國 立 交 通 大 學

電 控 工 程 研 究 所

碩 士 論 文

基於時序影像差之人體輪廓擷取與頭部偵測

Video Human Silhouette Extraction and Human Head Detection

Based on Temporal Difference

研 究 生：陳 嘉 臨

指 導 教 授：張 志 永

中 華 民 國 九十九 年 七 月

基於時序影像差之人體輪廓擷取與頭部偵測

Video Human Silhouette Extraction and Human Head Detection Based on Temporal Difference

學　　生：陳嘉臨　　Student : Chia-Lin Chen

指導教授：張志永　　Advisor : Jyh-Yeong Chang

國立交通大學

電機工程學系

碩士論文

A Thesis

Submitted to Department of Electrical Engineering

College of Electrical Engineering

National Chiao-Tung University

in Partial Fulfillment of the Requirements

for the Degree of Master in

Electrical Control Engineering

July 2010

Hsinchu, Taiwan, Republic of China

中華 民 國 九 十 九 年 七 月

基於時序影像差之人體輪廓擷取與頭部偵測

學生：陳嘉臨　　指導教授：張志永博士

國立交通大學電控工程研究所

摘要

前後景分離是將我們感興趣的物體(前景)從靜止的影像(背景)中分割出來。此分離技術經常是許多影像監控系統的第一個步驟，前後景分離的正確性高度影響了之後的步驟，例如：物體追蹤、姿態辨識、動作辨識，由此可見其重要性。在這篇論文中，我們提出了一個基於時序影像差的人體輪廓擷取技術，可以在**不事先建立背景模型**的情況下，從非完全控制的環境中(室外或是有光線變化的室內)將人體輪廓擷取出來，並且對於背景的亮度變化具有適應性。我們從影像中連續三個畫面得到時序影像差，並且結合邊緣資訊，將畫面中變動物體的邊緣取出。畫面中變動物體的邊緣可能不完整，為一個非封閉的曲線，因此我們提出背景區域成長演算法，可以在物體的邊緣輪廓不完整的情況下，擷取出影像中前景的輪廓。

人體的形狀經常與其他物體的形狀有很大的差異，因此可以視為人體偵測的重要線索，其中又以人體的頭部為最重要的人體特徵。我們將基於時序影像差之前後景分離法作為頭部偵測的前處理，以簡化複雜的背景並且縮減頭部的搜尋範圍。接著我們提出了基於模糊理論樣板比對的方法結合形狀與顏色的資訊將頭部定位。首先建立了人體頭部的左右邊緣模型、人體膚色模型與人體髮色模型，分別建立人體頭部的左右模型可以使得可偵測的人體頭部寬度具有可變動的範

圍，並對不同角度的人頭(例如：正面、側面和背面)具有適應性。利用人體頭部的左右邊緣模型與邊緣資訊做比對，偵測出搜尋範圍中各區域與模型相似的程度作為形狀分數，再利用人體膚色模型與人體髮色模型計算出區域中屬於膚色或髮色的程度作為顏色分數，最後結合形狀分數與顏色分數將人體的頭部定位。藉由頭部的定位可以確認所偵測的前景為人體，並對之後的人臉辨識、動作辨識提供有用的資訊。

Video Human Silhouette Extraction and Human Head Detection Based on Temporal Difference

STUDENT: Chia-Lin Chen ADVISOR: Dr. Jyh-Yeong Chang

Institute of Electrical Control Engineering
National Chiao-Tung University

ABSTRACT

Foreground-background segmentation is the process of separating the objects of interest (foreground) from the rest of the image (the background). It is often the first step in many visual-based surveillance systems and therefore a crucial process. The following processes such as tracking, pose estimation, and action recognition are highly dependent on the accuracy of the segmentation results. In this thesis, we propose a human silhouette extraction method based on temporal differencing for extracting complete human silhouette **without a pre-built background model**. The proposed method adapts quickly to changes in the scene and can extract human silhouette from incompletely controlled environment (outdoor or indoor with illumination change). We combine the temporal differencing from three successive video frames, current together with previous and the next, and the edge image to subtract the outline of moving object in the frame. The outline of the moving object could be incomplete therefore a non-closed curve. Hence, we propose a novel background region growing technique which grows the background region and then obtains the foreground silhouette from the incomplete edge image.

The shape of a human is often very different from the shape of other objects. Shape-based detection of humans can therefore be a powerful cue. Human head (including face) is the most important feature in the human shape. We take temporal differencing method as a pre-processing step before human head detection, which can simplify the complex backgrounds and reduce the detecting area. Then we propose a fuzzy theory based pattern-matching technique which combines the shape and color information to locate human head. We begin with building left head-shape model, right head-shape model, skin color model and hair color model. Detecting with two head-shape models gives somewhat size tolerance capability in human head width and adapts to different view angles, such as frontal view, lateral view, diagonal view, and so on. We compare the edge map of the given image with the pre-built left head-shape model and right head-shape model to detect head candidates. Then we use skin and hair color model to compute the belongness degree of each pixel within the head candidate area. Consequently, we combine the shape matching technique and color matching technique to better estimate the location of a human head. The resultant of human head detection can confirm the foreground extracted is human silhouette and is useful for face recognition, face tracking and motion recognition.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Jyh-Yeong Chang for his patient guidance and inspiration during these two years. Thanks are also given to all the teachers who once gave me more knowledge edification for valuable suggestions. In addition, I am thankful to all the people who assisted me in completing this research especially all members in Lab.

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.

Foreground-background segmentation is often applied as the first step in many visual-based surveillance systems and therefore a crucial process. One method of foreground-background segmentation is temporal differencing which adapts quickly to changes in the scene and does not need a pre-built background model. However, pixels from the foreground that have not moved or are similar to their neighbors are not detected. Therefore the detection result often falls into pieces. The incomplete detecting results cannot provide enough information to the following process.

The shape of a human is often very different from the shape of other objects. Shape-based detection of humans can therefore be a powerful cue. Human head (including face) is the most important feature in the human shape. The location of human head is useful for identification and motion recognition.

In summary, this motivates us to design a human silhouette extraction method based on temporal difference and edge information which having good adaptability to changes in the scene and coping with the incompletion of the detection results. Then we take human silhouette extraction method as a pre-processing step before human

head detection, which can simplify the complex backgrounds and reduce the detecting area. Finally, we propose a fuzzy theory based pattern-matching technique which combines the shape and color information to locate human head. The system flowchart is illustrated in Fig 1.1.

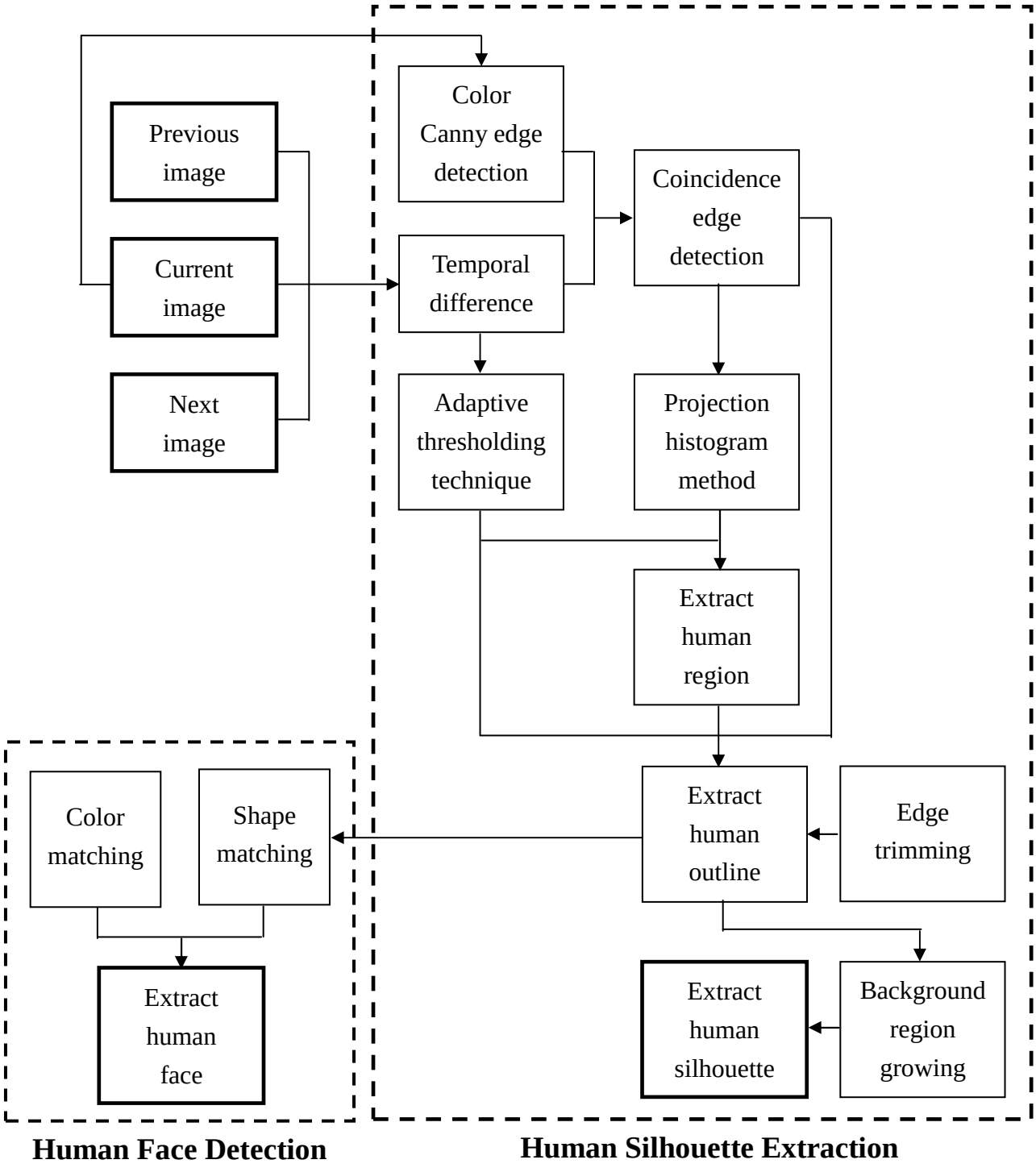


Fig. 1.1 System flowchart.

1.2. Human Silhouette Extraction

Foreground-background segmentation is the process of separating the objects of interest (foreground) from the rest of the image (the background). It is often applied as the first step in many visual-based surveillance systems and therefore a crucial process. This is evident by looking at the number of researches devoted to such topics which tend to achieve much faster and more precise results [1]. The following processes such as tracking, pose estimation, and movement recognition are highly dependent on the accuracy of the segmentation results. The results have a great effect on the whole system and even can decide the reliability and the precision of the system. Therefore the performance of the foreground-background segmentation could be a conclusive gauge of whether a surveillance system is good or bad.

The methods of foreground-background segmentation can be roughly divided into two categories: (1) background subtraction and (2) temporal differencing. Background subtraction is the process to detect foreground by subtracting each pixel in the current frame from those in the pre-built background model. Temporal differencing is the process to detect foreground by subtracting each pixel in the current frame from those in the previous frames.

Background subtraction was known as a powerful pre-processing step in controlled indoor environments which can be represented as a stable background model. Subtracting each pixel in the current frame from those in the background model could yield complete foreground silhouette. Stauffer and Grimson [2] presented the idea of representing each pixel by a mixture of Gaussians (MoG) and updating each pixel with new Gaussians during run-time. This allows background subtraction to be used in outdoor environments. Normally the updating was done recursively, which can model slow changes in a scene, but not rapid changes like

clouds.

Temporal differencing adapts quickly to changes in the scene. This approach assumes that the objects in the foreground are moving continuously. However, pixels from the foreground that have not moved or are similar to their neighbors are not detected. Therefore the detection result often falls into pieces. The incomplete detecting results cannot provide enough information to the following process.

Among all foreground objects, human extraction from the background is a highly active research area due both to the number of potential applications and its inherent complexity. Previously approaches were mostly tested in controlled environments and with only a few people present in the scene. Recently, algorithms have addressed more natural outdoor scenarios where multiple people and occlusions are present and have focused on detection of humans in still.

In this thesis, we propose an improvement to an existing temporal differencing method, and incorporate a novel technique for extraction of complete human silhouette without a pre-built background model.

1.3. Human Head Detection

The shape of a human is often very different from the shape of other objects in a scene. Shape-based detection of humans can therefore be a powerful cue. The advances are first of all to allow human detection and tracking in an uncontrolled environment on the premise that reliable silhouette outlines can describe the shape of the humans in the image sequence. Furthermore, advances are to allow human representation and segmentation in still images.

Many researches are specifically interested in tracking human heads or faces due to they are the most important feature in the human shape [3]. Because it is difficult to

automatically detect human heads or faces in images having complex backgrounds, much previous research dealt only with images having simple backgrounds. However, for many practical applications, automatically detection and tracking of human heads should not be limited to simple backgrounds.

Most of the previous research concentrated on quasi-frontal view faces [4][5][6]. This is because the prior knowledge of the geometric relation with regard to the facial topology of frontal view faces can help the detection of facial features and it also makes the face modeling with a generic pattern possible. However, the quasi-frontal view assumption limits the kind of faces that can be processed. Another disadvantage is that the facial-feature-based approaches rely on the performance of feature detectors. For small faces or low quality images, the proposed feature detectors are not likely to perform well.

In this thesis, we take temporal differencing method as a pre-processing step before human head detection, which can simplify the complex backgrounds and reduce the detecting area. We propose a fuzzy theory based pattern-matching technique, and use it to detect head silhouette outlines from the edge map, the extracted skin and hair regions.

1.4. Thesis Outline

The thesis is organized as follows. An improved method of temporal differencing which can extract complete human silhouette is described in Chapter 2. In Chapter 3, a human head extraction method by fuzzy shape matching and skin-color/hair-color extraction is introduced. In Chapter 4, the experiment results of our object segmentation and human head extraction system are shown. At last, we conclude this thesis with a discussion in Chapter 5.

Chapter 2 Human Silhouette Extraction

In this chapter, we extract human silhouette by two steps. First, we use the color Canny edge map and combine with the temporal difference from three successive frames to extract human outline. Then we propose a background region growing method to extract human silhouette from the human outline even when the outline is not a closed curve.

2.1. Human Outline Extraction

We combine color edge information with the temporal difference to obtain coincidence edges which are edges of the current image and also have great value in their temporal difference. We use the coincidence edges to capture the human outline.

2.1.1. Color Canny Edge

Most edge detection schemes are based on finding the maxima in the first derivative of the image function or zero-crossings in the second derivative of the image function. The difficulty in extending derivative approaches to color images arises from the fact that the image function is vector-valued. Whenever the gradients of the image components are computed, the question remains of how to combine them into one result. Several approaches already exist for color edge detection. Perhaps the simplest one is to apply an edge detector for grayscale to the three color channels independently and to combine the results using logical operation.

In this way, we run each color channel through the Canny edge detector separately to yield edge maps in R channel (E_R), in G channel (E_G), and in B channel (E_B). There are many ways to combine three edge images to one general

edge image. We choose “OR” operator to reserve most edge information, i.e., if there is an edge in any one of the three colored edge images, we add it to the general edge image (E).

$$E(i,j)=\begin{cases}1, & \text{if } E_R(i,j)=1 \cup E_G(i,j)=1 \cup E_B(i,j)=1 \\ 0, & \text{otherwise}\end{cases} \quad (1)$$

The human silhouette extraction method proposed in this thesis is based on the edge image. Hence if the edge information is lost, the silhouette we extract will be incomplete.

Some results of edge detection are shown in Fig. 2.1 and 2.2. It should be noticed that the outline of the human can not be extracted completely from the edge image. See Figs. 2.1(b) and 2.1(f) for examples, Fig. 2.1(b) shows a frame image of Daria performing “waving-two-hands” action and Fig. 2.1(f) is the edge image of Fig. 2.1(b). Comparing the frame image and the edge image, we can find that Daria’s hands are not detected in the edge image. Therefore, the silhouette of Daria’s hands will be lost in the silhouette extraction method.

The outline of the human is often a non-closed curve. See Figs. 2.1(d) and 2.1(h) for examples, Fig. 2.1(d) shows a frame image of Daria performing “galloping-sideways” action and Fig. 2.1(h) is the edge image of Fig. 2.1(d). Comparing these two images we can find that the sole of Daria’s left foot in the edge image is not detected, hence the outline of Daria’s body is a non-closed curve. If the outline of an object is a closed curve, extracting its silhouette is easy by applying connected component labeling [7]. However, it is a relatively challenging task to extract silhouette of an object whose outline is a non-closed curve.

The outline of a person in clothes which have similar color to background is especially detected incompletely as shown is Fig. 2.2. Shahar wears white trousers

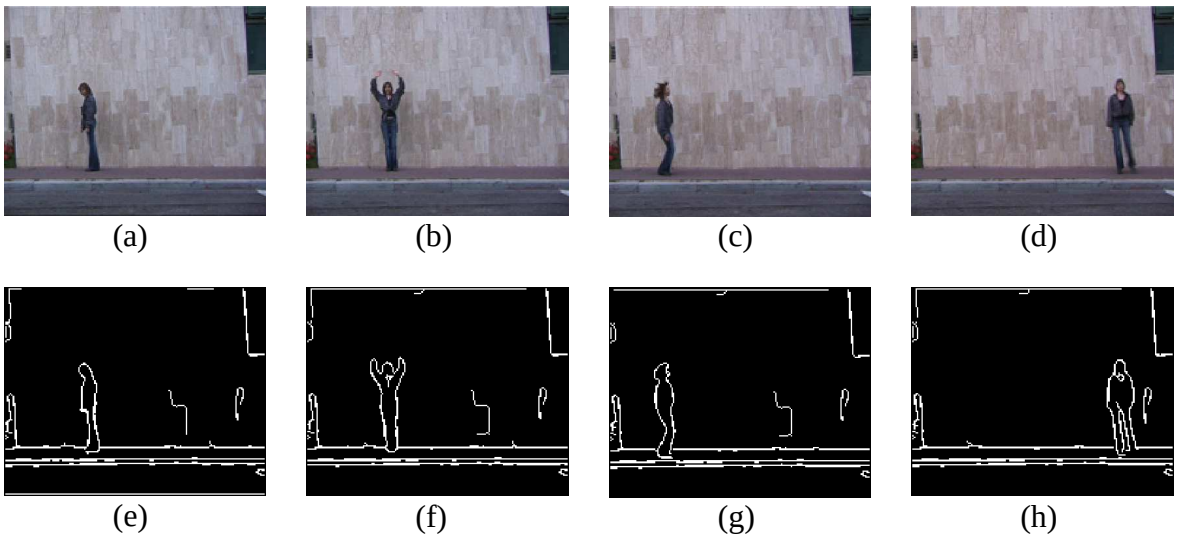


Fig. 2.1. (a)—(d) Daria performing “bending”, “waving-two-hands”, “jumping-forward-on-two-legs” and “galloping-sideways” actions; (e)—(h) corresponding edge images by color Canny edge detection.

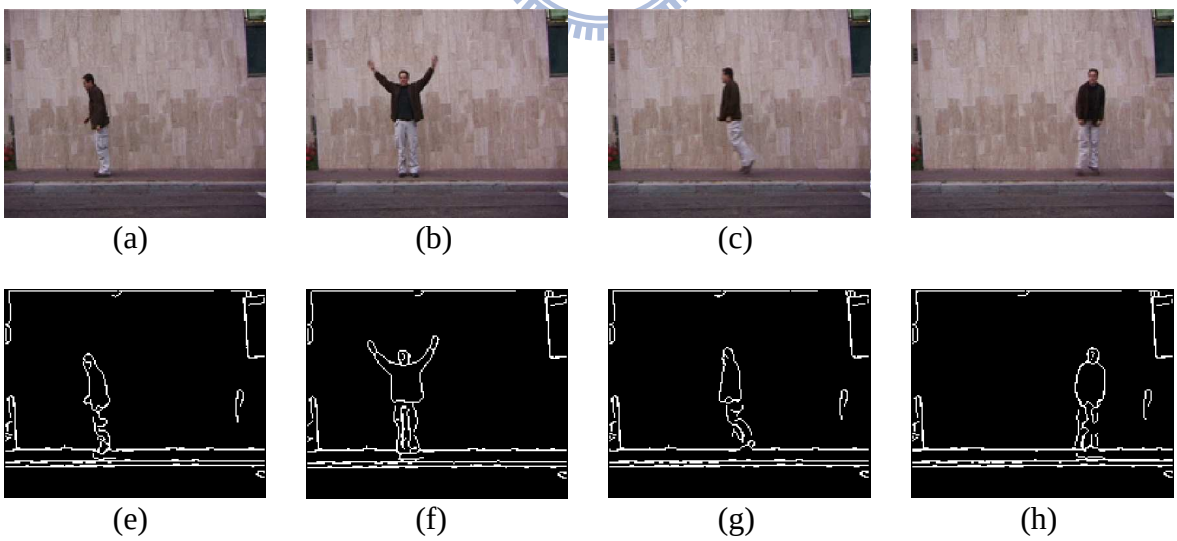


Fig. 2.2. (a)—(d) Shahar performing “bending”, “waving-two-hands”, “jumping-forward-on-two-legs” and “galloping-sideways” actions; (e)—(h) corresponding edge images by color Canny edge detection.

whose color is similar to the background; hence the outline of the trousers can not be detected completely by the edge detector. This problem is un-avoidable in every foreground-background segmentation method, however.

2.1.2. Temporal Difference Image

We extract the motive information by making a temporal difference image from successive frames in a video stream. First, we generate current image I_t, N previous images I_{t-N}, I_{t-N+1}, …, I_{t-1}, and N following images I_{t+1}, I_{t+2}, …, I_{t+N}, from successive frames in an video stream. That means when we apply this method in real time, it will delay for N frames. Then we summarize the absolute difference between current image and all previous and following images. We call a resultant image a temporal difference image $D(i, j)$ as given by

$$D(i,j) = \sum_{k=1}^{N} | I_{t-k}(i,j) - I_t(i,j) | + \sum_{k=1}^{N} | I_{t+k}(i,j) - I_t(i,j) | \qquad (2)$$

In the implementation, N is set to 1. It is to be noted that the system output has to delay one time frame for we employ image I_{t+1} in Eq. (2).

Fig. 2.3 shows an example of temporal difference image. Fig. 2.3(d) is absolute difference image between the current image (b) and the previous image (a), Fig. 2.3(e) is absolute difference image between the current image (b) and the following image (c) and Fig. 2.3(f) is the sum of (d) and (e) which called temporal difference image. Each difference image is shown in 8-bit grey level which maximum value is 255 and indicated by the lightest color in the figure. If the intensity of the difference image is larger than 255, it is also indicated by the lightest color.

In the difference image, the great values occur at the location occupied by the foreground in one image but occupied by the background in another image. Therefore

the human silhouette in the difference image is expanded.

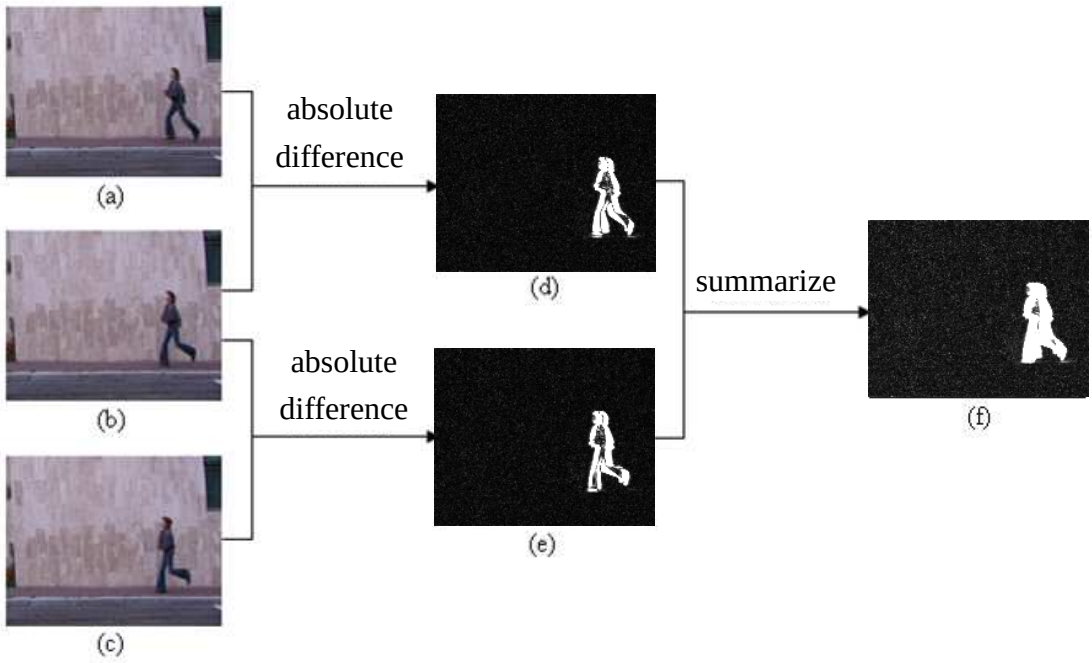


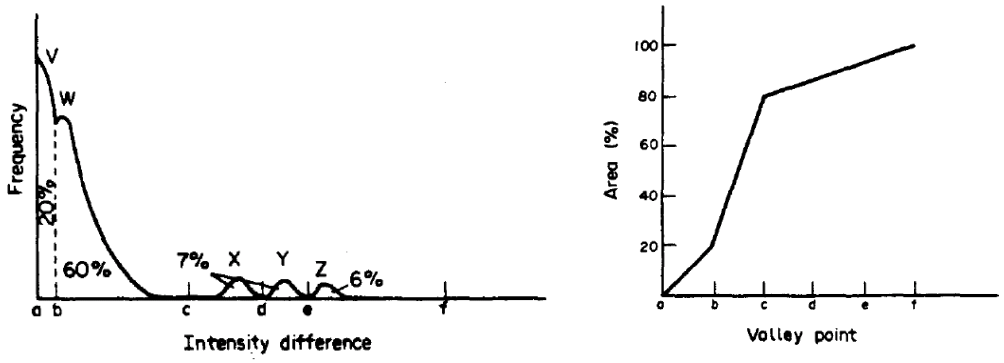
Fig. 2.3. (a)—(c) Successive three, previous, current and next, frames of Daria performing "running" action, (d) Absolute difference image between the current and previous image, (e) Absolute difference image between the current and the next image, (f) The resulting (from summing (d) and (e)) temporal difference image.

Intuitively, stationary regions can be eliminated through the subtraction process and only regions that have been moved can appear in the difference picture. However, in reality, the temporal difference image obtained often contains extraneous information because of changes in the illumination condition and noise (see Fig. 2.3(f), there is much salt-and-pepper noise in the background). Hence, stationary regions may frequently survive the differencing process.

A adaptive thresholding technique was developed to cope with the above mentioned problem by analyzing the shape of the temporal difference image histogram (i.e. occurring frequency versus intensity of the temporal difference image) [8]. It is assumed that (1) the area of the stationary regions is larger than or equal to

the area covered by the regions in motion and (2) the pixels within all the stationary regions undergo approximately the same intensity change with small variation. Consequently, the pixels from the stationary regions are grouped under a few peaks in the histogram with large area while the pixels within the moving regions are grouped under a number of peaks with relatively small area.

Fig. 2.4(a) illustrates an example with peaks V and W corresponding to the stationary regions and peaks X, Y and Z corresponding to the moving regions. The areas under the peaks are 20, 60, 7, 7 and 6% of the total area, respectively. Fig. 2.4(b) shows the accumulated area from valley point *a* to point *f* in Fig. 2.4(a). It should be noted that the curve is plotted versus the valley points which are spaced at equal interval. Also the area between two consecutive valley points in Fig. 2.4(a) is equal to the slope of the line between the two corresponding points in Fig. 2.4(b). The change in the slope at a valley point gives an indication of the difference of contribution due to the next peak. Because of the assumptions (1) and (2), the separation between the stationary regions and the moving regions occurs at the valley point with the largest slope change. This valley point is then chosen as the threshold value.



(a) (b)

Fig. 2.4. (a) An example of temporal difference image histogram distribution, (b) accumulative area chart of (a) [8].

Fig. 2.5 shows one real example of temporal difference image and the charts used to deciding its threshold. Fig. 2.5(a) shows a frame image of Daria performing "running" action and Fig. 2.5(b) shows the temporal difference images of (a). We can observe that (1) the background area is larger than or equal to the area covered by the regions in motion and (2) there is salt-and-pepper noise spread in the background area which means the assumptions of the adaptive thresholding technique are satisfied in our case. A part of the histogram distribution of Fig. 2.5(b) is shown in Fig. 2.5(d) and we mark the valley points with red dots. There are a few peaks in the histogram with large area while a number of peaks with relatively small area. A part of the accumulated area line chart of Fig. 2.5(d) is shown in Fig. 2.5(e) and we mark the valley point with the largest slope change which is supposed to be the separation between the background regions and the moving regions. The coordinate of the valley point (v1, 95.13) means the threshold value is chosen to be v1 and there are 95.13 percentage of area in the temporal difference image labeled as background region. Fig. 2.5(e) shows the temporal difference image after thresholding and major part of the extraneous information due to changes in the illumination condition and noise is eliminated.

All movements in the Weizmann human action database can roughly divided into "whole body movement" and "partial movement". When human perform "whole body movement" such as "running", "walking", "jumping-jack", "jumping-in-place-on-two-legs", "jumping-forward-on-one-leg", "jumping-forward-on-two-legs" and "galloping-sideways", each part of their bodies has displacement and there are relatively complete human shapes in temporal difference images (see Fig. 2.6). However when human perform "partial movement" just parts of their bodies move and there are incomplete human shapes in temporal difference image (see Fig. 2.7). When human perform "waving-two-hands" or "waving-one-hand" movement just

their hands have displacement and other parts of their body stay still. When human perform “bending” movement just upper part of their bodies has displacement and lower part of their bodies stay still.

2.1.3. Coincidence Edge

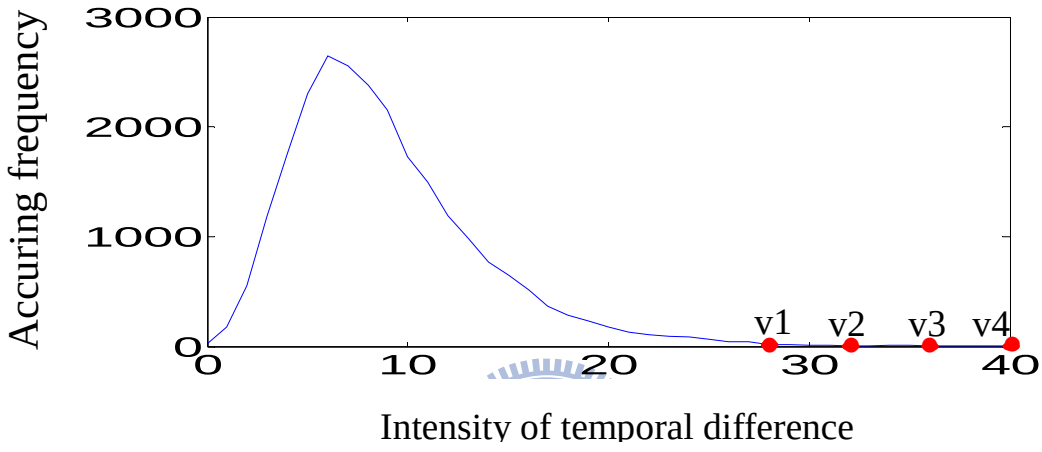
The coincidence edges are edges of the current image and also have non-zero values in temporal difference image. We use the coincidence edges to capture the edges of the moving objects. The moving edge detection technique can generate a more complete outline of the moving object because motion information is accumulated and tracked. The algorithm is based on the difference picture method developed by Jain and Nagel [9], [10] together with the coincidence edge accumulation process proposed in this thesis. The coincidence edge image is obtained by combining the edge information and the temporal difference image as follows:

$$CE(i,j)=\begin{cases} D(i,j), & \text{if } E(i,j)=1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

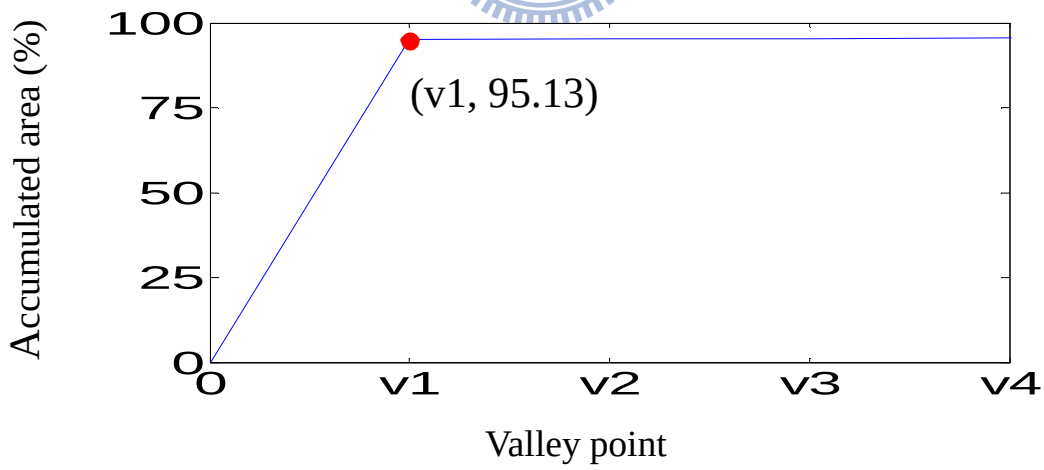
In the background subtraction method, the projection histogram is a commonly used method to roughly segment foreground region from the image. In this thesis, we try to segment foreground region from projection histogram of the edge image, the temporal difference image and the coincidence edge image, and we find the projection histogram of the coincidence edge image is a useful method which can segment foreground region precisely.

Fig. 2.8 shows an example of using projection histogram to extract motion region in background subtraction method. Let $I(x,y)$ be a result binary image of background subtraction which represents a detected motion region. The coordinate system of

(a) (b) (c)



(d)



(e)

Fig. 2.5. (a) Daria performing “running” image, (b) The temporal difference image of (a), (c) The temporal difference image after thresholding, (d) Histogram distribution of (b), the valley points marked with red dots, (e) Accumulative area chart of (d), valley point with the largest slope change marked with red dot.

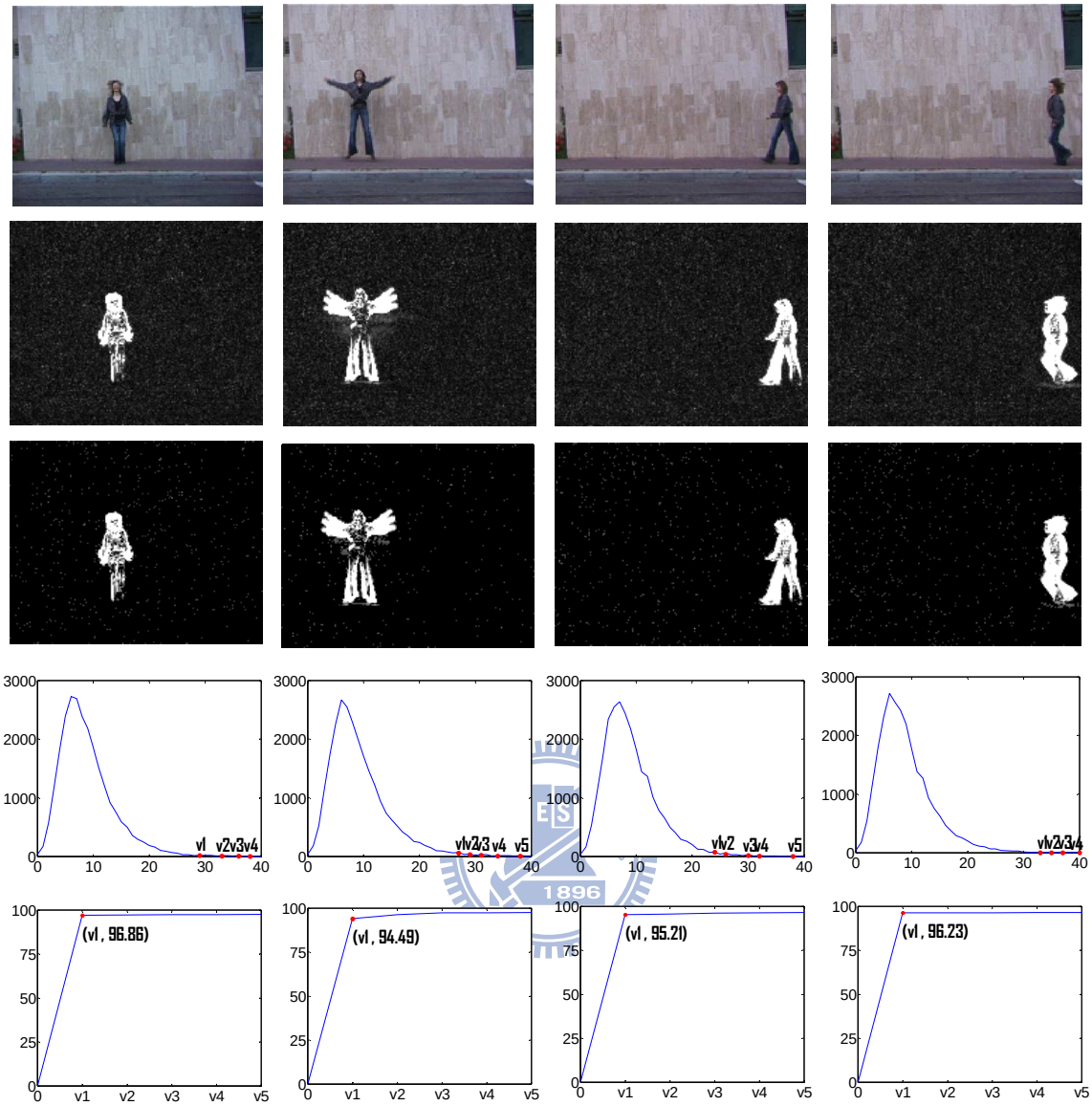


Fig. 2.6. Top row: Daria performing “jumping-in-place-on-two-legs”, “jumping-jack”, “walking” and “jumping-forward-on-one-leg” images; Second row: The corresponding temporal difference images; Third row: The temporal difference images after thresholding; Forth row: The partial histograms of the temporal difference image, intensity values changing from 0 to 40; Bottom row: The accumulated area line chart of the corresponding above histogram.

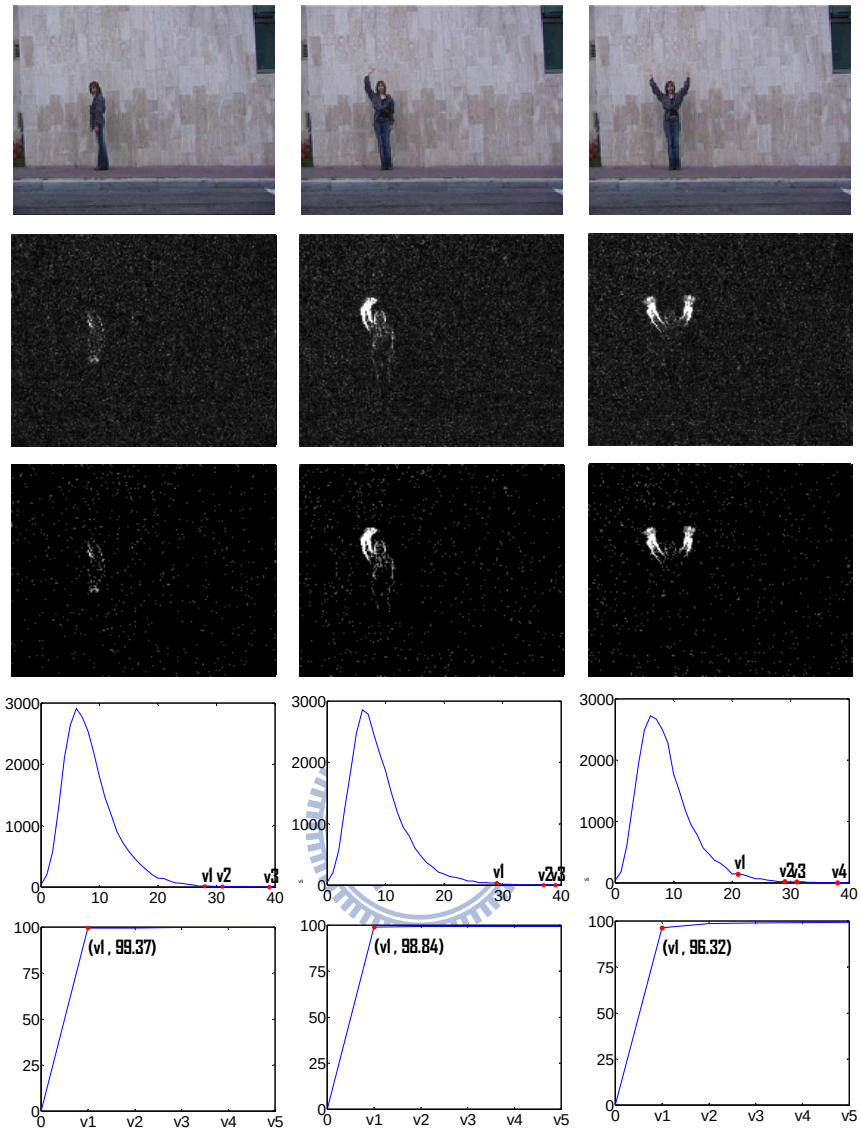


Fig. 2.7. Top row: Daria performing “bending”, “waving-one-hand” and “waving-two-hands” images. Second row: The corresponding temporal difference images. Third row: The temporal difference images after thresholding. Forth row: The partial histograms of the temporal difference image, intensity values changing from 0 to 40. Bottom row: The accumulated area line chart of the corresponding above histogram.

image I is expressed with x in the horizontal direction and y in the vertical direction. Let H and W be, respectively, the height and width of the image I. The vertical projection histogram of image I is acquired by projecting pixels onto the horizontal coordinate of the image as follows:

$$proj_v(x) = \sum_{y=1}^{H} I(x, y), \quad x \in [1, W] \tag{4}$$

$Proj_v(x)$ looks like a range of mountains with peaks and valleys. The region surrounding each mountain peak is likely to contain the motion region. The peaks and their surrounding areas of $proj_v(x)$ above some threshold are extracted to produce a vertical slice of the image. Next, a horizontal projection is created from the slice to determine the motion regions as follows:

$$proj_h(y) = \sum I(x, y), \quad y \in [1, H] \tag{5}$$

Extraction of the peaks and surrounding areas of $proj_h(y)$ results in an initial set of motion region.

In this thesis, we apply the projection histogram method to the edge image, the temporal difference image and the coincidence edge image. It should be noted that the edge image contains binary information while the temporal difference image and the coincidence edge image contain analog information.

Fig. 2.9 and Fig. 2.10 show examples of the vertical projection histogram from frames of Daria performing “whole body movement”. On the other hand, Fig. 2.11 and Fig. 2.12 show examples of the vertical projection histogram from frames of Daria performing “partial movement”. The red lines are marked manually as a

reference to locate the motive region.

It is obvious that the projection histogram of the edge image can extract human region in the simple background, but it also extract regions without human (see Fig. 2.9(e), Fig. 2.10(e), Fig. 2.11(e), and Fig. 2.12(e)). The projection histogram of the temporal difference image can extract the motive region while human perform "whole body movement". However, when human perform "partial movement", the separation between the stationary region and the moving region is blurred (see Fig. 2.11(f) and Fig. 2.12(f)). In Fig. 2.9(g) and Fig. 2.10(g), we can find that the projection histogram

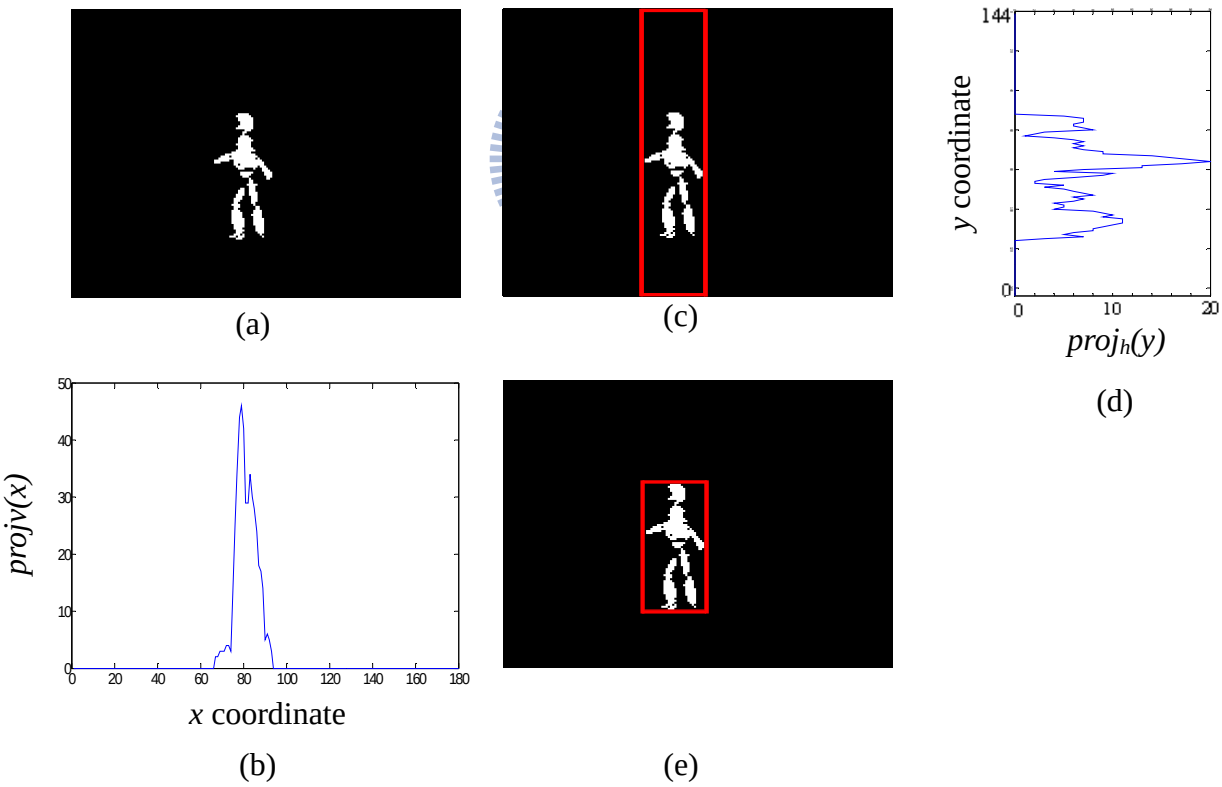


Fig. 2.8. (a) Resultant binary image $I(x,y)$ of background subtraction, (b) The vertical projection histogram of I, (c) The peaks and their surrounding areas of $proj_v(x)$ above some threshold extracted to produce a vertical slice of the image which is marked with a red frame, (d) The horizontal projection histogram of the vertical slice marked with a red frame in (c), (e) Resultant image confined by $proj_v(x)$ and $proj_h(y)$.

of the coincidence edge image can extract the motive region while human perform “whole body movement” as the projection histogram of the temporal difference imagecan do. Besides, in Fig. 2.11(g) and Fig. 2.12(g), the projection histogram of the coincidence edge image also can extract the human region due to the accumulation of small temporal differences on the human edge. To sum up, the projection histogram of the coincidence edge image is a useful method which can segment human region precisely no matter human performing “whole body movement” or “partial movement”.

The simple uniform threshold is marked by green lines in each projection histogram of the coincidence images (see Figs. 2.9(g)－2.12(g)). The peaks and their surrounding areas of the vertical projection histogram above the threshold are extracted to produce a vertical slice of the image. Next, a horizontal projection of the coincidence edge image is created from the slice to determine the human regions.

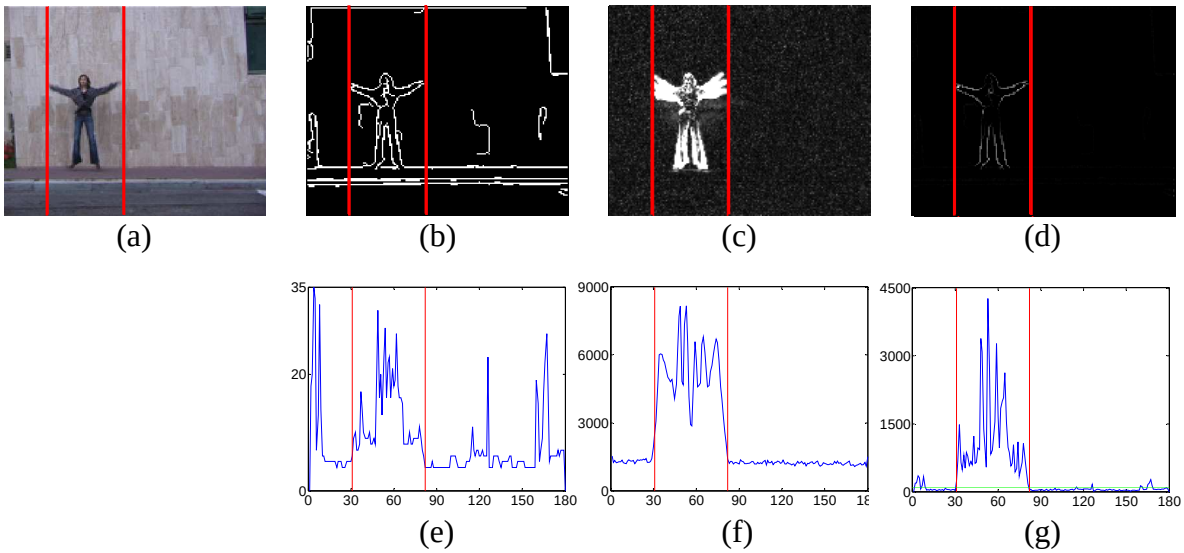


Fig. 2.9. (a) Daria performing “jumping-jack” which is a “whole body movement”, (b) The edge image, (c) The temporal difference image, (d) The coincidence edge image, (e)—(g) the corresponding projection histograms.

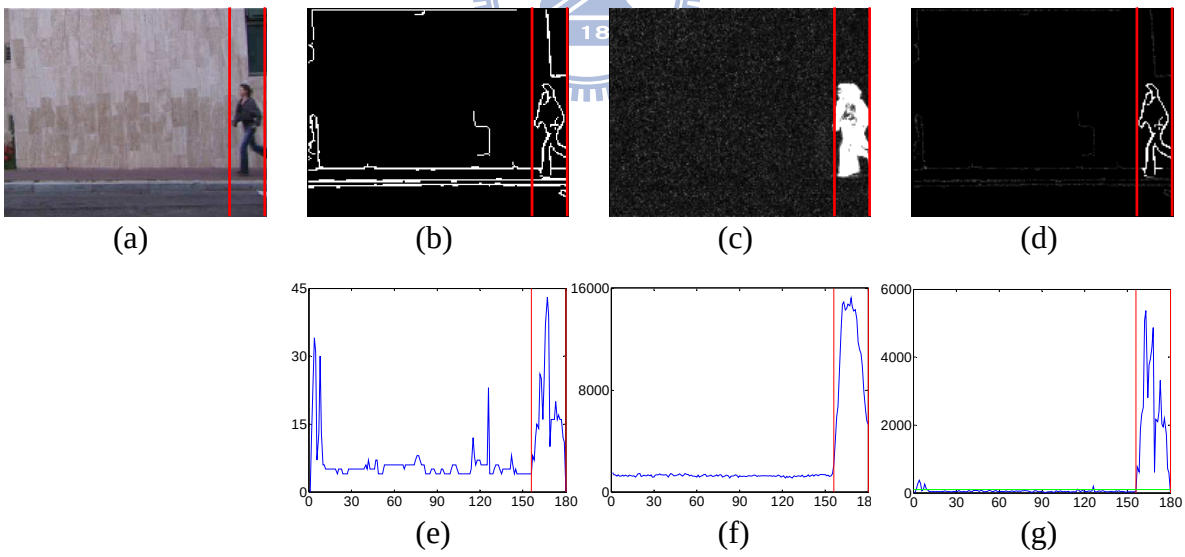


Fig. 2.10. (a) Daria performing “running” which is a “whole body movement”, (b) The edge image, (c) The temporal difference image, (d) The coincidence edge image, (e)—(g) the corresponding projection histograms.

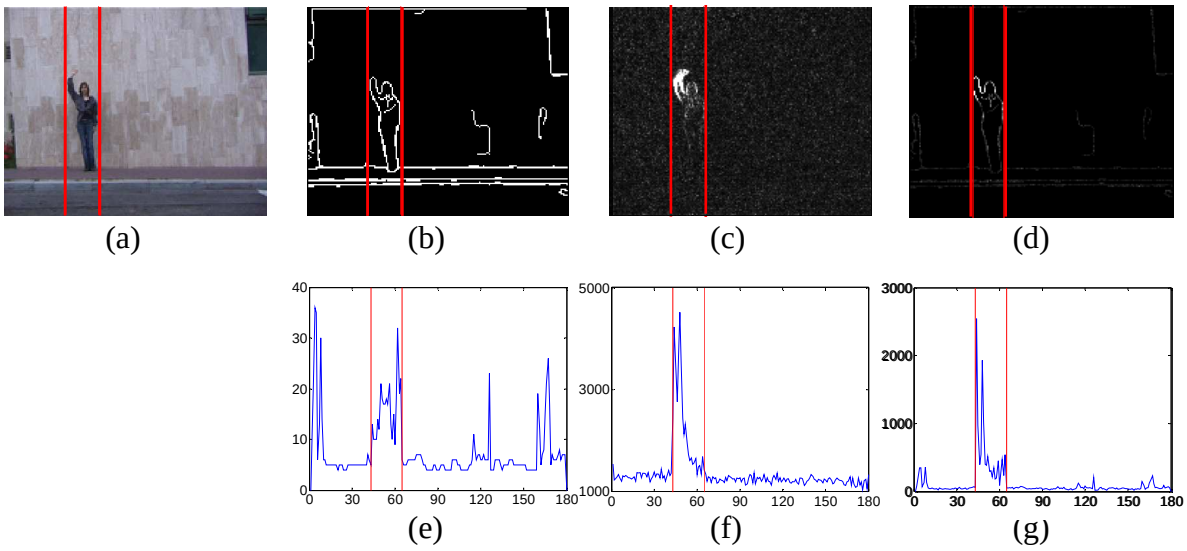


Fig. 2.11. (a) Daria performing “waving-one-hand” which is a “partial movement”, (b) The edge image, (c) The temporal difference image, (d) The coincidence edge image, (e)—(g) the corresponding projection histograms.

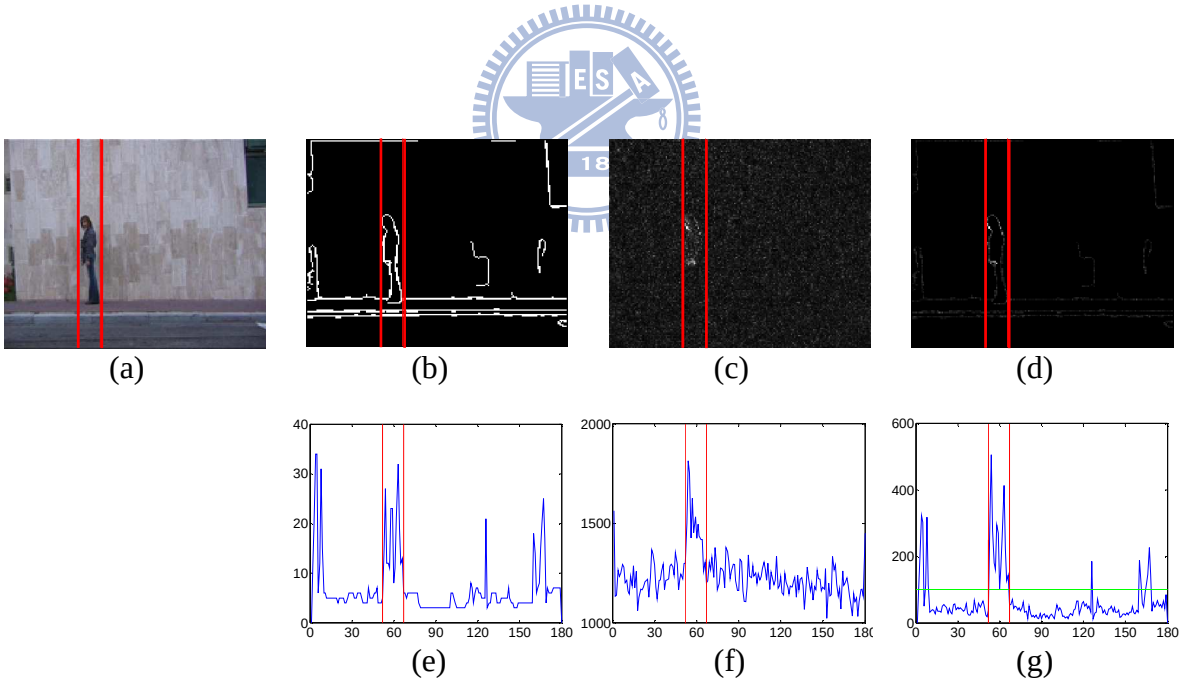


Fig. 2.12. (a) Daria performing “bending” which is a “partial movement”, (b) The edge image, (c) The temporal difference image, (d) The coincidence edge image, (e)—(g) the corresponding projection histograms.

2.1.4. Edge Trimming

Using horizontal and vertical projection histogram to extract human region is a rough method whose resultant region is bounded by a rectangle. In the resultant human region, there are still some edges not belong to human outline and we want to make out and trim away.

Most of the edges surrounding human outline are edges of the floor. The edge pixels of floor are often continuous on a vertical line. In the human region, if pixel number of a vertical line is more than half width of the human region, the vertical line is regarded as floor line. Every edge pixel on the floor line has lower temporal difference than threshold will be trimmed off.

Besides, the outline of an object (including human) often not a single line because the object has volume. Due to this, we designed four trimming matrixes T_1, T_2, T_3 and T_4 (shown is Fig. 2.13) which have pixel 0s in the center region and 1s surrounding, but a split pin in upper boundary, down boundary, left boundary and right boundary, respectively. If we define E as the edge image, and we calculate *T-score* as follows:

$$T\text{ - }socre = T_i * E \tag{6}$$

where * here denotes the 2-dimensional convolution operation and $1 \le i \le 4$. *T-score*=0 means there is a single edge line in the center region. Every edge pixel on the single edge line has lower temporal difference than threshold will be trimmed off..

Fig. 2.14 shows some resultant edge image after edge trimming. Images in the first column, the third column, the fifth column and the seventh are edges in the

human region extracted by projection histogram. The corresponding images below are the edge image after edge trimming.

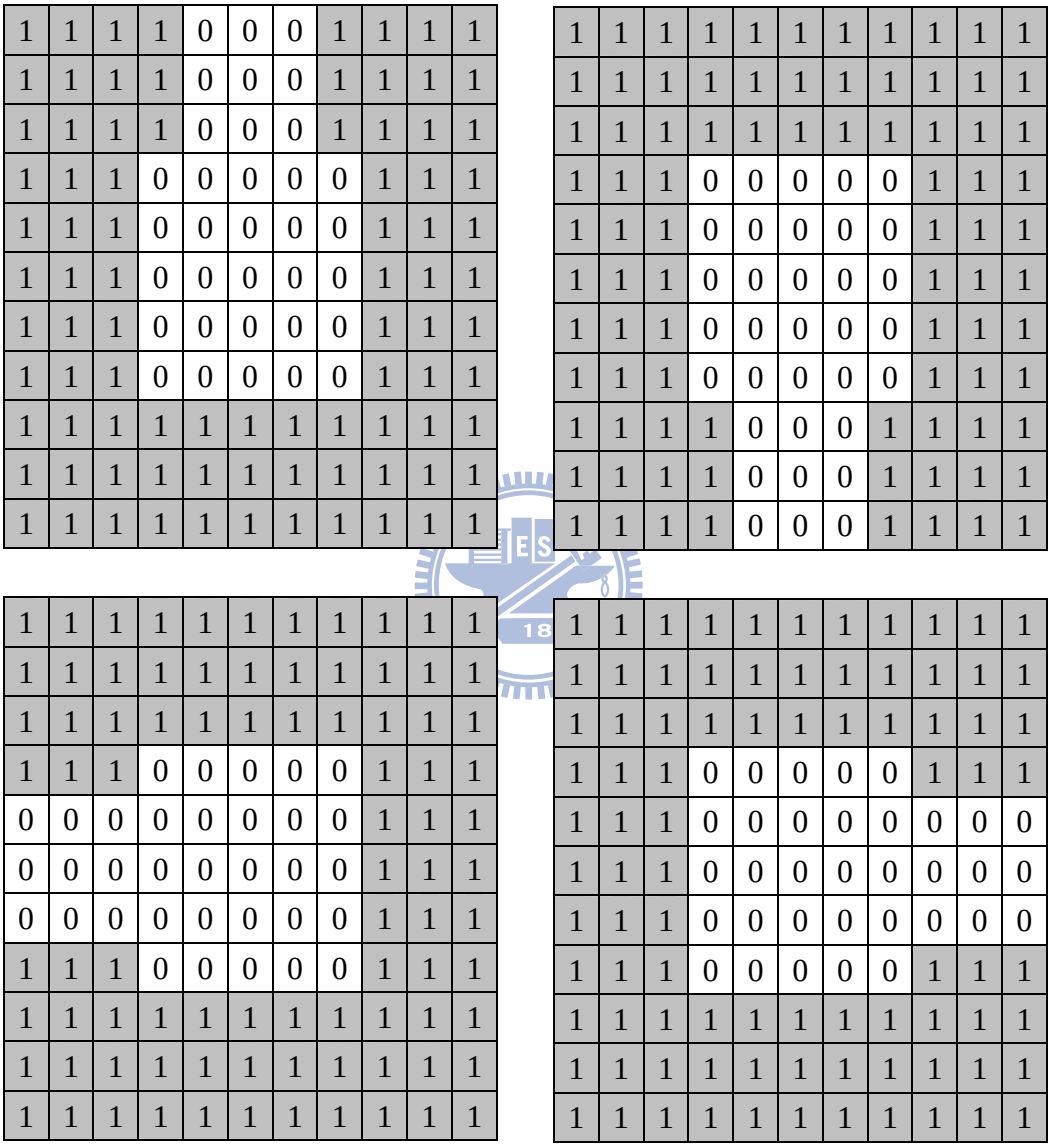


Fig. 2.13 Trimming matrixes to detect single edge line.

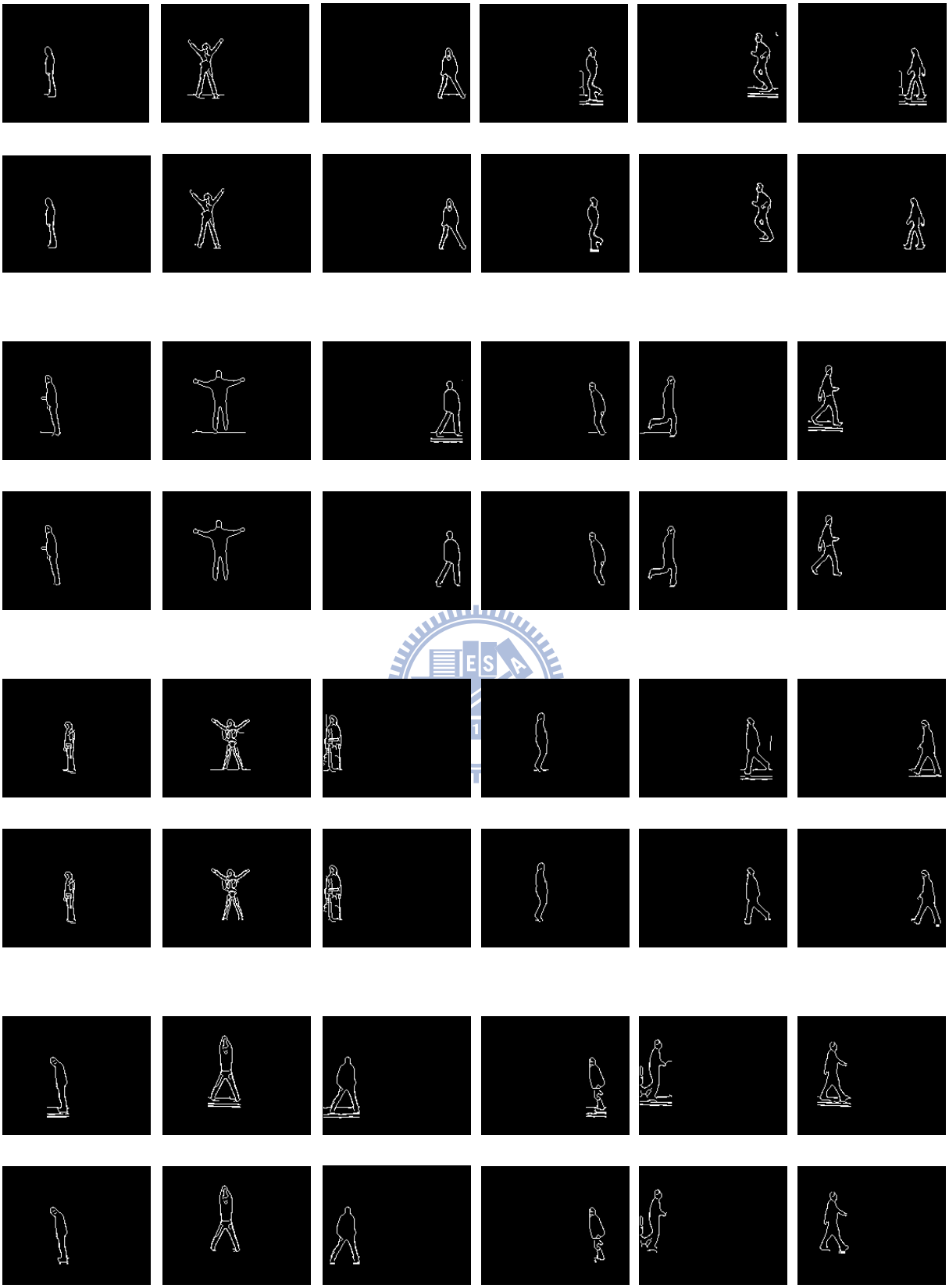


Fig. 2.14 The resultant edge images after edge trimming.

2.2. Background Region Growing

If the outline of human silhouette extracted is a closed curve, it is easy to mark the human silhouette which is the inner part of the closed curve. In the implementation, however, the outline of human silhouette extracted is often not continuous; it does not form a closed curve due to the color edge being not continuous. Consequently, we proposed a simple background region growing method so that the human silhouette can be better marked. The algorithm applies divide-and-conquer strategy and is described below:

Step 1: Prepare a new image whose pixels are all initialized to -1 which represents as background. The human region substitute for the corresponding region in the new image with pixel value 1 represented as edge and pixel value 0 represented as non-edge. An example is shown is Fig. 2.15(a), in which the pixels with value -1 being represent as gray color; the pixels with value 0 being represent as black color and the pixels with value 1 being represent as white color. The outer human region is set to be the first operating region which is marked with red frame in Fig. 2.15(b).

Step 2: For the operating region above, we will grow the background region from left red boundary to right until the vertical line containing an edge pixel. Similarly we will grow the background region from the bottom red boundary to upper, from the right red boundary to left and from the upper red boundary to bottom. After this showing process, we will obtain the new outer human region shown in Fig. 2.15(c).

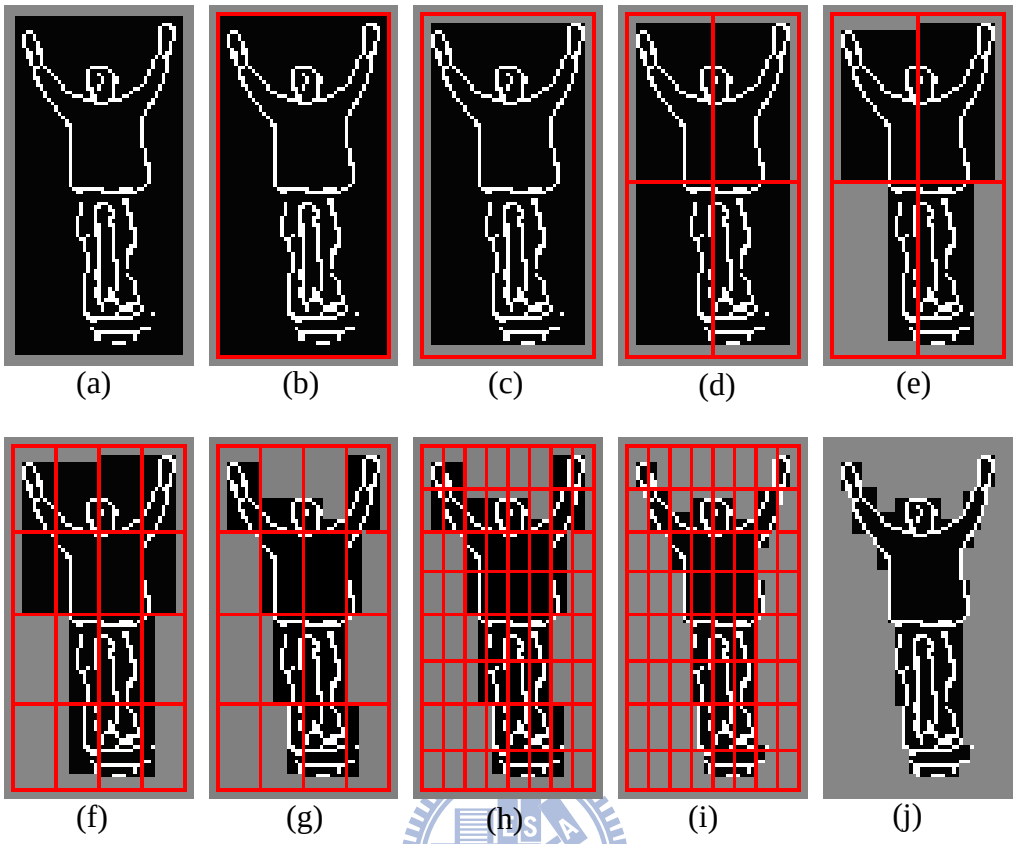


Fig. 2.15 Illustrative images to show the procedure of background region growing, (a) An example image in which background region marked in gray color, edge marked in white color and non-edge region marked in black color, (b) The operating region (whole human region) marked with red frame, (c) The resultant image after applying background region growing, (d) Dividing equally the operating region into four new operating regions. Repeating the above process, we can obtain (e) from (d), (g) from (f) and (i) from (h) respectively, (j) The resultant image.

Step 3: We further divide the operating region into four equal quadrants (see Fig. 2.15(d)), re-apply the above divide-and-conquer procedure, we will obtain the background grown image Fig. 2.15(e).

Step 4: Repeat Step 2 and Step 3 to grow background region we can obtain Figs. 2.15(f)－2.15(i) respectively, until all new operating regions have length or width shorter than threshold preset.

All 64 operating regions in Fig. 2.15(i) have width shorter than threshold which is preset to be 10 in this example, therefore the background region growing is completed and the resultant image is shown is Fig. 2.15(j).

Chapter 3 Human Head Detection

In this chapter, we propose a human head outline extraction method in color images that can be used to extract head outline in different view angles, such as frontal view, lateral view, diagonal view, and so on. Our method uses color Canny edge information in conjunction with skin and hair color to locate heads in the given images. Firstly, we compare the edge map of the given image with the pre-built left head-shape model and right head-shape model to detect head candidates. Detecting with the above two models gives somewhat size tolerance capability in one's head width. Secondly, we compute the occupying proportions of the skin area or hair area with the head candidate area. Namely, we combine the shape matching technique and color matching technique to better estimate the location of a human head.

3.1. Shape Matching

The shape of a human is often very different from the shape of other objects in a scene. Shape-based detection of humans can therefore be a powerful cue. The advances are first of all to allow human detection and tracking in the uncontrolled environments. Due to the method mentioned in the last chapter, reliable silhouette outlines can describe the shape of the humans in the image sequence. Furthermore, the silhouette outline can be advanced to help segmenting human of an image frame.

3.1.1. Building Head-Shape Model

Human heads in different view angles have an outline shape similar to "∩". Because our processing video signal captures a person somewhat far away, we ignore the details of facial features and only consider the outline of the head to build the

head-shape model which can also be adapted to different view angles.

The size of human head changes from person to person, different capturing distance and different view angle. To cope with these variations, we built left head-shape model and right head-shape model separately to give somewhat size tolerance capability in human head width.

The head-shape model is built based on fuzzy set theory. A fuzzy set is a class of points possessing a continuum of membership grades, where there is no sharp boundary among elements that belong to this class and those that do not [11]. We can express this membership grade by a mathematical function called membership function or characteristic function $\mu_A(x_i)$. This function assigns to each element in the set a membership grade in the interval [0, 1]. Let **X** be the universe of discourse, with a generic element denoted by $x_i : \mathbf{X} = \{x_1, x_2, ..., x_m\}$. A fuzzy set **A** in **X** is formally defined as follows:

$$\mathbf{A} = \{(x_i, \mu_{\mathbf{A}}(x_i))\}, \quad x_i \in \mathbf{X} \tag{7}$$

where **A** is characterized by the function $\mu_A(\cdot)$, which associates with each point $x_i \in \mathbf{X}$ a membership grade $\mu_A(x_i) \in [0,1]$. In this work, the S-function is used for modeling the characteristic function. Such a function is defined as follows:

$$\mu_{\mathbf{A}_S}(x) = S(x;a,b) = \begin{cases} 0, & x \le a \\ \dfrac{2(x-a)^2}{(b-a)^2}, & a < x \le \dfrac{(a+b)}{2} \\ 1 - \dfrac{2(x-b)^2}{(b-a)^2}, & \dfrac{(a+b)}{2} < x \le b \\ 1, & b < x \end{cases} \tag{8}$$

where $0 \le a \le 1$, $0 \le b \le 1$, and $a \le b$. The parameters *a* and *b* control the shape of the function (see Fig. 3.1). When *a* is close to *b*, the function will behave like a step function. If *a* is set to a big value, the output of the function will decrease, and if *b* is set to a small value, the output will increase.

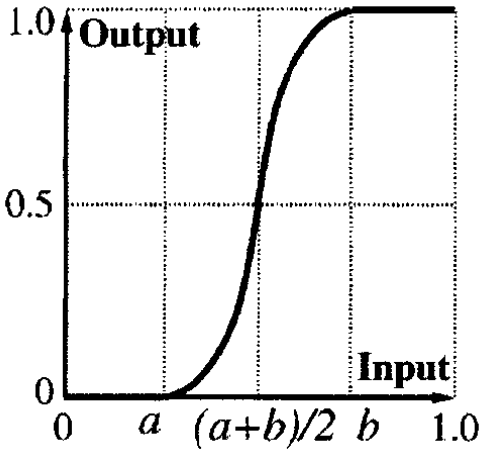


Fig. 3.1 The S-function used for modeling the characteristic function.

The procedure to build head-shape model is listed below:

Step 1: Manually select the head region from the edge map of training database (see Fig. 3.2). The number of all head regions is N_{head}. Each head region is a two dimension data with different length m_k and width n_k:

$$Head_k(i, j) = \begin{cases} 1, & \text{edge} \\ 0, & \text{non - edge} \end{cases} \tag{9}$$

where $1 \le k \le N_{head}$, $1 \le i \le m_k$, $1 \le j \le n_k$.

Step 2: Decide the size $m \times n$ of the head-shape model by computing the mean and standard deviation of the head region width and length. In the implementation, m is set to 8 and n is set to 6.

Step 3: Manually select the left part L_Head and the right part R_Head of head from the head region selected in step 1 (see Fig. 3.2). The boundary box of L_Head and R_Head is located to have maximum number of edge pixels in the second row and second column as well. Build the left head-shape model (or simply LHSM) by calculating pixel-by-pixel mean percent value as follows:

$$LHSM = \frac{1}{N_{head}} \sum_{k=1}^{N_{head}} L\_Head_k \tag{10}$$

Similarly build the right head-shape model (or simply RHSM) from the right part of all head regions:

$$RHSM = \frac{1}{N_{head}} \sum_{k=1}^{N_{head}} R\_Head_k \tag{11}$$

Step 4: To normalize the head-shape models, we use two S type standard functions to renew each LHSM and RHSM.

$$\begin{aligned} LHSM(i,j) &\leftarrow S(LHSM(i,j),a,b) \\ RHSM(i,j) &\leftarrow S(RHSM(i,j),a,b) \end{aligned} \tag{12}$$

Step 5: To ignore the details of facial features and only consider the outline of the head, we manually select the ignored region which is not outline of head, i.e. the inner region and the outer region. Assign a value (in the implementation, the value is set to -1) to each point in the ignored regions and ignore them when we estimate the similarity between the model and an $m \times n$ rectangle region in the edge map E of the input image.

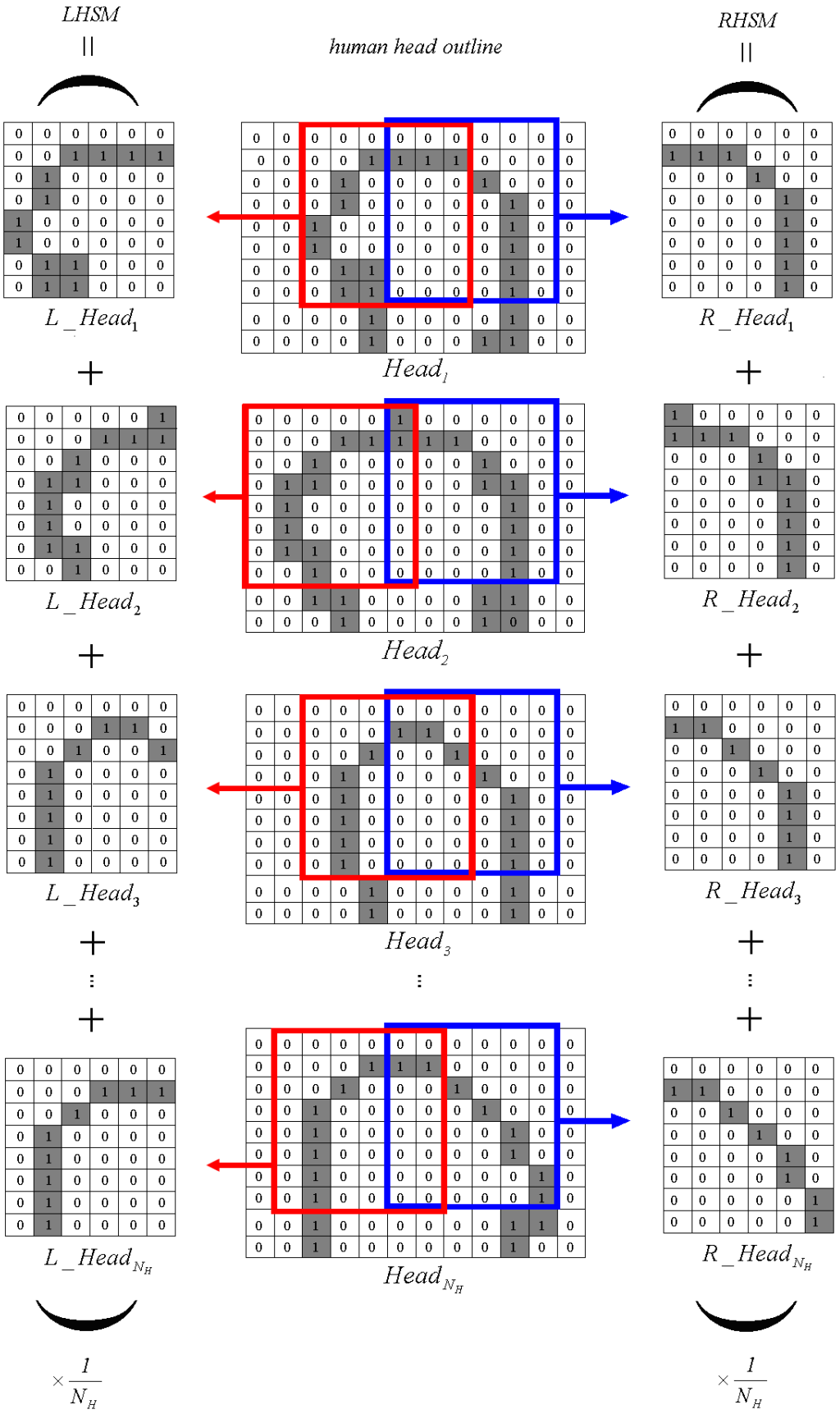


Fig. 3.2 Illustrative images to show the procedure to build left head-shape model and right head-shape model.

3.1.2. Shape Pattern Matching

Let $RG(u, v)$ denote a rectangular region of dimension 8×6 in the edge map E with its upper left pixel at (u,v). To estimate the similarity between the LHSM model and $RG(u, v)$. We compute the sum of absolute difference between each pixel in the LHSM and the corresponding pixel in $RG(u, v)$ except the pixels we want to ignore. The result is recorded to $left\_shape\_score(u,v)$.

$$left\_shape\_score(u,v)=\sum_{i=1}^{m}\sum_{j=1}^{n}\left| RG(u,v;i,\ j)-LHSM(i,\ j)\right| \tag{13}$$

Similarly we compute the sum of absolute difference between each pixel in the RHSM and $RG(u, v)$ except the pixels we want to ignore. The result is recorded to $right\_shape\_score(u,v)$.

$$right\_shape\_score(u,v)=\sum_{i=1}^{m}\sum_{j=1}^{n}\left| RG(u,v;i,\ j)-RHSM(i,\ j)\right| \tag{14}$$

For each $left\_shape\_score$, there is a corresponding right head search region $R(u,v)$ defined as follows:

$$R(u,v)=\{(u',v')\,|\,u-1\le u'\le u+1, v\le v'\le v+6\} \tag{15}$$

The maximum of all $right\_shape\_score$ within the right head search region $R(u,v)$ is called $max\_right\_shape\_score(u,v)$ and indicate the corresponding right head.

$$max\_right\_shape\_score(u,\ v)=\max_{(u',\ v')\in R(u,\ v)} right\_shape\_score(u',\ v') \tag{16}$$

Then the total shape matching score is decided by each $left\_shape\_score$ and $max\_right\_shape\_score$ as follows:

$$shape\_score(u,v)=\frac{1}{2}[left\_shape\_score(u,v)+max\_right\_shape\_score(u,v)] \quad (17)$$

An example of head-shape pattern matching is shown is Fig. 3.3. In Fig. 3.3(a) *RG*(*4*, *3*) denote the rectangular region marked with red frame whose upper left pixel is the pixel marked in pink at $(4,3)$. The corresponding right head search region $R(4,3)$ is marked with green frame. The maximum of all $right\_shape\_score$ within $R(4,3)$ indicates the corresponding right head.

When we built the head-shape model there are two parameters *a* and *b* in the S-function. We let a vary from 0 to 0.2 and b vary from 0.3 to 0.7 and use them to

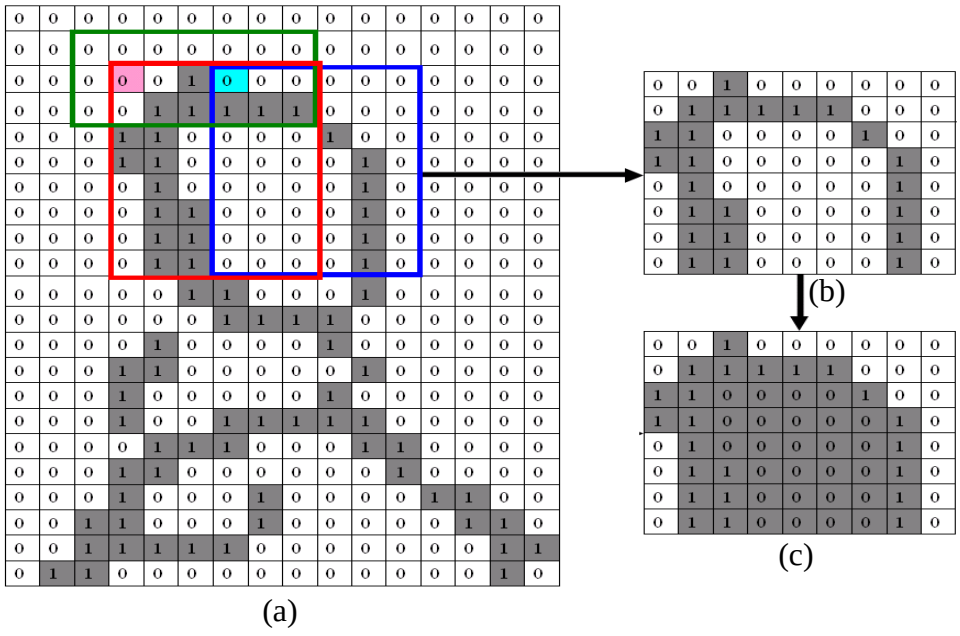


Fig. 3.3 (a) Illustrative images to show the procedure of head shape pattern matching, (b) the head edge found, (c) the head region found.

estimate the training database. The accuracy rate is listed in Table Ⅰ.

TABLE Ⅰ

THE ACCURACY RATE OF HUMAN HEAD DETECTION

FROM SHAPE PATTERN MATCHING

b\a	0	0.1	0.2
0.3	0.75	0.78	0.71
0.4	0.80	0.77	0.70
0.5	0.79	0.74	0.69
0.6	0.79	0.73	0.67
0.7	0.78	0.71	0.67

According to the result listed in Table Ⅰ, the parameters *a* and *b* in the S-function are set to be 0 and 0.4 respectively. From Table Ⅰ, the highest accuracy rate of human head detection using head-shape matching is 0.8.

3.2. Color Matching

Color information is also an important feature for human head extraction. Due to the goal of this research is to detect human head in different view angle, we not only use a skin color detector but also a hair color detector color analysis and the fuzzy theory to combine and then extract the head region candidates. Several color spaces suitable for segmenting the skin-color and hair-color in an image have been proposed. Choosing the representative and discriminative color space for the 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 [12]. In [13], 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 follows:

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \quad (18)$$

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.

3.2.1. Building Skin and Hair Color Models

The terms skin color and hair color are subjective human concepts. Because of this, the color representation should be similar to the color sensitivity of human eyes to obtain a stable output similar to the one given by the human visual system. Such a color representation is called the perceptually uniform color system or simply UCS.

In conventional methods, all visible colors are divided into two groups: skin color and non-skin color. However, consider two colors near the boundary of the skin part. Although the difference between them is almost unnoticeable by a human viewer, one is regarded as ‘skin color’ and the other is not. This is unnatural and is considered as

one of the reasons of instability in conventional methods for skin color detection. We assign a value within [0.0, 1.0] to each point in the color space to indicate how much a visible color looks like the skin color. We call this value as skin color likeness and use a table to describe the skin color likeness of all visible colors. We call it the Skin Color Distribution Model, or simply SCDM. The SCDM is a fuzzy set of skin color. We use a large image set from Weizmann dataset containing faces to derive the distribution of color of the human skin region in order to build the SCDM. The procedure to build the SCDM is as follows:

Step 1: Manually select skin regions in each training image .

Step 2: Prepare a table to record the two dimensional chromatic histogram of skin regions, and initialize all the entries with zero.

Step 3: Convert the chromaticity value of each pixel in the skin regions from RGB color space to YCbCr color space, and then increase the entry of the chromatic histogram corresponding to it by one.

Step 4: Normalize the table by dividing all entries with the greatest entry in the table.

We use a model similar to SCDM to describe the hair color. We call it the Hair Color Distribution Model, or simply HCDM. The HCDM describes the hair color likeness of all visible colors.

3.2.2. Color Pattern Matching

We use SCDM and HCDM to extract the skin color region and the hair color region, respectively. The average in each pixel are the skin/hair color likeness of each pixel in the input image. We call them the Skin Color Similarity Map (or SCSM) and Hair Color Similarity Map (or HCSM).

$$\begin{cases} SCSM = SCS(p) = SCDM(C_b(p), C_r(p)) \\ HCSM = HCS(p) = HCDM(C_b(p), C_r(p)) \end{cases} \quad (19)$$

where $C_b(p)$ and $C_r(p)$ are the chromaticity of pixel p in the input image, $SCS(p)$ and $HCS(p)$ are the skin color likeness and the hair color likeness of pixel p, respectively.

In the case that the skin (or hair) color regions are represented in binary images, the skin (or hair) color area can be estimated by counting the number of skin (or hair) color pixels. Here, we apply a method based on the fuzzy theory to estimate the skin (or hair) proportion from the average SCS and the average HCS of the head region candidates. We called the result $color\_score$ described below:

$$color\_score = \frac{\sum_{p \in region} \max(SCS(p), HCS(p))}{n} \quad (20)$$

where n is the number of pixels in the head region candidate.

We use the method of color pattern matching to estimate the accuracy of the training database and obtain a testing accuracy of 0.57.

3.3. Human Head Detection

We combine the shape and color matching performance in proportion to their respectively testing accuracies. In this way, we can hence locate the human head by the following equation:

$$total\_score = w \times shape\_score + (1-w) \times color\_score \tag{21}$$

The pixels in the possible region which has the highest total score is defined as the head region.

The weight w between *shape_score* and *color_score* is in proportional to the accuracy rates of shape pattern matching and color pattern matching, respectively, as follows:

$$w = \frac{\text{accuracy rate of shape pattern matching}}{\text{accuracy rate of shape pattern matching} + \text{accuracy rate of color pattern matching}} \tag{22}$$

This leads to $w = \frac{0.8}{0.8+0.57} = 0.58$, which is the proportional constant or weight for shape feature. Namely, this implies 0.42 proportional constant for color feature in computing the head matching score of a region.

Chapter 4 Experimental Results

There are two parts in this chapter. The first part deals with human silhouette extraction. The second part is the human head detection. We present experimental results on the Weizmann human action database reported in [14]. The database contains 80 low resolution (180 × 144 pixels resolution at 25 fps) video sequences depicting eight persons, namely, Daria, Denis, Ido, Ira, Lyova, Moshe and Shahar, performing ten actions, i.e., “running”, “walking”, “bending”, “galloping-sideways”, “jumping-forward-on-two-legs”, “jumping-forward-on-one-legs”, “jumping-jack”, “jumping-in-place-on-two-legs”, “waving-two-hands” and “waving-one-hand”. Samples of the successive frames of the activity categories are shown in Fig. 4.1.

4.1. Human silhouette extraction

It should be noted that these database videos are taken with static camera and simple background in outdoor environments with changes in the illumination. Therefore there is no unique background video for all database videos. The database also contains some background videos and provides a lookup table, in which each movement videos are assigned a corresponding background video.

In order to calculate the person segmentation accuracy rate, the ground truth images are obtained by manually extracting human silhouette from each image frame to be tested. Each video contains about 70 frames on average. Due to most actions in the video database are periodical, we produce 20 ground truth images from 20 successive frame images for each person and each action. There are totally 1600 ground truth images produced and some of them are shown in Fig. 4.2.

To calculate the accuracy rate, a minimum region surrounding the human silhouette of the ground truth image and the segmented human silhouette of the

corresponding image in order to void the bias leaning to the background area, which constitute a major portion of an image in general. True positive pixels are these pixels classified to human silhouette both in the ground truth image and in the segmented image; while true negative pixels are these pixels classified as background both in the ground truth image and in the segmented image.

Hence, the true positive rate, true negative rate and the accuracy rate can now be calculated as follows:

$$true\_positive\_rate = \frac{number\ of\ true\ positive\ pixels}{number\ of\ pixels\ classified\ as\ human\ silhouette\ in\ the\ ground\ truth\ image} \tag{23}$$

$$true\_negtive\_rate = \frac{number\ of\ true\ negtive\ pixels}{number\ of\ pixels\ classified\ as\ background\ in\ region\ disscussed\ of\ the\ ground\ truth\ image} \tag{24}$$

$$accuracy\_rate = \frac{true\_positive\_rate + true\_negtive\_rate}{2} \tag{25}$$

We compare our method with the method of median background subtraction which described in Section 4.1.2, W^4 method in gray scale which described in Section 4.1.3 and W^4 method in color scale which described in Section 4.1.4. The results of median background subtraction and W^4 method underwent a noise filter and a shadow filter which described in Section 4.1.5.

Fig. 4.1 Example images from video sequences in the Weizmann human action database [14] which depicting eight persons performing ten actions.

Fig. 4.2 Some ground truth images obtained by manually extracting human silhouette from the image frame to be tested.

4.1.1. Results of our method

In this thesis, we propose a human silhouette extraction method based on temporal differencing, and incorporate a novel background region growing technique for extraction of complete human silhouette without a pre-built background model. We combine the temporal differencing from three successive video frames and the edge image to subtract the outline of motive object in the frame. Some examples of outline images extracted is shown in the second column in Fig. 4.3. The outline of the motive object could not be complete and is a non-closed curve. Hence, we propose a novel background region growing technique which growing the background region and then obtain the human silhouette from incomplete edge image. The resultant human silhouette images are shown in the third column in Fig. 4.3. The average accuracy rate is listed in Table II.

TABLE II

THE ACCURACY RATE OF HUMAN SILHOUETTE EXTRACTION USING OUR METHOD

	bend	jack	jump	pjump	side	skip	run	walk	wave1	wave2	average
daria	0.94	0.93	0.93	0.94	0.91	0.92	0.91	0.90	0.94	0.93	0.92
denis	0.89	0.89	0.92	0.87	0.87	0.90	0.88	0.89	0.92	0.94	0.90
eli	0.93	0.91	0.95	0.90	0.92	0.94	0.92	0.91	0.92	0.92	0.92
ido	0.94	0.92	0.92	0.91	0.91	0.91	0.92	0.92	0.92	0.92	0.92
ira	0.94	0.91	0.90	0.91	0.89	0.89	0.87	0.88	0.92	0.91	0.90
lyova	0.90	0.91	0.89	0.92	0.92	0.92	0.88	0.90	0.94	0.93	0.91
moshe	0.93	0.89	0.92	0.91	0.89	0.89	0.89	0.90	0.72	0.93	0.89
shahar	0.90	0.93	0.84	0.88	0.86	0.93	0.92	0.82	0.86	0.91	0.88
average	0.92	0.91	0.91	0.91	0.90	0.91	0.90	0.89	0.90	0.92	**0.91**

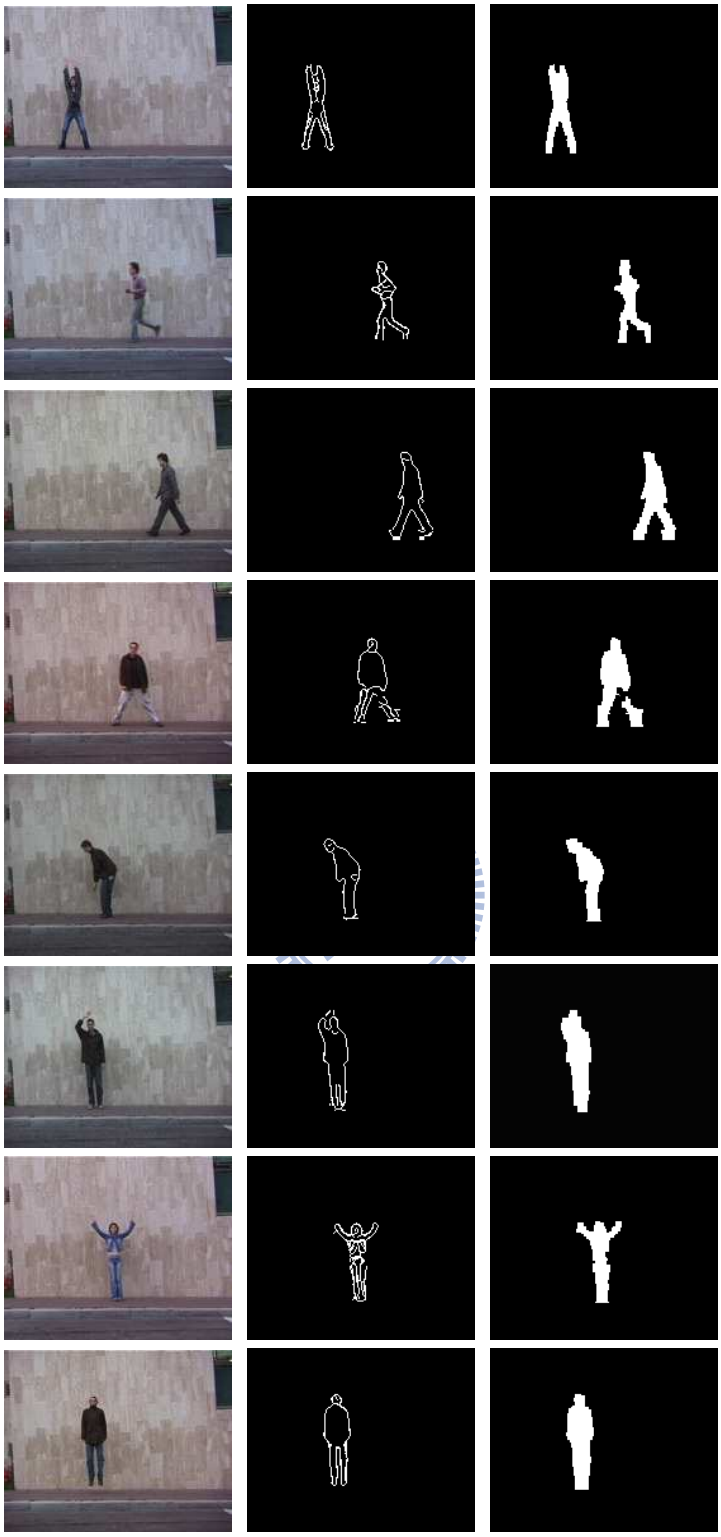


Fig. 4.3 Examples of the resultant images using our method. First column: sample image frames from the Weizmann dataset. Second column: human edge image extracted. Third column: resultant images after background region growing of human edge images extracted.

4.1.2. Median background subtraction

The foreground-background separation method of Weizmann is described briefly in [14] as follows: “To obtain space-time shapes of the actions, we subtracted the median background from each of the sequences and used a simple thresholding in color-space. The resulting silhouettes contained “leaks” and “intrusions” due to imperfect subtraction, shadows and color similarities with the background (see Fig. 4.4 for examples).” According to the statements in [14], we built the median background model and used a simple thresholding in color-space. Let V be an array containing N consecutive images, $V_t(i, j, k)$ be the *k-th* color channel intensity of a pixel (i, j) in the t-th image of V, $\lambda(i, j, k)$ be median value of k-th color channel intensities at pixel (i, j) in all images in V, respectively. The initial background model for a pixel (i, j) is formed by $\lambda(i, j, k)$.

After the training period, an initial background model is obtained. Then, each input image I_t of the video sequence is compared to the background model, and a pixel $I_t(i, j, k)$ is classified as a background pixel if:

$$\begin{aligned} &\left|I_t(i, j, 1) - \lambda(i, j, 1)\right| \le k_m \text{ and} \\ &\left|I_t(i, j, 2) - \lambda(i, j, 2)\right| \le k_m \text{ and} \\ &\left|I_t(i, j, 3) - \lambda(i, j, 3)\right| \le k_m \end{aligned} \tag{26}$$

where k_m is a fixed constant .

The resultant images underwent noise filter and shadow filter described in Section 4.5. Fig. 4.5. shows some resultant images for different threshold values k_m. It is noted that the human silhouette extracted differs from threshold values and there is no unique threshold value suitable for all videos. The line chart of accuracy rate versus threshold value for person and for action is plotted is Fig. 4.6. and Fig. 4.7. The

peak of each curve in the chart pointed to the suitable threshold value which differs from person to person and from action to action. The accuracy rate calculated under k_m=30, a better value from several trials, is listed in Table III.

Fig. 4.4 Examples of video sequences and extracted silhouettes from Weizmann's database [14].

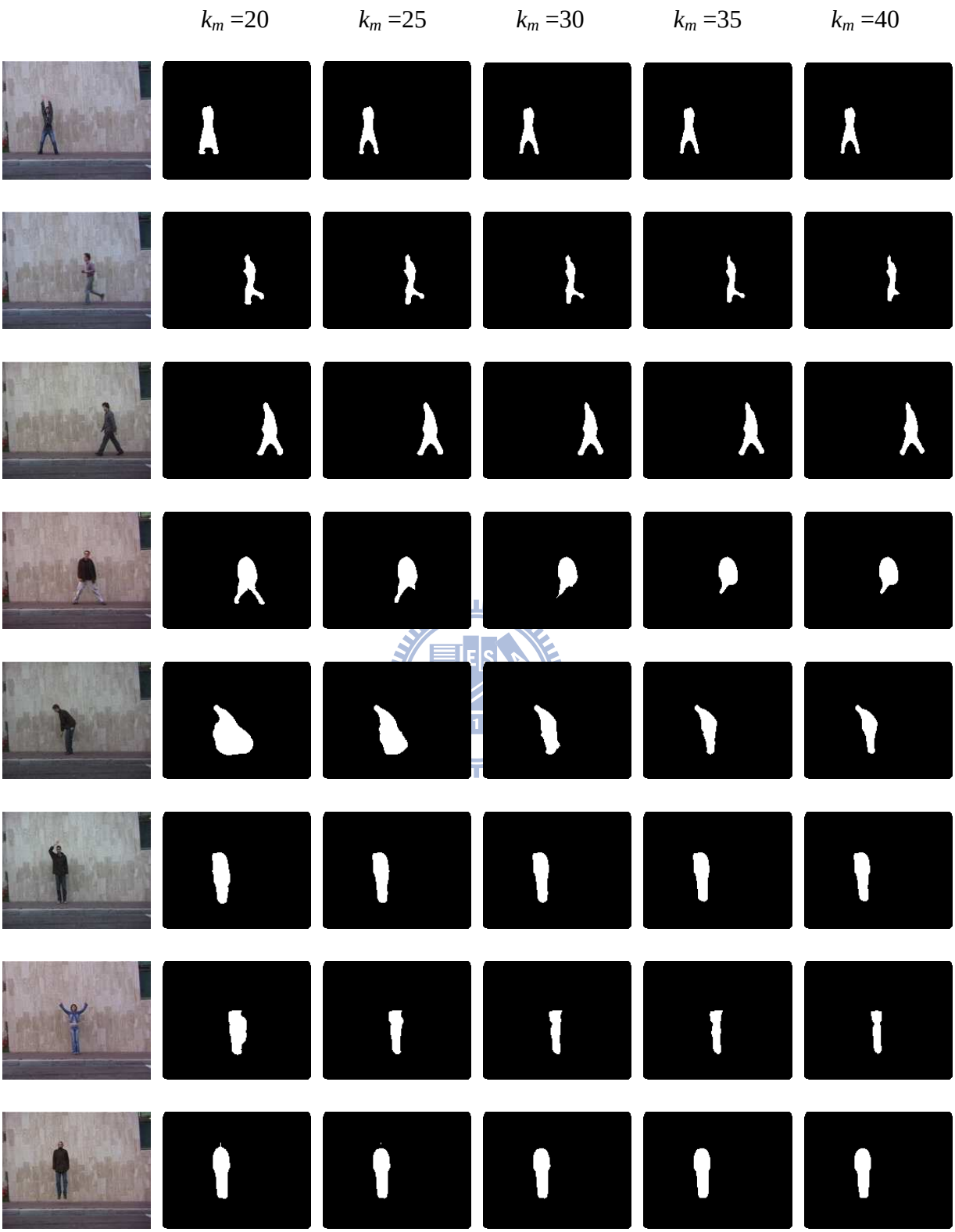


Fig. 4.5 Examples of the resultant images using the median background model and undergoing noise filter and shadow filter for different threshold values k_m.

TABLE III

THE ACCURACY RATE OF HUMAN SILHOUETTE EXTRACTION

USING MEDIAN BACKGROUND MODEL AND k_m =30

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2	average
daria	0.91	0.87	0.94	0.92	0.89	0.91	0.90	0.90	0.89	0.85	0.90
denis	0.89	0.80	0.87	0.80	0.84	0.87	0.84	0.87	0.85	0.83	0.85
eli	0.92	0.88	0.94	0.81	0.90	0.91	0.92	0.90	0.78	0.73	0.87
ido	0.87	0.84	0.90	0.88	0.88	0.91	0.91	0.90	0.86	0.83	0.88
ira	0.90	0.84	0.86	0.87	0.86	0.86	0.81	0.81	0.84	0.81	0.85
lyova	0.88	0.85	0.87	0.90	0.89	0.89	0.89	0.89	0.87	0.85	0.88
moshe	0.89	0.87	0.89	0.88	0.91	0.89	0.91	0.90	0.89	0.86	0.89
shahar	0.79	0.90	0.76	0.80	0.79	0.93	0.93	0.91	0.79	0.77	0.84
average	0.88	0.86	0.88	0.86	0.87	0.90	0.89	0.88	0.85	0.82	**0.87**

4.1.3. W^4 method in gray scale

W^4 uses a model of background variation that is a bimodal distribution constructed from order statistics of background values during a training period, obtaining robust background model even if there are moving foreground objects in the field of view, such as walking people, moving cars, etc [15]. It uses a two stage method based on excluding moving pixels from background model computation. In the first stage, a pixel wise median filter over time is applied to several seconds of video (typically 20-40 seconds) to distinguish moving pixels from stationary pixels (however, our experiments showed that 50 frames ≈ 2 seconds are typically enough for the training period, if not too many moving objects are present). In the second stage, only those stationary pixels are processed to construct the initial background model. Let V be an array containing N consecutive images, $V_k(i, j)$ be the intensity of a pixel (i, j) in the k-th image of V, $\sigma(i, j)$ and $\lambda(i, j)$ be the standard deviation and

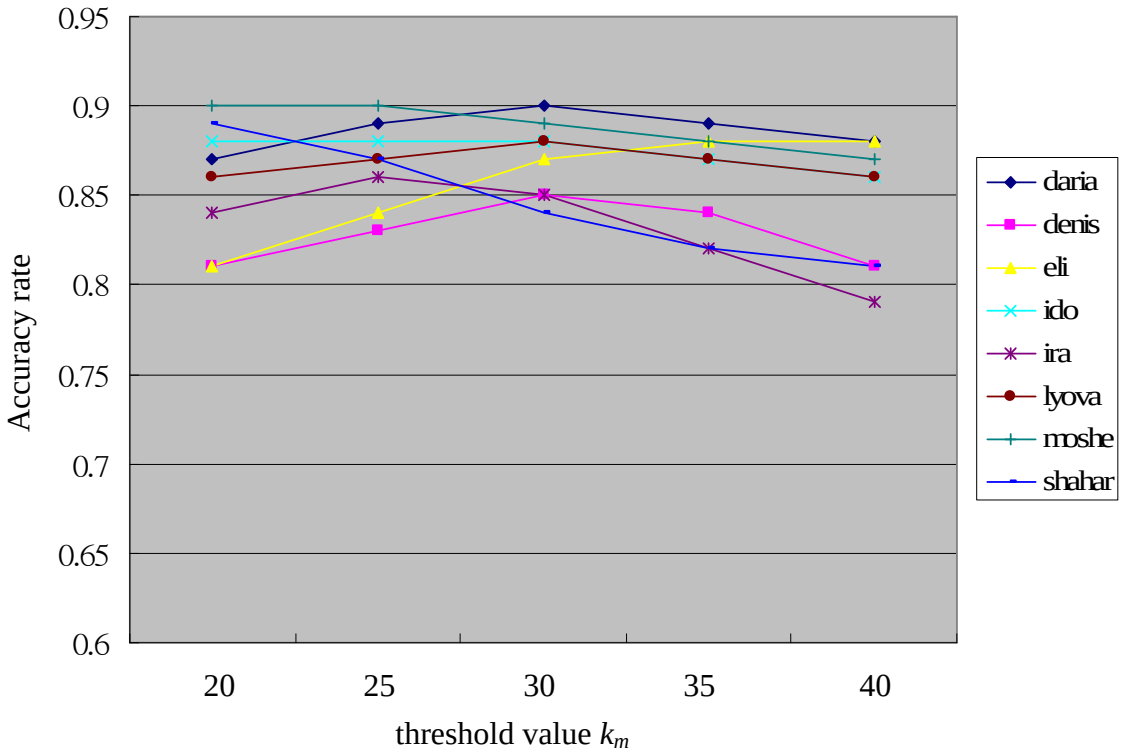


Fig. 4.6 The line chart of accuracy rate versus threshold value for action using median background model.

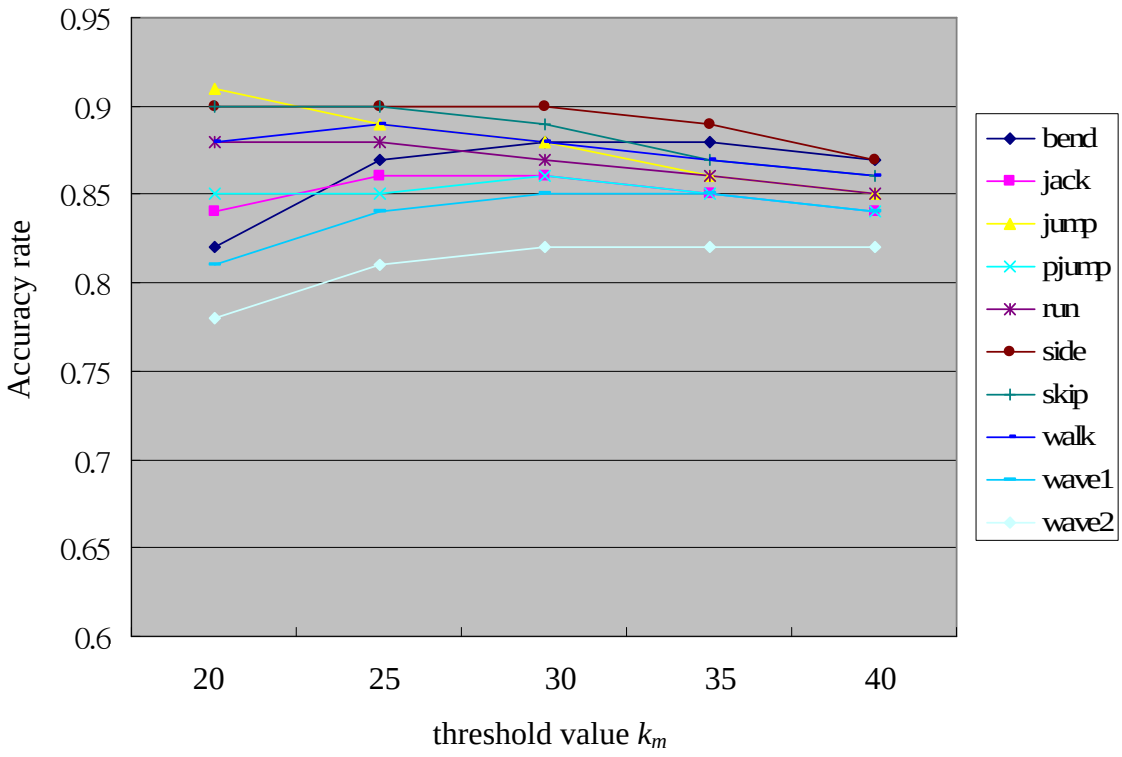


Fig. 4.7 The line chart of accuracy rate versus threshold value for person using median background model.

median value of intensities at pixel (i, j) in all images in V, respectively. The initial background model for a pixel (i, j) is formed by a three-dimensional vector: the minimum $m(i, j)$ and maximum $n(i, j)$ intensity values and the maximum intensity difference $d(i, j)$ between consecutive frames observed during this training period. The background model $\mathbf{B}(i, j) = [m(i, j), n(i, j), d(i, j)]$, is obtained as follows:

$$\begin{bmatrix} m(i,j) \\ n(i,j) \\ d(i,j) \end{bmatrix} = \begin{bmatrix} \min\limits_{z} V_z(i,j) \\ \max\limits_{z} V_z(i,j) \\ \max\limits_{z} \left| V_z(i,j) - V_{z-1}(i,j) \right| \end{bmatrix} \tag{27}$$

where z are frames satisfying $\left| V_z(i,j) - \lambda(i,j) \right| \le 2\sigma(i,j)$. This condition guarantees that only stationary pixels are computed in the background model, i.e., $V_z(i, j)$ is classified as a stationary pixel.

After the training period, an initial background model $\mathbf{B}(i, j)$ is obtained. Then, each input image I_t of the video sequence is compared to $\mathbf{B}(i, j)$, and a pixel $I_t(i, j)$ is classified as a background pixel if:

$$|I_t(i,j) - m(i,j)| \le k_g \mu \quad \text{or} \quad |I_t(i,j) - n(i,j)| \le k_g \mu \tag{28}$$

where μ is the median of the largest interframe absolute difference image $d(i, j)$, and k_g is a fixed constant (the authors suggested the value $k_g = 2$).

The improvement method classify $It(i, j)$ as a foreground pixel if:

$$I_t(i,j) > (m(i,j) - k_g \mu) \quad \text{and} \quad I_t(i,j) < (n(i,j) + k_g \mu) \tag{29}$$

The resultant images underwent noise filter and shadow filter described in

Section 4.5. Fig. 4.8. shows some resultant images for different threshold values k_g. The line chart of accuracy rate versus threshold value for person and for action is plotted is Fig. 4.9. and Fig. 4.10. The peak of each curve concentrated in the chart and pointed to a suitable threshold value. The accuracy rate calculated under k_g=4, a better value from several trials, is listed in Table IV.

TABLE IV

THE ACCURACY RATE OF HUMAN SILHOUETTE EXTRACTION USING W^4 BACKGROUND MODEL IN GRAY SCALE AND k_g=4

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2	average
daria	0.94	0.93	0.96	0.96	0.95	0.95	0.94	0.95	0.94	0.92	0.94
denis	0.89	0.87	0.88	0.86	0.85	0.90	0.88	0.89	0.89	0.88	0.88
eli	0.95	0.93	0.94	0.89	0.93	0.95	0.95	0.94	0.93	0.92	0.93
ido	0.92	0.92	0.91	0.92	0.90	0.91	0.91	0.91	0.92	0.90	0.91
ira	0.92	0.90	0.84	0.90	0.86	0.87	0.84	0.85	0.91	0.90	0.88
lyova	0.94	0.91	0.89	0.94	0.77	0.92	0.91	0.92	0.93	0.92	0.91
moshe	0.93	0.90	0.94	0.93	0.90	0.93	0.94	0.93	0.95	0.93	0.93
shahar	0.89	0.93	0.82	0.88	0.83	0.96	0.96	0.95	0.90	0.89	0.90
average	0.92	0.91	0.90	0.91	0.87	0.92	0.92	0.92	0.92	0.91	**0.91**

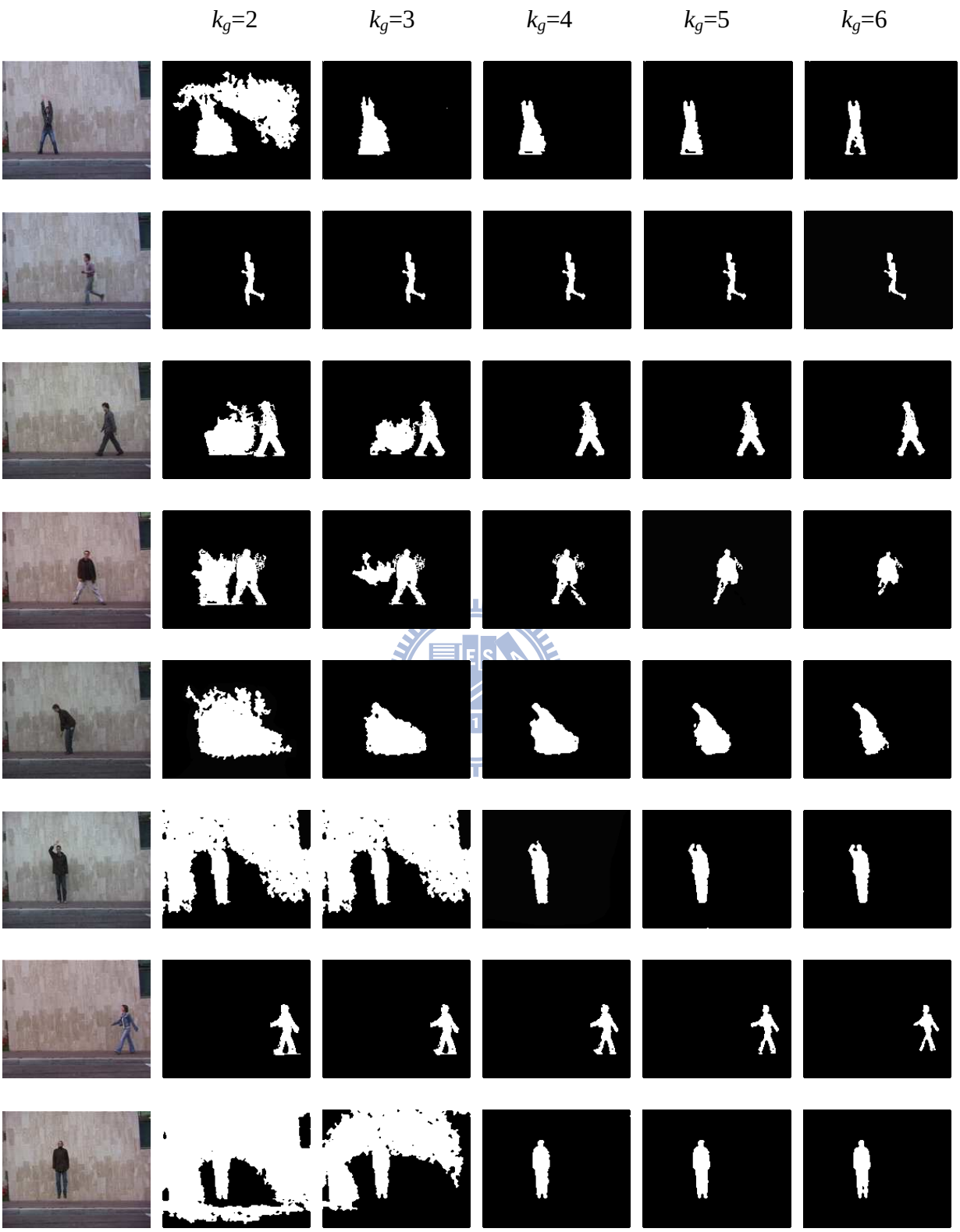


Fig. 4.8 Examples of the resultant images using the W^4 background model in gray scale and undergoing noise filter and shadow filter for different threshold values k_g.

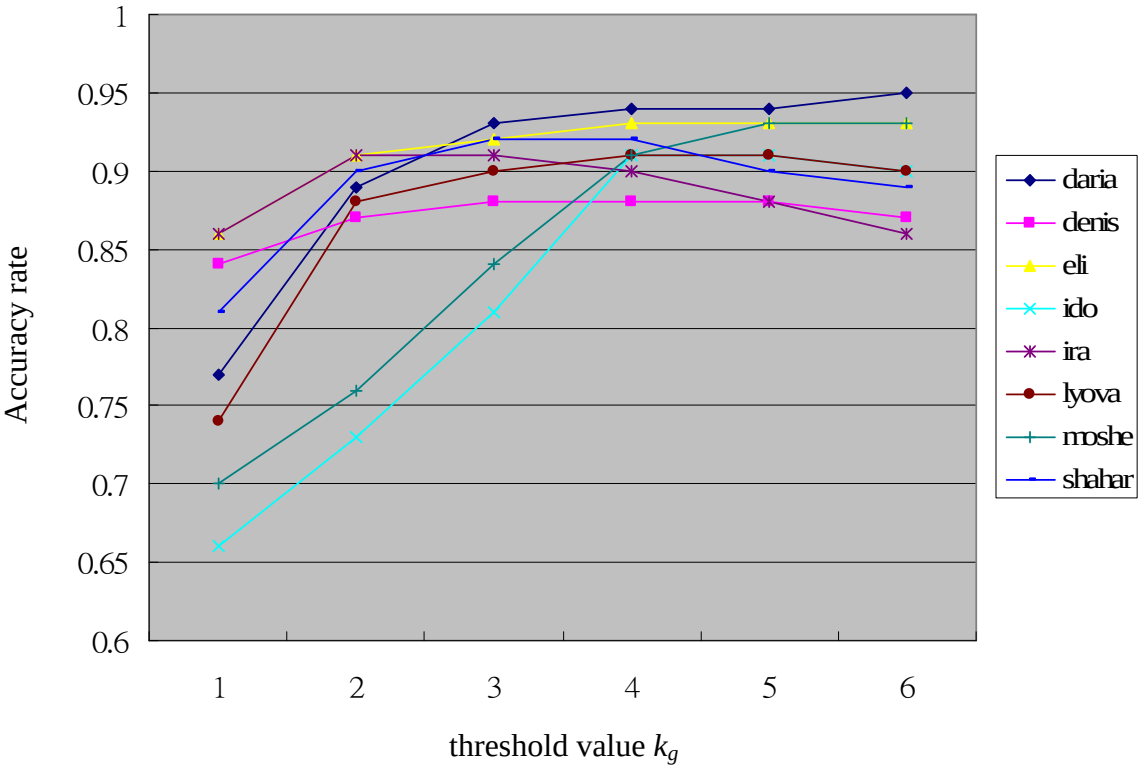


Fig. 4.9 The line chart of accuracy rate versus threshold value for person using W^4 background model in gray scale.

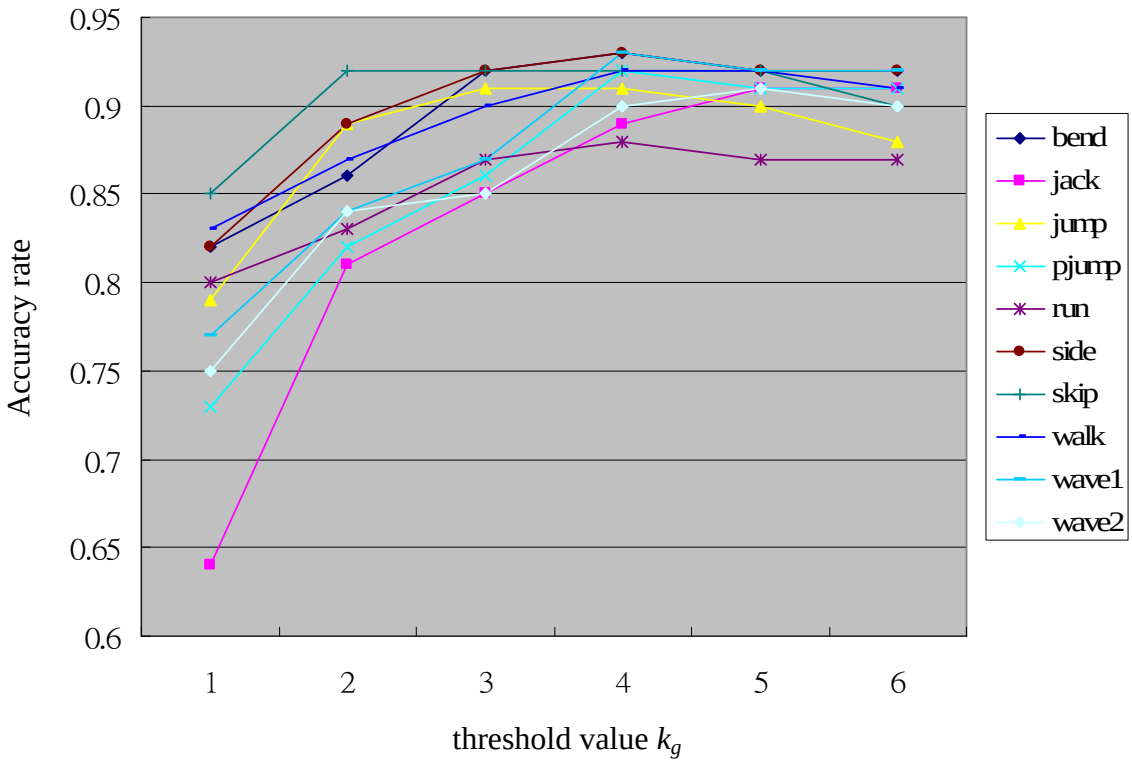


Fig. 4.10 The line chart of accuracy rate versus threshold value for action using W^4 background model in gray scale.

4.1.4. W^4 method in color scale

We spread W^4 method to color scale whose background model contains three color channels. Let V be an array containing N consecutive images, $V_t(i, j, h)$ be the h-th color channel intensity of at pixel (i, j) in the t-th image of V, $\sigma(i, j, h)$ and $\lambda(i, j, h)$ be the standard deviation and median value of the h-th color channel intensities at pixel (i, j) in all images in V, respectively. The initial background model for a pixel (i, j) is formed by the minimum $m(i, j, h)$ and maximum $n(i, j, h)$ intensity values and the maximum intensity difference $d(i, j, h)$ between consecutive frames observed during this training period. The background model $\mathbf{B}(i, j, h) = [m(i, j, h), n(i, j, h), d(i, j, h)]$, is obtained as follows:

$$\begin{bmatrix} m(i,j,h) \\ n(i,j,h) \\ d(i,j,h) \end{bmatrix} = \begin{bmatrix} \min_z V_z(i,j,h) \\ \max_z V_z(i,j,h) \\ \max_z \left| V_z(i,j,h) - V_{z-1}(i,j,h) \right| \end{bmatrix} \tag{30}$$

where z are frames satisfying $\left|V_z(i, j, h) - \lambda(i, j, h)\right| \le 2\sigma(i, j, h)$. This condition guarantees that only stationary pixels are computed in the background model.

After the training period, an initial background model $\mathbf{B}(i, j, h)$ is obtained. Then, each input image I_t of the video sequence is compared to $\mathbf{B}(i, j, h)$, and a pixel is classified as a background pixel if:

$$I_t(i, j, h) > (m(i, j, h) - k_c\mu(h)) \quad \text{and} \quad I_t(i, j, h) < (n(i, j, h) + k_c\mu(h)) \tag{31}$$

where $\mu(h)$ is the median of the largest interframe absolute difference image $d(i, j, h)$, and k_c is a fixed constant.

The resultant images underwent noise filter and shadow filter described in Section 4.5. Fig. 4.11. shows some resultant images for different threshold values k_c. The line chart of accuracy rate versus threshold value for person and for action is plotted is Fig. 4.12. and Fig. 4.13. The peak of each curve concentrated in the chart and pointed to a suitable threshold value. The accuracy rate calculated under k_c=5, a better value from several trials, is listed in Table V.

TABLE V

THE ACCURACY RATE OF HUMAN SILHOUETTE EXTRACTION USING W^4 BACKGROUND MODEL IN COLOR SCALE AND k_c=5

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2	average
daria	0.80	0.82	0.94	0.93	0.91	0.93	0.91	0.91	0.83	0.76	0.88
denis	0.76	0.78	0.82	0.75	0.81	0.84	0.80	0.82	0.77	0.77	0.79
eli	0.82	0.51	0.64	0.79	0.78	0.91	0.91	0.84	0.78	0.73	0.77
ido	0.92	0.90	0.83	0.92	0.72	0.80	0.79	0.77	0.91	0.89	0.85
ira	0.81	0.81	0.82	0.83	0.83	0.83	0.80	0.77	0.82	0.79	0.81
lyova	0.82	0.91	0.83	0.90	0.64	0.84	0.82	0.80	0.84	0.82	0.82
moshe	0.94	0.89	0.88	0.93	0.88	0.88	0.88	0.87	0.95	0.87	0.90
shahar	0.83	0.62	0.76	0.91	0.74	0.74	0.58	0.88	0.89	0.91	0.79
average	0.84	0.78	0.81	0.87	0.79	0.85	0.81	0.83	0.85	0.82	**0.83**

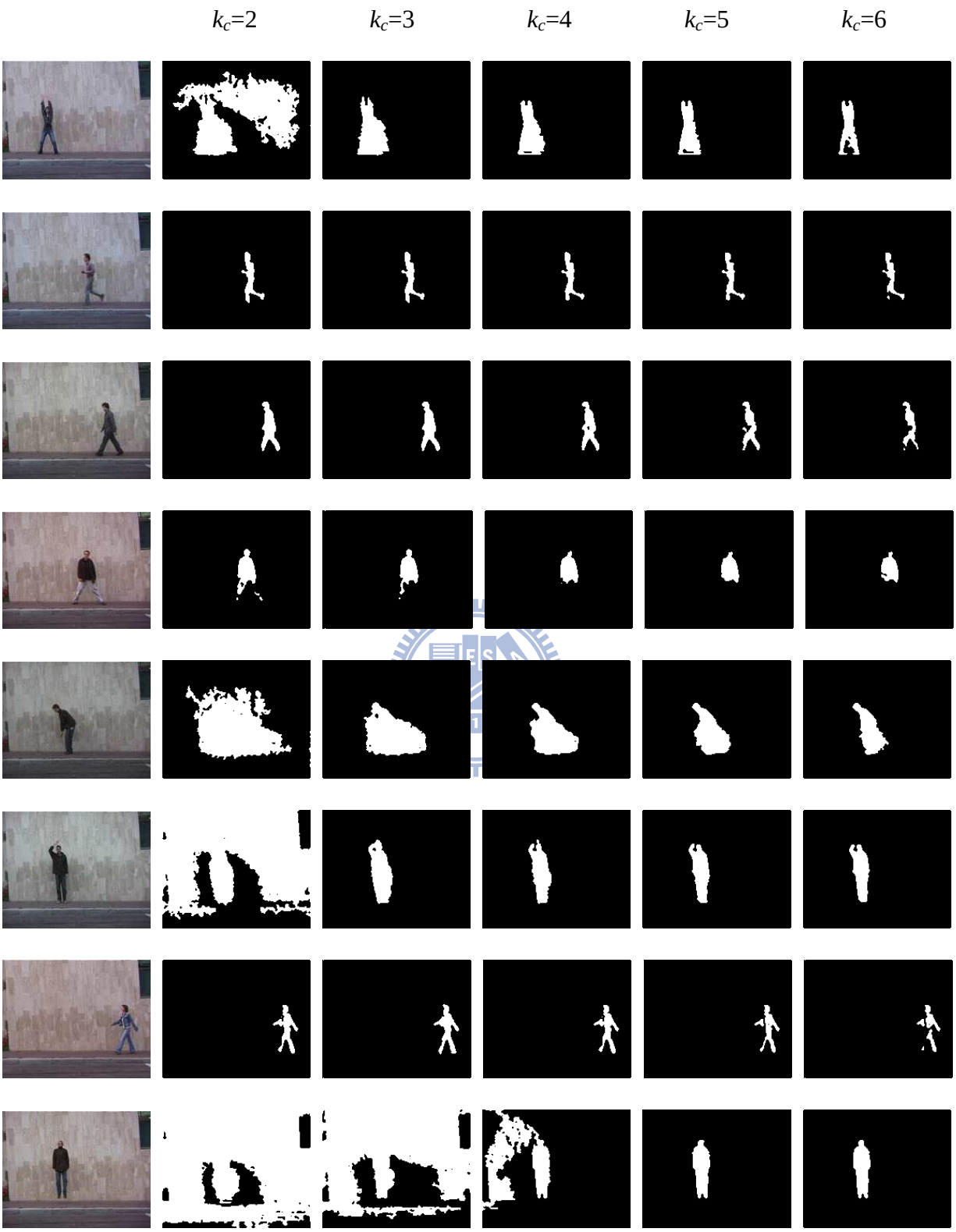


Fig. 4.11 Examples of the resultant images using the W^4 background model in color scale and undergoing noise filter and shadow filter for different threshold values k_c.

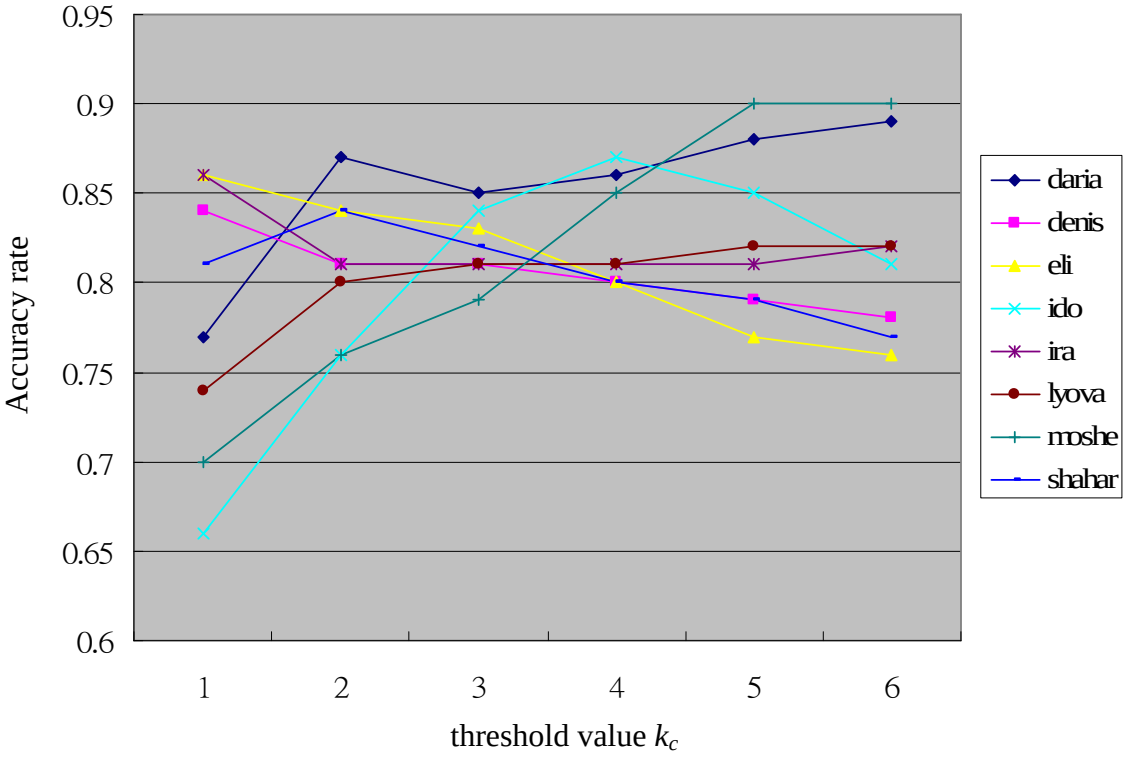


Fig. 4.12 The line chart of accuracy rate versus threshold value for person using W^4 background model in color scale.

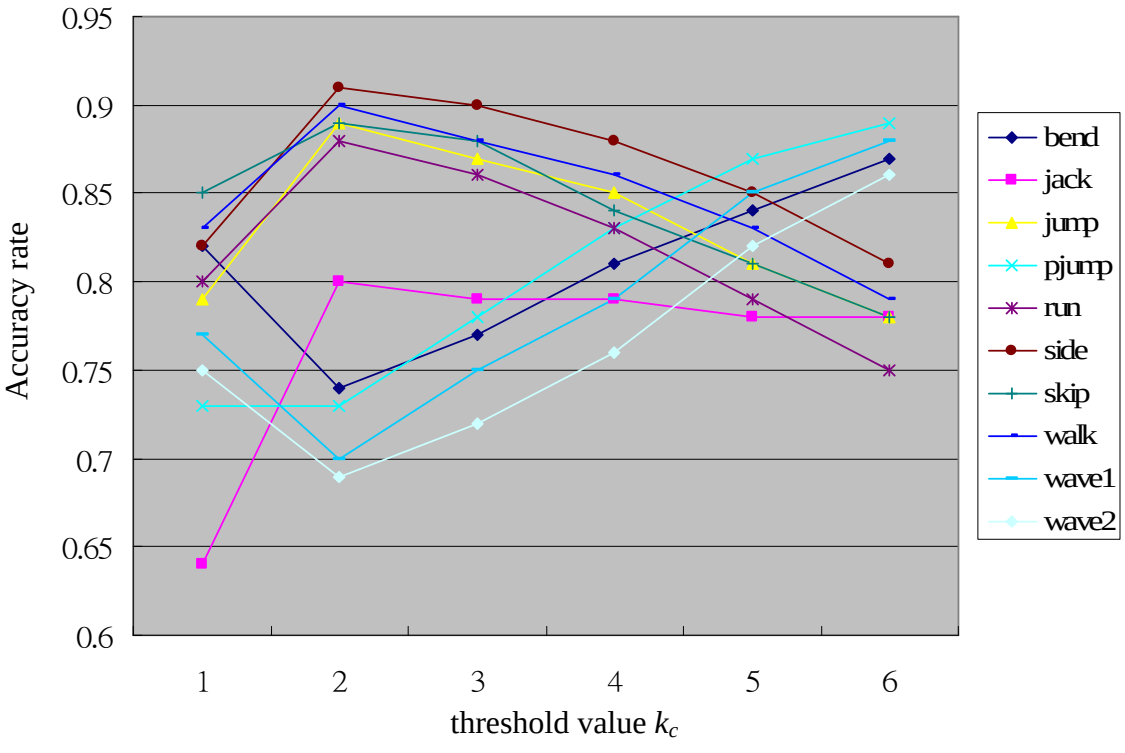


Fig. 4.13 The line chart of accuracy rate versus threshold value for action using W^4 background model in color scale.

4.1.5. Noise filter and shadow filter

The resultant silhouettes of median background subtraction and W^4 method contained “leaks” and “intrusions” due to imperfect subtraction, shadows and color similarities with the background. Therefore, the resultant images underwent a noise filter and a shadow filter described below.

In noise region, we apply a “majority vote” method to remove salt-and-pepper noise and fill the leaks inside human silhouettes. Let $I(i, j)$ be an resultant image in which $I(i, j)=1$ if pixel (i, j) belonging to the foreground and $I(i, j)=0$ if pixel (i, j) belonging to the background. For each pixel (i, j), consider a $(2N+1)\times(2N+1)$ template N_{ij} such that $N_{ij}(n,m) = I(i+n, j+m)$, for $-N \le n \le N$, $-N \le m \le N$ (i.e. N_{ij} corresponds to a neighborhood of pixel (i, j)). If the sum of every elements in N_{ij} is larger than $(2N+1)\times(2N+1)\times0.5$ (i.e. foreground pixels are the majority in the neighborhood of pixel (i, j)), the value of $I(i, j)$ is set 1 which means pixel (i, j) belonging to the foreground. Similarly, if the sum of every elements in N_{ij} is less than $(2N+1)\times(2N+1)\times0.5$ (i.e. background pixels are the majority in the neighborhood of pixel (i, j)), the value of $I(i, j)$ is set 0 which means pixel (i, j) belonging to the background. In the implementation, N is set to 1.

In shadowed regions, it is assumed that the observed intensity of shadow pixels is directly proportional to incident light; consequently, shadowed pixels are scaled versions (darker) of corresponding pixels in the background model. The normalized crosscorrelation (NCC) is used as an initial step for shadow detection, and refine the process using local statistics of pixel ratios [16].

Let $B(i, j)$ be the background image formed by temporal median filtering, and $I(i, j)$ be an image of the video sequence. For each pixel (i, j) belonging to the foreground, consider a $(2N+1)\times(2N+1)$ template T_{ij} such that $T_{ij}(n,m) = I(i+n, j+m)$, for $-N \le n$

$\leq N$, $-N \leq m \leq N$ (i.e. T_{ij} corresponds to a neighborhood of pixel (i, j)). Then, the NCC between template T_{ij} and image B at pixel (i, j) is given by:

$$NCC(i, j) = \frac{ER(i, j)}{E_B(i, j)E_{T_{i,j}}}, \tag{32}$$

where

$$\begin{aligned} ER(i, j) &= \sum_{n=-N}^{N} \sum_{m=-N}^{N} B(i+n, j+m)T_{ij}(n, m), \\ E_B(i, j) &= \sqrt{\sum_{n=-N}^{N} \sum_{m=-N}^{N} B(i+n, j+m)^2}, \quad \text{and} \\ E_{T_{if}} &= \sqrt{\sum_{n=-N}^{N} \sum_{m=-N}^{N} T_{ij}(n, m)^2}. \end{aligned} \tag{33}$$

For a pixel (i, j) in a shadowed region, the NCC in a neighboring region T_{ij} should be large (close to one), and the energy $E_{T_{if}}$ of this region should be lower than the energy $E_B(i, j)$ of the corresponding region in the background image. Thus, a pixel (i, j) is pre-classified as shadow if:

$$NCC(i, j) \geq L_{\text{ncc}} \quad \text{and} \quad E_{T_{if}} < E_B(i, j), \tag{34}$$

where L_{ncc} is a fixed threshold (the authors suggested the value $L_{\text{ncc}} = 0.95$ and $N = 4$).

The NCC provides a good initial estimate about the location of shadowed pixels, by detecting pixels for which the surrounding neighborhood is approximately scaled with respect to the reference background. However, some background pixels related to valid moving objects may be wrongly classified as shadow pixels. To remove such false positives, a refinement stage is applied.

The proposed refinement stage consists of verifying if the ratio $I(i, j)/B(i, j)$ in a neighborhood around each shadow pixel candidate is approximately constant, by computing the standard deviation of $I(i, j)/B(i, j)$ within this neighborhood. More specifically, we consider a region R with $(2M+1)\times(2M+1)$ pixels (we used $M = 1$ in all experiments) centered at each shadow pixel candidate (i, j), and classify it as a shadow pixel if:

$$\mathrm{std}_R\left(\frac{I(i, j)}{B(i, j)}\right) < L_{\mathrm{std}} \quad \text{and} \quad L_{\mathrm{low}} \le \left(\frac{I(i, j)}{B(i, j)}\right) < 1, \tag{35}$$

where $\mathrm{std}_R\left(\frac{I(i, j)}{B(i, j)}\right)$ is the standard deviation of quantities $I(i, j)/B(i, j)$ over the region R, and L_{std}, L_{low} are thresholds suggested to be 0.05 and 0.5 respectively.

4.2. Human head detection

We propose a human head outline extraction method in color images that can be used to extract head outline in different view angles, such as frontal view, lateral view, diagonal view, and so on. The human silhouette extraction method is taken as a pre-processing step before human head detection, which can simplify the complex backgrounds and reduce the detecting area. Then we propose a fuzzy theory based pattern-matching technique which combines the shape and color information to locate human head.

We present experimental results of human head detection on the same testing images as human silhouette extraction which contains eight persons performing ten actions. There are totally 1600 testing images and 1491 of them detected head correctly. The accuracy rate for head detection is $\frac{1491}{1600} = 93\%$.

Chapter 5 Conclusion

In this thesis, we propose a human silhouette extraction method based on temporal differencing, and incorporate a novel background region growing technique for extraction of complete human silhouette without a pre-built background model. The proposed method adapts quickly to changes in the scene and can extract human silhouette from incompletely controlled environment (outdoor or indoor with change of illumination). We combine the temporal differencing from three successive video frames and the edge image to subtract the outline of motive object in the frame. The outline of the motive object could not be complete and is a non-closed curve. Hence, we propose a novel background region growing technique which gradually grows the background region and then obtain the foreground silhouette from incomplete edge image.

We also propose a human head outline extraction method in color images that can be used to extract head outline in different view angles, such as frontal view, lateral view, diagonal view, and so on. We take temporal differencing method as a pre-processing step before human head detection, which can simplify the complex backgrounds and reduce the detecting area. Then we propose a fuzzy theory based pattern-matching technique which combines the shape and color information to locate human head.

Experiment results have shown that our approach can extract human silhouette without pre-built background model and have good accuracy rate competitive to those by the background subtraction methods. Experiment results have also shown that our approach can also obtain good results on human head detection.

To investigate further, extracting multiple (occluded) people from more complicated scene without a pre-built background model is our future work.

References

[1] Thomas B. Moeslund , Adrian Hilton , Volker Krüger, “A survey of advances in vision-based human motion capture and analysis,” *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, November 2006.

[2] C. Stauffer, W.E.L. Grimson, “Adaptive background mixture models for real-time tracking,” in *Proc. Computer Vision and Pattern Recognition*, vol. 2, pp. 246–252, August 1999.

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

[4] S. Y. Lee, Y. K. Ham and R. -H. Park, “Recognition of human front faces using knowledge-based feature extraction and neurofuzzy algorithm,” *Pattern Recognition*, vol. 29, issue 11, pp. 1863–1876, November 1996.

[5] Choong Hwan LEE, Jun Sung Kim and Kyu Ho Park, “Automatic human face location in a complex background using motion and color information,” *Pattern Recognition*, vol. 29, issue 11, pp. 1877–1889, November 1996.

[6] Eli Saber and A. Murat Tekalp, “Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions,” *Pattern Recognition*, vol. 19, issue 8, pp. 669–680, June 1998.

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

[8] Maylor K. Leunga and Yee-Hong Yang, “Human body motion segmentation in a complex scene,” *Pattern Recognition*, vol. 20, issue 1, pp. 55–64, 1987.

[9] R. Jain, “Extraction of motion information from peripheral processes,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 3, no. 5, pp. 489–503,

1981.

[10] R. Jain and H.-H. Nagel, “On the analysis of accumulative difference pictures from image sequences of real-world scenes,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 206–214, 1979.

[11] L. A. Zadeh, “Fuzzy sets,” *Inform. Control*, vol. 8, pp. 338–353, 1965.

[12] 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.

[13] 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.

[14] B. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, “Actions as space-time shapes,” in *Proc. Computer Vision*, vol. 2, pp. 1395–1402, Oct , 2005.

[15] I. Haritaoglu, D. Harwood, and L. Davis, “W4: Real-time surveillance of people and their activities,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, August 2000.

[16] Julio Cezar Silveira Jacques Jr, C. R. Jung, and S. R. Musse, “Background subtraction and shadow detection in grayscale video sequences,” in *Proc. Computer Graphics and Image Processing*, 2005.