ELSEVIER

Signal Processing: *Image Communication* 14 (1999) 359–388

SIGNAL PROCESSING: IMAGE COMMUNICATION

Automatic location and tracking of the facial region in color video sequences

N. Herodotou[a,*], K.N. Plataniotis[b], A.N. Venetsanopoulos[a]

[a] *Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Rd, Toronto, Canada, Ont. M5S 3G4*
[b] *Department of Mathematics, Physics & Computer Science, Ryerson Polytechnic University, 350 Victoria Street, Toronto, Canada, Ont. M5B 2K3*

Received 6 June 1997

Abstract

A novel technique is introduced to locate and track the facial area in videophone-type sequences. The proposed method essentially consists of two components: (i) a color processing unit, and (ii) a knowledge-based shape and color analysis module. The color processing component utilizes the distribution of skin-tones in the HSV color space to obtain an initial set of candidate regions or objects. The second component in the segmentation scheme, that is, the shape and color analysis module is used to correctly identify and select the facial region in the case where more than one object has been extracted. A number of fuzzy membership functions are devised to provide information about each object's shape, orientation, location and average hue. An aggregation operator finally combines these measures and correctly selects the facial area. The suggested approach is robust with regard to different skin types, and various types of object or background motion within the scene. Furthermore, the algorithm can be implemented at a low computational complexity due to the binary nature of the operations involved. Experimental results are presented for a series of CIF and QCIF video sequences.

Keywords: Face localization and tracking; Color image segmentation; Video processing; Facial features; Fuzzy membership functions

1. Introduction

Digital video is an integral part of many newly emerging multimedia applications. Recent advances in the area of mobile communications and the tremendous growth of the *Internet* have placed even greater demands on the need for more effective video coding schemes. However, future coding techniques must focus on providing better ways to represent, integrate and exchange visual information in addition to efficient compression methods. These efforts aim to provide the user with greater flexibility for 'content-based' access and manipulation of multimedia data. Numerous video applications such as portable videophones, videoconferencing, multimedia databases, and video on demand can greatly benefit from better compression schemes and this added 'content-based' functionality.

* Corresponding author. Fax: +1 416 978 4425; e-mail: nicos@dsp.toronto.edu

0923-5965/99/$ – see front matter
PII: S0923-5965(98)00018-6

International video coding standards such as H.261, and more recently recommendation H.263, are widely used for very low bit-rate applications such as those described above. These existing standards including MPEG 1 and 2 are all based on the same framework, that is, they employ a block-based motion compensation scheme and the discrete cosine transform for intra-frame encoding. However, this block-based approach introduces blocking artifacts and motion 'jerkiness' in the reconstructed sequences. Furthermore, the existing standards deal with video exclusively at the frame level, thereby preventing the manipulation of individual objects within the bitstream. Second generation coding algorithms have focused on representing a scene in terms of 'objects' rather than square blocks [15,10]. This approach not only improves the coding efficiency and alleviates the blocking artifacts, but it can also support the content-based functionalities mentioned previously by allowing interactivity and manipulation of specific objects within the video stream. These are some of the objectives and issues addressed within the framework of the MPEG 4 and future MPEG 7 standards [4].

In order to obtain an object-based representation, an input video sequence must first be segmented into an appropriate set of arbitrarily shaped regions (termed the *video object planes* in the *MPEG 4 verification model*), where each of the regions may represent a particular content of the video stream [25]. The features of each 'object' such as shape, motion, and texture information can subsequently be coded into the so called *video object layer* for transmission or storage. Thus, the success of any object-based approach depends largely on the segmentation of the scene based on its image contents. In a videophone-type application for example, an accurate segmentation of the facial region can serve two purposes: (1) it can allow the encoder to place more emphasis on the facial region since this area (i.e. the eyes and mouth in particular) is the focus of attention to the human visual system of an observer, and (2) it can also be used to extract features so that higher-level descriptions can be generated (i.e. personal characteristics, facial expressions, and composition information). In a similar fashion, the contents within a video database can be segmented into individual objects, where the following features can be supported: (1) sophisticated query and retrieval functions, (2) advanced editing and compositing, and (3) better compression ratios.

In this paper, we focus on the automatic location and tracking of the facial region of a head-and-shoulders videophone-type sequence using color and shape information. The method we present utilizes the skin-tone distribution of the histograms in the HSV color space to initially extract the facial region. The segmentation results are then refined using a series of post-processing operations which include median filtering, region filling and removal, and morphological opening and closing operations. A series of fuzzy membership functions are finally used to correctly classify and retain the facial area in the case of additional falsely included regions. The feature vector obtained from this last step can be used to augment a further feature extraction stage which can support the aforementioned 'content-based' functionalities. Our approach is robust with regards to facial shape, size, skin color, orientation, motion, and lighting conditions. Furthermore, it can be implemented at a relatively low computational complexity due to the binary nature of the operations performed.

The organization of the paper is as follows. In Section 2 the color attribute is investigated as one of the two visual cues to be used in the segmentation process. The distributions of various skin-types are first examined within the framework of the HSV color space. The technique utilized to extract the skin-tone clusters within an image is later introduced. A series of post-processing operations used to refine the shape of the facial region are then discussed in the last part of the section. In Section 3 the second part of the localization scheme is described which consists of the shape and color analysis module. More specifically, the shape attribute is discussed along with the fuzzy membership functions used to form this knowledge-based decision module. Aggregation operators used to select the facial region from the set of candidate objects, are finally examined at the end of the section. In Section 4, experimental results are presented and analyzed for several videophone-type sequences. Finally, in Section 5 the conclusions are drawn up.

2. Color image segmentation

2.1. Motivation and related work

The recognition of human faces is currently an active area of research in computer vision [1,13,16,31]. The task of recognizing human faces is essentially a two-step process: (1) the detection and automatic location of the human face, and (2) the automatic identification of the face based on the extracted features. Most of the research to date has been directed towards the latter identification phase, with less emphasis being placed on the initial localization stage. However, the first step is critical to the success of the second and the overall recognition system. Thus, the importance of obtaining an accurate localization of the face is clear and vital in numerous applications including human recognition for security purposes, human-computer interfaces and, more recently, for video coding, video databases, and video on demand. Nevertheless, determining the location of a face of unknown size, in a scene with a complex or moving background still remains a difficult problem that is relatively unexplored.

Several techniques based on shape and motion information have been proposed recently for the automatic location of the facial region [5,14,22]. The former two are related to video coding applications while the latter is part of a facial recognition system. The shape-based approach in [5] models the contours of the face as an ellipse. The location of the facial region is determined by performing an ellipse fitting task to a thresholded binary edge image. In [22], a generic 3-D face model is adapted to the extracted facial outline from a videophone type scene for the case where only one person is talking against a stationary background. In this application, a hierarchical localization scheme is utilized to isolate the facial area. The technique is based on the shape of the extracted head-and-shoulders silhouette which is obtained using the thresholded frame differences. Finally, in [14], a motion detection algorithm is used to segment the facial area from a complex background. The proposed method locates the facial region by assuming that the object having the greatest motion in the video sequence is the face to be detected. This assumption, however, may limit the success of the approach in applications with non-stationary backgrounds (i.e. mobile videophones) and/or other moving objects in the scene. The authors also acknowledge potential problems caused by noise or other objects moving in the background and also suggest a modification in their technique to better handle the case of tilted or turned faces.

Color is a key feature used to understand and recollect the contents within a scene. It is found to be a highly reliable attribute for image retrieval as it is generally invariant to translation, rotation, and scale changes [12]. In our approach we use color as the primary tool in detecting and locating the facial region in a scene with a complex or moving background. In certain cases, however, the use of a single image attribute such as color may lead to additional falsely detected objects. This situation may occur when other objects in the scene have colors similar to those of skin tone regions. In these cases, a feature vector based on a number of shape characteristics is constructed from a series of fuzzy membership functions to provide the necessary discriminatory information. The aggregation of these features within the framework of a knowledge-based decision system provides the mechanism of selecting the facial area from the set of candidate regions.

The segmentation of a color image is the process of classifying the pixels within the image into a set of clusters with a uniform color characteristic. The objective in our approach is to detect and isolate the color clusters that correspond to the skin areas of the facial region. However, the shape or distribution of the clusters that form depend on the chosen color space [29]. Therefore, the most advantageous color space must first be selected (i.e. one which produces distinct and clearly separated clusters) in order to obtain the most effective results in the segmentation process. In [14], the color distribution of a facial image was examined using four different color coordinate systems, which included the RGB space, HSI, CIE $L^*u^*v^*$, and the Karhunen–Loeve transformation. The HSI color space was found to be the most suitable as it produced clusters that were clearly separated, allowing them to be detected and readily extracted. The other three spaces showed ambiguity in the partitioning

of these clusters. We have found similar results by examining the RGB, HSV and the L*a*b* color spaces. The HSV space (similar to HSI) appeared to be the most advantageous due to the distribution of the clusters formed. The perceptually uniform L*a*b* space did not exhibit a global compactness in the different skin clusters, making it difficult to derive a uniform distance metric for segmentation purposes. The set of transformation equations that relate the different color coordinate systems can be found in [21]. A more detailed account of the selected HSV space is presented below.

2.2. HSV color space

Color information is commonly represented in the widely used RGB coordinate system. This basis is hardware oriented and is suitable for acquisition or display devices but not particularly applicable in describing the perception of colors. On the other hand, the HSV (hue, saturation, value) color model corresponds more closely to the human perception of color. The HSV color space is conveniently represented by the hexcone model shown in Fig. 1. The *hue* (H) is measured by the angle around the vertical axis and has a range of values between 0 and 360 degrees beginning with red at 0°. It gives us a measure of the spectral composition of a color. The saturation (S) is a ratio that ranges from 0 (i.e. on the V axis), extending radially outwards to a maximum value of 1 on the triangular sides of the hexcone. This component refers to the proportion of pure light of the dominant wavelength and indicates how far a color is from a gray of equal brightness. The value (V) also ranges between 0 and 1 and is a measure of the relative brightness. At the origin, $V = 0$ and this point corresponds to 'black'. At this particular value, both H and S are undefined and meaningless. As we traverse upwards along the V axis we perceive different shades of gray until the endpoint is reached (where $V = 1$ and $S = 0$) which is considered to be 'white'. At any point along the V axis the saturation component is zero and the hue is undefined. This singularity occurs whenever $R = G = B$. The set of equations below can be used to transform a point in the RGB coordinate system to the appropriate value in the HSV space.

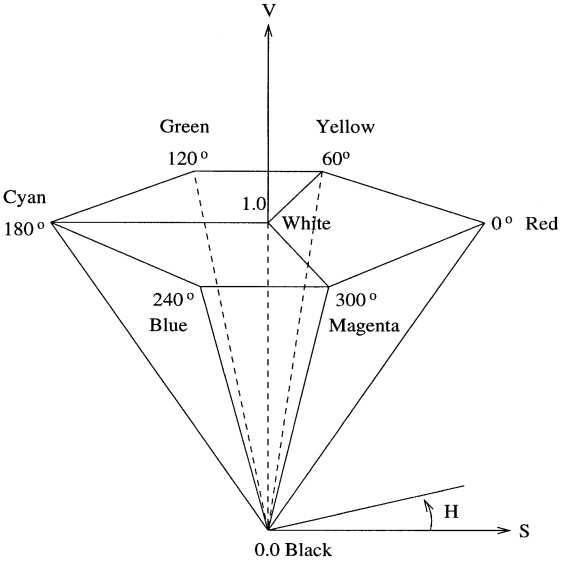


Fig. 1. HSV hexcone color model.

$$H_1 = \cos^{-1}\left\{\frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right\}, \quad (1)$$

$$H = H_1 \quad \text{if } B \leqslant G, \quad (2)$$

$$H = 360^\circ - H_1 \quad \text{if } B > G, \quad (3)$$

$$S = \frac{\text{Max}(R,G,B) - \text{Min}(R,G,B)}{\text{Max}(R,G,B)}, \quad (4)$$

$$V = \frac{\text{Max}(R,G,B)}{255}. \quad (5)$$

In the expressions above, the Max and Min operators select the maximum and minimum values of the operand, respectively, and R, G and B range between 0 and 255. A fast algorithm used here to convert the set of RGB values to the HSV color space is provided in [7]. The HSI (hue, saturation, intensity) color space mentioned earlier, is analogous to the HSV model and is conveniently represented by the biconical color solid [11]. In this biconical model, the effectiveness of the hue is limited at both, the low and high Intensity values [28]. This additional limitation at the high end may affect the robustness of our proposed hue-based

color segmentation approach and, thus, the HSV model is selected. The computational complexity of the HSV transformation equations are also advantageous over the corresponding trigonometric HSI relations.

Having defined the selected HSV color space, we must subsequently devise a technique to determine and extract the color clusters that correspond to the facial skin regions. This requires an understanding of where these clusters form in the space just outlined above. We examine the distribution of these clusters next.

2.3. Skin-tone distribution

Human skin is composed of several layers of tissue which consist essentially of blood cells, and a yellow pigment called melanin [6]. The appearance of the skin is affected by a number of factors which include the degree of pigmentation (varies amongst individuals and different races), the concentration of blood, and the incident light source. The combination of all of these factors give rise to a variation in skin color which spans over the range of red, yellow and brownish-black. Nevertheless, this corresponds to a restricted range of hue values as will be shown below. In [27], a hue range that is representative of skin regions has also been proposed.

A large sample of head-and-shoulders type images were collected to observe the distribution of skin colors in the HSV color space. The test images contained several MPEG 4 test sequences (i.e. well-behaved lighting), as well as numerous still images obtained from the Internet that contained random lighting conditions (i.e. poor and well lit). The test set consisted of facial images from different races, in order to model a wide range of skin colors. These included Caucasian, Asian and African–American skin-types. Over three hundred images were used as the training set for each category so that an adequate sample set could be obtained at a suitably feasible complexity. The following scheme was used to generate the histograms for the H, S and V components of each category. The facial skin region was manually selected in each sample image, and the H, S and V values were determined for each pixel within this area. The histograms were subsequently formed by compiling the results from all of the images within each category. The normalized histograms obtained from this procedure are shown in Fig. 2. It is clear that in all three categories the hue component consists of a limited range of values. The hue values of Caucasian and Asian samples fall predominantly between 0° (Red) and 60° (Yellow) while those of African–American are shifted closer towards 0° with a small portion of the distribution in the red–magenta hue sector. One may also note that the hue values between 180° and 360° can be represented by their equivalent negative values (i.e. $340^\circ = -20^\circ$). In the figures, the Saturation component ranges from about 10 to 100% in all cases, with the majority falling in the 20–60% range. This suggests that the skin colors for all races are somewhat saturated but not deeply saturated. Finally, we see in Fig. 2 that the Value or brightness component for both Caucasian and Asian distributions ranges from approximately 40% to the maximum value of 100%. The Asian test images are shifted even more so towards the maximum value of V (i.e. top of the hexcone model) signifying a high level of brightness in the facial skin regions of these samples. The African–American test set on the other hand, has a wider value range but is shifted towards lower values. The mean, m, and standard deviation, σ (both given in degrees), of the three hue distributions are conveniently summarized in Table 1. The tabulated values indicate that the Asian test samples have the highest mean value of the three distributions, $m = 28.9^\circ$ (i.e. greater shift towards Yellow) with the lowest standard deviation, σ. The Caucasian sample set has similar statistics with a slightly smaller mean value, $m = 25.3^\circ$ and a slightly larger value of σ. The African–American distribution has the smallest mean value of the three, $m = 8.6^\circ$ (shift towards red) and the largest

Table 1
Statistics of the hue distribution categorized by race

Caucasian		African–American		Asian	
m (°)	σ (°)	m	σ	m	σ
25.3	6.8	8.6	8.2	28.9	5.1

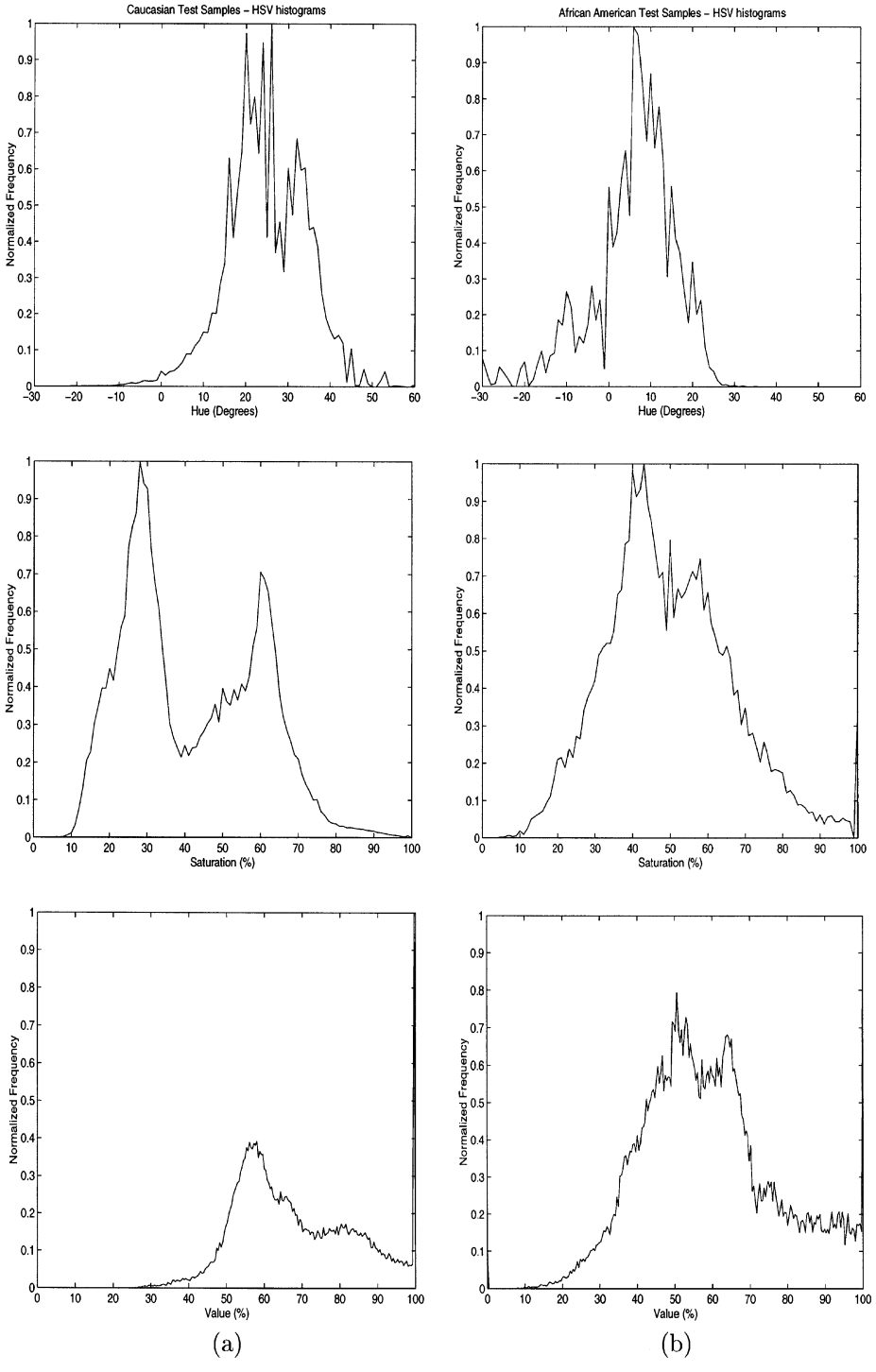
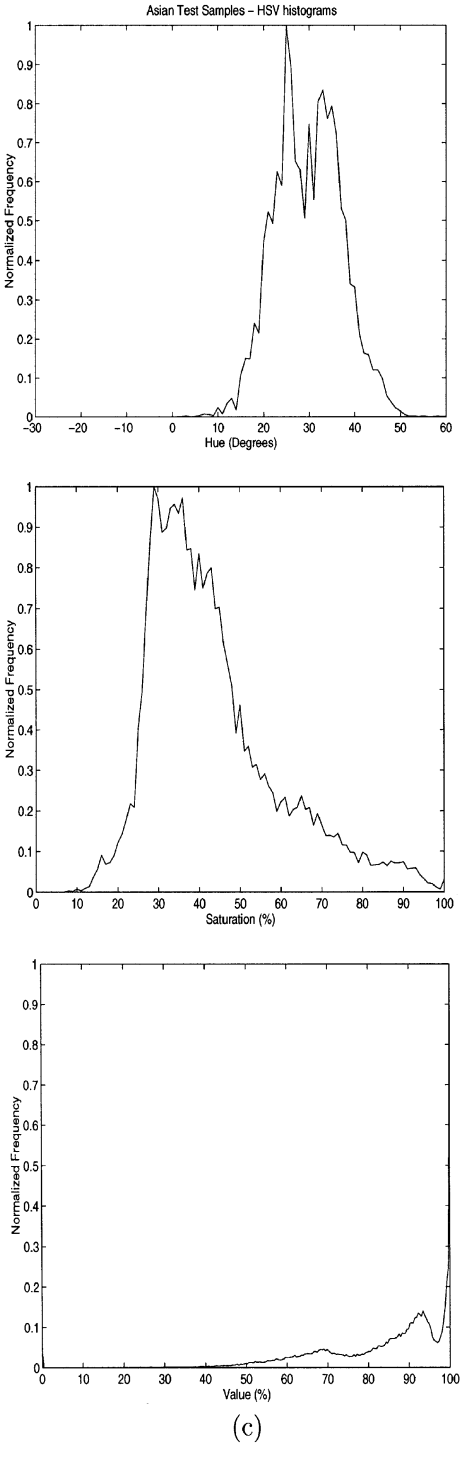


Fig. 2. Skin color distributions of different races in the HSV space: (a) Caucasian, (b) African–American and (c) Asian test samples.



(c)

Fig. 2. Continued.

standard deviation. The large value in σ can be attributed to the variation in skin colors within the African–American sample set.

Having obtained the distribution for this wide range of skin colors, we must devise an appropriate scheme to segment the facial skin area in any given image. We outline the proposed technique below.

2.4. Extraction of skin-tone regions

The basis of segmenting an image by color lies in the extraction of a set of regions that satisfy some homogeneity criterion using the spectral components of the image. The approach in any technique depends on the way these regions to be extracted are defined and formed. Four fundamental approaches can be identified and are categorized as follows: (1) pixel-based techniques, (2) area-based methods, (3) edge-based schemes, and (4) physics-based vision models. In the first of these techniques, the regions to be segmented are determined by operating directly in the color space domain. The set of pixels that form each region are determined by a class membership function which is defined in the selected color space. Histogram-based techniques, clustering and fuzzy clustering methods all fall into this first category. In the area-based schemes of the second category, the regions of uniformity are determined by operating spatially in the image domain. Region growing, and split and merge algorithms belong to this class of techniques. In edge-based segmentation, a color contour is created by connecting a set of edge pixels determined by various color edge detectors. Finally, the fourth category belongs to a relatively new class of computer vision methods which employ physical models to partition the image. The aim in this latter approach is to segment the image at the object boundaries rather than the edges of highlights or shadows of the image. An extensive survey of the various techniques can be found in [26].

The method we propose here falls into the first of the four categories described above. The objective in pixel-based segmentation techniques is to partition or divide the color space rather than segment the spatial domain of the image. In histogram-based

approaches this partitioning can be accomplished by determining the significant peaks and valleys of the computed histograms and setting the thresholds accordingly. A variety of multi histogram-based thresholding schemes have been suggested to divide multichannel data as in color images [17,18]. Alternatively, the color space can be divided by using a technique known as clustering [2,29]. In this scheme, the partitioning is a function of the input vectors (i.e. vector values of the color pixels) and is based on a criterion of optimality such as the least sum of squares. This is closely related to the vector quantization problem of mapping the set of input vectors to a finite number of weight vectors which form the Voronoi tessellation. The computational complexity of these latter techniques can become quite demanding. Either of the two approaches just described can be utilized as general purpose segmentation schemes. However, a scene that consists of an unknown number of homogeneous regions or objects is, in general, very difficult to segment. In many cases, the techniques involve some human interaction in which certain thresholds are manually selected or where assumptions are made regarding the number of distinct regions or clusters in the scene. In our particular application, we utilize the apriori knowledge of the skin-tone distributions found previously to identify and extract the facial skin regions. A polyhedron is defined in the HSV hexcone model which contains the skin-colored clusters. The proper selection of this polyhedron is the key to obtaining successful segmentation results.

The hue component is the most significant feature in defining the desired polyhedron. The histograms of Fig. 2 indicate that the hue values can be represented by a limited range 340–360° (magenta–red) and 0–50° (red–yellow) for all skin types. This range is very effective in extracting skin colored regions under higher levels of illumination and sufficiently saturated colors. However, the hue can be unreliable when the following two conditions arise: (1) when the level of brightness (i.e. value) in the scene is low, or (2) when the regions under consideration have low saturation values. The first condition can occur in areas of the image where there are shadows or, generally, under low lighting levels. In the second case, low values of saturation correspond to achromatic regions. As mentioned previously, saturation values of zero lie on the V axis in the hexcone model and appear as gray areas. Many objects, by nature, are achromatic (i.e. white clouds, gray asphalt roads, etc.), however, shadows or conditions of non-uniform illumination (i.e. specular reflection) can cause chromatic regions such as skin areas to appear achromatic. Thus, we must define thresholds for the value and saturation components where the hue attribute is reliable. Incidentally, this will also define the desired polyhedron.

The HSV hexcone model of Fig. 1 and the distributions of Fig. 2 were used in the threshold selection process. A lower bound threshold of $T_{\mathrm{val}} = 35\%$ was chosen for the value component. Pixels with values less than T_{val} were not considered in the segmentation process as the hue becomes unreliable for values below this threshold. This can be seen visually by observing the hexcone model as the value component is varied. Fig. 3 illustrates four different cases: (1) when the brightness value is at its maximum, $V = 100\%$, (2) at $V = 63\%$, (3) at the threshold value, where $V = T_{\mathrm{val}} = 35\%$, and (4) below the threshold value at $V = 20\%$. The figure gives us an indication of the discriminatory power of the hue component at four different slices (i.e. hexagons) of the hexcone model. Radial supersets of the hexagons are shown in the figure for the sake of simplicity. In each circular plot, the saturation varies radially from 0%, at the center of each circle, to 100%, at the outer edges of the circle. The effectiveness of the hue is evident in parts (a) and (b) where the value is at its maximum, and at $V = 63\%$, respectively. Part (d) clearly illustrates that the hue is meaningless when the brightness in the scene is low. On the other hand, the threshold value of $V = 35\%$ in part (c) is a break point where the hue component starts to become ineffective. Experimental results also indicated that the selection of a lower threshold led to erroneously detected regions. The importance of intensity information for color image segmentation has also been emphasized in [8,9,29].

A saturation threshold, T_{sat}, is also very important in obtaining reliable segmentation results. We have found that the hue is reliable when the saturation is greater than 20% and meaningless when it is

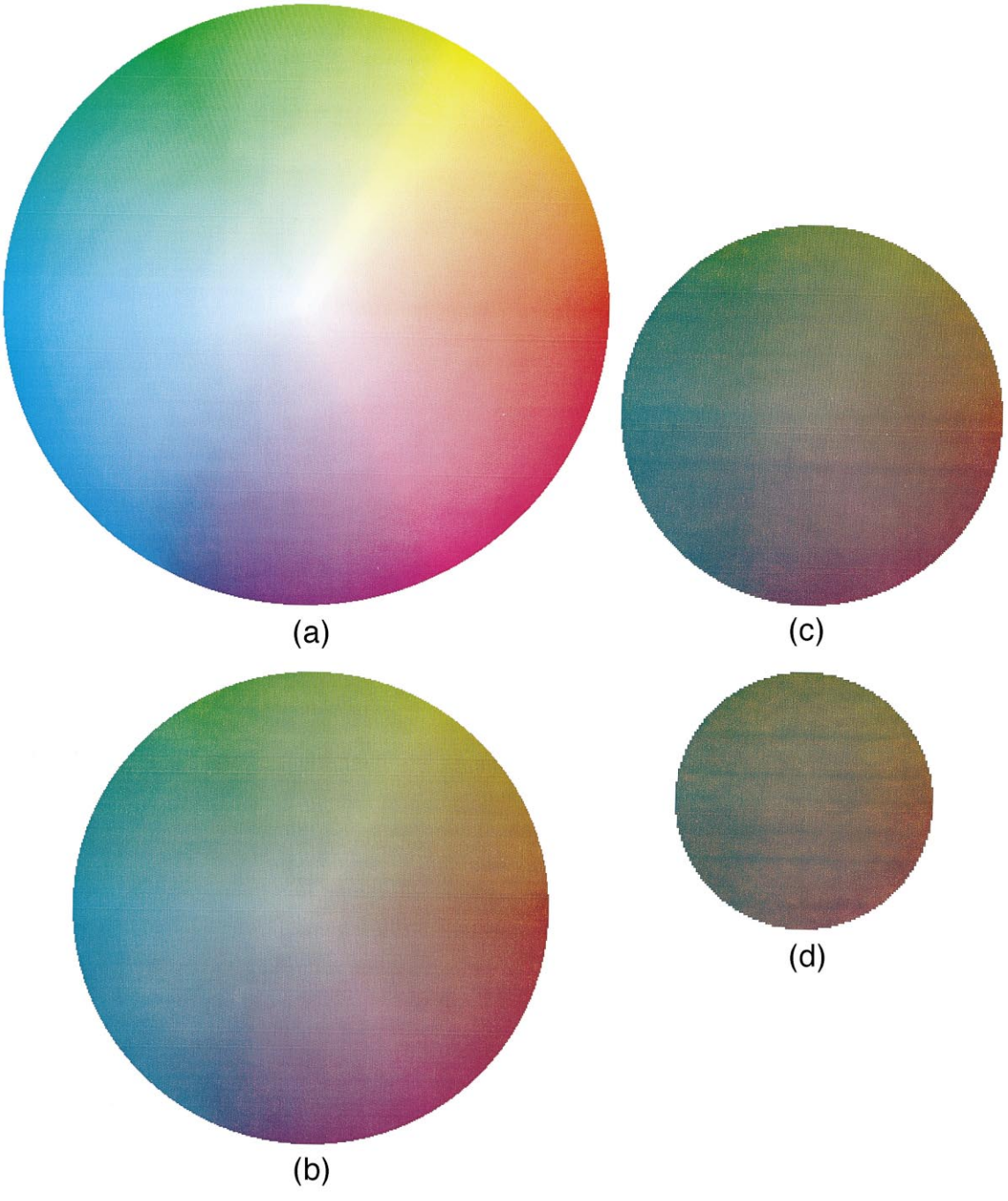


Fig. 3. HSV hexcone model at different values of the V component: (a) $V = 100\%$, (b) $V = 63\%$, (c) $V = T_{\mathrm{val}} = 35\%$ and (d) $V = 20\%$.

less than 10%. Similar results have been determined in [8]. The sector between 0% and 10% corresponds to the achromatic sector of a particular hexagonal slice in the HSV model. The range between 10% and 20% represents a sort of transition from the achromatic to the chromatic areas. Selecting $T_{\text{sat}} = 20\%$ as a lower bound yields satisfactory segmentation results, however, we have found that the addition of a select number of pixels within the 10–20% range can improve the results. This procedure is outlined below.

A principal polyhedron, PP, that corresponds to skin colored clusters with well-defined saturation components is formed by the selection of the following four thresholds:

$$T_{\text{hue1}} = 340^\circ \leqslant H \leqslant T_{\text{hue2}} = 360^\circ, \tag{6}$$

$$T_{\text{hue3}} = 0^\circ \leqslant H \leqslant T_{\text{hue4}} = 50^\circ, \tag{7}$$

$$S \geqslant T_{\text{sat1}} = 20\%, \tag{8}$$

$$V \geqslant T_{\text{val}} = 35\%. \tag{9}$$

Although this polyhedron is successful in extracting the skin-tone regions, an improvement can be realized if an additional number of pixels are selected from a second polyhedron, SP. This second polyhedron corresponds to the 10–20% transitional range and is determined adaptively as described below.

The histogram of all saturation values that lie within the bounds of Eqs. (6), (7) and (9) is first formed. The analysis of this histogram allows the threshold, T_{sat2}, to be selected which is essentially used to separate the chromatic and achromatic regions within the transitional region. Having determined the saturation histogram, we search for the first peak, Pk_1, beginning the search from the 0% saturation level. If the first peak exists at a value greater than 20% then the scene consists of mainly chromatic areas and a choice of $T_{\text{sat2}} = 10\%$ can safely be made. If Pk_1 is within the range 0–20% then the image also contains some achromatic regions (more so if Pk_1 is between 0% and 10%) which must be separated. In this case, the second peak, Pk_2 is detected (i.e. as we move away from Pk_1 in the direction of increasing saturation) and the in-between valley, $\text{Vl}_{1,2}$ is found. If $\text{Vl}_{1,2}$ lies in the range 0–20% then the selection $T_{\text{sat2}} = \text{Max}(10\%, \text{Vl}_{1,2})$ is made, where Max selects the maximum value of the operand. In certain images, the scene may consist of mainly chromatic regions where the saturation component gets slightly shifted due to the lighting conditions. This may result in the first peak being just under 20% with the valley being greater than 20%. In this case, we would like to include the chromatic component, and thus a saturation threshold of 10% is chosen. A similar procedure has been proposed in [8] for determining an adaptive threshold value in the saturation component. Thus, the selection of T_{sat2} is summarized as follows:

$$T_{\text{sat2}} = 10\%, \quad \text{if } \text{Pk}_1 > 20\%, \tag{10}$$

$$T_{\text{sat2}} = \text{Max}(10\%, \text{Vl}_{1,2}), \quad \text{if } \text{Pk}_1 < 20\% \cap 0\% \leqslant \text{Vl}_{1,2} \leqslant 20\%, \tag{11}$$

$$T_{\text{sat2}} = 10\%, \quad \text{if } \text{Pk}_1 < 20\% \cap \text{Vl}_{1,2} > 20\%. \tag{12}$$

In order to extract the significant peaks and valleys, then the histograms above must be smoothed to remove any meaningless local extrema. For this purpose, we apply the well-known scale space filter [3,30] where the 1-D saturation histogram, $f_s(x)$, is convolved with the Gaussian function, $g(x, \tau)$, of zero mean, m, and standard deviation, τ:

$$F_s(x,\tau) f_s * g(x,\tau) = \int_{-\infty}^{\infty} f_s \frac{1}{\sqrt{2\pi}\tau} \exp\left[\frac{-(x-u)^2}{2\tau^2}\right] \mathrm{d}u. \tag{13}$$

The peaks and valleys can subsequently be determined by examining the first and second derivatives of F_s above. The procedure just described is effective in separating the chromatic and achromatic regions.

The second polyhedron, SP, can now be formed by using the saturation threshold, T_{sat2}, that was just determined and this is defined by

$$T_{\text{hue1}} = 340^\circ \leqslant H \leqslant T_{\text{hue2}} = 360^\circ, \tag{14}$$

$$T_{\text{hue3}} = 0^\circ \leqslant H \leqslant T_{\text{hue4}} = 50^\circ, \tag{15}$$

$$T_{\text{sat2}} \leqslant S \leqslant 20\%, \tag{16}$$

$$V \geqslant T_{\text{val}} = 35\%. \tag{17}$$

The two polyhedra, PP and SP, expressed by Eqs. (6)–(9) and Eqs. (10)–(17), respectively, can

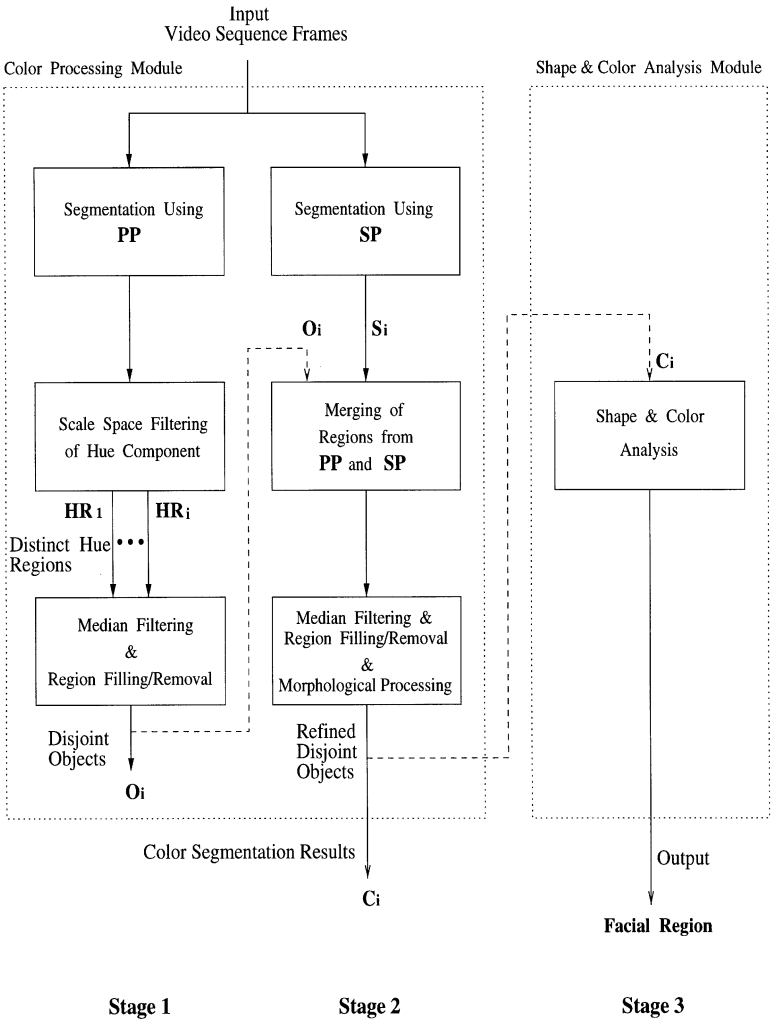


Fig. 4. Overall segmentation scheme using color and shape attributes.

now be used to extract the areas that correspond to the skin-tone clusters. The initial color segmentation using the defined polyhedra is summarized in the next section.

2.5. Segmentation using the color attribute

The overall segmentation technique that we propose is shown in the block diagram of Fig. 4. It consists essentially of two components: (1) a two-stage color processing module, and (2) a shape and color analysis module which is implemented in the third and final stage.

The first stage of the procedure is composed of three fundamental blocks as shown in Fig. 4. In the first of these, all of the pixels in the input image or frame that lie within the principal polyhedron, PP, are extracted and passed on to the next block. The transformation equations of Eqs. (1)–(5) can be used to convert RGB pixel values to the HSV color space. The extent of the hue range as defined in the principal polyhedron was chosen to be quite wide so that a variety of skin types could be modeled. As

a result of this, other objects in the scene with 'skin-like' colors (i.e. reddish-brown shirt) may also be extracted by this first block. Thus, the function of the second block is to separate these objects by color if this case arises. This is accomplished by analyzing the hue histogram of the extracted pixels. Scale space filtering, as described earlier, is used to smoothen the histogram and obtain the meaningful peaks and valleys. The valley between two peaks is used to separate two objects that possess different hue ranges (i.e. the facial region and a different colored object). This process partitions the initial segmentation into distinct hue regions, HR_i, as shown in the output of the second functional block of Fig. 4. Incidentally, in the remote case that another object matches the skin color of the facial area (i.e. separation is not possible by the scale space filter), then the shape analysis block in stage 3 will again provide the necessary discriminatory functionality. The binary representation of the pixels within each hue region are subsequently taken and passed on to the last block of Stage 1. A binary median filter and a region filling and removal step is applied to each hue region independently, to generate a set of objects of significant size. The objects within each region are finally combined to obtain the distinct objects, O_i. Further post-processing and shape analysis of these objects takes place on the bi-level images (i.e. binary representation of the objects) which we refer to as the object silhouettes. Details of the final post-processing block are described in the next subsection.

The output from the first stage is next passed on to the second stage of the color processing module. As mentioned earlier, the purpose of this second stage is to refine the segmentation results of the initial stage. In most cases, very reasonable results may be obtained even if this second stage is bypassed. In the first block, the secondary polyhedron, SP, is now used to extract the set of pixels that lie within this solid (the first block of both stages 1 and 2 can actually be implemented in one pass of the image). The extracted pixels S_i are subsequently merged with the results from stage 1. The merging process is performed as follows. Each pixel S_i is taken, and the distance d_{cs} to the centroid of each object, O_{i_c}, is computed. The centroid of each object is easily determined from the bi-level image using the object silhouette [11]. If the distance to the closest object is less than a certain threshold, then the pixel under consideration is added to that particular object. If it does not fall within the threshold, then the candidate pixel is discarded (i.e. not part of any object). The threshold chosen here is that d_{cs} must be within a certain factor, f_d, of the distance from the centroid of the object to the most distant point of the object, d_{cp}. In other words, $d_{cs} \leqslant f_d \times d_{cp}$, where a reasonable selection of f_d is made if this factor ranges between 1.0 and 1.5. The merging block just described consists of binary operations (i.e. performed on the object silhouettes) which can be implemented at a very low computational complexity.

The output from the merging block of the second stage is finally passed on to a post-processing block similar to the one in Stage 1 with the exception of an additional morphological operator. This block essentially refines the shape of the objects in the image and produces the final results from the color processing module (i.e. a set of refined objects, C_i). The details of this post-processing block are presented next.

2.6. Median filtering and region filling/removal

The median filter has found its way into numerous applications and has been particularly successful in the filtering of noise corrupted images and video sequences [19]. Here, the median filter is applied in the third block of each stage of the color processing module. Once again, we operate on the binary image frames which consist of the object silhouettes. The purpose of this median operation is to smoothen these silhouettes and also eliminate any isolated misclassified pixels that may appear as impulsive type noise from the initial color extraction stage (i.e. the output from block 2 of either stage). The two-dimensional median filter is given by

$$y_{k,l} = \mathrm{med}\{x_{k+r,l+s};\ (r,s) \in A\}, \tag{18}$$

where A defines the size and structure of the filter window about the central pixel (k,l). A set of n o $\cdots$ bservations, x_i, for $i = 1, \ldots, n$ are obtained from the filter window, and the median value is

computed as follows:

$$y_{k,l} = \text{med}(x_i) = x_{(v+1)}, \tag{19}$$

where $n = 2v + 1$ and $x_{(i)}$ denotes the ith order statistic. Square filter windows of size 5×5 and 7×7 provide a good balance between adequate noise suppression, and sufficient detail preservation. The binary output, $y_{k,l}$ above, can also be determined by a simple counting procedure which leads to a fast implementation.

The result of the median operation is successful in removing any misclassified 'noise-like' pixels, however, small isolated regions and small holes within object areas may still remain after this step. Thus, we follow the application of median filtering by region filling and removal. This operation fills in small holes within objects which may occur due to color differences (i.e. eyes and mouth of the facial skin region), extreme shadows, or any unusual lighting effects (specular reflections). At the same time, any erroneous small regions are also eliminated as candidate object areas.

This second post-processing step involves boundary extraction and contour tracing/counting of the median filtered binary image. The boundaries or edges of the binary image are easily determined by identifying the black pixels with at least one white nearest neighbor. We note that in the bi-level image, the black pixels correspond to skin-colored regions while the white space represents the non-skin-colored areas. The edge points of each contour formed are subsequently followed (under eight connectivity) and counted. If the contour boundary is less than a pre-determined threshold then the region is either filled or removed. If the region is surrounded by neighboring skin pixels then it is filled otherwise it is eliminated. The selection of the contour boundary threshold is based on the size of the image (i.e. smaller thresholds for smaller size images).

2.7. Morphological processing

The result of median filtering and region filling/removal yields one or more objects of significant size in which one of these is the facial region. In certain video sequences, however, we have found gaps or holes around the eyes of the segmented facial area. This occurs in sequences where the forehead is covered by hair and as a result the eyes fail to be included in the segmentation. Two morphological operators are used in the final block of the color processing module to account for this problem and also to smoothen the facial contour.

Most morphological operations can be defined in terms of two basic operations, *erosion* and *dilation* [24]. The erosion of an object $\boldsymbol{X}$ with the structuring element $\boldsymbol{B}$ is defined as the set of all points x such that $\boldsymbol{B}_x$ (the translation of $\boldsymbol{B}$ so that its origin is located at x) is included in $\boldsymbol{X}$,

$$\boldsymbol{X} \ominus \boldsymbol{B} = \{x: \boldsymbol{B}_x \subset \boldsymbol{X}\}. \tag{20}$$

Similarly, the dilation of $\boldsymbol{X}$ by $\boldsymbol{B}$ is the set of all points x such that $\boldsymbol{B}_x$ hits x, that is, they have a non-empty intersection [11],

$$\boldsymbol{X} \oplus \boldsymbol{B} = \{x: \boldsymbol{B}_x \cap \boldsymbol{X} \neq \emptyset\}. \tag{21}$$

The erosion outlined above uniformly reduces the size of an object whereas dilation performs the inverse and expands the object size. When combined, these two operations form the familiar morphological opening

$$\boldsymbol{X}_{\boldsymbol{B}} = (\boldsymbol{X} \ominus \boldsymbol{B}) \oplus \boldsymbol{B} \tag{22}$$

and closing

$$\boldsymbol{X}^{\boldsymbol{B}} = (\boldsymbol{X} \oplus \boldsymbol{B}) \ominus \boldsymbol{B}. \tag{23}$$

Here, we use these last two operations in the final post-processing stage. The closing operation is first used to fill in small holes and gaps followed by an opening operation which is used to remove small spurs and thin channels. Both of these operations maintain the original shapes and sizes of the object. A compact structuring element such as a circle or square without holes can be used to implement these operations and at the same time it can also help to smoothen the object contours. A semi-circular structuring element was used here as it provided adequate smoothing and a two-fold reduction in the computation time over its circular counterpart. Furthermore, these binary morphological operations can be implemented by low complexity 'hit or miss' transformations [24].

The output from the last block of the second stage, C_i, is the final result that is obtained from the color

processing module. At this point, the segmented results may contain one or more objects, C_i, in which one of these consists of the facial area. The shape and color analysis module of Stage 3 provides the mechanism to correctly select and classify the facial region. More details of this third stage are provided in the following section on shape and color analysis.

3. Shape and color analysis

3.1. Introduction

The output from the color processing module may contain objects other than the facial area. In this case, additional processing is needed to guarantee that the actual face will be extracted rather than an object with similar hue characteristics. In order to achieve this, a number of expected facial characteristics such as the shape, symmetry and facial location within the image should be used to determine the correct facial region. These facial characteristics will be fuzzified so that they become less sensitive to variations in the feature values. Although we apply the knowledge-based methodology to the problem of face location, it should be noted that feature-based recognition systems can be used to identify arbitrary objects. Such systems are based on the development of an object description from examples that are available to the designer. In the actual operating phase, the knowledge-based system associates a *membership value* with every feature for each one of the objects. These values give us an indication of the *goodness of fit* with an ideal prototype of the corresponding feature. An overall '*goodness of fit*' value can finally be derived for each object by combining the measures obtained from the individual primitives.

In most cases, the description of an object cannot be characterized by some unique or ideal value. However, fuzzy set theory can be used to quantify the acquired knowledge about the object. A number of fuzzy membership functions can be utilized to transform the physical measurements of the object into a set of values in the interval $[0, 1]$. The value of a particular membership function quantifies the degree to which the object fits the corresponding primitive. Depending on the construction of the knowledge-based system many of these membership values can be fused together to generate an overall *goodness of fit* measure for the object under consideration.

In conventional knowledge-based face recognition systems, features such as the width of the eyes, nose and mouth, the distances between pairs of facial components, and the geometry of the human face are used as primitives. In Stage 3 of our segmentation scheme we utilize a set of features that are suitable for our application purposes. In most videoconferencing or videophone-type sequences, the scene consists of front-view faces which are relatively close to the center of the image. Thus, we utilize features such as the location of the face, its orientation from the vertical axis, and its aspect ratio to assist us with the location/recognition task. Values from these primitives are used to construct the membership functions using a set of examples that are available during the training phase. In the evaluation phase, the corresponding membership function values are used to determine the degree to which each object satisfies the particular invocation of the facial feature.

In the methodology we propose, each segmented object, C_i (i.e. obtained from the color processing module of Stage 2), is examined to determine the degree to which it satisfies the selected facial primitives. More specifically, we consider the following four features (primitives) in our face localization system:

1. *Deviation from the average hue value of the different skin-type categories.* The average hue value for different skin-types varies amongst humans and depends on the race, gender and the age of the person. However, it was shown in Section 2 that the facial region exhibits regular properties in the HSV color space. In particular, the hue values of skin fall within a specific range for all skin-type categories (Table 1). The average hue of the different skin-types forms a range that represents the most probable hue for human skin tones. The deviation of an object's expected hue value from this defined range gives us an indication of its similarity to skin tone colors.
2. *Face aspect ratio.* Given the geometry and the shape of the human face, it is reasonable to expect that the ratio of height to width falls

within a specific range. If the dimensions of a segmented object fit the commonly accepted dimensions of the human face then it can be classified as a facial area.

3. *Vertical orientation.* The location of an object in a scene depends largely on the viewing angle of the camera, and the acquisition devices. In video sequences intended for videoconferencing, videophone or multimedia mail applications, it is assumed that:
   3.1. The head is not tilted forwards, or backwards so that the face becomes occluded.
   3.2. Only reasonable rotations of the head are allowed in the image plane. This corresponds to a small deviation of the facial symmetry axis from the vertical direction. This is a logical assumption for the intended applications, as the head will not be parallel to the horizontal axis in a video communication scenario.

   This primitive is utilized so that an object is excluded as a valid facial area when its orientation axis (i.e. least moment of inertia) exhibits a large deviation from the vertical axis.
4. *Relative position of the facial region in the image plane.* By similar reasoning to 3 above, it is more probable that the face will be located in a region that is relatively close to the center rather than the edges of the image. This feature is used so that any segmented objects which are located near the edges and corners of the image plane are less likely to be classified as facial regions.

3.2. Fuzzy membership functions

The four features described above are used to define the membership functions required in calculating an appropriate evaluation measure for the invocation of the different primitives. In our segmentation scheme, each membership function provides the degree of similarity of the given object to the facial primitive in question. Thus, the membership values are used to quantify the deviation from the expected or ideal feature value.

A number of membership function models can be constructed and empirically evaluated. A simplified function model is utilized here in order to keep the complexity of the overall scheme to a minimum. A trapezoidal shape was selected as the membership function for each of the primitives described above. The general form of the function is defined below and is also shown schematically in Fig. 5:

$$\mu(x) = \begin{cases} \dfrac{(x-c)}{(a-c)} & \text{if } c \leqslant x \leqslant a, \\ 1 & \text{if } a \leqslant x \leqslant b, \\ \dfrac{(d-x)}{(d-b)} & \text{if } b \leqslant x \leqslant d, \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

This type of membership function attains the maximum value only over a limited range of input values. Symmetric or asymmetrical trapezoidal shapes can be obtained depending on the selected parameter values of a, b, c and d. The membership

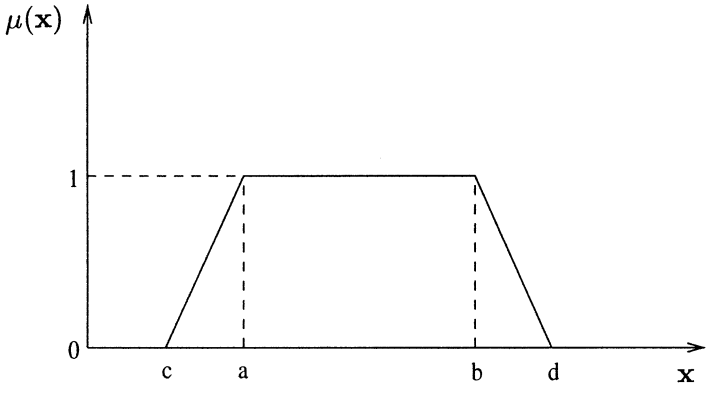


Fig. 5. General trapezoidal membership function.

function can assume any value in the interval [0, 1], including both of the extreme values. A value of 0 in the definition above, indicates that the event is impossible. On the contrary, the maximum membership value of 1 represents total certainty. The intermediate values are used to quantify variable degrees of uncertainty. The estimates for the four membership functions are obtained by a collection of physical measurements of each primitive from our extensive database. The values of the trapezoidal parameters in the four membership functions are set so that each function accurately represents the physical primitives observed.

The image database that was constructed for the analysis of the skin-tone distributions was also used in devising the ranges of the trapezoidal membership functions. The hue characteristics of the facial region were used to form the first membership function. The extent of our proposed hue range $[-20°, 50°]$ (i.e. discrete universe of discourse) has purposely been designed to be quite wide in order to adequately model the different skin-types and the varying illumination conditions. However, the mean hue values of the different skin-type categories fall within an even narrower range. The lower bound of the average hue observed in the image database is approximately 8° (African–American distribution) while the upper bound average value is around 30° (Asian distribution). A range is formed using these values, where an object is accepted as a skin-tone color with probability 1 if its average hue value falls within these bounds. Objects C_i with average hue values outside this range (i.e. closer towards the extremes of the defined range) are assigned a smaller weighting and are also less likely of being classified as a facial region. Thus, the membership function associated with the first primitive is defined as follows:

$$\mu_1(x_1) = \begin{cases} \dfrac{(x_1 + 20)}{28} & \text{if } -20° \leqslant x_1 \leqslant 8°, \\ 1 & \text{if } 8° \leqslant x_1 \leqslant 30°, \\ \dfrac{(50 - x_1)}{20} & \text{if } 30° \leqslant x_1 \leqslant 50°. \end{cases} \tag{25}$$

Experimentation with a wide variety of facial images has led us to the conclusion that the aspect ratio (height/width) of the human face has a nominal value of approximately 1.5. This finding confirms previous results reported in the open literature [14]. However, in certain video sequences we must also compensate for the inclusion of the neck area which has similar skin-tone characteristics to the facial region. This has the effect of slightly increasing the aspect ratio. Using this information along with the observed aspect ratios from our database, we can tune the parameters of the trapezoidal function for this second primitive. The final form of the function is given by

$$\mu_2(x_2) = \begin{cases} \dfrac{(x_2 - 0.75)}{0.5} & \text{if } 0.75 \leqslant x_2 \leqslant 1.25, \\ 1 & \text{if } 1.25 \leqslant x_2 \leqslant 1.75, \\ \dfrac{(2.25 - x_2)}{0.5} & \text{if } 1.75 \leqslant x_2 \leqslant 2.25, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

The vertical orientation of the face in the image is the third primitive used in our shape recognition system. As mentioned previously, the orientation of the facial area (i.e. deviation of the facial symmetry axis from the vertical axis) is more likely to be aligned towards the vertical due to the type of applications considered. The following range of values were observed for the orientation of the facial region in over 100 frames from several video sequences: (i) (0–10.5°) in *Foreman*, (ii) (0–4.75°) in *Akiyo*, (iii) (3.5–21°) in *Carphone* and (iv) (0.5–6.75°) in *Claire*. A reasonable threshold selection of 30° can be made for valid head rotations as confirmed in the observed sequences above. Thus, a membership value of 1 is returned if the orientation angle is less than this threshold. The membership function for this primitive is defined as follows:

$$\mu_3(x_3) = \begin{cases} 1 & \text{if } 0 \leqslant x_3 \leqslant 30°, \\ (90 - x_3)/60 & \text{if } 30° \leqslant x_3 \leqslant 90°. \end{cases} \tag{27}$$

The last primitive used in our knowledge-based system refers to the relative position of the face in the image. Due to the nature of the applications considered, we would like to assign a smaller weighting to objects that appear closer to the edges and corners of the images. For this purpose, we

construct two membership functions. The first one returns a confidence value for the location of the segmented object with respect to the X-axis. Similarly, the second one quantifies our knowledge about the location of the object with respect to the Y-axis. The discrete universe of discourse for these membership functions depends on the dimensions of the image. Since our system supports variable size images, the following membership function has been defined for the position of the segmented object with respect to either the X or Y-axis:

$$\mu_4(x_4) = \begin{cases} \dfrac{(x_4 - (d))}{d/2} & \text{if } d \leqslant x_4 \leqslant 3d/2, \\ 1 & \text{if } 3d/2 \leqslant x_4 \leqslant 5d/2, \\ \dfrac{((3d) - x_4)}{d/2} & \text{if } 5d/2 \leqslant x_4 \leqslant 3d, \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

The membership function for the X-axis is determined by letting $d = D_x/4$ where D_x represents the horizontal dimensions of the image (i.e. in the X-direction). In a similar way, the Y-axis membership function is found by letting $d = D_y/4$ where D_y represents the vertical dimensions of the image (i.e. in the Y-direction). Thus, the X- and Y-axis membership functions are assigned the maximum value if the centroid of the object is within a window that is relatively central to the image. The parameter values have been appropriately chosen for the intended applications.

3.3. Aggregation operators

In the end, the individual membership functions must be combined to form an overall decision. A nonlinear operator is used to arrive at this final decision by appropriately combining the information from the different features. The function of the operator is to reduce the imprecision and uncertainty in the decision-making process. A number of fuzzy operators can be used to combine or fuse together the various sources of information. Conjunctive type of operators represent a consensus between the different sources of information. Such operators search for a simultaneous satisfaction of the various primitives or objectives by weighting more heavily the criterion with the smallest membership value. On the contrary, disjunctive operators express redundancy between information by assigning the most weight to the criterion with the largest membership value. Compromise operators, such as weighted mean operators or fuzzy integrals provide a trade-off among different and possibly incompatible objectives.

The latter approach is followed in this paper. An aggregator (fuzzy connective) whose shape is defined a priori, is used to combine the four elemental membership functions resulting from the primitives discussed above.

The *compensative operator* selected mixes both conjunctive and disjunctive behavior. Following the results in [32], the operator is defined as the weighted mean of a (logical AND) and a (logical OR) operator

$$A \odot_\gamma B = (A \cap B)^{1-\gamma} \cdot (A \cup B)^\gamma, \tag{29}$$

where A and B are sets defined on the same space and represented by their membership functions. Different *t-norms* and *t-conorms* can be used to express a conjunctive or a disjunctive attitude. If the product of membership functions is utilized to determine the intersection (logical AND) and the possibilistic sum for the union (logical OR), then the form of the operator becomes [32]

$$\mu_c = \prod_{j=1}^{m} \mu_j^{(1-\gamma)} \left(1 - \prod_{j=1}^{m} (1 - \mu_j)\right)^\gamma, \tag{30}$$

where μ_c is the overall membership function which combines all the knowledge primitives for a particular object, and μ_j is the jth elemental membership value associated with the jth primitive. The weighting parameter γ is interpreted as the *grade of compensation* taking values in the range of $[0, 1]$ [32]. The product and the possibilistic sum are not the only operators that can be used in Eq. (29). A simple and useful *t-norm* function is the min operator while the corresponding one for the *t-conorm* is the max operator. In this paper, we utilize this *t-norm* to represent intersection. In this case, the compensative operator of Eq. (29) has the following form:

$$\mu_c = \left(\min_{j=1}^{m} \mu_j\right)^{(1-\gamma)} \left(\max_{j=1}^{m} \mu_j\right)^\gamma. \tag{31}$$

The form of the compensative operator in Eq. (30) is not unique. A number of other mathematical models can be used to represent the AND aggregation. An alternative operator, which combines the averaging properties of the arithmetic mean (member of the averaging operator class) with a logical AND operator (conjunctive operator) was also proposed in [32]

$$\mu_c = \gamma \min_{j=1}^{m} \mu_j + (1-\gamma)\left(m^{-1}\sum_{j=1}^{m} \mu_j\right), \tag{32}$$

where μ_c is again the overall membership function and the parameter $\gamma \in [0,1]$ is interpreted as the grade of compensation. In this equation the min *t-norm* stands for the logical AND. Alternatively, the product of membership functions can be used instead of the min operator in the above equation. The arithmetic mean is used to prevent higher elemental weights with extreme values to dominate the final outcome.

Compensatory operators are intuitively attractive and provide a simple yet powerful method to express the interactions between different knowledge primitives. For this reason, our shape and color analysis module utilizes these operators in correctly selecting the facial area from a set of candidate objects.

In this work we define $\gamma = 0.5$. Therefore, the compensative operator assumes the form of a weighted product. The min and max operators were selected to model the corresponding *t*-norm and *t*-conorm functions [20]. Thus, the overall fuzzy membership function can be defined as

$$\mu_c = \left(\left(\min_{j=1}^{m} \mu_j\right)\left(\max_{j=1}^{m} \mu_j\right)\right)^{0.5}. \tag{33}$$

In general, additional weighting factors must be used in the generalized function above in order to absorb possible scale differences in the definition of the elemental membership functions. However, all the elemental membership functions used here are within the interval $[0,1]$, and thus no such weighting factors are required.

The aggregation operator defined in Eq. (33) can be used to form the final decision based on the four primitives under consideration. However, in order for our results to be meaningful, the nonlinear operator applied must satisfy some properties that will guarantee that its application will not alter in any manner the elemental decisions about the knowledge primitives. In the literature, there are a number of properties that all the aggregation or compensative operators must satisfy. We will try to examine if the operator which we intend to use in the calculation of the final membership function satisfies these properties [23]. These properties are listed below:

1. *Convexity*. The convexity of the operators allows for a compromise among the different elemental membership functions. The weighted operator in Eq. (33) is convex since it is known from statistics that

$$\mu_c^a = \left(\min_{k=1,j} \mu_k \max_{k=1,j} \mu_k\right)^{0.5}, \tag{34}$$

$$\min_k \mu_k \leqslant \mu_c^a \leqslant \max_k \mu_k, \tag{35}$$

where $k = 1,2,\ldots,j$ is the number of elemental membership functions to be fused together.

2. *Neutrality* (*symmetry*). The operator used here is symmetric. The property guarantees that the order of presentation for the elemental membership functions does not affect the overall final membership value. It is not hard to see that by simply interchanging the order of presentation for the max and the min value the same result will occur

$$\mu_c = \left(\left(\min_{j=1}^{m} \mu_j\right)\left(\max_{j=1}^{m} \mu_j\right)\right)^{0.5} = \left(\left(\max_{j=1}^{m} \mu_j\right)\left(\min_{j=1}^{m} \mu_j\right)\right)^{0.5}. \tag{36}$$

3. *Monotonicity*. The property of monotonicity guarantees that the stronger piece of evidence (larger elemental membership value) generates a stronger support in the final membership function.

Let us assume that $\mu_i \leqslant \mu_l$, with $A = \min_{k=1}^{j} \mu_k$ and $B = \max_{k=1}^{j} \mu_k$.

By the definition of the min and max operators

$$\min(A, \mu_i) \leqslant \min(A, \mu_l) \tag{37}$$

and

$$\max(A, \mu_i) \leqslant \max(A, \mu_l). \tag{38}$$

Therefore,

$$(\min(A, \mu_i)\max(A, \mu_i))^{0.5} \leqslant (\min(A, \mu_l)\max(A, \mu_l))^{0.5}. \tag{39}$$

4. *Idempotence*. The operator considered in Eq. (33) is idempotent. The property guarantees that the outcome of the overall function generates the same value with each elemental value if all of them report the same result. Given the form of the operator

$$\mu_c = (\mu_a \mu_b)^{0.5} = (\mu^* \mu^*)^{0.5} = \mu^*, \tag{40}$$

with

$$\mu_a = \left(\min_{j=1}^{m} \mu_j \right) = \mu^* \tag{41}$$

if $\mu_1 = \mu_2 = \cdots = \mu_j = \mu^*$.

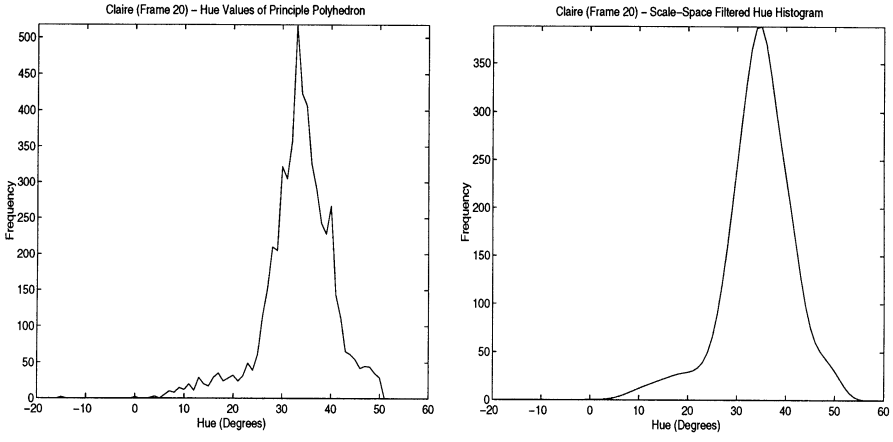
In summary, it is proven that the compensatory operator that we intend to utilize for our shape and color analysis module in Stage 3 corresponds to an aggregation class which satisfies a number of natural properties, such as neutrality and monotonicity.

4. Experimental results

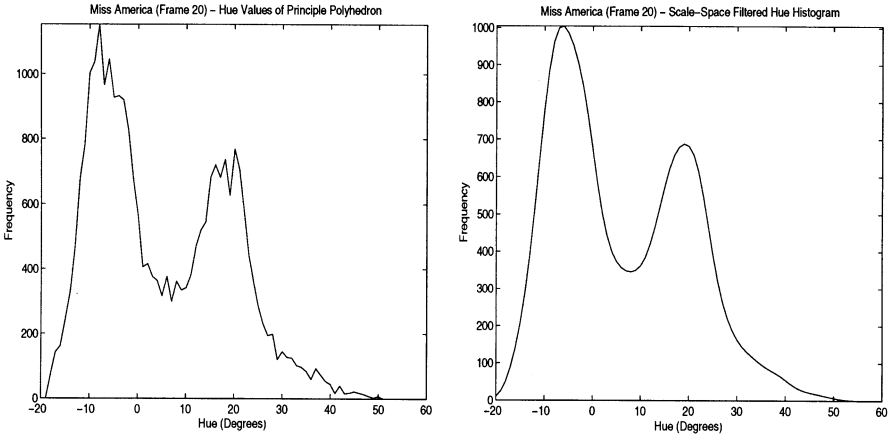
The steps outlined in Fig. 4 were used to locate and track the facial region of several videophone-type sequences. The results from 3 CIF and 2 QCIF sized sequences, as well as one still image, are presented below: (1) Claire, (2) Miss America, (3) Akiyo, (4) Foreman, (5) Carphone and (6) an African–American sample image, respectively. The aforementioned test sequences were chosen so as to represent all skin-type categories (i.e. Caucasian, African–American and Asian) and various types of motion within the scene (camera pan and zoom, head rotations and tilts, and moving complex backgrounds). The segmentation results in Fig. 10 illustrate the robustness of the technique to the various cases of motion and skin color mentioned above. The facial region is successfully located and tracked when the head is rotated as in Fig. 10(a) or in the cases of head tilts as in Fig. 10(c) and (f). The technique is successful even when the facial area undergoes various deformations caused by different facial expressions (i.e. Claire, Miss America, Carphone, and Akiyo). The Foreman sequence of Fig. 10(d) demonstrates that the extraction process is invariant to pans and zooms within the scene while the Carphone sequence in Fig. 10(f) illustrates the effectiveness of the algorithm under conditions of a complex and moving background. The latter scenario may be the case in an environment where mobile videophones are employed. Finally, we observe that successful results are obtained for the complete range of skin colors (Akiyo, African–American sample and Miss America, i.e. Fig. 10(c), (e) and (b), respectively).

In Fig. 6(a)–(f) we present the hue histograms of Frame 20 from the different video sequences. These are obtained by passing each of the images through the principal polyhedron (PP) as defined previously in Section 2.4. The smoothed scale-space filtered versions of these histograms are also shown alongside the former, and are derived from the second block in Stage 1 of the color processing module (Fig. 4). A standard deviation of $\tau = 2$ in the Gaussian function, $g(x, \tau)$, provided adequate smoothing of the histograms and was found to be appropriate for the different skin-tone distributions which had standard deviations, σ that ranged from 5.1 to 8.2°. In the Claire sequence of Fig. 6(a), the histograms indicate that one distinct hue range exists which has a mean value around 34°. In turn, this range contains only one distinct object, O_1, which is the facial region. The observed hue values are shifted towards the yellow spectrum which is also evident visually from the results in Fig. 10(a). Thus, we see that in the case of the Claire sequence, the shape and color analysis module (SC module) of Stage 3 need not be invoked.

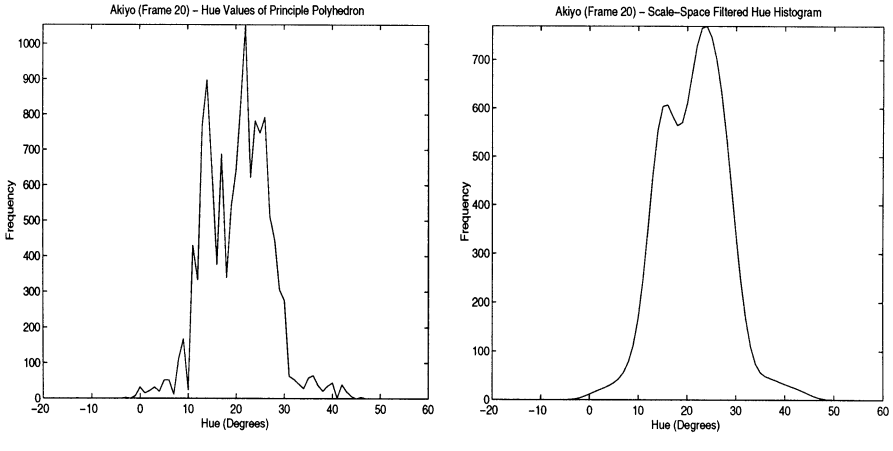
The histograms in both, Fig. 6(b) and (c) indicate that two different hue ranges exist and we refer to each of these as hue regions, HR_1 and HR_2. Incidentally, each hue region, HR_i, may contain one or more disjoint areas which we refer to as objects, O_i (C_i is the post-processed version of O_i). The latter are processed by the SC module in the selection of the facial region. Each hue range in both, Fig. 6(b)



(a)

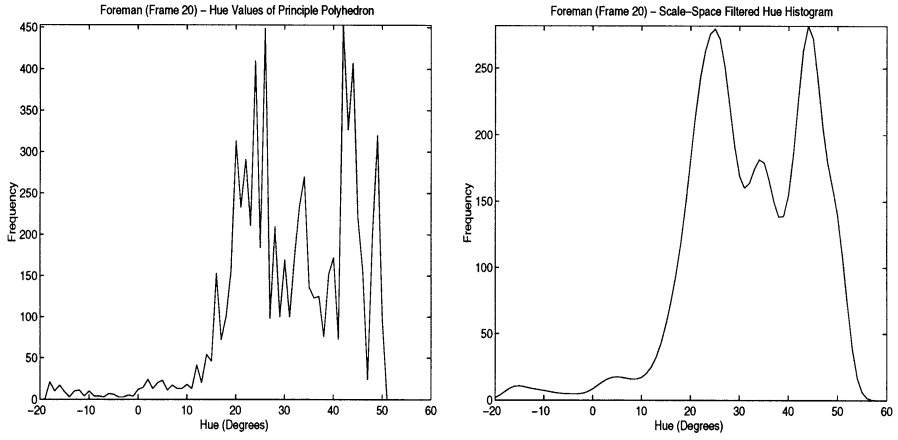


(b)

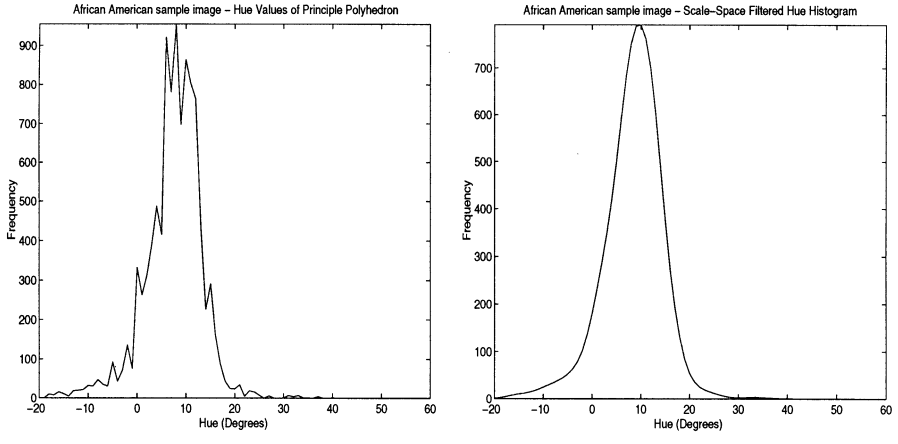


(c)

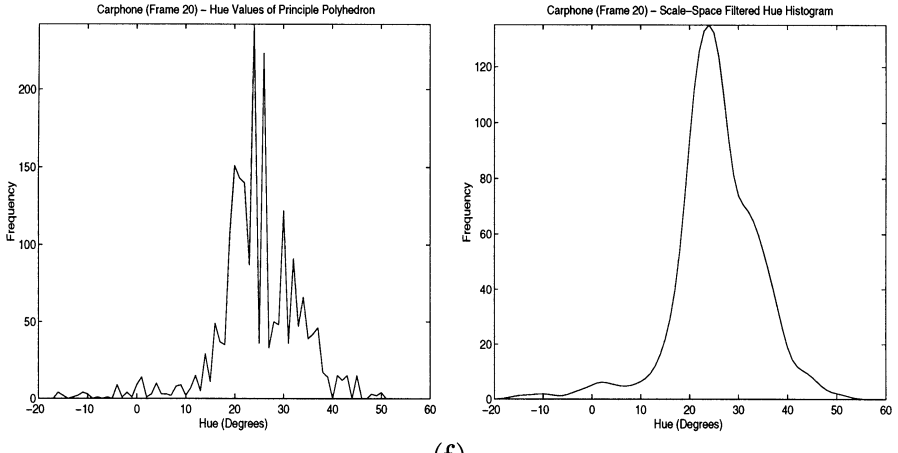
Fig. 6. Hue values of the principal polyhedron along with its scale-space filtered version for Frame 20 of the following video sequences: (a) Claire, (b) Miss America, (c) Akiyo, (d) Foreman, (e) African–American sample image and (f) Carphone.



(d)



(e)



(f)

Fig. 6. Continued.

and (c), is determined by utilizing the hue histogram. Each range, in turn, may contain one or more objects which are analyzed by the shape module in order to correctly select the one which corresponds to the facial area. A more detailed analysis of these two sequences is presented in the subsections below.

In Fig. 6(d), the Foreman sequence contains three different hue regions (i.e. 3 distinct peaks), which are separated by the valleys at 32° and 38°. The first two regions $-20° \leqslant HR_1 \leqslant 32°$, $32° \leqslant HR_2 \leqslant 38°$, each contain one object, C_1 and C_2, respectively, while the third region $38° \leqslant HR_3 \leqslant 50°$ contains two objects, C_3 and C_4. In the final analysis, the shape module correctly selects the first object, C_1 as the facial area of the sequence. The mean value of C_1 (i.e. first peak in Fig. 6(c)) is approximately 25° which is also in accordance with the mean hue value of the Caucasian skin-type distribution ($\mu = 25.3°$). A summary of the results from the SC module are presented in Table 4.

Next, in Fig. 6(e), the African–American sample image consists of only one hue region, HR_1, containing one object, C_1, which has a mean of approximately 10°. This value is close to the mean of 8.6° which was found earlier for the African–American distribution. Once again, the shape module is not necessary in the segmentation of this image.

Finally, in the Carphone sequence of Fig. 6(f) we can only identify one region with a distinct hue range, HR_1 (i.e. only one distinct maximum or minimum point), and this has a mean value of approximately 24°. This value is also in accordance with the expected value of the Caucasian distribution and the pixels about the peak belong to the facial area. Some of the values in the tail of the distribution (i.e. 35–50°) do not correspond to the facial area, however, these pixels are scattered and thus are removed by median filtering and region removal. As a result, only one object, C_1, remains which is the extracted facial region. Incidentally, we have found that this sequence benefits by using the secondary polyhedron (SP in Stage 2 of Fig. 4). The segmentation results are refined around the chin area of the facial region which contains less saturated pixel values. A more detailed account of the Akiyo and Miss America sequences is given next to illustrate the overall procedure and the use of the shape and color analysis module in the selection of the facial region from the set of candidate objects. The remaining sequences (excluding *Foreman*) do not require the discriminatory functionality of the SC module.

4.1. Akiyo sequence

The set of results in Fig. 8 illustrate the details of the segmentation process for the Akiyo sequence (Frame 20) through the different stages of the facial extraction scheme outlined in Fig. 4. The segmentation process begins by first passing the input image through the principal polyhedron, PP. The histogram of the hue values within PP is formed and subsequently smoothened by the scale-space filter, $g(x, \tau)$. The results of the histograms obtained from this step are shown in Fig. 6(c). From the scale-space filtered version we can identify two hue regions. The minimum value at $H = 18°$ is used to separate the two regions as follows $18° \leqslant HR_1 \leqslant 50°$ and $-20° \leqslant HR_2 < 18°$. The local maxima and minima in Fig. 6(c) are determined automatically by the scale-space filtering technique described earlier. The images in Fig. 8(a)–(f) illustrate the results obtained from the remaining steps in the segmentation process for the region HR_1 which incidentally corresponds to the facial area. A similar procedure is also carried out for the second region, HR_2.

In Fig. 8(a) the results are shown for HR_1 after the primary extraction process using PP. Most of the facial skin area is extracted from this initial step, with the addition of some erroneous regions from the jacket area. Fig. 8(b) illustrates the output after median filtering which is used to remove the isolated 'noise-like' pixels. A filter window of 7×7 was chosen for the CIF size images while a 5×5 mask was utilized for the smaller QCIF size format. In Fig. 8(c) the results are shown after region filling/removal. This step eliminates small misclassified regions and also fills in holes within the larger regions (i.e. the eyes and mouth). A region was removed if its perimeter was less than a threshold value of 200 pixels for the CIF images and 100 for the QCIF sequences. These values were found to be

appropriate choices for the videophone-type applications that were considered. A selection of smaller thresholds can also be made for other applications (i.e. where the face occupies a very small area within the image), however, this requires the SC module to analyze a greater number of objects. At this point in the segmentation process, we are left with one object, O_1, within the first hue region, HR_1.

The steps highlighted above mark the completion of Stage 1 of the color processing module. In the first block of Stage 2, the pixel values within the secondary polyhedron, SP, are extracted. As mentioned earlier, this is done to include pixels which lie in the transition range of chromatic and achromatic regions. The threshold T_{sat2} found in Eqs. (14)–(17) defines SP, and its selection is made from the saturation histogram formed by Eqs. (6), (7) and (9). This histogram is shown in Fig. 7(a) for the region of HR_1, while the one in Fig. 7(b) illustrates that for HR_2. A choice of $T_{sat2} = 18\%$ is made for SP of HR_1 according to the conditions in Eq. (11). Fig. 8(d) shows the pixels extracted by the Secondary Polyhedron while Fig. 8(e) displays the result of the merging block in the second step of Stage 2. A factor of $f_d = 1.1$ was used in merging the results from the two polyhedra. As we can see, in the Akiyo sequence, the SP extraction process has virtually no effect in refining the results obtained from the principal polyhedron. Finally, in Fig. 8(f) the segmentation results are shown for the region HR_1 after the final post-processing block in Stage 2. A semi-circular structuring element (SCSE) of radius 15 was utilized for the morphological closing operation and a radius of 5 for the opening operation. This was found to be quite effective in accomplishing the desired objectives. Furthermore, the SCSE performed equally as well as its circular counterpart while requiring only half the number of operations. In Fig. 8(f), only one object remains, C_1 (i.e. C_i is the post-processed version of O_i) and this happens to be the facial region. A similar step-by-step procedure was repeated for the second hue region, HR_2, and in this case the object C_2 in Fig. 8(h) was obtained.

The two objects, C_1 and C_2, obtained from the color module in Fig. 4 were subsequently passed on to the shape and color analysis module for the selection process. The shape module analyzes each

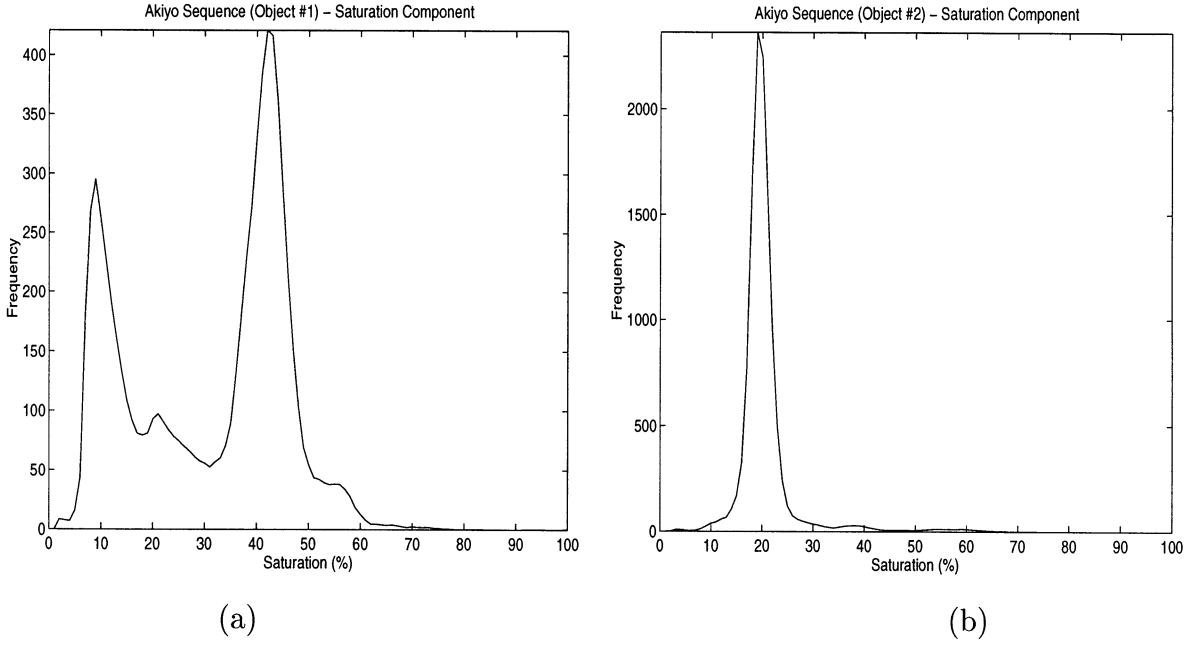


Fig. 7. Saturation components of the principal polyhedron for the two hue regions, HR_1 and HR_2, of the Akiyo sequence (Frame 20).

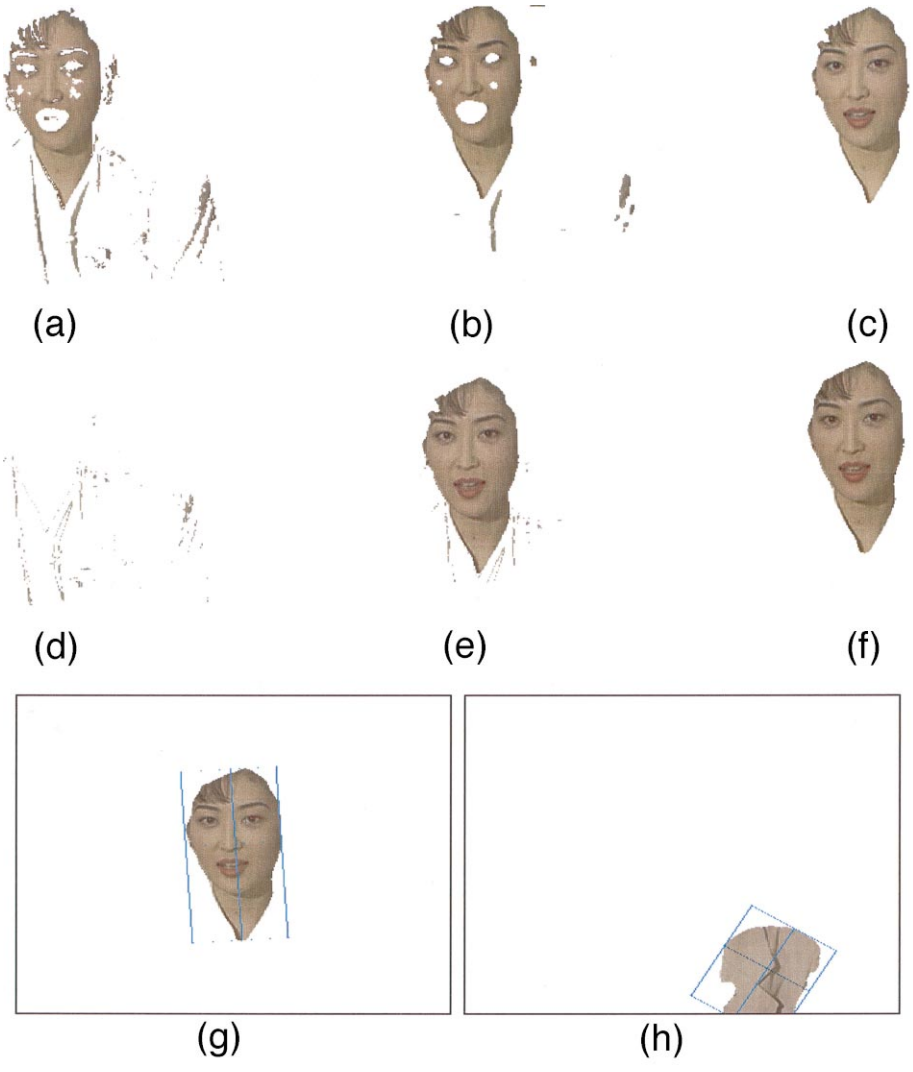


Fig. 8. Extraction of the facial region for the Akiyo sequence (Frame 20) through the various stages: (a) initial extraction by the principal polyhedron for the hue region, $8^\circ \leqslant \mathrm{HR}_1 \leqslant 50^\circ$, (b) median filtered result, (c) region filling and removal, (d) extraction of the secondary polyhedron for HR_1, (e) region merging of the extracted PP and SP regions, (f) final result of HR_1 after morphological processing (i.e. C_1), (g) shape processing of object, C_1 (i.e. facial region), and (h) shape processing of object C_2 found from the second hue region $-20^\circ \leqslant \mathrm{HR}_1 \leqslant 8^\circ$.

Table 2
Akiyo (width × height = 352 × 288): shape and color analysis

Object	Centroid location				Orientation		Object ratio		Mean hue		Aggregation
C_i	X	μ_1	Y	μ_2	θ°	μ_3	r	μ_4	H_m (°)	μ_5	μ_c
1	178	1	134	1	3.10	1	1.97	0.56	23	1	**0.75**
2	246	0	245	0.2	31.70	0.98	1.18	0.86	14	1	0.0

Table 3
Miss America (width × height = 360 × 288): shape and color analysis

Object	Centroid location				Orientation		Object ratio		Mean hue		Aggregation
C_i	X	μ_1	Y	μ_2	θ°	μ_3	r	μ_4	H_m (°)	μ_5	μ_c
1	177	1	188	1	4.92	1	1.61	1	20	1	**1.0**
2	245	0	120	1	47.74	0.7	1.16	0.82	− 6	0.5	0.0
3	244	0	269	0.02	44	0.77	1.32	1	− 5	0.54	0.0

object and computes a set of values for the different primitives considered. Table 2 summarizes the results of the five primitives along with the membership function values, μ_i, for $i = 1, \ldots, 5$ for each of these features. The aggregation of these functions, μ_c is computed by Eq. (33) and is shown in the final column of the table. The first object, C_1, scored the highest aggregate value and, therefore, was selected as the facial region. A high membership value was obtained for every primitive, except for the object ratio which begun to exceed the bounds of the allowable facial aspect ratio. Object C_2 scored reasonably well in orientation, object ratio and mean hue, however, its poor location in the image brought its aggregation value down. Both of these objects C_1 and C_2 can be seen in Fig. 8(g) and (h), respectively.

4.2. Miss America sequence

The detailed procedure just described was also applied to the Miss America sequence. The scale-space filtered hue histogram in Fig. 6(b) indicates that two hue regions exist just as in the Akiyo sequence. These are easily separated into the following two ranges: $8^\circ \leqslant \mathrm{HR}_1 \leqslant 50^\circ$ and $-20^\circ \leqslant \mathrm{HR}_2 < 8^\circ$. In the first region, HR_1, only one object remains (i.e. the facial region) and this is shown in Fig. 9(b). However, the second region, HR_2, contains two objects, C_2 and C_3, and these are illustrated in Fig. 9(c) and (d), respectively. These latter two correspond to the jacket area and are in the red–magenta sector of the hue hexagon. Fig. 9(a) (image before the morphological operation) is shown simply to illustrate the importance of the morphological operation in filling the holes around the eye regions in cases where the hair is close to these areas.

The shape and color feature values are provided in Table 3 for each of the three objects in the Miss America sequence (Fig. 10). Once again, object number 1, C_1 is correctly chosen as the facial region based on the computed aggregation value. The objects C_2 and C_3 scored poorly in their location and Mean hue value, and also had lower membership values in the orientation primitive. The net effect of this was to bring the aggregation value down to zero in each case.

The results in Table 4 illustrate the details of the color and shape analysis for the Foreman sequence. As mentioned previously, four objects are identified in the sequence as a result of three hue regions. The first object corresponds to the facial region and this one is also selected by our knowledge-based system due to its aggregation value.

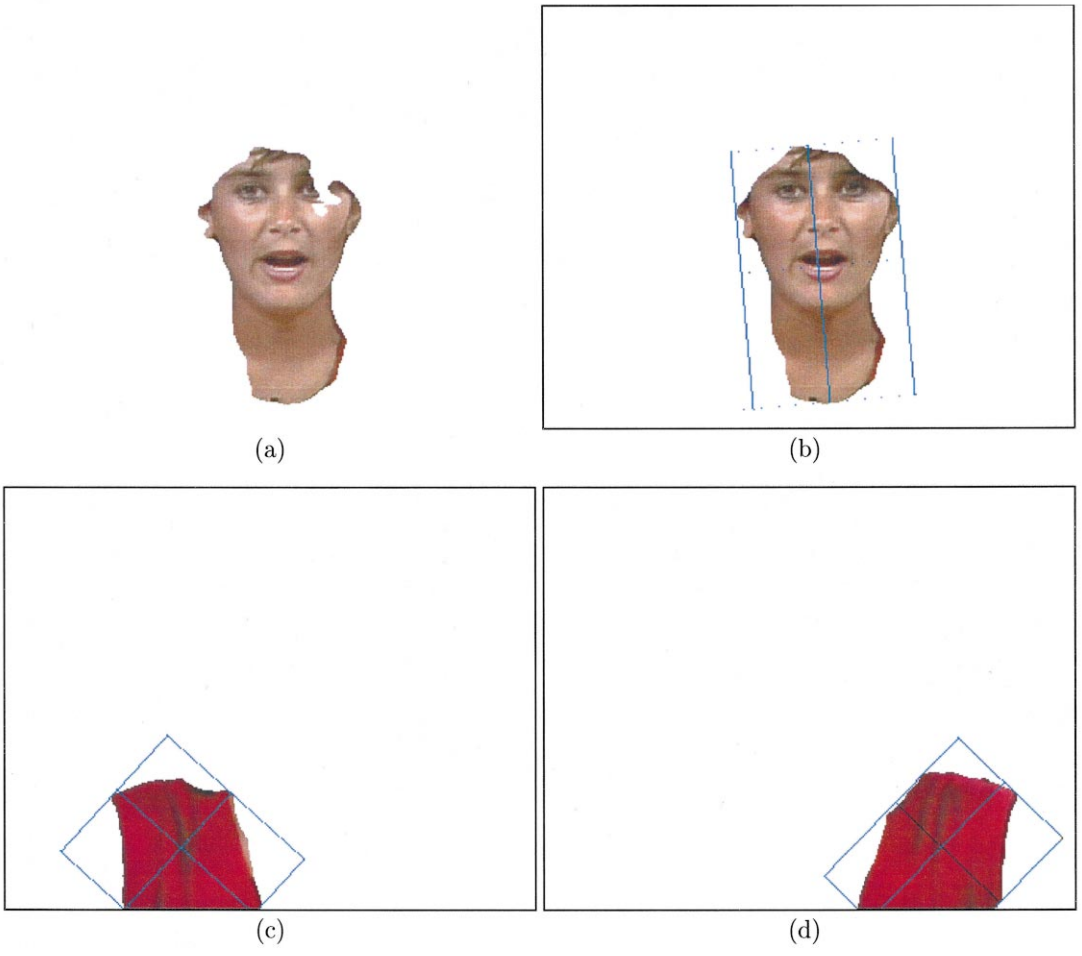


Fig. 9. Extraction of the facial region for the Miss America sequence (Frame 20): (a) extraction of the facial region prior to morphological processing, (b) shape processing of object, C_1 (i.e. facial region) in hue region, HR_1, (c) shape processing of object, C_2 in hue region, HR_2, and (d) shape processing of object, C_3 in HR_2.

Table 4
Foreman (width × height = 176 × 144): shape and color analysis

Object	Centroid location				Orientation		Object ratio		Mean hue		Aggregation
C_i	X	μ_1	Y	μ_2	θ°	μ_3	r	μ_4	H_m (°)	μ_5	μ_c
1	76	1	82	1	12.8	1	1.46	1	25	1	**1.0**
2	130	0	159	0	81.6	0.14	1.20	0.9	32	0.9	0.0
3	59	0	36	0	82.0	0.13	1.75	1	45	0.25	0.0
4	94	0	137	0	64.43	0.43	2.14	0.22	45	0.25	0.0

(a) Frame 20

Frames 20-130

(b) Frame 20

Frames 20-120

(c) Frame 20

Frames 20-110

Fig. 10. Location and tracking of the facial region for the following video sequences: (a) Claire, (b) Miss America, (c) Akiyo, (d) Foreman, (e) African–American sample image and (f) Carphone.

(d) Frame 20 Frames 20-95

(e)

(f) Frame 20 Frames 20-95

Fig. 10. Continued.

5. Conclusions

In this paper, a novel technique was proposed for the automatic location and tracking of the facial area in color video sequences. The attributes of color and shape were utilized in devising a three-stage segmentation scheme which consisted of a two-stage color processing unit, and a single-stage shape/color analysis module. The suggested method led to a consistent and accurate localization of the facial region and performed robustly for different skin types and various cases of object or background motion within the scene. The first stage of the color processing module was used to

extract the regions in the image that matched the hue characteristics of skin tones. This extraction process was formulated in the perceptual HSV color space by utilizing the a priori knowledge of the skin-tone distributions for various skin-type categories. The second stage in the color module was essentially used to refine the results of the initial extraction stage. In most cases, it was found that reasonable output could be obtained by excluding this second stage, thereby, decreasing the overall execution time of the algorithm. A number of binary post-processing operations were also included in the color processing unit to refine the shape of the segmented facial region. The computational complexity of these steps were minimal due to the binary nature of the operations. In many cases, only the facial area was extracted from the image, since no other objects in the scene possessed hue characteristics that were similar to the face. In a situation where more than one object was detected, then the final shape and color analysis stage provided the mechanism to correctly select the facial area. A compensative aggregation operator was used to combine the results from a series of fuzzy membership functions that were tuned for videophone-type applications. A number of features such as object shape, orientation, location and average hue were used to form the appropriate membership functions. The three-stage segmentation process appears to be quite promising and can be used with an additional feature extraction stage to provide higher level descriptions in future video coding environments.

References

[1] R. Brunelli, T. Poggio, Face recognition: Features versus templates, IEEE Trans. Pattern Anal. Mach. Intell. 15 (10) (1993).

[2] J. Bryant, On the clustering of multidimensional pictorial data, Pattern Recognition 11 (1979) 115–125.

[3] M.J. Carlotto, Histogram analysis using a scale-space approach, IEEE Trans. Pattern Anal. Mach. Intell. 9 (1) (1987) 121–129.

[4] L. Chiariglione, MPEG and multimedia communications, IEEE Trans. Circuits and Systems for Video Technol. 7 (1) (1997) 5–18.

[5] A. Eleftheriadis, A. Jacquin, Automatic face location detection for model-assisted rate control in H.261-compatible coding of video, Signal Processing: Image Communication 7 (4–6) (1995) 435–455.

[6] R.M. Evans, An Introduction to Color, Wiley, New York, 1948.

[7] J. Foley, A. van Dam, S. Feiner, J. Hughes, Computer Graphics, Principles and Applications, 2nd ed., Addison-Wesley, Reading, MA, 1990.

[8] Y. Gong, M. Sakauchi, Detection of regions matching specified chromatic features, Comput. Vision and Image Understanding 61 (2) (1995) 263–269.

[9] N. Herodotou, A.N. Venetsanopoulos, Image segmentation for facial image coding of videophone sequences, in: 13th Internat. Conf. on Digital Signal Processing, Santorini, Greece, July 1997.

[10] M. Hötter, Object-oriented analysis–synthesis coding based on moving two-dimensional objects, Signal Processing: Image Communication 2 (4) (1990) 409–428.

[11] A.K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[12] A.K. Jain, A. Vailaya, Image retrieval using color and shape, Pattern Recognition 29 (8) (1996) 1233–1244.

[13] M. Kirby, L. Sirovich, Application of the Karhunen–Loeve procedure for the characterization of human faces, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1) (1990).

[14] C.H. Lee, J.S. Kim, K.H. Park, Automatic human face location in a complex background using motion and color information, Pattern Recognition 29 (11) (1996) 1877–1889.

[15] H.G. Musmann, M. Hötter, J. Ostermann, Object-oriented analysis–synthesis coding of moving objects, Signal Processing: Image Communication 1 (2) (1989) 117–138.

[16] O. Nakamura, S. Mathur, T. Minami, Identification of human faces based on isodensity maps, Pattern Recognition 24 (1991) 263–272.

[17] R. Ohlander, K. Price, D.R. Reddy, Picture segmentation using a recursive region splitting method, Comput. Graphics and Image Process. 8 (1978) 313–333.

[18] Y. Ohta, T. Kanade, T. Sakai, Color information for region segmentation, Comput. Vision, Graphics and Image Process. 13 (1980) 222–241.

[19] I. Pitas, A.N. Venetsanopoulos, Nonlinear Digital Filters, Kluwer Academic Publishers, Massachusetts, 1990.

[20] K.N. Plataniotis, D. Androutsos, A.N. Venetsanopoulos, Multichannel filters for image processing, Signal Processing: Image Communication 9 (2) (1997) 143–158.

[21] C.A. Poynton, A Technical Introduction to Digital Video, Wiley, Toronto, 1996.

[22] M.J.T. Reinders, P.J.L van Beek, B. Sankur, J.C.A. van der Lubbe, Facial feature localization and adaptation of a generic face model for model-based coding, Signal Processing: Image Communication 7 (1) (1995) 57–74.

[23] F.S. Roberts, Measurement Theory with Applications to Decision-Making, Utility and the Social Sciences, Addison-Wesley, Reading, MA, 1979.

[24] J. Serra, Image Analysis and Mathematical Morphology, Academic Press, New York, 1982.

[25] T. Sikora, The MPEG-4 video standard verification model, IEEE Trans. Circuits and Systems for Video Technol. 7 (1) (1997) 19–31.

[26] W. Skarbek, A. Koschan, Colour image segmentation – A survey, Technical Report 94-32, Technical University of Berlin, Dept. of Computer Science, October 1994.

[27] K. Sobottka, I. Pitas, Face localization and facial feature extraction based on shape and color information, in: 1996 IEEE Internat. Conf. on Image Processing, Vol. 3, Lausanne, Switzerland, September 1996, pp. 483–486.

[28] D.C. Tseng, C.H. Chang, Color segmentation using perceptual attributes, in: Proc. 11th Internat. Conf. on Pattern Recognition, Vol. 3, 1992, pp. 228–231.

[29] T. Uchiyama, M.A. Arbib, Color image segmentation using competitive learning, IEEE Trans. Pattern Anal. Mach. Intell. 16 (12) (1994) 1197–1206.

[30] A. Witkin, Scale-space filtering, in: Proc. IJCAI-83, August 1983, pp. 1019–1022.

[31] A.L. Yuille, Deformable templates for face recognition, J. Cognitive Neurosci. 3 (1) (1991) 59–70.

[32] H.J. Zimmermann, P. Zysno, Latent connectives in human decision making, Fuzzy Sets and Systems 4 (1980) 37–51.