EYES 'N EARS FACE DETECTION

B. Kapralos[1,3], M. Jenkin[1,3], E. Milios[2,3] and J. K. Tsotsos[1,3]

[1] Dept. of Computer Science, York University, Toronto, Ontario, Canada. M3J 1P3
[2] Dept. of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada. B3H 1W5
[3] Centre for Vision Research, York University, Toronto, Ontario, Canada. M3J 1P3

{billk, jenkin, eem, tsotsos}@cs.yorku.ca

ABSTRACT

We present a robust and portable visual-based skin and face detection system developed for use in a multiple speaker teleconferencing system, employing both audio and video cues. An omni-directional video sensor is used to provide a view of the entire visual hemisphere, thereby allowing for multiple dynamic views of all the participants. Regions of skin are detected using simple statistical methods, along with histogram color models for both skin and non-skin color classes. Regions of skin belonging to the same person are grouped together, and using simple spatial properties, the position of each person's face is inferred. Preliminary results suggest the system is capable of detecting human faces present in an omni-directional image despite the poor resolution inherent with such an omni-directional sensor.

1. INTRODUCTION

Existing teleconferencing systems utilize conventional cameras, thereby providing a limited number of static or manually tracked views. Consequently, in a multiple speaker setting, either the speaker must move into the camera's view or a camera operator is used to manually locate and track and choose between speakers. This is both bothersome and inconvenient for the participants and has deterred many from using such systems. In addition to the cost overhead, the presence of a camera/equipment operator during a teleconferencing session may interfere with the group dynamics [10]. Although visual based systems capable of detecting and tracking human faces exist, they also employ conventional camera lenses, which capture only a narrow field of view.

Teleconferencing systems must be able to capture and transfer audio (e.g. the speaker's voice). As a result, in a multiple speaker setting, the teleconferencing system must be able to localize a speaker in the audio domain as well. Although sound localization systems exist, most rely on extensive audio arrays which require very expensive and specialized equipment, are computationally intensive and are non-portable.

This paper describes the video component of the Eyes 'n Ears teleconferencing system, which is responsible for automatically detecting the face of all participants, estimating the position of each face in the "real world" and providing this information to the audio system as possible sound source positions (or directions).

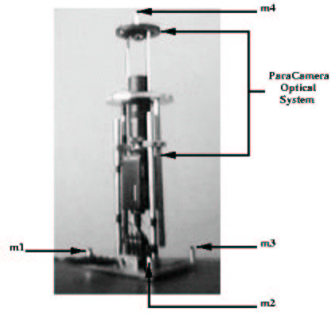


Fig. 1. Eyes 'n Ears Sensor.

1.1. Overview

Figure 1 illustrates the hardware components comprising the Eyes 'n Ears sensor. The sensor is comprised of a ParaCamera optical system coupled with a microphone array. The sensor is compact, lightweight, portable and is meant to be placed in the middle of a table with the participants of the teleconference session seated around it. The following sections describe the audio and video components in greater detail.

2. AUDIO COMPONENT

As shown in figure 1, four omni-directional microphones ($m_1 \ldots m_4$), mounted in a static pyramidal shape about the base of the ParaCamera provide an economical and portable acoustic array. Given the initial position (or direction) estimate of each person's face, using beam-forming and sound detection techniques [4], the audio system detects each speaker and focuses on their speech, thereby reducing unwanted noise and sounds emanating from other locations. Greater details regarding the Eyes 'n Ears system may be found in [6].

3. VIDEO COMPONENT

Rather than requiring a user move into the camera's field of view, or manually moving the camera to locate a speaker, Eyes 'n Ears utilizes Cyclovision's ParaCamera omni-directional optical system [7]. The ParaCamera consists of a high precision paraboloidal mirror and a combination of special purpose lenses. By aiming a video

camera to the face of the paraboloidal mirror, the combination of these optics permit the ParaCamera to capture a 360^o (*hemispherical*) view of potential speakers from a single viewpoint. Once the hemispherical view has been obtained, it may be easily un-warped [8], producing a panoramic view. From this panoramic, perspective views of any size corresponding to different portions of the scene may be extracted easily.

3.1. Algorithm Overview

In the Eyes 'n Ears system, the ParaCamera is used to identify potential speakers in the environment. Using a statistical color model similar to that described by Jones *et al*. [5] but suitably modified for use with the inherent low resolution ParaCamera images, the video system locates regions of human skin present within the ParaCamera's view (hemispherical image). These skin regions correspond to faces, arms and other exposed skin regions. Regions of skin in the camera image, which are spatially close, are then grouped together to form a cluster. Assuming there is a reasonable amount of space between participants in view, each cluster of skin regions is assumed to correspond to one particular person. With the assumption that the face of a person in the ParaCamera image is further away from the center of the image relative to the other skin regions, one can easily locate the region of each cluster corresponding to the participant's face. Once each face has been found, an estimate of its position (or direction) in the real world is made and provided to the audio system as a potential position of (or direction to) a sound source.

3.2. Skin and Non-Skin Histograms

Two-dimensional hue-saturation histograms for both skin and non-skin color classes were constructed by manually classifying portions of images obtained with the ParaCamera, as either skin or non-skin. The RGB color pixels from each sample were converted to their corresponding HSV values. Value was ignored, while hue and saturation values were discretized to 32 and eight discrete values respectively (e.g. each histogram contains a total of $32 \times 8 = 256$ possibles hue saturation pairs or *bins*). 88,888 skin pixels were obtained from 30 subjects of various ethnic groups to ensure a wide variety of skin colors. For each subject, samples of their hands, face (and in several cases legs) were obtained over several images. The subjects were asked to change both their pose and their distance relative to the ParaCamera in each image to ensure samples in different lighting and orientation conditions were obtained. Similarly, the 223,728 non-skin pixel samples comprising the non-skin histogram were obtained by sampling portions of images obtained with the ParaCamera which did not contain human skin. Figure 2 illustrates the resulting hue-saturation histograms for both the skin and non-skin classes. For each "bin", the axis labeled *Distribution*, records the number of pixels from the samples with corresponding hue-saturation values.

As shown in Figure 2a, human skin color in hue-saturation space forms a tight cluster. Over half of the skin samples (52.6%) are distributed among three bins with hue saturation values of (3,1), (3,2) and (3,3). On the contrary, as illustrated in Figure 2b, there is a scattering of the non-skin pixel colors throughout a larger region in hue-saturation space.

3.3. Finding Skin Regions

Given the skin and non-skin histograms described above, the estimated probability that a particular hue and saturation pair (referred to as "HS") belongs to either the skin, $P(HS|skin)$ or non-skin $P(HS|\neg skin)$ classes may be found using basic probability theory (see [9] for a review of probability theory and [5] for an earlier application of probability theory to the skin detection problem);

$$P(HS|skin) = \frac{skin[HS]}{T_s} \tag{1}$$

$$P(HS|\neg skin) = \frac{\text{non-skin}[HS]}{T_n} \tag{2}$$

where $skin[HS]$ and non-skin$[HS]$ is the count in bin HS of the skin and non-skin histograms respectively. T_s is the total number of pixels contained in the skin histogram while T_n is the total number of pixels contained in the non-skin histogram.

When $skin[HS]$ is less than a pre-defined threshold value δ, the probability that a pixel is skin color $P(skin|HS)$, in the hue-saturation color space, is set to zero. Otherwise, $P(skin|HS)$ is determined using Bayes' rule

$$P(skin|HS) = \frac{P(HS|skin)P(skin)}{P(HS|skin)P(skin) + P(HS|\neg skin)P(\neg skin)} \tag{3}$$

where $P(skin) = \frac{T_s}{T_s+T_n}$ and $P(\neg skin) = \frac{T_n}{T_s+T_n}$ is the probability of a pixel belonging to the skin and non-skin color classes respectively. Once the probability that a particular HS pixel value corresponds to skin has been calculated, the pixel is classified as skin if the probability is less than a pre-defined threshold.

Figure 2c provides a graphical representation of $P(skin|HS)$ for each of the hue-saturation pairs given the skin and non-skin color models illustrated in figures 2a,b.

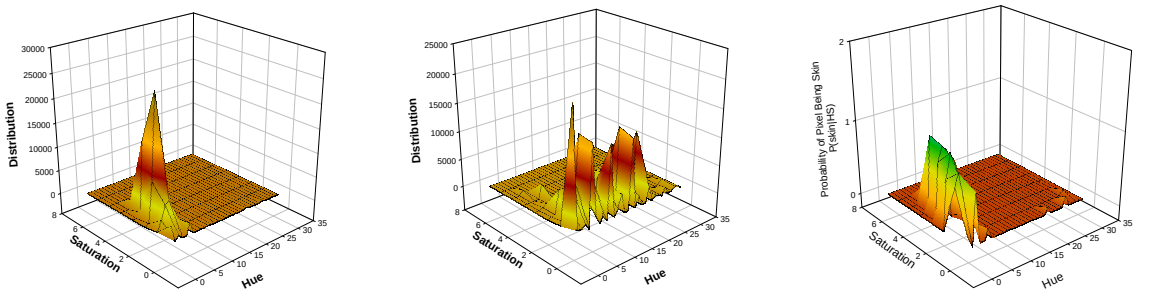
3.4. Temporal Coherence Process

To limit the probability of incorrectly classifying non-skin pixels as skin, change detection with an adaptive background reference image is used. During system initialization, a reference image I_{ref} without any participants present is obtained. Once the system has started, prior to applying the skin classification process to an incoming image I_{crnt}, a difference image I_{dif}, between I_{ref} and I_{crnt} is calculated. The difference image records the changes in intensity between corresponding pixels of both images. A pixel will only be classified as skin if and only if, its corresponding pixel in the difference image also registers as "changed".

To limit the number of incorrectly detected changes in the difference image, *temporal integration* using equation 4 [1], is used to periodically update the reference image.

$$I_{ref_{new}} = (\alpha \times I_{crnt}) + ((1-\alpha) \times I_{ref}) \tag{4}$$

where α is the update factor and $I_{ref_{new}}$ is the updated reference image. Temporal integration allows for both adaptation to changes in lighting (intensity) and ensures background objects remain part of the background even if moved from their initial position, blending such objects back into the background. To limit the number of computations performed, background adaptation is performed once every 10 frames.



(a) Skin Histogram. (b) Non-Skin Histogram. (c) Probability of a Pixel Being Skin.

Fig. 2. Skin Classification Using Skin and Non-Skin Color Models.

3.5. Finding Human Faces

Once pixels are classified as skin or non-skin, erosion and dilation operators are applied to remove isolated skin classified pixels. The remaining pixels are are then grouped into labeled regions using an 8-neighbor connected components operator and any components smaller than a pre-defined threshold size are eliminated. A search is then conducted to cluster connected regions which are spatially close. Assuming there is a reasonable amount of space between participants in view (the exact spacing between participants has not been calculated but informal surveys suggest a reasonable amount of space is approximately $1m$), each cluster of skin regions is assumed to correspond to a particular person. Given the geometry of the ParaCamera, the region of each cluster furthest from the center of the ParaCamera image is chosen as the face. Once each face has been found, an estimate of its position (or direction) in the real world is made and provided to the audio system.

3.6. Converting to World Coordinates

To convert positions from a single ParaCamera image to position x, y, z in the real world, a ground-plane perpendicular to the optical axis of the ParaCamera is assumed [2]. Informal lab surveys indicate the height of most people seated in a chair falls within a small region. The average height of the face of 10 people seated in a chair was found to be $1.20m$ above the floor. As a result, the height of the ground-plane ("h") for this application is chosen to be $1.20m$ above the floor. Furthermore, it is assumed that the face of each person intersects this ground-plane.

As illustrated in Figure 3, since the focus of the paraboloid is also the center of projection, the line rl (reflection line), between the focus of the paraboloidal mirror and the point of the reflection of the head (face) on the mirror's surface intersects the head's location in the real world (the point on the ground-plane). In addition, the azimuth angle in this ground-plane is the same as the azimuth angle in the image. As a result, by extending this line so that it intersects the ground-plane, the world (x, y) coordinates at this point of intersection may be found [3].

World coordinates of a pixel in the ParaCamera image with coordinates i, j are obtained by first converting Cartesian image coordinates (i, j) into polar coordinates (r, θ). Once the polar coordinates have been determined, the x and y world coordinates of

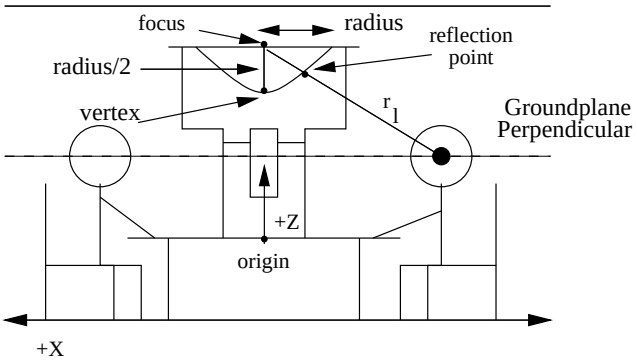


Fig. 3. Converting Image Coordinates to World Coordinates.

the point of intersection obtained by extending line rl until it hits the ground-plane are found as done in [3] using

$$z(r) = \frac{r_{\text{mirror}}^2 - r^2}{2 \times r_{mirror}} \tag{5}$$

$$x_{\text{world}} = \frac{hr \sin\theta}{z} \qquad y_{\text{world}} = \frac{hr \cos\theta}{z} \tag{6}$$

where $z(r)$ is the surface equation of the paraCamera's paraboloidal mirror, r_{mirror} is the radius in pixels of the ParaCamera image and x_{world}, y_{world} are the x and y world (room) coordinates respectively and $z_{world} = 1.20m$.

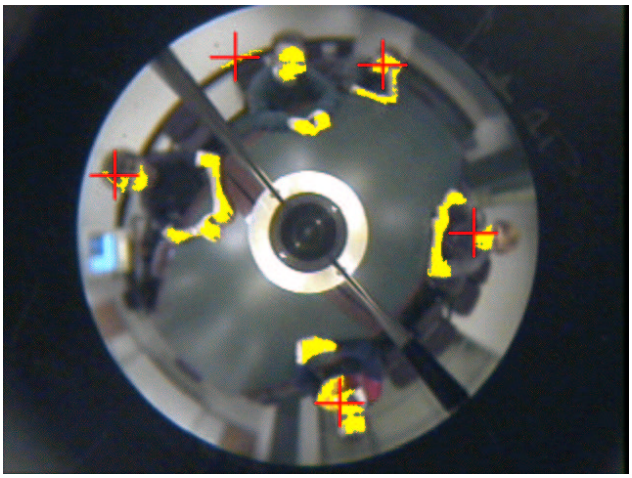
The ground-plane assumption can be relaxed when a direction to the participant, as opposed to a position, is of interest. This corresponds to the *far* field acoustical model. However, the position is required (and therefore the ground-plane assumption) for the *near* field acoustical model.

4. EXPERIMENTAL RESULTS

4.1. Overall Accuracy of the Video System

The skin and face detection process as described earlier was applied to 22 images (obtained with the ParaCamera), where each

(a) Original ParaCamera Image



(b) ParaCamera Image After Skin Classification

Fig. 4. An Example of the Face Detection Process. Yellow regions of image (b) correspond to the detected skin regions. The red "cross hairs" denote the position of the face as detected by the system.

image contained between zero and five people. Of the 108 regions of skin present in the 22 images, 100 (92%) were correctly detected while 33 non-skin regions were incorrectly classified as skin. Similarly, of the 37 faces present in the images, 32 (86%) were correctly detected, four were not detected and three non-face skin regions were incorrectly classified as a face. The skin regions corresponding to the four faces which were not detected were correctly classified as skin. However incorrectly classified regions of skin (which happen to fall within the cluster of skin regions corresponding to each person and were slightly farther from the ParaCamera image center than the actual face region) were chosen as the face.

A sample of the face detection technique is provided in Figure 4. Figure 4b illustrates the result of applying the skin pixel classification and face detection process to the image of figure 4a.

4.2. Position of Faces in the "Real World"

Each subject was seated in a chair at a known world position (x, y coordinates). The subject's height (and therefore the height of the ground-plane perpendicular) was kept constant at $1.20m$ above the floor (e.g. the height of the ground-plane perpendicular was $1.20m$ above the floor.). An image of the subject was then obtained with the ParaCamera and the skin and face detection process described in Section 3 was performed. The mean error and standard deviation associated with the $x-axis$ estimation was $0.06m$ and $0.04m$ respectively. Similarly the $y-axis$ estimation produced a mean error of $0.16m$ and a standard deviation of $0.23m$.

5. CONCLUSIONS

A simple, portable and robust visual based skin and face detection system for use in a teleconferencing application has been presented. Using model histograms for both skin and non-skin color classes along with simple statistical methods, temporal and spatial properties, the video component is capable of detecting exposed regions of human skin and faces present in images obtained with the ParaCamera with an accuracy of 92%. Similarly, the video component is capable of providing an estimate of the position of each detected face in the "real world" with a mean error rate of $0.06m$ and $0.16m$ in the x, y axis respectively.

References

[1] T. Boult, R. Michaels, P. Gao, C. Lewis, W. Yin, and A. Erkan. Frame rate omni-directional surveillance and tracking of camouflaged and occluded targets, 1998. http://www.eecs.lehigh.edu/ tboult/TRACK/LOTS.html.

[2] K. Danilidis. Personal communication.

[3] D. Gutchess, A. Jain, and S. Cheng. Automatic surveillance using omni-directional and active cameras. In *Proc. Asian Conf. Comput. Vis.*, 2000.

[4] D. Johnson and D. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Prentice Hall, USA, 1993.

[5] M. Jones and J. Rehg. Statistical color models with applications to skin detection. Technical Report CRL 98/11, Compaq Computer Corp., Cambridge, MA USA, 1998.

[6] B. Kapralos. Eyes 'n ears: A system for attentive teleconferencing. Master's thesis, Department of Computer Science, York University, Toronto, Ontario, Canada, April 2001.

[7] S. Nayar. Omnidirectional video camera. In *Proc. DARPA Image Understanding Workshop*, pages 235–241, New Orleans, LA, 1997.

[8] V. Peri and S. Nayar. Generation of perspective and panoramic video from omnidirectional video. In *Proc. DARPA Image Understanding Workshop*, pages 243–245, New Orleans, LA USA, 1997.

[9] J. Pitman. *Probability*. Springer Verlag, New York, NY USA, 1993.

[10] R. Yong, A. Gupta, and J. Cadiz. Viewing meetings captured by an omni-directional camera. In *ACM Trans. Comput.-Hum. Interact.*, March 2001.