Pattern Analysis & Applications (2001)4:93–107

Facial Image Indexing in Multimedia Databases

N. Tsapatsoulis, Y. Avrithis and S. Kollias

Image, Video and Multimedia Systems Laboratory, Computer Science Division, Department of Electrical and Computer Engineering, National Technical University of Athens, Zographou, Greece

Abstract: Pictures and video sequences showing human faces are of high importance in content-based retrieval systems, and consequently face detection has been established as an important tool in the framework of many multimedia applications like indexing, scene classification and news summarisation. In this work, we combine skin colour and shape features with template matching in an efficient way for the purpose of facial image indexing. We propose an adaptive two-dimensional Gaussian model of the skin colour distribution whose parameters are re-estimated based on the current image or frame, reducing generalisation problems. Masked areas obtained from skin colour detection are processed using morphological tools and assessed using global shape features. The verification stage is based on a template matching variation providing robust detection. Facial images and video sequences are indexed according to the number of included faces, their average colour components and their scale, leading to new types of content-based retrieval criteria in query-by-example frameworks. Experimental results have shown that the proposed implementation combines efficiency, robustness and speed, and could be easily embedded in generic visual information retrieval systems or video databases.

Keywords: Content-based retrieval; Face detection; Face indexing; Multimedia databases

1. INTRODUCTION

Multimedia applications have attracted increasing interest during recent years, leading to a growing demand for efficient storage, management and browsing in multimedia databases. More and more audiovisual information is becoming available represented in various forms of media, such as still pictures, video, graphics, 3D models, audio and speech. On the other hand, there is an increasing number of applications, such as image understanding, media conversion, information retrieval or filtering, where audiovisual information is created, exchanged, retrieved and re-used by computational systems. As a consequence, new tools for summarisation, content-based query, indexing and retrieval have received considerable attention, especially for browsing digital video databases, due to the huge amount of information involved. Such tools are of major importance in the context of the emerging MPEG-4 [1] and MPEG-7 [2] multimedia standards.

Several frameworks have been proposed in the recent literature for content-based indexing in image or video databases [3,4] and a lot of prototype systems have emerged, providing content-based image query and retrieval capabilities. Some of these systems, for example including VIRAGE, QBIC, Photobook and VisualSEEk, are already in the stage of commercial exploitation, and have been successfully used during the past years. In most cases, content information is handled by video object modelling and segmentation, and subsequent extraction of object attributes including colour, motion, texture, shape as well as spatial and temporal relation between objects [5,6].

The experience gained from the usage of such content-based indexing and retrieval systems clearly indicates that it is necessary to develop forms of audiovisual information representation that go beyond the simple frame-based, or even the object-based, representation of MPEG-4. For this reason, an integral part of the MPEG-7 standardisation process is to specify a set of standard multimedia content Descriptors and Description Schemes by means of a special language, the Description Definition Language (DDL) [7]. Although these descriptors do not depend upon the ways the described content is obtained, coded, stored or used, the key to the success of the future

Received: 30 May 2000
Received in revised form: 15 December 2000
Accepted: 15 February 2001

multimedia systems is the degree to which high-level, semantic information can be automatically extracted in order to speed up the audiovisual document characterisation process.

It has often been pointed out that existing systems lack the ability to extract and retrieve semantic information, which is naturally due to the fact that such capabilities always require a priori knowledge and can only be achieved in the context of specific applications. Examples of such applications with increasing interest include retrieval of images containing human faces, as well as subsequent face detection, segmentation or recognition. Since pictures and video sequences showing human faces play a significant role in the description of multimedia content, *face detection* has been established as an important tool in the framework of applications like video indexing and retrieval, video scene classification and video/news summarisation [8].

In the past the term 'face detection' was strongly related to the *face recognition* task; this fact had a deep impact in the developed algorithms. To achieve the required accuracy of detection rigorous constraints had been posed in the digital image environment [9]. Moreover, the great majority of the corresponding algorithms were based in greyscale images utilising face template matching, image invariants or low level features for the detection of principal facial features like eyes, nose and mouth [10,11]. Modern applications, however, tend to employ colour characteristics and require fast implementations with sufficient accuracy rather than exhaustive procedures providing higher precision. As a result, well established algorithms which had been successfully used for face recognition are not appropriate or require re-development, while novel approaches have emerged, effectively decoupling the face detection task from face recognition.

In particular, the work presented in Wang and Chang [12] received considerable attention and inspired many researchers for implementing colour-based face detection algorithms, as it combines very fast implementation with promising results. The basic idea of Wang and Chang [12] is the use of *colour thresholding* for face detection by means of a skin colour model based on the chrominance components of the *YCrCb* colour space and a suitable *skin colour distribution*. Most of the studies based on this idea reveal that considerable effort is required in post-processing steps to achieve remarkable results [13,14]. It is not clear, however, whether the post-processing steps are sufficient for face detection based on skin colour characteristics. Although the skin colour subspace indeed covers a small area of the *Cr-Cb* chrominance plane, it cannot be modelled in such a general way to be efficient for all images that include faces. To improve the generalisation ability, the skin colour model should be 'relaxed' leading to an increased number of *false alarms*. On the other hand, a 'rigorous' model increases the number of *dismissals*. Moreover, the influence of the luminance channel *Y* is not totally negligible.

For this purpose, a technique for dynamically updating a colour model to accommodate changes in apparent colour due to varying lighting conditions has been presented in Raja et al [15], where a Gaussian mixture model is employed to estimate colour probability densities for skin, clothing and background. Moreover, a Markov model is used in Sigal et al [16] to predict the evolution of a skin colour histogram over time in video sequences. The histogram spans all three components of the *HSV* colour space in contrast to *YCrCb*-based approaches, where only the *Cr-Cb* chrominance plane is used. A similar skin colour histogram detector has been proposed in Jones and Regh [17] using the *RGB* colour space directly; its performance on a large image dataset is shown to be superior to mixture models. All the above works tackle the problem of skin detection, whereas in Sun et al [18] skin colour features are combined with face outline and local symmetry information to locate potential facial feature points, in order to detect the presence of human faces. A general purpose colour segmentation approach, combined with a Gaussian skin colour model and shape filtering is also proposed in Tsapatsoulis et al [19] for face detection.

In this study, we combine skin colour and shape features with template matching in an efficient way for the purpose of facial image indexing. In particular, we propose an adaptive *two-dimensional Gaussian model* of the skin colour distribution whose parameters are re-estimated based on the current image in the case of still pictures, and on the previous frame in the case of video sequences. Masked areas obtained from skin colour detection are then processed using morphological tools and assessed using global *shape features*, in order to discard segments that possess extremely irregular shape. The content of the remaining skin segments is further examined in the verification stage, which is based on a *template matching* variation using a sample face pattern. Searching for entire face pattern provides superior performance to local facial feature detection in cases of low-resolution images or small facial scales.

The overall approach improves the efficiency of face detection in two main ways: the generalisation ability of the skin colour model is significantly improved by re-estimating the distribution's parameters in a per frame basis, and the verification stage is much more reliable since it takes into account texture information, apart from colour and shape characteristics. This reliability gain is obtained at an additional computational cost that is limited, since template matching is only employed on the detected skin segments.

The semantic characterisation of audiovisual programs usually requires a huge amount of human effort. Since human faces are the most common kind of semantic objects in programs such as news, debates, television series or movies, multimedia databases can benefit from the use of the proposed approach as a fully automated face indexing tool. In cases where accuracy is of prime importance, or face annotation and additional meta-information is required, it can also be used as a semi-automatic procedure assisting human operators in an interactive framework.

Moreover, an efficient scheme is proposed for *indexing facial images/video sequences* according to the number of faces included, their average colour components and their scale, leading to new types of content-based retrieval criteria in query-by-example frameworks. This scheme can be either used as a standalone *face retrieval* application, or embedded in generic visual information retrieval systems in order to enhance their performance and provide them with high-level, semantic indexing capabilities.

The paper is organised as follows. In Section 2 we describe the adopted skin colour model and the procedure for detecting image areas containing skin segments, based on their colour characteristics. Section 3 introduces the procedure used to extract and separate the various skin segments based on morphological operators and shape features. Section 4 presents the verification procedure where template matching is employed within the detected skin segments, while in Section 5 we discuss the criteria upon which face images can be indexed in a video database. Finally, experimental results are given in Section 6, and conclusions are drawn in Section 7.

2. SKIN COLOUR MODELLING AND DETECTION

This section describes the method followed to locate the probable skin segments, i.e. image/frame areas that contain objects with colour similar to that of the human skin. It is stated in some classic studies [20,21] that skin-tone colours are spread over a small area of the *Cr-Cb* chrominance plane of the *YCrCb* colour model. Wang and Chang [12], based on that idea, presented a fast face detection algorithm that inspired many researchers. In a similar way, we approximated skin-tone colours distribution using a *two-dimensional Gaussian density function*. Expanding the idea of Wang, we use a feedback model for re-estimating skin colours based on the current image or the previous video frame.

Although a mixture of Gaussian functions is more appropriate in many cases [15], the use of model parameter re-estimation compensates for the multi-modal skin colour distribution, while the template matching verification stage reduces potential false alarm cases due to colour model mismatches. Moreover, a Gaussian mixture model is in general more difficult to estimate, and still requires adaptation. On the other hand, Bayesian detectors based on skin colour histograms have been reported to produce higher classification rates [16,17], but their adaptation involves increased computational cost.

The mean vector μ_0 and the covariance matrix $\mathbf{C}$ of the human skin chrominance components are initially estimated from training data containing facial areas of different races, obtained from regular TV clips, colour images and personal video cameras. Then, according to the Gaussian distribution

$$P(\mathbf{x}|\mu_0,\mathbf{C}) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\mathbf{C}^{-1}(\mathbf{x}-\mu_0)\right\}}{(2\pi)^{\frac{k}{2}}\cdot|\mathbf{C}|^{\frac{1}{2}}} \tag{1}$$

where $k = 2$ is the number of chrominance components, the likelihood of an input pattern $\mathbf{x}$ (chrominance components of a particular pixel or average chrominance components of an image block) can be approximated by the quantity

$$p(\mathbf{x}) = \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\mathbf{C}^{-1}(\mathbf{x}-\mu_0)\right\} \tag{2}$$

Although using the above Gaussian model is rather efficient for the classification of pixels as skin and non-skin ones, especially in a controlled illumination environment, better performance is obtained by *re-estimating* the mean vector μ_0 based on the current image/previous video frame. In particular, the initial Gaussian distribution is used for a first pass classification, employing a threshold value which is based on the statistics of the likelihood on the current image, so that it adapts to varying lighting conditions. Let I be the current frame of dimensions $M \times N$, and I_{Cr} and I_{Cb} its Cr and Cb components, respectively. Also, let $p(\mathbf{x}(s))$ be the likelihood of pixel s with chrominance components $\mathbf{x}(s) = [I_{Cr}(s)I_{Cb}(s)]$. Then the threshold value is selected as $T_I = \mu_I + \sigma_I$ where

$$\mu_I = \frac{1}{MN}\sum_{s\in I} p(\mathbf{x}(s))$$

is the average likelihood of image I and

$$\sigma_I = \frac{1}{MN-1}\sqrt{\sum_{s\in I}\{p(\mathbf{x}(s)) - \mu_I\}^2}$$

is the corresponding standard deviation. The pixels that are classified as skin ones are used for the re-estimation of μ_0 according to the following equation:

$$\mu_0 = (1-m)\cdot\mu + m\cdot\mu_0 \tag{3}$$

where μ is the mean chrominance vector of image segments classified as skin ones, estimated from of the current image, and m is a memory tuning constant. The value of m defines the amount of model's readjustment. A high value prevents the initial model from being significantly altered, while a small one quickly adapts the model to the current data. It can be estimated by measuring the false alarm/dismissal rates or the determinant of the confusion matrix as a function of m and selecting a suitable point on the corresponding graph (ROC curve) [16]. From our experiments $m = 0.7$ has been shown to be a good compromise. Note also that the adaptation of the covariance matrix **C** is also possible; however, its effect on classification performance turns out to be insignificant.

For the final classification the *adapted Gaussian model* combined with a minimum risk threshold, estimated using the maximum likelihood criterion on the training set, is applied to the input images/video frames producing a binary image mask. In case there are no skin regions detected, the adaptation of μ_0 is considered as erroneous, and its previous value is restored. Thresholding can either be directly applied to image pixels, or to larger image blocks. The latter approach usually gives more robust results.

Equation (3) is also used to adapt the model in a fully dynamic environment. In video sequences the model is re-estimated in a per frame basis, giving a robust framework for skin segment tracking. In particular, the first-pass classification is performed in the first video frame, using threshold T_I. This scheme guarantees that there will always be some image parts classified as skin ones, and their average chrominance vector μ is used for model adaptation. In subsequent frames only the final classification is employed, using the minimum risk threshold; the adaptation of the model is performed based on the skin areas of the previous frame. If there is at least one skin segment detected, its average chrominance vector is used for adaptation; if, however, no skin segments were found in the previous frame, the first-pass classification is also performed in the current frame, the model is adapted and final classification is employed again. If the final binary mask produced after adaptation still contains no skin segments, the previous value of μ is restored, as in the case of static images.

The overall scheme manages to keep track of illumination changes in video sequences, and at the same time handles the appearance of faces with new colour characteristics without deviating from the generic skin colour model, thus risking the possibility of a high false alarm rate. Moreover, in an interactive content-based retrieval environment, face segments selected by the user can be exploited for the re-estimation of the model parameters [14]. This procedure is further examined in Section 5.

3. EXTRACTION OF FACE SEGMENTS

In the second stage, morphological operators like *opening* and *closing* are applied to spatially filter the obtained image masks, while the morphological *distance transform* and size distribution techniques [22] are used to isolate the disconnected areas and provide separate skin segments. *Shape features* are then employed to discard skin segments that possess irregular shape, and the bounding rectangles of the remaining segments are used in the subsequent verification stage.

3.1. Segment Isolation

The colour thresholding procedure described in the previous section does not take spatial information into account. For this reason, morphological operators like *opening* and *closing* [23] are applied to spatially filter the image mask, resulting in compact skin segments with smooth contours while eliminating very small ones. This is desired in the context of content-based retrieval, since small segments are not of high importance even if they do contain human faces. An example is given in Fig. 1, where for the input colour image of Fig. 1(a), the image mask obtained is presented in Fig. 1(c). It is evident that the faces of the crowd in the image background have been removed, while the hand of the person in the foreground does match the skin colour probabilistic model, and can only be eliminated by the template matching procedure of Section 4.

To isolate the disconnected areas of the binary image mask according to their size, providing separate skin segments, the morphological *distance transform* and size distribution techniques are employed [22]. Isolation is necessary for three main reasons: (i) the template matching verification stage has to be applied separately on each potential face segment; (ii) apart from separating segments, it is often required to filter them according to size and discard small

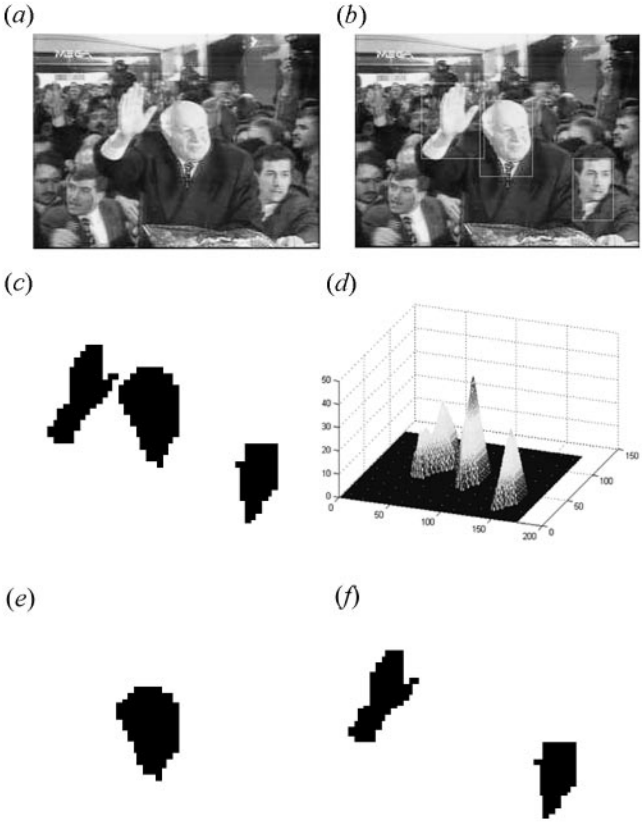


Fig. 1. Face detection based on colour and shape features. (a) Original image, (b) bounding rectangles of detected facial areas, (c) image mask of skin areas, (d) distance transform of the image mask, (e) isolation of the largest segment, (f) remaining segments.

ones; (iii) a per-segment comparison to ground-truth image masks is preferable for evaluation purposes, as explained in the experiments. Let B denote the binary image mask and $D(B)$ its distance transform. The following steps are then performed:

Step 1: Compute the set M which consists of the points of B whose distance transform value is maximum, i.e. $M = \{b \in B : D(b) = D_{max}\}$, where $D_{max} = \max\{D(B)\}$.

Step 2: Compute the set $O \subseteq B$ which corresponds to the reconstructed part of set B by applying *opening by reconstruction* with M as the marker set [23].

Step 3: Store set O as a skin segment.

Step 4: Remove the reconstructed part from set B, i.e. $B = B \cap O^c$ where O^c is the complement of O.

Step 5: If $B \neq \emptyset$ go to Step 1.

The binary sets O that have been produced in Step 3 of the above procedure correspond to separate skin segments obtained from the whole image mask. For example, the mask of Fig. 1(c) gives the distance transform of Fig. 1(d), while the isolation procedure results in the largest face segment of Fig. 1(e) and the remaining two segments of Fig. 1(f). More details about the distance transform and the opening by reconstruction procedures can be found in Maragos [23]. Note that a classic connected components algorithm with lower computational cost could also be used. The proposed technique, however, locates segments in decreasing order of size, and is possible to terminate before separating the smallest ones. Besides, its complexity is low compared to that of the template matching procedure, so it does not affect the overall speed of detection.

3.2. Shape Features

Since objects unrelated to human faces but whose colour chrominance components are still similar to those of the skin might be present in an image or video sequence, the contour shape of the isolated skin segments is also taken into account. Ideally, the similarity of segment shape to an ellipsis (or a more accurate face shape template) can be calculated, since face shape is supposed to be elliptical. Shape matching under arbitrary deformations can be achieved using, for example, active contour models (snakes) or deformable templates [24,25]. Moreover, rigid motion or affine transformations such as translation, rotation, scaling and skew can be efficiently removed by employing affine invariants or curve normalisation [26].

In most realistic cases, however, the object contours obtained through colour segmentation are far from the ideal. Even if contour curves are simplified or smoothed, using splines for instance, shape matching usually gives poor results. For this reason, only global shape features are taken into account. In particular, the contour of each isolated skin segment is calculated, and its *compactness* is obtained from its perimeter and the area of the corresponding segment:

$$g_X = 4\pi \frac{a_X}{r_X^2} \tag{4}$$

where r_X denotes the perimeter (number of contour points) and a_X is the area (total number of pixels) of segment X. Note that the maximal compactness is achieved for a circular shape and is equal to one, hence g_X is always normalised in the interval [0,1]. Naturally, this holds only for a Euclidean space, and the discrete image space where perimeter and area are measured is not such a space. This means that g_X may take values higher than one, but this is not a concern since it is further transformed using a non-linear function as described below.

The shape *elongation* is then obtained through its Hotelling, or discrete Karhunen–Loeve transformation. Let the $L \times 1$ vectors $\mathbf{x}$ and $\mathbf{y}$ denote the coordinates of L contour points representing the closed shape boundary of segment X. The 2×2 covariance matrix $\mathbf{C}_X$ of these points with respect to their center-mass point (μ_x, μ_y) is given by

$$\mathbf{C}_X = \frac{1}{N}[\mathbf{x}-\mu_x\mathbf{e}\ \ \mathbf{y}-\mu_y\mathbf{e}]^T[\mathbf{x}-\mu_x\mathbf{e}\ \ \mathbf{y}-\mu_y\mathbf{e}] \tag{5}$$

where $\mathbf{e}$ is the $L \times 1$ vector $[1\ \ 1\ \ 1]^T$. The two eigenvectors of this covariance matrix express the principal – minor and major – axes of the shape, while the ratio of its eigenvalues defines the elongation of the shape of segment X:

$$l_X = \sqrt{\lambda_2(\mathbf{C}_X)/\lambda_1(\mathbf{C}_X)} \tag{6}$$

where $\lambda_1(\mathbf{C}_X)$, $\lambda_2(\mathbf{C}_X)$ are the maximum and minimum eigenvalues of $\mathbf{C}_X$, respectively. The above global shape features are fairly robust to segmentation noise and invariant to translation, scaling and rotation.

Experimental results have shown that typical values corresponding to face segments range from 0.44 to 0.79 for shape compactness and from 0.59 to 0.91 for elongation. Consequently, shape matching is achieved by transforming compactness and elongation with appropriate non-linear functions taking values in the range [0,1], similarly to *fuzzy membership functions*: $g'_X = \mu_g(g_X)$, $l'_X = \mu_l(l_X)$. The nonlinear functions $\mu_g(\cdot)$ and $\mu_l(\cdot)$ used in our experiments are shown in Fig. 2. The transformed compactness/elongation are always in the range [0,1] and express the similarity to the compactness/elongation of typical face outlines.

Finally, the transformed compactness g'_X and elongation l'_X are combined with the skin-colour likelihood p_X (obtained from Eq. (2) and averaged over the whole area of segment X) using a weighted geometric mean, and an overall *face likelihood* is obtained, denoted as f_X. Weights are empirically estimated based on experiments, and depend on the values of g'_X and l'_X. In particular, since p_X is usually more reliable for face detection – skin colour is far more characteristic for a human face than its shape – it is assigned a higher weight, except if g'_X and l'_X take values close to zero. This means that shape features are in effect used only to discard face segments that possess extremely irregular shape, although they match the skin-colour probabilistic model.

Skin segments are assessed against their shape by employing a suitable threshold on the overall face likelihood.

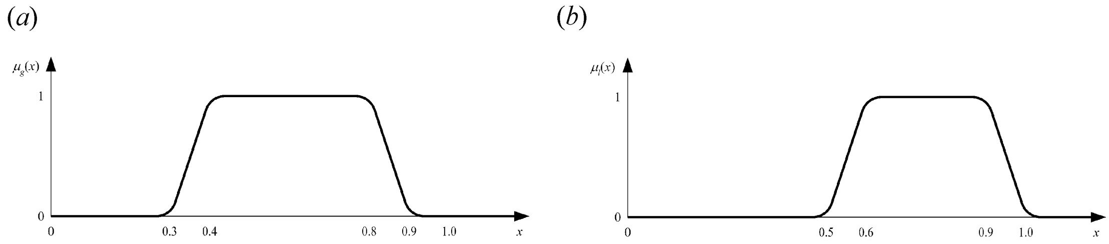


Fig. 2. Nonlinear functions for the transformation of (a) compactness and (b) elongation into the range [0,1].

Each of the remaining segments is marked as a probable face segment, and its bounding rectangle is cropped and used to feed the subsequent verification stage. For example, the bounding rectangles of the probable face segments corresponding to the input image of Fig. 1(a) are illustrated in Fig. 1(b). Note that the right hand of the person in the foreground is not discarded based on its shape; face template matching is employed for this purpose.

4. VERIFICATION STAGE: TEMPLATE MATCHING

In the third stage, the texture of all isolated probable face segments modelled by their bounding rectangles is assessed by means of a face template matching variant. Although this is a time-consuming procedure, the overall face detection complexity is not severely affected, since template matching is only applied on small areas of the whole image/video frame. In particular, let $M(u,\theta)$ be a face pattern at scale $u(h,v)$, described by horizontal scaling h, vertical scaling v, and rotation θ. The face pattern $M(u,\theta)$ is an averaged face area computed over a training set of face images. Let F be an image area which possibly contains a face at arbitrary scale, location and rotation, and A a sub-area of F. We use the following metric to find the minimum correlation between A and M at scale u and orientation θ:

$$r(u,\theta) = \min_{A \subset F} \left\{ \frac{|A - M(u,\theta)|}{\rho \cdot a \cdot b} \right\} \tag{7}$$

where

$$\rho = 1 - c \cdot \left| \frac{h}{v} - \frac{2}{3} \right|$$

is used to account for the face anatomy, c is a constant $(0 < c < 0.5)$ and a,b are the mean greyscale values of area A and pattern M, respectively, used to account for illumination variations [27]. If the value of r is lower than a given threshold, then F is accepted as an area that contains a face. The best scale and orientation are obtained by $[U,\Theta] = \arg\min\{r(u,\theta)\}$, and the final detection is performed using the template $M(U,\Theta)$ and the metric in Eq. (7). The corresponding detected area $A*$ is the required facial area corresponding to the image area F.

By using information gained from Section 3, h is upper bounded by rectangle's width while $v < 2h$ according to the face anatomy, the best scale U is efficiently estimated and the computational complexity is significantly reduced. An example of face detection and verification based on template matching in given in Fig. 3, corresponding to the input image of Fig. 1(a) and the bounding rectangles of Fig. 1(b). It is clear that of the three probable face segments, only two are retained, while the hand segment is discarded. Moreover, the exact location of the face segments is more accurately estimated. The benefit of substantial robustness and accuracy thus justifies the additional computational cost of the verification stage.

It should be mentioned that the above procedure has been used as a standalone face detection method [27], especially for greyscale images where colour information is not available and skin detection cannot be applied. The main disadvantage is its computational complexity, since it requires searching in the entire image/video frame. Moreover, detection of multiple face segments demands either *a priori* knowledge of the number of faces, or the empirical estimation of a threshold. Nevertheless, a huge amount of audiovisual material, especially from historical archives, is already stored in several databases. For this reason, our experiments include cases of greyscale images and video sequences.

In the case of video sequences, the temporal correlation of consecutive video frames is taken into account, mainly

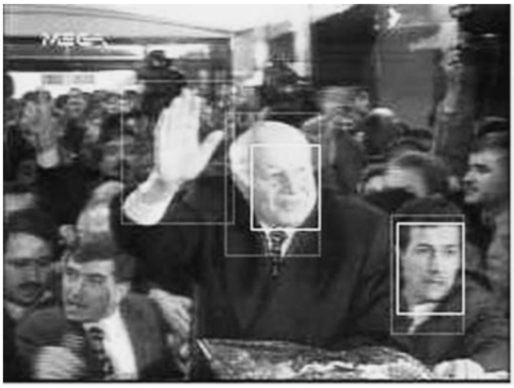


Fig. 3. Multi-face detection and verification. The outer rectangles are located by skin colour detection; the inner rectangles are detected by template matching.

for speed up purposes. In particular, the entire search procedure of Eq. (7) is only applied in the first frame; in subsequent frames, information gained about the size, location and orientation of face segments detected is used to determine the starting point of the search procedure, minimising the detection effort. Apart from reducing the computational cost, temporal correlation can also be exploited for the enhancement of detection performance. For example, since position parameters of detected face segments are known for all previous frames, their values can be *predicted* for subsequent frames, and estimated values can be compared to the predicted ones for evaluation purposes. Moreover, if optical flow information is already available (e.g. motion vectors in MPEG sequences), it can be exploited for tracking face segments. Those issues are currently under investigation.

5. INDEXING CRITERIA

Once facial segments have been successfully detected and verified, we exploit the derived information, including number of face segments, scale and average chrominance components. This information is stored in a multimedia database and employed for content-based indexing and retrieval purposes, leading to new types of indexing criteria in a *query-by-example* framework. Input images are analysed in real-time, and features obtained from the face segments detected are employed for face matching and retrieval in the database.

5.1. Facial Segment Features

Each image in the database is indexed by the following features, resulting from the proposed face detection approach:

- *Number of face segments*.
- *Average chrominance components of each face segment*. Let I_Y, I_{Cr} and I_{Cb} be the Y, *Cr* and *Cb* components of image *I*, respectively, and S_i its *i*th face segment. Then

  $$m_{Cr}(S_i) = \frac{1}{\|S_i\|}\sum_{s\in S_i} I_{Cr}(s)$$

  is the average *Cr* component and

  $$m_{Cb}(S_i) = \frac{1}{\|S_i\|}\sum_{s\in S_i} I_{Cb}(s)$$

  the average *Cb* component of S_i.
- *Percentage $A(S_i)$ of image area covered by each face segment*.

Additional features can be employed in the indexing process, including face location, rotation, shape features (compactness, elongation, etc.) as well as texture descriptors (e.g. moment invariants, FFT coefficients, etc.). Moreover, these can be combined with features used in generic image retrieval environments, such as colour characteristics of the input image and motion patterns of objects, resulting in a variety of retrieval scenarios.

5.2. Retrieval

In the context of this work, three types of indexing criteria are used for image retrieval, which have been examined experimentally with promising results.

5.2.1. Average Colour of Face Segments. The user is interested in retrieving images or video shots in which face segments have similar colour properties as the one they present to the system. A whole face segment is then automatically detected and assessed using the proposed scheme; if more than one face is present, the user can interactively select the particular face of interest. Let *F* be the selected face segment. The average *Cr* and *Cb* components of *F*, calculated as

$$m_{Cr}(F) = \frac{1}{\|F\|}\sum_{s\in F} I_{Cr}(s)$$

and

$$m_{Cb}(F) = \frac{1}{\|F\|}\sum_{s\in F} I_{Cb}(s)$$

respectively, are used to modify the mean vector parameter of the model of Eq. (1) according to

$$\mu_0 = (1-m)\cdot\mu + m\cdot\mu_0$$

where $\mu = [m_{Cr}(F)\ m_{Cb}(F)]^T$ and constant *m* has a relatively small value (typically 0.4), so that the model is efficiently adapted to the selected face segment, while maintaining the general skin colour characteristics. Now let $S_i^{(k)}$ be the *i*th face segment of the *k*th image in the database, and $\mathbf{x}(S_i^{(k)}) = [m_{Cr}(S_i^{(k)}) m_{Cb}(S_i^{(k)})]^T$ the corresponding mean chrominance vector. Then the degree of similarity of segment $S_i^{(k)}$ to the selected face segment is given by

$$c(S_i^{(k)}) = \exp\left\{-\frac{1}{2}(\mathbf{x}(S_i^{(k)})-\mu_0)^T\mathbf{C}^{-1}(\mathbf{x}(S_i^{(k)})-\mu_0)\right\} \qquad (9)$$

and the colour similarity between the entire *k*th image and the input image is

$$c_k = \max_i\{c(S_i^{(k)})\} \qquad (10)$$

Retrieved images are ranked with respect to the above colour similarity measure; images/video shots containing faces with the highest similarity to that presented are depicted first. Note that direct comparison between the average chrominance components of the selected input segment and the face segments in the database could have been used; however, it is important that retrieved images contain segments whose colour still satisfies the generic skin colour model.

5.2.2. Facial Scale. This criterion is intended for users interested in retrieving images or video shots containing faces whose scale is similar to that of the face segment they present to the system. This case can be useful, for example, to detect face close-ups or people in the background. Let $A(F)$ be the percentage of the input image area covered by the selected face segment F, normalised in the range [0,1]. Then the degree of similarity of segment $S_i^{(k)}$ (the ith face segment of the kth image in the database, as in the previous case) to the selected face segment is given by

$$r(S_i^{(k)}) = 1 - |A(S_i^{(k)}) - A(F)| \tag{11}$$

and the scale similarity between the entire kth image and the input image is

$$r_k = \max_i \{r(S_i^{(k)})\} \tag{12}$$

Retrieved images or video shots are thus ranked according to the above facial scale similarity, which is also normalised in the range [0,1], while $r_k = 1$ corresponds to maximum similarity, i.e. same scale. Images/video shots containing faces with the highest similarity to that presented are depicted first.

5.2.3. Number of Face Segments. Finally, users may be interested in retrieving images or video shots where a specific number of human faces are present. In this case, images/video shots that contain the required number of face segments are retrieved from the database, while ranking of the retrieved images is performed according to either the colour similarity of the face segments to the presented example, the facial scale similarity, or a combination of both measures.

6. EXPERIMENTAL RESULTS

Three classes of experiments have been performed to demonstrate the efficiency of the proposed system. In the first, we concentrate on skin detection and evaluate the proposed skin colour modelling procedure, as well as segment isolation. In the second, face detection is examined, also including shape filtering and template matching. Finally, in the third we present content-based retrieval experiments based on the proposed indexing criteria.

6.1. Skin Detection

Skin detection is evaluated using a subset of the Compaq research database for skin colour evaluation, presented by Jones and Regh [17]. The database consists of 4670 skin images, containing skin segments and faces, and 8964 non-skin images, not containing any skin segments. All images were obtained from a parallel crawl of the world wide web, and are available in GIF and JPEG format in resolutions ranging from very low (50 × 50 pixels) to moderately high (800 × 600 pixels), and in 24 bit RGB colour resolution. In skin images, all skin regions are labelled in a semi-manual way using a software tool that enables interactive skin region segmentation by controlling a connected-components algorithm. The resulting binary mask identifying skin pixels for each skin image is included in the database.

In this experiment, we randomly selected 213 skin images along with their binary masks and 107 non-skin images. The proposed skin colour modelling and detection procedure was applied on the total of 320 images, and skin segments were isolated using the approach presented in Section 3.1. Shape filtering and face template matching is not included in the current experiment, since all kinds of skin segments are labelled. Considering the database's binary masks as the 'ground truth' for skin detection, we evaluated our results using *precision* (P) and *recall* (R) measurements, as depicted in Table 1.

Four types of measurements are actually provided in this table. In the first, precision is measured as the average ratio of the common area covered by both masks (ours and ground truth) to the area of our masks, while recall is the average ratio of the common area to the area of the ground truth masks. This measurement is denoted as TAO in Table 1 (Total Area, Original). In the second, the ground truth masks are first processed through morphological closing. This gives considerably better results for precision, since the original masks do not include details like eyes, hair and mouth openings; recall is not remarkably affected. The second measurement is denoted as TAP (Total Area, after Processing).

In the third measurement, denoted as SAT (Segment Area, with Thresholding), we first separate each one of our masks into individual connected skin segments. Each segment is considered as 'correct' if its area is covered by at least 55% by skin segments of the ground truth mask. Replacing the common area with the total area of correct segments gives a new precision measurement, while recall is calculated through a similar procedure on the ground truth masks. Due to thresholding, both precision and recall are much higher. Moreover, these measurements are more reliable, since a skin segment should be considered as correct even when its mask does not exactly match the ground truth, and incorrect even when it partially overlaps a ground truth mask. Finally, precision and recall are measured based on the number of skin segments instead of the respective areas. This is denoted as NST in Table 1 (Number of Segments, with Thresholding). Slightly worse results are obtained, since missing a very small segment is now equivalent to missing a large one.

The distribution of precision and recall measurements over the test image set are presented in the histograms of Fig. 4 for the TAP case. It can be seen that the majority of the images have high values of both measurements showing the robustness of the skin detection procedure. In Fig. 5 some examples are given that clarify the above evaluation scheme. Binary masks are only included, as we are not

Table 1. Skin detection evaluation based on the ground truth masks provided with the Compaq database.

Measurement	P (%)	R (%)
TAO	66.42	85.69
TAP	77.82	85.36
SAT	87.61	93.57
NST	85.77	88.46

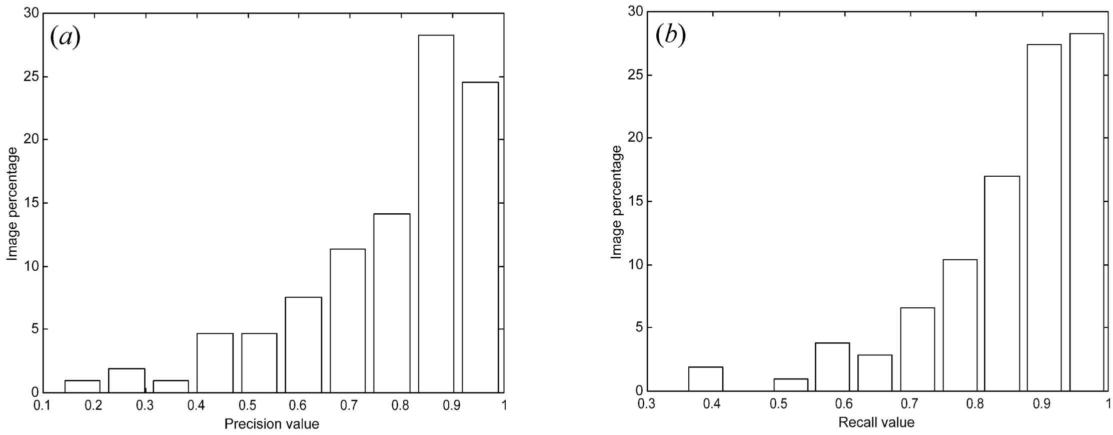
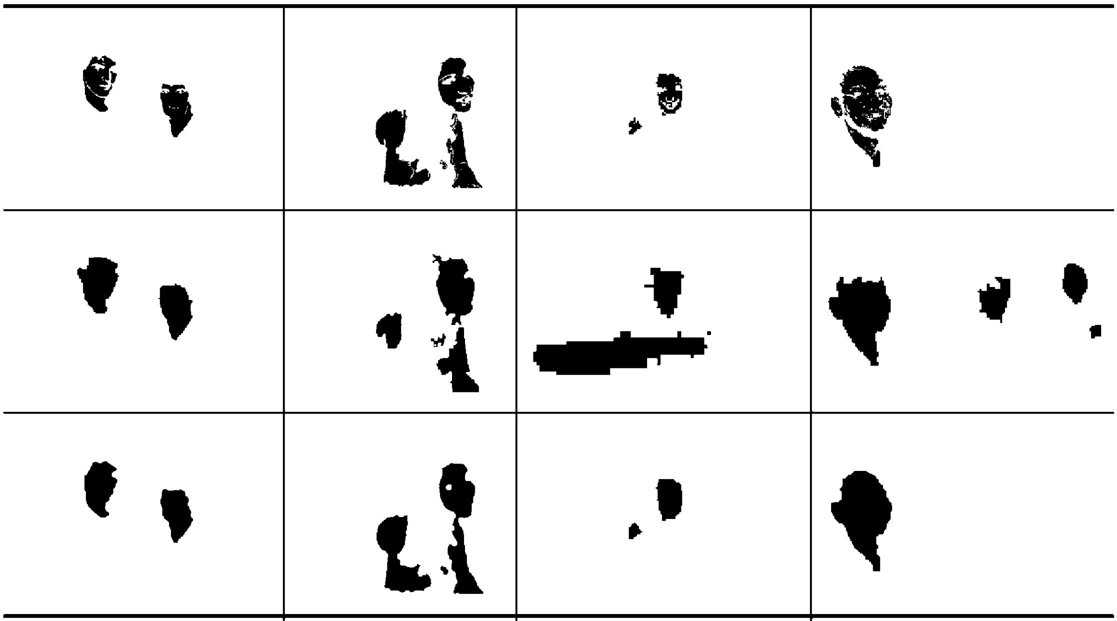


Fig. 4. Histograms of (a) precision and (b) recall values for the case of TAP measurement.



	P (%)	R (%)		P (%)	R (%)		P (%)	R (%)		P (%)	R (%)
TAO	79.76	98.14	TAO	71.53	63.76	TAO	10.54	73.20	TAO	45.01	84.56
TAP	90.23	98.18	TAP	80.03	64.38	TAP	13.37	75.21	TAP	60.09	86.49
SAT	100	100	SAT	99.26	55.71	SAT	18.40	87.50	SAT	65.46	100
NST	100	100	NST	75	66.67	NST	50	50	NST	25	100

Fig. 5. Characteristic examples of skin detection. First row: ground truth masks of the Compaq database. Second row: skin areas detected by the proposed algorithm. Third row: ground truth masks after applying morphological closing. Fourth row: precision/recall measurements.

permitted to publish original images of the Compaq database due to copyright issues. Figure 5(a) presents a typical successful case. Although a high precision value is expected, the TAO measurement is only 79.76% due to the areas of the eyes and nose, which are not included in the original ground truth masks. After applying morphological closing (TAP measurement), the precision value is increased to 90.23%. Comparison with the closed masks is more appropriate, since in our masks closing has already been applied. For the same case, when considering entire skin segments (SAT and NST cases), perfect performance is achieved. This is natural, since both ground truth skin segments have been detected.

Figure 5(b) shows an extreme case of a low recall rate; almost the entire segment of a human arm is missed due to the lighting conditions (coloured light source). Note that the SAT recall measurement is considerably lower, showing that the thresholding procedure is not always in favour of the proposed technique. Similarly, Fig. 5(c) depicts the worst case of our data set in terms of precision rate. A large object is mistakenly classified as skin due to colour similarity, introducing a false alarm. Such cases are unavoidable, and demonstrate the importance of the template matching verification stage for the purpose of face detection. Finally, Fig. 5(d) illustrates a case where the ground truth mask fails to include three skin segments – two faces and a hand – that do exist in the original image, leading to false measurements.

6.2. Face Detection

Face detection is evaluated through four different experiments. In the first two, we examine the efficiency of the face detection procedure in greyscale images and video sequences, while the last two refer to colour images and video sequences. Since the skin detection scheme is based on colour information, we do not concentrate on the greyscale case, which is only included for comparisons regarding speed.

6.2.1. Greyscale Images. In this case there is no evidence for the location of the face within the image, nor for its scale and orientation. The only way to detect the face is to apply the template matching algorithm on the whole image, resulting in increased complexity. The visual content used in this experiment consists mainly of images from the CMU expression database, which contains pictures of 20 different males and females. There are 32 different images for each person showing various expressions, looking straight to the camera, left, right, or up. The images with the highest resolution (up to 120 × 128 pixels) and straight facial orientation were used for the detection procedure. We have also included images from TV shots with a complex background to test the robustness of the algorithm.

Examples of face detection are depicted in Fig. 6, while the performance of the algorithm for a total number of 100 images in terms of *precision* and *recall* rates is depicted in the first row of Table 2. Note that precision is the ratio of correctly detected faces to the total number of ground truth faces (i.e. the opposite of the *false alarm* rate), while recall is the ratio of correctly detected faces to the total number of ground truth faces (i.e. the opposite of the *dismissal* rate). The ground truth was manually estimated on each of the 100 images. Since in all images there is only one face, we adopt as the face region the area that presents the highest similarity with the template. For this reason, the dismissal rate is the same as the false alarm rate: in five cases, the best match was found at an incorrect location, and that counted as both a dismissal and a false alarm. The *Mean Execution Time* (MET) was measured on a Pentium II/233 MHz.

Table 2. Evaluation of face detection applied in greyscale images and video sequences.

	Total images/frames	Detected faces	P/R (%)	MET (sec)
Images	100	95	95	131.7
Sequences (1st frame)	10	9	90	225.4
Sequences (subs. frames)	290	261	90	5

P: precision rate; R: recall rate; MET: mean execution time.

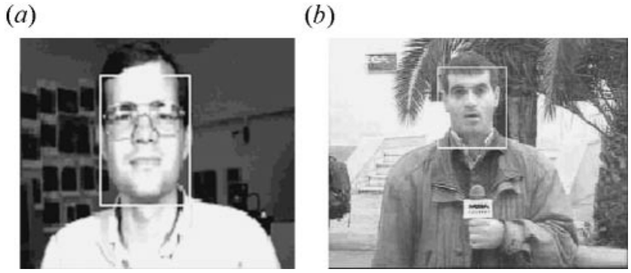


Fig. 6. Examples of face detection in greyscale images using template matching only.

6.2.2. Greyscale Video Sequences. This case is similar to the previous one, except that the template matching procedure is applied on the whole image only for the first frame of the sequence. In subsequent frames, there is enough evidence for the face scale as well as for its location and orientation. Assuming that no shot change occurs during each sequence, face *tracking* actually takes place instead of detection, speeding up computations significantly. The material of this experiment consists of broadcast TV programs recorded from three Hellenic TV channels at CIF resolution and 10 frames per second, while colour information was discarded to provide greyscale sequences.

A face detection example for two consequent frames of a news program is presented in Fig. 7, while the performance of the algorithm for a total of 300 images (10 sequences of 30 frames each), containing exactly one face each, is depicted in the last two rows of Table 2. Only one face is detected and tracked in each sequence, hence the precision rate is the same as the recall rate, as in the previous case.

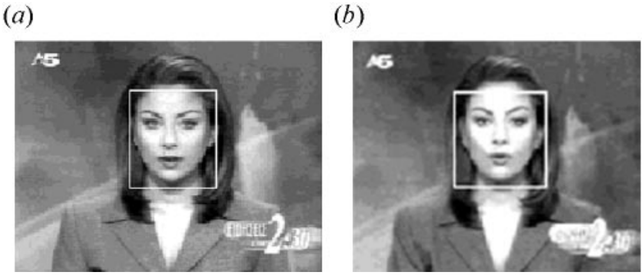


Fig. 7. Examples of face detection in greyscale sequences using template matching in the first frame and tracking in subsequent frames.

It is clearly shown that by exploiting the information gained in the first frame, the mean execution time drops about 40 times in subsequent frames. Correct detection in the first frame guarantees successful detection in subsequent frames, while failure in the first frame, leads to the opposite result, justifying the higher dismissal and false alarm rates compared to the previous experiment. In one of the sequences, the algorithm failed to detect the face in the first frame, and was unable to recover it in the following frames.

6.2.3. Colour Images. Since colour information is available in this case, the whole algorithm is applied, as described in Sections 3 through 5, and several faces can be detected in a single image, since template matching is applied in each bounding rectangle located by skin colour detection. The visual content used in this experiment consists of 100 selected frames recorded from TV shots at CIF resolution, as well as the 320 images of the Compaq database used in the experiment of Section 6.1. The image of Fig. 3 discussed in Section 4 is an example of a colour image in which three probable face segments have been located, and two of them have been verified by template matching. Note that small face segments in the background have been discarded at the stage of morphological filtering due to their small size.

The performance of the algorithm for a total of 420 images containing 409 human faces is depicted in the first row of Table 3. Since some images contain no faces and others contain several faces, precision and recall rates are different. Failure to locate all probable skin segments is the major factor for dismissals, while false alarms are due to template matching. It should be noted that in some cases, the combination of skin colour detection and template matching increases the overall detection rate. The second (right) image of Fig. 8 presents such an example. Template matching applied in the greyscale case fails to locate the strongly illuminated face; however, the outer bounding rectangle located by skin colour detection restricts the search area and prevents false alarms. Moreover, as shown in Table 3, there is a significant drop of the mean execution time for the combined algorithm.

6.2.4. Colour Video Sequences. This last case is similar to that of the static colour images. The difference is that the execution time is further reduced by performing a full search only in the first frame, while tracking is employed in subsequent frames for skin segments that were verified in the first frame using template matching. The visual content consists of 20 sequences of 30 frames each, recorded from TV news shots. It also contains two sequences of the Boston University IVC database [16], namely sequence #3 (52 frames) and #12 (49 frames). Two representative frames of those sequences are shown in Fig. 9, along with the ground truth skin masks (included in the IVC database) and the detected ones. The performance of the algorithm for the total of 701 frames, containing 736 faces, is depicted in the second row of Table 3. Since the number of faces varies in each frame, the precision and recall rates are different, as in the previous experiment. Both, however, are close to those obtained in the colour image experiment. Exploitation of the temporal correlation between the frames of the sequence results in further reduction of the mean execution time.

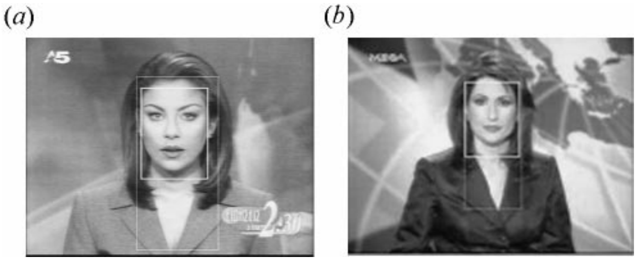


Fig. 8. Examples of face detection in colour images using skin colour detection combined with template matching. The outer rectangles are located by skin colour detection; the inner rectangles are detected by template matching.

6.3. Content-Based Retrieval

Content-based retrieval using the proposed indexing criteria has been evaluated on a video database. Its content was created using a total of 200 video clips of 120–850 frames each, recorded from news programs of five Hellenic TV channels at CIF resolution and 10 frames per second. Of the 10,850 total video frames, approximately 28% contain no faces (or faces at very small scale), 49% contain exactly one face, while 23% contain two or more faces. The whole face detection procedure, comprising skin colour detection, morphological processing, shape filtering and template matching, was applied off-line to the entire database; colour, scale and shape features of the face segments detected were then stored as indexing information in the database.

Table 3. Evaluation of face detection applied in colour images and video sequences.

	Total images/frames	Images with faces	Total faces	Detected faces	Correctly detected faces	P (%)	R (%)	MET (sec)
Images	420	305	409	379	368	97.10	89.98	1.9
Sequences	701	643	736	681	667	97.94	90.63	1.5

P: precision rate; R: recall rate; MET: mean execution time.

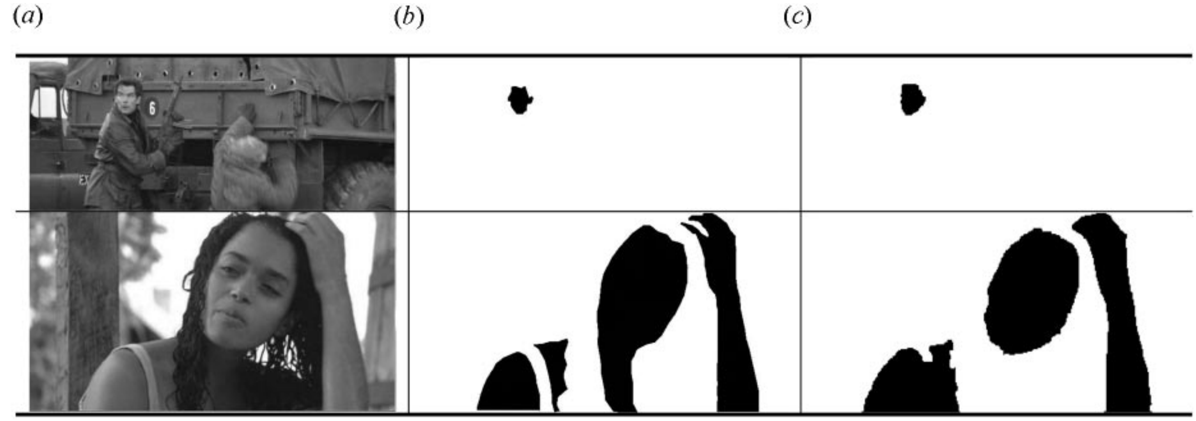


Fig. 9. Example of skin detection in video sequences. (a) Two representative frames from the IVC database, (b) ground truth skin masks, and (c) areas detected by the proposed algorithm.

Content-based retrieval was evaluated in a query-by-example framework. Input images presented to the system were analysed in exactly the same way using the proposed scheme, and features of the face segments detected were compared to those of the video clips existing in the database, according to the three different proposed indexing criteria. For example, Fig. 10 depicts retrieval results based on colour similarity, according to the evaluation protocol described in Section 5. The single face segment present in the input image is correctly detected, and its chrominance colour components are compared to those available in the database, retrieving the following eight video frames presented by decreasing colour similarity. It is observed that frames containing faces of a similar colour are retrieved, regardless of face size or number of faces.

Figure 11 presents retrieval based on the facial scale criterion. The main face segment is again extracted from the input image, but now similarity is measured in terms of facial scale, i.e. the percentage of the frame area that the corresponding face segments cover. Retrieval is now independent of face colour, and results are ranked according to decreasing facial scale similarity. Finally, Fig. 12 illustrates retrieval based on the number of face segments. Since two face segments are detected in the input image, the database is searched for frames containing two human faces; however, similarity is measured in terms of facial scale in this case.

Although the results presented are only qualitative, and it is not straightforward to produce a numerical evaluation, mainly due to the subjective nature of the problem, it can be claimed that a high retrieval efficiency has been achieved. It is clear that it would be hard to achieve such results employing only generic image features such as colour composition or object shape and texture, without knowledge of the human facial structure. More reliable experiments that take human perception into account can be carried out by having a team of humans evaluate the relevance of the retrieved images to the queries posed.

In the special case of queries based on the number of faces, numerical evaluation has been performed on a subset of 500 frames of the video database. Those frames have been manually classified into four classes, according to the number of faces they contain, as shown in Table 4. Four corresponding queries were executed, and the retrieved images were labelled as correct if they contained exactly the same number of faces as requested. Note that a similar numerical evaluation for retrieval based on colour and scale similarity is not straightforward, as in these cases ground truth is not easily defined.

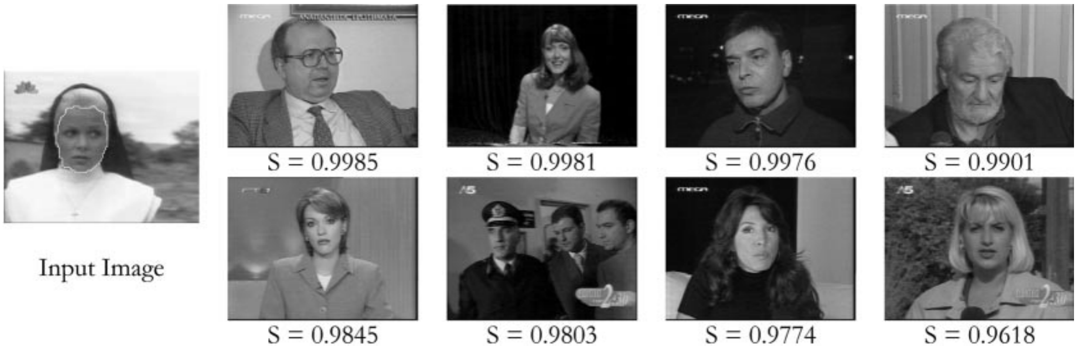


Fig. 10. Retrieval based on facial colour (S: colour similarity).

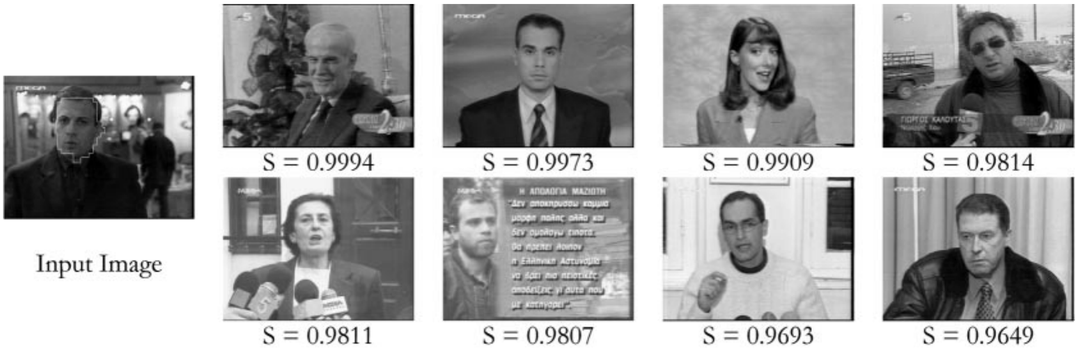


Fig. 11. Facial scale based retrieval (S: scale similarity).

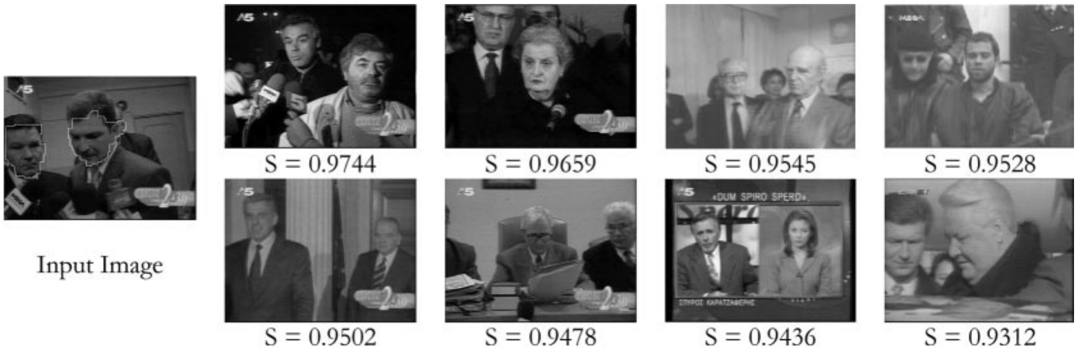


Fig. 12. Retrieval based on required face segments (S: scale similarity).

Table 4. Numerical evaluation of retrieval based on number of segments.

	Total images	Retrieved	Correctly retrieved	P (%)	R (%)
No faces	62	66	59	89.34	95.16
1 face	275	282	273	96.81	99.27
2 faces	112	106	101	93.57	87.61
3 faces	51	46	42	91.30	82.35

It can be observed from Table 4 that the best results are obtained for the case of one face; this is expected, since a single face is usually shown in high spatial resolution and is easier to detect. Moreover, many such frames present frontal views of newscasters in indoor recordings, where lighting conditions are controlled. Decreased resolution is the main factor of lower recall rates in the cases of two or three faces, as it leads to an increased number of face dismissals. The precision corresponding to the 'no faces' class is lower for the same reason.

7. CONCLUSION

It is shown in the experiments that, apart from automating the process of human face indexing, the use of the derived facial features in multimedia databases can also lead to new types of searching criteria in a query-by-example framework. Three such criteria are proposed, namely, facial colour similarity, desired number of faces and facial scale, with promising retrieval results. These criteria can also be combined with existing ones based on generic colour, motion, shape or texture features, resulting in a variety of retrieval scenarios. The proposed framework thus provides a powerful tool that can be embedded in existing content-based indexing and retrieval systems.

The face detection approach presented in this study can also be used in a variety of multimedia applications, such video partitioning, browsing and summarisation, indexing of TV news, interactive content-based retrieval and multimedia database management. The use of template matching significantly improves its performance at the cost of increased computational complexity. Nevertheless, the latter is reduced up to 100 times, since only those image/frame areas that belong to skin segments are verified. Skin colour modelling is re-estimated in a per frame basis, providing the necessary generalisation ability to efficiently mask skin areas. An overall experimental evaluation suggests that the proposed method successfully tackles the trade-off between speed and efficiency for face detection.

Acknowledgements

This work has been partially funded by the project PHYSTA of the Training Mobility and Research Program of the European Community (1998–2001), dealing with emotion analysis based on speech, facial and psychological cues. The authors would like to thank Dr Michael Jones for giving us permission to use the image database of the Compaq Cambridge Research Lab.

References

1. Sikora T. The MPEG-4 Video Standard Verification Model. IEEE Trans Circuits and Systems for Video Technology 1997; 7(1): 19–31
2. ISO/IEC JTC1/SC29/WG11. MPEG-7: Context and Objectives (v.5). Doc. N1920, October 1997
3. Androutsos D, Plataniotis KN, Venetsanopoulos AN. Extraction of detailed image regions for content-based image retrieval. Proc IEEE ICASSP, Seattle, WA, May 1998
4. Special Issue on Segmentation, Description and Retrieval of Video Content. IEEE Trans Circuits and Systems for Video Technology 1998; 8(5)
5. Alatan A, Onural L, Wollborn M, Mech R, Tuncel E, Sikora T. Image sequence analysis for emerging interactive multimedia services – the European Cost 211 framework. IEEE Trans Circuits and Systems for Video Technology 1998; 8(7): 802–813
6. Avrithis Y, Doulamis A, Doulamis N, Kollias S. A stochastic framework for optimal key frame extraction from MPEG video databases. Computer Vision and Image Understanding 1999; 75(1/2): 3–24
7. ISO/IEC JTC1/SC29/WG11. MPEG-7 Overview (v. 1.0). Doc. N3158, December 1999
8. Avrithis Y, Tsapatsoulis N, Kollias S. Broadcast news parsing using visual cues: a robust face detection approach. Proc IEEE Int Conf on Multimedia and Expo (ICME), New York, NY, July 2000
9. Samal A, Iyengar PA. Automatic recognition and analysis of human faces and facial expressions: a survey. Pattern Recognition 1992; 25(1): 65–77
10. Yang G, Huang TS. Human face detection in complex background. Pattern Recognition 1994; 27(1): 55–63
11. Yow KC, Cipolla C. Feature-based human face detection in complex background. Image or Vision Computing 1997; 15(9): 713–735
12. Wang H, Chang S-F. A highly efficient system for automatic face region detection in MPEG video. IEEE Trans Circuits and Systems for Video Technology 1997; 7(4): 615–628
13. Garcia C, Tziritas G. Face detection using quantized skin color regions merging and wavelet packet analysis. IEEE Trans Multimedia 1999; 1(3): 264–277
14. Avrithis Y, Tsapatsoulis N, Kollias S. Color-based retrieval of facial images. Proc EUSIPCO, Tampere, Finland, September 2000
15. Raja Y, McKenna SJ, Gong S. Tracking and segmenting people in varying lighting conditions using color. Proc 3rd Int Conf on Automatic Face and Gesture Recognition, Nara, Japan, April 1998
16. Sigal L, Sclaroff S, Athitsos V. Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. Proc Int Conf on Computer Vision and Pattern Recognition (CVPR), Hilton Head Island, SC, June 2000
17. Jones MJ, Regh MR. Statistical color models with application to skin detection. Compaq Cambridge Research Lab Technical Report CRL 98/11, 1998
18. Sun QB, Huang WM, Wu JK. Face detection based on color and local symmetry information. Proc 3rd Int Conf on Automatic Face and Gesture Recognition, Nara, Japan, April 1998
19. Tsapatsoulis N, Avrithis Y, Kollias S. Face detection for multimedia applications. Proc IEEE ICIP, Vancouver, BC, Canada, September 2000
20. Rzeszewski T. A novel automatic hue control system. IEEE Trans Consumer Electronics 1975; 21(2): 155–163
21. Harwood LA. A chrominance demodulator IC with dynamic flesh correction. IEEE Trans Consumer Electronics 1976; 22(2): 111–117
22. Tsapatsoulis N, Doulamis N, Doulamis A, Kollias S. Face extraction from non-uniform background and recognition in compressed domain. Proc IEEE ICASSP, Seattle, WA, May 1998
23. Maragos P. Morphological signal and image processing. Madisetti V, Williams D (eds) The Digital Signal Processing Handbook. CRC Press, 1997
24. Jain AK, Zhong Y, Lakshmanan S. Object matching using deformable templates. IEEE Trans Pattern Analysis and Machine Intelligence 1996; 18(3): 267–278
25. Bimbo AD, Pala P. Visual image retrieval by elastic matching of user sketches. IEEE Trans Pattern Analysis and Machine Intelligence 1997; 19(2): 121–132
26. Avrithis Y, Xirouhakis Y, Kollias S. Affine-invariant curve normalization for shape-based retrieval. Proc Int Conf on Pattern Recognition (ICPR), Barcelona, Spain, September 2000
27. Tsapatsoulis N, Avrithis Y, Kollias S. On the use of radon transform for facial expression recognition. Proc Int Conf on Information Systems Analysis and Synthesis (ISAS), Orlando, FL, August 1999

Nicolas Tsapatsoulis was born in Limassol, Cyprus in 1969. He graduated from the Department of Electrical and Computer Engineering, the National Technical University of Athens in 1994, and received his PhD degree in 2000 from the same university. His current research interests lie in the areas of machine vision, image and video processing, neural networks and biomedical engineering. He is a member of the Technical Chambers of Greece and Cyprus and a member of IEEE Signal Processing and Computer societies. Dr Tsapatsoulis has published six papers in international journals and 23 in proceedings of international conferences. Since 1995 he has participated in seven research projects at Greek and European level.

Yannis S. Avrithis was born in Athens, Greece in 1970. He received the Diploma degree in Electrical and Computer Engineering from the National Technical University of Athens (NTUA), Greece, in 1993, the MSc degree in Electrical and Electronic engineering (Communications and Signal Processing) from the Imperial College of Science, Technology and Medicine, London, UK, in 1994, and the PhD degree in Electrical and Computer Engineering from NTUA in 2001. His research interests include digital image and video processing, image segmentation, affine-invariant image representation, content-based indexing and retrieval, and video summarisation. He has published six articles in international journals and books, and 17 in proceedings of international conferences.

Stefanos D. Kollias was born in Athens, Greece in 1956. He received the

Diploma degree in Electrical Engineering from the National Technical University of Athens (NTUA) in 1979, the MSc degree in Communication Engineering from the University of Manchester (UMIST), England in 1980, and the PhD degree in Signal Processing from the Computer Science Division of NTUA in 1984. In 1982 he received a ComSoc Scholarship from the IEEE Communications Society. Since 1986, he has been with the NTUA, where he is currently a Professor. From 1987 to 1988 he was a Visiting Research Scientist in the Department of Electrical Engineering and the Center for Telecommunications Research, Columbia University, NY, USA. Current research interests include image processing and analysis, neural networks, image and video coding and multimedia systems. Stefanos Kollias is the author of more than 140 articles in the aforementioned areas.

Correspondence and offprint requests to: N. Tsapatsoulis, Image, Video and Multimedia Systems Laboratory, Department of Electrical and Computer Engineering, National Technical University of Athens, Heroon Polytechniou 9, 157 73 Zographou, Greece. E-mail: ntsap@image.ntua.gr