

# Detection and Tracking of Faces in Real-Time Environments

R.C.K Hua, L.C. De Silva and P. Vadakkepat  
Department of Electrical and Computer Engineering  
National University of Singapore  
4 Engineering Drive 3 Singapore 117576

**Abstract**--We propose a new approach to detect and track faces in a real-time video feed, that is computationally-inexpensive and does not require any specialized hardware. A 2-dimensional color model is built to capture the inherent chrominance of the human skin color. The 2-dimensional nature of the color model, with some additional morphological processing and filtering, ensures real-time responses when it is used to delineate facial regions in a real-time video feed. The result is a PC-based system, that is capable of detecting multiple faces and tracking a dominant face to keep it in view, given an arbitrary real-time video feed, regardless of the size, orientation and viewpoint of the faces. We were able to attain face detection accuracies as high as 87.5% and pan-tilt tracking errors as low as 15 pixels in a video sequence with 640 x 480 resolution.

**Keywords**--Camera motion control, face detection, face tracking, image processing, real-time.

## 1. INTRODUCTION

Recent heightened security concerns have made the need for robust face detection and tracking more imperative, especially in the areas of security and surveillance. In addition, face detection and tracking are also crucial in applications such as teleconferencing, face and facial gesture recognition, telecommunications, robotics as well as human-computer interactions (HCI).

The goal of face detection is to determine the presence of any human face in an image, or a sequence of images, and return its location and spatial extent. Since faces of different sizes may appear in arbitrary locations with a variety of orientations in an image, or a sequence of images, this task greatly involves scale, space, orientation, and viewpoint analysis. Many methods have been proposed to solve this problem and these have been best summarized by Ahuja et al. in [1], which categorized face detection into 4 main methods: knowledge-based

techniques, feature-invariant approach, template-matching and appearance-based approach. The systems implemented in [2] and [3] are examples of realized systems that utilize knowledge-based techniques, employing rules derived from our knowledge of human faces to detect them. Yang et al. [2] uses a hierarchical knowledge-based method while Kotropoulos et al. [3] seeks to extend the work done in [2], producing a success rate of 86.5%. Proponents of the feature-invariant methods, such as Waibel et al. [4], Stiefelhagen et al. [5] and Jebara et al. [6], seek to localize feature invariants of faces, such as face skin color, for detection and have done so with reasonable successes. Template-matching methods such as those employed in [7], [8] and [9], on the other hand, utilize the correlation values of an input image with standard patterns of a human face, to perform the task of face detection. Appearance-based methods, unlike template-matching methods, employ templates that are not pre-defined and that require statistical analysis or machine learning from examples to find the relevant face characteristics. Examples of systems using such methods include [10] and [11], of which the later achieved a success rate of 90% using 2-dimensional pseudo Hidden Markov Models.

For our purpose of implementing a video-based face-detection and tracking system, it is decided that the use of the feature invariant approach with facial color as the invariant feature to aid in the detection of faces will be most appropriate. The usage of skin color as a feature for tracking a face has several advantages. One advantage is the fact that processing facial skin color is much faster than processing any other facial features. Another advantage is that facial skin color is orientation invariant under certain lighting conditions. But however, color is not a physical phenomenon. Instead, it is a perceptual phenomenon that is related to the spectral characteristics of electro-magnetic radiation in the visible wavelengths striking the retina. This poses numerous problems. Firstly, the color representation of a face obtained by a camera is

influenced by many factors such as ambient light, object movement etc. Secondly, different cameras produce significantly different color values even for the same person under the same lighting conditions. Thirdly, facial skin colors differ from person to person, varying from dark facial skin tones to lighter ones. And lastly, with the camera moving and tracking faces, the face detection method must be tolerant of the optical flow resident within the successive image frames that are grabbed. In order to use color as a criteria for face detection and tracking, these problems have to be solved. Our proposed method seeks to overcome all these to successfully detect and track faces, regardless of their sizes, skin tones and orientations.

## 2. PROPOSED METHOD

### 2.1 Facial Skin Color Model

Although skin color appears to vary, it is our hypothesis that there exists underlying similarities in the chromatic properties of all faces and that all major differences lie in intensity rather than in the facial skin color itself. As such, using the feature invariant method described earlier, we adopted a skin color model based on the YUV color space but with the non-critical illuminance value discarded. This disregard of the illuminance value will not affect the successful detection of facial skin color but will more importantly reduce the color space substantially from a bulky 3-dimensional YUV color space to a manageable 2-dimensional UV map. This helps reduce the computational processing required for each image frame.

To determine the definition of facial skin color on this UV map, repeated experimentations were carried out to find a way to encapsulate the inherent properties of chrominance that is common to all faces, regardless of the degrees of variation in skin tones. The results from these experimentations are the best-fit regions for 3 separate categories of facial skin tones, namely the light skin tone, the medium skin tone and the dark skin tone. Each category is represented by a specific ethnic group. The Caucasian ethnic group is judged to have light skin tone while the Chinese ethnic group is judged to have a medium skin tone. The Indian ethnic group, on the other hand, is judged to have dark skin tone.

The best-fit region on the UV map that captures the most of the light facial skin tone is illustrated in Figure 1.

For the medium and dark skin tones, the corresponding best-fit regions are illustrated in Figure 2 and 3 respectively. As can be observed, these 3 facial skin tones share a common region on the UV map and this continuous region, which forms the final 2-dimensional facial skin color model, is illustrated in Figure 4.

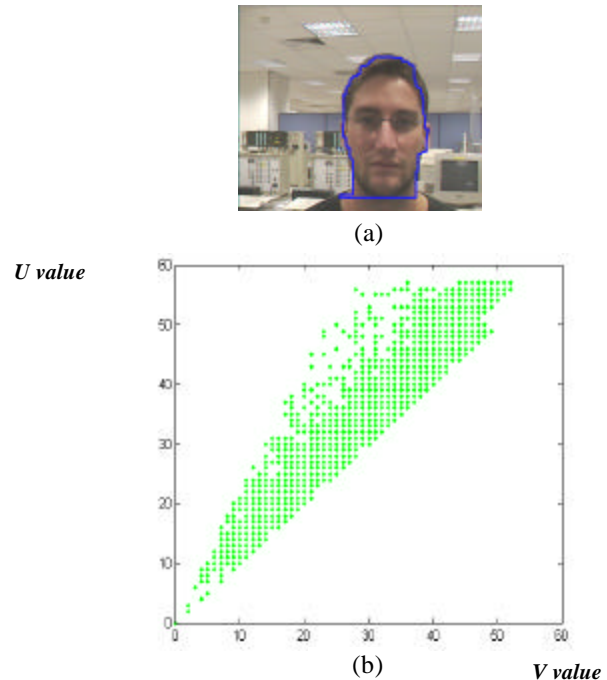


Figure 1: Best-fit region on UV map for light facial skin tones UV map

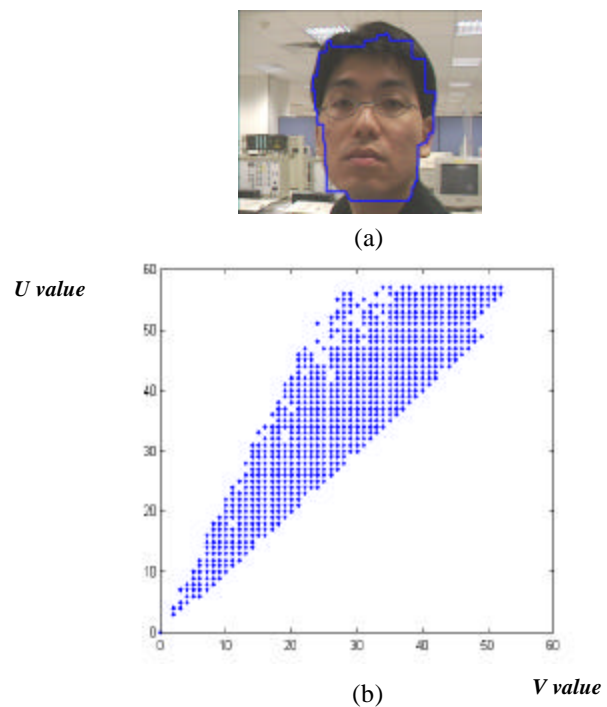
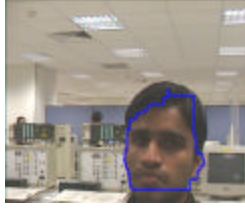


Figure 2: Best-fit region on UV map for medium facial skin tones UV map UV map



(a)

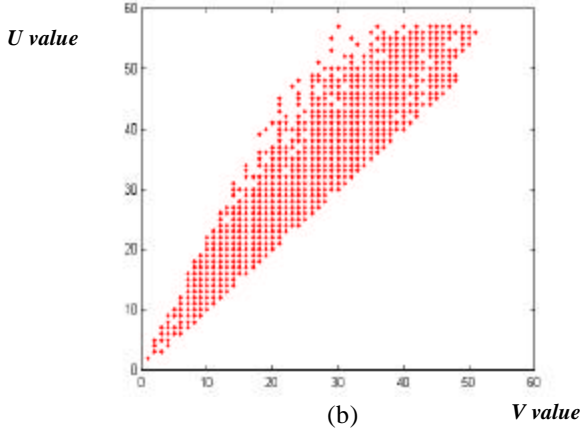


Figure 3: Best-fit region on UV map for dark facial skin tones UV map UV map

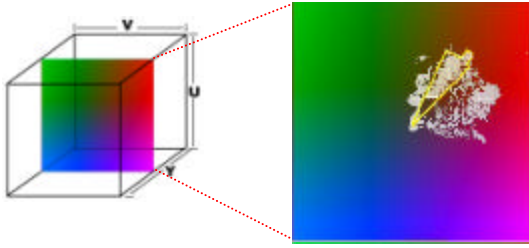


Figure 4: Conversion of 3D YUV color space to 2D UV map

This region, delineated in Figure 4, can be represented mathematically by the following sets of equations, (1) and (2) respectively.

$$0.2611\mathbf{p} \leq \tan^{-1}\left(\frac{u}{v}\right) \leq 0.3111\mathbf{p}$$

and  $43 \leq \sqrt{u^2 + v^2} \leq 78$

$$0.25\mathbf{p} \leq \tan^{-1}\left(\frac{u}{v}\right) \leq 0.3611\mathbf{p}$$

and  $0 \leq \sqrt{u^2 + v^2} \leq 70$

where  $u$  and  $v$  are the U and V color values of each image pixel.

A pixel is identified to have facial skin color if its corresponding position in the UV map fulfills either one of the 2 sets of equations. In this way, the facial skin color area of the input image can be extracted. Figure 5 shows a sample result after the facial skin color regions have been extracted from an image.



Figure 5: Preliminary results of skin color extraction

To eliminate random "salt & pepper" noise, median filtering is performed after facial skin color extraction. Subsequently, to remove all other extraneous background pixels that may have been erroneously identified as having facial skin color, morphological operations of dilations and erosions are employed. Erosions and dilations are then performed as part of an image closing operation to fill small and thin holes in the detected facial skin regions and to smooth their boundaries without significantly changing their area. Figure 6 illustrates the results after median filtering and morphology. Figure 6(a) shows the image before median filtering and morphology. Figure 6(b) illustrates the noise-removing effect of median filtering while Figure 6(c) shows the final output image after morphology. In Figure 6(c), it can be seen that the face has been clearly extracted.



(a)

(b)

(c)

(1a) Figure 6: Results of skin color extraction after processing

(1b) Subsequently, the detected facial region is subjected to heuristic rules, which are based on the geometric analysis of the general human face. As established in [14], "a merged skin color region [can be] a human face candidate if the ratio of major axis to minor axis is less than a threshold of 1.7". The application of this heuristic rule is the final part of the face detection and tracking module of the system. With the face area extracted, the result of the face detection is then displayed on screen by utilizing the smallest circle, that encloses the entire face

area extracted, to delineate the detected face's spatial location, as illustrated in Figure 7. In this way, faces can be detected despite variations in face orientations and partial face occlusions.

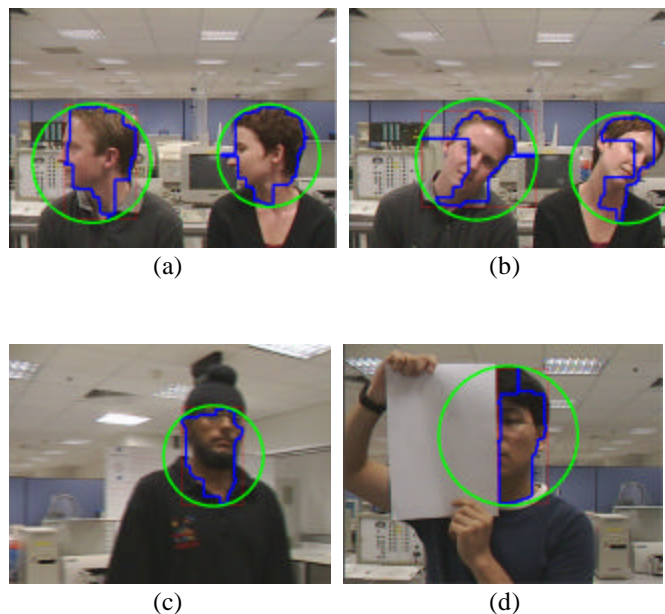


Figure 7: Test subjects from various ethnic groups in a variety of face orientations

## 2.2 Pan-Tilt Control of Camera

With the face region extracted, the next task is to manipulate the pan-tilt camera intelligently such that the dominant face detected is kept in view at all times in the center of the image frame. The logic flow of the camera motion control implemented is illustrated in Figure 8. The most important phase of this pan-tilt control of the camera lies in the use of a closed-loop feedback system, illustrated in Figure 9, in which a pair of proportional controllers is employed to correct the positional difference between the center of the detected face and the center of the image frame. Each basic proportional controller is implemented here in software to ensure that the camera pans or tilts at the appropriate degrees. One controller corrects the horizontal difference while the other resolves the vertical difference.

Their proportional control actions are similarly implemented, with the relationship between the output of each controller  $u(t)$ , and the actuating error  $e(t)$  to be

$$u(t) = K_p e(t) \quad (3)$$

where  $K_p$  is the proportional gain,  $u(t)$  is the extent of pan or tilt movement required and  $e(t)$  is the pixel difference along one axis.

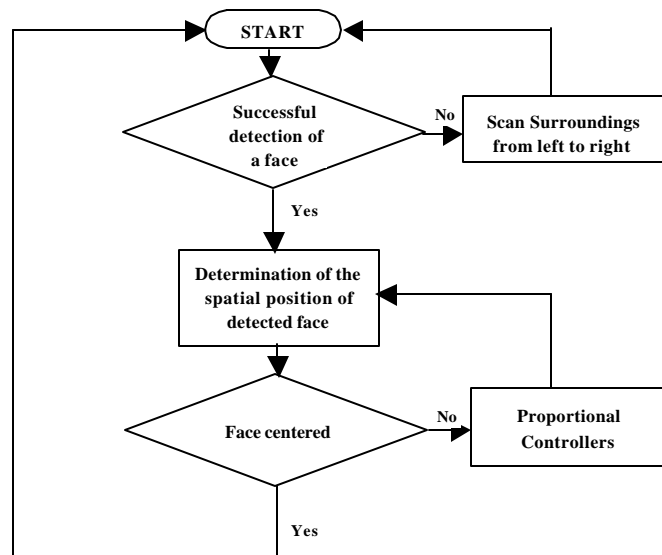


Figure 8: Flowchart for execution of pan-tilt control

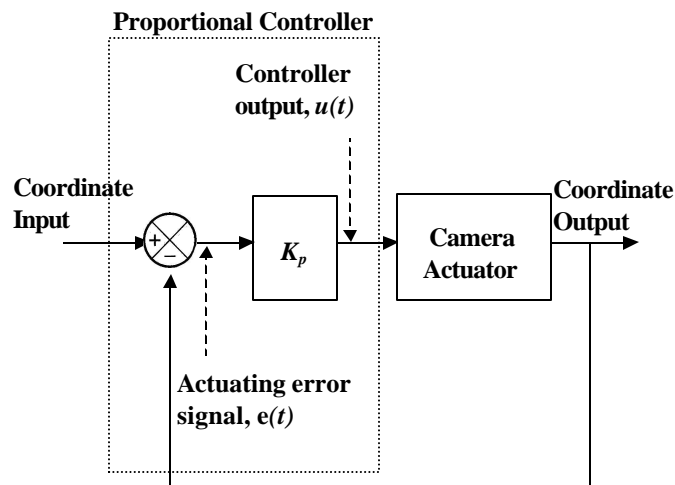


Figure 9: Block diagram of the closed-loop feedback camera control

The most appropriate  $K_p$  value for our purposes has been found to be 0.001. In this way, amplifiers with adjustable gains are realized to send the appropriate control signals to the pan-tilt camera to keep the detected face in the center of the image frame, realizing the tracking of the face. The detected face can thus be kept in view at all times. Results of the face tracking are illustrated in Figure 10 and 11.

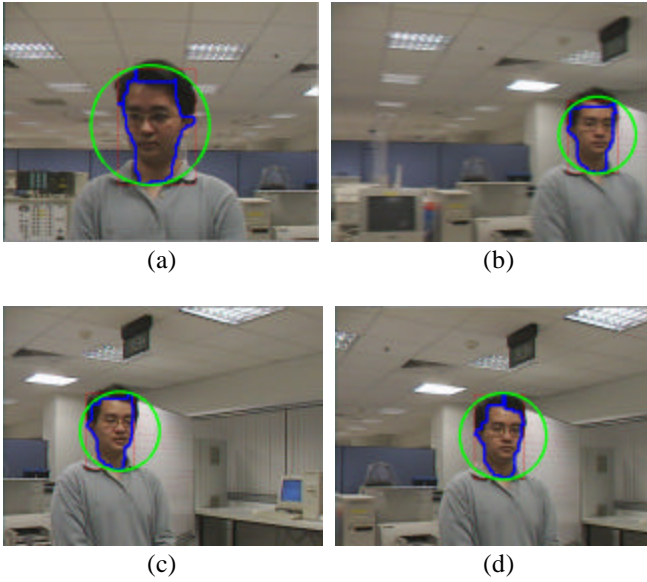


Figure 10: Image frames from the camera as it tracked a Chinese test subject moving out of view

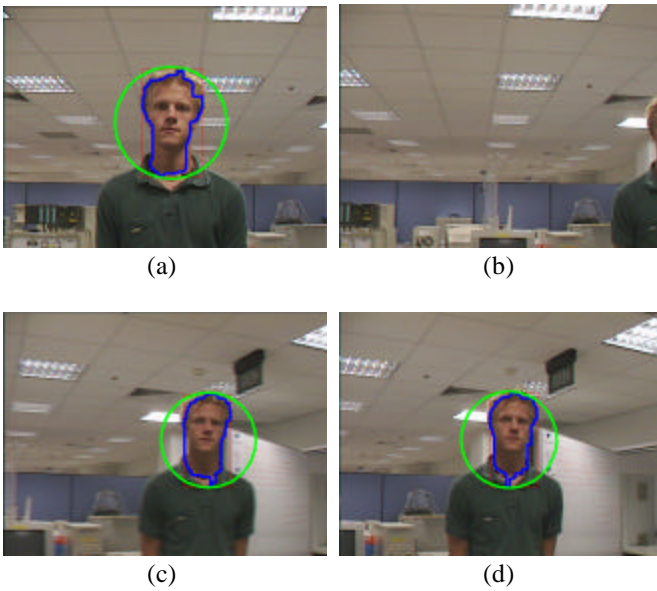


Figure 11: Image frames from the camera as it tracked a Caucasian test subject moving out of view

### 3. EXPERIMENT RESULTS

The system was comprehensively tested with 12 male and 4 female test subjects, of whom 8 are Chinese, 4 are Caucasians and 4 are Indians. This selection of diverse ethnic groups in the test group was to ensure that the system is robust enough to accommodate varying facial skin tones. 3 separate sets of tests were commissioned to establish the system's (i) face detection capability, (ii)

face tracking capability and (iii) pan-tilt motion control functionality respectively.

#### 3.1 Tests on Face Detection

In the first set of tests, each individual test subject varied his/her face orientation while remaining stationary in front of the camera. The face detection rates obtained are tabulated in Table 1.

Table 1: Face detection rates for singular faces

Face Orientation	Face Detection Rate (%)
Front Profile	87.50
Left Profile	87.50
Right Profile	87.50
Tilted 90 <sup>0</sup> to the left	87.50
Tilted 90 <sup>0</sup> to the right	87.50
Tilted upwards	68.75
Tilted downwards	75.00
Half of face occluded	50.00

Multiple face detection rates are also established by having 2 test subjects repeat their face-orientation variations simultaneously in front of the camera. In this scenario, the successful detection rates are slightly lower than those for individual test subjects. The multiple face detection rates obtained are tabulated in Table 2.

Table 2: Face detection rates for multiple faces

Face Orientation	Face Detection Rate (%)
Front Profile	81.25
Left Profile	75.00
Right Profile	87.50
Tilted 90 <sup>0</sup> to the left	75.00
Tilted 90 <sup>0</sup> to the right	75.00
Tilted upwards	75.00
Tilted downwards	81.25
Half of face occluded	34.38

#### 3.2 Tests on Face Tracking

The second set of tests establishes the face tracking capability of the system. Each individual test subject moves in a prescribed fashion in front of the camera and the error of the face track is determined. This error is defined as the pixel difference between the center of the test subject's face and the center of the circle that denotes the spatial location of the detected face. The face track errors for the 16 test subjects are tabulated in Table 3.

Table 3: Face track errors for 16 test subjects

Subject No.	Square root of Mean Square Error along x axis (pixels)	Square root of Mean Square Error along y axis (pixels)
1	10	38
2	18	19
3	20	19
4	6	39
5	19	11
6	26	5
7	35	26
8	23	29
9	9	12
10	17	20
11	24	13
12	13	46
13	12	16
14	11	10
15	23	35
16	33	49

From Table 3, the best-case and worst-case errors can be determined. The plots of the actual face track errors along the x-axis for these cases are given in Figure 12 while those along the y-axis are furnished in Figure 13.

### 3.3 Tests on Pan-Tilt Motion Control

The third set of tests establishes the accuracy of the system's pan-tilt motion control. Each individual test subject moves in a prescribed fashion such that each subject moves out of the view of the camera, activating the pan-tilt motion of the camera. The error of the pan-tilt motion control, defined as the pixel difference between the center of the image frame and the center of the detected face, is then determined. The pan-tilt motion control errors for the 16 test subjects are tabulated in Table 4. The plots of the actual pan motion errors for the best and worst cases are given in Figure 14 while those of the tilt motion errors are furnished in Figure 15.

## 4. HARDWARE AND SOFTWARE SPECIFICATIONS

Our system is based on a Pentium III 450MHz PC, which communicates control instructions to a Sony EVI-D31 Pan-Tilt camera via the serial communications port using the VISCA™ network protocol. The video feed from the camera is captured using a Matrox Meteor-II color analog frame-grabber.

Table 4: Pan-tilt motion control errors for 16 test subjects

Subject No.	Square root of Mean Square Error of Pan motion (pixels)	Square root of Mean Square Error of Tilt motion (pixels)
1	107	69
2	126	28
3	140	37
4	108	52
5	112	36
6	93	15
7	96	26
8	101	47
9	89	21
10	82	42
11	121	37
12	108	46
13	123	34
14	97	26
5	93	78
16	84	21

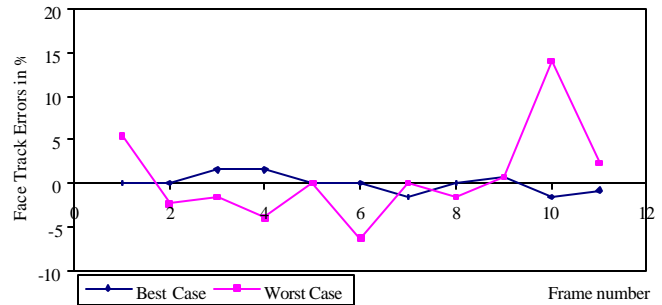


Figure 12: Plots of face track errors along x-axis

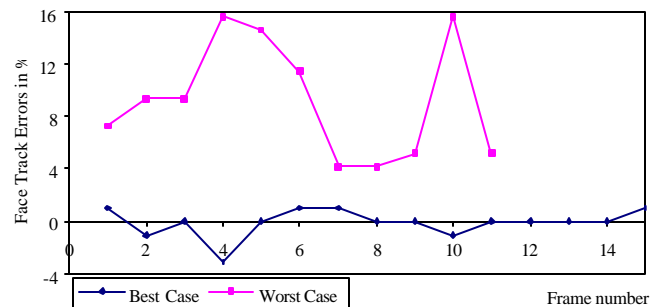


Figure 13: Plots of face track errors along y-axis

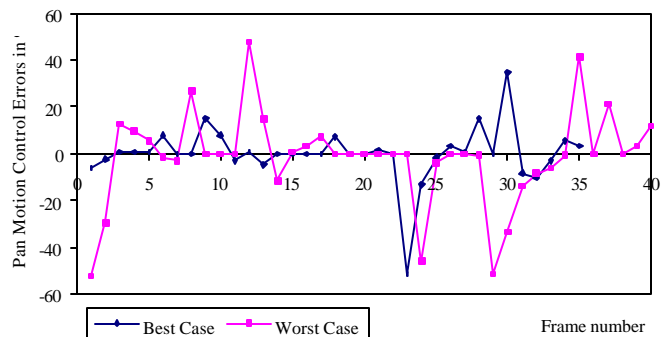


Figure 14: Plots of pan motion control errors

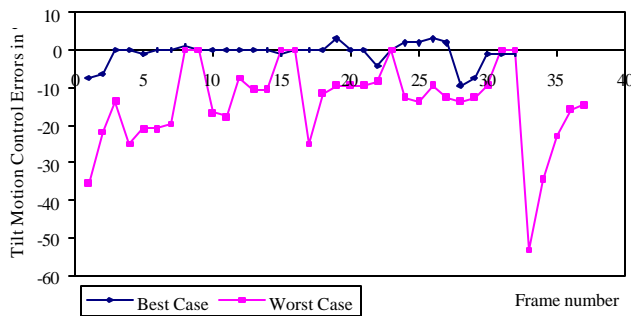


Figure 15: Plots of tilt motion control errors

## 5. CONCLUSION AND FUTURE WORK

In this paper, we have presented a new method to detect and track faces given an arbitrary video feed. The proposed method utilizes a 2-dimensional skin color model to extract the facial skin color regions. The delineated faces are then tracked by manipulating the pan-tilt camera to keep the face in view. Experimental results show that such a method is robust enough to ensure successful detection and tracking of faces even under conditions where the size, orientation and viewpoint of faces are varied. Future areas of development can include integrating face recognition into the system so that its propensity to be used a commercial security system can be fully realized. The system can also be incorporated into a teleconferencing system so as to allow the conferencing participants to move freely around the room while the system keeps the participants in view.

## ACKNOWLEDGEMENT

The support of the Philips CFT RoboCup Team, Netherlands is gratefully acknowledged.

## REFERENCES

- [1] Ming-Hsuan Yang, Narendra Ahuja and David Kriegman, "A Survey on Face Detection Methods", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, no. 1, pp. 34-58, 2002.
- [2] G. Yang and T.S. Huang, "Human face detection in complex background", *Pattern Recognition*, vol. 27, no. 1, pp. 53-63, 1994.
- [3] C. Kotropoulos and I. Pitras, "Rule-based face detection in frontal views", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol.4, pp. 21-24, 1997.
- [4] J. Yang and A. Waibel, "A real time face tracker", *Proceedings of the Third Workshop on Applications of Computer Vision*, pp. 142-147, 1996.
- [5] J. Yang, R. Stiefelhagen, U. Meier, and A. Waibel, "Visual tracking for multimodal human computer interaction", *Proceedings of SIGCHI 98*, pp. 140-147.
- [6] T.S. Jebara and A. Pentland, "Parameterized structure from motion to 3D adaptive feedback tracking of faces", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 144-150, 1997.
- [7] V. Govindara, D.B. Sher and R.K. Shari, "A computational model for face location", *Proceedings of the Third International Conference on Computer Vision*, pp. 718-721, 1990.
- [8] V. Govindara, "Locating human faces in photographs", *International Journal of Computer Vision*, vol. 19, no. 2, pp. 129-146, 1996.
- [9] V. Govindara, D.B. Sher and R.K. Shari, "Locating human face in newspaper photographs", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 549-554, 1989.
- [10] F. Samaria and S. Young, "HMM based architecture for face identification", *Image and Vision Computing*, 12, pp.537-583, 1994.
- [11] F.S. Samaria, "Face Recognition Using Hidden Markov Models", PhD thesis, University of Cambridge, 1994.
- [12] H. P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan, "Multimodal system for locating heads and faces", *Proceedings of the Second IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 88-93, 1996.
- [13] J. Yang, W. Lu, and A. Waibel, "Skin-color modeling and adaptation", *Technical Report CMU-CS-97-146*, School of Computer Science, Carnegie Mellon University, 1997.
- [14] Ming-hsuan Yang and Narendra Ahuja, "Detection of human faces in color images", *Proceedings of the International Conference on Image Processing*, 1998, pp. 127-130.