Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis

Christophe Garcia and Georgios Tziritas, *Member, IEEE*

***Abstract*— Detecting and recognizing human faces automatically in digital images strongly enhance content-based video indexing systems. In this paper, a novel scheme for human faces detection in color images under nonconstrained scene conditions, such as the presence of a complex background and uncontrolled illumination, is presented. Color clustering and filtering using approximations of the YCbCr and HSV skin color subspaces are applied on the original image, providing quantized skin color regions. A merging stage is then iteratively performed on the set of homogeneous skin color regions in the color quantized image, in order to provide a set of potential face areas. Constraints related to shape and size of faces are applied, and face intensity texture is analyzed by performing a wavelet packet decomposition on each face area candidate in order to detect human faces. The wavelet coefficients of the band filtered images characterize the face texture and a set of simple statistical deviations is extracted in order to form compact and meaningful feature vectors. Then, an efficient and reliable probabilistic metric derived from the Bhattacharrya distance is used in order to classify the extracted feature vectors into face or nonface areas, using some prototype face area vectors, acquired in a previous training stage.**

***Index Terms*— Bhattacharrya distance, color clustering, face detection, wavelet decomposition.**

I. Introduction

DETECTING human faces automatically is becoming a very important task in many applications, such as security access control systems or content-based indexing video retrieval systems like the Distributed audioVisual Archives Network (DiVAN) system [11]. The European Esprit project DiVAN aims at building and evaluating a distributed audio-visual archives network providing a community of users with facilities to store raw video material, and access it in a coherent way, on top of high-speed wide area communication networks. The raw video data is first automatically segmented into shots using techniques based on color histograms dissimilarity and camera motion analysis. From the content-related image segments and keyframes, salient features such as region shape, intensity, color, texture, and motion descriptors are extracted and used for indexing and retrieving information. In order to allow queries at a higher semantic level, some particular pictorial objects may be detected and exploited for indexing. The automatic detection of human faces provides users with powerful indexing capacities of the video material. In DiVAN, faces are detected in the extracted keyframes and stored in a meta-database. Without performing face recognition, frames containing faces may be searched according to the number, the sizes, or the positions of the detected faces within the keyframes, looking for specific classes of scenes representing a large audience (multiple faces), an interview (two medium size faces) or a close-up view of a speaker (a large size face). Face recognition may follow face detection when faces are large enough and in a semi-frontal position, using a method we developed and described in [16]. When detected faces are recognized and associated automatically with textual information like in the systems Name-it [33] or Piction [6], potential applications such as news video viewer providing description of the displayed faces, news text browser giving facial information, or automated video annotation generators for faces are possible.

Although face detection is closely related to face recognition as a preliminary required step, face recognition algorithms have received most of the attention in the academic literature compared to face detection algorithms. Considerable progress has been made on the problem of face recognition, especially under stable conditions such as small variations in lighting, facial expression and pose. Extensive surveys are presented in [42] and [5]. These methods can be roughly divided into two different groups: geometrical features matching and template matching. In the first case, some geometrical measures about distinctive facial features such as eyes, mouth, nose and chin are extracted [3], [8]. In the second case, the face image, represented as a two-dimensional (2-D) array of intensity values, is compared to a single or several templates representing a whole face. The earliest methods for template matching are correlation-based, thus computationally very expensive and require great amount of storage. In the last decade, the principal components analysis (PCA) method also known as Karhunen–Loeve transform, has been successfully applied in order to perform dimensionality reduction [22], [39], [29], [37], [1]. We may cite other methods using neural network classification [30], [9], using algebraic moments [18], using isodensity lines [28], or using a deformable model of templates [23], [43].

In most of these face recognition approaches, existence and location of human faces in the processed images are known *a priori*, so there is little need to detect and locate faces. In image and video databases, there is generally no constraint on the number, location, size, and orientation of human faces and the background is generally complex. Moreover, color information

Manuscript received April 16, 1999; revised July 13, 1999. The associate editor coordinating the review of this paper and approving it for publication was Prof. Alberto DelBimbo. This work was supported in part by the DiVAN Esprit Project EP 24956.

The authors are with the Institute of Computer Science, Foundation for Research and Technology-Hellas, GR 711 10 Heraklion, Crete, Greece (e-mail: cgarcia@csi.forth.gr; tziritas@csi.forth.gr).

Publisher Item Identifier S 1520-9210(99)07634-8.

is very useful for face detection, whereas this information is not used in face recognition approaches. Over the last five years, increasing activity has been noticed in developing algorithms to either detect or locate faces in "mugshot" style images, or detect faces in uncontrolled images [42], [31]. Existing methods may be roughly divided into three broad categories: local facial features detection, template matching and image invariants. In the first case, low-level computer vision algorithms are applied in order to detect facial features such as eyes, mouth, nose and chin. Then, statistical models of human face are used like in [4], [12], [44], [19], [45]. In the second case, several correlation templates are used to detect local subfeatures which can be considered as rigid in appearance (view-based eigenspaces [29]) or deformable (deformable templates [43]). In the third case, image-invariants schemes assume that there are certain spatial image relationships common and possibly unique to all face patterns, even under different imaging conditions [34]. Instead of detecting faces by following a set of human-designed rules, alternative approaches are based on neural networks [24], [32], [36] which have the advantage of learning the underlying rules from a given collection of representative examples of face images, but have the major drawback of being computationally expensive and challenging to train because of the difficulty in characterizing "nonface" representative images.

In this paper, we propose a novel algorithm for automatically detecting human faces in digital still color images, under nonconstrained scene conditions, such as presence of a complex background and uncontrolled illumination, where most of the local facial features based method are not stable. As a preliminary work, we presented, in [17], a face detector which had been developed in order to index a huge amount of video and images data and to cope with high-speed requirements. Only I frames (Intraframe transform coding) were analyzed from MPEG streams as we wanted to avoid costly decompression. Color segmentation of these I frames was performed at MPEG macro-block level (16×16 pixels). Skin color filtering was performed, providing a macro-block binary mask which was segmented into nonoverlapping rectangular regions containing contiguous regions of skin color macro-blocks (binary mask segment areas). Then, the algorithm searched for the largest possible candidate face areas and iteratively reduced their size in order to scan all the possible aspects ratios and positions into each binary mask segment area. In this paper, the proposed scheme has been substantially enhanced. It performs first color clustering of the original image, in order to extract a set of dominant colors and quantize the image according to this reduced set of colors. Then, a chrominance-based segmentation using an improvement of the method presented in [17] is performed. A merging stage is iteratively applied on the set of homogeneous skin color regions in the color quantized image, in order to provide a set of candidate face areas, without scanning all the different possible aspects ratios and the possible positions into binary segment areas. This improvement leads to a better precision in locating the faces and helps in segmenting the faces from the background, especially when parts of the surrounding background have a color that may be classified as skin color. Constraints related to shape and size of faces are applied, and face intensity texture is analyzed by performing a wavelet packet decomposition on each face area candidate in order to detect human faces. Each face area candidate is described by band filtered images containing wavelet coefficients. A set of simple statistical data is extracted from these coefficients, in order to form vectors of face descriptors, and a well-suited probabilistic metric derived from the Bhattacharrya distance is used to classify the feature vectors into face or nonface areas, using some prototype face area vectors, which have been built in a training stage.

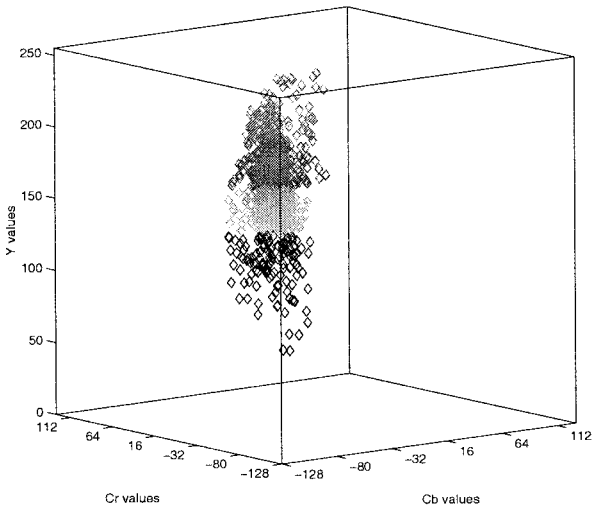
The remainder of this paper is organized as follows. In Section II, our method for segmenting skin color regions is presented. In Section III, we present the algorithm of detection of the candidate face regions, which performs an iterative merging of skin color regions and applies constraints related to human faces shape. In Section IV, we present the core component of our face detection scheme, by describing the extraction of face texture descriptors vectors using wavelet packet analysis and their classification using a distance derived from the Bhattacharrya dissimilarity measure. Experimental results and a comparison of the proposed scheme with the CMU face detector [32] are provided in Section V. Finally conclusions are drawn.

II. Detection of Skin Color Regions

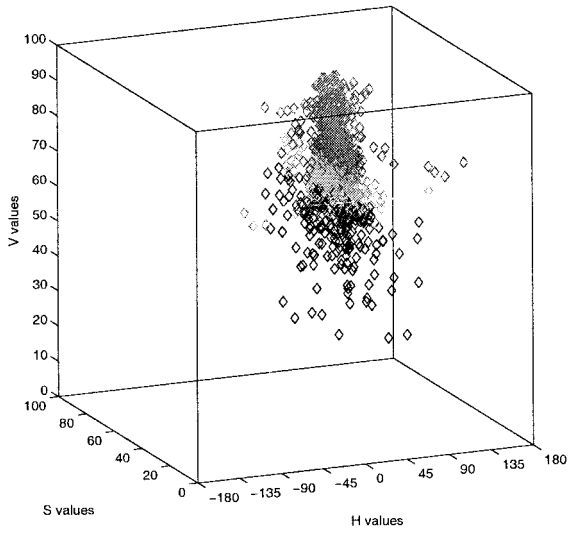
The first stage of the proposed scheme consists in locating the potential face areas in the image, using skin chrominance information, given that such an information strongly reduces the search space. Before proceeding with skin color classification, a color quantization of the original image is performed, in order to improve skin color segmentation by homogenizing the image regions.

A. Color Quantization of the Original Image

Vector quantization is applied to quantize the image colors to a reduced set of so-called dominant colors. The basic principle is to map a color vector x onto another color vector y_i, y_i being a *codebook* vector and representing a color cluster C_i in color space. y_i is a dominant color and is the mean value of color vectors belonging to C_i. The initial codebook vectors y_i are first determined by color histogram calculation. A complete three-dimensional (3-D) histogram is computed from the original color image, in hue-saturation-value (HSV) color space. Histograms bins are hierarchically merged according to the Euclidian distance denoted by $\mathcal{D}c$ in $(SV\cos(H), SV\sin(H), V)$, treating the HSV color space as a cone, like in [2], in order to build an initial set of color cluster C_i. In this implementation, a fixed number of 16 color clusters are considered. The final set of codebook vectors is obtained using an iterative algorithm that originates from pattern recognition, the K-means algorithm, also known as the Linde–Buzo–Gray algorithm (LBG) [25], which has been first applied to vector quantizer design for signal compression tasks. At each iteration, each color vector (pixel) is assigned to the closest color cluster according to its distance $\mathcal{D}c$ to the mean color vector of the cluster, and the mean color vector

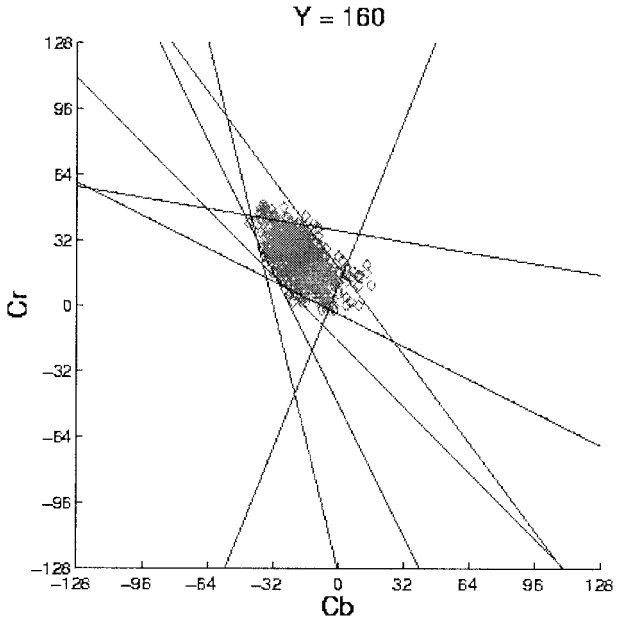


(a)

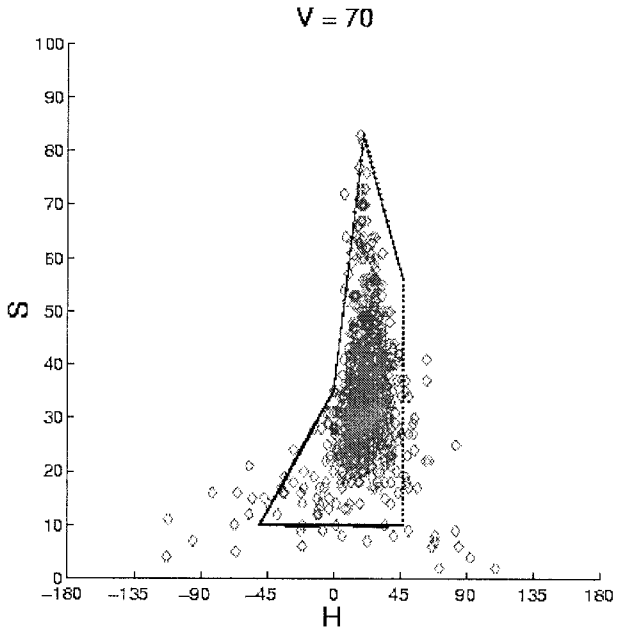


(b)

Fig. 1. Sample skin colors distribution in YCbCr color space (a) and HSV color space (b).



(a)



(b)

Fig. 2. Skin color bounded areas.

of the set of pixels which have been associated to one color cluster is updated. The algorithm stops after a given number of iterations. Then, color clusters are merged according to a predefined maximum distance. Finally, each pixel in the image receives the value of the mean color vector of the cluster it has been assigned to. The resulting image is therefore quantized according to the set of dominant colors.

B. Skin Color Segmentation

Our purpose is to segment the quantized color image according to skin color characteristics. Two color models have been evaluated and used. The YCbCr model is naturally related to MPEG and JPEG coding. The HSV (Hue, Saturation, Value) model is used mainly in computer graphics and is considered by many to be more intuitive to use, closer to how an artist actually mixes colors. Skin color patches had been used in order to approximate the skin color subspaces, in both models. The training data set is composed of 950 skin color samples which have been extracted from various still images and video frames, covering a large range of skin color appearance (different races, different lighting conditions). In Fig. 1(a) and (b), skin color samples are plotted in YCbCr space and HSV, respectively. One can observe that skin color samples form a single and quite compact cluster in both YCbCr and HSV color spaces. Our purpose is to approximate the 3-D envelop of these clusters.

In the case of the YCbCr color space, we noticed that the intensity value Y has little influence on the distribution in the CbCr plane and that sample skin colors form a small and very compact cluster in the CbCr plane. Wang and Chang in [40] performed skin colors classification directly in the

chrominance plane (CbCr) without taking the intensity value Y into account. A Bayesian decision rule for minimum cost is applied in order to classify a color into the skin color or the nonskin color class. 40 samples of face patches as well as false alarms results are used in order to apply this method. We found it was difficult to use the false alarms results, because there are usually quite a lot of image areas with colors similar to the skin colors, like over part of the human body, natural background, etc. We decided to estimate the envelop of the skin color subspace in YCbCr, using the intensity value Y in order to cope with strong lighting variations, as we noticed that the distribution turns out to be different for the extrema of Y values corresponding to dark (for values of Y around 50) and light (for values of Y around 240) lighting conditions. Given that quite a large amount of noise is contained in the skin color samples due to dark areas of the face and facial hair, we decided it was more efficient to approximate the skin color subspace borders using planar approximations that can be easily adjusted rather than using automatic clustering methods. Sets of planes have been found by successive adjustments according to the segmentation results.

In Fig. 2(a), the intersections of the adjusted bounding planes with the CbCr plane for $Y = 160$ are displayed. One may notice that some sample points are not contained in the skin color subspaces. These points have not been included during the successive adjustments given that they correspond to nonrepresentative samples of the skin colors and cause errors in the face detection process.

We report hereafter the equations defining the bounding planes that have been found. As one can notice, there are two sets of eight equations depending on two areas of the color space, separated by the horizontal plane $Y = 128$, in order to approximate the distribution borders in the light and dark extreme cases:

$$\begin{aligned}
&\text{if } (Y>128) \quad \theta_1 = -2 + \frac{256-Y}{16};\\
&\qquad\qquad\quad \theta_2 = 20 - \frac{256-Y}{16};\\
&\qquad\qquad\quad \theta_3 = 6; \quad \theta_4 = -8\\
&\text{if } (Y \le 128) \quad \theta_1 = 6; \quad \theta_2 = 12; \quad \theta_3 = 2 + \frac{Y}{32};\\
&\qquad\qquad\quad \theta_4 = -16 + \frac{Y}{16}\\
&Cr \ge -2(Cb+24); \quad Cr \ge -(Cb+17);\\
&Cr \ge -4(Cb+32); \quad Cr \ge 2.5(Cb+\theta_1);\\
&Cr \ge \theta_3; \quad Cr \ge 0.5(\theta_4 - Cb);\\
&Cr \le \frac{220-Cb}{6}; \quad Cr \le \frac{4}{3}(\theta_2 - Cb).
\end{aligned}$$

We noticed that the cluster of skin color is less compact in HSV space than in YCbCr. The projection onto the HS plane only is used by some authors like in [38] where skin color classification is performed by setting appropriate thresholds to Hue and Saturation. Using these thresholds, we found that the segmentation results are affected by variations in lighting conditions.

Like in the YCbCr case, we directly estimated the shape of the skin color subspace in HSV. A set of planes have been found by successive adjustments according to segmentation results. In Fig. 2(b), the intersections of the adjusted bounding planes with the HS plane for $V = 70$ are drawn. We report hereafter the equations defining the six bounding planes that have been found in the HSV color space case.

$$\begin{aligned}
&S \ge 10; \quad V \ge 40; \quad S \le -H - 0.1V + 110;\\
&H \le -0.4V + 75;\\
&\text{if } H \ge 0\ S \le 0.08(100 - V)H + 0.5V\\
&\text{else } S \le 0.5H + 35
\end{aligned}$$

Extensive experiments have shown that segmentation results are quite equivalent using both color models. However, we noticed that even if the skin color cluster is more compact in the CbCr plane than in the HS plane, the bounding planes are more easily adjusted using the HSV model, because of a direct access to H (Hue) which mainly encodes skin colors.

In Fig. 3, four examples of skin color-based image segmentation are shown. The first line displays the original images. The second line shows the color quantized image obtained after dominant colors detection. The third line displays the skin color areas of the quantized color image, using the skin color subspaces approximation. The first and the second examples demonstrate the efficiency of the skin color-based segmentation process. The two last examples show that parts of the background may have a color similar to skin color. It is obvious that chrominance information on its own is not sufficient and therefore that another face detection stage based on shape and texture analysis is required.

III. Detection of Candidate Face Regions

In this section, we will discuss how to locate candidate face areas in the skin color filtered image, denoted as SCF image. Skin color areas have to be segmented in order to form potential face areas which will be classified in the latest stage of the proposed scheme, by performing wavelet packet analysis.

In our previous work [17], skin-color filtering was applied on I frames at MPEG macro-block level (16×16 pixels), providing a macro-block binary mask which was segmented into nonoverlapping rectangular regions containing contiguous regions of skin color macro-blocks (binary mask segment areas). Then, the algorithm had to search for the largest possible candidate face areas and iteratively reduced their size in order to scan all the possible aspects ratios and positions into each binary mask segment area.

In the scheme presented in this paper, a merging stage is iteratively applied on the set of homogeneous skin color regions in the color quantized image, in order to provide a set of candidate face areas, without scanning all the possible aspects ratios and the possible positions into binary segment areas. This improvement leads to a better precision in locating the faces and helps to segment the faces from the background, especially when parts of the surrounding background have a color close to skin color.

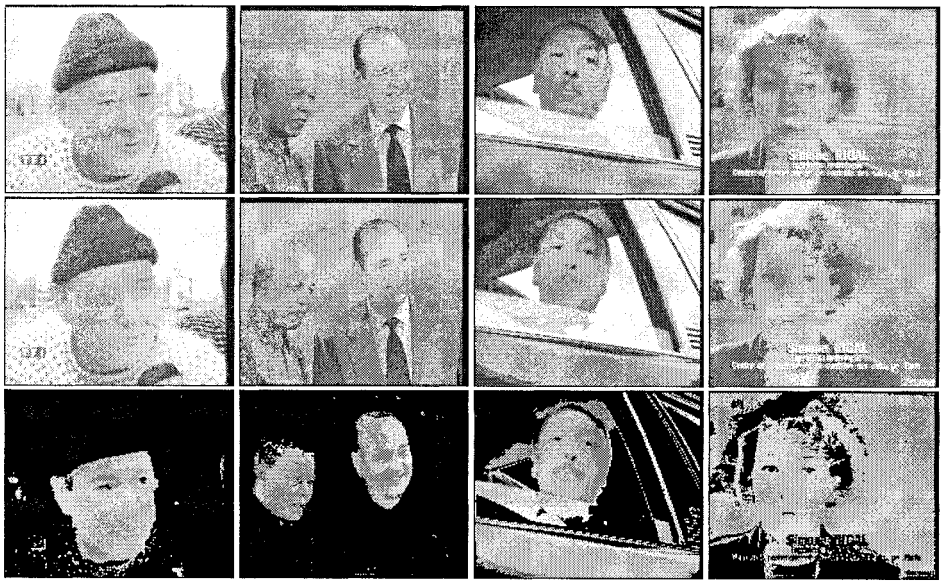


Fig. 3. Original images (first line), color quantized images (second line), and skin color filtered images (third line).

The SCF image is composed of a relatively small set of homogeneous regions, denoted as $\mathcal{R}_i$, due to color quantization. We aim at building potential candidate face areas by iteratively merging adjacent homogeneous skin color regions. Therefore, we do not treat the skin color areas as a whole binary mask that we have to divide *a posteriori*. By merging adjacent similar and homogeneous skin color regions, we are able to iteratively construct candidate face areas while distinguishing between the different skin tones in the skin color areas. This approach allows to segment faces from a surrounding skin color background.

The skin color regions merging algorithm starts by computing a region adjacency graph, where each node represents a homogeneous skin color region of the quantized image. The criterion of connectivity is based on a minimum distance computed between the associated bounding boxes of each homogeneous region. Evaluating connectivity using a distance between the vertices of the bounding boxes is much faster than estimating the connectivity in the pixel domain.

Let $\mathcal{C}_0$ be the set of homogeneous skin color regions $\mathcal{R}_i$, $i \in [1 \cdots N]$. First, the criterion of connectivity is applied to build the region adjacency graph. Given that we are dealing with bounding boxes, a fast algorithm has been implemented, which considers the normal distance between the vertices of the disconnected bounding boxes or determines if the bounding boxes are overlapping. Therefore, adjacent regions are defined. The merging criterion among the bounding boxes of adjacent regions $\mathcal{R}_i$ and $\mathcal{R}_j$ is based on a maximum allowed distance $\mathcal{D}r(\mathcal{R}_i, \mathcal{R}_j)$, which encodes color dissimilarity in the bounding boxes. We noticed that the different color regions over a given face are stable according to H (Hue) and that color differences appear mainly on the V and S components due to variations of the light incidence over the face surface. For that reason, the color dissimilarity measure has been chosen to be the following:

$$\mathcal{D}r(\mathcal{R}_i, \mathcal{R}_j) = \alpha|H_i - H_j| + \beta|S_i - S_j| + \gamma|V_i - V_j| \quad (1)$$

where (H_i, S_i, V_i) is the average skin color vector in the bounding box of region $\mathcal{R}_i$. The value of α has to be chosen much bigger than the values of β and γ in order to take into account mainly Hue differences.

The set of potential face areas is built iteratively, starting from the $\mathcal{C}_0$ set and its adjacency graph. The set $\mathcal{C}_1$ is obtained by merging the compatible adjacent regions of $\mathcal{C}_0$. Then, each new set of merged region $\mathcal{C}_k$ $(k \in [1 \cdots K])$ is obtained by merging iteratively the set of merged region $\mathcal{C}_{k-1}$ with the original set $\mathcal{C}_0$ in the following way (note that a region adjacency graph is computed before each new iteration):

$$\forall k \in [1 \cdots K], \quad \forall \mathcal{R}_i \in \mathcal{C}_0, \quad \forall \mathcal{R}_j \in \mathcal{C}_{k-1}$$
$$\mathcal{C}_k = \bigcup \mathcal{R}_i \cup \mathcal{R}_j \quad (2)$$

if $\mathcal{D}r(\mathcal{R}_i, \mathcal{R}_j)$ is allowed and $\{\mathcal{R}_i, \mathcal{R}_j\}$ are adjacent regions. Note that the doublets, which may appear while combining the merged regions are discarded by a simple test on insertion.

In the current implementation, iterative merging is performed with a maximum number of iterations $K = 3$, as we want to avoid time-consuming combinatory. We also noticed that the merging process provides the expected candidate face areas within these three iterations. Finally, the candidate face area set $\mathcal{CF}$ contains the original regions set $\mathcal{C}_0$ and the interatively built sets of merged region $\mathcal{C}_k$ $(k \in [1 \cdots K])$, i.e., $\mathcal{CF} = \bigcup_{k=0}^{K} \mathcal{C}_k$.

The merging process may be more easily described in the graphical example of Fig. 4. In this simple example, five different skin color homogeneous regions have been labeled with letters ranging from A to E. In that case, the initial partition is $\mathcal{C}_0 = \{A, B, C, D, E\}$. Let suppose that all regions, except region E, have a similar color, i.e., the distance $\mathcal{D}_r$ is allowed. Given that E is not color-compatible, some of the potential merged sets will not be built. According to the connectivity and merging criteria, the first interation of the iterative merging process will produce the set $\mathcal{C}_1 = \{AB, AD, AC, BC, CD\}$, where XY stands for "minimal

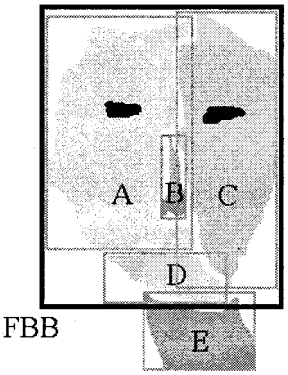


Fig. 4. Merging process of homogeneous skin color regions.

bounding box containing regions X and Y". Note that different cases appear when this merged set is built. DE and CE have not been built because although D and E as well as C and E are adjacent regions, their colors are incompatible. The bounding box of B being included in the bounding box of A, AB turns to be the bounding box of A. AC is built because the bounding boxes of A and C overlap, which is also the case of BC and CD. AD is built because the distance computed between the associated bounding boxes of regions A and D is allowed. The second and third iterations leads to $\mathcal{C}_2 = \{ABC, ABD, ACD, BCD\}$ and $\mathcal{C}_3 = \{ABCD\}$. Finally, the candidate face area set is: $\mathcal{CF} = \bigcup_{k=0}^{K} \mathcal{C}_k = \{A, B, C, D, E, AD, AC, BC, CD, ABC, ABD, BCD, ABCD\}$. Each of these bounding boxes will be processed in the last step of our scheme, for face texture classification. In this simple example, the expected result of face detection would be obtained for the candidate face areas $ABCD$, drawn as FBB (Face Bounding Box) in the figure. Note that ACD is equivalent to $ABCD$, given that B is included in the bounding box of A.

In Fig. 5, the merging process of the skin color areas is presented in two real examples. For each example, the four first images contain the homogeneous regions corresponding to four dominant skin colors. The regions corresponding to the considered skin color are displayed in white. For reasons of readability, the bounding boxes have not been drawn around each skin color (white) area. In the fifth image, the bounding boxes of the candidate face areas built by the merging process and contained in the $\mathcal{CF}$ are drawn. The final detected faces are shown in the last image. As one can notice in the fifth image, some bounding boxes in $\mathcal{CF}$ are small or have an aspect ratio which is not compatible with the one of a human face. Therefore, constraints related to shape, acceptable size range, color homogeneity and texture are then applied to the candidate face areas of $\mathcal{CF}$ in order to classify these regions into faces or nonfaces. Face texture analysis will be described in the next section.

First, constraints relative to shape and size are applied in order to remove most of the non potential candidate face areas. The shape of human faces can be approximated and can be considered as a discriminant information. In order to perform fast computation, we analyze the face shape by considering the aspect ratio of its bounding box. The allowed aspect ratio range has been set to [0.9, 2.1], which was chosen quite large in order to cope with different orientations and poses. Moreover, constraints related to allowed sizes are applied. The range of allowed size has been lower-bounded by 80×48 pixels. It is naturally upper-bounded by the size of the whole frame image. By applying these two constraints, an important number of regions are removed from the set of candidate face areas $\mathcal{CF}$.

Then, a constraint of color homogeneity is applied by computing the density of skin color pixels in each candidate face area $\mathcal{R}_i$. Two areas, respectively the *outer* and the *inner* areas, are defined into each $\mathcal{R}_i$, the *outer* area corresponding to the border area of $\mathcal{R}_i$ (whose width is a percentage of the bounding box width, typically 15%) and the *inner* area corresponding to the remaining central part of $\mathcal{R}_i$. A $\mathcal{R}_i$ area is accepted as a potential face area, if the density of skin color pixels in the whole $\mathcal{R}_i$ area, denoted as p is above a threshold set to 0.4 and the density in the *inner* area, denoted as p_{in}, is above a threshold defined as $0.7p + 0.3$. Homogeneity in the *inner* area has to be higher than in the whole $\mathcal{R}_i$ area given that the *inner* area is the central part of the face with no background (especially in the case of small images), whereas in the whole $\mathcal{R}_i$ area, hair and parts of background have to be taken into consideration.

IV. Classification of Candidate Face Areas Using Wavelet Packet Analysis

The main purpose of the last stage of the proposed algorithm is to classify the $\mathcal{R}_i$ areas, which are compatible with the constraints of shape, size and color homogeneity, into faces or nonfaces, according to a textural analysis. The classification aims at removing false alarms caused by objects with color similar to skin color and similar aspect ratios, such as other exposed parts of the body or part of background. For this purpose a wavelet packet analysis scheme is applied using a method which is similar to the one we proposed for face recognition in [16]. A wavelet packet decomposition is performed on the intensity plane of the image and feature vectors are extracted from a set of wavelet packet coefficients in each $\mathcal{R}_i$ area. The intensity image is the Y plane of the original color image, which has been transformed from HSV to YCbCr color space. This intensity image conveys information about the texture of the faces, which will be retained into the wavelet coefficients.

Wavelet packet decomposition for analyzing the face texture has been found suitable given that this scheme provides a multiscale analysis of the image in the form of coefficient matrices with a spatial and a frequential decomposition of the image at the same time. Strong arguments for the use of multiscale decomposition can be found in psychophysical research, which offers evidence that the human visual system processes the images in a multiscale way.

In the last decade, wavelets have become very popular, and new interest is rising on this topic. The main reason is that a complete framework has been recently built [26], [10] in particular for what concerns the construction of wavelet bases and efficient algorithms for the wavelet transform computation.

An interesting characteristic of wavelets is flexibility: several bases exist, and one can choose the basis that is more suitable for a given application. In [7], an entropy-based

Fig. 5. Homogeneous skin color regions, candidate face bounding boxes, detected face.

functional is used for the selection of the best bases, in a general case. We think that this is still an open problem, in particular for face images. Experimental considerations may rule the choice of a wavelet form.

Computational complexity of wavelets is linear with the number (N) of computed coefficients $(O(N))$, while other transformations, also in their fast implementation, lead to $N \times \log(N)$ complexity. Thus, wavelets are adapted also for dedicated hardware design (discrete wavelet transform). The possibility of easily embedding part of the process in hardware allows real time computation, like in compression tasks [13]. Moreover, a wavelet compression algorithm (increasing compression efficiency by 30%) will be included in the future JPEG2000 coding model, after the results observed during the 12th WG1 meeting in Sydney in November 1997. The resulting wavelet coefficients may be used directly in our scheme, while processing JPEG2000 images.

A. Wavelet Packet Decomposition of Face Images

The discrete wavelet series for a continuous signal $s(t)$ is defined by the following equation:

$$c_{n,k} = \frac{1}{2^{n/2}} \int s(t)\psi^*(2^{-n}t - k)\, dt. \tag{3}$$

The series use a dyadic scale factor, n being the scale level and k being the localization parameter. The function $\psi(t)$ is the mother wavelet which satisfies some admissibility criteria and ensures a complete, nonredundant, and orthogonal representation of the signal. The discrete wavelet transform (DWT) results from the above series, and is equivalent to the successive decomposition of the signal by a pair of filters $h(\cdot)$ and $g(\cdot)$, as shown in Fig. 6. The low-pass filter $h(\cdot)$ provides the approximation of the signal at coarser resolutions, while the high-pass filter $g(\cdot)$ provides the details of the signal at coarser resolutions.

The extension to the 2-D case is usually performed by applying these two filters separately in the two directions. The convolution with the low-pass filter results in a so-called *approximation* image and the convolutions with the high-pass filter in specific directions result in so-called *detail* images.

In classical wavelet decomposition, the image is split into an approximation and details. The approximation is then split into a second-level of approximation and details. For a n-level

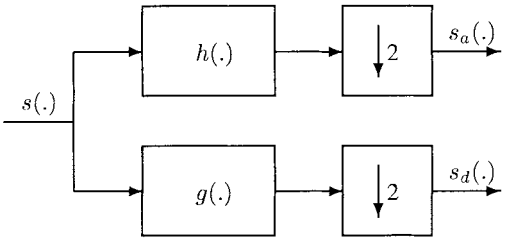


Fig. 6. Dyadic decomposition of a signal.

decomposition, the signal is decomposed in the following way:

$$\begin{aligned} A_n &= [h_x * [h_y * A_{n-1}]_{\downarrow 2,1}]_{\downarrow 1,2} \\ D_{n1} &= [h_x * [g_y * A_{n-1}]_{\downarrow 2,1}]_{\downarrow 1,2} \\ D_{n2} &= [g_x * [h_y * A_{n-1}]_{\downarrow 2,1}]_{\downarrow 1,2} \\ D_{n3} &= [g_x * [g_y * A_{n-1}]_{\downarrow 2,1}]_{\downarrow 1,2} \end{aligned} \tag{4}$$

where $*$ denotes the convolution operator, $\downarrow 2,1$ $(\downarrow 1,2)$ subsampling along the rows (columns), and $A_0 = I(x,y)$ is the original image. A_n is obtained by low-pass filtering and is the approximation image at scale n. The detail images D_{ni} are obtained by bandpass filtering in a specific direction $(i = 1,2,3,$ for vertical, horizontal and diagonal directions respectively) and thus contain directional details at scale n. The original image I is thus represented by a set of subimages at several scales: $\{A_n, D_{ni}\}$.

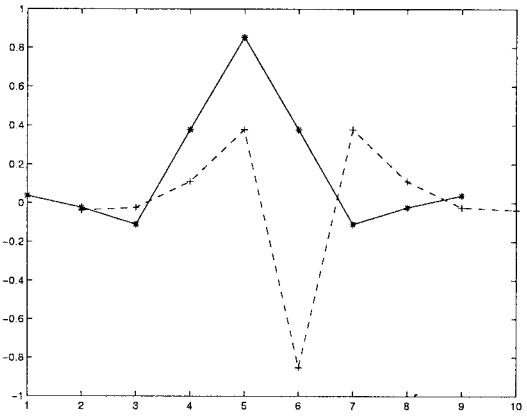
Fig. 7(a) shows the h and g filters that have been applied in the proposed scheme. As stated before, the choice of the most suitable h and g filters for a given application is mainly ruled by experimental results. The constraints that have been considered in order to build these filters are closely related to the work of Smith and Barnwell concerning exact reconstruction techniques for tree structured subband coders [35]. According to this work and [41], h and g are built as conjugate quadrature filters. The coefficients values and the filter support size have been chosen during series of experiments.

The Z-transform of the pair of filters is

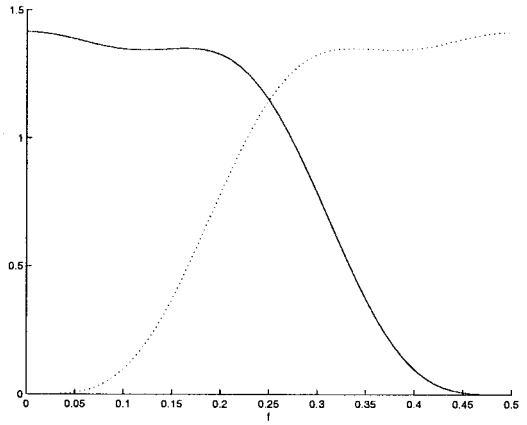
$$\begin{aligned} H(z) = {} & 0.853 + 0.377(z + z^{-1}) - 0.111(z^2 + z^{-2}) \\ & - 0.024(z^3 + z^{-3}) + 0.038(z^4 + z^{-4}) \end{aligned} \tag{5}$$

$$G(z) = -z^{-1}H(-z^{-1}). \tag{6}$$

The Fourier transform of these filters is given in Fig. 7(b). The low-pass filter is symmetric and the filter pair is orthogonal,



(a)



(b)

Fig. 7. (a) h (solid line) and g (dashed line) filters and (b) their Fourier transforms.

that is

$$\sum_n h(n-2k)g(n-2l) = 0, \quad \forall (k,l). \tag{7}$$

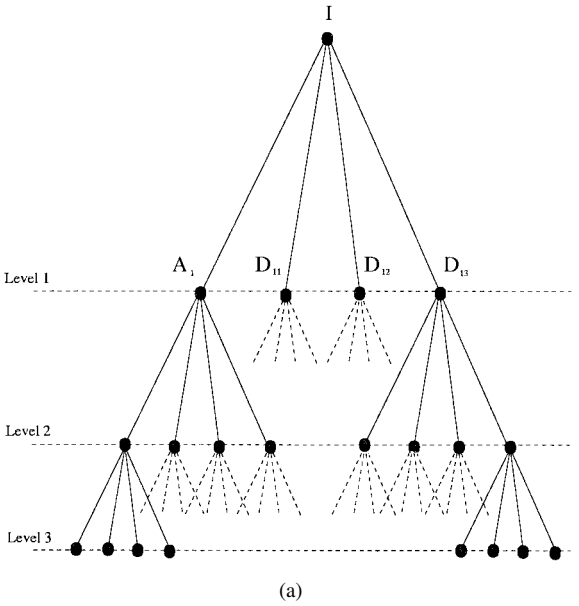
Since we are interested in having decorrelated filter responses, we consider two measures of this property. The antialiasing coefficient is defined and calculated as

$$\eta = \frac{\int_0^{1/4} |\mathcal{H}(f)|^2 \, df}{\int_0^{1/2} |\mathcal{H}(f)|^2 \, df} = 0.87 \tag{8}$$

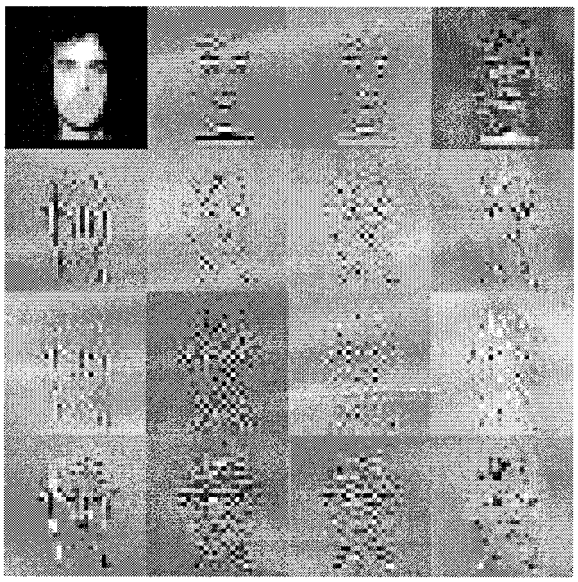
where $\mathcal{H}(f)$ is the Fourier transform of $h(n)$ and the ideal value being equal to 1. We also define a correlation coefficient between the approximation and the detail signals as follows:

$$\rho = \frac{\sum_{k,l} h(k)g(l)\gamma_s(k-lt)}{\sqrt{\sum_{k,l} h(k)h(l)\gamma_s(k-l)}\sqrt{\sum_{k,l} g(k)g(l)\gamma_s(k-l)}} \tag{9}$$

where $\gamma_s(\cdot)$ is the autocovariance function of the input signal. For a noncorrelated (white noise) or fully correlated input signal, it is obvious that ρ is zero valued. We calculated ρ under the assumption of a first-order Markov process for $s(t)$ and we found that the maximum value of ρ is 0.06. It is sufficiently small for considering that the hypothesis of interband decorrelation is valid in practice.



(a)



(b)

Fig. 8. (a) Wavelet packet tree and (b) Level 2 coefficients of the wavelet packet tree.

The *wavelet packet decomposition*, that is performed in the proposed approach, is a generalization of the classical wavelet decomposition that offers a richer signal analysis (discontinuity in higher derivatives, self-similarity, etc.). In that case, the details as well as the approximations can be split. This results in a wavelet decomposition tree as shown in Fig. 8(a), with the correspondent wavelet coefficients of level 2 displayed in Fig. 8(b).

Usually, an entropy-based criterion is used to select the deepest level of the tree, while keeping the most meaningful information. In the present case, the depth of wavelet packet decomposition tree is ruled by the size of the $\mathcal{R}_i$ areas

which are processed. A database of 50 manually extracted face areas which contains a large number of different cases (size, chrominance, intensity, position and orientation) has been built. The processed frames have a size of 288 × 352 pixels. These samples have been classified into two categories, according to their respective sizes. They are used in order to estimate prototype vectors. A face is classified as *medium*, if its height is smaller than 128, and as *large*, if its height is bigger than 128. Candidate faces intensity images are decomposed with the discrete wavelet packet analysis until level 3 for *large* areas and until level 2 for *medium* areas. In both cases, a deeper decomposition will not provide more valuable information, the coefficient images becoming too small. According to its size, a candidate face image is described by a set of n wavelet coefficient matrices belonging to the deepest level of decomposition, ($n = 16$, i.e., one approximation image and 15 detail images for *medium* areas and $n = 64$, i.e., one approximation image and 63 detail images for *large* areas) which, in all cases, represent quite a huge amount of information (equal to the size of the input image). Dimensionality reduction has to be performed by extracting discriminatory information from these images.

In our approach for face recognition [16], we reported good results, obtained by extracting statistical information from the wavelet coefficients in some different specific areas of the approximation image of the face, and statistical information from the wavelet coefficients of each whole detail image. We first located the face bounding box and then we divided it into two areas, the top part and the bottom part, separated by the nose baseline. Like in [17], we follow a slightly different approach by considering each $\mathcal{R}_i$ area as a bounding box, and by dividing it into four parts: a left top part (top_1), a right top part (top_2), a left bottom part ($bottom_1$) and a right bottom part ($bottom_2$), all of equal size, where the wavelet packet analysis is performed. By extracting moments from wavelet coefficients in these areas, we obtain information about the face texture, related to different facial parts, like the eyes, the nose and the mouth, including facial hair. Standard deviations have been chosen as face texture descriptors coefficients. Therefore, from the top and bottom areas, we extract the corresponding standard deviations σ_{top1}, σ_{top2}, $\sigma_{bottom1}$, and $\sigma_{bottom2}$ of the wavelet coefficients contained in the approximation image of the selected level of decomposition. From the m detail images ($m \in \{15, 63\}$), the corresponding standard deviations σ_i ($i = 4, \cdots, n$; $n = m + 3$) are extracted from the whole $\mathcal{R}_i$ area.

Thus, extracted feature vectors contain a maximum of $4+m$ components (four standard deviations for the approximation image and m standard deviations for the detail images) and are described as follows: $\mathcal{V} = \bigcup_{i=0}^{m+3} \{\sigma_i\}$ where indices $i = 0, 1, 2, 3$ stand respectively for the top_1, top_2, $bottom_1$, and $bottom_2$ standard deviations. Finally, a set of vectors of size 19 and 67 for *medium* and *large* $\mathcal{R}_i$ areas, respectively, is built.

B. Feature Vectors Classification

In this section, we will describe how a $\mathcal{R}_i$ feature vector is classified into the two possible classes: face or nonface. From the database of manually extracted face areas, that we previously classified into two size categories, features vectors are extracted and an average prototype feature vector has been retained for each of these two categories.

When solving a pattern recognition problem, the ultimate objective is to design a recognition system which will classify unknown patterns with the lowest possible probability of misrecognition. In the feature space defined by a set of features $X = [x_1, \cdots, x_n]$ which may belong to one of the possible m pattern classes $\omega_i, i = 1, \cdots, m$, an error probability can be defined. It is easy to show that, in the two-classes case, the error probability e can be written

$$e = \frac{1}{2}\left[1 - \int |p(X|\omega_1)P(\omega_1) - p(X|\omega_2)P(\omega_2)|\, dX\right]. \tag{10}$$

As discussed in ([14]), e can not be easily evaluated and a number of alternative feature evaluation criteria have been suggested in the literature. One of these criteria is based on probabilistic distance measures. According to (10), the error e will be maximum when the integrand is zero, that is, when density functions are completely overlapping, and it will be zero when they do not overlap. The integral in (10) can be considered as the probabilistic distance between the two density functions.

An important issue is the choice of the probabilistic distance between the density functions. This choice may be done according to the model of the density functions. Concerning our classification problem, statistical analysis of experimental results have shown that the probability distribution of the wavelet images could be the generalized Gaussian [26]

$$p(x) = \frac{c}{2\sigma\Gamma\left(\frac{1}{c}\right)} e^{-(|x|/\sigma)^c} \tag{11}$$

where the parameter σ is the standard deviation and c reflects the sharpness of the probability density function. For $c = 2$, we obtain the Gaussian function, and for $c = 1$, the Laplacian function.

One of the most popular probabilistic distances in the field of pattern recognition is the *Bhattacharrya* distance [14], [20], [21]. It is related to the well-known *Chernoff bound* and therefore, has an explicit expression for a generalized Gaussian distribution. Since we are dealing with such a distribution, we make use the Bhattacharyya distance $\mathcal{B}$ as our probabilistic distance. The Bhattacharyya distance is given as

$$\mathcal{B}_c = \frac{1}{c} \ln \left(\frac{\sigma_1^c + \sigma_2^c}{2\sqrt{\sigma_1^c \sigma_2^c}}\right) \tag{12}$$

with c described above.

As stated before, interband values are decorrelated, and therefore the distance between two feature vectors $\mathcal{V}_k$ (a candidate face feature vector) and $\mathcal{V}_l$ (a prototype vector) may be computed on a component-pair basis, that is, the distance is considered as a sum of distances relative to each of these component pairs.

TABLE I
CHARACTERISTICS OF THE FACES IN THE TEST DATA SET

Type	Number
Frontal	48
Semifrontal $\pm 15^o$	25
Side$\pm 45^o$	18
Tilted $\pm 25^o$	13
Total	104

In our wavelet packet analysis scheme, we make the possible assumption that approximation images are distributed according to the Gaussian law, while the detail images are distributed according the Laplacian law. We do not consider the mean values differences of the approximation images in our scheme in order to have a distance measure independent of the lighting conditions. Due to the design of the filters, the mean values are zero valued for the detail images.

According to (12), the resulting distance $\mathcal{D}$ between two feature vectors $\mathcal{V}_k$ and $\mathcal{V}_l$ is

$$\mathcal{D}(\mathcal{V}_k,\mathcal{V}_l)=\frac{1}{2}\sum_{i=0}^{3}\ln\left(\frac{\sigma_{ik}^2+\sigma_{il}^2}{2\sigma_{ik}\sigma_{il}}\right)+\sum_{i=4}^{m+3}\ln\left(\frac{\sigma_{ik}+\sigma_{il}}{2\sqrt{\sigma_{ik}\sigma_{il}}}\right). \tag{13}$$

Therefore, classification is performed by evaluating the distance $\mathcal{D}$ from each $\mathcal{R}_i$ feature vector $\mathcal{V}_k$ to the prototype feature vector $\mathcal{V}_l$ of the corresponding size category.

Each $\mathcal{R}_i$ feature vector has to be classified as face or nonface according to distance $\mathcal{D}$ to the average prototype vector of the corresponding size category. $\mathcal{R}_i$ is classified as a face area, if $\mathcal{D}$ is below a threshold T_{HD}, and rejected otherwise. T_{HD} values of 6.0 for *medium* face areas and 8.0 for *large* face areas have been found to be optimal after a large number of experiments for both size categories.

Finally, problems of overlapping selected $\mathcal{R}_i$ areas are solved as follows. The set of face areas that overlap by more than 25% in both dimensions is sorted in ascending order according to normalized distances $\mathcal{D}/hw$, related to the size of $\mathcal{R}_i$. Then, the first ranked face area is selected and the other ones are rejected.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed algorithm has been evaluated using the same test data set as in [17]. This test data set contains images that have been extracted as key-frames from various MPEG videos and especially from the test videos used in the DiVAN project evaluation phase [11]. The video material has been kindly provided by the Institut National Audiovisuel, France and by ERT Television, Greece. The test data set contains 100 images, most of them being extracted from advertisements, news, movies and external shots. This set of 100 images contains 104 faces (with sizes above the minimal one) and ten images which do not contain faces. They cover most of the cases that the algorithm has to deal with. In Table I, we give a detailed description of the content of this set, according to facial characteristics.

In Fig. 9, we present some results of the proposed face detection scheme for 32 images of the test data set. These examples include color images with multiple faces of different sizes, different colors, different positions and images which do not contain any face. False alarms and false dismissals examples are presented as well.

We compared the proposed scheme with the well-known face detector developed by Rowley *et al*. [32] at Carnegie Mellon University. The CMU face detector operates in two stages. It first applies a set of neural network-based filters to an image and then uses a network arbitration and bootstrap algorithm to detect mostly frontal faces of various sizes and at various locations. The filters examine each location in the image at several scales, looking for locations that might contain a face. The arbitrator then merges detections from the individual filters and eliminates overlapping detections. This face detector can handle pictures of people (roughly) facing the camera in an (almost) vertical orientation. The faces can be anywhere inside the image, and range in size from at least 20 pixels height to covering the whole image. Our test data set of 100 images have been processed using the interactive World Wide Web demonstration of this system at http://www.ius.cs.cmu.edu/demos/facedemo.html, which allows anyone to submit an image for processing in batch mode and to retrieve the result image with bounding boxes overlaid on the detected faces.

Table II shows the comparative results of the proposed algorithm to the CMU face detector on our test data set. A classification according to the quality of the detected faces (position of the bounding box) is proposed. Fig. 10 presents some comparative examples. In each pair of images, labeled from (a) to (l), the left image corresponds to the output of the CMU face detector and the right image corresponds to the output of our scheme.

Before commenting the comparative results, a few remarks have to be done. The CMU face detector has been designed to detect faces with a minimal size of 20 $\times$ 20 pixels, which corresponds to the size of the input layers of the neural network filters, whereas our scheme provides results for faces with a minimal size of 80 $\times$ 48 pixels. This can be observed in Fig. 10(a) and (b), where the CMU face detector found several small size faces while our method was not designed to detect them. Therefore, the CMU face detector was able to detect small faces which have not been taken into consideration when labeling the 104 reference faces that we retained in our test data set. Another main difference between the two methods is that the CMU face detector does not use color information. Only the intensity face texture discrimination, corresponding to the last step of our scheme, is performed. Therefore, a set of candidate face areas cannot be built and the neural network filters have to be applied at every pixel location in each image of the multiscale pyramid. Even if a fast implementation is proposed in [32] by introducing a step while moving the search window, the CMU face detector is still more time consuming than our face detector. Moreover, when applied to color images, the CMU face detector may find false alarms in nonskin color areas, whereas our scheme does not have this drawback. On the other hand, as we may see in Fig. 10, our scheme will fail in the case of extreme lighting conditions leading to bad skin color segmentation.

Fig. 9. Some detection results of our method for a subset of the test data set.

TABLE II
RESULTS OF FACE DETECTION ON THE TEST DATA SET

Quality of detection	Results The proposed scheme	Results The CMU face detector
Well framed faces	90 (91.83%)	86 (96.62%)
Moderate framed faces	8 (8.17%)	3 (3.38%)
Detected faces	**98 (94.23%)**	**89 (85.57%)**
False dismissals	6 (5.76%)	15 (14.43%)
False alarms	20	9

As shown in Table II, our algorithm detects 98 of the 104 faces which means a successful detection rate of 94.23%, whereas the CMU detector detects 89 faces, leading to a successful detection rate of 85.57%. We observed that the CMU face detector failed mainly for faces of large size. This can be observed in Fig. 10(c) and (d), even if these faces appears in a upright, semi-frontal positions, whereas our method is very efficient in that case. The main reason could be that neural networks filters are trained for a fixed window

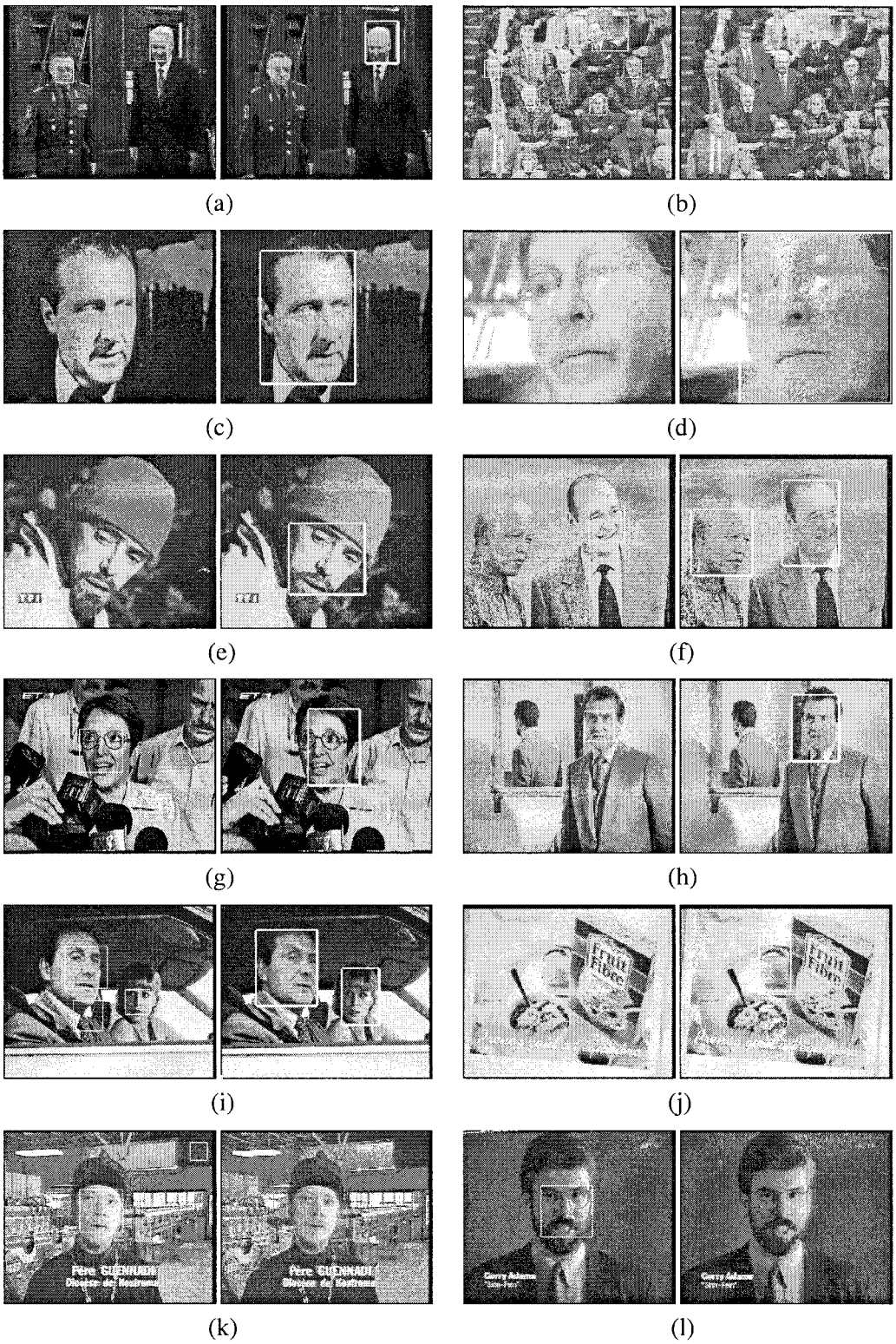


Fig. 10. Some comparative results for the test data set. For each pair of results, the left image corresponds to the output of the CMU face detector and the right image corresponds to the output of our scheme.

size of 20 $\times$ 20 pixels, and in order to detect faces larger than the window size, the input image is repeatedly reduced in size by a factor of 1.2. We believe that some details in the face texture are lost during this subsampling process. This makes the neural networks fail to detect these facial areas. Moreover, due to its design (learning phase), the CMU face detector is more sensitive to face poses than our system, especially for tilted faces. Related examples can be observed in Fig. 10(e) and (f).

Considering the quality of detection, 91.83% (84.53% in our previous work [17]) of the detected faces have been accurately framed, while 8.17% (15.47% in [17]) have been framed

with less accuracy. With the proposed scheme, precision in framing the face has been improved. This is mainly due to the macro-block approach used in [17], which provides a rough approximation of the $\mathcal{R}_i$ areas. Color clustering and segmentation leads to a better precision in locating the faces and helps in segmenting the faces from the background especially when parts of the surrounding background have a color which may be classified as skin color. The CMU face detector performs slightly better than our scheme if quality of detection is considered, with 96.62% of well-framed faces. In that case, the neural network learning of the inner part of the face (around eyes and mouth) gives more accurate results than our color segmentation approach especially in the case of partial occlusion or extreme lighting conditions. Two examples of framing results are presented in Fig. 10(g) and (h).

Less false alarms are obtained using the CMU face detector. Twenty false alarms and six false dismissals (5.76%) are obtained with our method whereas nine false alarms and 15 false dissmissals (14.43%) are obtained using the CMU face detector. After the chrominance segmentation step, a number of potential false alarms appear but the wavelet packet analysis scheme reduces them in a very efficient way, using intensity face texture. By tuning the threshold T_{HD} used for classification according to the prototypes face features vectors, it is possible to control the final false alarms rate, at the cost of having false dismissals, especially if faces are very tilted. Using the CMU interactive face detector demo, we were not able to tune the two parameters required by the "thresholding" heuristic for merging detection [32]. We believe that tuning these parameters may help the CMU face detector to avoid a few false dismissals, at the cost of increasing the false alarms rate. Using both methods, some false alarms will still appear because some regions have colors and textures of faces, without corresponding to human faces. False dismissals cannot be totally avoided, especially in scenes with many partially occluded faces or under extreme lighting conditions. An example of false alarm from the CMU face detector is presented in Fig. 10(i) and an example of false alarm from our scheme is presented in Fig. 10(j). Another false alarm from the CMU face detector surprisingly appears in Fig. 10(k). Among our six false dismissals cases, three faces were either very light or very dark and three faces were small, close to the lower bound allowed size. Such an example is shown in Fig. 10(k) and (l), where the faces have a color out of the range of our approximated skin-color subspace, due to lighting conditions.

VI. Conclusion

We have presented a novel scheme for human faces detection in color images under nonconstrained scene conditions, such as the presence of a complex background and uncontrolled illumination. First, a LBG color quantization is performed and the quantized images are filtering using approximations of the HSV or YCbCr skin color subspaces, providing skin color regions. A merging stage is then iteratively performed on the set of homogeneous skin color regions in the color quantized image, in order to provide a set of candidate face areas. Constraints related to shape and face texture analysis are applied, by performing a wavelet packet decomposition on each face area candidate and extracting simple statistical features such as standard deviation. These features provide an excellent basis for face detection, if an appropriate distance is used. The use of the Bhattacharyya distance proved to be very suitable for classifying these feature vectors into faces or nonfaces classes. For a data set of 100 images with 104 faces covering most of the cases of human faces appearance, a 94.23% good detection rate, 20 false alarms and a 5.76% false dismissals rate were obtained. A comparison with the CMU face detector system showed that our system is very efficient especially for large size faces detection and that skin color segmentation helps to improve the speed of face detection process by pre-selecting candidate face areas. Thus, the wavelet packet analysis provide a robust scheme for face detection, even if no constraint is imposed on the faces to be detected (except a minimal size). Moreover, very fast implementations of wavelet decomposition are available in hardware form, to cope with real time face detection needs. As an extension of this work, we believe that a slightly different feature extraction process may be performed, that extracts statistical information only from the wavelet images, which convey most information about faces texture.

Acknowledgment

The authors would like to thank S. Liapis for the implementation of the color quantization algorithm, and the INA's Innovation Department and ERT Television for having provided test video materials.

References

[1] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces versus fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 711–720, July 1997.

[2] S. Belongie, C. Carson, H. Greenspan, and J. Malik, "Color- and texture-based image segmentation using EM and its application to content-based image retrieval," *Proc. 6th IEEE Int. Conf. Computer Vision*, Jan. 1998.

[3] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 1042–1052, Oct. 1993.

[4] M. C. Burl, T. K. Leung, and P. Perona, "Face localization via shape statistics," presented at Int. Workshop on Automatic Face and Gesture Recognition, June 1995.

[5] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, pp. 705–740, May 1995.

[6] K. Chopra and R. K. Srihari, "Control structures for incorporating picture-specific context in image interpretation," in *Proc. Int. Joint Conf. Artificial Intell.*, 1995.

[7] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inform. Theory*, vol. 38, pp. 713–718, 1992.

[8] J. Cox, Y. J. Ghosen, and P. Yianilos, "Feature-based face recognition using mixture distance," in *Proc. Computer Vision Pattern Recognition*, 1996, pp. 209–216.

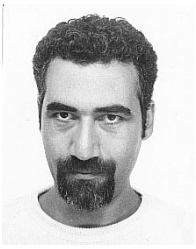
[9] Y. Dai and Y. Nakano, "Recognition of facial images with low resolution using a Hopfield memory model," *Pattern Recognit.*, vol. 31, no. 2, pp. 159–167, 1998.

[10] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 961–1005, May 1990.

[11] DiVAN: Distributed audioVisual Archives Network (European Esprit Project EP 24956). http://divan.intranet.gr/info, 1997.

[12] A. Eleftheradis and A. Jacquin, "Model-assisted coding of video teleconferencing sequences at low bit rates," in *Proc. IEEE Int. Symp. Circuits and Systems*, 1994, pp. 3.177–3.180.

[13] M. Ferretti and D. Rizzo, "Wavelet transform architectures: A system level review," in *Proc. 9th ICIAP'97*, 1997, pp. 77–84.
[14] Y. Fu. *Handbook of Pattern Recognition and Image Processing*. New York: Academic, 1986.
[15] C. Garcia, G. Zikos, and G. Tziritas, "A wavelet-based framework for face recognition," in *Int. Workshop on Advances in Facial Image Anal. Recognition Technology, 5th European Conf. Computer Vision*, 1998.
[16] ———, "Wavelet packet analysis for face recognition," *Image and Vision Computing*, to be published.
[17] ———, "Face detection in color images using wavelet packet analysis," *Proc. IEEE Intern. Conf. Multimedia Computing and Systems*, Florence, vol. I, pp. 703–708, June 1999.
[18] Z. Q Hong, "Algebraic feature extraction of image for recognition," *Pattern Recognit.*, vol. 24, no. 3, pp. 211–219, 1991.
[19] S.-H. Jeng, H. Y. M. Yao, C. C. Han, M. Y. Chern, and Y. T. Liu, "Facial feature detection using geometrical face model: An efficient approach," *Pattern Recognit.*, vol. 31, no. 3, pp. 273–282, 1998.
[20] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun.*, vol. COM-15, pp. 52–60, 1967.
[21] D. Kazakos, "The Bhattacharyya distance and detection between Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 747–754, June 1978.
[22] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure and the characterization of human faces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 103–108, Jan. 1990.
[23] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 743–756, July 1997.
[24] S.-H. Lin, S.-Y. Kung, and L.-J. Lin, "Face recognition/detection by probabilistic decision-based neural network," *IEEE Trans. Neural Networks*, vol. 8, pp. 114–131, Jan. 1997.
[25] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, 1980.
[26] S. Mallat, "A theory of multi-resolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 674–693, 1989.
[27] B. Moghaddam and A. P. Pentland, "Probabilistic visual learning for object detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 696–710, July 1997.
[28] O. Nakamura, S. Mathur, and T. Minami, "Identification of human faces based on isodensity maps," *Pattern Recognit.*, vol. 24, no. 3, pp. 263–272, 1991.
[29] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," in *Proc. SPIE, Storage and Retrieval and Video Databases II*, 1994.
[30] J. L. Perry and J. M. Carney, "Human face recognition using a multilayer perceptron," in *IJCNN*, 1990, pp. 413–416.
[31] P. J. Phillips, R. McCabe, and R. Chellappa, "Biometric image processing and recognition," in *Proc. IX European Signal Processing Conference*, vol. I, 1998.
[32] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–28, 1998.
[33] S. Satoh and T. Kanade, "Name-it: Association of face and name in video," in *Proc. Computer Vision and Pattern Recognition*, 1997, pp. 368–373.
[34] P. Sinha, "Object recognition via image invariants: A case study," *Invest. Ophthal. Vis. Sci.*, vol. 35, pp. 1.735–1.740, 1994.
[35] M. Smith and T. Barnwell, "Exact reconstruction techniques for tree structured subband coders," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 434–441, Mar. 1986.
[36] K. K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 39–51, 1998.
[37] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 831–836, Aug. 1996.
[38] S. Tsekeridou and I. Pitas, "Facial features extraction in frontal views using biometric analogies," in *Proc. IX European Signal Processing Conference*, 1998, vol. I.
[39] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Sci.*, vol. 3, no. 1, pp. 71–86, 1991.
[40] H. Wang and S.-F. Chang, "A highly efficient system for automatic face region detection in MPEG video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 4, pp. 615–628, 1997.
[41] P. H. Westerink, J. Biemond, and D. E. Boekee, "Subband coding of color images," *Subband Image Coding*, J. W. Woods, Ed. Boston, MA: Kluwer, 1991, pp. 193–228.
[42] C. L. Wilson, C. S. Barnes, R. Chellappa, and S. A. Sirohey, "Face recognition technology for law enforcement applications," NISTIR 5465, U.S. Dept. Commerce, 1994.
[43] L. Wiskott, J. M. Fellous, N. Kruger, and C. Von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 775–779, July 1997.
[44] G. Yang and T. S. Huang, "Human face detection in a complex background," *Pattern Recognit.*, vol. 27, no. 1, pp. 55–63, 1994.
[45] K. C. Yow and C. Cipolla, "Feature-based human face detection," *Image Vis. Comput.*, vol. 15, pp. 713–735, 1997.



Christophe Garcia was born in L'Horme, France, on September 7, 1966. He received the M.S. and Ph.D. degrees from the University of Lyon, Lyon, France, in 1991 and 1994, respectively.

He has been involved in numerous computer vision projects during his stays of two years at the IBM Montpellier Vision Automation Group, France, and one year at the Computer Vision Center of the Universitat Autònoma de Barcelona, Spain. As a fellow of the European Consortium in Informatics and Mathematics (ERCIM), he has spent one year in the German National Research Center (GMD), working towards the development of autonomous and multi-agents robots. In 1998, he joined the Foundation for Research and Technology Hellas (FORTH, Greece) where he is currently involved in advanced European projects in the field of multimedia and image analysis. His research interests include neural networks, pattern recognition, video indexing and computer vision.

Georgios Tziritas (M'89) was born in Heraklion, Crete, Greece, on January 7, 1954. He received the Diploma of Electrical Engineering in 1977 from the Technical University of Athens, Athens, Greece, the Diplôme d'Etudes Approfindies (DEA) in 1978, the Diplôme de Docteur Ingénieur in 1981, and the Diplôme de Docteur d'Etat in 1985, all from the Institut National Polytechnique de Grenoble, Grenoble, France.

From 1982 to August 1985, he was a Researcher, Centre National de la Recherche Scientifique (CNRS), with the Centre d'Etudes des Phénomènes Aléatoires (CEPHAG), then with the Institut National de Recherche en Informatiques et Automatique (INRIA), until January 1987, and the Labroatoire des Signaux et Systèmes (LSS). In 1992, he joined the Computer Science Department, University of Crete, where he is currently an Associate Professor teaching digital image processing, digital signal processing, information and coding theory, image sequence analysis, and artificial neural networks. He is coauthor (with C. Labit) of the book *Motion Analysis for Image Sequence Coding* (Amsterdam, The Netherlands: Elsevier, 1994), and of more than 50 journal and conference papers on signal and image processing, and image and video communication. His research interests are in the areas of signal processing, image processing and analysis, computer vision, motion analysis, video compression, video indexing, and image and video communication.