not possible to achieve higher true positive rate even at high false positive rate. This is due to the fact that the true positive rate of the previous levels puts an upper bound on the achievable rate of the next level.

6.2. Experiment with the Bayesian network classifier

Fig. 7 shows the receiver operating characteristic (ROC) curves of the three different classifiers

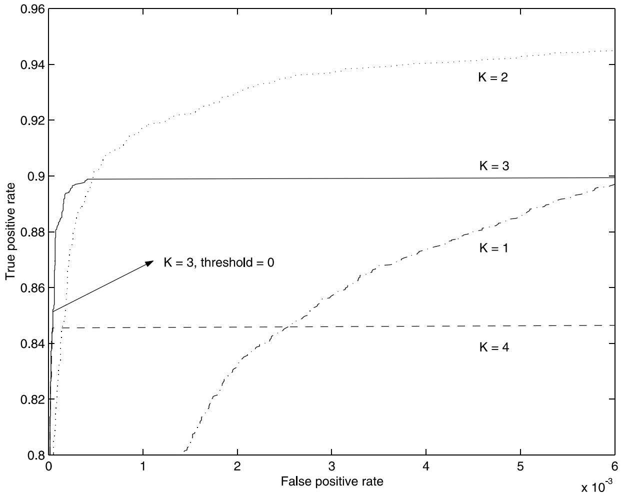

Fig. 6. The receiver operating characteristic (ROC) curves for φ_1, φ_2, φ_3 and φ_4 on the pattern test set $\mathscr{L}^{\mathrm{T}}$.

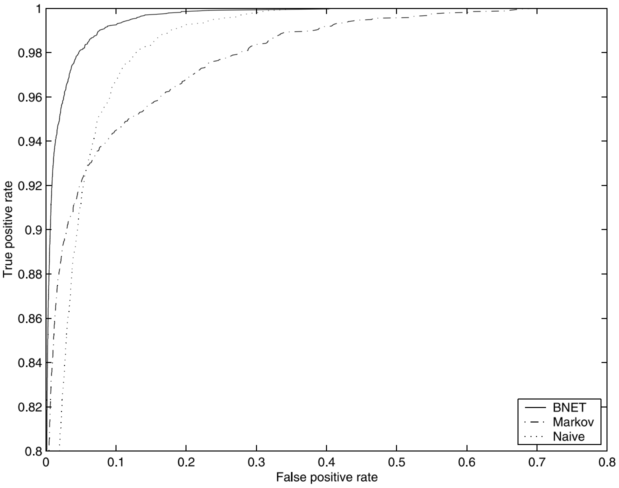

Fig. 7. The receiver operating characteristic (ROC) curves of three classifiers: the Markov process classifier (Colmenarez and Huang, 1997), the naive Bayes classifier (Duda and Hart, 1973) and the Bayesian network classifier φ_1.

on the pattern test set $\mathscr{L}^T$: the Markov process classifier (Colmenarez and Huang, 1997), the naive Bayes classifier (Duda and Hart, 1973) and our method, the Bayesian network classifier φ_1. Our method outperforms both the Markov process classifier and the naive Bayes classifier.

As an aside, it is interesting to see that the Markov process classifier performs better than the naive Bayes classifier only when the false positive rate is smaller than 6%.

6.3. Experiment on the full image test set

The system is evaluated using the CMU test set (Rowley et al., 1998). This test set consists of 130 images with a total of 507 frontal faces, including images of the MIT test set (Sung and Poggio, 1998). The images were collected from the World Wide Web, scanned from photographs and newspaper pictures, and digitized from broadcast television. There is a wide range of variation in image quality. It should be noted that some authors report their results on a test set excluding five images of line draw faces (Schneiderman and Kanade, 2000), which leaves this test set with 125 images with 483 labeled faces only. We use the groundtruth with 507 faces as in (Rowley et al., 1998).

Table 1 shows the performance of our face detection system in comparison with systems in (Rowley et al., 1998) on the CMU test sets. It can be seen that with an equivalent detection rate, Bayesian network based method gives about half the number of false detections in comparison with the neural network method (Rowley et al., 1998). Fig. 8 illustrates the detection result on some images of the CMU test set.

Table 1
Evaluation of the performance of the aggregated Bayesian networks, φ_3, as compared to the neural network, NN (Rowley et al., 1998) on the CMU test set (Rowley et al., 1998)

	MFs	Rate	FDs
Our system			
φ_3	47	90.7%	264
System in (Rowley et al., 1998)			
NN, System 5	48	90.5%	570
NN, System 6	42	91.7%	506
NN, System 7	49	90.3%	440
NN, System 8	42	91.7%	484

The criteria are: the number of missed faces (MFs), the true detection rate (rate) and the number of false detects (FDs).

7. Discussion and conclusion

In this paper we have considered the face detection task as a representative of the class/non-class classification problem where the class is subjected to many sources of variation. The sources of variation include position of the face relative to the camera, illumination condition, non-rigid characteristic of the face, and presence of other devices. The appearance variation is also caused by differences among races, and between male and female. In addition, the classification

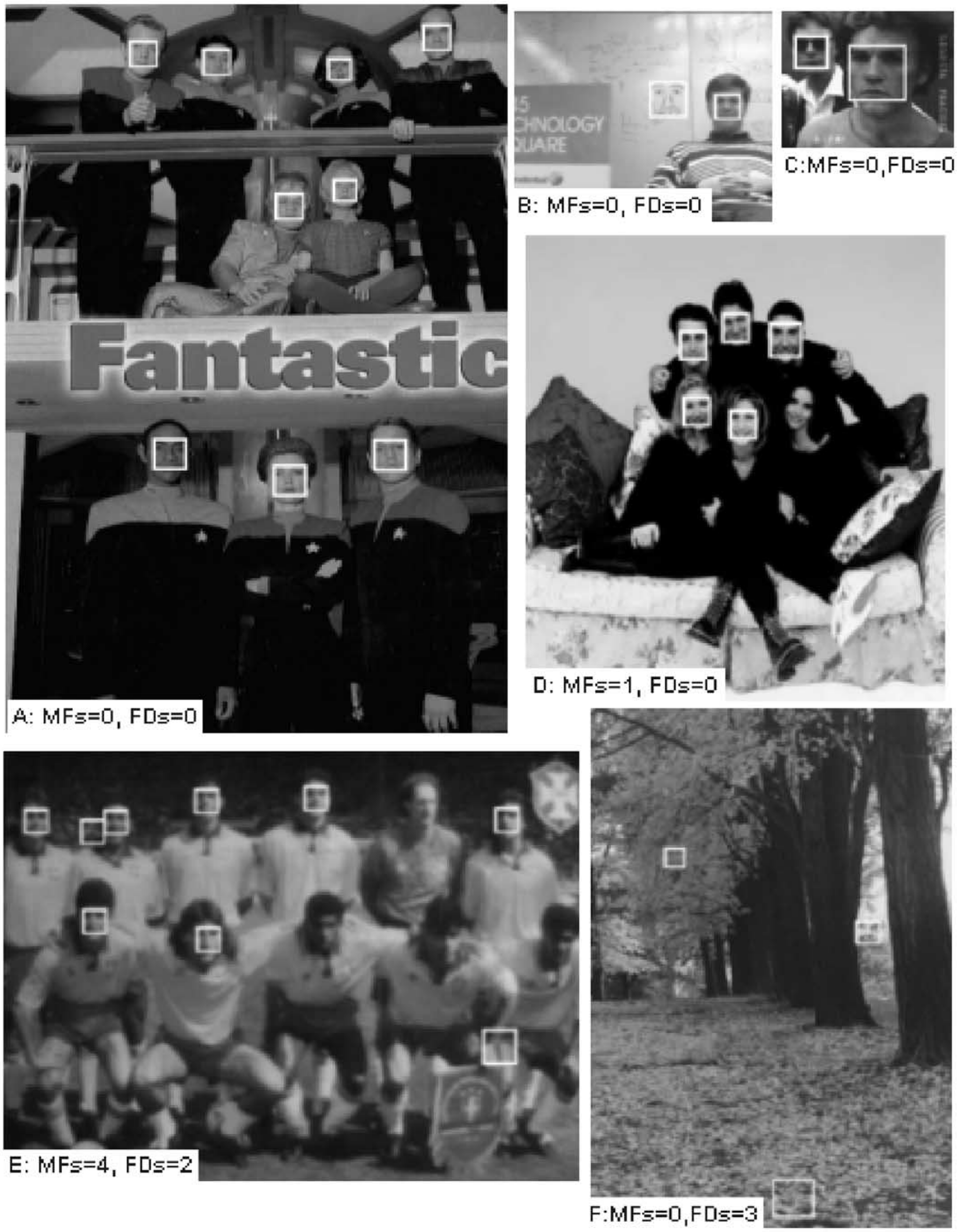

Fig. 8. Output of the system on some images of the CMU test set (Rowley et al., 1998). MFs is the number of missed faces and FDs is the number of false detections.

method must have a very low false positive rate and a fast classification step.

Our face detection system performs well. On the CMU test set it achieves detection rate of about 90% with an acceptable number of false alarms. In comparison with other methods, our classification method using Bayesian networks outperforms related methods (namely the Markov process method (Colmenarez and Huang, 1997) and the naive Bayes classifier (Duda and Hart, 1973), as shown in Fig. 7). On the CMU test set, our system performs better than the neural network method (Rowley et al., 1998). Our system gives about half the number of false alarms at an equivalent detection rate (see Table 1).

Approximately half of the missed detections are caused by rotated angles (see Fig. 8, image D). Large in-plane rotation or out-of-plane rotation are not handled with this method. When the subject has the intention of looking into the camera, false negatives are rare. In fact, the missed detection in image D is one of the very few cases. Poor image conditions, such as low brightness and strong shadows, account for about one third of the missed detections (see the three examples in image E). In order to resolve this a special image enhancement preprocessing step might help. The remaining missed detections are caused by various reasons including the sizes of the faces being too small. Among the false positives, in 30 cases out of 264, the patches do appear as human faces (see the false alarm in image E and the top two false alarms in image F). Other cases might be eliminated by further postprocessing. Given the large number of tested windows (Rowley et al., 1998), our method makes only one incorrect classification out of each 300 000 tests.

Table 2 summarizes the parameters used by our system. The number of patterns created from one example (parameter 1) becomes saturated, that is the performance of the system does not improve by increasing this value. All parameters from 2 to 6 are identical to (Rowley et al., 1998) to permit a fair comparison.

As concerns parameter 7, because our method uses a memory-based histogram for probability density estimation, there is a limitation on the number of discrete levels to be used. During the training process, at six discrete levels, each histogram takes up 44 MB of memory. At eight discrete levels, each histogram would take up about 78 MB. Discretization causes loss of information, but does not necessarily reduce the classification accuracy. With higher number of discrete levels, more training data are needed to characterize the distributions. Furthermore, we still can distinguish face patterns from non-face patterns at six discrete gray values. An experiment with four discrete levels (data not shown) indicates a slightly degraded performance. For the purpose of this paper, 6-level discretization is appropriate. A higher number of levels might improve the performance of the system.

As concerns parameter 8, setting the value of K to 4 does not help because the true positive rate is low, see Fig. 6. Regarding the last parameter, the thresholds of all classifiers are set to 0. Nevertheless, the threshold of the last classifier may vary, reflecting the tradeoff between true positive and false positive rate of the system.

Table 2
System parameters and their values

Phase		Parameter	Value
Learning	1	Number of patterns per one example	30
	2	Random rotating range (degree)	Uniform over $[-10, +10]$
	3	Random scaling range	Uniform over $[0.9, 1.1]$
	4	Random translating range (pixel)	Uniform over $[-0.5, 0.5]$
Learning and runtime	5	Window size (pixel)	20×20
	6	Scale step	20%
	7	Number of discrete levels	6
Runtime	8	K, number of classifiers	3
	9	Threshold per classifier	0

Our system makes use of the symmetry property of the human face only implicitly by the mirroring operation on the training face examples. It is interesting to investigate how symmetry can be encoded in the Bayesian network prior to the learning phase. It is important to note, however, that structural biases and lighting may affect the symmetry property.

In conclusion, this paper presents a face detection system using an aggregation of Bayesian network classifiers. The use of an aggregated classifier is well suited for the class/non-class classification problem in the visual domain, where a complex decision boundary is anticipated due to many sources of variation. In addition, aggregated classifiers allow a very low false positive rate and fast detection.

Acknowledgements

The authors would like to thank the reviewers for their helpful comments.

References

Colmenarez, A., Huang, T., 1997. Face detection with information-based maximum discrimination. In: Proc. CVPR'97, pp. 782–787.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Gondran, M., Minoux, M., 1984. Graphs and Algorithms. Wiley, New York.

Kullback, S., 1959. Information Theory and Statistics. Wiley, New York.

Landeweerd, G.H., Timmers, T., Gelsema, E.S., Bins, M., Halie, M.R., 1983. Classification of normal and abnormal samples of peripheral blood by linear mapping of the feature space. Pattern Recognition 16, 319–326.

Leung, T., Burl, M., Perona, P., 1995. Finding faces in cluttered scenes using random labeled graph matching. In: Proc. ICCV'95, pp. 637–644.

Moghaddam, B., Pentland, A., 1997. Probabilistic visual learning for object representation. IEEE PAMI 19 (7), 696–710.

Osuna, E., Freund, R., Girosi, F., 1997. Training support vector machines: an application to face detection. In: Proc. CVPR'97, Puerto Rico, pp. 130–136.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA.

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

Schneiderman, H., Kanade, K., 2000. A statistical method for 3D object detection applied to faces and cars. In: Proc. CVPR 2000, pp. 746–751.

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

Tarjan, R., 1977. Finding optimum branchings. Networks 7, 25–35.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York.

Wei, G., Sethi, I.K., 1999. Face detection for image annotation. In: Gelsema, E.S., Veenland, J.F. (Eds.), Pattern Recognition in Practice VI, Vlieland, June 99. Pattern Recognition Letters 20 (11–13), 1313–1321.

Yang, J., Waibel, A., 1996. A real-time face tracker. In: Proc. WACV'96, pp. 142–147.

Yow, K., Cipolla, R., 1996. Scale and orientation invariance in human face detection. In: Proc. 7th British Machine Vision Conference, pp. 745–754.