A FUZZY-BASED FACE TRACKING SCHEME

N. Herodotou[†] *, K. N. Plataniotis, and A. N. Venetsanopoulos*

Department of Electrical and Computer Engineering
University of Toronto
10 King's College Road
Toronto, Ont., M5S 3G4, CANADA
http://www.dsp.toronto.edu

ABSTRACT

In this paper, a technique is presented to locate and track the facial areas in videophone-type sequences. The proposed face localization method consists of essentially two components: i) a color processing unit, and ii) a fuzzy-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 latter shape and color analysis module is used to correctly identify the facial regions when falsely detected objects are 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 regards 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.

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 [1], [2].

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. 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.

2. COLOR IMAGE SEGMENTATION

The detection and automatic location of the human face is important and vital in numerous applications including human recognition for security purposes, human-computer interfaces, and more recently, for video coding, and content-based storage/retrieval in image and video databases. Several techniques based on shape and motion information have recently been proposed for the automatic location of the facial region [3, 4]. The former technique is based on fitting an ellipse to a thresholded binary edge image while the latter approach utilizes the shape of the thresholded frame differences. In our approach we use color as the primary tool in detecting and locating the facial areas in a scene with a complex or moving background.

Color is a key feature used to understand and recollect the contents within a scene. It is also found to be a highly reliable attribute for image retrieval as it is generally invariant to translation, rotation, and scale changes [5]. 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 [6]. 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 segmentation of the skin areas within an image is most effective when a suitable color space is selected for the task, as mentioned earlier. This is the case when the skin clusters are compact, distinct, and easy to extract from the color coordinate system.

[†]The author is now with Mediamatics of National Semiconductor Corporation, 48430 Lakeview Boulevard, Fremont, CA 94538.

The complexity of the algorithm must also be low to facilitate real-time applications.

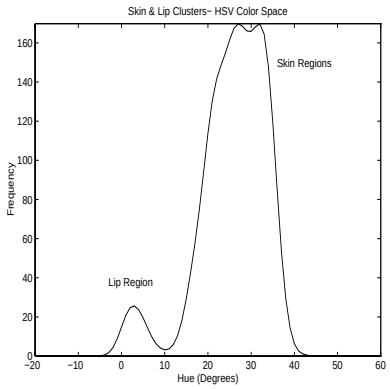

Figure 1: **Skin and Lip Hue Distributions in the HSV color space model.**

In Figure 1, the Hue component of the skin and lip clusters from the HSV space are shown. The graph illustrates that the spectral composition of the skin and lip areas are distinct, and compact. Skin clusters #1 and #2 are contained between the Hue range of 10 and 40° while the lip region lies at a mean Hue value of about 2° (i.e. close to the Red Hue value at 0°). Thus, the skin clusters are well partitioned allowing the segmentation to be performed by a thresholding scheme in the Hue axis rather than a more expensive multidimensional clustering technique. The HSV model is also advantageous in that the mean Hue of the skin values can give us an indication of the skin tone of the facial region in the image. Average Hue values closer towards 0° contain a greater amount of reddish spectral composition while those towards 60° contain greater yellowish spectral content. This can be useful for content-based storage and retrieval for MPEG 4 and 7 applications as well as multimedia databases. On the contrary, central cluster values in the other coordinate systems, (i.e. $[R_c\ G_c\ B_c]^T$ or $[L^*_c\ a^*_c\ b^*_c]^T$) do not provide the same meaningful description to a human observer.

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 in the previous section. We examine the distribution of these clusters next.

The identification and tracking of the facial region is determined by utilizing the apriori knowledge of the skin-tone distributions in the HSV color space outlined above. It has been found that skin-colored clusters form within a rather well defined region in chromaticity space [9], and also within the HSV hexcone model [10], for a variety of different skin types. In the HSV space in particular, the skin distribution was found to lie predominantly within the limited Hue range between 0°−50° (Red-Yellow), and in certain cases within 340°−360° (Magenta-Red) for darker skin types. The Saturation component suggests that skin colors are somewhat saturated, but not deeply saturated, with varying levels of intensity.

The Hue component is the most significant feature in defining the characteristics of the skin clusters. However, the Hue can be unreliable when: 1) the level of brightness (i.e. Value) in the scene is low, or 2) 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 are found in the achromatic regions of a scene. Thus, we must define appropriate thresholds for the Value, and Saturation components where the Hue attribute is reliable. We have defined the following polyhedron that corresponds to skin colored clusters with well defined Saturation and Value components, based on a large sample set [?]

$$T_{hue1} = 340^\circ \leq \mathrm{H} \quad \leq T_{hue2} \quad = 360^\circ \tag{1}$$

$$T_{hue3} = 0^\circ \leq \mathrm{H} \quad \leq T_{hue4} \quad = 50^\circ \tag{2}$$

$$\mathrm{S} \quad \geq T_{sat1} \quad = 20\% \tag{3}$$

$$\mathrm{V} \quad \geq T_{val} \quad = 35\% \tag{4}$$

The extent of the above Hue range is purposely designed to be quite wide so that a variety of different skin-types can be modeled. As a result of this, however, other objects in the scene with *skin-like* colors may also be extracted. Nevertheless, these objects can be separated by analyzing the Hue histogram of the extracted pixels. The valleys between the peaks are used to identify the various objects that possess different Hue ranges (i.e. facial region and different colored objects). Scale space filtering [11] is used to smoothen the histogram and obtain the meaningful peaks and valleys. This process is carried out by convolving the original Hue histogram, $f_h(x)$, with a Gaussian function, $g(x,\tau)$ of zero mean and standard deviation, τ as follows

$$F_h(x,\tau) = f_h(x) * g(x,\tau) \tag{5}$$

$$F_h(x,\tau) = \int_{-\infty}^{\infty} f_h(u) \frac{1}{\sqrt{2\pi}\,\tau}\, exp[\frac{-(x-u)^2}{2\tau^2}]\, du \tag{6}$$

where $F_h(x,\tau)$ represents the smooth histogram. The peaks and valleys are determined by examining the first and second derivatives of F_h above. 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 module that follows provides the necessary discriminatory functionality.

A series of post-processing operations which include median filtering, and region filling/removal are subsequently used to refine the regions obtained from the initial extraction stage. Median filtering is the first of two post-processing operations that are performed after the initial color extraction stage. The median operation is introduced in order to smoothen the segmented object silhouettes and also eliminate any isolated misclassified pixels that may appear as impulsive-type noise. Square filter windows of size 5×5, and 7×7 provide a good balance between adequate noise

suppression, and sufficient detail preservation. This operation is computationally inexpensive since it is carried out on the bi-level images (i.e. object silhouettes). 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. At this point, one or more of the extracted objects correspond to the facial regions. 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. We utilize two morphological operators to overcome this problem and at the same time smoothen the facial contours. A morphological closing operation is first used to fill in small holes and gaps, followed by a morphological opening operation which is used to remove small spurs and thin channels [13]. Both of these operations maintain the original shapes and sizes of the objects. A compact structuring element such as a circle or square without holes can be used to implement these operations and also help to smoothen the object contours. Furthermore, these binary morphological operations can be implemented by low complexity *hit or miss* transformations [13].

The results at this point contain one or more objects that correspond to the facial areas within the scene. The block diagram in Figure 2 summarizes the proposed face localization procedure. The shape and color analysis unit, described next, provides the mechanism to correctly identify the facial regions.

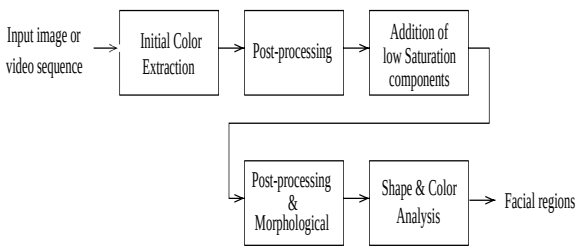

Figure 2: **Overall scheme to extract the facial regions within a scene.**

3. SHAPE AND COLOR ANALYSIS

The input to the shape and color analysis module may contain objects other than the facial areas. Thus, the function of this module is to identify the actual facial regions from the set of candidate objects. In order to achieve this, a number of expected facial characteristics such as shape, color, symmetry, and location are used in the selection process. Fuzzy membership functions are constructed in order to quantify the expected values of each characteristic. Thus, the value of a particular membership function gives us an indication of the *goodness of fit* of the object under consideration with 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 our segmentation and localization scheme we utilize a set of features that are suitable for our application purposes. In facial image databases (i.e. employees, models, etc.), or videophone-type sequences (video archives of newscasts, interviews, etc.), the scene consists of predominantly upright faces which are contained within the image (i.e. not typically at the edges 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 recognition task. These features can be determined in a simple and fast manner as opposed to measurements based on facial features such as the eyes, nose, and mouth which may be difficult to compute (i.e. in certain images the features may be small or occluded). More specifically, we consider (i) the dviation from the average Hue value of the different skin-type categories, (ii) the face aspect ratio, (iii) the vertical orientatio, and (iv) the relative position of the facial region in the image plane as primitives in our face localization system. A number of membership function models can be constructed and empirically evaluated. A trapezoidal function model is utilized here for each primitive in order to keep the complexity of the overall scheme to a minimum. The Hue characteristics of the facial region (for different skin-type categories) were used to form the first membership function. This function is built using the discrete universe of discourse $[-20^\circ, 50^\circ]$ (i.e. $-20^\circ = 340^\circ$). 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. Thus, the membership function associated with the first primitive is defined as follows

$$\mu(x) = \begin{cases} \frac{(x+20)}{28} & , \text{ if } -20^\circ \leq x \leq 8^\circ \\ 1 & , \text{ if } 8^\circ \leq x \leq 30^\circ \\ \frac{(50-x)}{20} & , \text{ if } 30^\circ \leq x \leq 50^\circ \end{cases} \quad (7)$$

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 [7]. However, in certain images 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

$$\mu(x) = \begin{cases} \frac{(x-0.75)}{0.5} & , \text{ if } 0.75 \leq x \leq 1.25 \\ 1 & , \text{ if } 1.25 \leq x \leq 1.75 \\ \frac{(2.25-x)}{0.5} & , \text{ if } 1.75 \leq x \leq 2.25 \\ 0 & , \text{ otherwise} \end{cases} \quad (8)$$

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. A reasonable threshold selection of 30° can be made for valid head rotations also observed within our database. 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

$$\mu(x) = \begin{cases} 1 & , \text{ if } 0^\circ \leq x \leq 30^\circ \\ \frac{(90-x)}{60} & , \text{ if } 30^\circ \leq x \leq 90^\circ \end{cases} \quad (9)$$

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 following membership function has been defined for the position of a candidate object with respect to either the X or Y-axis

$$\mu(x) = \begin{cases} \frac{(x-(d))}{\frac{d}{2}} & , \text{ if } d \leq x \leq \frac{3d}{2} \\ 1 & , \text{ if } \frac{3d}{2} \leq x \leq \frac{5d}{2} \\ \frac{((3d)-x)}{\frac{d}{2}} & , \text{ if } \frac{5d}{2} \leq x \leq 3d \\ 0 & , \text{otherwise} \end{cases} \quad (10)$$

The membership function for the X-axis is determined by letting $d = \frac{D_x}{4}$ 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 $d = \frac{D_y}{4}$ where D_y represents the vertical dimensions of the image (i.e. in the Y-direction).

3.1. Aggregation Operators

The individual membership functions expressed above, must be appropriately combined to form an overall decision. A number of fuzzy operators can be used to combine or fuse together the various sources of information. Conjunctive type of operators weigh the criterion with the smallest membership value more heavily while disjunctive ones assign the most weight to the criterion with the largest membership value. Here, we utilize a compensative operator which offers a compromise between conjunctive and disjunctive behaviour. This type of operator is defined as the weighted mean of a (*logical AND*) and a (*logical OR*) operator

$$A \bigodot_{\gamma} B = (A \bigcap B)^{1-\gamma} \cdot (A \bigcup B)^{\gamma} \quad (11)$$

where A, and B are sets defined on the same space and represented by their membership functions. 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 as follows

$$\mu_c = \prod_{j=1}^{m} \mu_j^{(1-\gamma)} \left(1 - \prod_{j=1}^{m} (1-\mu_j)\right)^{\gamma} \quad (12)$$

where μ_c is the overall membership function which combines all the knowledge primitives for a particular object, and μ_j is the j^{th} elemental membership value associated with the j^{th} primitive. The weighting parameter γ is interpreted as the *grade of compensation* taking values in the range of $[0, 1]$. The product and the possibilistic sum however, are not the only operators that may be used [14]. A simple and useful *t-norm* function is the *min* operator while the corresponding one for the *t-conorm* is the *max* operator. These operators were selected to model the compensative operator (i.e. overall fuzzy membership function), which assumes the form a weighted product as follows

$$\mu_c = \left(\left(\min_{j=1}^{m} \mu_j\right)\left(\max_{j=1}^{m} \mu_j\right)\right)^{0.5} \quad (13)$$

where the grade of compensation $\gamma = 0.5$ provides a good compromise of conjunctive and disjunctive behaviour [14]. The aggregation operator defined in (13) is used to form the final decision based on the designed primitives.

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

The scheme outlined in Figure 2 was used to locate and track the facial region in a number of still images and video sequences. The results from a videophone type sequences *Akiyo* (i.e. newscast or interview-type sequences) are reported here. A parameter selection of $\tau = 2$ was made in the Gaussian function of equation (5) in order to smoothen the histograms. This provided adequate smoothing, and was found to be appropriate for the skin-tone distribution models. A similar value [12] has also been suggested in the HVC space. The shape and color analysis module was used to identify the facial regions from the set of candidate objects. An object was classified as a facial region if its overall membership function, μ_c exceeded a predefined threshold of 0.75. In Figure 3 (b), the facial region was successfully identified and tracked for the *Akiyo* sequence. Two candidate objects were extracted in this case, and once again, the face was correctly selected based on the aggregation values.

In conclusion, this paper presents a technique to locate and track the facial areas within videophone type sequences. The attributes of color and shape were utilized. The technique was found to be of relatively low computational complexity due to the 1D histogram procedure, and the binary nature of the post-processing operations involved. In the case where more than one candidate object was detected, then the fuzzy-based shape and color analysis module 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 proposed fuzzy-based face tracking scheme 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.

5. REFERENCES

[1] H.G. Musmann, M. Hotter, J. Ostermann, 'Object-oriented analysis-synthesis coding of moving objects',

Signal Processing: Image Comm., Vol. 1, No. 2, pp. 117-138, Oct. 1989.

[2] L. Chiariglione, 'MPEG and Multimedia Communications', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 7, No. 1, pp. 5-18, February 1997.

[3] A. Eleftheriadis, A. Jacquin, 'Automatic face location detection for model-assisted rate control in H.261-compatible coding of video', *Signal Processing: Image Communication*, Vol. 7, No. 4 pp. 435-455, November 1995.

[4] 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*, Vol. 7, No. 1 pp. 57-74, March 1995.

[5] A.K. Jain, A. Vailaya, 'Image Retrieval Using Color and Shape', *Pattern Recognition*, Vol. 29. No. 8, pp. 1233-1244, 1996.

[6] T. Uchiyama, M.A. Arbib, 'Color Image Segmentation Using Competitive Learning', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 12, pp. 1197-1206, December 1994.

[7] C.H. Lee, J.S. Kim, K.H. Park, 'Automatic Human Face Location in a Complex Background using Motion and Color Information', *Pattern Recognition*, Vol. 29, No. 11, pp. 1877-1889, 1996.

[8] J. Foley, A. van Dam, S. Feiner, J. Hughes, *Computer Graphics, Principles and Applications*, Addison-Wesley Publishing Company, Inc., Second Edition, 1990.

[9] T.C. Chang, T.S. Huang, C. Novak, 'Facial Feature Extraction from Color Images', *Proc. of the 12th Int. Conf. on Pattern Recognition*, Vol. 3, pp. 39-43, 1994.

[10] N. Herodotou, A.N. Venetsanopoulos, 'Image Segmentation for Facial Image Coding of Videophone Sequences', *13th International Conference on Digital Signal Processing*, Vol. 1, pp. 233-236, 1997.

[11] M.J. Carlotto, 'Histogram Analysis Using a Scale-Space Approach', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 9, No. 1, pp. 121-129, 1987.

[12] Y. Gong, M. Sakauchi, 'Detection of Regions Matching Specified Chromatic Features', *Computer Vision and Image Understanding*, Vol. 61, No. 2, pp. 263-269, March 1995.

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

[14] K.N. Plataniotis, D. Androutsos, A.N. Venetsanopoulos, Multichannel filters for image processing, *Signal Processing: Image Communications*, Vol. 9, No. 2, pp. 143-158, 1997.

Frame 20

Frames 20-110

Figure 3: **Location and tracking of the facial region for Akiyo.**