國 立 交 通 大 學

電 機 與 控 制 工 程 研 究 所

碩 士 論 文

自動影像分割技術及其於人體偵測與深度估測之應用

Automatic Image Segmentation and its Applications to Human Detection and Depth Estimation

研 究 生：曾 筱 君

指 導 教 授：張 志 永

中 華 民 國 九十八 年 七 月

自動影像分割技術及其於人體偵測與深度估測之應用

Automatic Image Segmentation and its Applications to Human Detection and Depth Estimation

學　　生：曾筱君　　Student : Hsiao-Chun Tseng

指導教授：張志永　　Advisor : Jyh-Yeong Chang

國立交通大學

電機與控制工程學系

碩士論文

A Thesis

Submitted to Department of Electrical and Control Engineering

College of Electrical Engineering and Computer Science

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of Master in

Electrical and Control Engineering

July 2009

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 八 年 七 月

自動影像分割技術及其於人體偵測與深度估測之應用

學生：曾筱君　　指導教授：張志永博士

國立交通大學電機與控制工程研究所

摘要

影像分割技術在醫學影像處理、交通流量監測、人體偵測和多媒體等方面的應用中佔有主要的地位。深度估測也是多媒體與服務機器人應用的一個重要的課題之一。在無法事先取得任何背景資訊的情況下，從各種不同的場景中將人體或其他感興趣的物體抽取出來當成前景，且從單張影像所獲得的資訊估測人體，或者物體的深度將是一件非常具有挑戰性的工作。為了解決這個問題，我們結合了影像中特徵、輪廓和空間分佈等資訊來判別不同的分割區域是否代表著同一個物體，並將判別為同一個物體的不同區域做合併。接著我們使用了兩種不同的深度估測方法對已經從場景中被截取出來的人體作深度的估測。這兩種深度估測方法分別是以消失線及消失點為基礎的估測方式，和以固定相機之查表法的估測方式來重建二維平面影像的三維景深資訊。

在此篇論文中，我們結合了影像分割和人臉偵測技術，希望能在不同場景中將人體抽取出來。首先，我們利用了膚色資訊與橢圓樣板比對找出人臉在影像中的位置，接著我們提出了一套改良式自動種子區域成長演算法來分割影像。在我們提出的分割演算法中，初始種子將會自動產生，且剩餘未分類的像素將會歸到其最接近的區域。在影像初始分割完成後，任兩個相鄰的區域如果具有高相似性將被合併。再根據人體的形態找出人體的範圍，將分割結果屬於人臉與人體的區

域作合併，完成前景人物偵測，並確定人體的位置。最後，我們將偵測人體位置的垂直 y 座標值，並對其做深度估測。假如影像中的消失線或消失點可以被偵測，我們可採用 cross-ratio 的關係式來估測深度。另一方面，我們藉著建立相機深度查表法，我們可採用查表法來估測其深度。

Automatic Image Segmentation and its Applications to Human Detection and Depth Estimation

STUDENT: Hsiao-Chun Tseng ADVISOR: Dr. Jyh-Yeong Chang

Institute of Electrical and Control Engineering
National Chiao-Tung University

ABSTRACT

Image segmentation plays an essential role in applications such as medicine image processing, traffic flow magnitude monitored, human detection, multimedia applications, and many others. Depth estimation is one of the important topics in multimedia and service robot applications. It is a very challenging task to extract human or other objects of interested from scenes without any background information, and then to estimate the human depths from single camera view. To solve this, we adopt the method which combines the feature-based, shape, and space information of an image to recognize different segmented regions. Then we estimate the human depth based on vanishing line and point, or based on camera's depth look-up table.

In the thesis, we combine image segmentation techniques and face detection methods to extract the human from scenes. Firstly, skin regions are detected and an ellipse fitting method is employed to detect the face region and consequently locate the human position. Then we propose an improved automatic seeded region growing algorithm to segment the image. The initial seeds are generated automatically, and the remaining pixels are classified to the nearest region. After the region growing procedure, two neighboring regions with high similarity are merged. The human body

is determined by confining semantic human body region in segmented regions, and those belonging to the human face and human body are merged afterward. The human is extracted and the human position is also decided. Lastly, we will detect the human vertical *y*-coordinate values in the image, and the depths can then be estimated according to the cross-ratio formula or the depth look-up tables of the camera.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all the teachers who once gave me more knowledge edification for valuable suggestions, guidance, supports and inspirations. Thanks are also given to all the people who assisted me in completing this research.

Finally, I would like to express my deepest gratitude to my family for their concern, supports and encouragements.

Content

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Motivation

Multimedia applications in daily life become widespread used for education, security, entertainment and medicine, and provide more additional value for clients. Many image segmentation techniques are used in multimedia applications and service robot, can subdivide an image or video frame into its constituent regions or objects and then become an important topic in recent years. The objects of interested can be extracted from an image as the foreground and are then used in industrial inspection, autonomous target acquisition, medicine image processing, traffic flow magnitude monitored, human detection, depth estimation, and etc.

Face detection techniques develop rapidly and are also an important topic in computer vision. A robust face detection system can locate each human face from an image and are widely applications in face recognition, face tracking, automatic surveillance system, human-machine interface, and home care system.

The human visual system has a strange ability to recognize objects from single view. However, for computer vision, it is a very challenging task to extract objects from scenes without any background information. Therefore, color feature, texture feature, shape matching, and any other information acquired by analyzing original image are used to achieve more accurate segmentation in computer vision. Since the display applied on 2D image does not satisfy the user's requirement, we reconstruct the stereo image from single camera view by estimating the extracted subject depths.

Summary of above, this motivates us to design a robust method that adopts feature-based, shape-based, and space information of an image to recognize different

segmented regions as identical object and confirm object boundaries. The face detection techniques are utilized and located human face by extracting skin-color and then fitting by the ellipse templates. We also propose an improved automatic seeded region growing algorithm to simplify pre-process procedure. Furthermore, human bodies are extracted by using some semantic rules. After the human extraction procedure, the depths are estimated by analyzing the image information and acquiring the vertical *y*-coordinate values in the image.

The system flowchart is illustrated in Fig 1.1. Our system can be separated into three components. The first component is image segmentation. Edge detection methods are utilized initially. The detected edges are used for elliptical template matching and automatic seed selection. Then the color image transformed from RGB to YC_bC_r color space is used for skin-color extraction and region growing procedure. The second component is human recognition. The interesting human is extracted by using some semantic rules and the adjacent regions with high similarity are merged to accomplish the human recognition. The third component is depth estimation. Hough transform is utilized for detecting the intersection of two or more parallel lines in the world, and the vanishing lines and points are detected. Then the vertical *y*-coordinate values in the image are detected and the depths are estimated base on the vanishing line or the depth look-up table of the camera.

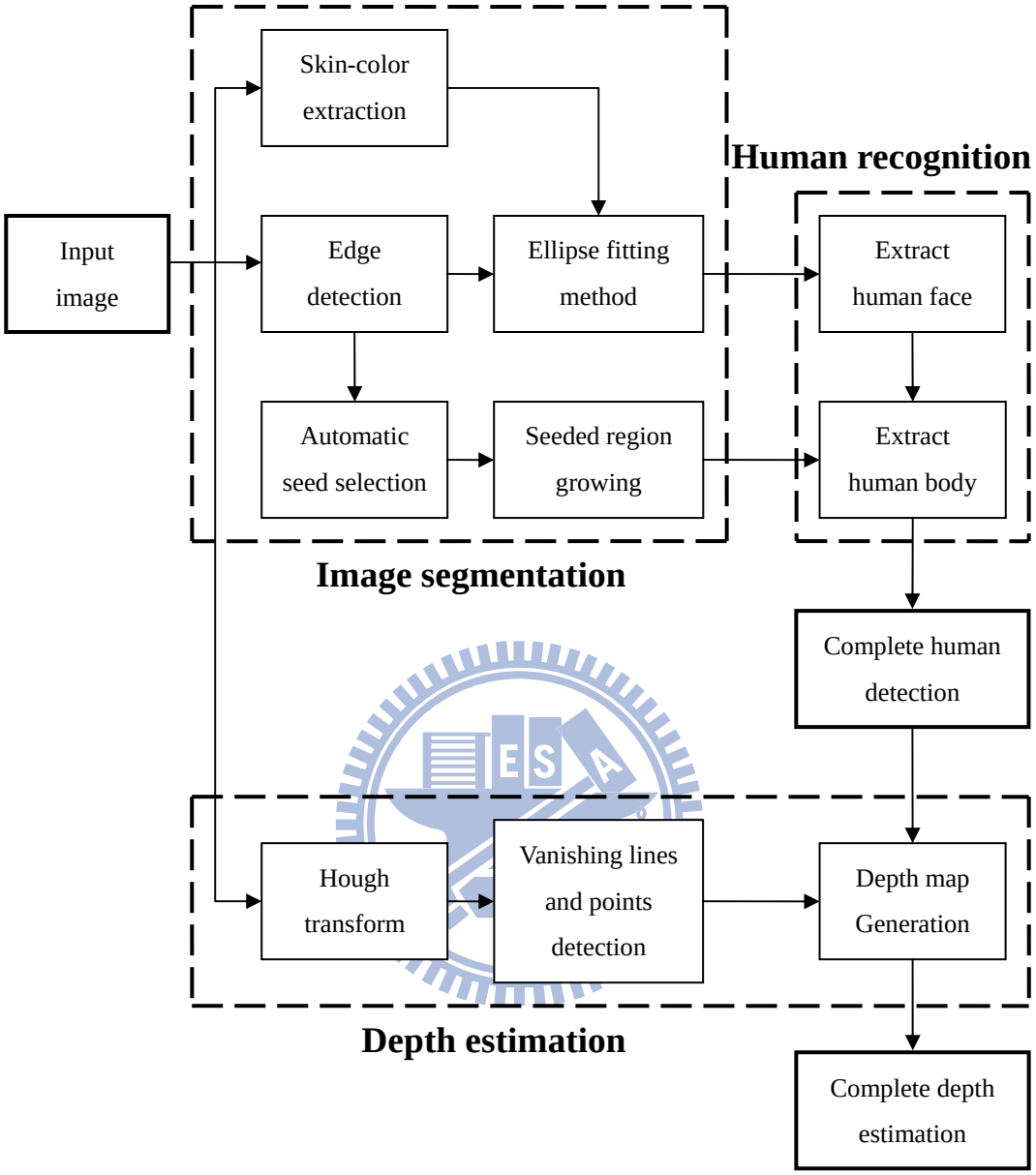


Fig 1.1 The block diagram of human recognition system.

1.2 Face Detection

The purpose of face detection is to localize and extract the face region from the background. Most of face detection techniques can be classified as feature-based and image-based approaches [1]. Feature-based approaches can be divided into low level

analysis, feature analysis and active shape model. Image-based approaches can be divided into linear subspace method, neural networks and statistical approach. A more detail techniques is shown in Fig. 1.2.

Color method is one of the low level analyses and a number of skin-color models have been constructed. A color image is specified in RGB components. The RGB model is suitable for color display, but is not good for color analysis because of its high correlation among *R*, *G*, and *B* components. Besides, the distance in RGB color space does not represent the perceptual difference in a uniform scale. In image processing and analysis, we often transform these components into other color spaces such as normalized RGB, HSV, CIE(Lab), HIS, YIQ, and YC_bC_r. In this paper, we detect the skin-region in the YC_bC_r color space for three reasons: First, an effective use of the chrominance information for modeling skin-color can be achieved in the YC_bC_r color space. Second, the chrominance components (C_b and C_r) are explicitly separated from the luminance (*Y*) component in the YC_bC_r model. Third, the YC_bC_r space is typically used in most image and video coding standards.

In most cases, the shape of the human face is similar to an ellipse. An ellipse fitting method is one of the feature analyses and is employed to extract the face region. We can detect the edges of the original image initially. Then, edge detection results are utilized to perform elliptical template matching. The parameters that describe the center of the ellipse, the major and minor radii, have high variations of the shapes of human faces. If dealing with the whole image and cover the high shape variations, the computation burden would be very high. We utilize instead an elliptical model to search for the human face proposed by Tang and Chen [2], which is much more computational efficient. Furthermore, we improve to do the elliptical template matching only to the possible region instead of the whole image.

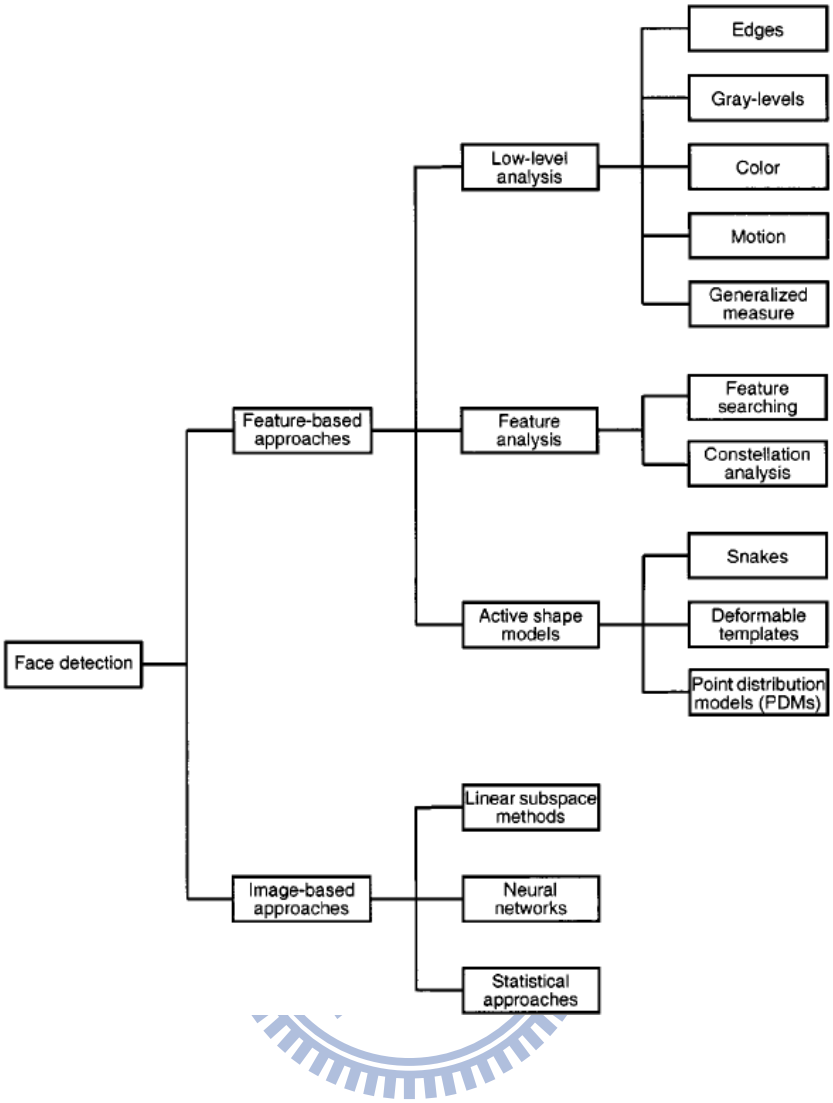


Fig. 1.2 Face detection divided into two approaches [1].

1.3 A Hybrid Method of Seeded Region Growing

Many image segmentation techniques have been proposed in the past few years. Most of image segmentation techniques can be classified into four main categories: thresholding, edge-based, region-based, and hybrid techniques. Thresholding methods are based on the assumption that adjacent pixels whose value lies within a certain range belong to the same class. Edge-based methods assume that the pixel values change abruptly at the boundary between two different regions. Region-based

methods assume that adjacent pixels within the same region should have similar visual features such as intensity, color, or texture. Hybrid methods tend to integrate the results of boundary detection and region growing together to achieve better segmentation.

Hybrid methods of the image segmentation integrate the results of boundary detection and region growing to improve the drawback of discontinuous edges detected by edge-based method only. A hybrid method of seeded region growing (SRG) is widely used for segmenting object regions or boundaries from images. It is first proposed by Adams and Bischof [3], starts with assigned seeds, and grows regions by merging a pixel into its nearest neighboring seed region. The initial seeds are manually selected. Then, Fan *et al.* [4] presented a color image segmentation algorithm automated the initial seed selection by integrating color-edge extraction and seeded region growing on the YUV color space. The *Y*, *U*, *V* components perform edge detection procedure individual and edges are decided by the fast entropic thresholding technique. Then, edge results for the three color components are integrated to obtain color edges. The centroids between adjacent edge regions are taken as the initial seeds. It is too sensitive and may cause over-generated. Shih *et al.* [5] proposed an automatic seeded region growing algorithm and initial seeds are generated if pixel has high similarity to its neighbors. The disadvantage is that the computation is much complicated.

In order to decrease the computation complexity, we develop a simpler method. It detects the edges using Canny operator [6] initially. If neighbors of one pixel have no edge or few detected edges, the pixel is considered as a seed candidate.

1.4 Depth Estimation

Depth estimation has become an important topic in multimedia applications and several methods have been proposed to reconstruct the stereo image from a 2D image. Xiong *et al*. [7] presented a method to obtain depth information from focus and defocus. Criminisi *et al*. [8] compute the 3D affine measurements from a single perspective view of a scene given only the vanishing line and point information determined from the image. Luong *et al*. [9] presented that correspondences between three images taken by the same camera with fixed internal parameters are sufficient to recover the internal parameters of the camera, to compute coherent perspective projection matrices, and to reconstruct 3D structure up to a similarity.

In the thesis, we estimated the human depth based on vanishing lines and points. Sometimes the vanishing lines and points are not available, we utilized another method based on depth look-up table of the camera to estimate the depths.

1.5 Thesis Outline

The thesis is organized as follows. The skin-color extraction method and the elliptical template matching technique are described in Chapter 2. In Chapter 3, an improved method of automatic seeded region growing is introduced. In Chapter 4, the experiment results of our object segmentation and depth estimation system are shown. At last, we conclude this thesis with a discussion in Chapter 5.

Chapter 2 Human Face Detection

In this chapter, we improve the face extraction method to enhance the location detection accuracy of subject interests. Firstly, we utilize skin-color extraction method in the YC_bC_r color space. Secondly, we utilize an elliptical model to match human face. Consequently, we can combine skin-color extraction and ellipse fitting to better estimate the location and size of a human face. Then, the size of the face image can be indirectly exploited to estimate the human distance from the camera to the human.

2.1 Skin-Color Extraction

2.1.1 YC_bC_r Color Space

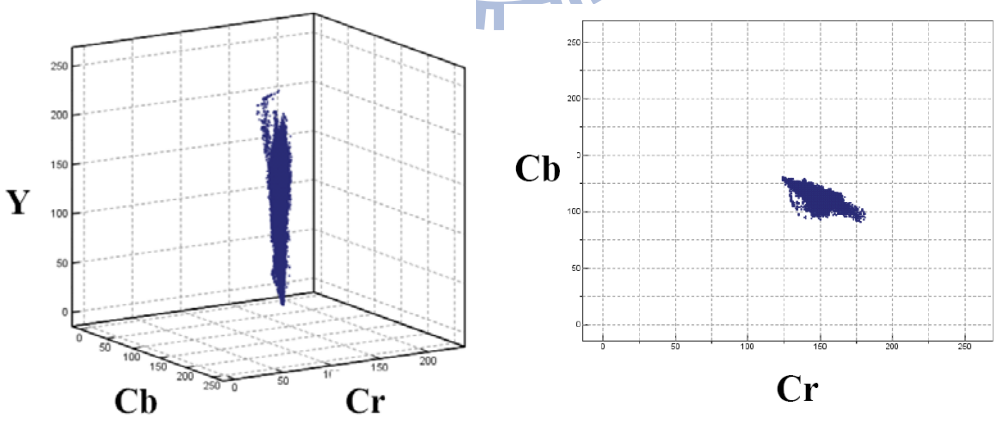
The first stage in the face detection algorithm makes use of skin-color extraction to distinguish the face-region from non-face-region. Several color spaces suitable for segmenting the skin-color in an image have been proposed. Choosing the representative and discriminative color space for the skin-color modeling becomes very important. Although different races have different skin colors, several studies have shown that the major difference lies largely between their luminance rather than their chrominance [10]. In [11], YC_bC_r and HSV color spaces for skin-color segmentation have been investigated. It was concluded that the skin color distribution in YC_bC_r color space is more centralized than HSV color space. The color space of YC_bC_r, which revises the color space of YUV, can divide luminance component (Y), and two chromatic blueness component (C_b), redness component (C_r). The transformation between YC_bC_r and RGB is linear and is represented as

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \quad (1)$$

The YC_bC_r model is naturally related to MPEG and JPEG coding. The skin color distribution in YC_bC_r color space is more centralized than other color spaces, and the advantage of converting the image to the YC_bC_r color space is that the effect of luminosity can be decoupled with coloring components during the image processing. For this reason, we utilize YC_bC_r color space for skin color region detection.

2.1.2 Robust Skin-Color Extraction Technique

We obtain a skin-color reference map in YC_bC_r color space by classifying the pixels of the input image into skin region and non-skin region. In Fig. 2.1, it has shown that the skin-color region is distributed consistently in the YC_bC_r color space and forms a very compact area in the C_bC_r space.



(a) (b)

Fig. 2.1 Distribution of samples of skin-color pixels in

(a) YC_bC_r 3D space, (b) a 2D projection in C_bC_r subspace [13].

The suitable ranges that Chai and Ngan [12] found for skin-color regions are

$R_{C_b} = [77, 127]$ and $R_{C_r} = [133, 173]$. Garcia and Tziritas [11] constructed a more complex skin color decision boundary up of eight planes in the YC_bC_r space.

Chi *et al*. [13] presented a robust skin-color extraction technique and resulted in less erroneous pixels. In Fig. 2.2, for each C_r in the skin-color region, the maximum and minimum C_b are used to estimate the upper and lower quadratic functions. A pixel is labeled as skin candidate if it falls within the locus. Skin pixels of the proposed approach are determined by (2)

$$Skin(C_b, C_r) = \begin{cases} 1, & (C_b < \Gamma_{up}(C_r)) \cdot (C_b > \Gamma_{bottom}(C_r)) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where $\Gamma_{up}(x) = \alpha_2 x^2 + \alpha_1 x + \alpha_0$ and $\Gamma_{bottom}(x) = \beta_2 x^2 + \beta_1 x + \beta_0$

For the upper bound, the quadratic coefficients found are $\alpha_2 = -0.0225$, $\alpha_1 = 6.1251$, and $\alpha_0 = -290$ while the lower bound coefficients are $\beta_2 = 0.0284$, $\beta_1 = -9.1477$, and $\beta_0 = 836$.

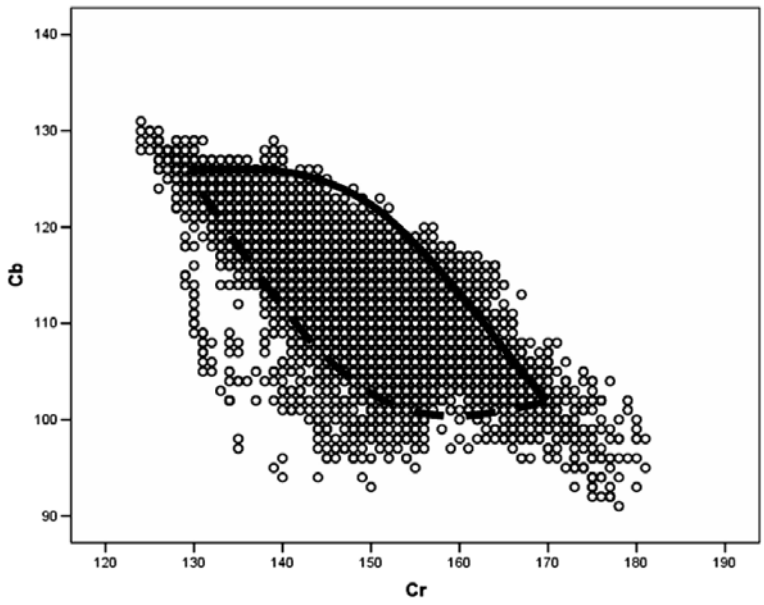


Fig. 2.2 Samples distribution of skin-color pixels in C_bC_r space [13].

2.2 Elliptical Face Matching

The second stage in the face detection algorithm uses ellipse to fit the human face. Under the most likely line of reasoning, we try several elliptical models with a hierarchical structure proposed by Tang and Chen [2] to search for the best-fit human face.

2.2.1 Elliptical Template

An ellipse can be described by the following equation:

$$\frac{(x-x_0)^2}{S_x^2}+\frac{(y-y_0)^2}{S_y^2}=1 \tag{3}$$

where $(x_0, y_0)^T$ is the center of the ellipse, S_y and S_x are one-half of the major and minor diameters satisfied by $S_y = \sigma \cdot S_x$. In this thesis, the face is modeled as a vertical ellipse with a fixed aspect ratio of $\sigma = 1.1, 1.2,$ or 1.3 in our experiments. The elliptical template model is shown in Fig. 2.3. Let $r_0 = (x_0, y_0)^T$, $s = S_x$, and $v_s^i, i = 1, 2, \ldots, N,$ denote the points located uniformly on an ellipse of size s centered at r_0.

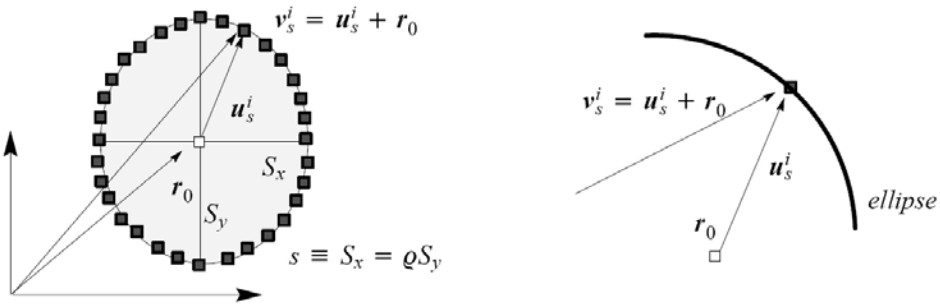


Fig. 2.3 One-pixel-width elliptical template [2].

In [2], the elliptical template is described as

$$T_{r_0,s}(r) = \sum_{i=1}^{N} h_i \delta(r - u_s^i - r_0) \tag{4}$$

where $\boldsymbol{r} = (x, y)^T$ are the image coordinates, $\delta(\cdot)$ is the delta function, and $\boldsymbol{h_i} = [h_{xi}, h_{yi}]^T_{,}\ i = 1, 2, \ldots, N,$ are the weighting factors.

Then, let g be an image of intensity gradients defined as the vector function $\boldsymbol{g} : \boldsymbol{D} \rightarrow \boldsymbol{R}^2,$ where $\boldsymbol{D} = \{ (x, y) \,|\, x = 1, \ldots, P, y = 1, \ldots, Q \}$ and $\boldsymbol{R}^2$ is the set of all possible intensity gradient vectors. We let $G = \{ g(r), r \in \boldsymbol{D} \}$ and assume that a noisy gradient image $\{ g(r), r \in \boldsymbol{D} \}$ containing an elliptical contour can be modeled as

$$g(r) \,|\, r_0, s = T_{r_0,s}(r) + \eta(r) \tag{5}$$

where the 2×1 noise vector $\eta(r)$ is assumed to be Gaussian.

The joint probability density function $p(G|r_0, s)$ is also Gaussian, and can be shown as

$$p(G \,|\, r_0, s) = \frac{1}{C} \exp\left\{ \frac{2\sum_{i=1}^{N} h_i^T g(u_s^i + r_0) - G^T G - \text{const}}{2\sigma_\eta^2} \right\} \tag{6}$$

where $C = (2\pi\sigma_\eta^2)^{N/2}$. The ML estimates the face position $r_0,$ and size s can be obtained by maximizing $p(G|r_0, s)$ with respect to $r_0, s,$ which is equivalent to maximizing $F(r_0, s)$ with respect to r_0 and $s,$ in which

$$F(r_0, s) \equiv \sum_{i=1}^{N} h_i^T g(u_s^i + r_0) \tag{7}$$

2.2.2 A Hierarchical Structure of Searching

An image pyramid containing k levels of images (level 1: $P \times Q$, level 2: $\frac{1}{2}P \times \frac{1}{2}Q$, level 3: $\frac{1}{4}P \times \frac{1}{4}Q$, …, level k : $\frac{1}{2^{k-1}}P \times \frac{1}{2^{k-1}}Q$) is constructed. For each image in the pyramid, using a template of size s for human face detection in the low-resolution image of level k is equivalent to perform a coarse template matching with size $s \times 2^{k-l}$ in the high-resolution image of level l. For example, the template matching using $s = 8, 9, 10, 11, 12$ in the 256×192 image can roughly handle the matching process of using $s = 64, 72, 80, 88, 96$ in the 2048×1536 image.

2.3 An Improved Method of Human Face Detection

Since not all skin color regions are human face regions, we will utilize elliptical template matching method to locate the position of the human face explicitly. It is computational efficient to utilize an elliptical model to search for the face from some potential facial regions instead of the entire image. For the image f' of size $\frac{1}{k}P \times \frac{1}{k}Q$ which is resized from the original image f of size $P \times Q$, the edge results G are detected by the Canny edge detector and the skin-region S is described as

$$S(x, y) = \begin{cases} 1, & \text{if } Skin(C_b(x, y), C_r(x, y)) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $x = 1, 2, \ldots, \frac{1}{k}P$ and $y = 1, 2, \ldots, \frac{1}{k}Q$.

For every pixel in G, its 4-neighbors are also marked as the edge pixel if the pixel is detected as edge pixel and are adapted as G'. The adapted edge results G' are considered as the extended edges. From Eq. (3), we can construct a hierarchical ellipse model E having a set of templates with different sizes. Because the best candidate of the elliptical face matching might be detected nearby the skin region, the center of the ellipse (x_0, y_0) should be located inside the detected skin region. We utilize a hierarchical ellipse model E to fit the human face with G'. Instead of searching for the entire image, the ellipse center (x_0, y_0) is located at (x, y) in which $S(x, y) = 1$.

After the elliptical templates have been generated, measuring the goodness of match between the ellipse model and the edge result nearby the skin region can be measured as

$$T_1(r_0, s) = \frac{1}{N} \cdot F(r_0, s) = \frac{1}{N} \sum_{i=1}^{N} | E(i) \cdot G'(i) | \tag{9}$$

where $E(i)$ is the perimeter pixel i of the ellipse model located at r_0, with size s. $G'(i)$ is the edge detected corresponding to $E(i)$–th pixel, and N is the number of pixels on the perimeter of an ellipse with size s. Then, we calculate the skin region ratio which is shown as

$$T_2(r_0, s) = \frac{N_s}{N_e} \tag{10}$$

where N_e is the total number of pixels within an ellipse at location r_0, with size s and N_s is the total numbers of skin color pixels detected in N_e.

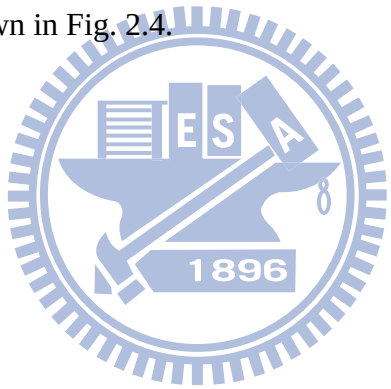
For the human face candidate, we define the following two conditions to be satisfied.

Condition 1. Human face candidate must have high enough edge points matching to the ellipse perimeter. It means that $T_1(r_0, s)$ must exceed a specified threshold value $Th_1 = 0.6$.

Condition 2. Human face candidate must also have high enough skin-color pixels within the region of ellipse to define. It means that $T_2(r_0, s)$ must exceed a specified threshold value $Th_2 = 0.7$.

A skin-color region is extracted as human face region if it satisfies the above two conditions.

We summarize the human face detection strategy by the case of an example sub-sampled image, as shown in Fig. 2.4.



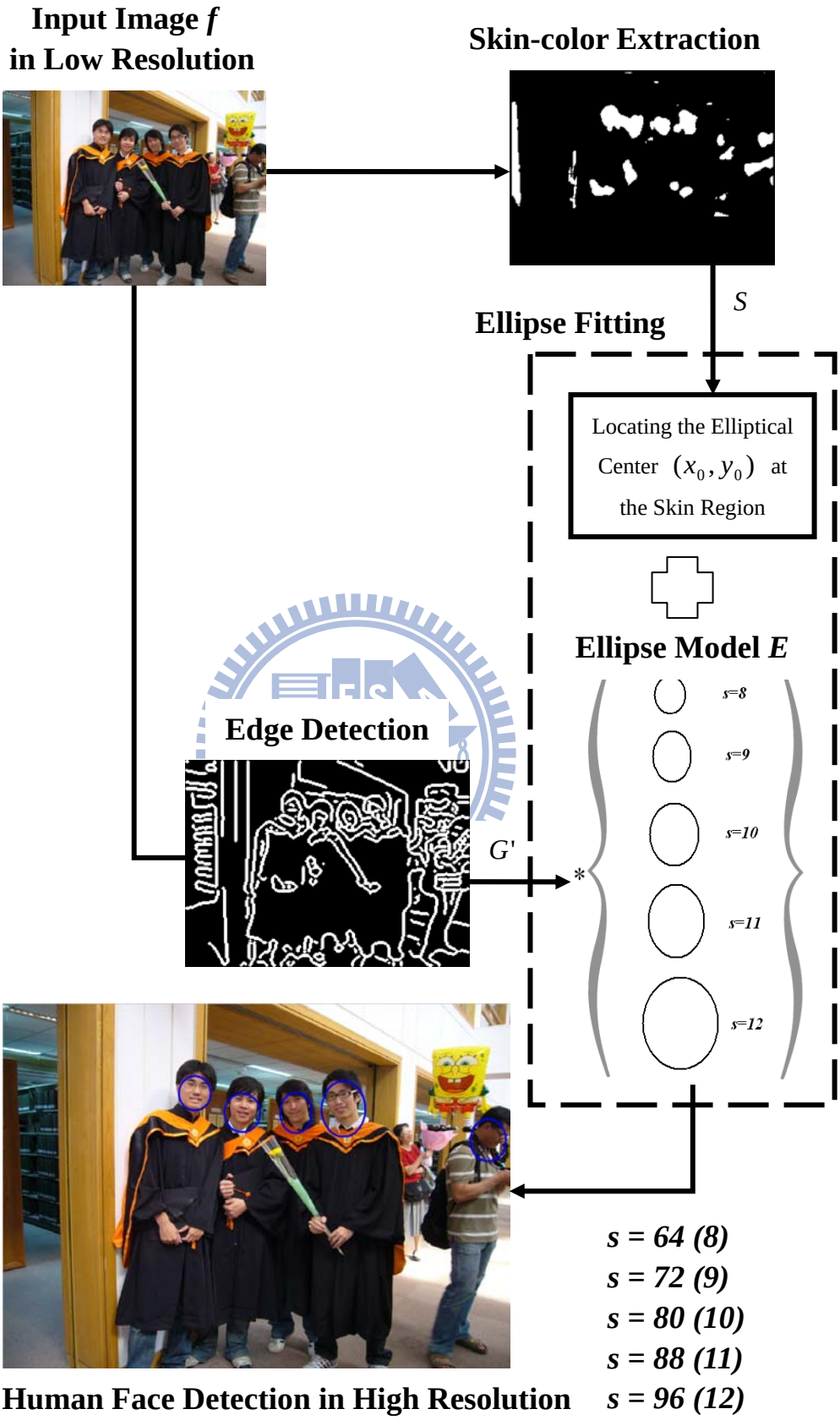


Fig. 2.4 Procedures of the human face detection.

Chapter 3 Automatic Seeded Region Growing

The first stage of human extraction system is image segmentation. We exploit an image segmentation technique for subject extraction. To date, there are many well-known segmentation methods developed. Among them, seeded region growing (SRG) is a typical method which integrates boundary edges and region growing. It generates a number of initial seeds by manual or automated selection, and then the remaining pixels are classified. The region growing procedure is not completed until all pixels are classified. Shih and Cheng [5] proposed an automatic seeded region growing algorithm in the YC_bC_r color space. The initial seeds are selected automatically if a seed candidate has high similarity to its neighbors. Because of the over-segmentation problem, it will conduct merging the regions with high similarity or the too small region after the initial region growing procedure. In our approach, we propose an improved automatic seeded region growing algorithm in the HSI color space which can simplify the complexity and reduce the computation burden.

3.1 HSI Color Space

RGB color space is convenient for display devices but is not good for image processing due to the high correlation. The HSI (hue, saturation, intensity) color space is compatible with the vision psychology of human eyes, and its three components are relatively independent. There are some variants of HSI systems, such as HSB (hue, saturation, brightness), HSL (hue, saturation, lightness), and HSV (hue, saturation, value) [14].

The HSI system decouples the intensity information from the color information.

Hue is a color attribute that describes a pure color, whereas saturation gives a measure of the degree to which a pure color is diluted by white light. Brightness is a subjective descriptor and embodies the achromatic notion of intensity [15]. In this thesis, we implement the segmentation method in HSI space because hue can be useful for separating objects with different colors.

Viewed from the circular side of the cone, the hue is represented by the angle of each color in the cone relative to the 0^o line assigned traditionally to be red, and increases counterclockwise from there. The saturation is represented as the distance from the vertical axis of the circle. Highly saturation colors are on the outer edge of the cone. The intensity is determined by the vertical axis in the cone. The vertical axis of the cone describes the gray levels, for instance, zero intensity is black, full intensity is white. Each slice of the cone perpendicular to the intensity axis is a plane with the same intensity [15]. Fig. 3.1 shows the HSI model based on color circles.

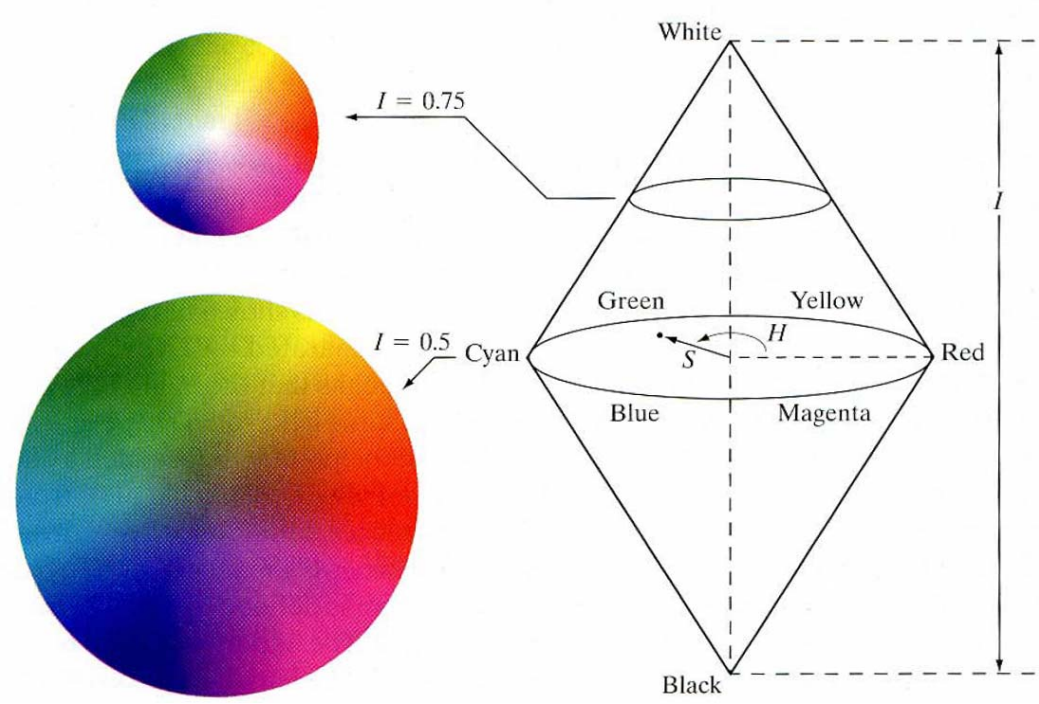


Fig. 3.1 The HSI color model based on circular color planes [15].

To convert RGB color space to HSI color space, the H component of each RGB pixel is obtained using the equation

$$H = \begin{cases} \theta, & \text{if } B \le G \\ 360 - \theta, & \text{if } B > G \end{cases} \tag{11}$$

with

$$\theta = \cos^{-1}\left\{ \frac{\frac{1}{2}\left[(R-G) + (R-B) \right]}{\left[(R-G)^2 + (R-B)(G-B) \right]^{1/2}} \right\}.$$

The saturation component is given by

$$S = 1 - \frac{3}{(R+G+B)}\left[\min(R, G, B) \right]. \tag{12}$$

Finally, the intensity component is given by

$$I = \frac{1}{3}(R+G+B), \tag{13}$$

It is assumed that the RGB values have been normalized to the range $[0,1]$ and that angle θ is measured with respect to the red axis of the HSI space, as indicated in Fig. 3.1. Hue can be normalized to the range $[0,1]$ by dividing by $360°$ all values resulting from Eq. (11). The other two HSI components already are in this range if the given RGB values are in the interval $[0,1]$.

3.2 Automatic Seeded Region Growing

The overview of an improved automatic seeded region growing algorithm is presented in Fig. 3.2. The initial seeds are chosen automatically if a pixel inside the edge region. After the initial seeds are generated, the remaining pixels are classified to the nearest region. The region growing procedure is completed if no pixel is unclassified. Finally, two neighboring regions with the similar hue and intensity values are merged. Some details are shown below.

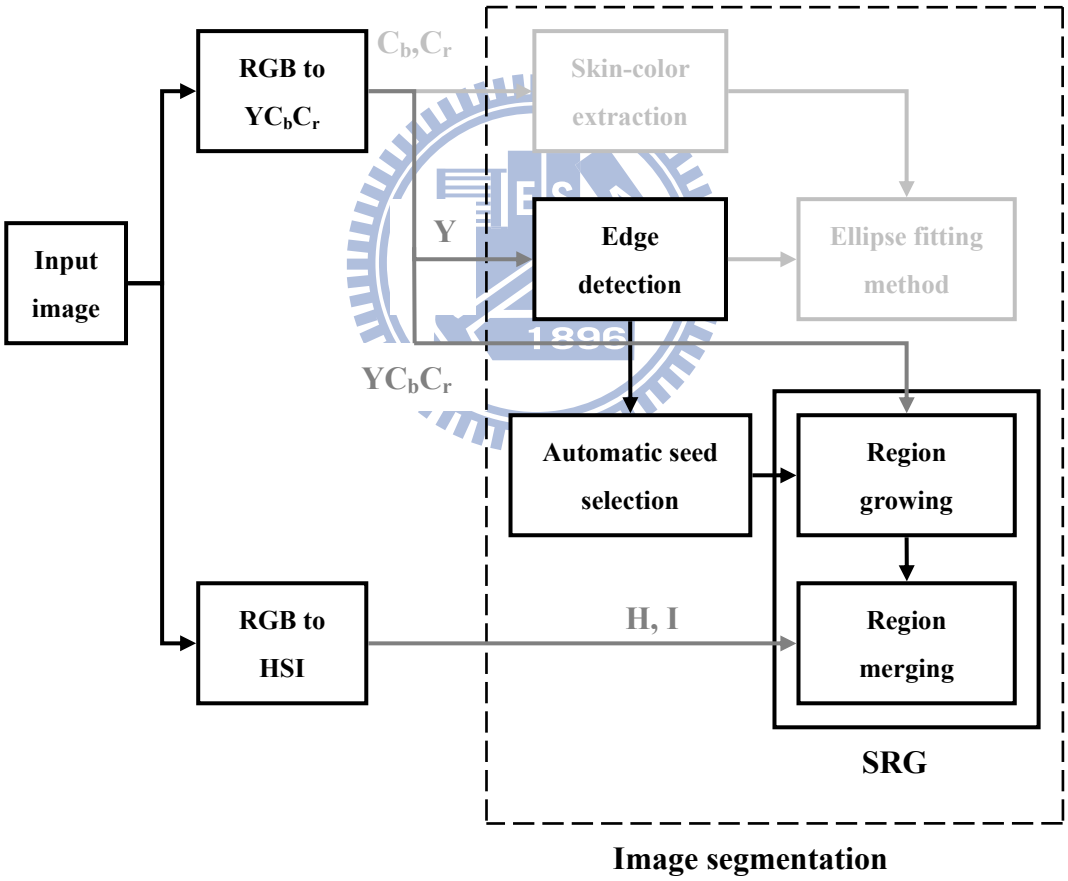


Fig. 3.2 Block diagram of the proposed algorithm.

3.2.1 Automatic Seed Selection

A good image segmentation technique could provide homogeneous image regions with accurate and closed boundaries. The edge detection is utilized frequently in dealing with regions and boundaries. Because edges are intensity discontinuities and local property, it is possible to find boundaries by linking edge segments or executing region-based segmentation.

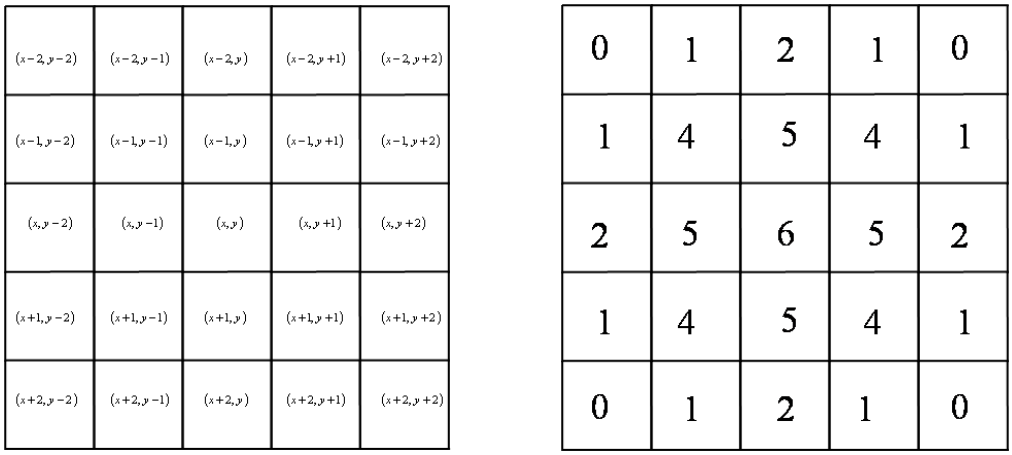
A pixel can become the seed candidate if it satisfies the following criteria. First, a pixel must have high similarity to its neighbors. Second, a pixel does not be located on the boundary of two regions. For these reasons, we utilize the result of the edge detection as the roughly confined boundaries for seed selection. Then the initial seed is chosen if a pixel within the edge region and should be farther from an edge point.

To this end, we propose a method to automate the initial seed selection procedure as follows. A 5×5 general mask is constructed and is shown in Fig. 3.3(a). It describes the relationship between the current pixel (x, y) and its neighboring pixels. The response R, an estimate of edge boundary closeness to the current pixel (x, y), of the edge result G with a predefined mask is given by the expression

$$R(x, y) = \sum_{u=-2}^{2}\sum_{v=-2}^{2} w(u, v)G(x+u, y+v) \tag{14}$$

where the w's are mask coefficients and are defined in Fig 3.3(b).

The process consists simply of moving the mask from point to point. After the mask process, we define the restriction to judge whether a pixel is chosen as the seed candidate. A pixel could be considered as an initial seed if $R(x, y) \le 3$. Some adjacent regions having the similar intensity but are different parts in real life may be detected no edge and cause discontiguous edges. The restriction described above can solve such problem and avoid regions over-integrated.



(a) (b)

Fig. 3.3 (a) A general 5×5 mask located at (x, y), (b) a predefined mask coefficients $w(u, v)$, $u = -2, \ldots, 2$, $v = -2, \ldots, 2$.

3.2.2 Seed Labeled

The initial seeds with 8-connectivity are grouped as one set and assigned the same label. We utilize the connectivity concept to find the connected seed regions.

Connectivity between pixels is a fundamental concept that simplifies the definition of numerous digital image concepts, such as regions and boundaries. To establish if two pixels are connected, it must be determined if they are neighbors and if their gray levels satisfy a specified criterion of similarity. For instance, in a binary image with values 0 and 1, two pixels may be 4-neighbors, but they are said to be connected only if they have the same value [15].

Let V be the set of gray-level values used to define adjacency. If the pixel is chosen as a seed, we set the pixel value as 1; otherwise the pixel is set as 0. Because we are referring to connectivity of pixels with value 1, the set is described $V = \{1\}$. There are three types of connectivity as shown below [15].

- 4-connectivity: Two pixels p and q with values from V are 4-connectivity if q is in the set $N_4(p)$.

- 8-connectivity: Two pixels p and q with values from V are 8-connectivity if q is in the set $N_8(p)$.
- m-connectivity: Two pixels p and q with values from V are m-connectivity if

  (i) q is in $N_4(p)$, or

  (ii) q is in $N_D(p)$ and the set $N_4(p) \cap N_4(q)$ has no pixels whose values are from V.

The $N_4(p)$, $N_8(p)$, and $N_D(p)$ mean 4-neighbors, 8-neighbors, and four diagonal neighbors of p. Fig. 3.4 shown a binary array and its three types of connectivity.

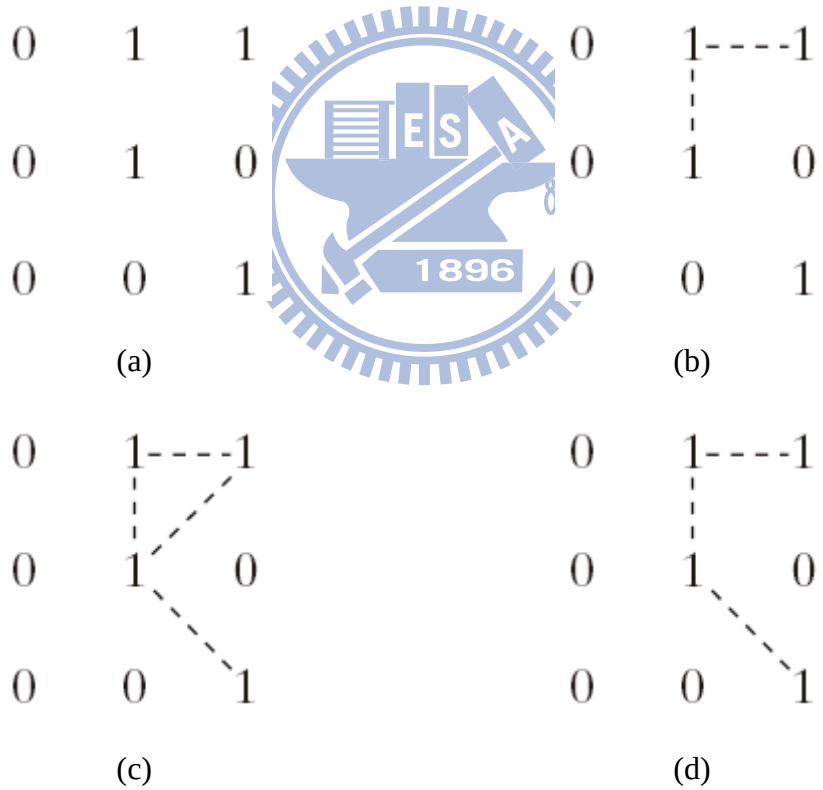


Fig. 3.4 (a) Arrange of pixels, (b) pixels that are 4-connectivity, (c) pixels that are 8-connectivity, (d) pixels that are m-connectivity [15].

3.2.3 Seeded Region Growing (SRG)

After the initial seed regions are generated and labeled, the regions start growing from the unclassified pixel located along region boundary with the minimum color distance to one of its neighbors. This procedure is repeated until all pixels are classified and labeled. The seeded region growing algorithm is described as follows:

- Step 1: Assign a label to each seed region after the initial seeds are selected automatically. Let $S_1, S_2, \ldots, S_n$ be the initial seeds which have been grouped into n sets and $A_1, A_2, \ldots, A_n$ be their corresponding regions.

- Step 2: T denotes the set of all unclassified pixels which are neighbors of at least one of the labeled region [3].

$$T = \left\{ (x, y) \notin \bigcup_{i=1}^{n} A_i \mid N(x, y) \cap \bigcap_{i=1}^{n} A_i \neq \phi \right\} \tag{15}$$

where $N(x, y)$ is the set of 4-neighbors, $(x-1, y), (x, y-1), (x+1, y), (x, y+1)$, of the pixel (x, y).

For $(x, y) \in T$, we have that $N(x, y)$ meets just one of the A_i and define $\gamma(x, y) \in \{1, 2, \ldots, n\}$ to be that index such that $N(x, y) \cap A_{\gamma(x, y)} \neq \phi$.

- Step 3: The relative Euclidean distance $d(x, y, A_i)$ between the pixel (x, y) and its adjacent labeled region is calculate as [5]

$$d(x, y, A_i) = \frac{\sqrt{(Y(x, y) - \overline{Y_i})^2 + (C_b(x, y) - \overline{C_{bi}})^2 + (C_r(x, y) - \overline{C_{ri}})^2}}{\sqrt{Y^2(x, y) + C_b{}^2(x, y) + C_r{}^2(x, y)}} \tag{16}$$

where $i \in \gamma(x, y)$, and $(\overline{Y_i}, \overline{C_{b\,i}}, \overline{C_{r\,i}})$ are the mean values of Y_i, $C_{b\,i}$, and $C_{r\,i}$ components in that region A_i.

Then a sorted list T records neighbors of all regions which satisfy Eq. (15) in a decreasing order of distances obtained from Eq. (16).

- Step 4: While T is not empty, remove the first point (x_1, y_1) with the minimum distance value and check its 4-neighbors. If all labeled neighbors of (x_1, y_1) have the same label, set (x_1, y_1) to this label. If the labeled neighbors of (x_1, y_1) have two or more labels, calculate the distances between (x_1, y_1) and all labeled neighbors and classify (x_1, y_1) to the nearest region.

$$d(\gamma(x, y)) = \min_{(x, y) \in T} \{ d(x, y, A_i) \mid i \in \gamma(x, y) \} \tag{17}$$

Then update the mean of this region, and add 4-neighbors of (x_1, y_1), which are neither classified yet nor in T, to T in a decreasing order of distances.

Note that previous entries in the T are not updated to reflect their differences from the new region mean. This leads to negligible difference in the results, but greatly enhance speed [3].

3.2.4 Region Merging

It is possible that a region is split into several small ones. To avoid the over-segmented problem, we perform region merging procedure. In most cases, the intensity can distinguish roughly from different objects. In the following, we merge some regions in the HSI color space due to above reasons. Hue (H) and intensity (I)

are the most important ones that we considered. The difference between two adjacent region R_i and R_j is defined as

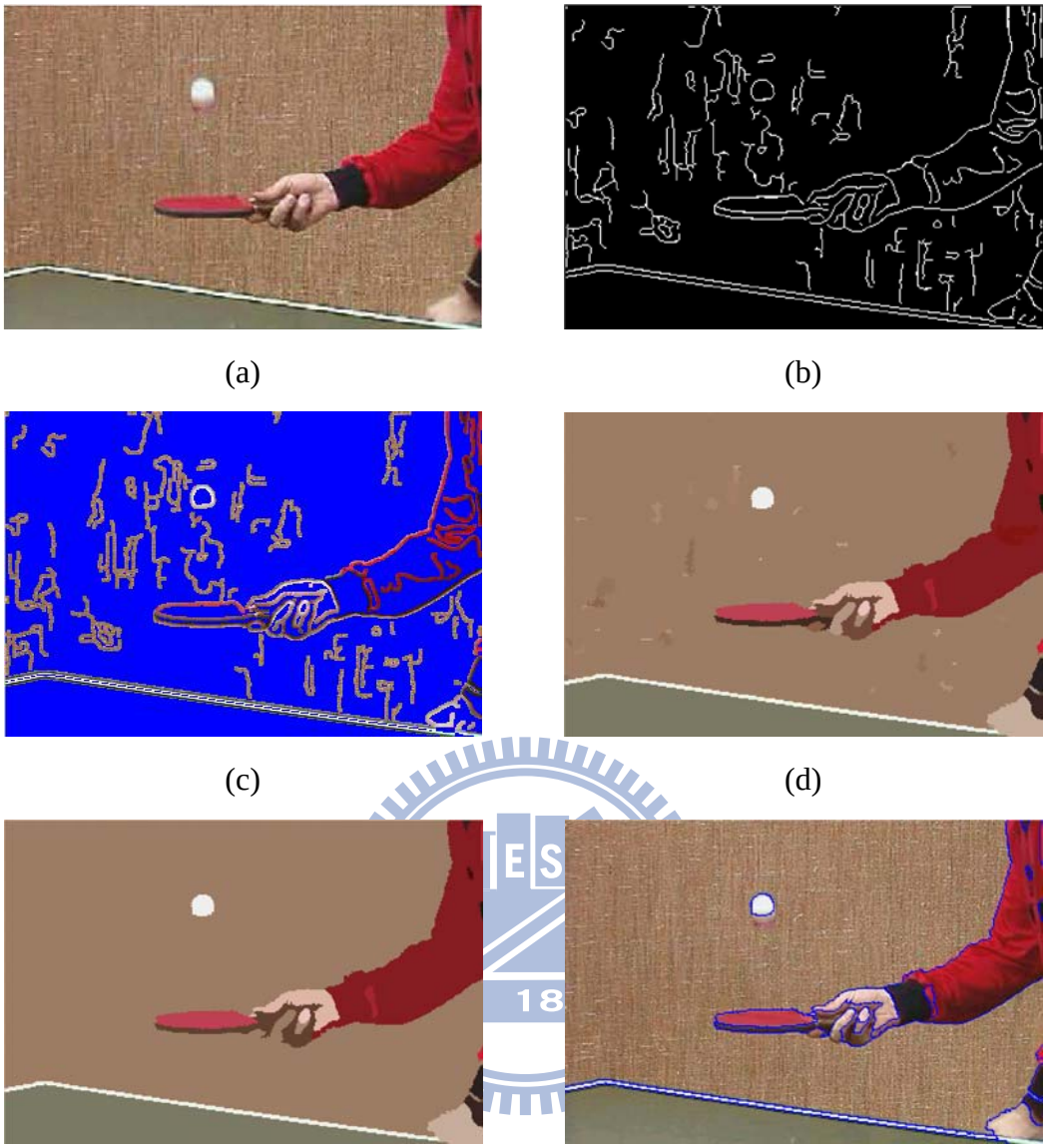
$$d(R_i, R_j) = \frac{\sqrt{(\overline{H_i} - \overline{H_j})^2 + (\overline{I_i} - \overline{I_j})^2}}{\min(\sqrt{H_i^{\,2} + I_i^{\,2}}, \sqrt{H_j^{\,2} + I_j^{\,2}})}, \tag{18}$$

where *H* and *I* components are in this range $[0,1]$.

If any two adjacent regions have distance $d(R_i, R_j)$ less than a threshold value, the merge procedure would be repeated.

Fig. 3.5(a) gives a color image, and Fig. 3.5(b) shows the extracted edges by performing Canny edge detector [6]. Fig. 3.5(c) shows the detected seeds marked in blue color. Note that the connected seed pixels are considered as one seed S_i, and A_i is the region corresponding to S_i. Fig. 3.5(d) shows the seeded region growing result. Fig. 3.5(e) shows the result of merging adjacent regions with similar hue and intensity values. The final segmented result is given in Fig. 3.5(f).

In Fig. 3.6, we show the results of different segmentation algorithm. Fig. 3.6(a) gives a color image. Fig. 3.6(b) shows the result of JSEG algorithm proposed by Deng *et al*. [16] and Fig. 3.6(c) shows the result of segmentation algorithm proposed by Shih *et al*. [5]. The result of our proposed segmentation algorithm is given in Fig. 3.6(d).

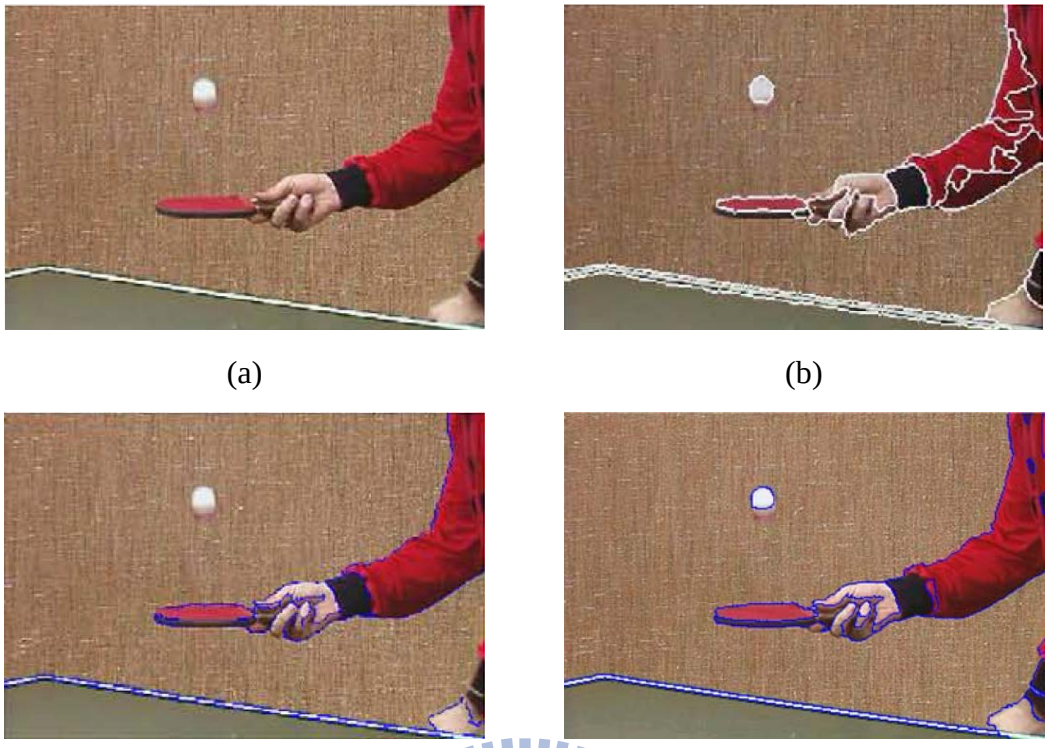


(a) (b)

(c) (d)

(e) (f)

Fig. 3.5 (a) Original color image, (b) edges obtained by Canny edge detector, (c) the initial seeds found in blue color, (d) seeded region growing result, (e) the merging result, and (f) final segmented result.



(a) (b)

(c) (d)

Fig. 3.6 (a) Original color image, (b) the result of JSEG algorithm [16], (c) the result of Shih [5], and (d) the result of our algorithm.

Chapter 4 Experimental Results

There are two parts in this chapter. The first part deals with human extraction. The second part is the depth estimation. In our experiment, we test our system on images with some static subjects ahead the camera. Human with different depths are extracted firstly by combining the face detection method and semantic human body model, then the depths are estimated.

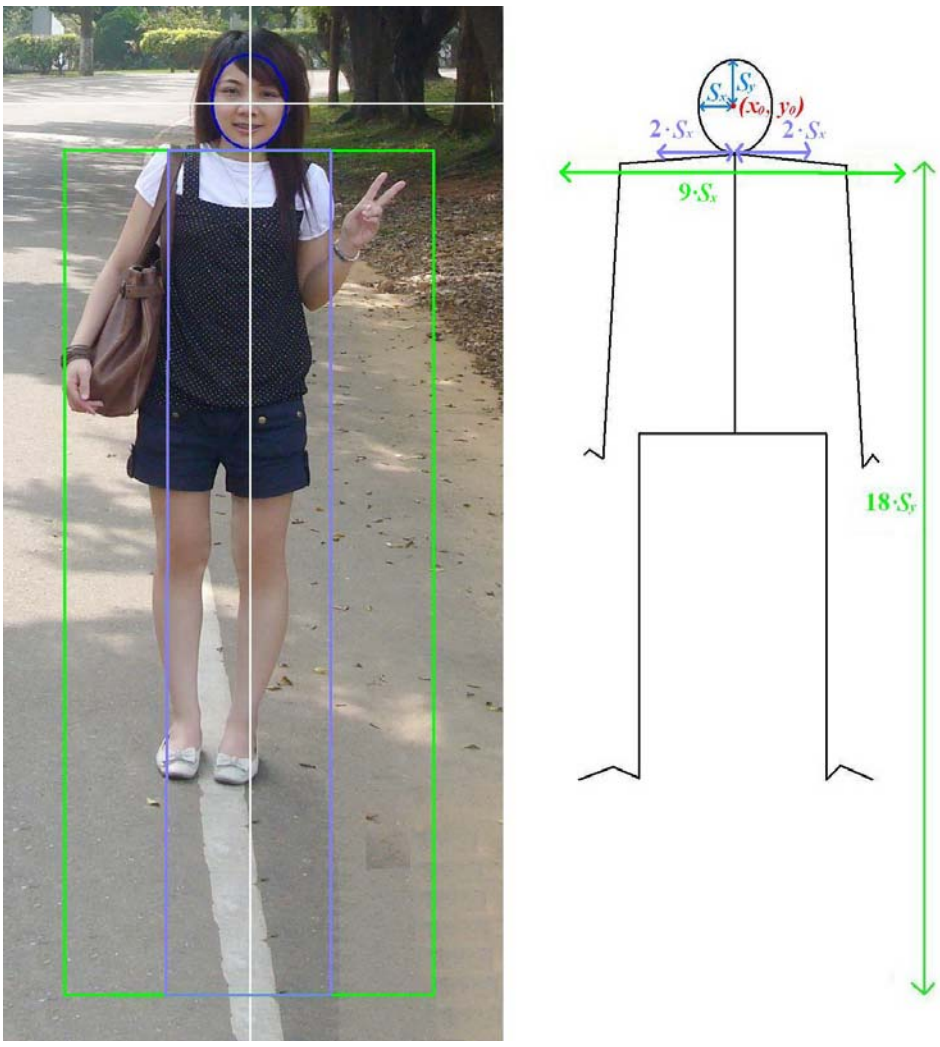
4.1 Human Extraction

In this section, two sets of experimental results show the effectiveness of our proposed algorithm on detecting the homogeneous regions, containing human faces and semantic human bodies. Without the use of background model, we extract the foreground subjects we interested by searching for some features related to human and then locate the subjects.

The human body is determined by analyzing semantic human bodies. We utilize a skin-color map and the elliptical shape to detect the human face and to locate the human position. Then the human body is extracted. In Fig 4.1(a), the white line represents the human face location estimated, the green rectangle confines the body ranges to be searched, and the purple rectangle confines the ranges of body region candidates.

A simplified graph of human body ratio is shown in Fig 4.1(b), and the rough body ranges are confined as below:

- The ellipse's minor radius S_x of detected face and the width of human body are defined to be of ratio 1:9, experimentally.
- The ellipse's major radius S_y of detected face and the height of human body are defined to be of ratio 1:18.



(a) (b)

Fig. 4.1 (a) Human face, body ranges, and the ranges of body region candidates are represented as a blue ellipse, a green and purple rectangle, and (b) the human body ratio.

Firstly, our proposed seed region growing method is utilized to segment the image. To avoid the over-segmented problem, we merge the regions with high similarity of hue and intensity values. It is to be noted that two similar regions are merged only when they are both inside, or both outside the defined body ranges. When the merge procedure is completed, all pixels in the same segmented region are labeled with the same color.

Not all regions within the defined body ranges are the body regions. If the *x*-coordinate value of a pixel is within the range $[x_0 - 2 \cdot S_x, x_0 + 2 \cdot S_x]$ and its *y*-coordinate value is within the range $[y_0 + S_y, y_0 + 19 \cdot S_y]$, where (x_0, y_0) is the face ellipse's center, the pixel's label is marked as the body region candidate. For all body region candidates, if the smallest and the largest *x*-coordinate values of the corresponding region are within the defined body ranges, the region could be considered as a part of the body regions.

These body regions are then merged to form a human body. Finally, the human is extracted by combining the detected human face and human body. Defining an explicit range is necessary to well segment the foreground subjects we interested from the background.

Fig. 4.2 shows an example of human body extraction. Fig. 4.2(a) is an image "Salesman." Fig. 4.2(b) shows the extracted edges by performing Canny edge detector. Figs. 4.2(c) and 4.2(d) show the skin-color extraction and ellipse fitting. The human location is determined from Figs. 4.2(c) and 4.2(d). Figs. 4.2(e)–(g) are the seeded region growing result. The extracted human is shown in Fig. 4.2(h). Similar results for "Akiyo" are also given in Fig. 4.3.

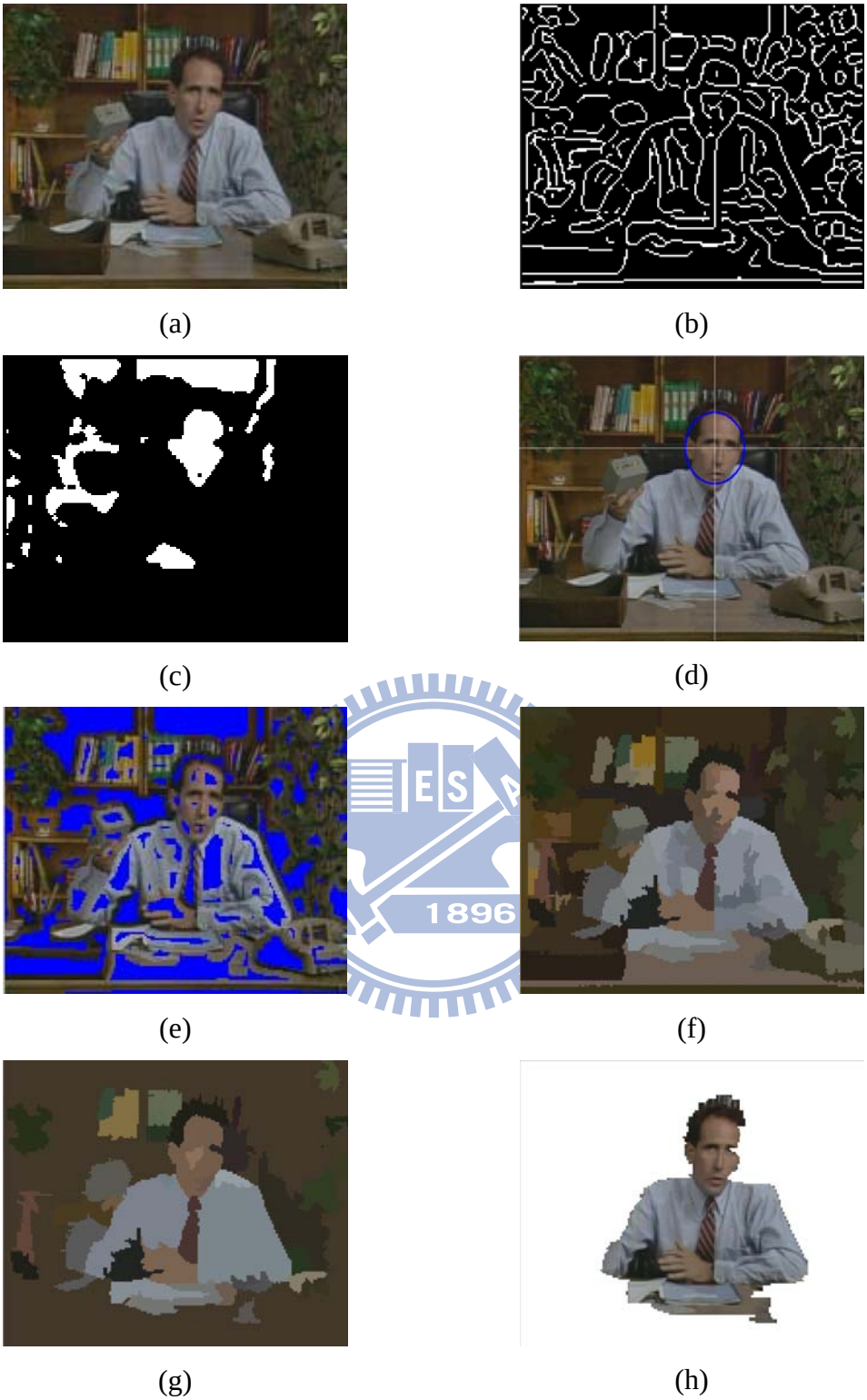


Fig. 4.2 An example of foreground region extraction. (a) An image frame, (b) edges obtained by Canny edge detector, (c) skin-color extraction, (d) ellipse fitting, (e) the initial seeds found in blue color, (f) seeded region growing result, (g) the merging result, and (h) the extracted human.

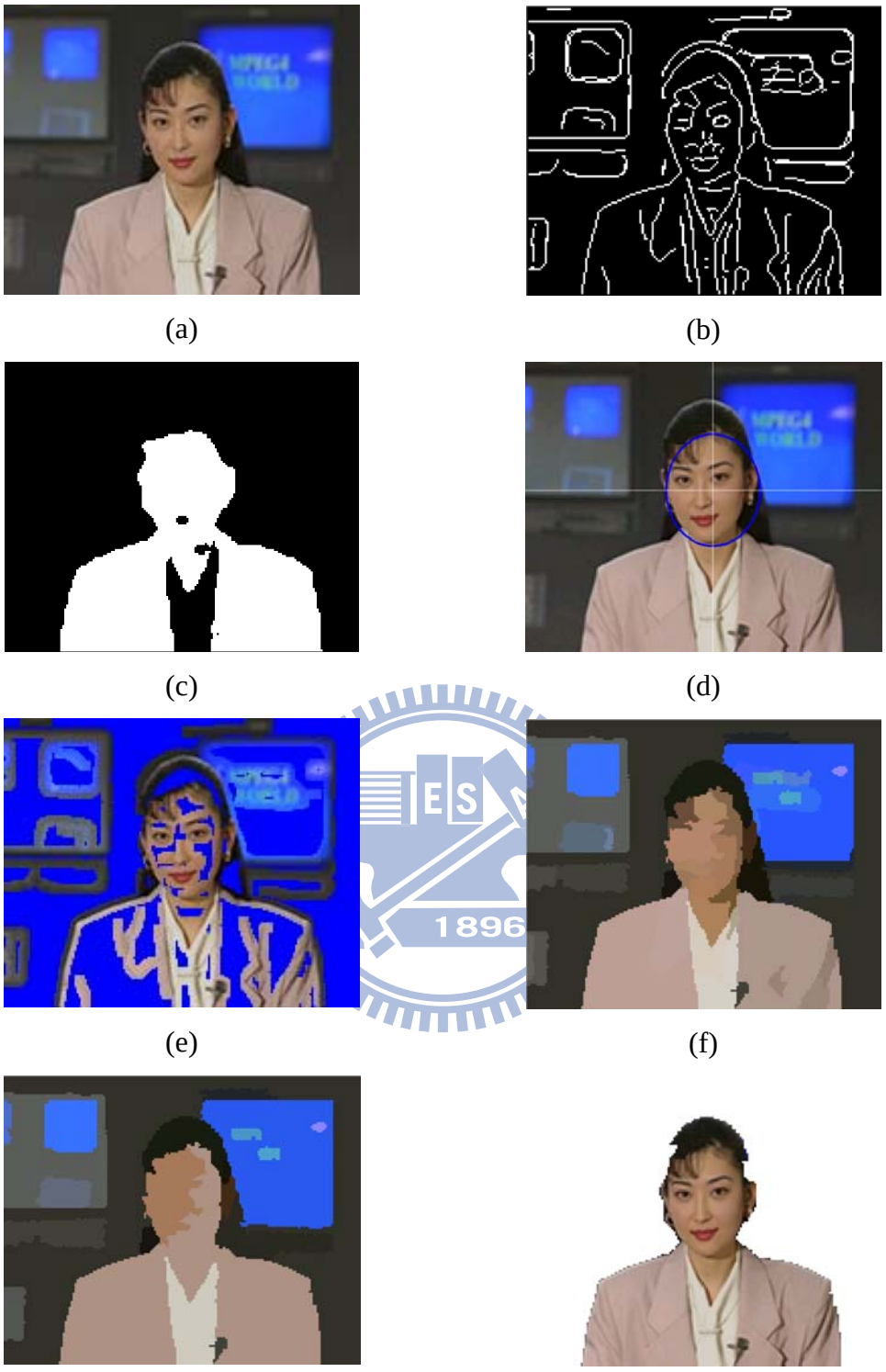


Fig. 4.3 An example of foreground region extraction. (a) An image frame, (b) edges obtained by Canny edge detector, (c) skin-color extraction, (d) ellipse fitting, (e) the initial seeds found in blue color, (f) seeded region growing result, (g) the merging result, and (h) the extracted human.

4.2 Depth Estimation Based on Vanishing Line and Point

4.2.1 Geometry

In this section, we describe how the depths may be estimated from a single perspective view. Firstly, the vanishing lines and points are detected. The vanishing line of the reference plane (the ground) is the projection of the line at infinity of the reference plane into the image. The vanishing point is the image of the point at infinity [8].

All world lines parallel to each other are imaged as lines which intersect in the same vanishing point. Therefore two or more such lines are sufficient to define the vanishing point. Measuring heights could be seen as measuring the vertical distance from a reference plane in the world. Any scene point which projects onto the vanishing line is at the same distance from the plane as the camera center; if it lies "above" the line it is farther from the plane, and if "below" the vanishing line, then it is closer to the plane than the camera centre [8]. Consequentially, measuring depths could be seen as measuring the horizontal distance from a subject position to the camera position projects onto the reference plane in the world. Any scene point position lies "on" or lies "above" the vanishing line is at the farthest distance from the camera position, and if "below" the vanishing line, it is closer to the camera position. Then, the depth estimation is based on the scene point position with the detected vanishing line location. The relation between depths and the vanishing line is shown in Fig. 4.4.

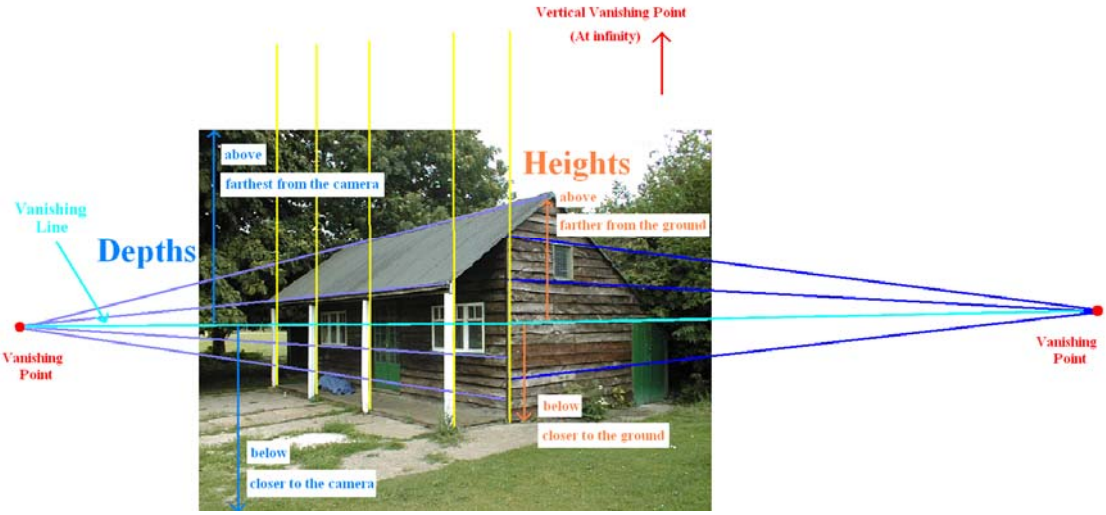


Fig. 4.4 The relation between depths and the vanishing line.

4.2.2 The Vanishing Lines and Points Detection

Because all world lines parallel to each other are imaged as lines which intersect in the same vanishing point, these parallel lines are detected and their intersection could be seen as the vanishing point. Firstly, the edges are detected by Sobel edge detector. Not all lines are useful for the vanishing point detection, so we find several the most representative lines in the image using the Hough transform. Therefore these lines are sufficient for the vanishing point detection.

4.2.3 Cross-Ratio

Cross-ratio is utilized in projective geometry because of the projective invariant in the sense. Let $\{L_i,\ i=1,\ldots,4\}$ be four distinct lines in the plane passing through the same point O. Then any line L not passing through O intersects these lines in four distinct points P_i, is shown in Fig. 4.5. It turns out that the cross-ratio of these points does not depend on the choice of a line L, and hence it is an invariant of the 4-tuple of lines $\{L_i\}$.

The cross-ratio is given by the formula [8]

$$CR(P_1, P_2, P_3, P_4) = \frac{\overline{P_1P_3}}{\overline{P_2P_3}} \frac{\overline{P_2P_4}}{\overline{P_1P_4}} \tag{19}$$

If four points p_1, p_2, p_3, p_4 indicate image quantities, and P_1, P_2, P_3, P_4 indicate quantities in the world. Then, we can write

$$\frac{\overline{p_1p_3}}{\overline{p_2p_3}} \frac{\overline{p_2p_4}}{\overline{p_1p_4}} = \frac{\overline{P_1P_3}}{\overline{P_2P_3}} \frac{\overline{P_2P_4}}{\overline{P_1P_4}} \tag{20}$$

Since p_4 is the vanishing point in the image, it means that P_4 is at infinity and $\frac{\overline{P_2P_4}}{\overline{P_1P_4}} = 1$. The right hand side of Eq. (20) reduces to $\frac{\overline{P_1P_3}}{\overline{P_2P_3}} = \frac{\overline{P_1P_3}}{\overline{P_1P_3} - \overline{P_1P_2}}$. Simple algebraic manipulation on Eq. (20) yields

$$\frac{\overline{P_1P_2}}{\overline{P_1P_3}} = 1 - \frac{\overline{p_2p_3}}{\overline{p_1p_3}} \frac{\overline{p_1p_4}}{\overline{p_2p_4}} \tag{21}$$

An example is shown in Fig. 4.6. Some depths in the world and their corresponding *y*-coordinate values in the image are marked as blue and green words and shown in Table I. The two world lines parallel to each other are found and its intersect point in the image is described as vanishing point p_4, with *y*-coordinate value $p_4 = 160$. Let P_1 = 5M, P_2 = 10M in the world, and their corresponding *y*-coordinate values in the image are $p_1 = 679$, $p_2 = 421$. If we let $p_3 = 339$, its corresponding depth in the world could be estimated as $P_3 = 14.6$M from Eq. (21). Similar estimated results are shown in Table II. It is shown that the unknown depths (P_3) in the world could be estimated if any two image points' *y*-coordinate value (p_1 and p_2) and their corresponding depths in the world (P_1 and P_2) are pre-measured and

the vanishing lines and points (p_4) are pre-detected, which are satisfied Eq. (21).

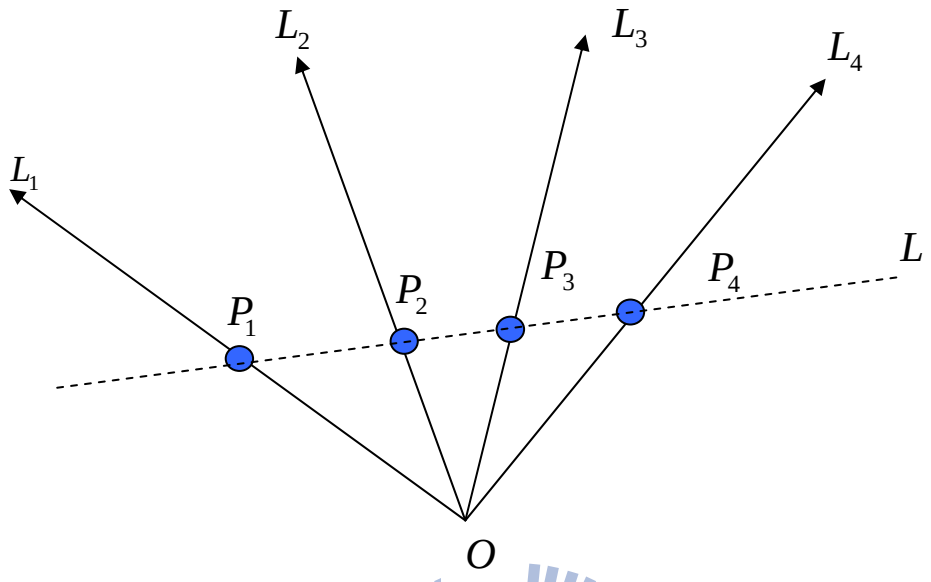


Fig. 4.5 The Cross-Ratio relation.

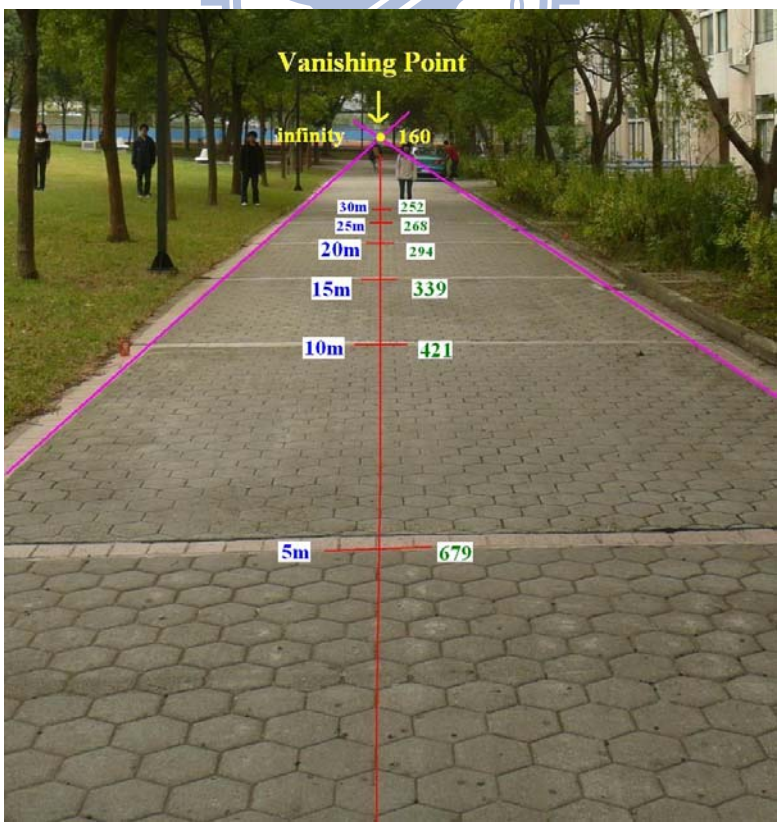


Fig. 4.6 Measuring depths in the world and in the image.

TABLE Ⅰ

THE DEPTHS IN THE WORLD AND THEIR CORRESPONDING *Y*-COORDINATE VALUE IN THE IMAGE

The *y*-coordinate value in the image (p_i)	Distance from camera in the world (P_i)
160	infinity
252	30M
268	25M
294	20M
339	15M
421	10M
679	5M

TABLE II

THE ESTIMATED DEPTHS IN THE WORLD AND THEIR CORRESPONDING *Y*-COORDINATE VALUE IN THE IMAGE

$(p_1, P_1) = (679, 5\text{M});\ (p_2, P_2) = (421, 10\text{M});\ p_4 = 160$	
The *y*-coordinate value in the image (p_3)	Estimated depths in the world (P_3)
252	28.5M
268	24.2M
294	19.5M
339	14.6M

4.2.4 Depth Planes Construction

There are several rules to represent the depth by a figure, as shown in Fig. 4.7, shown below [17]:

1. Higher depth level corresponds to lower gray values, often called depth gradient planes.
2. The vanishing point is the most distant point from the observer / camera (This assumption is almost true).

Because objects (buildings, cars, etc.) front the camera with various angles, the depths are generated corresponding to the direction and distance between the objects and the camera, and are general represented in Fig 4.7(a)–(d). In our experiment, one object could be seen with one depth and the ground depth is defined as the depth of the camera to the object interested, shown in Fig 4.7(d).

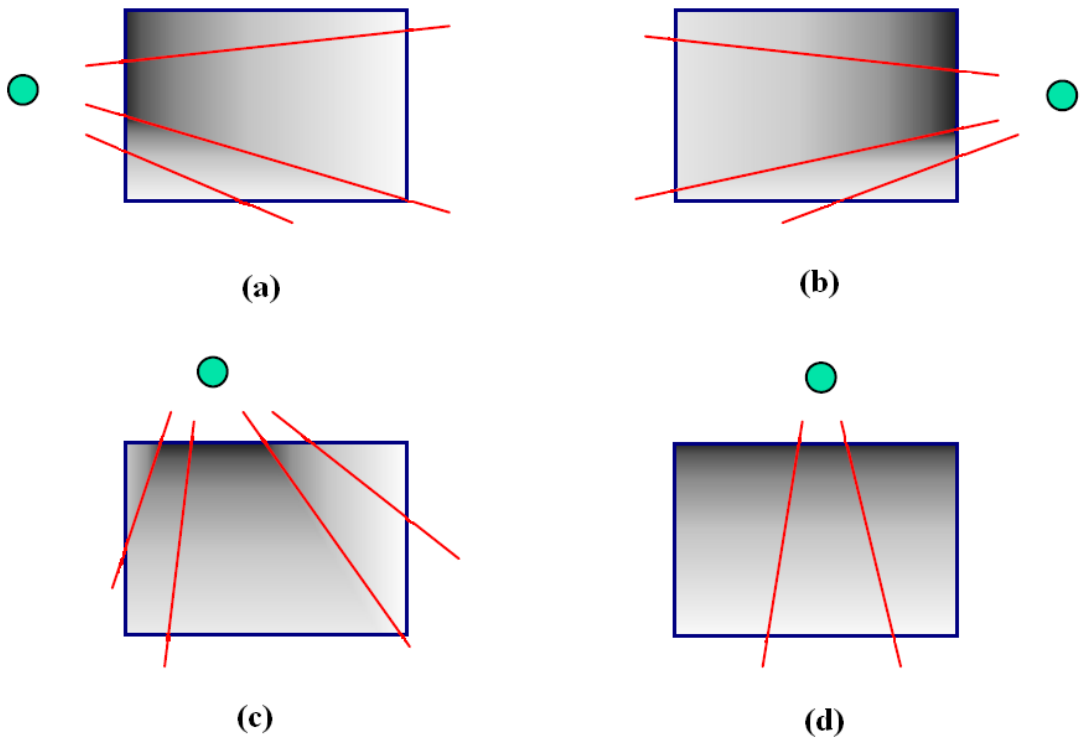


Fig. 4.7 Examples of the heuristic rules to generate depth gradient planes: the green circle represents the vanishing point [17].

Object locations projecting the same value onto Y direction in the image do not mean having the same depth in the world. It is not apparent if the distance is distant from the camera. Fig. 4.8 shows that the same depths in the ground are marked as a curve in the image. When the depth is distant, these points which have the same depth are connected almost as a straight line. To simplify the problem, we can ignore the erroneousness and consider that the same depth is represented as a straight line, shown in Fig. 4.7(d).

Fig. 4.8 The same depths in the ground are shown as a curve in the image.

4.3 Depth Estimation Based on Fixed Camera Parameters

There are several depth estimation methods which have been proposed in the past few years. One of the methods is depth estimation based on vanishing line and points as described in Sec. 4.2, which is better applicable for outdoor scene. Sometimes the vanishing line and points are not apparent enough to detect. One way

to simplify the depth estimation is based on fixed camera parameters. If the camera parameters are fixed and the camera does not pan / tilt, we can take some pictures and construct a depth look-up table of the camera containing pixel locations and their corresponding depth values. For example, Tables III and IV describe the relationship between the marked vertical pixel locations of an image and the corresponding depth distances in the real world of Panasonic TZ-2.

TABLE III

THE DEPTH LOOK-UP TABLE CORRESPONDING TO THE VERTICAL PIXEL LOCATIONS OF IMAGE WITH SIZE 1024×768

Distance from camera in the world	The Y axis value in the horizontal image
3M	715
4M	635
5M	575
6M	535
7M	512
8M	495
9M	484
10M	474
12M	456
15M	444
20M	435
30M	428
≥ 40M	424

TABLE IV

THE DEPTH LOOK-UP TABLE CORRESPONDING TO THE VERTICAL PIXEL LOCATIONS OF IMAGE WITH SIZE 768×1024

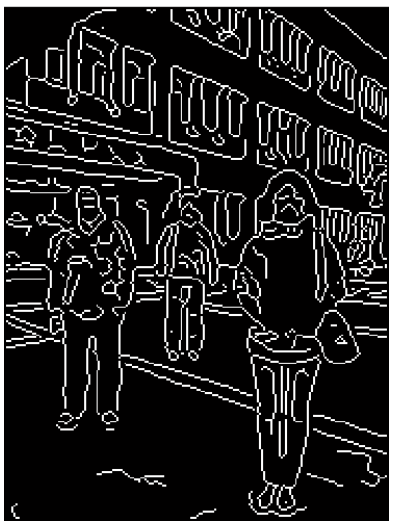
Distance from camera in the world	The Y axis value in the vertical image
2M	1000
3M	820
4M	736
5M	676
6M	636
7M	612
8M	596
9M	584
10M	572
12M	556
15M	544
20M	535
30M	528
≥ 40M	524

4.4 Depth Estimation Results

We have performed and reported two sets of experimental results to show the effectiveness of our proposed algorithms on detecting the human objects and estimating the depths. Using Panasonic TZ-2, we will illustrate the performance of all the techniques presented in Chapters 2, 3, and 4. Two scene images are analyzed below.

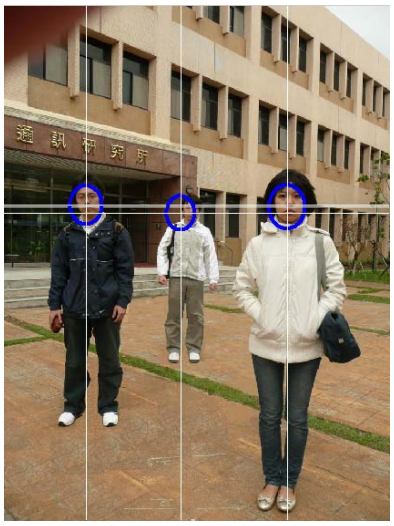
Fig. 4.9 shows an example of human extraction and depth estimation. Fig. 4.9(a) is a color image. Fig. 4.9(b) shows the extracted edges by performing Canny edge detector. Figs. 4.9(c) and 4.9(d) show the skin-color extraction and ellipse fitting, and human location is determined. Figs. 4.9(e)–(g) are the seeded region growing result. The extracted human is shown in Fig. 4.9(h). Fig. 4.9(i) is the obtained vanishing line by utilizing the Hough transform method. Figs. 4.9(j) and 4.9(k) show the ground extraction and the depth map. The extracted human depth map and the ground are shown in Fig. 4.9(l).

(a)



(b)

(c)



(d)

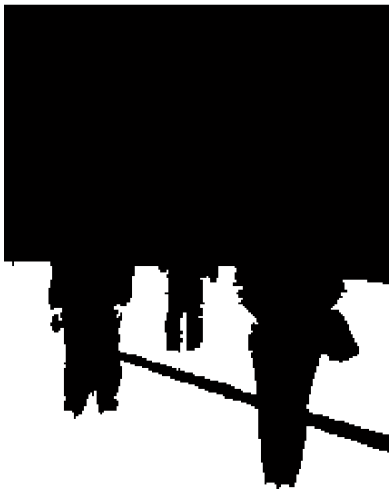
(e)

(f)

(g)

(h)

(i)



(j)

(k)

(l)

Fig. 4.9 An example of foreground region extraction and depth estimation. (a) An image frame, (b) edges obtained by Canny edge detector, (c) skin-color extraction, (d) ellipse fitting, (e) the initial seeds found in blue color, (f) seeded region growing result, (g) the merging result, (h) the extracted human, (i) the vanishing line is shown in blue line, (j) the ground extraction, (k) the ground depth map, and (l) the ground and extracted human depth map.

According to Eq. (21), let the distances in the world P_1 = 2M, P_2 = 2.5M are pre-measured, and their corresponding vertical *y*-coordinate values in the image are $p_1 = 1000,\ \ p_2 = 897.$ The vanishing line in the image is detected, with vertical *y*-coordinate value $p_4 = 485.$ If we have $p_3 = 964,$ which represents the detected vertical *y*-coordinate position of person 1, its corresponding depth in the world could be estimated as $P_3 = 2.15\text{M}$ from Eq. (21). Moreover, if we have $p_3 = 824$ and $p_3 = 728,$ which represent the detected vertical *y*-coordinate positions of person 2 and person 3, their corresponding depth in the world could be estimated as $P_3 = 3.03\text{M}$ and 4.24M, respectively, as shown in Table V. Table VI shows the accuracy rate of the depth estimated.

TABLE V

THE DEPTH ESTIMATION BASED ON CROSS-RATIO FORMULA

$(p_1, P_1) = (1000, 2\text{M});\ (p_2, P_2) = (897, 2.5\text{M});\ p_4 = 485$		
Detected vertical *y*-coordinate value of human position in the image	The estimated depths in the world	The actual depths in the world
p_3 = 964 (person 1)	2.15 M	2M
p_3 = 824 (person 2)	3.03 M	3M
p_3 = 728 (person 3)	4.24 M	4.5M

TABLE VI

THE ACCURACY RATE OF THE DEPTH ESTIMATION

	Person 1	Person 2	Person 3	Average
Depth estimation accuracy (%)	92.50%	99.00%	94.22%	95.24%

On the other hand, the human depths can be estimated using the look-up table according to the vertical *y*-coordinate values as shown in Tables VII, and the accuracy rate of the depth estimated is shown in Tables VIII.

TABLE VII

THE DEPTH ESTIMATION BASED ON THE LOOK-UP TABLE

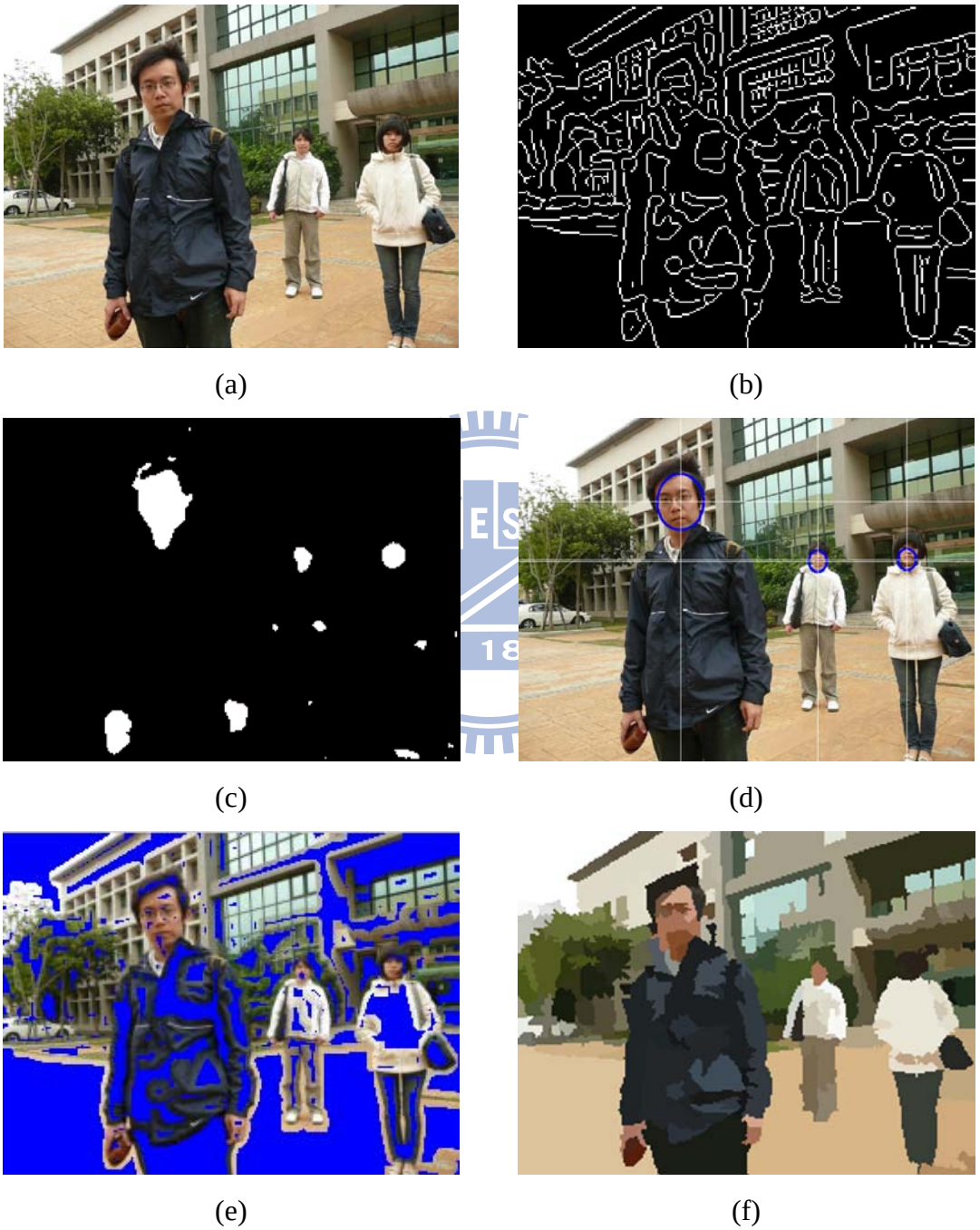
Detected vertical *y*-coordinate value of human position in the image	The estimated depths in the world	The actual depths in the world
964 (person 1)	2.20M	2M
824 (person 2)	2.98M	3M
728 (person 3)	4.13M	4.5M

TABLE VIII

THE ACCURACY RATE OF THE DEPTH ESTIMATION

	Person 1	Person 2	Person 3	Average
Depth estimation accuracy (%)	90.00%	99.33%	91.78%	93.70%

Similar results for another image are shown in Fig. 4.10. The depth estimation based on look-up table and the accuracy rate of the depth estimated are shown in Tables IX and X, respectively.



(a) (b) (c) (d) (e) (f)

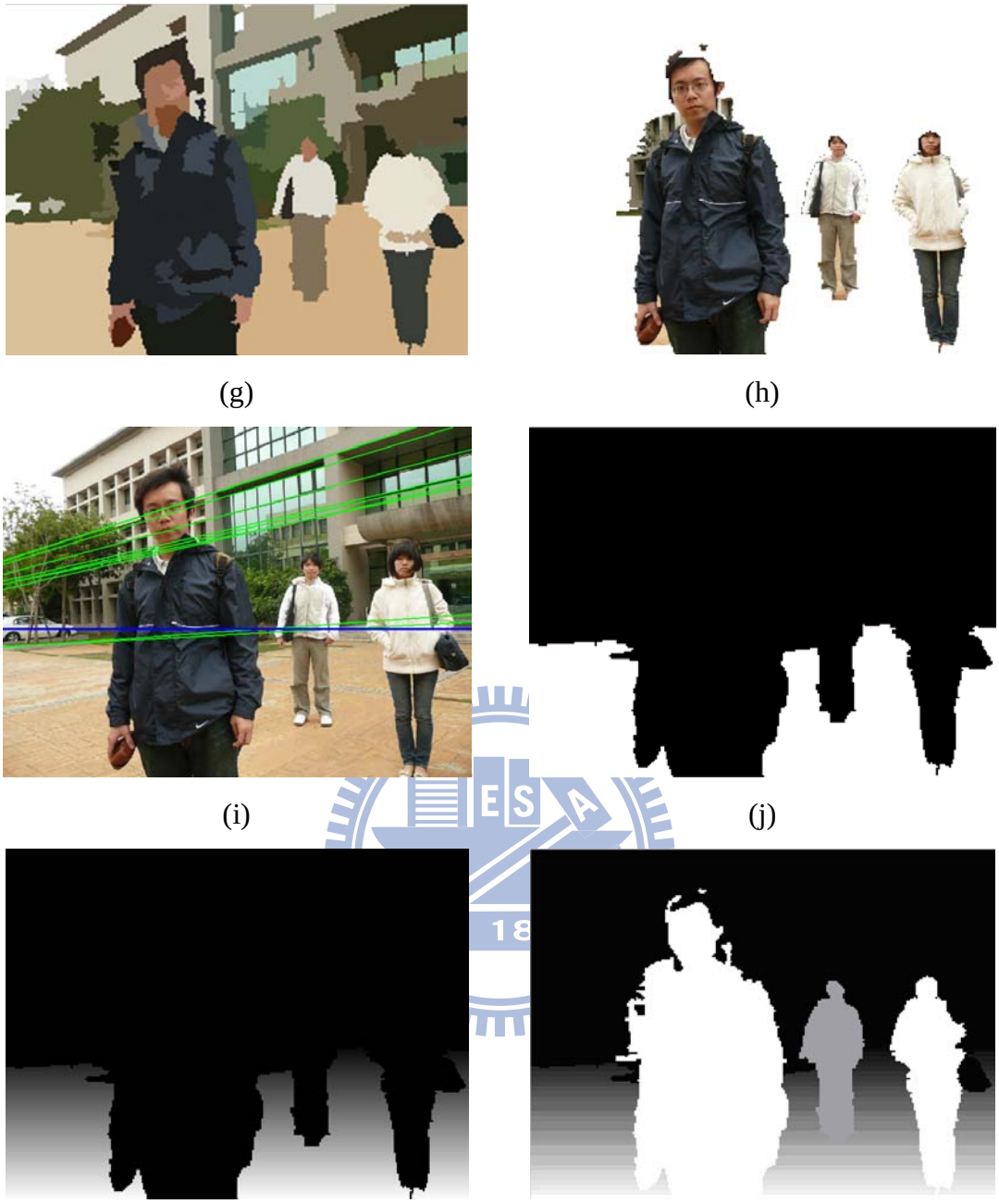


Fig. 4.10 An example of foreground region extraction and depth estimation. (a) An image frame, (b) edges obtained by Canny edge detector, (c)skin-color extraction, (d)ellipse fitting, (e) the initial seeds found in blue color, (f) seeded region growing result, (g) the merging result, (h) the extracted human, (i) the vanishing line is shown in blue line, (j) the ground extraction, (k) the ground depth map, and (l) the ground and extracted human depth map.

TABLE IX

THE DEPTH ESTIMATION BASED ON THE LOOK-UP TABLE

Detected vertical *y*-coordinate value of human position in the image	The estimated depths in the world	The actual depths in the world
768 (person 1)	≤ 3M	1.5M
764 (person 2)	≤ 3M	3M
648 (person 3)	3.84M	4M

TABLE X

THE ACCURACY RATE OF THE DEPTH ESTIMATION

	Person 1	Person 2	Person 3	Average
Depth estimation accuracy (%)	—	100%	96.00%	98%

4.5 Face Tracking and Depth Estimation by PTZ Camera

In this section, we use the Pan-Tile-Zoom (PTZ) camera Sony EVI-D100/P and present diverse experimental results to track human face and estimate the depth. Firstly, we utilize a skin-color map in the YC_bC_r color space to detect the face region. Some skin regions such as hand and leg would be rejected in the extracted image. Then we calculate the center and the size of the detected face region, in which facial size is used to compute the depth from the camera to the human. The camera panned and tilted according to the *x*-direction and *y*-direction distances from the image center, and zoomed if the occupying ratio of the human face in the image is too large or too small.

In our system, the camera operates at three zoomed modes, tracks the human

face and estimates the depth automatically. The flowchart of the Sony EVI-D100/P camera tracking and zooming control is shown in Fig. 4.11. The relationship between the occupying ratio of the human face in the image at three zoomed modes and the person's depth from the camera is constructed as the three modes' look-up tables, and the unknown depth is estimated from the tables of facial size, as shown in Table XI. Two examples of face tracking are shown in Figs. 4.12 and 4.13, respectively, with the distances between the human and the camera are respectively 3.3M and 5.6M, and their corresponding depth estimations are shown in Table XII.

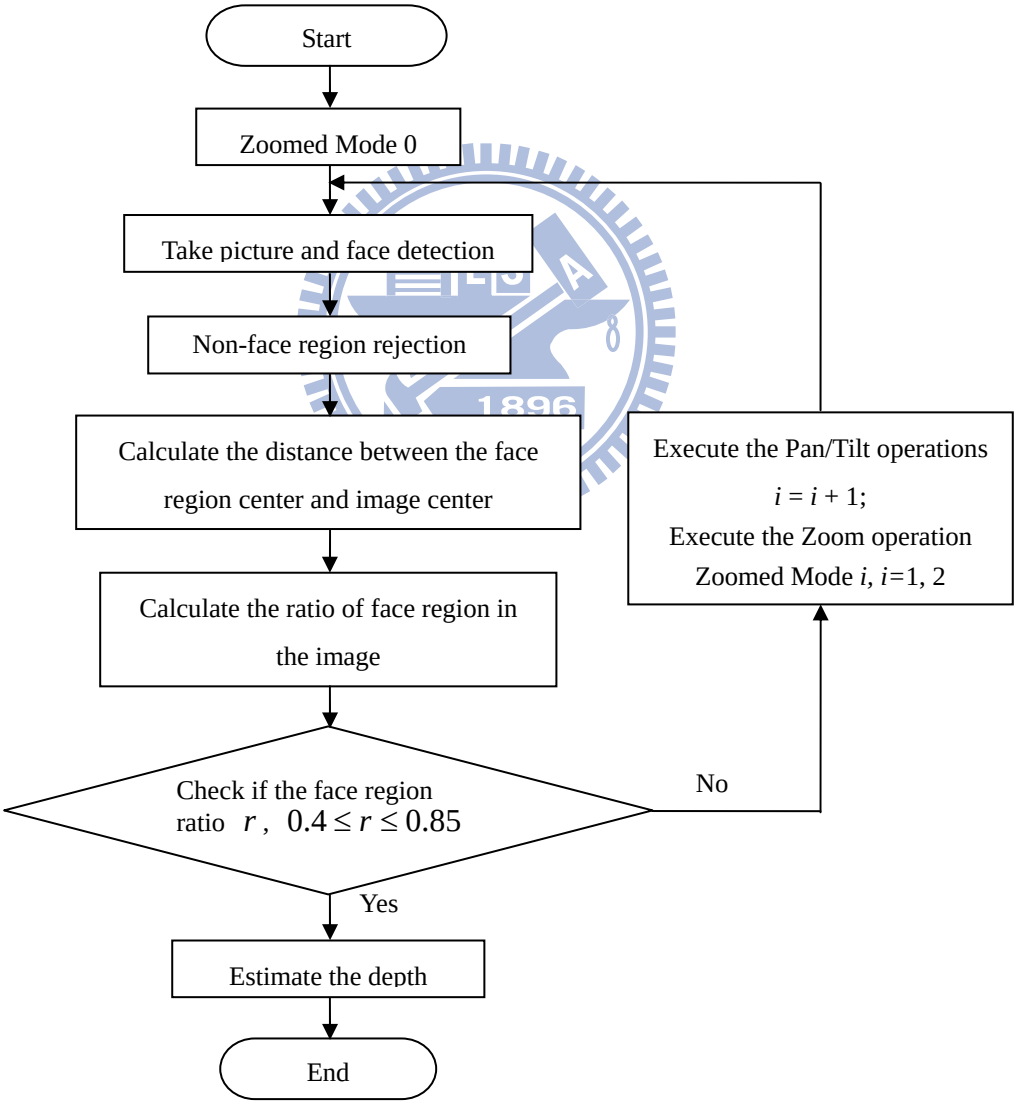


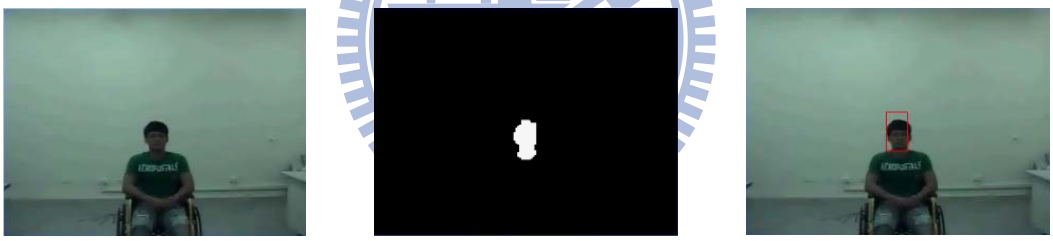
Fig. 4.11 The human tracking and zooming flowchart of the Sony EVI-D100/P camera.

TABLE XI

THE LOOK-UP TABLES CORRESPONDING TO THE OCCUPYING RATIO OF THE HUMAN FACE IN THE IMAGE AT THREE ZOOMED MODES AND DEPTH FROM THE CAMERA

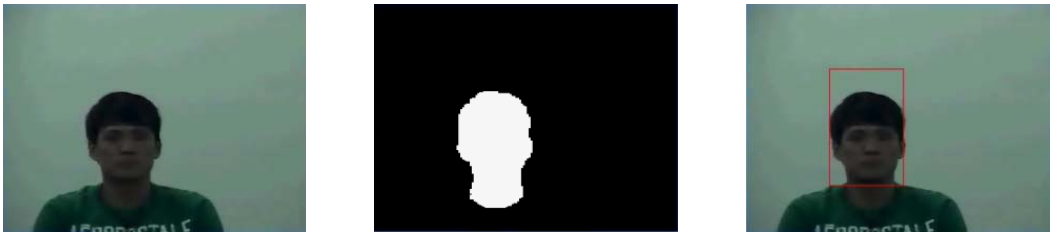
Depth from the camera	Zoomed Mode 0 (Initial)	Zoomed Mode 1	Zoomed Mode 2
1M	0.475	—	—
2M	—	0.775	—
3M	—	0.585	—
4M	—	0.468	—
5M	—	(0.374)	0.743
6M	—	(0.314)	0.612

Zoomed Mode 0



(a_1) (a_2) (a_3)

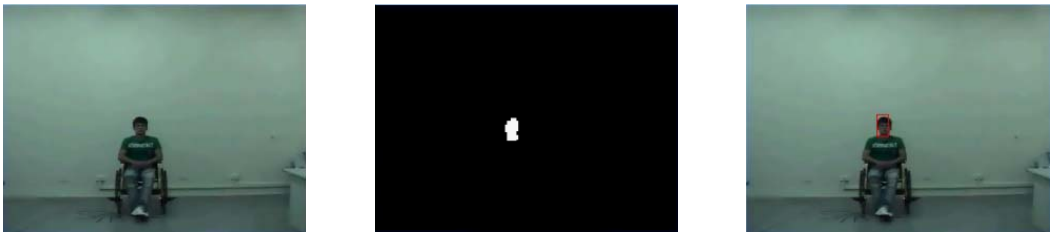
Zoomed Mode 1



(b_1) (b_2) (b_3)

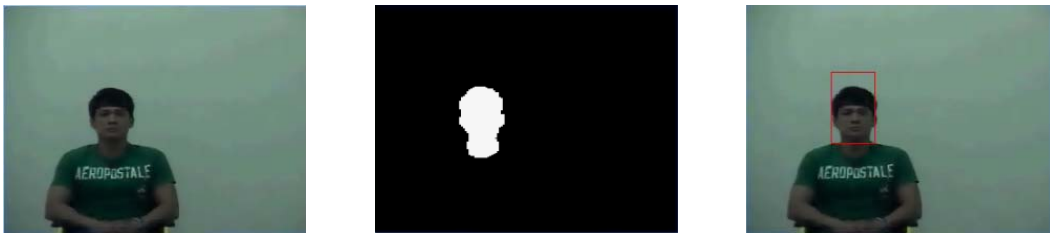
Fig. 4.12 The first example of face tracking and zooming process by PTZ camera if the distance between the person and the camera is 3.3M. (a_1)–(b_1) Input image. (a_2)–(b_2) Face detection by YC_bC_r skin color segmentation method. (a_3)–(b_3) The result of face tracking process.

Zoomed Mode 0



(a_1) (a_2) (a_3)

Zoomed Mode 1



(b_1) (b_2) (b_3)

Zoomed Mode 2

(c_1) (c_2) (c_3)

Fig. 4.13 The second example of face tracking and zooming process by PTZ camera if the distance between the person and the camera is 5.6M. (a_1)–(c_1) Input image. (a_2)–(c_2) Face detection by YC_bC_r skin color segmentation method. (a_3)–(c_3) The result of face tracking process.

TABLE XII

THE HUMAN DEPTH ESTIMATION FROM SONY PTZ CAMERA

	Zoomed Mode	The occupying ratio of the human face in the image (%) (* : final)	The estimated depths	The actual depths	Depth estimation accuracy (%)
Example 1	Zoomed Mode 0	17.50%	3.62M	3.3M	90.30%
	Zoomed Mode 1	51.25%*			
	Zoomed Mode 2	—			
Example 2	Zoomed Mode 0	10.00%	5.84M	5.6M	95.71%
	Zoomed Mode 1	31.25%			
	Zoomed Mode 2	63.33%*			

Chapter 5 Conclusion

In this thesis, we have presented an automatic image segmentation approach in subject extraction. In our approach, the initial seeds are chosen automatically if a pixel is inside the edge region. Then the initial seed regions are generated and labeled, and the regions start growing from the unclassified pixel located along region boundary with the minimum color distance to one of its neighbors. After all pixels are classified and labeled, the region merging procedure is performed if any two neighboring regions have high similarity. Application of our proposed image segmentation algorithm to human extraction and depth estimation is discussed. Firstly, the skin-color detection and elliptical template matching are utilized to extract the human face. The human body is determined by analyzing semantic human body rules. Then the human is extracted by combining the detected human face and human body. At last, the relative depth is described as the depth gradient map and the absolute depth can be estimated based on either the cross-ratio formula or the look-up table of a camera.

Experiment results have shown that our approach can simplify the automatic seed generation procedure, reduce the computation burden, and obtain good results on human extraction and depth estimation.

To investigate further, we shall detect different objects such as buildings, cars, and etc. In addition, extending semantic human body rules to detect human with different activities and extracting objects we interested from more complicated scene are our future work.

References

[1] E. Hjelmas and B. K. Low, "Face detection: A survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[2] C. Y. Tang, Z. Chen, and Y. P. Hung, "Automatic detection and tracking of Human Heads using an active stereo vision system," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 14, no.2, pp. 137–166, 2000.

[3] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, no. 6, pp. 641–647, June 1994.

[4] J. Fan, D. K. Y. Yau, A. K. Elmagarmid, and W. G. Aref, "Automatic image segmentation by integrating color-edge extraction and seeded region growing," *IEEE Trans. Image Processing*, vol.10, no.10, pp.1454 – 1466, October 2001.

[5] F. Y. Shih and S. Cheng, "Automatic seeded region growing for color image segmentation," *Image and Vision Computing*, vol. 23, pp. 877–886, 2005.

[6] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, no. 6, pp. 679–698, November 1986.

[7] Y. Xiong and S. A. Shafer, "Depth from focusing and defocusing," *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 68 – 73, 1993.

[8] A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology," *International Journal of Computer Vision*, vol. 40, no.2, pp. 123–148, 2000.

[9] Q. T. Luong and O. D. Faugeras, "Self-calibration of a moving camera from point correspondences and fundamental matrices," *International Journal of Computer Vision*, vol. 22, no.3, pp.261–289, 1997.

[10] Y. Dai and Y. Nakano, "Face-texture model based on SGLD and its application in face detection in a color scene," *Pattern Recognition*, vol. 29, no. 6, pp.

1007–1017, 1996.

[11] C. Garcia and G. Tziritas, “Face detection using quantized skin color regions merging and wavelet packet analysis,” *IEEE Trans. Multimedia*, vol. 1, no. 3, pp. 264–277, September 1999.

[12] D. Chai and K. N. Ngan, “Face segmentation using skin-color map in videophone applications,” *IEEE Trans. On Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 551–564, June 1999.

[13] M. C. Chi, J. A. Jhu, and M. J. Chen, “H.263+ region-of-interest video coding with efficient skin-color extraction,” in *Proc. IEEE Conference on Consumer Electronics*, pp. 381–382, 2006.

[14] H. D. Cheng, X. H. Jiang, Y. Sun, and J. L. Wang, “Color image segmentation: Advances and prospects,” *Pattern Recognition*, vol. 34, no. 12, pp. 2259–2281, December 2001.

[15] R. C. Gonzales and R. C. Woods, *Digital image processing*. Prentice Hall, 2002.

[16] Y. Deng and B. S. Manjunath, “Unsupervised segmentation of color-texture regions in images and video,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 8, pp. 800–810, 2001.

[17] S. Battiato, S. Curti, M. La Cascia, M. Tortora, and E. Scordato, “Depth-map generation by image classification,” *Proceedings of SPIE*, vol. 5302, pp. 95–104, April 2004.