IDIAP RESEARCH REPORT

Multiview Face Detection

Tiffany Sauquet, Sébastien Marcel and Yann Rodriguez

IDIAP–RR 05-49

September 2005

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-49

Multiview Face Detection

Tiffany Sauquet, Sébastien Marcel and Yann Rodriguez

September 2005

Abstract. In this document, we address the problem of multiview face detection. This work extends the frontal face detection system developed at the IDIAP Research Institute to multiview face detection. The main state-of-the art techniques are reviewed and a novel architecture is presented, based on a pyramid of detectors that are trained for different views of faces. The proposed approach robustly detects faces rotated up to $\pm 67.5^\circ$ in the image plane and up to $\pm 90^\circ$ out of the image plane. The system is real-time and achieves high performances on benchmark test sets, comparable to some state-of-the art approaches.

Contents

List of Figures

List of Tables

1 Introduction

Since these last ten years, face detection has become an important research area in computer vision, due to its wide range of possible applications such as human-computer interaction, video surveillance systems, biometrics, content-based video indexing or image retrieval for multimedia indexing and retrieval. The goal of face detection is to determine whether or not there are any faces in the image and, if present, their location. It is the crucial first step of any application that involves face processing systems including face recognition, face tracking, pose estimation or expression recognition. Thus, accurate and fast human face detection is the key to a successful operation.

Face detection is a challenging problem because faces highly vary in size, shape, color, texture and location. Their overall appearance can also be influenced by lighting conditions, facial expression, occlusion or facial features, such as beards, mustaches and glasses. Another challenging problem comes from the orientation (upright, rotated) and the pose (frontal to profile) of the face. Many real applications need the ability to deal with faces of varying head poses, called multiview faces, since most faces in the real world are not frontal. The background of an image (everything that is not a face) is also an important factor contributing to the difficulty of face detection. As most applications must be independent to background conditions, face detection systems should be able to detect faces in a complex background.

Researchers have essentially focused on upright frontal face detection, mainly because detecting profile views is a more difficult task. Indeed, profile views have fewer stable features and contain more background pixels. Detecting faces across multiple views is however becoming a topic of growing interest, and some researchers are starting to adapt their frontal detection system to the multiview framework, while others are proposing new approaches.

Recently, the IDIAP Research Institute has developed a face detection system which detects, in real-time, multiple upright frontal faces in complex backgrounds. The objective of this project is thus to extend the system to multiview face detection.

In the next Section, the main state of the art approaches, both in frontal and non-frontal detection, are reviewed shortly. Section 3 presents the baseline of the frontal face detection system developed at the IDIAP Research Institute. Section 4 introduces the proposed multiview face detection system. Experiments and results are described in Section 5. A conclusion and some directions for future research are presented in Section 6.

2 State of the Art

Numerous techniques have already been proposed, many of which are reviewed by Yang et al. [1], and by Hjelmas and Low [2] in their survey. These methods can be classified in two categories : feature-based approaches and image-based approaches. Feature-based approaches make explicit use of face knowledge. Some of them are based on local facial feature detection and classification such as edges [3], gray information [4], color [5] or motion [6] (if video is available). On the contrary, most of image-based methods rely on statistical analysis and machine learning to find the relevant characteristics of face and nonface images. Face detection is thus treated as a pattern recognition problem where a training procedure classifies examples into face and non-face prototype classes, so that face patterns can then be recognized.

Image-based approaches have proven to be more robust than feature-based approaches. Consequently, I will only consider, in this section, the most popular and successful image-based methods reported in the literature, which differ on the pattern classification techniques they apply. Among these methods, I will distinguish frontal face detection systems from non-frontal face detection systems, which deal with multiview faces.

2.1 Frontal Face Detection

2.1.1 Eigenface approach

Some methods can be understood in a probabilistic framework. Their objective is to estimate the density functions $p(x|face)$ and $p(x|nonface)$, where x is a random variable representing an image or a feature vector derived from an image. The dimension of the image space is usually very high, so it has to be reduced using, for example, principal component analysis (PCA). One possible approach is to use an eigenspace decomposition as in [7]. Performing PCA on a training set of face images generates the so-called Eigenfaces which span a subspace of the image space that best represent the set of face patterns, the face space. The distance between the input pattern and the face space can then be computed. Bayesian classification or maximum likelihood is usually used to classify a candidate image as face or nonface.

These methods are robust to noise, occlusion, facial expression and illumination. However, they are sensitive to variation in position and scale. The background is also a serious issue, because eigenface analysis does not distinguish the face from the background.

2.1.2 Sung and Poggio approach

Sung and Poggio have developed a distribution-based system considered as the first advanced image-based face detection system [8]. Their system consists of two components, distribution-based models for face and nonface patterns, and a multilayer perceptron classifier (Fig. 1). First, each face and nonface example is normalized and preprocessed to a 19×19 pixel image via lighting correction (a best fit linear function is subtracted from the original signal), followed by histogram equalization to enhance the contrast. The training patterns are then classified into face and nonface clusters using a modified k-means algorithm. Each cluster is represented as a multidimensional Gaussian function. Two distance metrics are computed between an input image pattern and the prototype clusters (a Mahalanobis-like distance and the Euclidean distance). Finally, a multi-layer perceptron is used to classify face window patterns from nonface patterns using the distances to each face and nonface cluster.

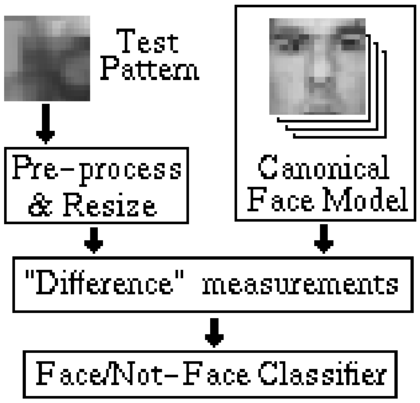


Figure 1: Classification in Sung and Poggio's face detection system.

2.1.3 Rowley approach

Rowley et al. [9] presented the first advanced neural network-based frontal face detection system. Their system operates in two stages. First a neural network-based filter examines each location in the image at several scales and looks for face-like locations. Then, the detections from individual filters are merged to eliminate the overlapping detections.

The input of the filter is a 20×20 pixel region of the image. To detect faces at any location and any scale in the image (and not only 20×20 faces), the filter is applied at every pixel position in the image and at several scales (with a subsampling factor of 1.2). The filtering algorithm, shown in Fig. 2, starts with a preprocessing step. The input window is preprocessed using the same algorithms as in the system of Sung and Poggio. The preprocessed window is then passed through a neural network. There are three types of hidden units, chosen to be able to detect local features such as mouths, pairs of eyes, a single eye or the nose. The weights are learned via standard back-propagation. The network has a single, real-valued output, ranging from -1 (nonface pattern) to 1 (face pattern).

To train the neural network, a large number of faces and non faces are needed. Rowley et al. use 1050 face examples, collected from several face databases (CMU,

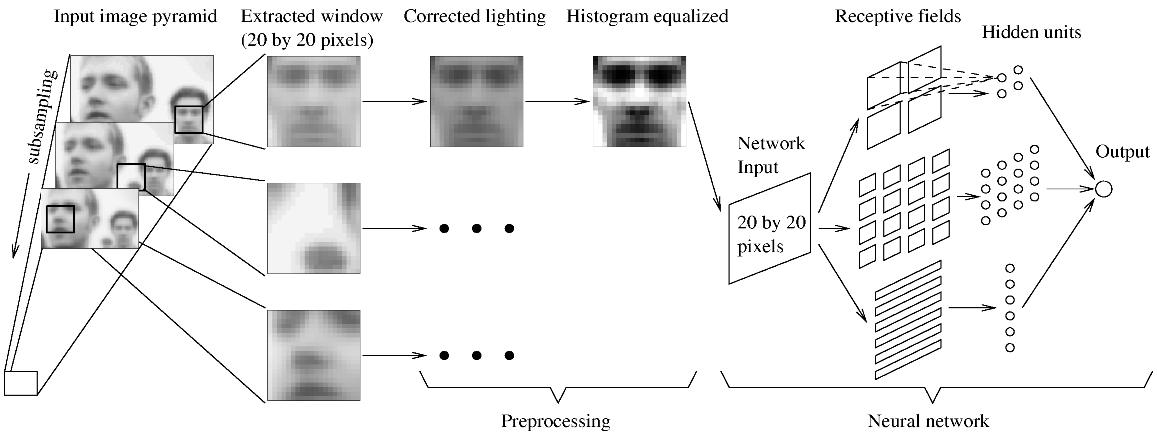


Figure 2: Algorithm used by Rowley et al. for their face detection system. In the neural network, a first hidden layer is composed of 4 units connected to 10×10 pixel subregions, a second of 16 units connected to 5×5 subregions and a third of 6 units connected to 20×5 subregions.

Harvard) and from Internet. The eyes, the tip of the nose, the corners and the center of the mouth of each face are manually labelled for geometric normalization. Each face example is then mapped to a 20×20 pixel window. Fifteen face examples are generated from each original image of the training set by randomly rotating, scaling, translating and mirroring the images to make the detector more robust to position and scale.

Since collecting a representative set of images not containing faces is a difficult task due to the huge size of the nonface space, they choose to use a "bootstrapping" method. This method consists in iteratively retraining the system with an updated training set containing false alarms produced after face detection has been performed on a set of scenery images that do not contain faces. Rowley et al. use 120 images of scenery for collecting nonfaces examples where approximately 8000 nonface images are selected from the 140 millions available.

As the network has some invariance to position and scale, multiple detections can occur around some faces and nonfaces. Thus, overlapping detections have to be merged into single detections. The number of detections within a specified neighborhood of that location are counted. If this number is above a threshold, the location is classified as a face and a single detection occurs.

To further reduce the number of false positives, Rowley et al. train multiple neural networks. Each network is trained in a similar manner, but with random initial weights and random initial nonface images. The detection and false alarm rates of the individual networks are close, but due to the different training conditions, the networks have different biases and make different errors. Their output are combined with an arbitration strategy (ANDing, ORing, voting, or a separate arbitration neural network). This algorithm however increase the computational cost.

2.1.4 Féraud approach

Féraud et al. [10] proposed another neural network model, based on the Constrained Generative Model (CGM). CGMs are auto-associative connected MLPs with three large layers of weights, trained to perform a non-linear PCA. Classification is obtained by considering the reconstruction errors of the CGMs.

The detector is composed of four filters. The first two stages are a motion filter (in the case of video sequences) and a color filter (in the case of color images). These first filters are fast and remove most of the non-face regions to reduce the global computation time. The third level is a multi-layer perceptron that filters more than 90% of the remaining subwindows. The last stage is based on a combination of CGMs (Fig. 3). The best results are reported using a conditional mixture of CGMs and a MLP gate network.

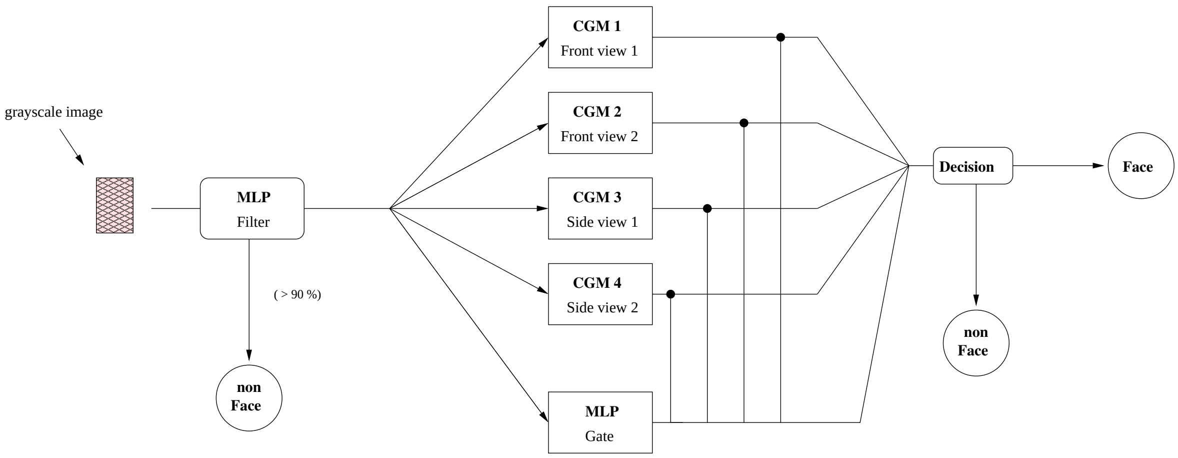


Figure 3: Structure of Féraud's face detector (without motion and color filters).

2.1.5 SNoW approach

Besides these neural network-based approaches, Roth et al. reported in [11] one of the most accurate face detector based on SNoW learning architecture (Sparse Network of Winnows). SNoW is a single layer neural network of linear functions that uses a variant of the Littlestone's Winnow update rule[1]. The number of required patterns for training grows linearly with the number of relevant features and only logarithmically with the total number of features. This is particularly interesting in the face detection domain, where the number of features may be very large, but only a relatively small number of them is relevant.

2.1.6 Viola and Jones approach

Viola and Jones introduced in [12] the first real-time frontal face detection system. It is based on AdaBoost algorithm. It achieves high detection rates with an acceptable number of false alarms. They reported that it is able to process an image of 384×288 pixels in approximately 0.07 seconds on a 700 MHz Pentium III. Viola and Jones proposed three main contributions. First, they use an image representation, called "Integral image", widely used in computer graphics, to compute very rapidly the features used by the detector. Then, they build a classifier using a variant of AdaBoost learning algorithm to select a small number of discriminant features from a very large set of possible features. Finally, the classifiers are combined in a "cascade" to discard rapidly background regions.

Instead of directly using the pixel information, the face detection procedure classifies images based on the value of simple features that are reminiscent of Haar basis functions[2]. Five kinds of features are used, as shown in Fig. 4.

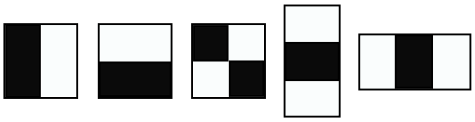


Figure 4: Five types of Haar-like features used in the system of Viola and Jones. The value of a feature is defined as the sum of the pixels within the white part(s) minus the sum of the pixels within the black part(s).

These features correspond to the difference between the sum of the pixels within adjacent rectangular regions of the image. They can be computed very rapidly at any scale and position in constant time using the integral image. Indeed, the integral image at location x, y contains the sum of the pixels above and to the left of x, y, inclusive :

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y')$$

where $ii(x, y)$ is the integral image and i(x,y) is the original image. Any rectangular sum can thus be computed in four array references (see Fig. 5).

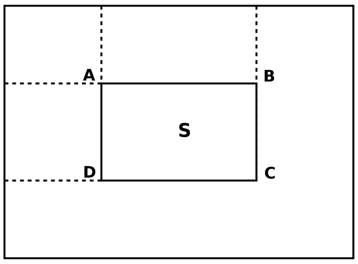


Figure 5: The sum of the pixels S within the rectangle ABCD can be computed with four array references: $S = ii(x_C, y_C) + ii(x_A, y_A) - (ii(x_B, y_B) + ii(x_D, y_D))$.

Like most face detection systems, the face detector scans the input image at many scales. The conventional approach is to compute a pyramid of subsampled images like Rowley et al. [9]. A fixed scale detector is then scanned across each of these images. The computation of the pyramid, while straightforward, requires however significant time. In contrast, each of the Haar-like features can be evaluated at any scale and location in few operations. A multi-scale detector is thus scanned across the original image, reducing the computational cost.

Given a training set of positive and negative images (face and nonface), all possible features are computed. The set of rectangle features thus obtained provides a rich image representation which supports effective learning. Viola and Jones formulate the hypothesis that a very small number of features can be combined to form an effective classifier. They use a variant of AdaBoost both to select the features and to train the classifier. The AdaBoost learning algorithm was mainly developed by Freund and Schapire [13]. It is a general method for improving the classification performance of any given learning algorithm : it boosts the classification performance of a simple learning algorithm, called a weak learner, by combining a collection of weak classification functions to form a stronger classifier. AdaBoost is an effective procedure for searching out a small number of features which have significant variety. For each feature, the algorithm computes the number of examples in respect of the feature value. The first feature chosen will thus be the most discriminant one and so on. The weak learner determines the optimal threshold classification function, such that the minimum number of examples are misclassified. A weak classifier $h(x, f, p, \theta)$ consists of a feature f, a threshold θ and a polarity p indicating the direction of the inequality :

$$h(x, f, p, \theta) = \begin{cases} 1 & \text{if } p \cdot f(x) < p \cdot \theta \\ 0 & \text{otherwise} \end{cases}$$

where x is a 24×24 pixel image.

Strong classifiers are then organized into a cascade structure of increasing complexity, illustrated in Fig. 6. Each stage is composed with a certain number of weak classifiers that depend on only a single feature :

$$H(x) = \sum_{i=0}^{n} w_i h_i(x) \in [-1; 1]$$

where w_i is the weight of the classifier h_i.

The threshold of the stage is determined to detect close to 100% of the faces. The objective of this cascade is to eliminate as many negative examples as possible at the earliest stage possible : simple classifiers are used to reject the majority of the subwindows while detecting almost all of the positive instances. More complex classifiers can then focus on a reduced number of subwindows. Thus, the construction of a cascade of classifiers reduces the computation time.

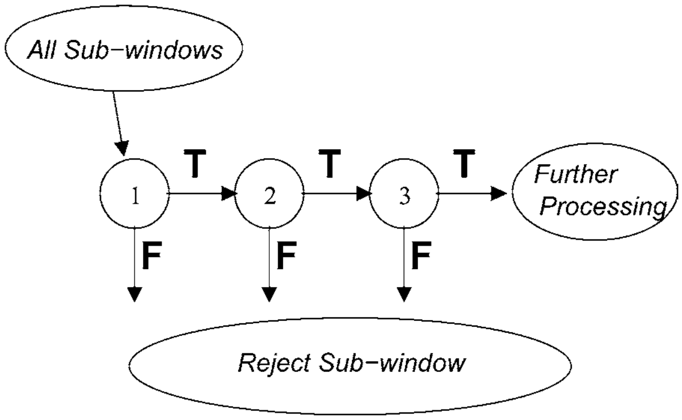


Figure 6: Schematic description of the detection cascade.

However, the overall training process is very difficult to optimize. Indeed, classifiers with more features will achieve higher detection rates with less false alarms, increasing at the same time the computational cost. Thus, finding a trade off between the number of stages in the classifier, the number of features of each stage and the threshold of each stage is a very difficult task. Nothing concrete has been proposed yet to solve this problem.

As an extension of Viola and Jones work, Lienhart et al. proposed in [14] an extended set of haar-like features, including 45° rotated features and center-surround features (Fig. 7), adding additional knowledge of the domain to the learning framework. These features can be computed rapidly using a scheme similar to the integral image in [12]. At a given detection rate, the authors report a 10% decrease of the false alarm rate with this extended feature set.

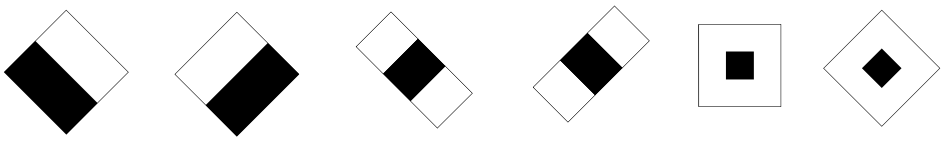


Figure 7: New Haar-like features introduced by Lienhart et al.

In addition to this extended feature set, they also compare different boosting algorithms : Discrete (used by Viola and Jones), Real and Gentle AdaBoost variants. Their experiments show a slight improvement with the Gentle AdaBoost training.

2.1.7 Li approach

Li et al. introduced in [15] a new feature set to construct the weak classifiers and a new learning algorithm, called FloatBoost, to select the features used to build a unique strong classifier. Li et al. use also Haar-like features, but different than those used by Viola and Jones, as shown in Fig. 8. The value of each feature is given by the weighted sum of the pixels within the rectangles. As previously, these features can be computed rapidly using the integral image, like in [12].

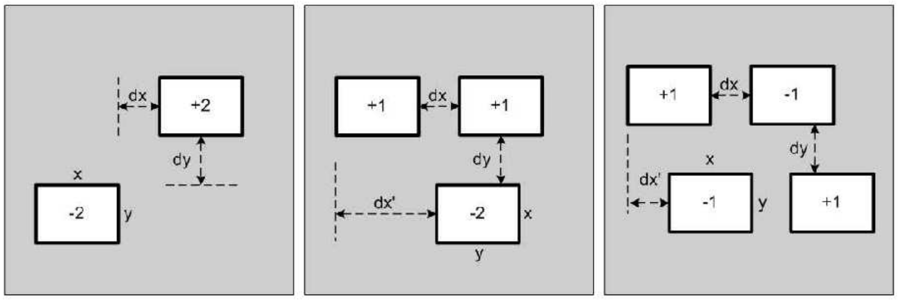


Figure 8: Three types of Haar-like features used in the system of Li et al. The size of the rectangles is $x \times y$ and they are at a distance (dx, dy) from each other. The values ($\pm$1,2) correspond to the weights assigned to the pixels in the rectangles.

FloatBoost backtracks unfavorable weak classifiers to remove them from the existing classifiers to achieve a low error rate. As in AdaBoost, the next best classifier is added one at a time to the already selected weak classifiers. But FloatBoost removes the least significant weak classifier if this removal leads to a lower error rate. This step is repeated until no more removals can be done. For the same error rate, FloatBoost selects thus fewer weak classifiers than AdaBoost. However, this improved feature selection method increases the training time (about 5 times longer than AdaBoost).

2.2 Non-Frontal Face Detection

Non-frontal face detection, or multiview face detection, deals with all possible in-plane and out-of-plane rotations of the face, as shown in Fig. 9. Most researchers have at first distinguished in-plane and out-of-plane rotations. Only recently, some researchers have started to propose fully integrated multiview face detection systems.

Rowley et al. extended in [16] their upright, frontal neural network-based face detection system to a fully in-plane rotation invariant system. They proposed to use another network, called a “router”, to determine the orientation of the input window before processing it by the face detector. Their overall algorithm is given in Fig. 10. After preprocessing (histogram equalization), the input window is given to a router network that returns the in-plane rotation angle (36 output units covering the entire 360° range). This angle is then used to rotate the original input window to an upright position. Finally, the “derotated” window is preprocessed in the same way and passed to the neural-network based face detector described in the previous section, which decides whether or not the window contains a face. The router network does not

(a)

(b)

Figure 9: Example of (a) in-plane rotations and (b) out-of-plane rotations.

require a face as an input. If a nonface is given, it will return a meaningless rotation and the window will be rejected by the face detector.

Rowley et al. reported good results, but the system is very slow (several seconds), which is the main drawback of this approach. The router network only estimates the pose of the input window, without eliminating any window. A possible improvement could be to add another output to the router network that would reject nonfaces. Less subwindows would thus be processed by the face detector, improving the computational cost.

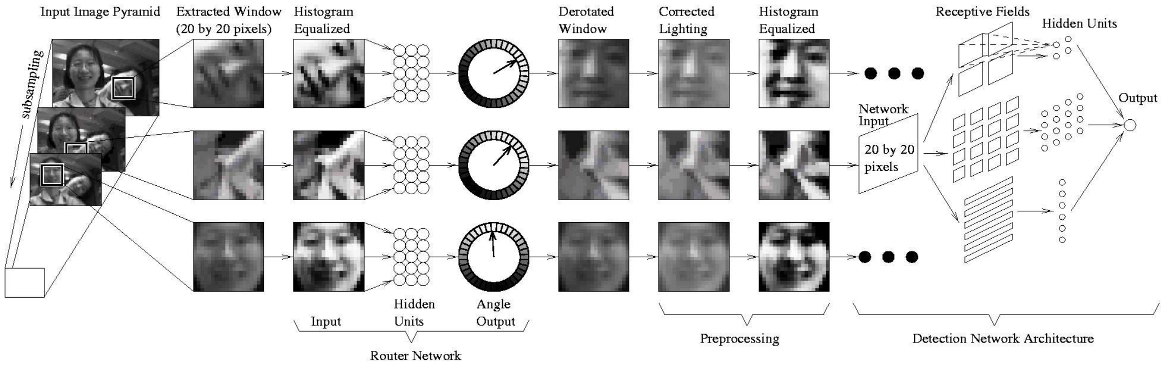


Figure 10: Algorithm used by Rowley et al. for their face detection system.

Viola and Jones extended their face detection system in [17] to handle profile views and in-plane rotated faces. Two systems with the same framework are proposed to detect either non-frontal faces (rotated out of the image plane) or non-upright faces (rotated in the image place).
Like Rowley et al., they proposed a two stage approach. They first estimate the pose of the window and then, instead of derotating the window, pass it directly to the rotation

specific detector that classifies it. The space of poses is divided into various classes and the different detectors are trained for each pose class.

The pose estimator is a decision tree trained to detect 12 different poses in the in-plane case. The 12 corresponding frontal face detectors are trained in 12 different rotation classes, covering the full 360 degrees of possible rotations, since each rotation class covers 30 degrees of in-plane rotation (the Viola and Jones upright frontal face detector handles approximately ± 15 degrees of in-plane rotation). Each detector is constructed using the same framework as in [12], that is a cascade of increasingly more complex classifiers.

In the out-of-plane case, the pose estimator is trained to detect 2 poses, left and right. The left and right profile detectors handle rotations from about 3/4 view to full profile, and they are constructed using the same framework as in the in-plane case.

Their method preserves the speed advantage of the original frontal face detector, while achieving high detection rates, both for in-plane rotated faces and for profile faces. Their detector processes a 320×240 pixel image in about 0.12 seconds on a 2.8 GHz Pentium IV.

Li and Zhang reported in [18] the first real-time multiview face detection system. Their system is based on the FloatBoost algorithm for learning face and nonface classifiers and they use Haar-like features to build the weak classifiers (see Section 2.1.7). Li and Zhang introduced a novel coarse-to-fine, simple-to-complex architecture, called detector-pyramid.

The multiview face detection system deals with three types of head rotations, which are out-of-plane rotations in the range of [-90°,+90°], in-plane rotations in the range of [-45°,+45°] and up-and down nodding rotations approximately in the range of [-20°,+20°]. Out-of-plane rotations are handled by the detector-pyramid. In-plane rotations are dealt with by applying the detector-pyramid on two images in-plane-rotated by ± 30 degrees as well as on the original image. The face detectors are trained to be tolerant to the up-and-down nodding rotations.

The detector-pyramid (Fig. 11) generalizes the cascade structure of the system of Viola and Jones [12] to suit the multiview case. The full range of out-of-plane rotations is divided into increasingly narrower ranges. The implementation of the detector-pyramid consists of 13 detectors distributed on three levels. They are learned independently from each other, using face examples for the corresponding view range and bootstrapped nonface examples. The detector on the top is for the coarsest classification of faces in the whole range of out-of-plane rotation. It is trained with face examples in the view range [-90°,+90°]. The second level is composed of three detectors, respectively trained with face examples in the subranges of [-90°,-30°], [-30°,+30°] and [+30°,+90°]. At the third level, the full range of [-90°,+90°] is partitioned into nine view groups : [-90°,-70°], [-70°,-50°], [-50°,-30°], [-30°,-10°], [-10°,+10°], [+10°,+30°], [+30°,+50°], [+50°,+70°] and , [+70°,+90°]. The resulting detectors are for the finest classification. The final result is obtained after merging the subwindows that pass the nine channels at the bottom level.

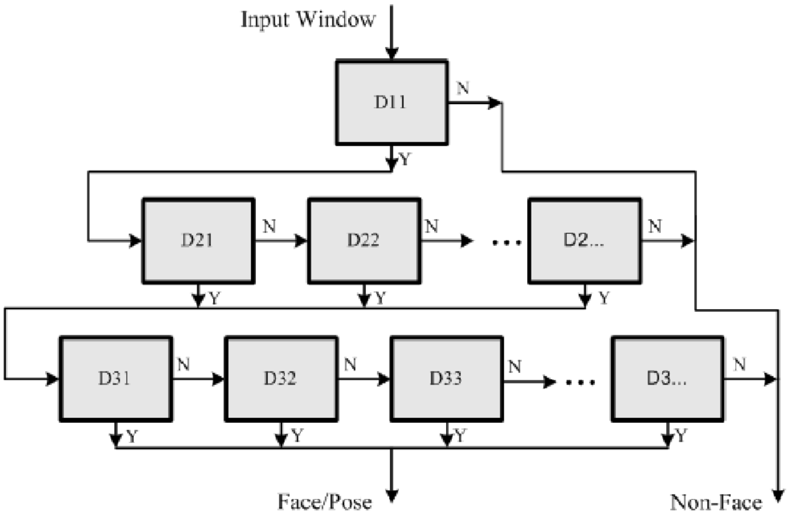


Figure 11: Detector-pyramid for the multiview face detection system of Li and Zhang.

The detector-pyramid discards as many nonface subwindows as possible at the earliest possible stage, so that only a few subwindows are processed further by the later stages. Indeed, the detectors in the early stages are simpler so as to reject a vast number of nonface subwindows with little computation, whereas those in the later stage are more complex and spend more time on each subwindow. Therefore, each detector in the pyramid consists of a cascade of strong classifiers for efficient classification, following the idea of Viola and Jones [12]. The top-level detector consists of a cascade of three strong classifiers. At the second level, each detector is a cascade of six strong classifiers and at the bottom level, each detector is a cascade of about 20 strong classifiers.

Their multiview face detection system is able to process an image of 320×240 pixels in approximately 0.2 seconds on a 700 MHz Pentium III. However, it is not specified if the detector is only applied on the original image, or if it is also applied on the two in-plane rotated images to deal with in-plane rotations, since they only use frontal and profile test sets to test their algorithm.

Garcia and Delakis presented in [19] a novel multiview face detection approach based on a convolutional neural network architecture. It is designed to detect face patterns of variable size and appearance, rotated up to ± 20 degrees in the image plane and turned up to ± 60 degrees out of the image plane. Their system acts like a pipeline of simple convolution and subsampling modules that treat the raw input image as a whole, without requiring any costly local preprocessing before classification.

The Convolutional Neural Network, that they call Convolutional Face Finder, consists in six layers, that receives as an input an image area of size 32×36 pixels to be classified as face or nonface. The first four layers contain a series of planes where successive convolutions and subsampling operations are performed. These planes extract and combine a set of appropriate features. The last two layers carry out the classification task using the features extracted in the previous layers. The detector is trained using highly variable face patterns artificially rotated by ± 20 degrees, covering the range of ± 60 degrees out-plane. The nonface patterns are collected via an iterative bootstrapping procedure.

They reported high detection rates with a particularly low level of false alarms, compare to other state-of-the art approaches, while processing an image of 384×288 pixels in approximately 0.25 seconds on a 1.6 GHz Pentium IV.

3 The Baseline Frontal Face Detection System

The frontal face detection system developed at the IDIAP Research Institute is inspired by the work of Viola and Jones [12]. The main difference lies in the feature set. By contrast to the Haar-like features used by Viola and Jones, IDIAP's features are invariant to illumination and convey only structural object information.

3.1 Feature Space

The feature space is defined as a set of 3×3 kernels, called Local Structure Kernels [21], which contain the local spatial image structure. The information is coded as binary information {0,1}. There exist $2^9 - 1 = 511$ such kernels (the kernel with all elements 0 and the one with all elements 1 convey the same information).

A modified version of the census transform[3] is used to obtain the index of the best kernel. The Modified Census Transform (MCT) is a non-parametric local transform, that captures the local spatial image structure. It maps the neighborhood of a pixel, including this pixel, to a bit string, comparing the intensity values of the pixels with the intensity mean on this neighborhood. Let a comparison function $\zeta(I(x), I(x'))$ be 1 if $I(x) < I(x')$ and let $\otimes$ denote the concatenation operation, the MCT at x is defined as

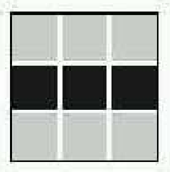
$$\Gamma(x) = \otimes\, \zeta(\bar{I}(x), I(y)) \quad \forall y \in N$$

where $\bar{I}(x)$ is the intensity mean on the neighborhood N of the pixel x, including x. $\Gamma(x)$ may be interpreted as the index of a structure kernel defined on $N(x)$. To illustrate the MCT, let us consider a region of an image (see Fig. 12). The pixel intensities of a 3×3 region are given in the first column and the structure kernel assigned by the MCT is displayed in the second column. The intensity mean in the example is: $\bar{I}(x) = \frac{27}{9} = 3$. The pixels of the first and last rows are all below or equal to the intensity mean, whereas the pixels in the middle row are all above. We can notice indeed that the MCT captures the local image structure correctly.

As no intensity value is encoded in the kernel, the representation is invariant to illumination. This is an interesting property, since illumination is one of the biggest

$I(x)$ $\Gamma(x)$

$$\begin{matrix} 1 & 2 & 1 \\ 7 & 5 & 6 \\ 1 & 1 & 3 \end{matrix}$$



(a) (b)

Figure 12: Illustration of the Modified Census Transform. (a) 3×3 region of an image with the corresponding pixel intensities and (b) kernel assigned by the MCT.

challenge in face detection. Preprocessing, such as histogram equalization, is thus not needed, saving costly computation.

3.2 Training and Classification

About 5500 frontal face patterns were collected to train the face detector. These face examples come from various databases available on the Web, such as Feret, Banca, Essex, Stirling, Yale and CMU frontal training databases. Fig. 13 shows some images extracted from these databases. Besides the faces, the images contain a lot of background. The first step is thus to extract the faces from the images to build the training set. This requires to build a model of the face.

3.2.1 Face Modelling

The corner of the eyes of each face were manually labelled. From these ground truth labels, the faces were automatically extracted, scaled and aligned to a 19×19 pixel window, according to the measures given by Farkas et al. in [23]. The bounding box of the faces is computed in the following way (see Fig. 14):
A ratio is calculated between the distance between both eyes in the original image, EE, and the distance between both eyes in the 19×19 pixel window, D_EYES, in order to map the original face to the 19×19 pixel window. EE is computed using the center coordinates of the eyes, given from the labels of the corner of the eyes. Knowing that the width of the face corresponds to the width of the window, D_EYES is computed as follows:

$$\begin{aligned} D\_EYES &= \frac{2 \cdot pupil\_se \cdot model\_width}{zy\_zy} \\ &= \frac{2 \cdot 33.4 \cdot 19}{139.1} \\ &\simeq 10 \text{ pixels} \end{aligned}$$

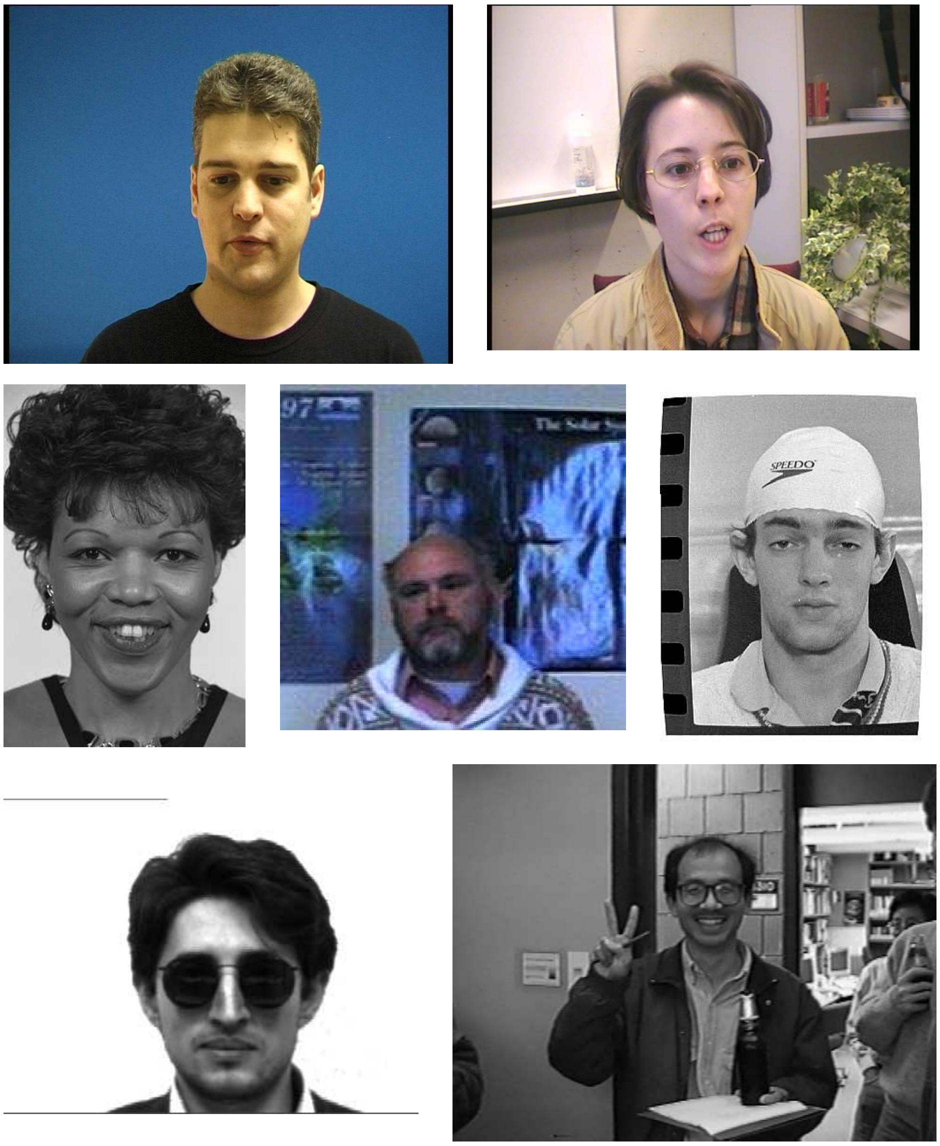


Figure 13: Example of face images from different databases.

where $pupil\_se$ is the pupil-facial midline distance, zy_zy the width of the face and $model\_width$ the width of the window. We have thus:

$$ratio = \frac{EE}{D\_EYES}$$

The upperleft x coordinate of the face is given by the middle of the face (center of the eyes) minus the half size of the face:

$$upperleft_x = c_x - (ratio * model\_width/2))$$

where c_x is the x coordinate of the middle of the face.

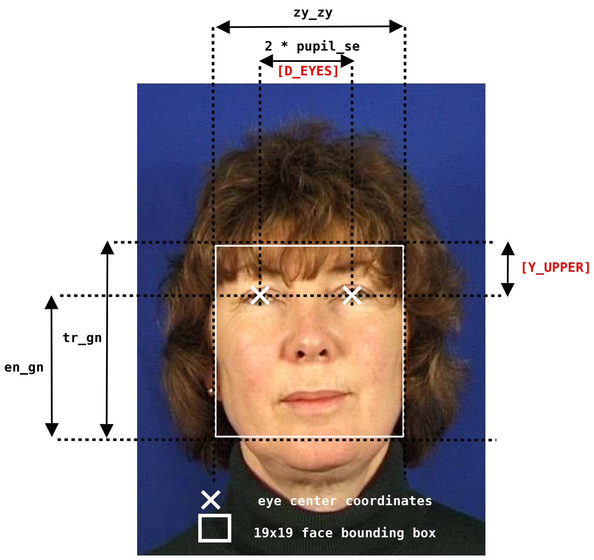


Figure 14: Face modelling using eyes center coordinates and facial anthropometry measures.

The upperleft y coordinate of the face is given by the constants en_gn and tr_gn, respectively the lower half of the craniofacial height and the physiognomical height of the face. In a similar way to D_EYES, the distance between the eye and the top of the window, Y_UPPER, is computed as follows:

$$\begin{aligned} Y\_UPPER &= \frac{\frac{tr\_gn - en\_gn}{2} \cdot model\_width}{zy\_zy} \\ &= \frac{\frac{187.2-117.7}{2} \cdot 19}{139.1} \\ &\simeq 5 \text{ pixels} \end{aligned}$$

The upperleft y coordinate of the face is thus given by:

$$upperleft_y = c_y - (ratio * Y\_UPPER)$$

where c_y is the y coordinate of the eyes center.

3.2.2 Face and Nonface Datasets

In addition to these cropped faces, fifteen face patterns were generated from each image of the training set by randomly translating, rotating and scaling the images, leading to a set of almost 90000 face patterns. Perturbations are generated to make the detector more robust to small perturbations in position and scale and to increase the number of available face patterns. Within this set, about 11000 faces were randomly chosen to compose the training set and another 11000 to compose the validation set.

Both sets contain faces of different individuals in order to obtain an unbiased estimation of the parameters[1]. Some of the face patterns used for training are shown in Fig. 15.

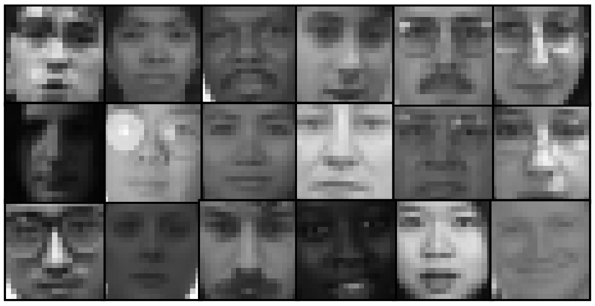


Figure 15: Example of face patterns composing the training set.

1614 images were used to collect nonface patterns. Most of the images represent various backgrounds. Some images also represent people in front of complex backgrounds, so that some nonface patterns can contain parts of faces (the faces are labelled, so they are excluded from the training set). About 400 million nonface patterns are available at all locations and scales in the training images. Some of them are shown in Fig. 16.

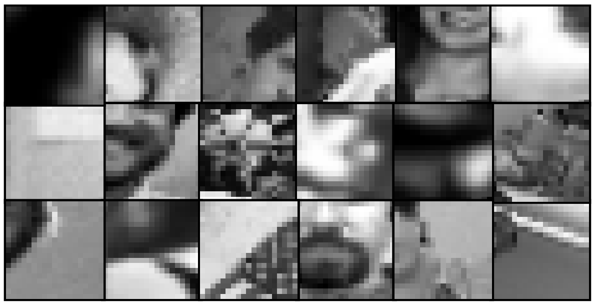


Figure 16: Example of nonface patterns composing the training set.

3.2.3 Cascade Training

As in the approach of Viola and Jones [12], both the feature selection and the classifier training were done using a variant of AdaBoost learning algorithm (see Section 2.1.6). Each weak classifier depends on a single MCT feature. They are combined into a cascade of strong classifiers, so-called MCT cascade, which are a weighted combination of weak classifiers. At each stage, 11000 nonfaces are randomly selected among the nonface set (400 millions patterns are available at the first stage). All features are computed on the face training set and on the nonface training subset. The features chosen are those which minimize the classification error. The number of features and the detection rate of each stage are defined by the user and the threshold is determined

[1]In this work, the training set and the validation set will always contain faces of different individuals.

by testing the detector on the face validation set. The classifier is then passed on all the nonface set and only the nonface patterns misclassified as faces are kept for the following stage.

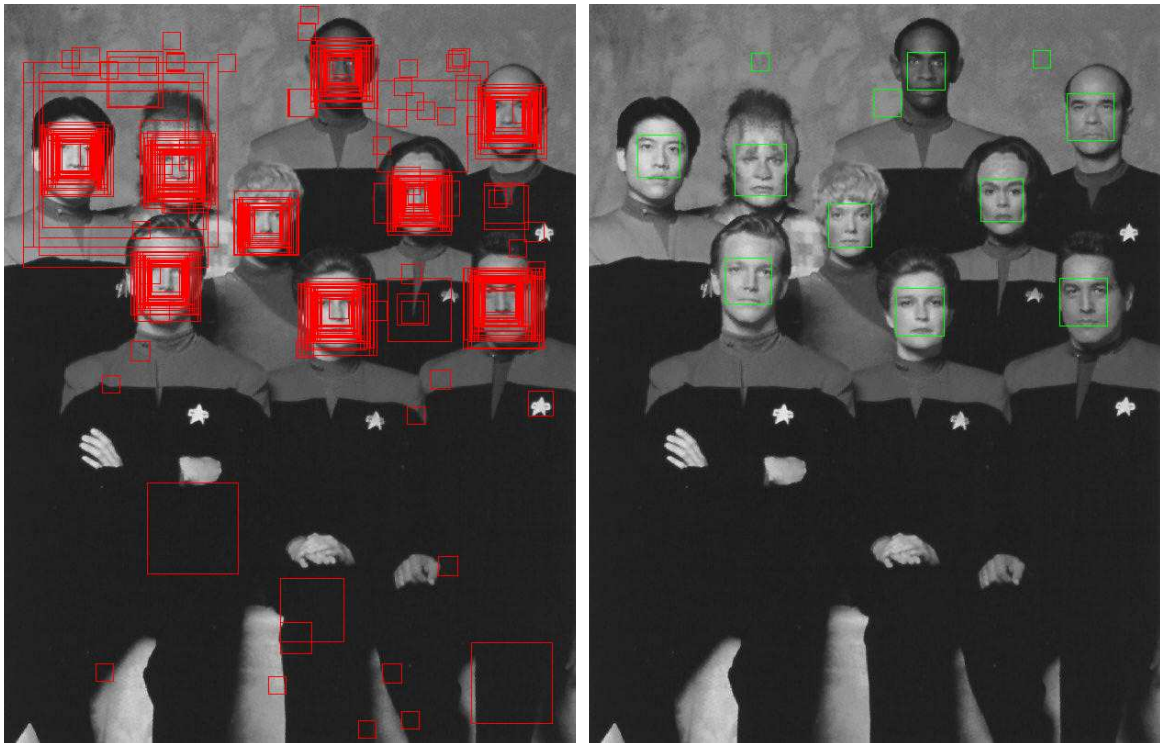
The final frontal face detector is a 4 stage cascade of classifiers which includes a total of 262 features. Each stage is trained to detect 99.5% of the faces. The first classifier in the cascade is constructed using two features and rejects about 50% of nonfaces. The second classifier has 10 features and rejects another 40% of nonfaces. The third classifier has 50 features and the last one 200.

The face detector can process a 384×256 pixel image in approximately 0.05 seconds on a 3.2 GHz Pentium IV.

3.3 Merging Overlapping Detections

Since the face detector has some invariance to position and scale, multiple detections usually occur around faces and nonfaces. The overlapping detections are merged to obtain one final detection per face and to reduce the number of false alarms.

First, the set of subwindows that passed all the stages of the detector is partitioned into disjoint subsets of overlapping detections. Then, if a detection does not overlap with another detection and/or if its score is below a threshold, the detection is rejected. It is indeed more likely that this detection is a false alarm (see Fig. 17). In each partition, only the detections which have above 60% of their surface in common are kept. Finally, the final detection is the mean of these detections.



(a) (b)

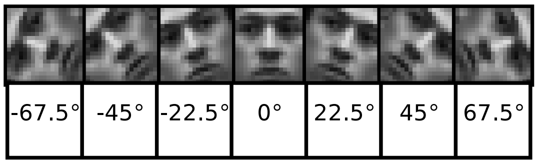
Figure 17: Output of the frontal face detector (a) before merging overlapping detections and (b) after.

4 The Proposed Multiview Face Detection System

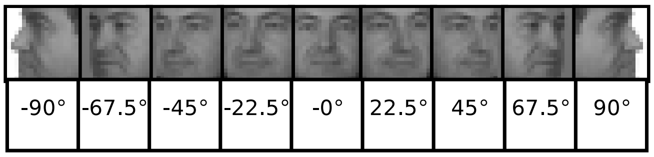
The purpose of this project is to extend the frontal face detection system described in Section 3. The proposed architecture generalizes the cascade structure of the frontal face detector. It is inspired by the detector-pyramid of Li and Zhang [18].

4.1 Dealing with Head Rotations

The system deals with two types of head rotations, out-of-plane and in-plane. It is trained to detect faces in the range of [-90°,+90°] out-of-plane and in the range of [-67.5°,+67.5°] in-plane. The ranges were chosen to suit real world images. The face space is divided into 7 subspaces in the in-plane case and into 9 subspaces in the out-of-plane case, as shown in Fig. 18. Thus, the multiview face detector deals with 16 different poses.



(a)



(b)

Figure 18: (a)In-plane and (b) out-of-plane view partitions.

Three steps were needed to obtain the multiview face detector. First, a classifier was build dealing only with out-of-plane rotations. Then, another classifier was build dealing only with in-plane rotations. Finally, both classifiers were integrated into a multiview classifier. These three detectors all have a pyramid-like structure.

4.2 Detector-Pyramid

4.2.1 Out-of-plane Face Detector

The detector-pyramid architecture of the out-of-plane detection system is illustrated in Fig. 19. It consists of 13 detectors distributed on 3 levels. This architecture is directly inspired by Li and Zhang [18]. However, it differs in the structure of the bottom level. The detector on the top is trained with face examples in the view range [-90 $^\circ$,+90 $^\circ$]. The second level is composed of three detectors, respectively trained with face examples in the subranges of [-90 $^\circ$,-45 $^\circ$], [-22.5 $^\circ$,+22.5 $^\circ$] and [+45 $^\circ$,+90 $^\circ$]. At the third level, the full range is divided into nine view groups, according to the partitions in Fig. 18 (b). Each view corresponds to one detector, so there are nine detectors which are trained independently from each other. Each detector in the pyramid is a MCT cascade, constructed in a similar way as the classifier in the frontal face detection system.

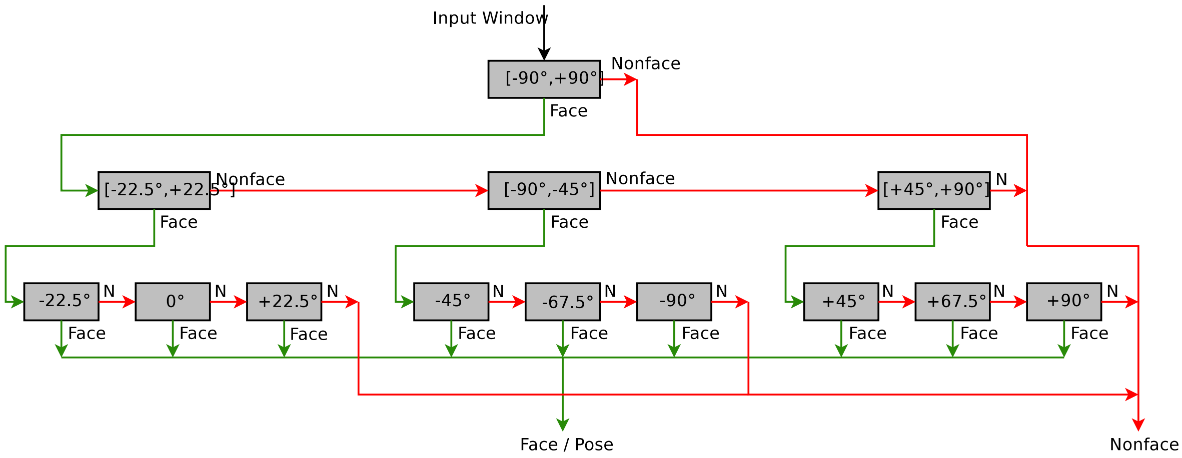


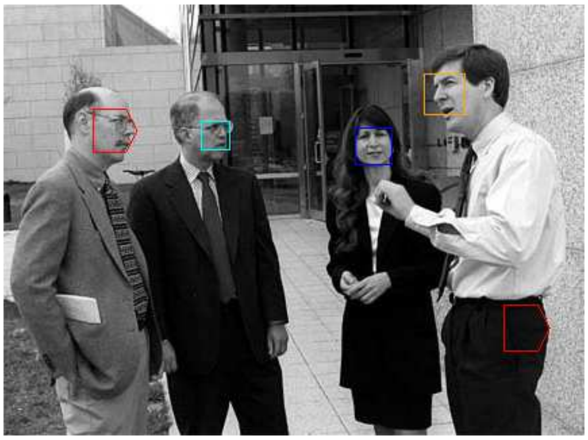
Figure 19: Detector-pyramid for the out-of-plane face detection system.

A 19×19 input window is first processed by the top-level detector of the detector-pyramid. If not rejected as a nonface, it goes through the first detector of the second level, the [-22.5 $^\circ$,+22.5 $^\circ$] detector. The window is processed by the second detector of the stage, the [-90 $^\circ$,-45 $^\circ$] detector, only if it was previously rejected as a nonface, and so on with the third detector of the stage. Otherwise, if the window is classified as a face by one of the three detectors, it goes directly to the third level and is processed in the same way by the three corresponding bottom detectors. For example, if the window is classified as a face by the [-90 $^\circ$,-45 $^\circ$] detector, it is then processed by the -45 $^\circ$ detector. If this detector rejects the window, it is processed by the -67.5 $^\circ$ detector. If the window is rejected again, the window will be processed by the -90 $^\circ$ detector. If at one point, one of the three detectors classifies the window as a face, the processing stops and a detection occurs. The pose of the face is the view of the detector which classified the windows as a face. The window will thus not be processed by the detectors in the range [+45 $^\circ$,+90 $^\circ$]. The final result is obtained after merging the subwindows that pass the nine detectors at the bottom level, as illustrated in Fig. 20.

(a)

(b)

(c)



(d)

Figure 20: Outputs of the bottom detectors before merging, respectively (a) -90 $^\circ$, -67.5 $^\circ$, -45 $^\circ$ (b) -22.5 $^\circ$, 0 $^\circ$, +22.5 $^\circ$ (c) +45 $^\circ$, +67.5 $^\circ$, +90 $^\circ$ and (d) final result after merging.

4.2.2 In-plane Face Detector

Instead of turning the image to handle the in-plane rotations like in [18], we use a similar architecture than the out-of-plane detector, as illustrated in Fig. 21. It consists of 8 detectors distributed on 2 levels followed by 3 multi-layer perceptrons, or MLPs, used for postprocessing. The top-level detector is trained with face examples covering the view range [-67.5 $^\circ$,+67.5 $^\circ$]. At the second level, the full range is divided into seven view groups, according to the partitions in Fig. 18 (a). As previously, each view corresponds to one detector, independently trained. Each detector in the pyramid is also a MCT cascade.

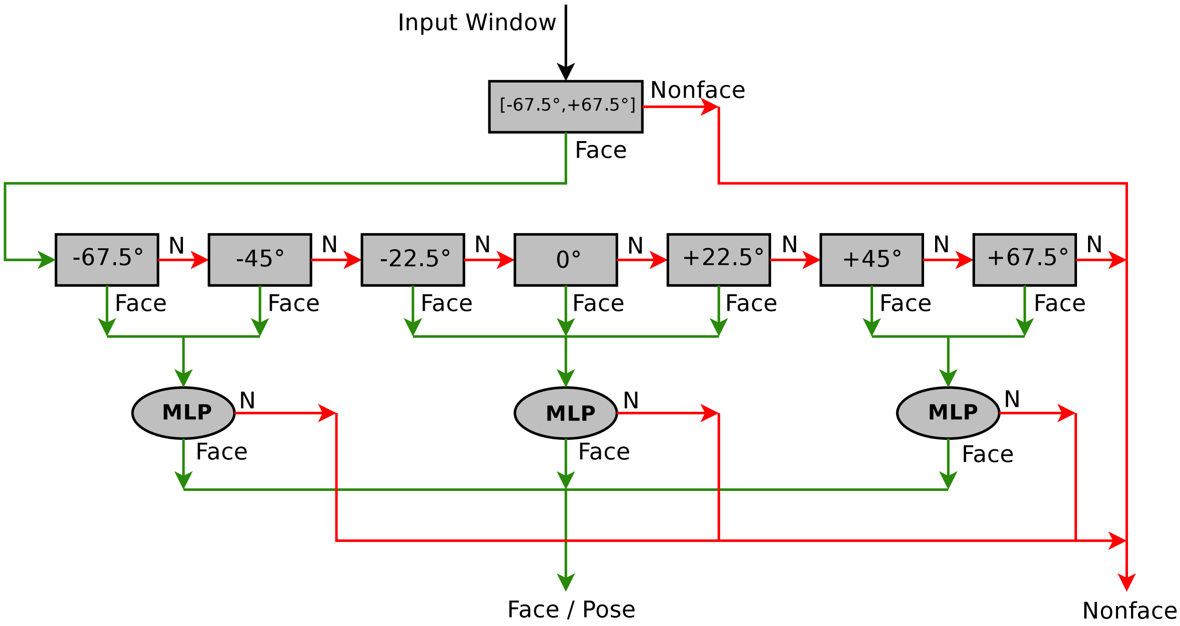
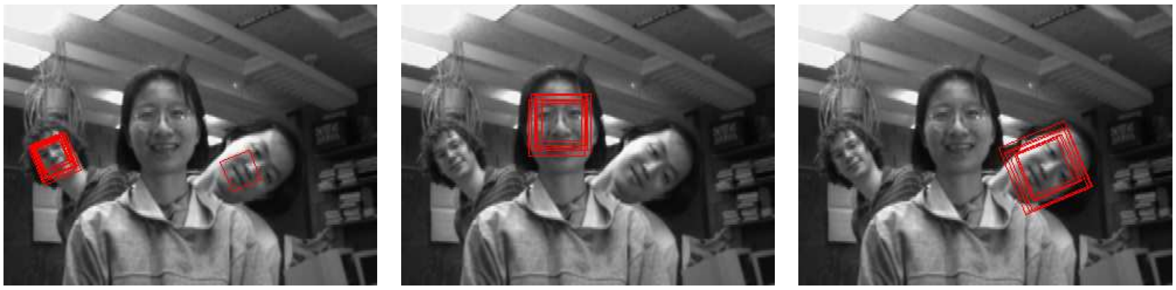
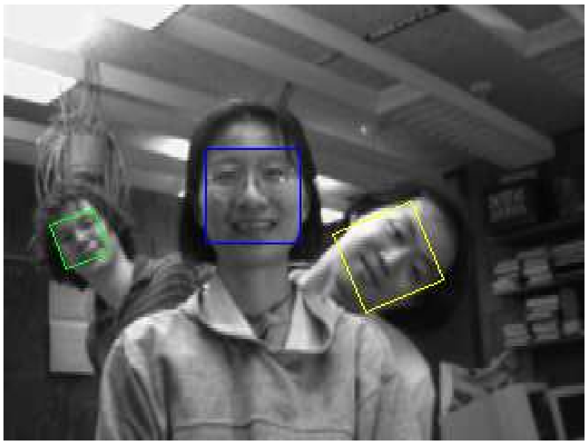


Figure 21: Detector-pyramid for the in-plane face detection system.

As in the out-of-plane detector, a 19×19 input window is first processed by the top-level detector of the detector-pyramid. If not rejected as a nonface, the window goes to the second level and is processed by the 67.5 $^\circ$ detector. The window goes to the next detector only if it is rejected by the current one. When one of the bottom detectors classifies the window as a face, the window is postprocessed by a multiclass multi-layer perceptron, or MLP, that checks if the window is a face. A detection occurs if the MLP classifies the window as a face. As previously, the final result is obtained after merging the subwindows that pass the 3 MLPs, as illustrated in Fig. 22.



(a)



(b)

Figure 22: Outputs of the MLPs (a) before merging, (b) after merging.

4.2.3 Multiview Face Detector

The architecture of the final detector is illustrated in Fig. 23. It is composed of both out-of-plane and in-plane detectors.

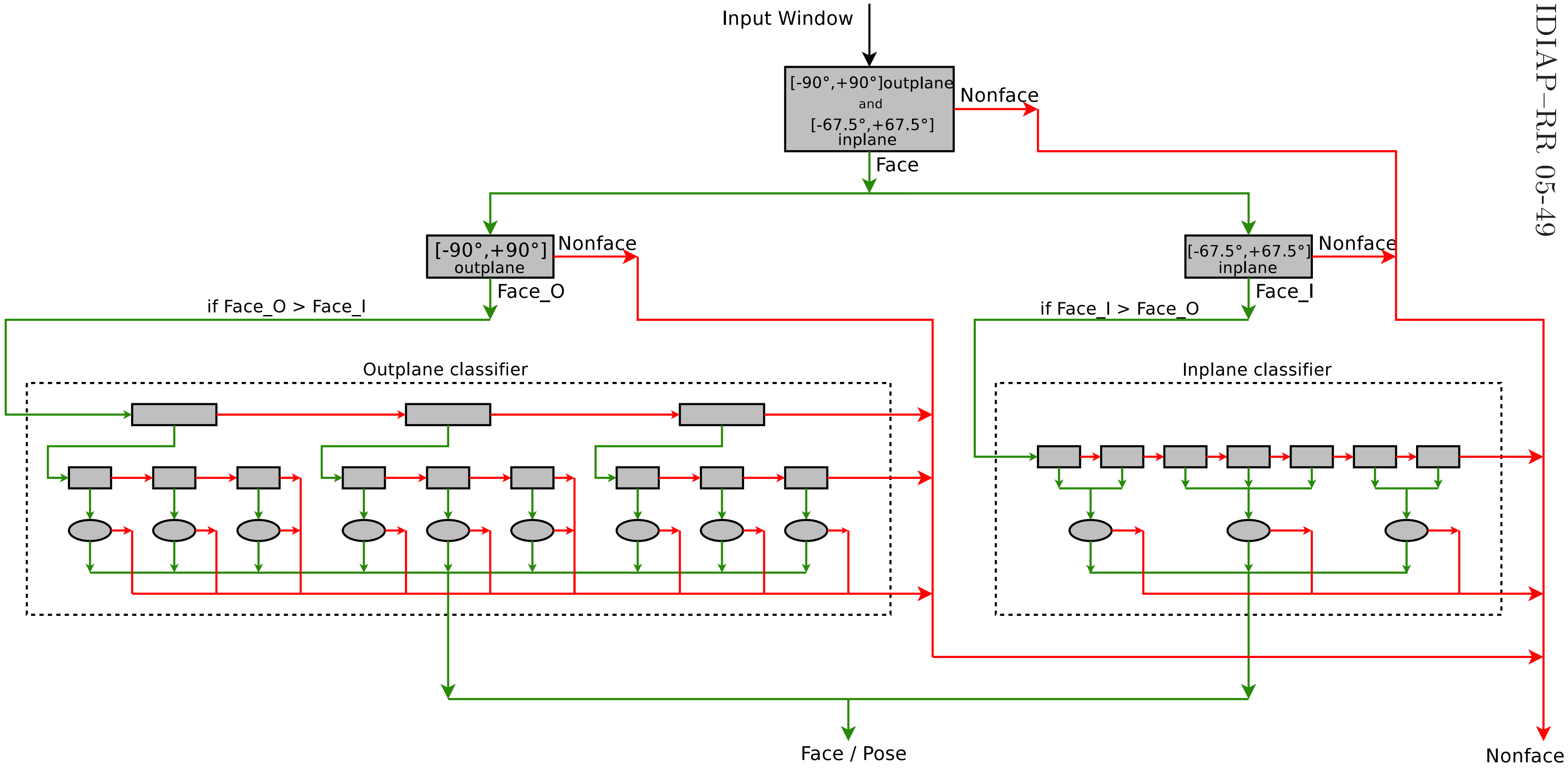


Figure 23: Detector-pyramid for the multiview face detection system.

It consists of 22 detectors distributed on 3 levels followed by 12 MLPs for postprocessing. The top-level detector is trained with all the available faces (faces covering the view range [-90°,+90°] out-of-plane and [-67.5°,+67.5°] in-plane). The second level consists of 2 detectors, which are the top-level detectors of the out-of-plane and in-plane detector-pyramids. The third level corresponds to the following levels of the out-of-plane and in-plane detectors. In addition to the 3 MLPs used in the in-plane case, there are 9 MLPs at the output of each bottom detectors of the out-of-plane detector-pyramid.

A 19×19 input window is first processed by the top-level detector. If not rejected as a nonface, the window goes to the second level and is processed by the two detectors. If they both reject the window, the processing stops and the window is classified as a nonface. If one of them classifies the window as a face, the window is processed only by the corresponding detector-pyramid (out-of-plane or in-plane) in the way described in sections 4.2.1 and 4.2.2. If both detectors classify the window as a face, the score of the outputs is compared and the window is processed by the detector-pyramid whose top-level detector has the maximum score. When one of the bottom detectors classifies the window as a face, the window is given to the corresponding MLP that checks if it is a face. The final result is obtained after merging the outputs of all the bottom detectors that pass the MLPs.

4.3 Postprocessing

Two postprocessing strategies are applied on the outputs of the face detector to reduce the number of false alarms and to eliminate the overlapping detections.

4.3.1 Multi-Layer Perceptron

As seen previously, each output of the multiview face detector is given to a MLP corresponding to the view of the output. In the out-of-plane case, an MLP is not able to classify the different poses in a sufficiently robust way, but is capable of separating the two classes (face versus nonface) with a low classification error rate. There are thus 9 MLPs, each of them trained with a set of faces rotated by a given angle and a set of nonfaces. In the in-plane case, on the contrary, an MLP can robustly separate the different poses. We could use one MLP that would separate the 7 in-plane poses and the nonface class with a low classification error rate. But the processing time would be too long. Consequently, three MLPs are used instead of one. Two of them classify two different poses in addition to the nonface class and the third one classifies three different poses also in addition to the nonface class.

The MLPs are trained using k-fold cross-validation in order to select the appropriate number of hidden units. The data is divided here into 3 subsets of approximately equal size. The neural network is trained 3 times, each time leaving out one of the subsets from training, but using only the omitted subset to compute the error criterion.

4.3.2 Merging Overlapping Multiview Detections

As in the frontal face detection system, multiple detections usually occur around faces and nonfaces. However, the problem is in this case not exactly the same, since each detection has a specific pose. Merging all the overlapping detections by taking the mean detection like in the frontal face detection system, would not mean anything, especially merging in-plane with out-of-plane detections. The merging process is thus divided in two steps, illustrated in Fig. 25.

First, the set of detections is partitioned into 16 subsets, each subset corresponding to one view. In each subset, the overlapping detections are merged like in the frontal face detection system, as explained in section 3.3 (see Fig. 25 (b)). Then, the new set of detections is partitioned into disjoint subsets of overlapping detections, where two detections overlap if they have at least 40 % of their surface in common. Finally, for each subset, the final detection is the maximum of the overlapping detections (see Fig. 25 (c)).

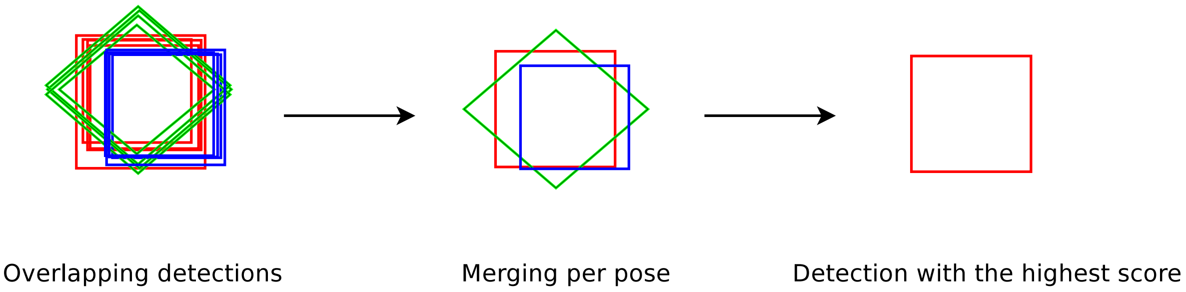
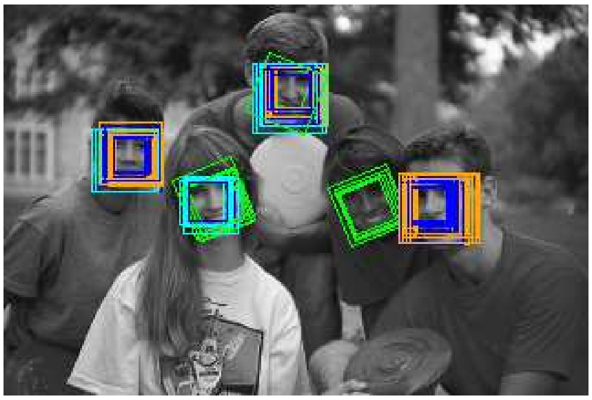
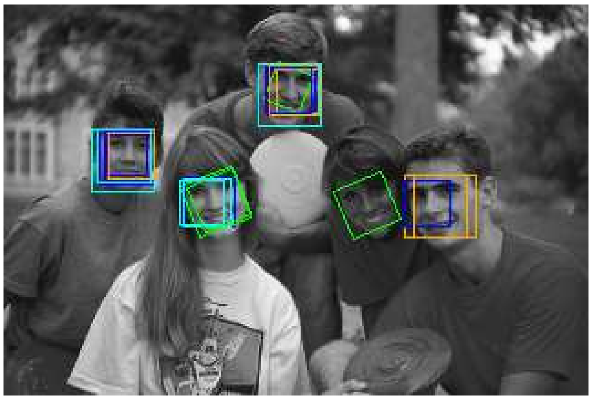


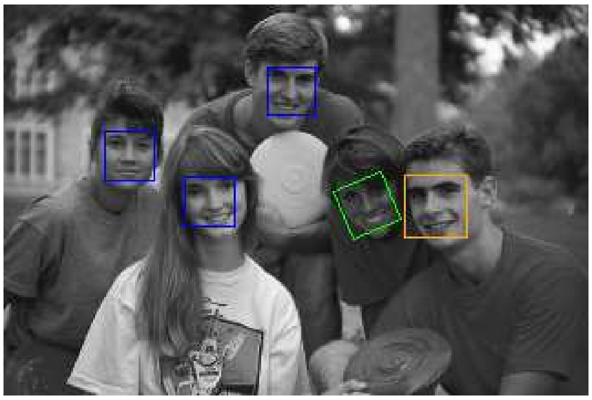
Figure 24: Merging overlapping multiview detections.



(a)



(b)



(c)

Figure 25: Output of the multiview detector-pyramid (a) before merging, (b) after merging the overlapping detections per pose and (c) after final merging.

5 Experimental Results

5.1 Databases

Several databases are commonly referenced in the literature. They have been compiled in order to compare the level of performance of the different face detection methods. Databases are here classified in two categories: training databases and testing databases.

5.1.1 Training Databases

- **FERET** - This database consists of 14051 grayscale images of 1199 individuals in front of a uniform background, with views ranging from frontal to left and right profiles [24]. It was designed at first to develop and evaluate face recognition algorithms, but it also can be used to train face detection algorithms.

- **PIE** - This database is known as the CMU Pose, Illumination, and Expression Database [25]. It consists of 41368 color images of 68 people. Each person is taken under 13 different poses, 43 different illumination conditions and with 4 different expressions in front of a complex background.

- **Prima Head Pose**[4] - This database consists of 15 sets of color images and was collected at INRIA Rhone-Alpes [26]. Each set contains 2 series of 93 images of the same person in front of a uniform background at different poses in the range of [-90°,+90°] out-of-plane and [-90°,+90°] up-and-down.

- **Stirling**[5] - This database comes from the Psychological Image Collection at Stirling (PICS) and was collected by the University of Stirling Psychology Department. The subset used contains 1161 frontal images.

- **Yale**[6] - This database contains 165 grayscale images of 15 subjects. There are 11 frontal views per subject, corresponding to different facial expressions and lighting conditions.

- **Essex**[7] - This database consists of 7900 color images of 395 people and was collected by Libor Spacek at Essex University. The subset used contains 392 images taken in front of simple and complex backgrounds.

5.1.2 Testing Databases

- **CMU-130**[8] - This data set is, so far, the most widely-used face detection benchmark database. It was collected by Henry A. Rowley, Shumeet Baluja, and Takeo Kanade at CMU and contains 130 grayscale images with a total of 507 frontal human faces, as well as hand-drawn faces. It includes images from the Web, scanned from photographs and newspaper pictures, that are of varying size and quality. It contains many images with faces against complex backgrounds and some images without any faces. This data set also includes 23 images of the data set used by Sung and Poggio [8], referred to as the MIT test set.

- **CMU Profile face set**[9] - This data set consists of 208 images with 441 faces of which 347 are profile views. They were collected from various news Web sites at CMU by Henry Schneiderman and Takeo Kanade.

- **CMU Rotated face set**[8] - This data set contains 50 grayscale images with a total of 223 faces, of which 207 are rotated by an angle in the range [-67.5°,+67.5°] in-plane.

- **Web** and **Cinema** - These two data sets were collected by Garcia and Delakis [19]. The Web test set contains 215 images with 499 faces. The images come from a large set of images that have been submitted to the interactive demonstration of their system, available on the Web. The Cinema test set consists of 162 images extracted from various movies, containing 276 faces. It contains a high number of faces with extreme facial expressions as well as faces which are partially occluded or heavily shadowed, surrounded by very complex backgrounds.

- **Sussex**[10] - This face database was collected by Jonathan Howell at the University of Sussex. It is composed of 10 individuals with 10 orientations between 0 to -90 degrees, leading to a total of 100 grayscale images with 100 faces. The faces are surrounded by a simple background.

5.2 Performance Evaluation

Several measures are used to evaluate the performance of the different face detection systems:

- The *detection rate* (DR) is defined as the ratio between the number of faces correctly detected and the number of faces determined by a human. But the definition of what is a correctly detected face is often missing in the literature, leading to some uncertainty in the systems performance [27].

- Two types of errors can be made : a *false alarm* (FA) is a background region classified as a face and a *false rejection* (FR) is a face classified as background and thus discarded. From these measures are defined the *false alarm rate* (resp. *false rejection rate*) which is the number of FA (resp. FR) over the total number of subwindows tested.

- The *Receiver Operating Curve* (ROC) represents the detection rate in function of the false alarm rate or the number of false alarms. It allows a better visualization of the performance of a face detector and ease the comparison between several systems.

The objective is thus to maximize the detection rate while minimizing the number of false alarms. But obtaining a high detection rate with a small number of false alarms is a difficult task, since increasing the detection rate usually increases the number of false alarm. Consequently, a trade off has to be found between both parameters.

5.3 Experimental Setup

5.3.1 Training Datasets

- Out-of-plane face sets

4700 face patterns were collected to train the out-of-plane face detector. These face examples come from Feret, PIE and Prima Head Pose databases. Table 1 shows the distribution of the face patterns depending on the head pose. More 90° faces were collected because full profile faces contain a lot of background and are thus more difficult to classify.

Head pose	# faces
22.5°	1000
45°	1000
67.5°	1000
90°	1700

Table 1: Distribution of the face patterns in the out-of-plane training datasets.

The faces were manually labelled, depending on the pose. For the 22.5°, 45° and 67.5° faces, the eyes and the chin were labelled. For the 90° faces, the eye, the nose and the chin were labelled. Given these ground truth labels, the faces were automatically extracted and scaled to a 19×19 pixel window according to a non-frontal face model depending on the pose of the face (see Fig. 26). In frontal faces, the distance between the eyes is constant and is thus used to compute a ratio between the original image and the 19×19 pixel window. In non-frontal faces, on the contrary, the distance between the eyes is not constant and can not be used for the ratio. Instead, the ratio is calculated using the distance between the eye and the chin in the original image, EC, and the distance between the eye and the chin in the the 19×19 pixel window, EYE_CHIN (the measures are given by [23]). EYE_CHIN is computed as follows:

$$\begin{aligned} EYE\_CHIN &= \frac{en\_gn \cdot model\_width}{zy\_zy} \\ &= \frac{117.7 \cdot 19}{139.1} \\ &\simeq 16 \text{ pixels} \end{aligned}$$

where *en_gn* is the lower half of the craniofacial height, *zy_zy* the width of the face and *model_width* the width of the window. We have thus:

$$ratio = \frac{EC}{EYE\_CHIN}$$

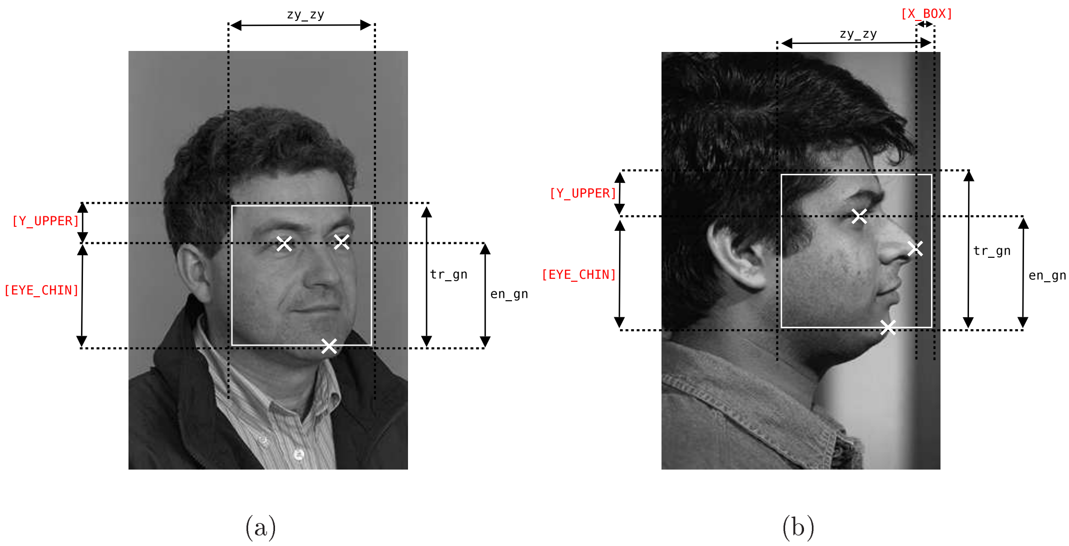


Figure 26: Non-frontal face models (a) 22.5 $^\circ$, 45 $^\circ$ and 67.5 $^\circ$, (b) 90 $^\circ$.

In the frontal case, the bounding box is centered on the middle of the eyes. However, in the non-frontal case, the center of the eyes does not really correspond to the center of the bounding box any more.
The upperleft x coordinate of the face is thus given by:

$$upperleft_x = c_x - (ratio * model\_width/2))$$

where c_x is the x coordinate of the center of the bounding box. In the 22.5 $^\circ$ case, it is defined as the center of the eyes. In the 45 $^\circ$ and 67.5 $^\circ$ cases, the center of the eyes is shifted to the left and then taken as the center in x of the bounding box. In the 90 $^\circ$ case, there is only one eye on the face, so the upperleft x coordinate of the face is redefined:

$$upperleft_x = n_x + (ratio * X\_BOX) - model\_width$$

where n_x is the x coordinate of the nose and X_BOX is fixed at 1 pixel. X_BOX corresponds to the space chosen between the nose and the boundary of the bounding box.
For all the poses, the upperleft y coordinate of the face is given by Y_UPPER, like in the frontal face model:

$$upperleft_y = c_y - (ratio * Y\_UPPER)$$

where c_y is the y coordinate of the point c defined above. In the 90° case, c_y corresponds to the y coordinate of the eye.

The faces were mirrored to obtain the -22.5°, -45°, -67.5° and -90° face sets. In addition to the cropped faces, fifteen face patterns were generated from each image of each training set by randomly translating, rotating and scaling the images leading to a set of 16000 face patterns / pose (resp. 27200 for the $\pm$90° poses). The face sets were divided in two to compose a training set and a validation set of 8000 face patterns each (resp. 13600) for each pose. 8 bottom detectors of the out-plane detector-pyramid were thus trained independently with the 8 corresponding training and validation sets. The frontal face detector was used as the 0° detector. The 3 second-level detectors were trained with a training set and a validation set of 4842 faces each, covering the view range [-90°, -45°], [-22.5°, +22.5°] and [+45°,+90°]. The top-level detector was trained with a training set and a validation set of 14416 faces each, covering the view range [-90°,+90°].

- In-plane face sets

5575 frontal face patterns were collected to train the in-plane face detector. These face examples come from Feret, Stirling, Essex and Yale databases. The faces were automatically extracted from the images using the frontal face model described in section 3.2 and scaled to a 19×19 pixel window. In addition to these faces, face patterns were generated by randomly translating, rotating and scaling the images leading to a training set of 8744 images and a validation set of 9232 images. In order to obtain the 7 face sets corresponding to each pose, the frontal training and validation sets were rotated by 22.5° to obtain the 22.5° face set, by 45° to obtain the 45° face set and so on. The 7 bottom detectors of the in-plane detector-pyramid were thus trained independently with the 7 corresponding training and validation sets. The top-level detector was trained with a training set of 61208 faces and a validation set of 64624 faces, corresponding to all the faces available in the range [-67.5°,+67.5°].

The top-level detector of the multiview detector pyramid was trained with a training set of 28989 faces and a validation set of 29802 faces covering the view range of [-90°,+90°] out-of-plane and [-67.5°,+67.5°] in-plane.

Each detector of the multiview detector pyramid was trained with a nonface set composed of over 800 million patterns. The nonface patterns come from 5038 background images and face images, like the nonface set used to train the frontal face detector.

5.3.2 Structure of the detectors

Each detector in the multiview detector-pyramid is a MCT-cascade. Table 2 describes the structure of each cascade, that are the number of stages and the number of weak classifier per stage. Each stage of each detector is trained to detect 99.5% of the faces.

Top-level detector		
View range	# Stages	# WC/Stage
All views	3	10, 25, 50
Out-of-plane detectors		
View range	# Stages	# WC/Stage
[-90°,+90°]	2	10, 50
[-90°,-45°]	2	10, 50
[-22.5°,+22.5°]	2	10, 50
[+45°,+90°]	2	10, 50
-90°	4	30, 60, 100, 200
-67.5°	3	10, 50, 100
-45°	3	10, 50, 100
-22.5°	3	10, 50, 100
0°	4	2, 5, 10, 50
22.5°	3	10, 50, 100
45°	3	10, 50, 100
67.5°	3	10, 50, 100
+90°	4	30, 60, 100, 200
In-plane detectors		
View range	# Stages	# WC/Stage
[-67.5°,+67.5°]	2	10, 50
-67.5°	4	2, 5, 10, 50
-45°	4	2, 5, 10, 50
-22.5°	4	2, 5, 10, 50
0°	5	2, 5, 10, 50, 100
22.5°	4	2, 5, 10, 50
45°	4	2, 5, 10, 50
67.5°	4	2, 5, 10, 50

Table 2: Structure of the detectors in the multiview detector-pyramid.

5.4 Comparative Results

Comparing face detection methods is a difficult task, even though they are evaluated on the same databases. Indeed, only a few researchers give their definition of *what is a correctly detected face*, such as [19], or use some error measures, like the one introduced by Jesorsky et al. [28]. Moreover, the correct detections and false alarms are usually counted by hand, because no automatic evaluation system dealing with multiview faces exist yet. In this work, a detection is considered as correct if the bounding box contains the eyes and the mouth without too much background (see Fig. 27). The following results are all obtained with the same thresholds (each classifier in each detector has a specific threshold). The final threshold (used at the end of the detector-pyramid) was set to the same value as in the frontal face detector.

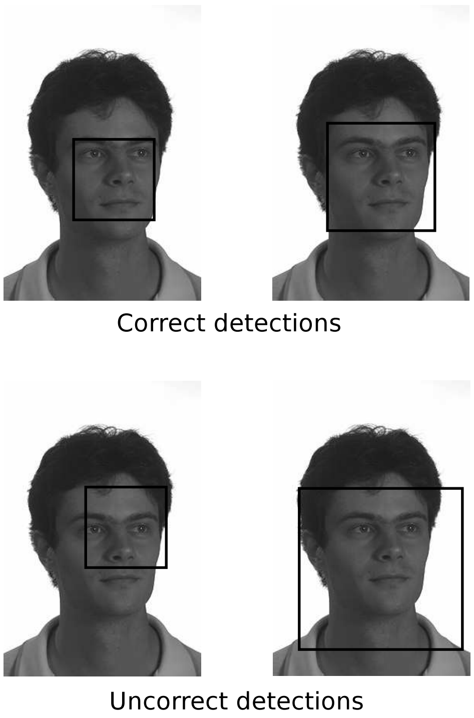


Figure 27: Examples of correct and uncorrect detections.

In the following experiments, different colors are used to draw the bounding boxes indicating the detections, as described in Table 3. Ideally, each color should represent a pose, but there was not enough colors available to differentiate the 16 poses.

Frontal	
Bounding Box Color	Pose
Blue	0°

In-plane	
Bounding Box Color	Pose
Green	-67.5°, -45°, -22.5°
Yellow	$+22.5^\circ$, $+45^\circ$, $+67.5^\circ$

Out-of-plane	
Bounding Box Color	Pose
Seagreen	-90°
Orange	-67.5°, -45°, -22.5°
Cyan	$+22.5^\circ$, $+45^\circ$, $+67.5^\circ$
Red	$+90^\circ$

Table 3: Colors of the detections bounding boxes.

5.4.1 Comparison between the Orginal Frontal System with the Proposed Multiview System

Table 4 compares the detection rate and the number of false alarms between the frontal face detector developed at the IDIAP Research Institute and the multiview face detector, on the CMU Frontal Test Set. The proposed system achieves a significantly higher detection rate (91.7%) than the frontal detector (84.6%) with a similar number of false alarms. Indeed, even though the CMU Frontal Test Set contains only frontal faces, some of them can be slightly rotated in-plane or out-of-plane (see Fig. 28 for some examples). The multiview face detector is more robust to variation in orientation and pose, so it can detect more easily those kinds of faces. However, it is two times slower than the frontal face detector to process a 384×256 image.

System	**CMU Frontal Test Set**	
	DR	**FA**
IDIAP Frontal Face Detector	84.6%	435
Multiview Face Detector	91.7%	441

Table 4: Comparison between the frontal face detector and the multiview face detector on the CMU Frontal Test Set.

Figure 28: Some results obtained on the CMU Frontal Test Set.

5.4.2 Comparison to State-of-the Art In-plane and State-of-the Art Out-of-plane Face Detectors

Table 5 compares the detection rate and the number of false alarms between the multiview face detector and two state-of-the art detectors on the CMU Rotated Test Set and the CMU Profile Test Set. The chosen state-of-the art detectors where those which obtained the best results on these test sets.

In the in-plane case (see Table 5 (a)), the chosen detector is the one introduced by Viola and Jones [17]. Our multiview face detector achieves a higher detection rate (92.3%) than Viola and Jones detector (89.7%), but with a higher number of false alarms. However, we can notice on their ROC curve that they obtain the same detection rate with a higher number of false alarms, which is around 700. Viola and Jones

System	CMU Rotated Test Set	
	DR	FA
Multiview Face Detector (in-plane and out-plane)	92.3%	342
State-of-the Art Face Detector (in-plane only)	89.7%	221

(a)

System	CMU Profile Test Set	
	DR	FA
Multiview Face Detector (in-plane and out-plane)	53.1%	416
State-of-the Art Face Detector (out-plane only)	92.8%	700

(b)

Table 5: Comparison between the multiview face detector and state-of-the art face detectors on (a) the CMU Rotated Test Set and (b) the CMU Profile Test Set.

trained their detector to detect 12 different poses covering the full 360 degrees in-plane range, whereas the multiview face detector was trained to detect 16 different poses in-plane and out-of-plane, covering only 135 degrees in-plane. Thus, the results can not be really compared since both detectors are not trained to detect the same type of faces. Some examples are presented in Fig. 29.

In the out-of-plane case (see Table 5 (b)), the chosen detector is the one proposed by Schneiderman and Kanade [20]. Our multiview face detector achieves a much lower detection rate (53.1%) than Schneiderman and Kanade detector (92.8%) with a lower number of false alarms. The low performance of our multiview face detection system on this test set can have several reasons. First, Schneiderman and Kanade trained their detector to cover the full 180° out-of-plane range, when the multiview face detector also detects [-67.5°,+67.5°] in-plane. Moreover, their detector only distinguishes frontal, from left or right profile, whereas the proposed system estimates a lot more precisely the pose: 16 poses are tested where Schneiderman and Kanade only test 3 poses. Then, many faces in the test set are very small, since their size is close to 19×19 pixels, corresponding to the limit of the detector. Finally, the proposed approach is a lot faster. Indeed, it takes about 1 minute to process a 320×240 pixel image with their detector, whereas the multiview face detector is real-time. However, as previously with the CMU Rotated Test Set, the results can not be really compared since both detectors are not trained to detect the same type of faces. Some examples from the CMU Profile Test set are shown on Fig. 30.

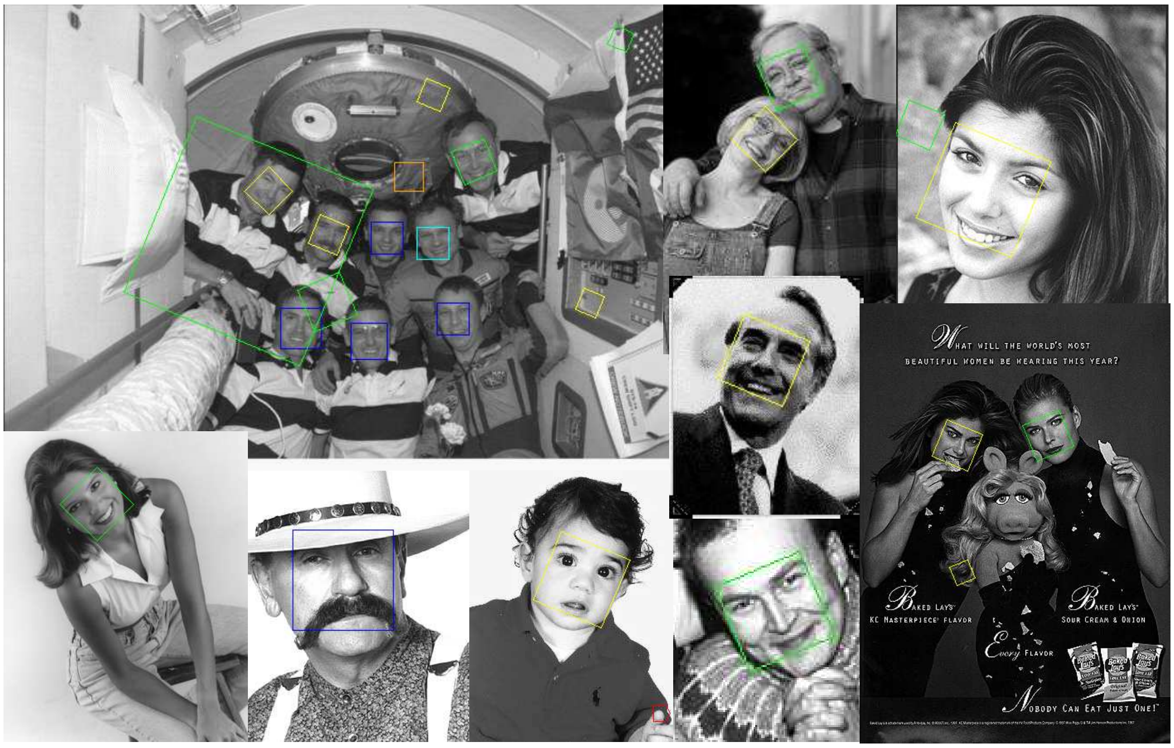


Figure 29: Some results obtained on the CMU Rotated Test Set.

Figure 30: Some results obtained on the CMU Profile Test Set.

5.4.3 Comparison to State-of-the Art Multiview Face Detector

Table 6 compares the detection rate and the number of false alarms between the multiview face detector and the one proposed by Garcia and Delakis [19] on the Web and Cinema Test Sets. Both detectors are here multiview detectors, by contrast to Viola and Jones and Schneiderman and Kanade detectors. This allows a better comparison. Our multiview face detector achieves a lower detection rate on the Web Test Set (94%) then Garcia and Delakis detector (98%), but a similar detection rate on the Cinema Test Set (95.3%). On both test sets, our multiview face detector obtains many more false alarms. However, Garcia and Delakis trained their detector on [-20°,+20°] in-plane and [-60°,+60°] out-of-plane. The view range covered is thus narrower than with our multiview detector, since the proposed system covers [-67.5°,+67.5°] in-plane and [-90°,+90°] out-of-plane. Moreover, they do not estimate the pose and our system is faster. Fig. 31 shows some results obtained on these test sets.

System	Web		Cinema	
	DR	**FA**	**DR**	**FA**
Multiview Face Detector	94%	743	95.3%	682
Garcia and Delakis Multiview Face Detector	98%	108	95.3%	104

Table 6: Comparison between the multiview face detector and Garcia and Delakis multiview face detector on the Web and Cinema Test Sets.

5.4.4 Pose Estimation

For each detection, the multiview face detector also estimates the pose. The pose estimation is evaluated on the Sussex face database (see Fig. 32). However, only the estimation of the out-of-plane pose can be evaluated on this database, since it only contains faces rotated out of the image plane. The 100 images were mirrored (see Fig.32 (b)) to obtain faces covering the view range [-90°,+90°], leading to a total of 200 images with 10 images per pose (20 for the frontal pose). The system achieves a detection rate of 98.5% with 10 false alarms and approximately 75% of the poses are correctly estimated. Table 7 describes in detail the number of correct detections per detector and the percentage of correctly estimated pose for each pose represented in the database. The pose of a face is correctly estimated if the difference between its angle and the one given by the detector does not exceed 22.5°. As each bottom detector of the detector-pyramid is trained to be robust to pose, the ranges that they cover overlap. Thus, the poses between $\pm60^\circ$ and $\pm30^\circ$ are those where there are the most errors.

(a)

(b)

Figure 31: Some results obtained on (a) Web and (b) Cinema Test Sets.

	Number of correct detections / detector view									Correct
Pose	-90 $^\circ$	-67.5 $^\circ$	-45 $^\circ$	-22.5 $^\circ$	0 $^\circ$	22.5 $^\circ$	45 $^\circ$	67.5 $^\circ$	90 $^\circ$	Pose Estimation
-90 $^\circ$	**10**	-	-	-	-	-	-	-	-	100%
-80 $^\circ$	**9**	**1**	-	-	-	-	-	-	-	100%
-70 $^\circ$	**6**	**2**	1	-	-	-	-	-	-	80%
-60 $^\circ$	5	**3**	**1**	1	-	-	-	-	-	40%
-50 $^\circ$	2	**3**	**4**	-	1	-	-	-	-	70%
-40 $^\circ$	-	1	**4**	**3**	1	-	-	-	-	70%
-30 $^\circ$	-	1	**4**	**1**	4	-	-	-	-	50%
-20 $^\circ$	-	1	1	**1**	**7**	-	-	-	-	80%
-10 $^\circ$	-	-	-	**1**	**9**	-	-	-	-	100%
0 $^\circ$	-	-	1	**1**	**16**	**1**	1	-	-	90%
10 $^\circ$	-	-	-	1	**7**	-	2	-	-	70%
20 $^\circ$	-	-	-	-	**6**	**2**	2	-	-	80%
30 $^\circ$	-	-	-	-	6	-	**4**	-	-	40%
40 $^\circ$	-	-	-	-	2	-	**6**	1	1	60%
50 $^\circ$	-	-	-	-	1	-	**1**	**5**	3	60%
60 $^\circ$	-	-	-	-	-	-	**2**	**4**	4	60%
70 $^\circ$	-	-	-	-	-	-	1	**1**	**8**	90%
80 $^\circ$	-	-	-	-	-	-	1	**1**	**8**	90%
90 $^\circ$	-	-	-	-	-	-	-	-	**9**	90%

Table 7: Out-of-plane pose estimation on Sussex Face Database. The bold numbers correspond to the poses considered as correctly estimated.

(a)

(b)

Figure 32: Example of out-of-plane pose estimation on one individual of the Sussex Database (a) left profile (b) right profile.

6 Conclusion and Future Work

In this report, a novel approach for multiview face detection was proposed. It extends the frontal face detection system developed at the IDIAP Research Institute to deal with faces rotated in the image plane and out of the image plane. After reviewing the main state of the art approaches, we have introduced the baseline of the existing frontal face detection system. Then, the proposed multiview face detection system was presented, inspired by the detector-pyramid architecture introduced by Li and Zhang. The system robustly detects faces rotated in the range of [-67.5°,+67.5°] in-plane and [-90°,+90°] out-of-plane, and estimates their pose. Finally, some experimental results obtained on several benchmark test sets were reported and compared to state-of-the art systems.

The multiview face detector handles most of the faces that one can find in real world images and yields good results on several benchmark test sets. Moreover, it improves the detection rate of the frontal face detector on the CMU Frontal Test Set while conserving a similar number of false alarms. However, the number of false alarms is still very high compared to other state-of-the art systems. To cope with this problem, some work could be done to improve the postprocessing step. Indeed, instead of training the MLPs with the face and nonface images, we could train them directly with the MCT-features. MLPs could also be associated to Principal Component Analysis and/or Linear Discriminant Analysis to be more robust.

The multiview face detector tests 16 poses whereas the frontal face detector tests only 1 pose. However, the multiview face detector is only two times slower than the frontal face detector and not 16 times slower. The real-time property is thus conserved. Nevertheless, speed could be improved to reach the same processing time than the frontal face detector.

The system was implemented using the *Torch3vision* package that can be downloaded from the Web at:
`http://www.idiap.ch/~marcel/en/torch3/torch3vision.php`

Notes

[1] The Littlestone's Winnow update rule is used to update the weights assigned to each feature, only if a mistake is made. If the classifier predicts that the pattern is not a face for a face pattern, then the weights are promoted:

$$\forall f \in F, w_f \leftarrow \alpha . w_f$$

where $\alpha > 1$ is a promotion parameter and F the active feature space. If the classifier predicts that the pattern is a face for a nonface pattern, then the weights are demoted:

$$\forall f \in F, w_f \leftarrow \beta . w_f$$

where $0 < \beta < 1$ is a demotion parameter. All other weights remain unchanged.

[2] Haar basis functions are an example of discrete wavelet transform. They are scale-varying basis functions. The mother scaling function is defined by :

$$\phi(t) = \begin{cases} 1 & \text{if} \quad 0 \leq t < 1 \\ 0 & \text{otherwise} \end{cases}$$

From the mother scaling function, we define a family of shifted and stretched scaling functions $\phi_{k,n}(t)$ according to :

$$\begin{aligned} \phi_{k,n}(t) &= \forall k, n, k \in Z, n \in Z : 2^{-k/2}\phi(2^{-k}t - n) \\ &= 2^{-k/2}\phi\left(\frac{1}{2^k}(t - 2^k n)\right) \end{aligned}$$

[3] The Census Transform was first proposed by Zabih and Woodfill [22]. They give the following definition: "The transform maps the local region surrounding a pixel to a bit string representing which pixel have lesser intensities". The value of the bit corresponding to the center pixel is set to 0. The other bits are set to 1 only if the intensity of the corresponding pixel is higher than the intensity of the center pixel.

[4] Prima Head Pose Database can be downloaded from the Web at:
`http://www-prima.inrialpes.fr/perso/Gourier/Faces/HPDatabase.html`

[5] PICS can be downloaded from the Web at:
`http://pics.psych.stir.ac.uk/`

[6] Yale Face Database can be downloaded from the Web at:
`http://cvc.yale.edu/projects/yalefaces/yalefaces.html`

[7] Essex Face Database can be downloaded from the Web at:
`http://cswww.essex.ac.uk/mv/allfaces/`

[8] CMU Frontal and Rotated Face Set can be downloaded from the Web at:
`http://vasc.ri.cmu.edu/idb/html/face/frontal_images/index.html`

[9] CMU Profile Face Set can be downloaded from the Web at:
`http://vasc.ri.cmu.edu/idb/html/face/profile_images/index.html`

[10] Sussex Database can be downloaded from the Web at:
`http://www.cogs.susx.ac.uk/users/jonh/`

References

[1] Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, *Detecting Faces in Images: A Survey*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 1, January 2002.

[2] Erik Hjelmas, Boon Kee Low, *Face Detection: A Survey*, Computer Vision and Image Understanding, Vol. 83, pp. 236-274, 2001.

[3] Venu Govindaraju, *Locating Human Faces in Photographs*, International Journal of Computer Vision, Vol. 19, No. 2, pp. 129-146, 1996.

[4] Guangzheng Yang, Thomas S. Huang, *Human Face Detection in a Complex Background*, Pattern Recognition, Vol. 27, No. 1, P. 53-56 , 1994.

[5] Ming-Hsuan Yang, Narendra Ahuja, *Detecting Human Faces in Color Images*, Proceedings of the IEEE International Conference on Image Processing, Vol. 1, pp. 127-130, 1998.

[6] Stephen J. McKenna, Shaogang Gong, Heather Liddell, *Real-Time Tracking for an Integrated Face Recognition System*, 2nd Workshop on Parallel Modelling of Neural Operators, Faro, Portugal, November 1995.

[7] Baback Moghaddam, Alex Pentland, *Probabilistic Visual Learning for Object Representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, July 1997.

[8] Kah Kay Sung, Tomaso Poggio, *Example-Based Learning for View-Based Human Face Detection*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 1, pp. 39-51, January 1998.

[9] Henry A. Rowley, Shumeet Baluja, Takeo Kanade, *Neural Network-Based Face Detection*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 1, pp. 23-38, January 1998.

[10] Raphaël Féraud, Olivier J. Bernier, Jean-Emmanuel Viallet, Michel Collobert, *A Fast and Accurate Face Detector Based on Neural Networks*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 1, pp. 42-53, January 2001.

[11] Ming-Hsuan Yang, Dan Roth, Narendra Ahuja, *A SNoW-Based Face Detector*, Advances in Neural Information Processing Systems, S.A. Solla, T.K. Leen and K.-R. Muller (eds), pp. 855-861, MIT Press, 2000.

[12] Paul Viola, Michael J. Jones, *Robust Real-Time Face Detection*, International Journal of Computer Vision, Vol. 57, No. 2, pp. 137-154, 2004.

[13] Yoav Freund, Robert E. Schapire, *A Short Introduction to Boosting*, Journal of Japanese Society for Artificial Intelligence, Vol. 14, No. 5, pp. 771-780, September 1999.

[14] Rainer Lienhart, Jochen Maydt, *An Extended Set of Haar-like Features for Rapid Object Detection*, Proceedings of the IEEE International Conference on Image Processing, pp. 900-903, September 2002.

[15] Stan Z. Li, ZhenQiu Zhang, Heung-Yeung Shum, HongJiang Zhang, *FloatBoost Learning for Classification*, Proceedings of the Neural Information Processing Systems, December 2002.

[16] Henry A. Rowley, Shumeet Baluja, Takeo Kanade, *Rotation Invariant Neural Network-Based Face Detection*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 38-44, 1998.

[17] Michael J. Jones, Paul Viola, *Fast Multi-view Face Detection*, IEEE Conference on Computer Vision and Pattern Recognition, June 2003.

[18] Stan Z. Li, ZhenQiu Zhang, *FloatBoost Learning and Statistical Face Detection*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 9, September 2004.

[19] Christophe Garcia, Manolis Delakis, *Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 11, November 2004.

[20] Henry Schneiderman, Takeo Kanade, *A Statistical Method for 3D Object Detection Applied to Faces and Cars*, IEEE Conference on Computer Vision and Pattern Recognition, June 2000.

[21] Bernhard Fröba, Andreas Ernst, *Face Detection with the Modified Census Transform*, International Conference on Automatic Face and Gesture Recognition, pp. 91-96, May 2004.

[22] Ramin Zabih, John Woodfill, *A Non-Parametric Approach to Visual Correspondance*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996.

[23] Leslie G. Farkas, Tania A. Hreczko, Marko J. Katic, *Craniofacial Norms in North American Caucasians from Birth (One Year) to Young Adulthood*, Anthropometry of the Head and Face, Leslie G. Farkas (ed), 1994.

[24] P. Jonathon Phillips, Hyeonjoon Moon, Syed A.Rizvi, Patrick J. Rauss, *The FERET evaluation methodology for face recognition algorithms*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 10, October 2000.

[25] Terence Sim, Simon Baker, Maan Bsat, *The CMU Pose, Illumination, and Expression Database*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 12, December 2003.

[26] Nicolas Gourier, Daniela Hall, James L. Crowley, *Estimating Face Orientation from Robust Detection of Salient Facial Features*, Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures, Cambridge, UK.

[27] Yann Rodriguez, Fabien Cardinaux, Samy Bengio, Johnny Mariéthoz, *Estimating the Quality of Face Localization for Face Verification*, Proceedings of the IEEE International Conference on Image Processing, Vol. 1, pp. 581-584, October 2004.

[28] Oliver Jesorsky, Klaus J. Kirchberg, Robert W. Frischholz, *Robust Face Detection Using the Hausdorff Distance*, Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication, pp. 90-95, 2001.