REAL-TIME HUMAN FACE TRACKING

WENBO PAN

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Real-Time Human Face Tracking

By

© Wenbo Pan, B.Sc., M.Sc.

A thesis submitted to the School of Graduate Studies
in partial fulfillment of the requirements for
the degree of Master of Science

Computational Science Programme
Memorial University of Newfoundland
August, 2000

St. John's Newfoundland Canada

Abstract

A real-time human face tracking system has been developed at the Multimedia Communication Laboratory of Memorial University of Newfoundland to investigate fast, efficient, reliable and flexible face finding and tracking techniques. By subtracting a well-maintained background from an incoming image, an object and background segmentation map can be constructed. The foreground object is outlined by a "draping" operation on the segmentation map. Once the drape is settled, an innovative head identification method consisting of exhausting head searching followed by head merging achieves accurate head extraction and identification. In order to tackle the problems associated with variations in lighting, local and global background movements and shadows in the background scenes, a multi-state background self-generating/adjusting method is applied. This allows the system to switch automatically between background formation and simple face tracking. The draping is applied on the inter-frame variance of incoming images to identify moving areas and thus to generate the background. Median filtering, multiple direction draping and polynomial interpolation are developed and incorporated into this system to overcome the possible pitfalls in the resultant drape. The background is updated automatically in real time once the changes in the background are detected to exceed a given threshold. Experiments show that the new real-time system is a robust and effective tool for extracting human heads from a very complex non-stationary background.

Acknowledgements

I would like to express my deep appreciation to my supervisor, Dr. John A. Robinson, for his great support, enduring patience, conscientious guidance and encouragement in the past two years, which brought my study to fruition.

I would like to say thanks to Li-Te Cheng for his wonderful MCLGallery and his kind help. I also want to say thanks to Qing Song and Charles Robertson for sharing their experience with me. I want to say thanks to Reza who took part in my experiments and everyone else at the MCL lab. I do enjoy the time I have spent with them.

I would like to take this chance to express my special appreciation to my husband, Guanjun, for his constant support; my mother and my daughter, for accompanying me during my study; my father, for his oversea moral encouragement.

I also want to say thanks to Dr. Lagowski for her generous support through my two years of study.

The financial support jointly provided by Graduate Studies of Memorial University of Newfoundland and Northern Telecom (NORTEL) is also greatly appreciated.

Table of Contents

List of Figures

Chapter One

Introduction

Real-time face tracking is currently an active research area in the computer vision community due to the possibility of being able to tackle more complex problems using readily available fast computers. A robust real-time face tracking system has many practical applications such as advanced video communication, virtual reality interfaces, smart rooms, very low bandwidth video compression, human computer interaction and security monitoring. All of these have in common the need to track and interpret the human body, especially the human face (Wren et al. 1996). Face tracking can be used to reduce communication bandwidth by locating and transmitting only the fraction of a video frame containing the speaker's face. In a human computer interaction application, face tracking can be used to direct the computer's attention to a user and increase the probability of the computer correctly recognizing the user's facial expressions, gesture, or speech. In video conferencing, it is desirable to have a computer-controlled camera to isolate the image of a special talker within a frame, adjusting for orientation and range as well as compensating for any source motion. A video surveillance and monitoring system

requires face tracking to be achieved in various situations such as lighting change, shadows, presence of moving elements in the background scene, and so on.

To address this need, a fast and reliable real-time face tracking system was developed in the Multimedia Communication Laboratory (MCL) at Memorial University of Newfoundland (MUN). Figure 1.1 presents the system setup. The computerized camera system consists of a digital camera and a personal computer equipped with an image grabbing and image-processing system.

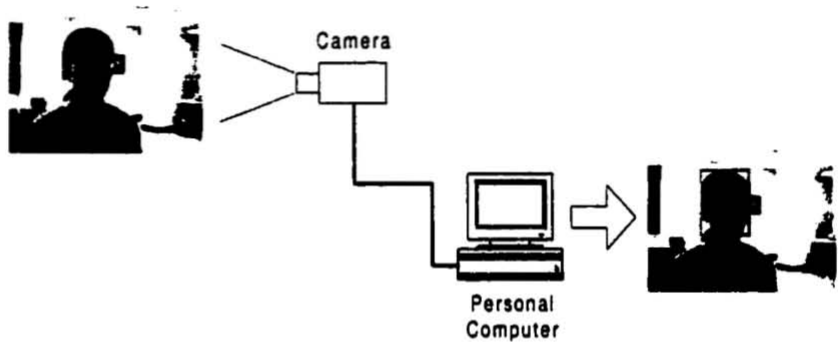


Figure 1.1 Schematic diagram of MCL real-time face tracking system

1.1 Problem Statement

This research aims to create a robust, adaptive real-time face tracking system that is flexible enough to handle variations in lighting, global background movement, moving objects in background scene, and any other arbitrary changes in the observed scene. Regular human faces in normal standing/sitting postures are of interest in this study.

Background subtraction is an effective and fast tool to locate foreground objects and hence it is used as the major technique of the face tracking system. It functions by subtracting a pre-selected background from the incoming image to recover the foreground objects. The significant advantage of this method is its low computational cost and therefore it can be easily implemented for real-time application. This method works accurately for stationary background scenes. Problems arise when the background becomes non-stationary due to changes in the background scene and similarities between the foreground and the background present. The background changes can be categorized as follows:

- Global background movements due to camera shifting, shaking, zoom in and out
- Local background changes due to local movement in the background scene
- Lighting changes
- Shadows

These problems presented in the background scene make it very difficult to accurately recover the desired foreground by using simple background subtraction and may lead to a catastrophic failure of the whole system. In addition, the system is required to track multiple heads at various postures and layouts from a complex background scene.

1.2 Structure of Thesis

This thesis describes a real-time face tracking system that can switch back and forth automatically between two states depending on the background information. The two states are head tracking and background formation. The first state applies when a

stable background estimate has been acquired. The second state applies at start-up or when the background changes, when it works to progressively build a background estimate by analyzing foreground movement. In chapter one, the entire system setup is introduced, together with its basic working principles. Problems associated with applications of the system are also addressed. Chapter two is a review of previous work in the relevant areas. Draping is identified as a useful tool for outlining the foreground object. Based on the reviewed works, techniques for background maintenance are proposed. Chapter three describes the principle of draping and its modification and improvement to suit our use. Chapter four presents an innovative head searching and merging technique developed in this study for head tracking, as well as gives some typical experimental results. Chapter five describes the multi-state real-time system, along with techniques for tackling the changing background. Additionally, typical experimental results are presented and discussed. Chapter six concludes my thesis and offers directions for further study.

Chapter Two

Literature Review

A system for object tracking must consist of a camera to capture images, a frame grabber, a processor and a set of techniques to process the image, in order to search the image for important features, and then use these features to determine the location of the object. Techniques for people tracking have been studied extensively. Some of the relevant works from literature are reviewed, analyzed and discussed in this chapter. This review begins with the introduction of some real-time people tracking systems, and is followed by descriptions of face tracking based on background subtraction and various methods for background maintenance and modeling to recover a complete foreground object. Skin color filtering is also an effective tool for identifying human faces from a color image. Skin color based face finding methods are reviewed in section 2.3. Three-dimensional head tracking is a newly emerging research activity and has been growing rapidly in the last decade because of the exponential advance in computer power. In order to extract three-dimensional parameters, a model that can encode head orientation and

position is often required. Some three-dimensional face models for facilitating three-dimensional face tracking are presented in section 2.4.

2.1 People Tracking Systems

A variety of people tracking systems have been developed in the past. Wren et al. (1996) proposed a real-time system called Pfinder (person finder) for searching arbitrarily complex scenes for single people using a fixed camera. Pfinder employs a multi-class statistical model of color and shape to segment the people from the background scene. It can find and track a human's head and hands under a wide range of viewing conditions. Pfinder can be regarded as a descendant of the vision routines originally developed for the ALIVE system (Darrel et al. 1994). It is also related to body-tracking research using both kinematic models (Rehg and Kanade 1994; Rohr 1994; Gavrila and Davis 1995) and dynamic models (Pentland and Horowitz 1991; Metaxas and Terzopolous 1993). Functionally, Pfinder is most closely related to the work of Bichsel (1994) and Baumberg and Hogg (1994) in which the person is segmented from the background in real time using only a standard workstation. Pfinder makes several domain specific assumptions to enable the vision task tractable. Its performance degrades when these assumptions do not hold. Also, due to the assumptions on which it is built, Pfinder cannot tackle large, sudden changes in the background, and it can only work for one user in the scene.

KidsRoom (Bobick et al. 1996; Intille et al. 1997) is a perceptually-based, multi-person and fully-automated people tracking system. It is built on a "closed world" assumption, which defines a region of space and time where the specific context of what

is in the region is assumed to be known. Its people tracking system uses an overhead camera view of the space in order to minimize the possibility of one object occluding another. Lighting is assumed to remain constant during the time when the tracker is operating. Background subtraction is used to segment objects from the background, and foreground pixels are clustered into two-dimensional blob regions. The system then maps each person known to be in the room with a blob in the incoming image frame. It uses colors, velocity estimation, and size information to disambiguate the match when the blobs later separate. These regions can be tracked in real-time domains where object motions are not smooth or rigid, and where multiple objects are interacting. The system is, however, overly reliant on blob data, which may not always be reliable. The second limitation of this system is that it has no mechanism for handling the slow variation of image features while objects are merged in a large closed world. The third limitation is that its matching algorithm can lead to some bad matches.

W^4 (Haritaoglu et al. 1998a) is a real-time system for tracking people and their body parts in monochromatic images. It constructs dynamic models of people's movements to detect what they are doing, and where and when they act. It employs a combination of shapes (shape, hands, feet, and torso) to create models of people's appearance so that they can be tracked through interactions such as occlusions (who is in the scene?). The models are constructed by using a "cardboard" human model of a person in a standard upright pose. This limits the application of this technique to tracking people in an upright-standing posture. Horprasert et al. (1998) later extended their W^4 method to allow operations on color images by using a new background subtraction technique.

Multiple cameras are used to observe a person in this method. Silhouette analysis and template matching then achieve a real-time three-dimensional estimation of human posture. The estimated body postures are reproduced in a three-dimensional graphical character model by deforming the model according to the estimated data. Dynamics and kinematics models of human body and linear Kalman filtering are utilized to help the tracking process as well as to interpolate some joint locations. The real-time three-dimensional computer vision system provides the user with control over the movement of a virtual computer graphics character. The application of this technique is, however, still limited to the upright-standing posture.

In an effort to incorporate other generic postures, Haritaoglu et al. (1998b) developed a monocular system called Ghost, which functions under the control of W^4. It can estimate a variety of human body postures and detect body parts in real time. It constructs a silhouette-based body model to determine the location of the six main body parts (head, two hands, two feet and the torso) while a person is in a number of postures. It combines hierarchical body pose estimation, a convex hull analysis of the silhouette, and a partial mapping from the body parts to the silhouette segments using a distance transform method that does not violate the topology of the human body. The algorithm developed works not only in the upright-standing posture but also in other generic postures. The hierarchical posture representation includes main postures (standing, sitting, crawling-bending, and laying-down) which are further sub-classified into one of three view-based appearances (front-view, left-side, and right-side). Shadows appearing

in the silhouette might give rise to difficulty in locating body parts that are too close to the ground.

Turk (1996, 1998) proposed a real-time head tracking system based on draping. Draping simulates a row of point masses connected to each neighbor by a spring. Gravity pulls the drape down over the thresholded foreground object while the foreground pixels collectively hold the drape in place. The draping is applied on the people and background segmentation map to produce a "head and shoulders" silhouette. Once the people outline ("drape") settles it can be used to locate people's heads. All these procedures can be done in real time on a standard low-end PC. The resultant drape can be used in a coarse posture and gesture recognition. The significance of this method is its tolerance to a reasonable amount of noise and holes presented in the segmented images.

Olson (1997) developed a general purpose system for moving object detection and event recognition in which the moving objects are detected using change detection and are tracked based on first-order prediction and nearest neighbor matching. Events are recognized by applying predicates to a graph formed by linking corresponding objects in successive frames.

2.2 Background Subtraction and Background Maintenance Techniques

Perhaps the most often used face tracking method is background subtraction, in which the foreground pixels are separated from the background pixels by a simple image subtraction. All the systems discussed in section 2.1 are based on background subtraction. Background subtraction is straightforward and conceptually simple. However, the

difficult part of this method is not the subtraction, but the maintenance of a background model (Toyama et al. 1999). The success of this method depends heavily on the accuracy of the modeled background. The background maintenance must be able to deal with the problems associated with lighting change and both regional and global movements of the background. A standard method of background maintenance is background averaging. In this method, the images are averaged over time to produce a background approximation that is similar to the current static scene except where motion occurs. While this is effective in situations where objects move continuously and the background is visible for a significant portion of the time, it is not robust to scenes with many moving objects, particularly if they move slowly. It also cannot handle bimodal backgrounds; it recovers slowly when the background is uncovered, and has a single predetermined threshold for the entire image.

Wren et al. (1996) used a mean and covariance method to model the image background as a texture surface. In this method, the mean and covariance pixel values are continuously updated to adapt the changing background. Each point on the texture surface is associated with a mean color value and a distribution about that mean. The color distribution of each pixel is modeled with Gaussian distribution described by a full covariance matrix. In each frame, visible pixels have their statistics recursively updated using a simple adaptive filter. This allows compensation for changes in lighting and object movement. Stafford-Fraser (1996) proposed two methods for background formation. One way of constructing an evolving background frame is to use the average pixel values of a number of preceding images; the other is to capture a number of

"background" frames during the application. Stafford-Fraser calculated the mean and standard deviation of the values at each pixel position. The mean is used to initialize the frame and the standard deviations are combined with a global threshold value to give a threshold that is specific to that pixel position. The eigenbackground method (Pentland 1994; Oliver et al. 1999) collects images of motionless background and then uses principle component analysis (PCA) to determine the mean and variances over the entire sequence (whole images as vector). The incoming images are projected onto the PCA subspace. Differences between the projection and the current image greater than a threshold are considered as foreground.

Haritaoglu et al. (1998c) used a temporal derivative method to tackle changing backgrounds. In the training stage, both the minimum and the maximum values of each pixel are saved along with the maximum inter-frame change in intensity at each pixel. Any pixel that deviates from its minimum or maximum by more than the maximum inter-frame change is regarded as background. In addition, statistical tools are used to deal with changing backgrounds. In the mixture of Gaussian method (Wren et al. 1996; Friedman and Russell 1997; Grimson et al. 1998), a pixel-wise mixture of three Gaussians models the background, with each Gaussian weighted according to the frequency with which it explains ($\pm 2\sigma$) the observed background. The most heavily weighted Gaussians that explain over 50% of past data are considered as the background. The background maintenance method proposed by Nakai (1995) is based on Bayesian decision theory. Pixel value probability densities, represented as normalized histograms, are accumulated over time, and the background is determined by a maximum of a posteriori criterion.

The Wiener filter is a linear predictor based on a recent history of values. Any pixel that deviates significantly from its predicted value is regarded as foreground. The linear prediction method works well for periodically changing pixels. Its main advantage is that it reduces the uncertainty in a pixel's value by accounting for how it varies with time. To handle changing backgrounds, the prediction coefficients are recomputed for every new frame. Based on this concept, Toyama et al. (1999) developed Wallflower, a three-component system for background maintenance. In this technique, the pixel-level component performs Wiener filtering to make probabilistic predictions of the expected background; the region-level component fills in homogenous regions of foreground objects; and the frame-level component detects sudden, global changes in the image and swaps in better approximations of the background. The application of this method must satisfy the following requirements: (i) semantic differentiation of objects should not be handled; (ii) background subtraction should segment objects of interest when they first appear in a scene; (iii) an appropriate pixel-level stationary criterion should be defined; (iv) the background model must adapt to both sudden and gradual changes in the background; and (v) background models should take into account changes at differing spatial scales.

Dynamic contours, or snakes, provide an effective method for tracking complex moving objects for segmentation and recognition purposes. Snakes track object boundaries by minimizing the sum of an external force from a local image measure, and an internal force from a shape dynamics model. The external force drives the dynamic contour according to the current image appearance while the internal force increases the

spatial and temporal continuity of the tracked boundary. When the boundary to be tracked is an occluding boundary, the dynamic contour confuses background texture for the desired boundary. To compensate for this shortcoming, dynamic contours often rely on detailed object shapes or motion models to distinguish between the boundary of the tracked object and other boundaries in the background (Terzopolous and Szeliski 1992; Cootes et al. 1993; Blake and Isad 1994). An alternative solution proposed by Covell and Darrell (1999) uses simple contrast measures for the external energy term of dynamic contour models without detailed object models. The image model developed by them, called radial cumulative similarity (RCS), describes the local contrast pattern but is largely insensitive to the changes in background contrast. The use of RCS enables the occluding boundaries to be tracked in a cluttered scene, with the simplest of internal energy terms.

2.3 Skin Color Based Face Tracking Method

A different approach for locating and tracking a face is by searching for skin color. Color is a feature of the human face. Skin color based face tracking has several advantages. First, processing color is much faster than processing other facial features. Second, under certain lighting conditions, skin color is orientation invariant. Yang and Waibel (1996) developed a real-time face tracking system by incorporating three models. A stochastic model is developed to characterize the skin color distributions of the human face, which provides sufficient information for tracking a human face in various poses and views. A motion model is used to estimate image motion and to predict a search

window. A camera model is used to predict and to compensate for camera motion. The system can track a person's face while the person moves freely in a room. Qian et al. (1998) presented a statistical-based algorithm for estimating the position and size of a face in a complex background. The estimations are derived from two one-dimensional histograms in two orthogonal directions obtained by projecting the result of skin color filtering. The projection histograms can be interpreted as the spatial distributions of the skin pixels along the corresponding directions. The proposed method uses a linear Kalman filter and a simple nonlinear filter to perform smooth tracking and to remove jitter. This algorithm is computationally simple and is robust against cracks or gags within the face region.

Darrell et al. (1998) proposed a complete passive and non-invasive method for real-time person tracking in crowded and/or unknown environments using an integration of multiple visual modalities. This system combines stereo, color, and face detection modules. Dense, real-time stereo processing is used to isolate objects from other objects and people in the background. Skin-hue classification identifies and tracks likely body parts within the silhouette of a user. Face pattern detection discriminates and localizes the face within the identified body parts. Faces and bodies of users are tracked over short-term, medium-term and long-term, respectively. Short-term tracking is performed using simple region position and size correspondences, while medium and long-term tracking are based on the statistics of user appearance.

Another standard approach to finding faces in still images is based on some rigid features common to all face patterns, such as two dark eyes, bright nose ridge and the

spatial layout of facial organs (Qing and Robinson 1999). This technique can accurately identify human heads from stationary backgrounds. However, this approach is usually computationally expensive and hence it is difficult to use in real time.

2.4 Three-Dimensional Face Tracking

Extensive research has been conducted on locating and tracking human heads and recognizing poses and facial expressions. Most often face detection is considered as a two-dimensional problem where facial features, facial color, and the shapes of the face are obtained from the image plane for locating the head (Pentland et al. 1994; Rowley et al. 1998). To extract three-dimensional parameters, a model that can translate head orientation and position is often required. Azarbayejani et al. (1993) used feature point tracking projected on an ellipsoidal model to track the head position. A drawback of feature point tracking is that tracking fails when the feature points are lost due to occlusions or lighting variations. New feature points are required at the cost of excessive error accumulation. Jebara and Pentland (1997) also used feature point tracking, but with automatically located head features like eyes and mouth corners. The three-dimensional position of the feature points is estimated using a structure from a motion technique that pools position information over the image sequence with an extended Kalman filter. The estimate of the feature point position is filtered using Eigenfaces to restrict the measurements to match an expected facial geometry.

Basu et al. (1996) coupled an ellipsoidal model with general optical flow computation for tracking. The optical flow algorithm estimates the two-dimensional flow

field from image intensities. Optical flow is first computed independently of face position and orientation using a gradient-based method. In this method, the velocity is computed from a filtered version of the image. Then the motion of an ellipsoidal mesh regularizes the flow. The method's strengths are also its weakness. It copes well with large head rotations since it does not rely on any fixed features. For the same reason, it has no means to ground the model to the face; thus the error accumulates and the mesh slowly drifts off the face.

Black and Yacoob (1995) proposed a rectangular planar patch under affine transformation as a face model. In this model, similar patches are attached to the eyebrows and the mouth. The movements of the underlying facial patch are followed to detect differential movements of the facial parts. The limitation of using affine motion is that it has no concept of self-occlusion taking place at the sides of the head and around the nose. Affine transformations can also distort the frontal face image when they are used to model large rotation.

Cascia et al. (1998) employed a textured cylinder as a head model. The technique uses a dynamic texture for tracking. The lack of fixed features may lead to error accumulation although confidence maps are used to minimize this problem. DeCarlo and Metaxas (1996) used a polygonal head model that is hand-positioned on the subject's face. This technique extracts optical flow at some feature points and regularizes it by the model movements. The measurements are stabilized using a Kalman filter. The use of the optical flow leads to a similar error accumulation as does the method used by Basu et al.

(1996). Both techniques used face edge information to prevent a possible divergence. Face shape and facial expressions can also be extracted by both methods.

Schödl et al. (1998) developed a three-dimensional textured model of the human head under perspective projection to track a person's face. The system is hand-initialized by projecting an image of the face onto a polygonal head model. The head tracking is achieved by finding six translation and rotation parameters to register the rendered images of the textured model with the video images. These parameters are found by mapping the derivative of the error with respect to the parameters to intensity gradients in the image. An error minimum is found by using an estimator to pool the information and perform gradient descent. The limitation in this method is the model's inflexibility and its set of parameters fail to make it closely resemble reality.

In summary, this review of the previous work indicates that most of the systems are still vulnerable to sudden background changes due to illumination change and local/global background movements. The purpose of this study is:

1. To develop a real-time system that can accurately identify human heads from versatile background scenes.
2. To develop techniques that are able to automatically build, update and maintain the background. The method must be reliable and effective in tackling both gradual and sudden background changes, particularly due to lighting variations, local/global background movements and shadows.
3. To develop an effective, robust and accurate head tracking technique, which is also flexible and adaptive to various situations.

4. To test and verify the performance and robustness of the system under various background scenes.

Chapter Three

Basic Scheme - Draping

In this chapter, simple background and object segmentation and draping techniques are introduced. The noise in the image associated with light flickering, shadows, etc. presents difficulties for accurate head tracking through background subtraction. An innovative draping method was, therefore, developed to overcome these difficulties and highlight the foreground outline. By proper adjustment of the drape parameters, the influence of noise can be overcome to a great degree.

3.1 Background and Object Segmentation

Background subtraction subtracts the background from the incoming image and hence extracts the difference between the two images. This method is effective for head tracking if the background is relatively stationary and the foreground object is sufficiently different from the background. The particular feature of this technique is that it interprets the image difference as the foreground object, which makes it very difficult to use in

practice. Noises induced by global and local background movements, illumination change and light flickering, and holes resulting from the subtraction of similar colors in the images will be inevitably misinterpreted as the foreground objects. Background averaging can partially eliminate these influences. The averaged background is obtained from a sequence of pre-captured background images. Figure 3.1a and b show two images from the captured background image sequence. The author deliberately waved a book shown in the middle of Figure 3.1b. The background image (Figure 3.1c) averaged from nine similar images minimizes this movement significantly. The book is no longer visible in the average image (Figure 3.1c).

(a) (b) (c)

Figure 3.1 Background sequence and averaged image

Background and object segmentation can effectively separate the difference in the incoming image from the background. Figure 3.2a is the incoming image and Figure 3.2b is the resultant background and object segmentation map. Figure 3.2b shows that the moving foreground object is separated from the complex background along with some noise resulting from the illumination change in the room. Simple background subtraction

exhibits high sensitivity to background variation and illumination change. A more accurate and robust technique is, therefore, required to correctly identify the foreground object from the object and background segmentation map. Draping is a useful technique to construct a head and shoulder's silhouette on the foreground object (Turk 1998) due to its good acceptability of noise and holes in the image. The mathematical model of draping is described in this context, along with detailed implementation procedures.

(a) (b)

Figure 3.2 Incoming image and segmentation map

3.2 Principle of draping

Noise and holes may result in the object and background segmentation map because of changing background, varying illumination and similarities between images. This gives rise to considerable difficulty in extracting the correct foreground objects from the background by using a simple background subtraction because the noise and holes are inevitably misinterpreted as foreground objects. The literature review in chapter three indicates that the draping method can tolerate a reasonable amount of noise and holes presented in the segmented map (Turk 1998). This innovative technique is, therefore,

modified and improved in this context as the major technique to extract the foreground objects from the background. The development of this method is inspired by the mechanism of a physical drape. It functions by forming a flexible one-dimensional sheet (drape) with simulated point masses connected by springs as shown in Figure 3.3.

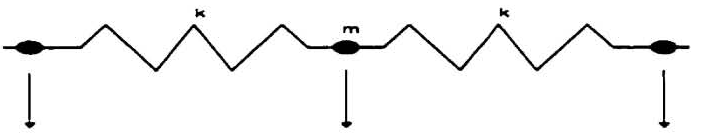


Figure 3.3 The initial drape

Every point mass is assumed to be subjected to a gravity force (W) induced by each point mass (m) and elastic forces (f_k) generated by springs provided that there is one pixel displacement and supporting force (f_S) provided by each foreground pixel. The mathematical representations of the spring force is as follows:

$$F_K = f_K \cdot N \tag{3-1}$$

where f_K is the induced spring force by one pixel displacement and N is the number of displacements measured in pixels due to the unaligned two neighboring point masses. When the drape is lowered from the top of the image due to gravity, the foreground object pixels collectively support the drape. In this case, these pixels underneath act like a solid supporting column. The pixels below the first contacting pixel collectively

contribute supporting force to that pixel. Only the continuously connected pixel chain with the first pixel is, however, included as the solid mechanical column. Meanwhile, the springs between the neighboring point masses are stretched and hence induce the spring force, which is linearly proportional to the spring displacement. If the supporting force of the object pixel exceeds the total effect of the downward gravitational and spring forces, the drape will rest on that pixel (Figure 3.4). The corresponding pixel trapped by the drape is regarded as the foreground object. Otherwise, the pixel fails to hold the drape at that location and the corresponding pixel is regarded as the noise. The drape will pass that pixel and drop along that column to the next closest foreground pixel. The process continues until the entire drape is eventually held stationary by the group effort of the foreground object. In this manner, the drape provides a clear outline of the moving foreground objects.

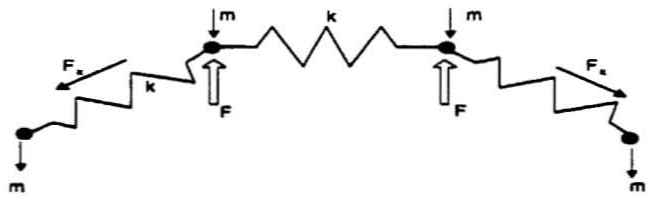


Figure 3.4 The deformed drape

3.3 Algorithm for draping

The original draping method proposed by Turk (1998) applies the draping on a thresholded background and foreground segmentation image. In this study, the draping is

extended and improved for use directly on background and segmentation images and on the variance of the incoming images. In order for the draping method to function as expected, the drape must be able to discriminate between noise and the desired foreground object. This is achieved by carefully selecting the values of point mass, spring constant and supporting force contributed by each pixel. A trial-and-error process is employed in this study to determine the optimal combination of these values. Once the proper parameters are determined, the mathematical model can be implemented through the following steps:

1. Suppose that there exist point masses (m) at the top of each column of pixels in the difference image, each connected to its neighbors by a spring with a spring constant (k) as shown in Figure 3.4.
2. For each point mass, calculate the vertical force exerted by its neighbors (it will be none if they are aligned) and the collective "force" exerted by the segmented binary foreground object (none if there is no foreground at that pixel, a constant value if there is foreground there).
3. If this vertical force is upwards above a small threshold (1.0) than the point mass, then the point mass does not move down to the next row in the next iteration (jump to step 5). Otherwise, the point mass moves down to the next row for the next iteration.
4. Once this calculation is performed for the entire row of point masses, update their positions (most of them will move down one row, as decided in step 3).
5. Go back to step 2 and repeat for the updated positions of the point masses.

6. Stop iterating when the point masses have all (or mostly) stopped moving, or when they hit the bottom of the image.

The point mass, the spring constant, the upward force of the foreground and the threshold value are determined experimentally to achieve a best draping result. Figure 3.5 shows a typical example obtained from the above operation. Figure 3.5a is the resultant draping resting on a person and Figure 3.5b shows the person outlined by the corresponding drape.

(a) (b)

Figure 3.5 Draping and the head and shoulder silhouette

3.4 Implementation of Draping

The implementation of draping can be best illustrated by referring to a typical example in which a user is sitting in front of the camera with some background and illumination changes. Figure 3.6 presents different stages of the consequent draping operation. The drape initially drops from the top of the image segmentation map due to the gravity (Figure 3.6a). At first, it forms a straight horizontal line as shown because it

has not made contact with the foreground object at this position. The noise presented in the image is only composed of a small cluster of floating pixels with little or no solid support from underneath. The drape eventually overcomes that noise because the noise is unable to develop sufficient supporting force to sustain the total effect of gravitational and spring forces. When the point masses of the drape hit the solid foreground object, the drape rests on the object because the collective supporting force developed at this stage exceeds the total effect of the gravity and elastic forces by the given threshold (Figure 3.6b). After several iterations, the drape eventually rests on the foreground object, providing a clear outline of the user (Figure 3.6c). This technique exhibits great tolerance of noise and holes presented in the image segmentation map because they cannot generate sufficient collective forces to support the drape.

(a) (b) (c)

Figure 3.6 Mechanism of draping

Experiments were conducted in order to determine the parameter values for achieving a good drape. Dimensional analysis was employed to analyze the problem. This analysis is based on the principle of Fourier's dimensional homogeneity theory. The

analysis results in a dimensionless functional equation. This use of this equation can tremendously simplify the problem and provide guidelines for the design of experiments and the presentation of results. The Buckingham method (Buckingham 1914) is used to analyze the problem. The initial step is to establish the original functional equation. The supporting force provided by unit pixel (f_S) can be related with the weight of unit point mass (W) and unit spring force (f_k) as follows:

$$f_S = g(W, f_K) \tag{3-2}$$

Equation (3-2) has the dimension of [*Newton/pixel*]. The dimensionless functional equation can be obtained by combining the terms in equation (3-2) to get the final dimensionless functional equation as follows:

$$\frac{f_S}{f_K} = \phi(\frac{W}{f_K}) \tag{3-3}$$

Dimensional analysis indicates that $\frac{W}{f_K}$ and $\frac{f_S}{f_K}$ are the two independent parameters. To investigate the parameter values for good draping, draping is attempted on a series of combinations of the two parameters and the results are plotted in Figure 3.7. The analysis conclude that as long as the parameters fall within this narrow band shown as a blue circle in Figure 3.7, a good drape will result.

The draping method is computationally simple and hence speedup can be achieved in tracking the head. This makes this system suitable for using in real time. For a 160×120 pixels image, it takes 0.05 seconds to process the draping. Comparing the frame rate of 3 frames per second, this processing speed allows the head to be tracked promptly in real time.

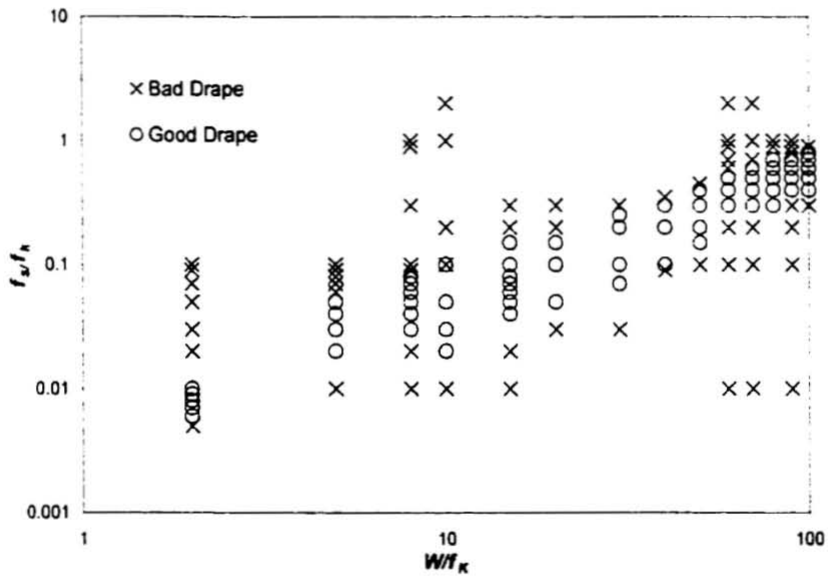


Figure 3.7 Selection of draping parameters

Chapter Four

From Drape to Description

The literature review reveals that the most commonly used techniques for head locating are color-based (Yang and Waibel 1996; Qian et al. 1998) and template-based face trackings (Pentland and Horowitz 1991; Wren et al. 1996; Haritaoglu et al. 1998a, 1998b, 1998c; Horprasert et al. 1998). The color-based technique uses color filtering to identify face skin, which is susceptible to variation in lighting conditions, skin color, background image, etc. The template-based method involves the creation of human models, which may not always match the postures and gestures of the people in practice. The significant advantages of the draping method over the above methods are its computation simplicity and good noise acceptability. The incorporation of this method is just the first step to a robust and effective real-time head tracking system.

Once the person's outline (drape) is settled, it can be used to locate the person's head. At the initial stage of the study, a simple head tracking method based on background subtraction and head ratio was developed. However, this method has difficulty in extracting heads from an incompletely recovered foreground. An innovative

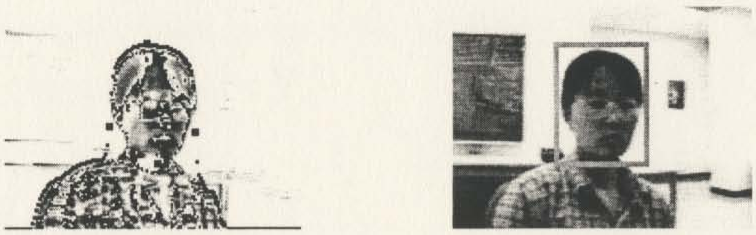
and simple head tracking technique consisting of a drape interpolation followed by a head merging was developed eventually to achieve an accurate head locating. Experiments were conducted on a variety of typical images. The results indicate that the new technique is effective, efficient and robust. In this chapter, both techniques are described along with several typical results.

4.1 Initial method for locating a person's head

At the initial stage of the study, a simple method for locating the head was developed based on the head geometry and the first- and second-order derivatives of the drape. The head tracking was achieved by generating a rectangular box circumscribing the tracked head. The procedure of this method is as follows:

1. The highest point in the drape is first identified to determine the top border.
2. Starting from this point, both the first- and second-order derivatives of each point along the two sides are calculated. The first point on each side whose second-order derivative changes the sign and first-order derivative is greater than a given threshold (4) is used to determine the left and right borders.
3. The head bottom border is estimated based on the averaged human head ratio. A rectangular box can, therefore, be generated based on this ratio.

A typical result arrived at from this method is shown in Figure 4.1. Figure 4.1a is the image segmentation map showing the resulting drape and the identified four border points (indicated as solid rectangles). Figure 4.2b shows the corresponding rectangular box determined by these four points in the original incoming image.

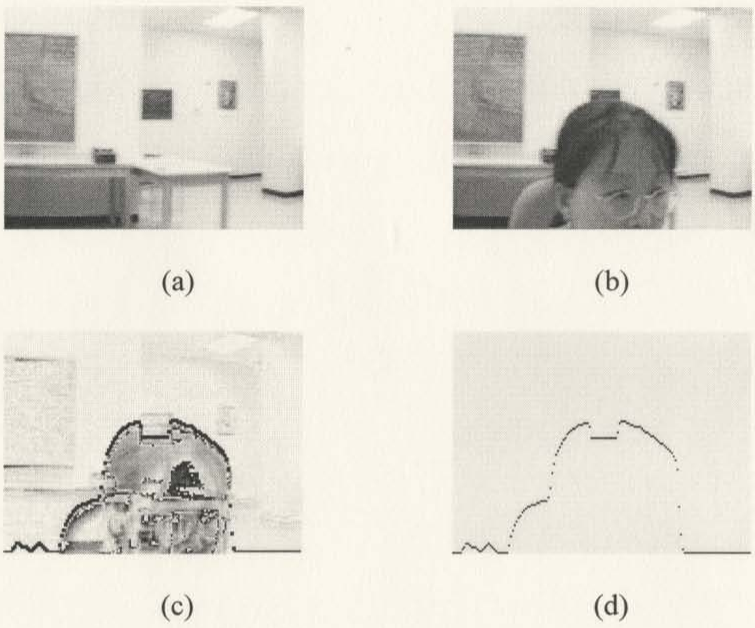


(a) (b)

Figure 4.1 Captured person's head from the drape using the initial method

This process is efficient in computation and simple in concept. These advantages, on the other hand, also limit its efficiency and accuracy in broader applications. For a particular circumstance (Figure 4.2), this method fails to track the head from the image segmentation map. Part of the drape penetrates the head, resulting in a notch on the top of the head. This can be explained by analyzing the mechanism of this method in detail. The background used by this method is the average of a sequence of pre-captured frames (Figure 4.2a). The foreground detection is carried out by simply subtracting the resultant background from the incoming image. Apparently, this operation will inevitably erase those parts of the foreground having similar colors to the corresponding background. As in Figure 4.2b, the painting on the wall has a similar color to the person's hair. Therefore, the part of the head overlapping with the painting has been erased by the subtraction as shown in Figure 4.2c. Figure 4.2d shows the resultant drape with a notch on the top of the head. A rectangular box cannot be located in this case because the drape profile does not match the head profile assumed in the scheme. In addition, this technique has difficulty in

identifying multiple heads from the drape. Further improvement of this system is, therefore, carried out to tackle these types of problems by using a drape interpolation and compounding method.



(a) (b)

(c) (d)

Figure 4.2 Image subtraction showing incompletely recovered head

4.2 Drape interpolation and compounding method

The initial derivative-based head tracking method and its limitations have been discussed in section 4.1. Techniques for identifying an incomplete head and multiple heads from a complex drape with more complicated background must be developed in order to improve its robustness and portability.

The undisturbed one-dimensional drape is a horizontal straight line with uniform intervals between the neighboring point masses. When the drape is settled on the foreground objects, these intervals are stretched unevenly according to the particular profile of the human head (Figure 4.3a). As can be observed, the drape is stretched significantly along the two sides of the head, along the two shoulders, and inside the notches of any incompletely recovered foreground objects. The resulting discontinuity in the drape makes it difficult to conduct an accurate image process. The drape line must, therefore, be connected and smoothed. This is achieved by using linear interpolation. The resulting drape makes the stretched parts smoothly connected (Figure 4.3b).

(a) (b)

Figure 4.3 Original drape and its corresponding smoothed drape

After the drape has been smoothed by linear interpolation, the following head searching scheme has been developed to track the person's head:

1. Scan the entire image from top to bottom (Y direction). At each row, the direction of the scanning is from left to right (X direction). This operation starts from the top-left corner of the image.
2. When any point, which does not belong to any existing head objects, from the smoothed drape is found, create a head object composed of the point itself and all its neighboring continuously connected points if any. For this head object, define the leftmost point as *HeadLeft* and the rightmost point as *HeadRight* for the current row. The newly created head object is regarded as a potential head. The criterion for classifying if the current points belong to any existing head is as follows:

   If the leftmost point of the considered points is zero or one pixel to the right of the *HeadRight* of the head created in the previous row, they are regarded as part of the head in the previous row. The current rightmost point is defined as the *HeadRight* of the current row. If the rightmost point of the considered points is zero or one pixel to the left of the *HeadLeft* of the head created in the previous row, they are regarded as the part of the head in the previous row. The current leftmost point is defined as the *HeadLeft* of the current row. Otherwise, create a new head object in the current row. The following pseudo code can be used to implement this process:

```
total = 0;
flag = 0;
for(headNum = 0; headNum < total; headNum++){
{
    if(heads[headNum].start[y-1] == end_point
    || heads[headNum].start[y-1] == end_point+1)
    {
        heads[headNum].start[y] = start_point;
        flag = 1;
```

```
        }

        if( heads[headNum].end[y-1] == start_point
        ||  heads[headNum].end[y-1] == start_point-1)
        {
            heads[headNum].end[y] = end_point;
            flag = 1;
        }
    }
    if(flag == 0){
        current_head = heads.NewHead();
        current_head->start[y] = start_point;
        current_head->end[y]   = end_point;
        total++;
    }
```

3. Repeat step 2 until the entire image has been scanned.

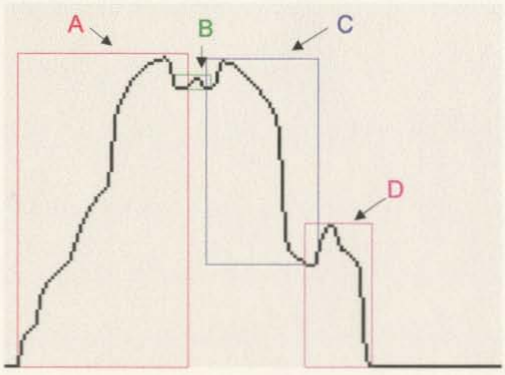


Figure 4.4 Head identification

This head searching technique can capture almost any wrinkles, notches and sudden variations that occurred in the drape. A number of potential heads can often be created as a list named *HeadList* through this exhausting head searching scheme. For the

image shown in Figure 4.3, four potential heads can be identified. They are labeled as A (red), B (green), C (blue) and D (pink) in Figure 4.4 to indicate intermediate head regions surrounded by rectangular boxes.

Most of the identified potential heads are, of course, not true heads, but may be only part of a true head. They need to be merged to form the true head. An innovative technique was developed to carry out the head merging. The merging process consists of two stages, with each stage tackling different situations.

Stage I: The first stage is designed to tackle potential heads with only a few points. It can also be regarded as the preparation operation for the second stage. We define a threshold (50 pixels) for head size. Any identified head that is less than this threshold is assumed as a false head. Therefore, if the dimension of a potential head is less than this threshold, it will be merged into its neighboring head. This in-process decision rules out the tiny false heads due to the wrinkles in the drape and heads that are too far away from the camera. The head size threshold can be adjusted according to image size. The newly merged head needs to update its ***HeadLeft*** and ***HeadRight*** for each row with the leftmost and rightmost points of the merged pixels at that row. Once the operations at this stage have been completed, the small potential heads such as wrinkles and notches have been merged to form large potential heads. By using the first stage merging, head B in Figure 4.4 is merged with its left neighbor, head A, to generate head AB shown in Figure 4.5 because its size is not sufficiently large (less than the threshold) to be considered as a head.

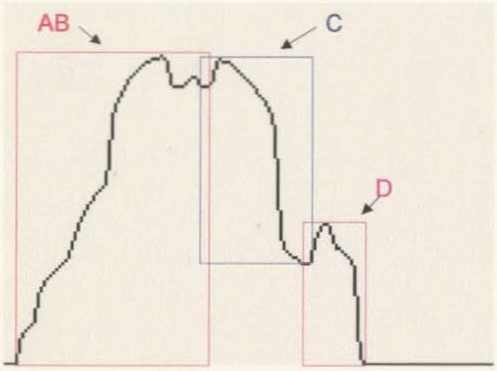


Figure 4.5 Stage I merging showing head B to be merged with head A

Stage II: The large heads resulting from stage I are considered at this step. Judgements on whether they need to be further merged are first carried out. The judgement is based on both the slope and the dimension of the head. The head slope is defined as the slope of the straight line connecting the starting and ending pixels of the potential head. Once the slopes are calculated for all potential heads, a merging operation will be followed. This process is carried out over the entire drape from left to right. A head with a negative slope will merge with its neighboring head with a positive slope if the dimension of either of the heads is less than a threshold (600 pixels). This merging generates a new potential head object in *Headlist*. This process will continue until no further merging is possible. Through stage II merging, head AB in Figure 4.5 merges with head C as a new potential head ABC. Head ABC then further merges with D to generate head ABCD. The two new heads are created as new head objects in *HeadList.* The resultant heads ABC and ABCD are shown in Figure 4.6, together with the previous heads.

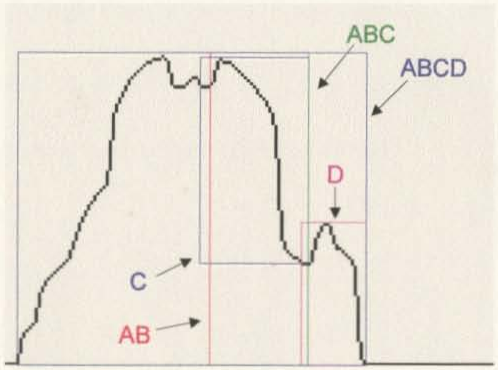


Figure 4.6 Stage II merging showing the newly merged heads

The above head searching and merging operations can often generate a number of potential heads. A method is needed to identify and extract the true heads from these potential heads. This can be achieved by the following head identification method.

1. Calculate the width of each row for each potential head by subtracting its *HeadLeft* from its *HeadRight* if both exist. The *HeadList* is updated with the resultant *HeadWidth*. Figure 4.7 shows the updated potential heads from Figure 4.6. Comparing with Figure 4.6, the areas occupied by heads AB, C, D and ABC are reduced significantly because the valid *HeadWidth* can only be found in these parts of the image.
2. The resulting width against row number function is then smoothed by five-point mean filtering and is plotted with the abscissa as the row number and the ordinate as the row width. Typical plots of this function are shown in upper part of Figure 4.8.

3. Compute the second-order derivatives of the resultant curves. The bottom part of Figure 4.8 shows the resultant second-order derivatives. Any two neighboring negative and positive peaks are considered as a pair.

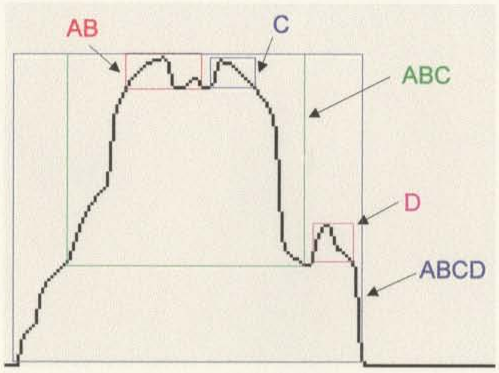
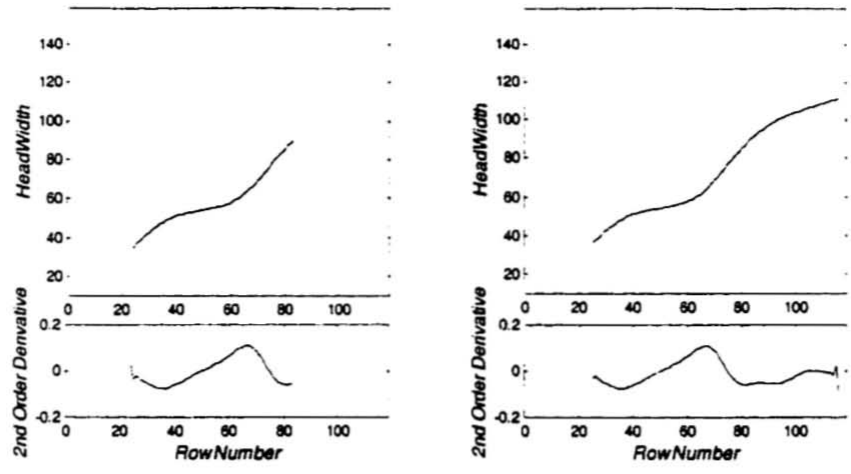


Figure 4.7 The updated potential heads

4. Marching from the left end of the *HeadWidth* curves, when a second-order derivative pair is found, define the value of width, where the positive peak is achieved, as *Width_of_Head*, and the distance from the top of the current head to where the positive peak is achieved as *Height_of_Head*.
5. Calculate the ratio of *Height_of_Head* to *Width_of_Head*. If the ratio is within a given threshold (1.5), and both the *Width_of_Head* and the *Height_of_Head* are greater than one given threshold (12 pixels), stop this search and generate a rectangular box based on the current *Height_of_Head* and *Width_of_Head*. Otherwise, search the next pair and repeat step 4 until the right end of the curve has been reached.



(a) (b)

Figure 4.8 *HeadWidth* functions and their second-order derivatives

6. Several such rectangular boxes may be located and some of them may overlap one another. An averaged rectangular box obtained from these overlapped boxes is eventually regarded as the head's location. Figure 4.9 shows the final head location box processed from Figure 4.8 by using this method.

The unique feature of the above technique is to leave the major decision for the true heads to the last step of the entire process. This can avoid losing any useful information. The head size by which we make final decision can be adjusted according to the size of the image used.

The method shows a significant improvement over the initial head tracking on both accuracy and robustness. As discussed previously in section 4.1, the initial method

failed to track the head from Figure 4.2a due to the notch which appeared on the top of the head. Using the improved method, the head has been successfully captured, as shown in Figure 4.10b. Figure 4.10a shows the interpolated drape. Figure 4.10c and d show the *HeadWidth*-row number curves, along with the corresponding second-order derivatives. Five potential heads can be found, each having its corresponding *HeadWidth*-row number curves. Only the curves for the two significant heads are shown in these figures.

Figure 4.9 Captured head through the improved method

This technique can also effectively track multiple heads from a complex background. For the case of two people in the scene, the interpolated drape is generated (Figure 4.11a). Initially, several potential heads are identified from the interpolated drape. The *HeadWidth*-row curves were produced for all the large heads (Figure 4.11c, d, e and f). Only two heads are identified (Figure 4.11b), with the corresponding rectangular boxes accurately capturing the two heads.

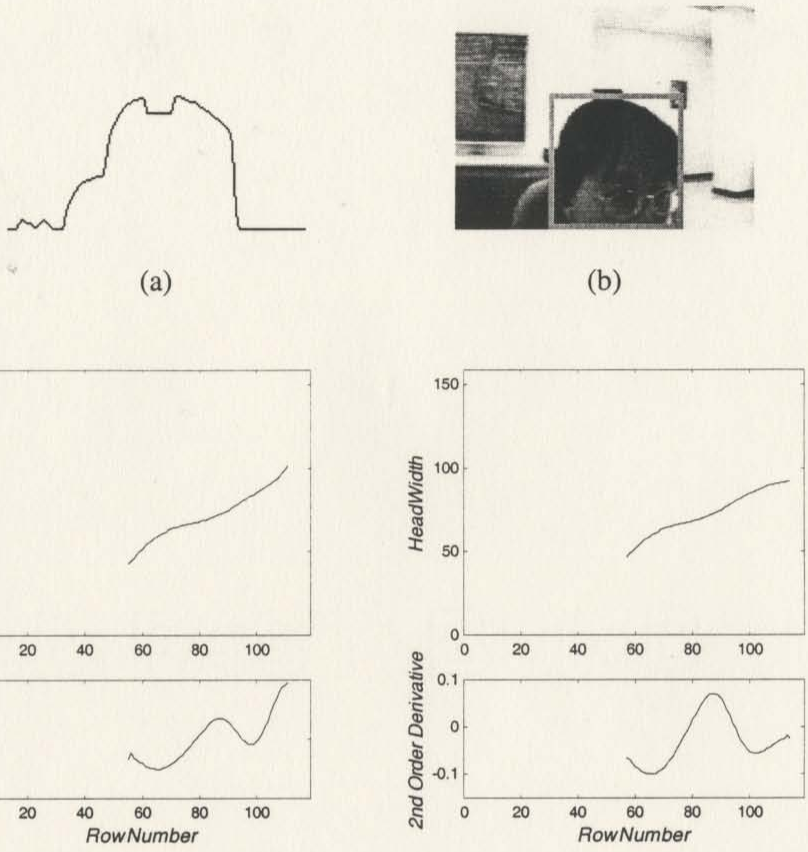


Figure 4.10 Tracking an incomplete head using the improved method

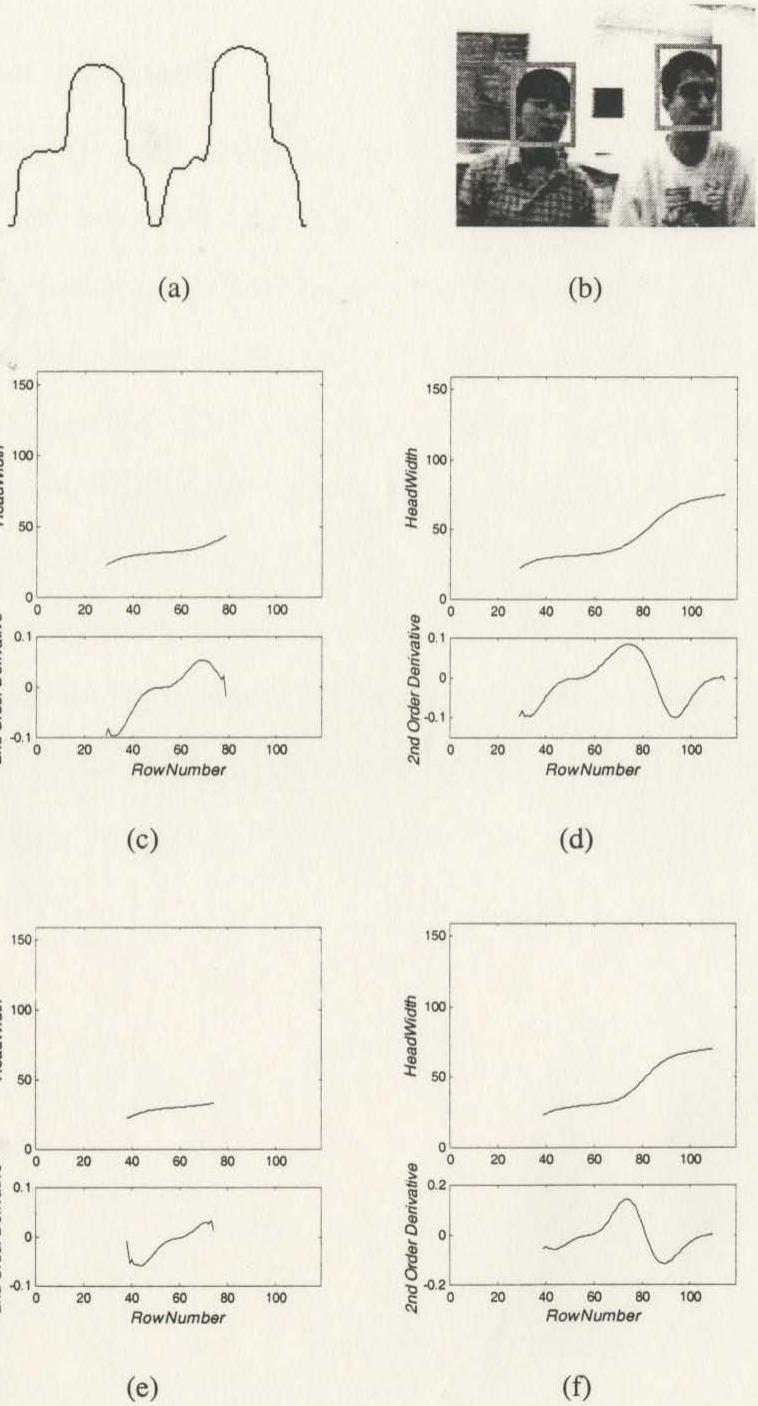
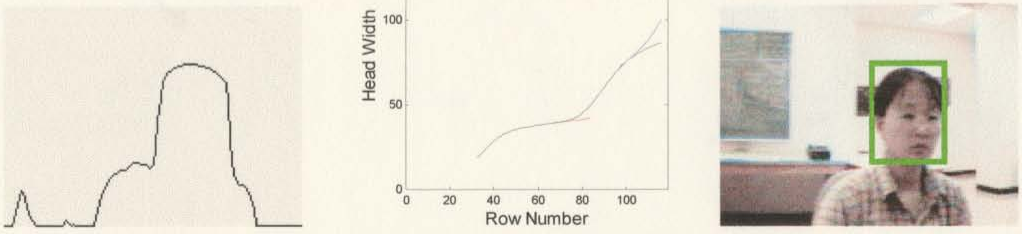


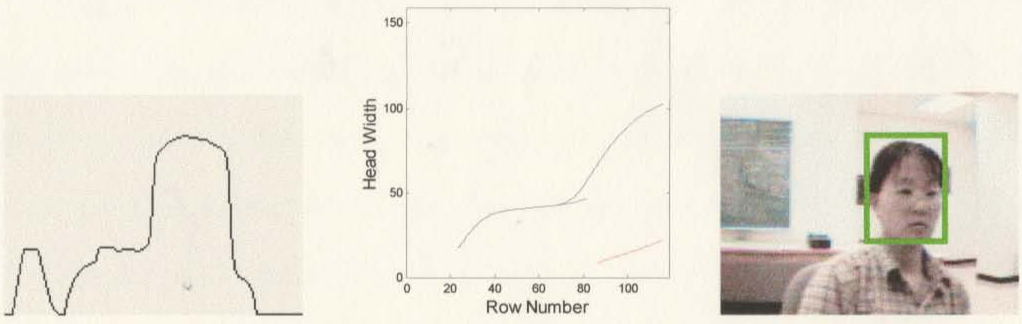
Figure 4.11 Tracking two heads using the improved method

4.3 Experiments and Results

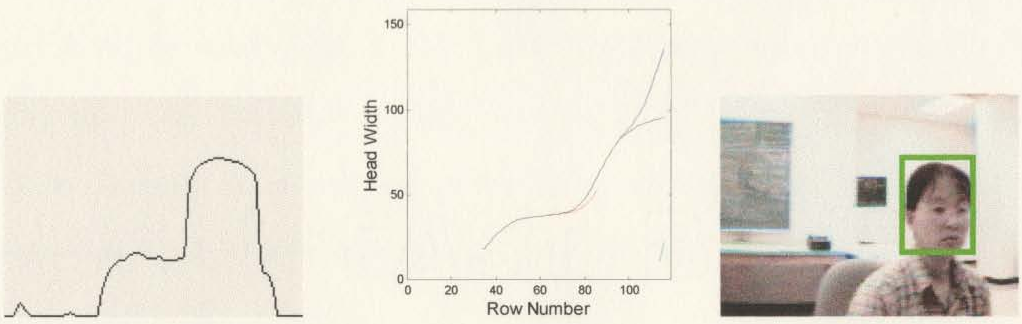
Experiments have been conducted on a variety of cases to test the effectiveness and accuracy of the improved method. Figure 4.12 illustrates the application of this method on various cases of single head tracking. The resultant drape, head width and the tracked head in the incoming image are shown for each case. The small peak in the left of image 4.12a was successfully identified as a false head. The local light flickering gives rise to a large plateau in the drape shown in Figure 4.12b. Although its magnitude is significant, the technique can identify it as a false head because its particular shape and parameters do not match those of a true head. The chair in Figure 4.12c was captured as the foreground object because the user deliberately moved it. The part of drape caused by the chair was successfully ruled out. This method can also successfully identify the true head even though there are substantial small peaks, notches and wrinkles presented in the drape (Figure 4.12d).



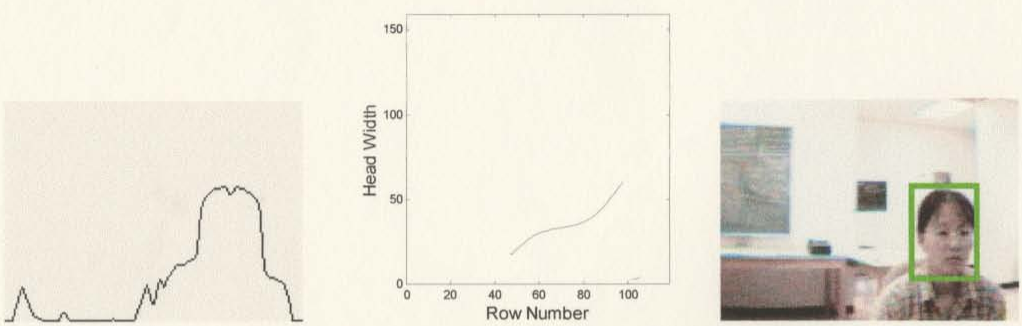
(a)



(b)



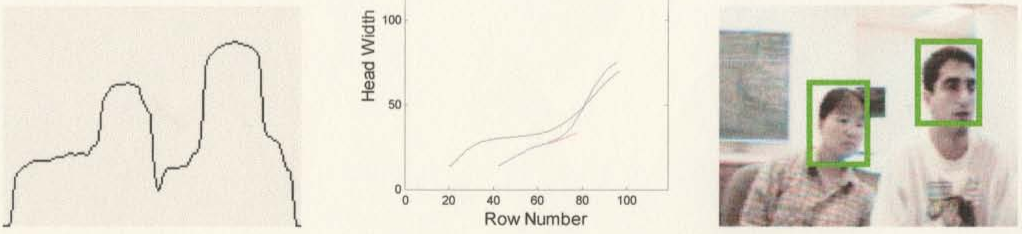
(c)



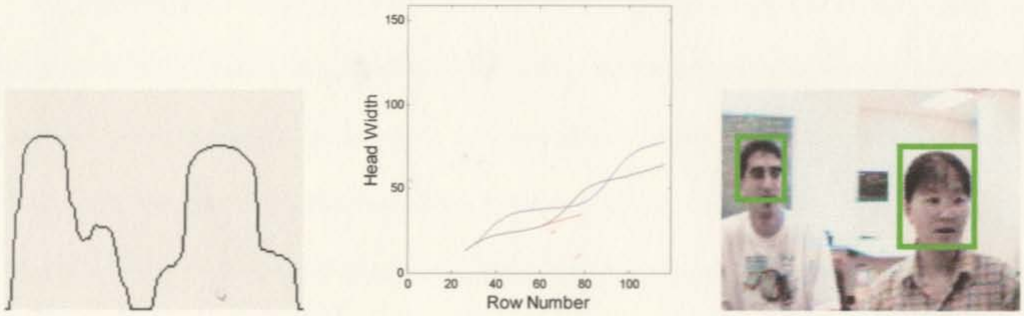
(d)

Figure 4.12 Experiments of tracking a single head using the improved method

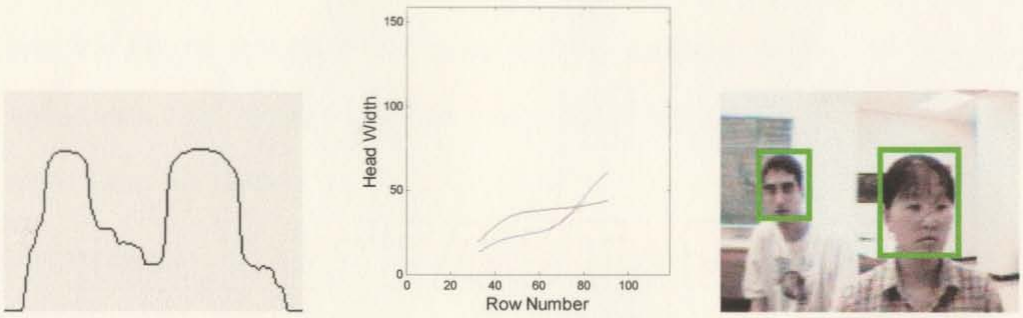
Figure 4.13 presents the application of this method for tracking two heads. Also, the resultant drape, head width and tracked heads are shown for each case. Figure 4.13a shows two users, one closer to the camera than the other with head aslant. The front user blocks part of the shoulder of the other. The chair was also moved. The head of the user behind overlaps with a black painting on the wall. The technique successfully identifies two true heads from the resultant irregular drape. The two users switched their position and part of the shoulder of the behind user is not captured in Figure 4.13b as expected. In figure 4.13c, the heads of the two users differ significantly because one user moved farther away from the camera. In addition, the front user blocked the behind user's shoulder. In figure 4.13d, one user moved out of the screen with the behind user's shoulder left. In all the above circumstances, the improved method has identified the true heads accurately and precisely.



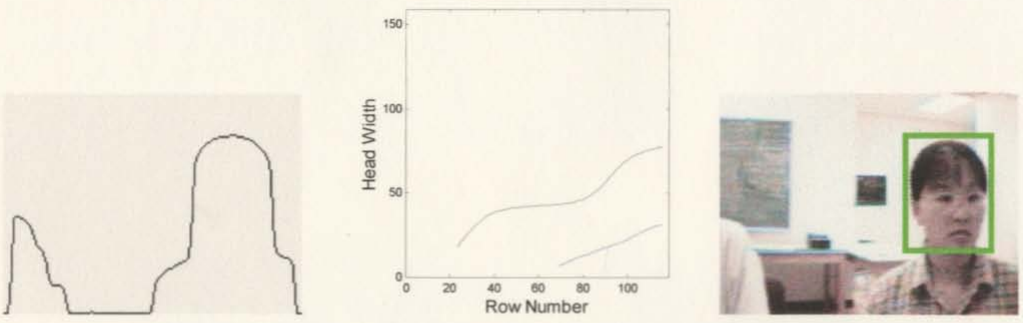
(a)



(b)



(c)



(d)

Figure 4.13 Experiments of tracking two heads using the improved method

However, in very few occasions (one out of a few hundred images), this system may lose its accuracy due to the uncertainty associated with the identification of the second-order derivative peak pairs. This misjudgment of the pairs may result in a head box either oversized or undersized from the true head. For the resultant drape shown in Figure 4.14a, the *HeadWidth*-row number curve has some small local fluctuations (peaks), which gives rise to uncertainty in determining the correct peak pairs. The misjudged peak pairs result in two oversized head boxes (Figure 4.14c). However, this problem can be overcome by increasing the iterations at the sacrifice of more computation time. Figure 4.14d shows the result at 40 iterations. In this case, the heads can be located accurately.

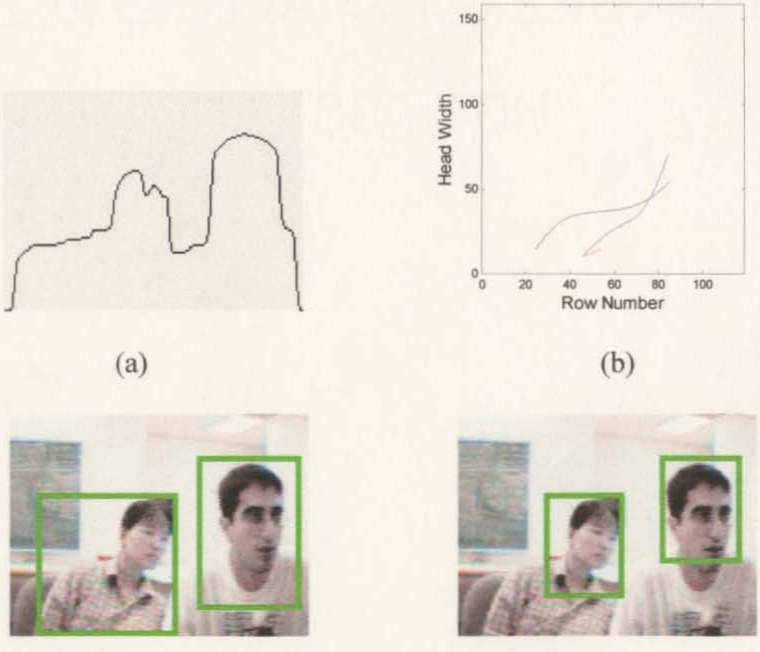
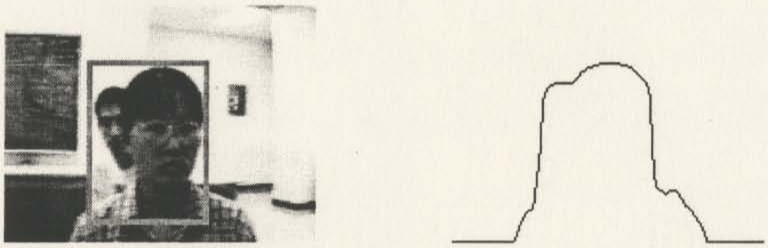


Figure 4.14 Bad case 1

This system can also lose its accuracy when the users in the scene are overlapped as shown in Figure 4.15a. In this case, only one true head is interpreted from the resulting drape (see Figure 4.15b). This is the common limitation of the background subtraction method itself. Two heads may be extracted if we incorporate face feature method into this system. However, this method is computational expensive and will slow down the performance of this system. This type of situation is, therefore, not tackled by this system.



(a) (b)

Figure 4.15 Bad case 2

Chapter Five

Dealing with Changes in Background

After the success of the first stage improvement, the focus of the study was then steered to tackling the problems associated with a change of background. The background image can change suddenly for a variety of reasons. Global movement of the background may be due to accidental camera movement, camera zoom in and out, etc. It is therefore inappropriate to use a pre-selected background to recover the foreground object through background subtraction. The background to be subtracted must be re-initialized automatically. To address the problem of background change that happens infrequently and instantaneously, we can use foreground movement to tentatively identify background regions. That is, rather than trying to estimate the amount of background shift (which may be impossible) we begin by assuming that the entire new background is unknown, and wait for foreground movement. Note that this method works only for rare, instantaneous background change, not for continuous camera movement, but such infrequent changes are exactly those that pose a problem for a background-subtraction tracking system. A multiple-state system of progressively generating background from

the inter-frame variance of incoming images has been developed. The system functions once changes in the background scene are detected. The draping is used on the resultant inter-frame variance map to extract the movement of a person's head and the part of the image outside the settled drape is assimilated to a background map. The background rebuilding process continues until the unassigned "Don't know" area is below a threshold (300 pixels). This system also incorporates median filtering, multiple draping and polynomial interpolation in an effort to tackle the problems associated with a bad drape.

In this chapter, the problem of assimilating a new background is addressed, together with the descriptions of the multiple-state system and the mechanisms of median filtering, multiple draping and polynomial interpolation. These three methods aim to improve the inter-frame variance map such that the draping can outline the desired foreground objects. The draping used in this chapter differs from that used in chapter four with the values of parameters. Therefore, these three methods can achieve good results when they are used in the inter-frame variance map but not in the foreground and background segmentation image.

5.1 Multiple State System

The problem of changing background handling was addressed by using a multiple-state background formation system. In the first state, the background is known, and face tracking happens by the simple background subtraction discussed in chapters three and four. Meanwhile, the incoming background is compared with the stored version, and if they differ significantly, the system assumes that the background has

changed (perhaps because the camera has moved, or the lighting has changed). Such a change forces the system into the next state, where it attempts to build a new background. The steps in this background-building process are as follows:

1. The system captures several incoming images – usually four or five frames. These frames possess sufficient information if the foreground object (assumed to be a person) is in motion. This requirement for "sufficient information" is verified by measuring the global variance between frames. If this is high enough. then local variances are calculated at each pixel and thresholded to produce a "variance map". The purpose of thresholding in this case is to rule out noise due to lighting changes to extract the person's movement. If the global variance is not high enough, then a further set of frames are captured and the process is repeated.
2. Draping is applied to the inter-frame variance map. Because the person moves, there will be a high-variance border surrounding the person. This can be found by draping, just as previously described in chapter three, so the difference between the foreground and the background is found. The parameters of draping are adjusted so that the drape can be held in place by the variance. The draping method is now extended and improved to apply on the variance of the incoming images to identify moving areas and thus separate foreground objects from the background.
3. If a good drape results, every point outside the drape can be assumed to belong to the background. These points are loaded into the background image setup area. The points below the drape are assigned "Don't know" values in the background image.

4. If the total area of "Don't know" points in the background image is smaller than the given threshold, the system switches back to its simple background-subtraction operating mode. If not, then background building continues: variance is re-calculated, a new variance map is generated, the drape is applied, and points outside are added to the background image. In this manner, as the foreground object moves, "Don't know" points in the background image are progressively replaced with new pixels of uncovered background.

Figure 5.1 illustrates the different stages of the background formation procedure described above. The inter-frame variance map is generated from several images and the variance is calculated as shown in Figure 5.1a. If the global variance is greater than a given threshold (30), draping is applied on this variance map (Figure 5.1b). The person's head can be located in the box through this process as shown in Figure 5.1c. In the meantime, this drape is used to identify the background areas. The part outside the drape is loaded as the background setup (Figure 5.1d).

After several iterations, the white ("Don't know") area becomes smaller and smaller until it is less than the given threshold. Potentially the white area can be completely eliminated if the iteration is sufficiently long. However, the threshold is taken as the stopping condition in order to be able to carry out this operation fast enough to be used in real time. Experiments demonstrate that the influence of this residual area on the final result can be negligible. At this stage, the background setup is regarded as finalized (Figure 5.1e). The person's head can then be extracted effectively by subtracting the resultant background map from the incoming image (Figure 5.1f).

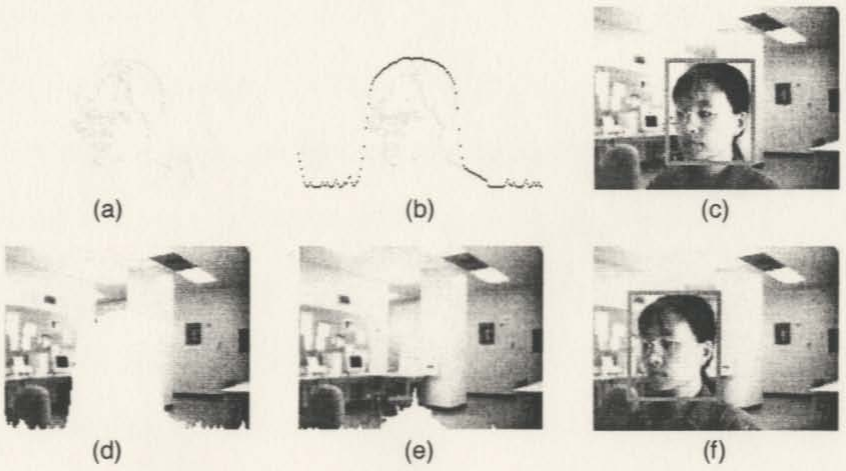


Figure 5.1 Multiple state face tracking system

5.2 System Improvement

The background-building scheme described in section 5.1 works satisfactorily if the borders of high variance around moving objects are well defined. Often, however, because of small movement, and similar shading between foreground and background, the borders are broken. This makes it difficult to recover the complete foreground object by simply applying the technique described in section 5.1. Part of the foreground object is inevitably misinterpreted as the background and is introduced to the background map. Several techniques have been investigated and attempted for correcting this problem. Among them, median filtering, multiple direction draping and polynomial interpolation yield satisfactory improvements. The difference among the three methods lies in their strategies for approximating the foreground outline from the drape.

5.2.1 Median filtering

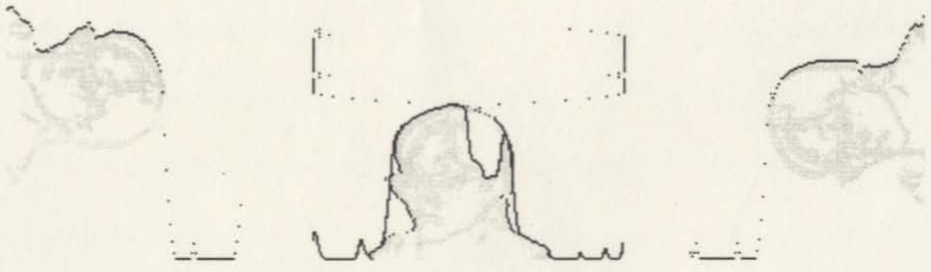
The median filter sorts the values of all points that are neighbors of the highest point of each column in the inter-frame variance map. The median value of the list is selected as the new value of this point. Seven points are used as the neighbors of this point. This strategy can remove peaks of both high and low values without flattening value steps, which separate variance value regions. The variance matrix is updated with these new values. This process often needs to be iterated several times. Draping is finally applied on the newly generated inter-frame variance map. This method can improve the system to a certain degree although it cannot completely solve the problem. Figure 5.1a shows the inter-frame variance map. Figure 5.1b is the original drape settling on the inter-frame variance map. The drape penetrates part of the head. A notch appears on the left of the head because part of the head has a similar color with the background and this means the head movement is not captured by the inter-frame variance map. By using the median filtering method, a new value of the highest point of each column is added to the inter-frame variance map. The result is improved as shown in Figure 5.2c.

(a) (b) (c)

Figure 5.2 Result of median filtering

5.2.2 Multiple direction draping

Multiple direction draping allows draping from top, left and right three directions of the inter-frame variance map respectively to yield a normal upper drape, a left drape and a right drape. The background map is generated from these three drapes. The part of the incoming image that is outside the outmost area of the three drapes is loaded as the background image. The residual area in the background is assigned as "Don't know" value. Figure 5.3 illustrates the mechanism of the multiple direction draping method. The drape is shown as the dashed line in the images. Figure 5.3a and Figure 5.3c show the draping operation from left and right respectively. Figure 5.3b is the combined draping from all three directions in the inter-frame variance map. Notice that the person's head cannot be fully recovered due to the notches on the top of the head if we only apply normal (up-down direction) draping. Consequently an accurate background map cannot be constructed because this part of the head is introduced to the background. The multiple direction draping overcomes this satisfactorily as shown in Figure 5.3b.



(a) (b) (c)

Figure 5.3 (a) Left draping (b) Combined draping (c) Right draping

5.2.3 Polynomial interpolation

This method starts by finding the highest point in each column of the inter-frame variance map. If the difference between any point with its neighbor in the highest point array is greater than a given threshold (10 pixels), it is regarded as an isolated point and is discarded. The remaining points form a new array. Applying polynomial interpolation on this new array results in a smooth curve. The inflection points are located and connected by straight lines. These straight lines are put back into the inter-frame variance map. Draping is applied on the new inter-frame variance map. We can modify the background image from the resulting drape. The portions of the incoming image whose positions are above the drape are regarded as the background image. The portions of the incoming image whose positions are under the drape are assigned as "Don't know" regions. The background image modification continues until the "Don't know" region is smaller than the given threshold. The object segmentation map can, therefore, be generated from the background image and the current incoming image. The head can then be located by applying another draping on this segmentation map.

Figure 5.4 shows the mechanism of the polynomial interpolation technique. Figure 5.4a is the original drape on the inter-frame variance map. The drape cannot cover the entire head. The dashed drape occupies part of the head area as shown in Figures 5.4a. Figures 5.4b and 5.4c show the results after the polynomial interpolation. In Figure 5.4b, we locate the high inflection points of the curve constructed from the highest points of each column in the inter-frame variance map. Interpolating the inflection points bridges the gaps of the head. The draping is then applied on Figure 5.4b to rule out the

unexpected interpolation points. Figure 5.4c is the improved drape. Meanwhile the background is generated accordingly.

The problem of background change is very difficult to solve because the situations resulting from the change are diverse and complex. The approaches developed so far in this context yield satisfactory results, enabling this human face tracking system to tackle problems with a substantial amount of illumination change and global movement.

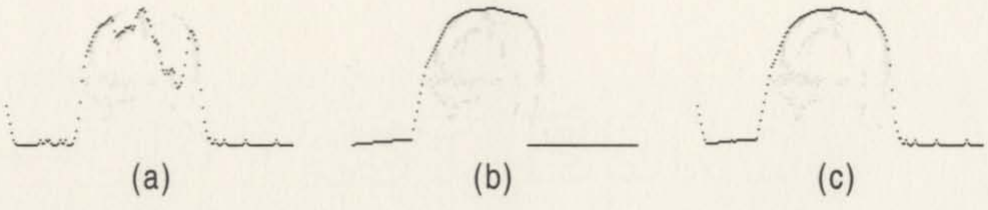


Figure 5.4 (a) Original draping (b) Polynomial interpolation (c) Improved draping

5.3 Interface Description

This real-time face tracking system was implemented using C++ and MCLGallery (Cheng and Robinson 1998) in a Pentium II 200MHz personal computer. A graphic user interface was developed to facilitate the visualization and control of the entire face tracking process. This interface consists of six image windows and five control buttons. The six image windows are *MCLBitmap* (displaying the captured head), *MCLVideo* (displaying the incoming image from camera), *Background formation image, Left drape,*

Combined drape and *Right drape*. The five control buttons are *Capture Background, Start People Detection, Load Background, Load File* and *Show Average,* which are all self-descriptive. Figure 5.5 shows this interface processing an example image. The background formation in this case is achieved by using multiple draping method. By pressing *Start People Detection button*, we can switch among polynomial interpolation, median filtering and multiple draping background formation methods. The dynamic states of the incoming image, modeling background, original draping, improved draping, foreground/background segmentation map and the tracked head are displayed in the corresponding windows.

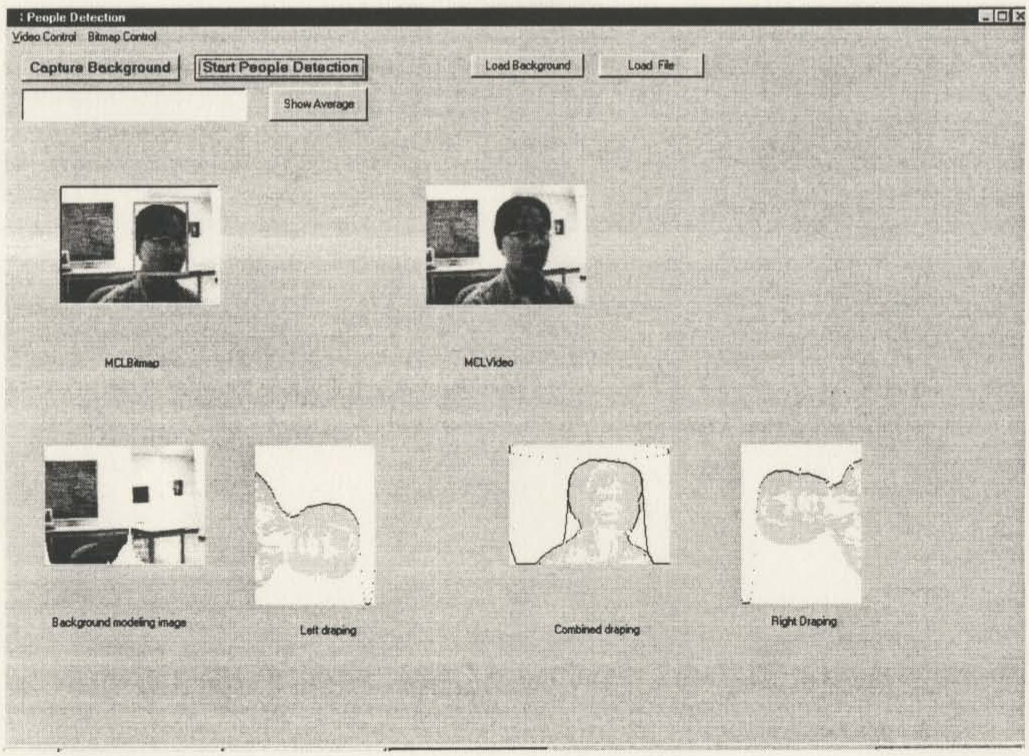


Figure 5.5 Graphic user interface showing example images

5.4 Experiments and Results

A series of experiments have been carried out to test the effectiveness and accuracy of the improved system. The three background formation methods are tested and compared against a common inter-frame variance map as shown in Figure 5.6a. There is a notch appearing on the top of the head because of a small movement, and similar shading between the foreground and background. The original drape penetrates into this notch as shown in Figure 5.6b. This part of object will be misinterpreted as background if the background is formed from the original drape. Median filtering using three iterations can partially recover this part of the object as foreground (Figure 5.6c). Better improvement can be achieved by more iterations at the sacrifice of the computation time. The result of multiple draping is shown in Figure 5.6e, together with its left drape (Figure 5.6d) and right drape (Figure 5.6f). The combined drape, which is the outmost part of the three drapes, can nearly recover the entire notch as the foreground (see Figures 5.6d, e and f). The smoothed drape by using polynomial interpolation is shown in Figure 5.6g, which also indicates a significant improvement. For the three methods used, median filtering can recover most of the notch back as the foreground. Both multiple draping and polynomial interpolation result in a significant improvement, and using these two methods, almost the entire background can be recovered.

The improved system can track the human head accurately and quickly in real time. The background formation is switched on whenever the difference between the consequent frames is sufficiently great. Figures 5.7, 5.8 and 5.9 are an image sequence showing the system performance at frames 5, 50 and 100. The inter-frame variance map

shown in Figure 5.7a is obtained from the first five consequent images. Figure 5.7b shows the effect of multiple draping applied on this inter-frame variance map. The resultant multiple drapes are shown in Figure 5.7b. The recovered "Don't know" region is shown as the white patch in Figure 5.7c, which occupies a significant portion of the image. Meanwhile, the incoming image (frame 5) subtracts the current background to yield a segmented image and then the draping is applied on it (Figure 5.7d). The head is identified as shown in Figure 5.7d. Since the "Don't know" area is very large at this stage, the background formation needs to be continued. The background formation is estimated from every five consequent frames, while the background subtraction and head tracking is performed at frame rate. Figure 5.8 shows the system performance at frame 50. Figure 5.8a is the inter-frame variance map resulting from frame 96 to frame 100. The multiple draping and the resultant "Don't know" area are shown in Figure 5.8b and c respectively. At this stage, the "Don't know" area shown in Figure 5.8c is reduced significantly after several times of background formation. The updated background is subtracted from the incoming image, followed by a draping (Figure 5.8d). The head border locates the head more accurately than it does at frame 5 because the updated background is much closer to the true background. This process repeats for frame 100 shown in Figure 5.9. The generated background shown in Figure 5.9c is very close to the true background with only a small patch of "Don't know" area. The system detects this to be less than the given threshold and regards it as the true background. It will switch on its background formation mode if it detects that the difference between the consequence

images is greater than the given threshold. Figure 5.9d shows the accurately tracked human head.

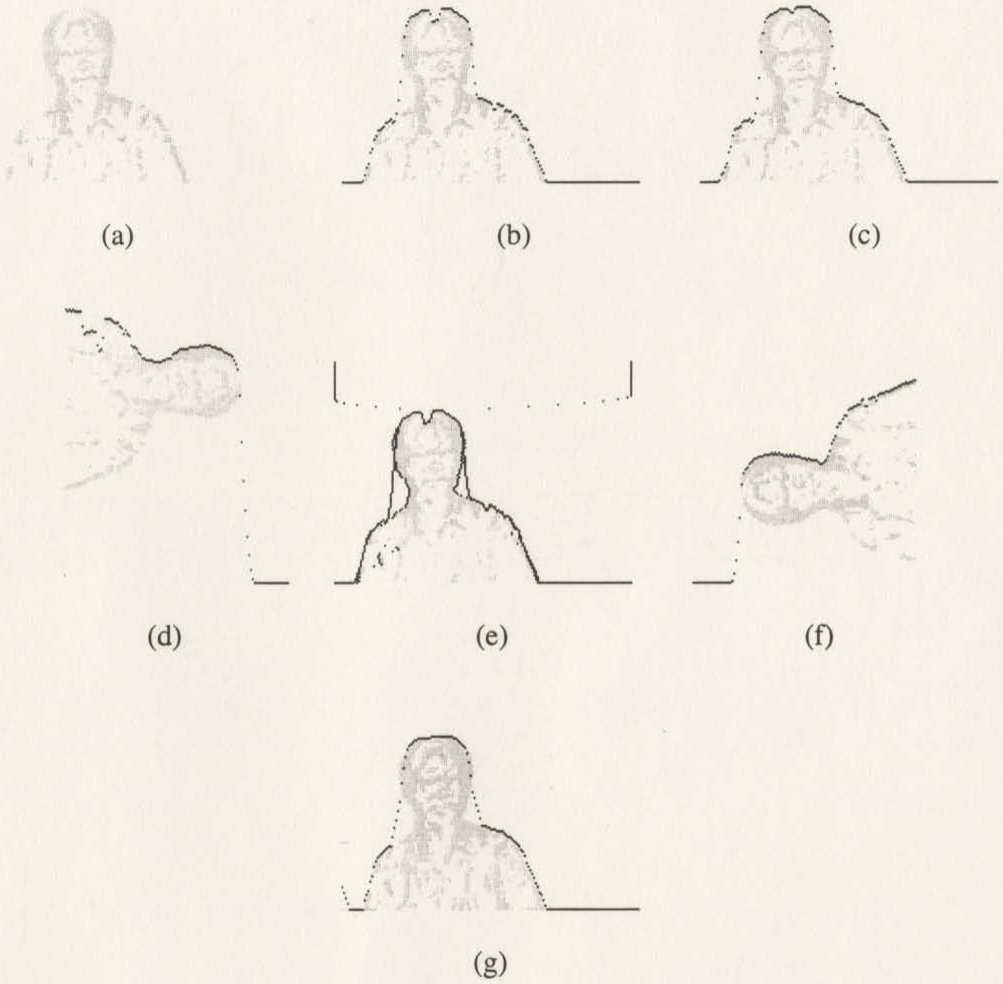
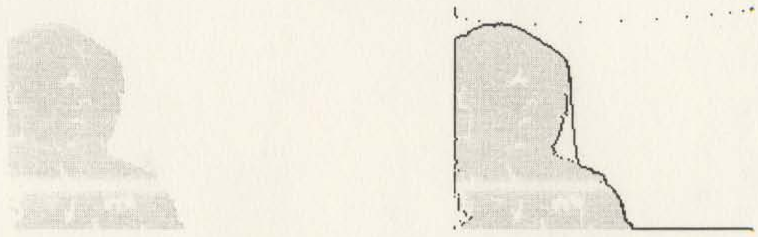


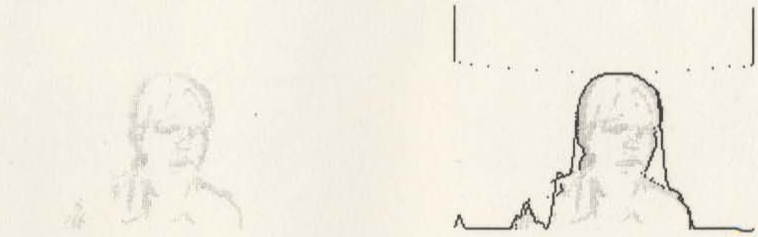
Figure 5.6 Comparison of three methods



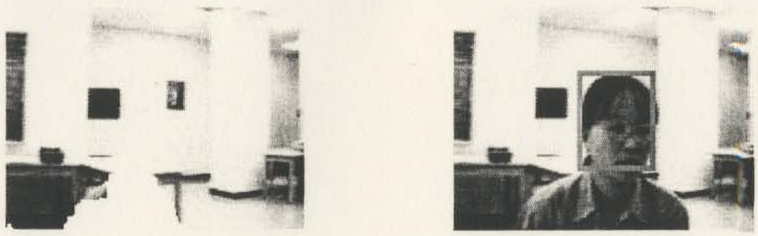
(a) (b)

(c) (d)

Figure 5.7 System performance at frame 5



(a) (b)



(c) (d)

Figure 5.8 System performance at frame 50

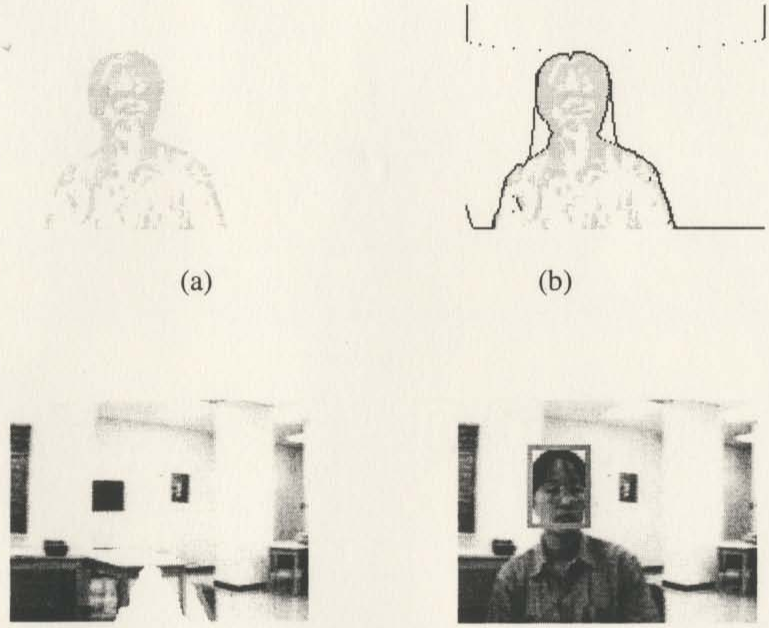


Figure 5.9 System performance at frame 100

Chapter Six

Conclusions and Future Work

A real-time face tracking system has been described in the preceding chapters. This chapter concludes the author's thesis work and points out the possible directions of further work.

6.1 Conclusions

A real-time face tracking system has been developed to track upright people's heads from various complex background scenes. The system setup consists of a digital camera, a personal computer and supporting face-tracking tools. A graphical user interface was also designed that allows the user to interact conveniently with the system. Combining with background subtraction, a draping technique was identified as the basic head tracking method. The draping method has been modified and extended throughout the study aiming at constructing head and shoulder silhouettes of the foreground objects. This enables accurate recovery of the desired foreground objects from the segmentation

map without including noise from the background scene. Once the drape is settled, an innovative head identification method consisting of a drape interpolation followed by a head merging was investigated to achieve an accurate head extraction. A significant amount of effort was focused on background formation because the accuracy of the system depends substantially on the background maintenance. A multiple state system was proposed to tackle the changes happening in the background scene. This system can switch automatically between background formation and face tracking depending on the detected background information. The dynamic background formation process continues until the "Don't know" area in the generated background is below the threshold and resumes whenever the difference between the incoming background and the generated background become significant. Median filtering, multiple draping and polynomial interpolation were also incorporated into the system to deal with problems associated with a bad drape. This system is implemented in a Pentium II 200 MHz personal computer. In the current setting, this system can process three frames (160×120 pixels image) per second. A variety of experiments have been carried out in the course of the study to test and verify the system's accuracy and robustness, which are also presented throughout the thesis. In summary, the major contributions of the thesis are:

1. Designing and developing a real-time face tracking system.
2. Modifying and extending the draping method so that it can be used on both an inter-frame variance map and the original segmentation map. The improved draping method is proved to be more flexible and robust than its predecessor.

3. Developing an innovative face tracking method consisting of an exhausting head search scheme followed by head merging. This method significantly improves the accuracy of face location and identification of the system.
4. Developing a multi-state fully automatic background formation technique to tackle both gradual and sudden background changes. This technique greatly improves the system's stability, efficiency and reliability of the background maintenance under diverse and complex situations, and hence allows accurate recovery of the correct foreground objects.
5. Conducting a significant number of experiments under various situations to test and verify the system performance. Experiments indicate that the real-time face tracking system can function well despite the variation in lighting, both local and global background movements and shadows in the background scene.

6.2 Future Work

Although the system performs successfully under a variety of background scenes, it may be further improved. The possible alternatives include:

1. Instead of using median filtering, some other filtering technique can be used to process the inter-frame variance map.
2. Using Gsnakes dynamic contour to extract the boundary of the foreground objects as it relies on detailed object-shape or object-motion models to distinguish between the foreground and the background.

3. A fixed number of iterations is used in smoothing the *HeadWidth*-row curve disregarding the results of the curve fitting. The system can be further optimized if it can automatically determine the iteration requirements depending on the different situations.
4. The system can be further updated by incorporating additional functions to recognize gestures, postures and body parts.

References

Azarbayejani, A., Starner, T., Horowitz, B., and Pentland, A., 1993. "Visually controlled graphics". IEEE Transactions on Pattern Analysis and Machine Intelligence,15(6), pp. 602-605.

Basu, S., Essa, I., and Pentland, A., 1996. "Motion regularization for model-based head tracking". ICPR.

Baumberg, A., and Hogg, D., 1994. "An efficient method for contour tracking using active shape models". Proceedings of the Workshop on Motion of Nonrigid and Articulated Objects, IEEE Computer Society.

Bichsel, M., 1994. "Segmenting simply connected moving objects in a static scene". Pattern Analysis and Machine Intelligence, 16(11), pp. 1138-1142.

Black, M.J., and Yacoob, Y., 1995. "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions". IEEE International Conference on Computer Vision, Cambridge, MA, pp. 374-381.

Blake, A., and Isad, M., 1994. "Three-dimensional position and shape input using video tracking of hands and lips". Proceedings. of SIGGRAPH'94, pp. 185-192.

Bobick, A., Davis, J., Intille, S., Baird, F., Cambell, L., Irinov, Y., Pinhanez, C., and Wilson, A., 1996. "KidsRoom: Action recognition in an interactive story environment". M.I.T. TR No: 398.

Buckingham, E., 1914. "On physically similar system". Phys. Rev. **4**, p. 345.

Cascia, M. La, Isidoro, J. and Sclaroff, S., 1998. "Head tracking via robust registration in texture map images". Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, pp. 508-514.

Cheng, L.T. and Robinson, J., 1998. "MCLGallery: A framework for multimedia communications research". In Proceedings of CCECE'98.

Cootes, T., Taylor, C., Lanitis, D., Cooper, D. and Graham, D., 1993. "Building and using flexible models incorporating grey-level information". Proc. ICCV, pp. 242-246.

Covell, M., and Darrell, T., 1999. "Dynamic occluding contours: A new external-energy term for snakes". Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 232-238.

Darrell, T., Gordon, G., Harville, M., and Woodfill, J., 1998. "Robust visual person tracking for interactive displays". Proceedings of the 1998 workshop on perceptual user interfaces, pp. 601-608.

Darrell, T., Maes, P., Blumberg, B., and Pentland, A., 1994. "A novel environment for situated vision and behavior". In Proceedings of CVPR-94 Workshop for Visual Behaviors, Seattle, WA, pp. 68-72.

DeCarlos, D. and Metaxas, D., 1996. "The integration of optical flow and deformable models with applications to human face shape and motion estimation". Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition San Francisco, CA, pp. 231-238.

Friedman, N., and Russell, S., 1997. "Image segmentation in video sequence: a probabilistic approach". In Proceedings of the thirteenth conference on uncertainty in artificial intelligence (UAI).

Gavrila, D. M., and Davis, L. S., 1995. "3-D model-based tracking of human upper body movement: a multi-view approach". Proceedings of the IEEE International Conference on Computer Vision, Coral Gables, FL, pp. 253-258.

Grimson, W.E.L, et al., 1998. "Using adaptive tracking to classify and monitor activities in a site". Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, pp. 22-29.

Haritaoglu, I., Harwood, D., and Davis, L.S., 1998b. "Ghost: A human body part labeling system using silhouettes". The Fourteenth international conference on pattern recognition, Brisbane, Australia.

Haritaoglu, I., Harwood, D., and Davis, L.S., 1998c. "W4s: a real-time system for detecting and tacking people in 2½ D". The Fifth European conference on computer vision.

Haritaoglu, I., Harwood, D., and Davis, L.S., 1998a. "W^4: A Real Time System for Detecting and Tracking People". Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, pp. 962.

Horprasert, T., Haritaoglu, I., Wren, C., Harwood, D., Davis, L., and Pentland, A., 1998. "Real-time 3D Motion Capture". Proceedongs of 1998 Workshop on Perceptual User Interfaces (PUI'98), San Francisco, CA.

Intille, S., Davis, J., and Bobick, A., 1997. "Real-Time Closed-Word Tracking". Proceedings of CVPR, June 1997, pp. 697-703.

Jebara, T.S., and Pentland, A., 1997. "Parametrized structure from motions for 3D adaptive feedback tracking of faces". Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, pp. 144-150.

Koller, D., Weber, J.W., and Malik, J., 1994. "Robust multiple car tracking with occlusion reasoning". In European Conference on Computer Vision.

Metaxas, D., and Terzopoulos, D., 1993. "Shape and non-rigid motion estimation through physics-based synthesis". T-PAMI, 15, pp. 580-591.

Nakai, H., 1995. "Non-parameterized Bayes decision method for moving object detection". In Asian Conference of Computer Vision, Singapore.

Oliver, N., Rosario, B., and Pentland, A., 1999. "A Bayesian computer vision system for modeling human interactions". International conf. On vision systems, Gran Canaria, Spain.

Olson, T., and Brill, F., 1997. " Moving objects Detection and Event Recognition algorithms for Smart Cameras". In Proc. DARPA Image Understanding Workshop, pp. 159-176

Pentland, A., and Horowitz. B., 1991. "Recovery of nonrigid motion and structure". IEEE Trans Pattern Analysis and Machine Intelligence, 13(7), pp. 730-742.

Pentland, A., Moghaddam, M., and Starner, T., 1994. "View-based and modular eigenspace for face recognition". Proceedings of IEE Conference on Computer Vision and Pattern Recognition, pp. 84-91.

Qian, R.J., Sezan, M.I., and Mathews, K., 1998. "Robust real-time face tracking algorithm". IEEE International Conference on Image Processing Chicago, IL, pp. 131-135.

Qing, S. and Robinson, J., 1999. "Optical flow used in face image classification". Proceedings of the Ninth Newfoundland Electrical and Computer Engineering Conference, St. John's, NF.

Rehg, J., and Kanade, T., 1994. "Visual tracking of high dof articulated structures: An application to human hand tracking". In ECCV94, pp. 35-46.

Rohr, K., 1994. "Towards model-based recognition of human movements in image sequences". CVGIPiu, 59(1), pp. 94-115.

Rowley, H.A., Baluja, S., and Kanade, T., 1998. " Human face detection in visual scene". CVPR.

Schödl, A., Haro, A.. and Essa, I., 1998. "Head Tracking using a Textured Polygonal Model". Georgia Tech, GVU Technique Report # 24.

Stafford-Fraser, Q., 1996. "Video-Augmented Environments". PhD. dissertation, Gonville & Caius College, University of Cambridge.

Terzopolous, D., and Szeliski, R., 1992. "Tracking with Kalman snakes". In Active Vision (Blake, A., and Yuille, A.,), Cambridge: MIT, MA, pp.3-20.

Toyama, K., Krumm, J., Brumitt, B. and Meyers, B., 1999. "Wallflower: Principles and practice of background maintenance". International conference on computer vision, Corfu, Greece.

Turk, M., 1996. "Visual interaction with lifelike characters". Proc. 2nd international conference on automatic face and gesture recognition, IEEE computer society press, Killington, VT, pp. 368-373.

Turk, M., 1998. "Moving from GUIs to PUIs". Symposium on Intelligent Information Media, Tokyo, Japan.

Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A., 1996. "Pfinder: real-time tracking of the human body". Proceedings of the International Conference on Automatic Face and Gesture Recognition Killington, VT, pp. 51-56.

Yang, J. and Waibul, A., 1996. "Real-time face tracker". IEEE Workshop on Applications of Computer Vision – Proceedings, Sarasota, FL, pp. 142-147.