International Journal of Information Acquisition
Vol. 4, No. 2 (2007) 1–22

World Scientific
www.worldscientific.com

ROBUST HUMAN DETECTION AND IDENTIFICATION BY USING STEREO AND THERMAL IMAGES IN HUMAN ROBOT INTERACTION

F. GUAN[†], L. Y. LI[‡], S. S. GE[†,*] and A. P. LOH[†]
[†]*Department of Electrical and Computer Engineering*
National University of Singapore
Singapore 117576
[‡]*Institute for Infocomm Research (I^2R)*
21 Heng Mui Keng Terrace
Singapore 119613

Received 2 April 2007
Accepted 25 July 2007

In this paper, robust human detection is investigated by fusing the stereo and infrared thermal images for effective interaction between humans and socially interactive robots. A scale-adaptive filter is first designed for the stereo vision system to detect human candidates. To eliminate the difficulty of the vision system in distinguishing human beings from human-like objects, the infrared thermal image is used to solve the ambiguity and reduce the illumination effect. Experimental results show that the fusion of these two types of images gives an improved vision system for robust human detection and identification, which is a most important and essential component of human robot interaction.

Keywords: Stereo images; thermal images.

1. Introduction

With the rapid developments in the fields of sensing technologies and robotics, there is an increasing need to employ intelligent robots with a friendly human robot interaction in a human environment for the purpose of surveillance, monitoring, services and so forth [Ge and Fua, 2005]. This has led to recent efforts by researchers worldwide in the area of social robotics [Breazeal, 2003]. Social robotics is the study of robots that are able to interact and communicate between themselves, with humans, and with the environment, within the social and cultural structure attached to its role [Ge, 2007]. Unlike industrial robots, these socially interactive robots are specifically developed to interact with humans socially. It is crucial that socially interactive robots understand, perceive, respond appropriately, and even adapt their behavior based on the cues from human partners. Therefore, effective human robot interaction is a critical component of socially interactive robots.

*Corresponding author. Tel: (65) 6516-6821; Fax. (65) 6779-1103; E-mail: elegesz@nus.edu.sg.

For effective interaction between humans and socially adept, intelligent service robots, a key capability required is the successful interpretation of visual data. One of the most important and essential aspects of human robot visual interaction is an intelligent visual perception system to robustly detect and identify human beings with a variable background. It involves human/face detection and the fusion of stereo and infrared vision on board social robots with greater flexibility and robustness [Loh *et al.*, 2004; Ge *et al.*, 2006], for the purposes of attention focusing and for synthesizing more complex social interaction concepts, like comfort zones, into the robots. As a crucial component of socially interactive robots, a surge of research on human robot interaction has focused on visual human detection activities in the literatures.

Without a prior knowledge and constraints on the background, it is difficult to detect human beings using monocular camera [Mohan *et al.*, 2001; Haritaoglu *et al.*, 2000]. Normally, a background image is used as a reference image to represent a monitored environment, where the objects in the environment are assumed to be stationary. Then, subsequent images are compared with the background image and the object of interest (OOI) can be extracted by using layered methods [Luca *et al.*, 2002]. The OOI is assumed to be contained in a subimage or layer that is superimposed to the background image. As such, in order to obtain the background knowledge, we need the camera to be stationary or a model of the background [Li *et al.*, 2004; Gavrila *et al.*, 1999; Wren *et al.*, 1997; McKenna *et al.*, 2000]. An alternative to detect human objects from images is face detection [Zhang *et al.*, 2004; Hsu *et al.*, 2002], which does not require a stationary background. However, it is only applicable to the front views of human beings. Thus, it still remains a challenge for a mobile robot to identify people around it. In comparison to monocular systems, stereovision systems are able to provide depth or scale information to the detection of human beings in a clustered environment, and enable the spatial object segmentation and also brings about the challenges, e.g. distinguishing humans from human-like objects, in the realization of human detection as a result of this new piece of information.

On the other hand, infrared thermal imaging has received much attention in the literatures [Kakuta *et al.*, 2002; Nanadhakumar and Aggarwal, 1988; Chan *et al.*, 2000; Maldague *et al.*, 1994; Trivedi *et al.*, 2004]. due to the fact that infrared thermal sensors are only sensitive to objects (e.g. humans and fire) that are able to generate heat, and are not affected by the level of illumination. In order to detect pedestrians and reduce accidents at night, a pair of infrared thermal cameras were used to develop a night-vision system in [Tsuji *et al.*, 2002]. Pedestrians are detected by processing the stereo infrared thermal images and their motions are estimated with respect to the vehicle where the night-vision system is mounted. Due to the lack of texture information about OOI, infrared thermal images can be combined with other sensory data to provide robust object detection such as the fusion of visual and thermal images [Waxman *et al.*, 1995]. A forest fire detection system was proposed in [Arrue *et al.*, 2000] by using infrared thermal image, visual image and expert knowledge about geographical information and human activities. The fire candidates are first provided by an infrared thermal image and then verified by using the expert knowledge and the extracted visual information.

In this paper, we are to investigate the human detection performance of a sensor suite, which consists of a stereo rig and an infrared thermal camera installed on a mobile robot. Since the visual cameras and the infrared thermal camera observe objects from different spectrum windows, these two types of sensors provide different features of the same objects simultaneously, which makes us obtain more comprehensive understanding of the objects. The method proposed in this paper fuses spectral, stereo and thermal features for robust human detection, which may provide an efficient way to deal with the challenges of human detection based on only one or two types of the features, e.g. human objects in clustered backgrounds, partial occlusion of multiple persons, great variations of human scales, and low false positives in complex environments. The main

contributions of the paper are as follows:

(i) The human features identified from the stereo-based techniques are fused with those identified from the thermal-based approach. As such, pseudo human objects can be filtered out easily and human beings are identified in virtual and the corresponding physical 3D environment.
(ii) A depth-oriented scale-adaptive filter is proposed to aggregate and enhance the salient features associated with human beings in a Y_I–D plane, which allows disparity transfer and spatial human segmentation.
(iii) Human objects are verified by using a deformable head-shoulder template based on the stereo and edge information.
(iv) Thermal images are filtered by an effective filter bank to highlight heat-generating objects such as face, hair. The calibration between the stereo rig and the infrared thermal camera is provided.

The rest of this paper is organized as follows: Sec. 2 describes the configuration of the proposed vision system. Human detection using stereo vision and infrared thermal imaging are analyzed in Secs. 3 and 4, respectively. A fusion algorithm is proposed in Sec. 5 to combine stereo-based human features with thermal-based human features. The fusion performance is verified in Sec. 6 and a brief conclusion is made in Sec. 7. For ease of presentation, the symbols used in this paper can be found in Table 1.

Table 1. Nomenclature.

$O_vX_vY_vZ_v$	reference frame attached to the stereo vision system;
$O_tX_tY_tZ_t$	reference frame attached to the infrared thermal camera;
$[x_v, y_v, z_v]^T$	coordinate of a scene point in the reference frame $O_vX_vY_vZ_v$;
$[x_t, y_t, z_t]^T$	coordinate of a scene point in the reference frame $O_tX_tY_tZ_t$;
(d_x, d_y, d_z)	offset of $O_tX_tY_tZ_t$ to $O_vX_vY_vZ_v$;
R	rotation matrix from $O_vX_vY_vZ_v$ to $O_tX_tY_tZ_t$;
T	translation vector from $O_vX_vY_vZ_v$ to $O_tX_tY_tZ_t$;
$O_l(O_r)$	center of projection for the left (right) visual camera;
O_t	center of projection for the infrared thermal camera;
$\mathrm{I}_l(\mathrm{I}_r)$	image plane of the left (right) visual camera;
d_b	length of the baseline O_lO_r;
$f_l(f_r)$	focal length of the left (right) visual camera;
f_t	focal length of the thermal camera;
$\mathcal{F}_l(\cdot)$ $(\mathcal{F}_r(\cdot))$	image formation process for the left (right) visual camera (the mapping from the image plane to the left (right) image);
$\mathcal{F}_t(\cdot)$	image formation process for the infrared thermal camera;
$[y_{n,l}, z_{n,l}]^T$	coordinate of a point in the left image plane;
$[y_{n,r}, z_{n,r}]^T$	coordinate of a point in the right image plane;
$[y_l, z_l]^T, [y_r, z_r]^T$	coordinate of a point in the left (right) visual image;
$[y_h, z_h]^T$	coordinate of a point in the thermal image;
d_n	disparity between the corresponding points on the two visual image planes;
$D(y, z)$ or D	disparity image;
d	disparity value;
L_m, L_n	width and height of a disparity image;
$z_h(\cdot)$	height limit of a human object in an image;
z_c	the vertical center of an image;
H_L, H_R	height limit of human objects, the height of the stereo rig in 3D space;
d_h	average thickness of human body in 3D space;
d_w	average width of human body in 3D space;
$w_b(\cdot)$	width of human body in an image;
$P(y, d)$	2D histogram transferred from $D(y, z)$;
$\hat{\Psi}(y, d)$	filtered 2D histogram;

Table 1. (*Continued*)

$d_{\max}, d_{\min}$	upper and lower bounds of image disparity;
S_h, H_h	average 2D size and height of human heads in 3D space;
y_k^*, $y_{k,h}^*$	horizontal center of the body and head of a human object in an image;
$G_y(u)$, $G_d(v)$	scale-adaptive filters for spatial (y) and depth (d) measures;

2. Vision System

To pave the way for further investigation, we firstly introduce the vision system setup. Then, the relationship between the image size of an object and its distance to the vision system is established for the design of the scale-adaptive filter and human spatial segmentation.

2.1. *System description*

In this paper, we investigate a vision system consisting of a stereo rig and an infrared thermal camera as shown in the left graph of Fig. 1. The stereo rig is a stereo head of SRI's (Stanford Research Institute) Small Vision System at a resolution of 320×240 and the infrared thermal camera is a Raytheon thermal-eye 300 digital thermal camera at the same resolution. The stereo head with an attached reference frame $O_vX_vY_vZ_v$ is composed of two conventional CCD cameras whose centers of projection, O_l and O_r, are on the axis Y_v and their optical axes point forward in parallel. Axis X_v passes through the middle point O_v of the baseline O_lO_r (with distance d_b) and points forward. Axis Z_v completes the coordinate system via the right hand rule. The left and right image planes of the CCD cameras are represented by I_l and I_r respectively. The thermal reference frame $O_tX_tY_tZ_t$ is located at an offset of (d_x, d_y, d_z) to $O_vX_vY_vZ_v$.

For a single camera, an image is obtained via an image formation process, by which a 3D point is perspectively projected onto the image plane of the camera to form a virtual image point, which is in turn mapped to a certain position of the image through a camera model [Trucco and Verri, 1998]. We define by $\mathcal{F}(\cdot)$ the image formation process in this paper. If the intrinsic parameters, i.e., $\mathcal{F}(\cdot)$, of the cameras are calibrated, they allow the derivation of the relationship between the image size of an object and its distance to the stereo vision system.

2.2. *Geometry relationship for stereo vision*

As discussed in the above subsection, $\mathcal{F}(\cdot)$ denotes the image formation process. Figure 2 illustrates the formation process by projecting

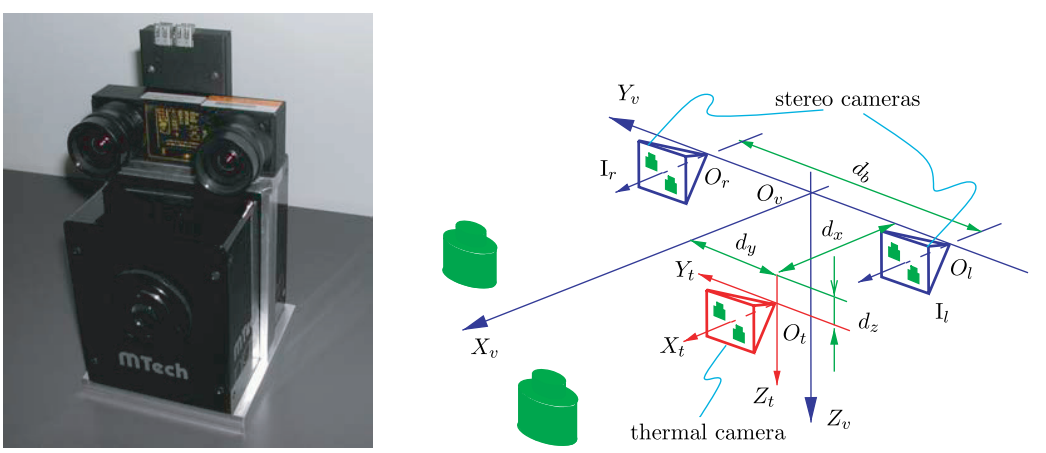
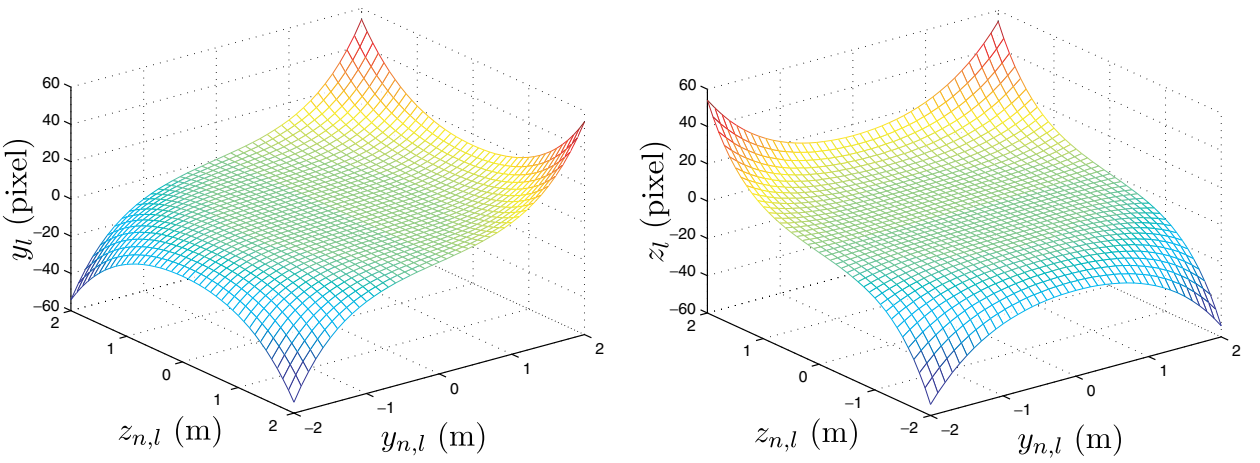


Fig. 1. Vision system setup.



(a) (b)

Fig. 2. Projection from $[y_{n,l}, z_{n,l}]^T$ to $[y_l, z_l]^T$.

a point, $[y_{n,l}, z_{n,l}]^T$, in the left image plane to a point, $[y_l, z_l]^T$ in the left image given that the left visual camera is calibrated. Figures 2(a) and (b) show the y_l and z_l responses against $[y_{n,l}, z_{n,l}]^T$ respectively. This result shows that y_l is monotonic with respect to $y_{n,l}$. The similarity is drawn to z_l (with respect to $z_{n,l}$). The data in Fig. 2 can be used to generate a lookup table with entries of y_l and z_l such that the coordinate $[y_{n,l}, z_{n,l}]^T$ in the image plane can be obtained. In other word, the inverse formation process, $\mathcal{F}^{-1}(\cdot)$, maps a pixel position in an image to a position in the image plane. Then, we can compute the disparity d_n from a pair of image planes by

$$d_n = \mathbf{Y}(\mathcal{F}_l^{-1}(y_l, z_l)) - \mathbf{Y}(\mathcal{F}_r^{-1}(y_l - d, z_l)), \tag{1}$$

where $\mathbf{Y}(\cdot)$ is the operator to extract the y component of "$\cdot$", $d = D(y, z)$ is the disparity value computed from a pair of stereo images, $\mathcal{F}_l^{-1}(\cdot)$ and $\mathcal{F}_r^{-1}(\cdot)$ are the inverse image formation processes of the left and right cameras respectively. If the focal lengths of these two cameras are identical, the depth of a 3D scene point can be computed by

$$x_v = f_v \frac{d_b}{d_n}, \tag{2}$$

where f_v is the identical focal length, $d_n = y_{n,l} - y_{n,r}$ is the disparity between the corresponding points in the two image planes and is obtained by Eq. (1), and $y_{n,r}$ is the y coordinate of the scene point in the right image plane.

It can be observed from Fig. 2 that the image formation enables a linear approximation in a restricted region, e.g. $y_{n,l} \in [-1.2\,\text{m}, 1.2\,\text{m}]$ and $z_{n,l} \in [-0.95\,\text{m}, 0.75\,\text{m}]$. This region is almost the field of view for the cameras. Therefore, the processes, $\mathcal{F}_l(\cdot)$ and $\mathcal{F}_r(\cdot)$, can be linearized and Eq. (1) becomes

$$\begin{aligned} d_n &= \mathbf{Y}(\mathcal{F}_l^{-1}(y_l, z_l)) - \mathbf{Y}(\mathcal{F}_r^{-1}(y_l - d, z_l)) \\ &= \overline{k} y_l - \overline{k}(y_l - d) = \overline{k} d, \end{aligned} \tag{3}$$

where $\overline{k}$ is a constant. Thus, the depth information can be obtained from Eq. (2)

$$x_v = f_v \frac{d_b}{d_n} = f_v \frac{d_b}{\overline{k} d} = \frac{k_1}{d}, \tag{4}$$

where $k_1 = f_v d_b / \overline{k}$ is a constant, which can be the mean of values from (4) by measuring x_v and the corresponding d for a set of 3D points. This approximation is essential in the derivation and realization of human detection in the next section. To achieve the robust human detection and identification, we restrict our attention to certain conditions, which are described by the following prerequisites:

A1: Human beings to be detected stand upright on the floor without a large degree of slant.
A2: Human beings are in the field of view of all three cameras.

A3: Human beings stand in the front of the vision system with appropriate distances to avoid a large or diminutive size of human beings in the captured images, and it should be ensured that the main features of human beings, i.e. head and shoulders, appear in the images.

3. Stereo-Based Human Detection and Identification

The stereo-based approach for human detection has three steps: (i) scale-adaptive filtering, (ii) human segmentation, and (iii) human verification. Firstly, a disparity image is transferred to a 2D histogram on a horizontal plane for ease of data manipulation. A model-driven (via scale analysis) method is then used to aggregate and enhance salient features of human beings in this histogram. Secondly, these enhanced features are extracted from the histogram for the purpose of segmenting human beings from other background objects. Finally, the image regions corresponding to the segments of human beings are verified by using a deformable head-shoulder template.

3.1. *Scale-adaptive filtering*

Given a pair of stereo images, we obtain the disparity image $D(y, z)$ with size $L_m \times L_n$ using cross correlation method. Prior to the segmentation of human beings, exploiting the depth measurement, we establish the spatial constraints for the human objects in an image from the physical properties of human beings as follows:

Body height constraint. As far as human beings are concerned, their heights are limited in 3D space such that their projections in an image are bounded. The height limit, a function over x_v or d, in the image is determined by

$$z_h(d) = z_c - f_v \frac{H_L - H_R}{\overline{k} x_v} = z_c - k_2 \Delta_H d, \quad (5)$$

where $z_c = L_n/2$ is the vertical center of the image, $k_2 = f_v/(k_1\overline{k}) = 1/d_b$, $\Delta_H = H_L - H_R$ is initially set, H_L is height limit of a person in 3D space and H_R is height of the stereo rig.

Body thickness constraint. Let d_h be the average thickness of a human body in 3D space, the disparity measurement is distributed in a range of $[d^-, d^+]$, if a person stands in front of the camera with distance $x_v = k_1/d$. The variables, d^- and d^+, are calculated by

$$d^- = \frac{k_1}{x_v + d_h/2} = \frac{2k_1 d}{2k_1 + d_h d},$$
$$d^+ = \frac{k_1}{x_v - d_h/2} = \frac{2k_1 d}{2k_1 - d_h d}. \quad (6)$$

Body width constraint. Let d_w be the average width of a human body in 3D space, the estimated width $w_b(d)$ of the human body in an image can be calculated by

$$w_b(d) = f_v \frac{d_w}{\overline{k} x_v} = k_2 d_w d. \quad (7)$$

These constraints are the relationship among the size of human body, body distance to the camera and its corresponding image size. To detect a human candidate by using these spatial constraints, we introduce a Y_I–D plane. All pixels of a disparity image are projected onto the plane Y_I–D according to their associated values, namely, y and d. Then, a 2D histogram, $P(y, d)$, is generated. Exploiting the fact that human objects stand and move on the ground surface, there is less overlap of the depth measures of different persons on the projected 2D histogram. Moreover, this projection facilitates the realistic implementation as Y_I–D is a flat plane and can be stored in an array. Mathematically, the transfer from $D(y, z)$ to $P(y, d)$ is obtained by

$$P(y, d) = \sum_{z=z_h(d)}^{L_n} \delta\left(D(y, z) - d\right), \quad (8)$$

where $z_h(d)$ is the computed human height constraint by (5). Next, the 2D histogram $P(y, d)$ is further filtered for the segmentation of each human individual. Human objects appear with different scales in the image. But the stereo disparity information allows us to estimate the appropriate scale for the person with the corresponding distance to the camera. Thus, associated points in $Y_I - D$ plane can be aggregated

by scale-adaptive filtering:

$$\hat{\Psi}(y,d) = (P * G)(y,d), \tag{9}$$

where "$*$" denotes the convolution and $G(\cdot,\cdot)$ is the scale-adaptive 2D kernel functions. Assume that the scale-adaptive filter in y and d directions are independent, the filter is thus chosen as

$$G(u,v) = G_y(u)G_d(v), \tag{10}$$

where $G_y(u)$ and $G_d(v)$ are the scale-adaptive filters for spatial (y) and depth (d) measures respectively. Substituting (10) into (9), the adaptively filtered 2D histogram can be expressed as

$$\hat{\Psi}(y,d) = ((P * G_d) * G_y)(y,d). \tag{11}$$

This equation indicates that the 2D convolution can be decomposed into two cascade 1D convolutions. By applying the body thickness constraint (6) to $G_d(v)$, the scale-adaptive filter in the depth direction is defined as

$$G_d(v) = e^{-\frac{v^2}{\sigma^2_{d^{\pm}(d)}}}, \quad (\text{`}+\text{'} : v \geq 0, \text{`}-\text{'} : v < 0), \tag{12}$$

where

$$\sigma_{d^+}(d) = \frac{d^+ - d}{2\sqrt{-\ln(0.5)}},$$
$$\sigma_{d^-}(d) = \frac{d - d^-}{2\sqrt{-\ln(0.5)}} \tag{13}$$

are the adaptive scale parameters of the filter for the depth, d, measure. This non-symmetry Gaussian-like kernel function is designed due to the non-linear relationship between x_v and d in (4). For each half range of $G_d(v)$, it equals to 0.5 for $v = (d^+ - d)/2$ or $v = (d^- - d)/2$, which indicate the half point for each half range. Applying it to the first convolution in equation (11), the 2D histogram $P(y,d)$ is first filtered in the depth direction as

$$\Psi(y,d) = \sum_{v=d^-}^{d^+} P(y,v)G_d(v-d). \tag{14}$$

By applying the body width constraint (7) to $G_y(u)$, the scale-adaptive filter in the spatial direction Y_I is defined as

$$G_y(u) = e^{-\frac{u^2}{\sigma^2_y(d)}}, \tag{15}$$

where $\sigma_y(d) = w_b(d)/(8\sqrt{-\ln(0.5)})$ is the adaptive scale parameter of the filter for the spatial y measure. To avoid merging adjacent human objects, the width of the filter is selected as the half of the estimated body width and $G_y(u) = 0.5$ at the half point of the range for each side of the filter, namely, $u = w_b(d)/8$. Applying the scale-adaptive filter $G_y(u)$ to each line, d, of $\Psi(y,d)$, the 2D histogram generated by using depth-oriented scale-adaptive filtering becomes

$$\hat{\Psi}(y,d) = \sum_{u=-w_\gamma}^{w_\gamma} \Psi(u+y,d)G_y(u), \tag{16}$$

where $w_\gamma = w_b(d)/4$.

Figure 3 shows the $\hat{\Psi}(y,d)$ obtained from (16). The first row of Fig. 3 shows three image samples wherein there are one, two and three persons respectively. The corresponding variations of $\hat{\Psi}(y,d)$ are shown below in the second row. It is observed from the computed $\hat{\Psi}(y,d)$ that there are one, two and three noticeable mounds correspondingly and their positions are proportional to those of human beings in the horizontal plane. This demonstrates that disparity image can be transferred to the $Y_I - D$ plane and the filtered 2D histogram may profit the subsequent human segmentation by detecting these mounds. In order to detect possible human candidates, the filtered 2D histogram, $\hat{\Psi}(y,d)$, is fed into a segmentation process for extraction.

3.2. *Human segmentation*

An algorithm to segment human objects from the filtered 2D histogram $\hat{\Psi}(y,d)$ is derived in this subsection. It extracts persons iteratively from close to far. In each iteration, k, it performs the operations as follows. First, a 1D disparity histogram $H_d(d)$ from $\hat{\Psi}(y,d)$ is generated as $H_d(d) = \sum_{y=1}^{L_m} \hat{\Psi}(y,d)$. The upper disparity edge of the closest human object is detected by searching an appropriate d from $d_{\max}$ to $d_{\min}$ such that $H_d(d) > T_d(d)$, where $d_{\max}$ and $d_{\min}$ are the upper and lower bounds of d. The adaptive threshold $T_d(d)$ is the image size of humans with distance $x_v = k_1/d$ to the camera and

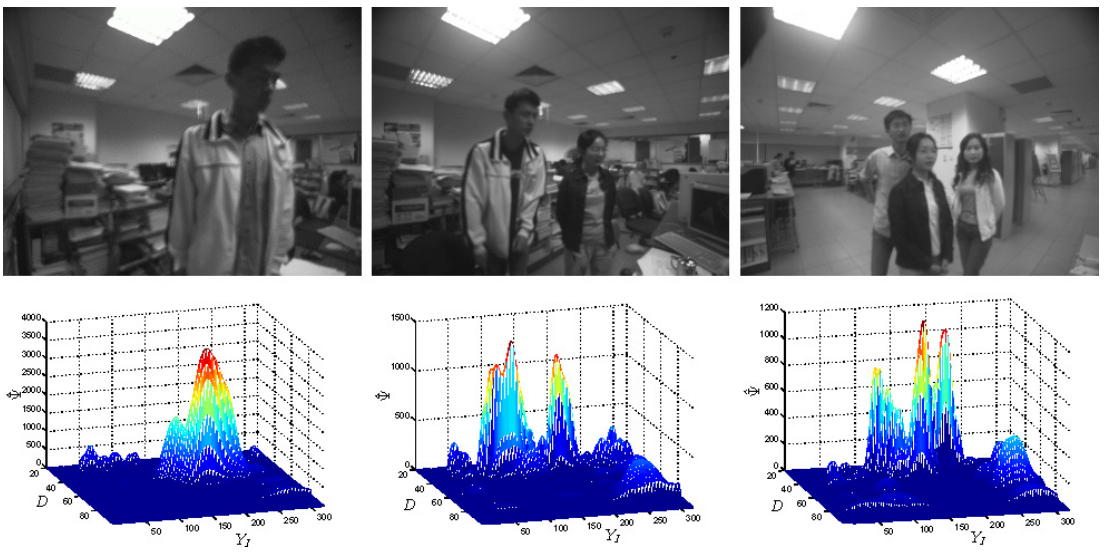


Fig. 3. Generation of $\hat{\Psi}(y,d)$.

calculated by

$$\begin{aligned} T_d(d) &= \int_{z_u}^{z_b}\int_{y_l(z)}^{y_r(z)} dydz \\ &= \int_{z_v^u}^{z_v^b}\int_{y_v^l(z_v)}^{y_v^r(z_v)} (k_2 d)dy_v(k_2 d)dz_v \\ &= k_2^2 d^2 \int_{z_v^u}^{z_v^b} dz_v \int_{y_v^l(z_v)}^{y_v^r(z_v)} dy_v \\ &= k_d k_2^2 S_b d^2, \end{aligned} \tag{17}$$

where z_u and z_b are the upper and lower bounds of a human object in an image, while $y_l(z)$ and $y_r(z)$ are the left and right bounds of him in the line, z, and

$$S_b = \mathbf{E}\left(\int_{z_v^u}^{z_v^b} dz_v \int_{y_v^l(z_v)}^{y_v^r(z_v)} dy_v\right) \tag{18}$$

is the average of 2D human sizes in the 3D space, k_d is a constant and $\mathbf{E}(\cdot)$ is the expectation. The virtual bounds of the human object in real 3D space

$$\begin{aligned} z_v^u &= \frac{z_u}{k_2 d}, \quad z_v^b = \frac{z_b}{k_2 d}, \\ y_v^l(z_v) &= \frac{y_l(z)}{k_2 d}, \quad y_v^r(z_v) = \frac{y_r(z)}{k_2 d}. \end{aligned} \tag{19}$$

If no such a peak of $H_d(d) > T_d(d)$ is found, the human segmentation process is terminated. Let d_k^+ be the upper edge of the kth detected peak. Thus, the depth of the major human body, d_k, is

$$d_k = \frac{2k_1 d_k^+}{2k_1 + d_h d_k^+}. \tag{20}$$

Applying (20) into (6), we have

$$d_k^- = \frac{2k_1 d_k}{2k_1 + d_h d_k} = \frac{k_1 d_k^+}{k_1 + d_h d_k^+}. \tag{21}$$

In this iteration, the person who is the kth closest to the camera is within the depth range of $[d_k^-, d_k^+]$. Next, the position of the person in the Y_I direction is located by examining a new histogram

$$H_y(y) = \sum_{d=d_k^-}^{d_k^+} \hat{\Psi}(y,d). \tag{22}$$

From $H_y(y)$, the peak position, y_k^*, namely, the center position of the person, is obtained by

$$y_k^* = \arg\max_{y\in[1,L_m]}\{H_y(y)\}. \tag{23}$$

Therefore, the major part of the kth peak is roughly within a $Y_I - D$ region of $[d_k^-, d_k^+]$ and $[y_{k,l}, y_{k,r}]$, where

$$\begin{aligned} y_{k,l} &= \begin{cases} 1, & \text{if } y_k^* - 0.4w_b(d_k) < 1 \\ y_k^* - 0.4w_b(d_k), & \text{otherwise} \end{cases} \\ y_{k,r} &= \begin{cases} L_m, & \text{if } y_k^* + 0.4w_b(d_k) > L_m \\ y_k^* + 0.4w_b(d_k), & \text{otherwise} \end{cases} \end{aligned} \tag{24}$$

Due to the fact that there may be several human objects staying close to each other in Y_I direction, the boundary of the region needs further refinement around its margins, e.g. $b_n(d_k) = 0.2w_b(d_k)$. The left and right margin of this peak is refined to a valley or the edge of the peak within a small neighborhood of $y_{k,l}$ and $y_{k,r}$ in $H_y(y)$. Let the refined edges of this peak be $y^*_{k,l}$ and $y^*_{k,r}$. Accordingly, the disparity histogram is refined by

$$H^*_d(d) = \sum_{y=y^*_{k,l}}^{y^*_{k,r}} \hat{\Psi}(y,d). \tag{25}$$

The refinement process is also applied to d^-_k and d^+_k. The two end points of the neighbor range are selected as follows:

$$\hat{d}^+_{k,u} = \frac{k_1 d^+_k}{k_1 - d'_h d^+_k}, \quad \hat{d}^-_{k,b} = \frac{k_1 d^-_k}{k_1 + d'_h d^-_k}, \tag{26}$$

where $d'_h = d_h/2$ is a constant. These two points are selected in such a manner that $\hat{d}^+_{k,u}$ is a little bit larger than d^+_k while $\hat{d}^-_{k,b}$ is a little bit smaller than d^-_k. The refined disparity range is obtained as follows:

$$d^{+*}_k = \begin{cases} d^+_t, \text{ if } d^+_t \le \hat{d}^+_{k,u} \text{ and} \\ \qquad H^*_d(d) > w_b(d_k)/8, \ \forall d \in (d^+_k, d^+_t] \\ d^+_k, \text{ otherwise} \end{cases}$$

$$d^{-*}_k = \begin{cases} d^-_t, \text{ if } d^+_t \ge \hat{d}^-_{k,b} \text{ and} \\ \qquad H^*_d(d) > w_b(d_k)/8, \ \forall d \in [d^-_t, d^-_k) \ . \\ d^-_k, \text{ otherwise.} \end{cases} \tag{27}$$

This refinement process allows the inclusion of salient depth measures that are likely associated with the kth human object. The disparity center for the kth iteration is thus refined by

$$d^*_k = \arg \max_{d \in [d^{-*}, d^{+*}]} H^*_d(d). \tag{28}$$

Then, the kth human object is segmented from an image by

$$O_k(y,z) = \begin{cases} k, & \text{if } y \in [y^*_{k,l}, y^*_{k,r}] \text{ and} \\ & \quad D(y,z) \in [d^{-*}, d^{+*}] \ . \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

Meanwhile, the present evidence of $O_k(y,z)$ is eliminated from $\Psi(y,d)$, $\hat{\Psi}(y,d)$ and $D(y,z)$ by setting

$$\begin{aligned} &\Psi(y,d) = 0, \quad \text{if } y \in [y^*_{k,l}, y^*_{k,r}] \text{ and } d \in [d^{-*}, d^{+*}] \\ &\hat{\Psi}(y,d) = 0, \quad \text{if } y \in [y^*_{k,l}, y^*_{k,r}] \text{ and } d \in [d^{-*}, d^{+*}] \\ &D(y,z) = 0, \quad \text{if } y \in [y^*_{k,l}, y^*_{k,r}], \ d \in [d^{-*}, d^{+*}] \text{ and } O_k(y,z) \neq 0. \end{aligned} \tag{30}$$

The obtained regions for the kth human-like object $O_k(y,z)$ are further smoothed by a morphological operation in order to erase small isolated regions and fill holes in certain regions [Maragos *et al.*, 1996]. Then, we have the new segmentation image $O^*_k(y,z)$.

3.3. *Human verification*

The segmented regions are confirmed as human objects by their head-shoulder contours. This process is described in this subsection. For the kth human object, $O^*_k(y,z)$, a histogram in z direction is obtained $H_z(z) = \sum_{y=1}^{L_m} \delta(O^*_k(y,z) - k)$. The peak of a possible head is obtained as z_{top} such that $H_z(z) = 0, \forall z < z_{top}$. If this object is valid, the sums of $H_z(z)$ over $[z_{top}, z_{top} + h_h]$ and $[z_{top}+h_h, z_{top}+2h_h]$ are larger than an adaptive threshold

$$T_h(d_k) = k_3 k_2^2 S_h d_k^2, \quad h_h = k_2 H_h d_k, \tag{31}$$

where S_h and H_h are the average 2D size and height of human heads, k_3 is a constant and the computation of $T_h(d_k)$ is similar to that in (17). Otherwise, this feature is eliminated. Another elimination condition is

$$\sum_{z=1}^{L_n} H_z(z) < S_b, \tag{32}$$

where S_b is given in (18). It means that the size of the object is too small to be a human object.

These conditions indicate that significant parts of head and shoulders are found. The position of the head is refined by firstly computing the histogram of upper portion of this feature. The histogram is obtained by

$$H_h(y) = \sum_{z=z_{top}}^{z_{top}+h_h} O_k^*(y,z). \tag{33}$$

Then, the horizontal center of the head is obtained by

$$y_{k,h}^* = \frac{\sum_{y=1}^{L_m} yH_h(y)}{\sum_{y=1}^{L_m} H_h(y)}. \tag{34}$$

Accordingly, the left and right bounds of human head, $y_{k,l}^h$ and $y_{k,r}^h$, are obtained by

$$y_{k,l}^h = \begin{cases} y_{k,h}^* - h_w/2, & \text{if } y_{k,h}^* - h_w/2 > 0 \\ 1, & \text{otherwise} \end{cases}$$

$$y_{k,r}^h = \begin{cases} y_{k,h}^* + h_w/2, & \text{if } y_{k,h}^* + h_w/2 < L_m \\ L_m - 1, & \text{otherwise} \end{cases}. \tag{35}$$

where $h_w = 0.4w_b(d_k)$ is the width of human head in an image according to the biometric measurement of human beings. If the size of the head portion is less than $h_w h_h/2$, it is discarded.

In order to extract the contour neighborhood of the kth human object, we dilate $O_k^*(y,z)$ k_s times as in [Maragos *et al.*, 1996] and have its dilated version, $\hat{O}_k^*(y,z)$. The contour neighborhood of this human object is obtained by

$$B_k(y,z) = \hat{O}_k^*(y,z) - O_k^*(y,z), \tag{36}$$

where $k_s = 0.5W_s$ and $W_s \times W_s$ is the window size for disparity calculation. Since the neighborhood $B_k(y,z)$ may not cover the bounding of the human object completely, a blur or averaging process is applied to it $l_k = 0.1w_b(d_k)$ times such that

$$\hat{B}_k^{(l)}(y,z) = \sum_{i=y-1}^{y+1} \sum_{j=z-1}^{z+1} \hat{B}_k^{(l-1)}(i,j)/9, \tag{37}$$

where $\hat{B}_k^{(0)}(y,z) = B_k(y,z)$, $\hat{B}_k(y,z) = \hat{B}_k^{(l_k)}(y,z)$. Indeed, $\hat{B}_k(y,z)$ provides a potential field for edge points that belong to the contour of the human object.

Since the dilation and average process deform the contour of the kth detected object, it is necessary to refine them using object edge information. The edge points are obtained by using Canny edge detector and represented by $M(y,z)$. Considering the spatial bias caused by stereo calculation, before matching of these two boundary information, we process $M(y,z)$ in the same manner as that of $B_k(i,j)$ and have

$$\hat{M}^{(l)}(y,z) = \sum_{i=y-1}^{y+1} \sum_{j=z-1}^{z+1} M^{(l-1)}(i,j)/9, \tag{38}$$

where $\hat{M}^{(0)}(y,z) = M(y,z), l = 1,\ldots,l_k$. Let $\hat{M}(y,z) = \hat{M}^{(l_k)}(y,z)$. Thus, the object image of the coincident evidence for object's contours is

$$E_k(y,z) = \min\{\hat{B}_k(y,z), \hat{M}(y,z)\}. \tag{39}$$

The feature contour is shown in Fig. 4, where the left image shows the present evidence of object's contour $E_k(y,z)$ refined by edge information and the left image from the stereo rig is shown right to it. However, there is one more step to go to verify if the object's contour presents a human candidate. It is further observed that there is coincidence of human-like head-shoulder contours from both stereo and edge features for human objects, while non-human objects have less inclinations towards human like appearance. Then, a model-driven approach based on a deformable head-shoulder template is used to distinguish humans from non-human objects.

To evaluate whether the upper part of $E_k(y,z)$ looks like a head-shoulder contour of human beings, a deformable template matching algorithm is designed. First of all, a normalized polygon head-shoulder template from the front

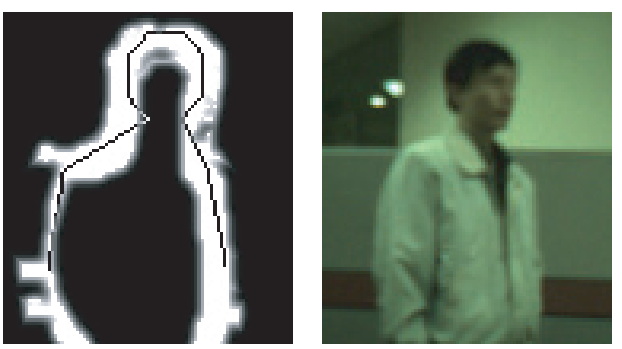
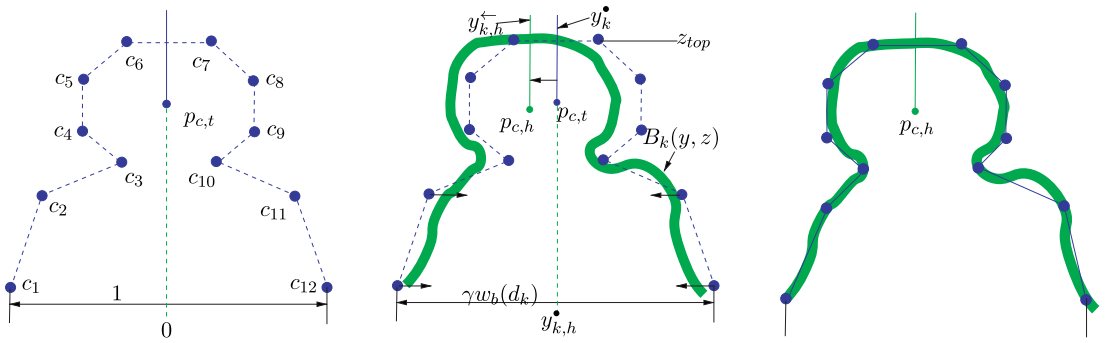


Fig. 4. Feature contour and human identification.



(a) Normalized template (b) Scaling (c) Deformation

Fig. 5. Deformable template

view of human beings is generated from the statistics of human images. The template has 12 control points, i.e., $c_i = (y_i, z_i), i = 1, \ldots, 12$, as shown in Fig. 5(a), where $p_{c,t}$ denotes the center of the gravity of head. Before matching the template to $E_k(y, z)$, three steps are applied to the template for the variation of scale and posture of human objects. First, the template is scaled up to the estimated width of the human candidate by $\gamma w_b(d_k)$:

$$c_i \leftarrow \gamma w_b(d_k) c_i, \quad i = 1, \ldots, 12. \tag{40}$$

As the size of a human being may deviate from the average size, the variable $\gamma = 1 \rightarrow 1.2$ is used to deal with the scale variation of persons. Secondly, for the scaled template, the head position is adjusted by

$$y_i \leftarrow y_i + (y_{k,h}^* - y_k^*), \quad i = 3, \ldots, 10, \tag{41}$$

where y_i is the y component of c_i, such that it is centered at $y_{k,h}^*$ by horizontally shifting the control points $c_3 - c_{10}$ as shown in Fig. 5(b) where $p_{c,h}$ denotes the center of the head for the human candidate. Finally, the control points of the shoulders are adjusted. For the left shoulder, the control points c_2 is first horizontally shifted to a point where $E_k(y, z)$ value is maximum on the left side of body. It is achieved by moving the control point horizontally to

$$y_2 = \arg \max_{y \in [y_s, y_e]} E_k(y, z_2), \tag{42}$$

where $y_s = y_{k,h}^* - 0.2 w_b(d_k)$ and $y_e = y_3$ are the two neighborhood points to set the searching bounds. Similar to c_2, the control point c_1 is horizontally adjusted. The same operation is also performed for the right shoulder, namely, c_{11} and c_{12}. After the deformation, a polygon template is generated by connecting the control points with straight lines as shown in the right graph of Fig. 5(c). The deformed template of scale γ can be represented by a point sequence $T_k^\gamma = (y_i^\gamma, z_i^\gamma)$ with n_k^γ points. For T_k^γ, the support from the evidence is computed from the vertical position $z_s = z_{top} - 0.2 w_b(d_k)$ to $z_e = z_{top} + 0.4 w_b(d_k)$. The best matching measure for the deformed template is obtained by

$$N_k = \max_{\gamma \in \{1, 1.2\}} \times \left\{ \max_{z \in \{z_s, z_e\}} \left\{ \frac{1}{n_k^\gamma} \sum_{i=1}^{n_k^\gamma} E_k(y_i^\gamma + y_c, z_i^\gamma + z) \right\} \right\} \tag{43}$$

If $N_k > T_m$, e.g. 0.5, the object is verified as a human object. One example of the contour image $E_k(y, z)$ and the deformed template is shown in Fig. 4 where the contour and the template are partially matched. Figure 6 shows the relationship between the threshold T_m and the rate of human detection. From this curve, $T_m = 0.5$ was chosen for the later experiments. However, if the person tilts greatly, the deformable temple would fail to identify him. If there are human-like objects, e.g. an advertisement board with human shape, the stereo rig would also generate false positive. Thus, we make use of thermal feature of human being to enhance the human verification.

Remark 1. Disparity maps usually come up with noise, which may affect the human

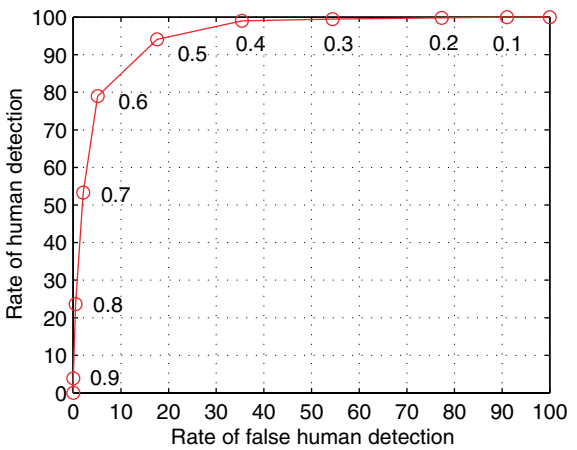


Fig. 6. Relationship between the threshold T_m and the rate of human detection.

detection performance if only hard thresholding is used. In this paper, a disparity map is processed through several coherent steps such as scale-adaptive filtering, human segmentation. Among them, there are several knowledge-based detection conditions such as (17), (31), (32) and so forth. The detected candidates are finally fed to a verification process. The curve shown in Fig. 6 demonstrates that human detection from a disparity map provide satisfactory results.

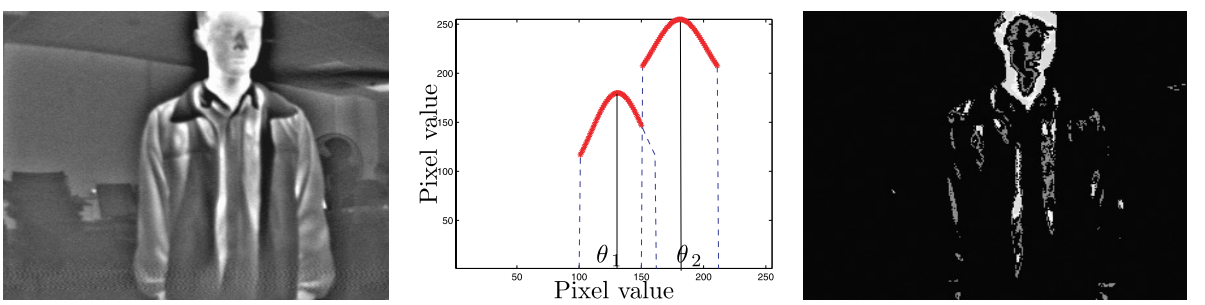
Remark 2. There are several state-of-the-art techniques available for human segmentation. However these approaches are usually used to detect segments in an image, but cannot tell if these detected segments belong to a human object. In this work, we firstly detect possible human candidates through scale-adaptive filtering, and then corresponding image areas can be found and verified. In this manner, human detection can be more robust.

4. Thermal Image Processing

This section describes the process of extracting the infrared thermal features of human being from the infrared thermal image for the enhancement of human verification. Due to the fact that the intensity value of a pixel in an infrared thermal image is proportional to the thermal radiation of the scene associated with it [Socolinsky *et al.*, 2001], pixel filter enables scene of interest such as human skin or face to be distinct. A filter bank for specified scenes is thus designed as

$$\mathcal{G}_k(I(i,j)) = \begin{cases} (g_h(k) - g_l(k))e^{-\frac{(I(i,j)-\mu_k)^2}{\sigma_k^2}} + g_l(k), \\ \qquad\qquad \text{if } |I(i,j) - \mu_k| < \sigma_k, \\ 0, \qquad\qquad \text{otherwise.} \end{cases} \tag{44}$$

where $\mathcal{G}_k(\cdot)$ is the kth filter, $I(i,j)$ is the intensity value of the pixel at the ith row and jth column, μ_k and σ_k are the kth mean and spread of the filter, $g_h(k)$ and $g_l(k)$ are the output upper and lower bounds of the gray values. The values of μ_k and σ_k can be set experimentally. The objective of each filter is to extract image portion associated with a specified thermal feature. Figure 7(b) shows the response of two combined thermal filters. The filtered image is obtained by $I(i,j) = \max(\mathcal{G}_1, \ldots, \mathcal{G}_k, \ldots, \mathcal{G}_N)$, $k = 1, \ldots, N$. The left image in Fig. 7 shows the thermal image taken in a laboratory condition, while the right image demonstrates its filtered



(a) Thermal image (b) Thermal filter (c) Filtered thermal image

Fig. 7. Thermal image filtering.

version by using the filter in Fig. 7(b). It is observed that the thermal features, such as face, arms are distinct, while other objects such as table and ceiling are filtered out.

5. Human Detection by Fusion

In order to fuse stereo and thermal features, a scene point in the pair of stereo images, $[y_l, z_l]^T, [y_r, z_r]^T$, should be associate with its corresponding point, $[y_h, z_h]^T$, in the infrared thermal image. This can be achieved in the following manner:

$$([y_l, z_l]^T, [y_r, z_r]^T) \xrightarrow[\text{step 1}]{\mathcal{F}_l^{-1}(\cdot), \mathcal{F}_r^{-1}(\cdot)} [x_v, y_v, z_v]^T \xrightarrow[\text{step 2}]{R,T} [x_t, y_t, z_t]^T \xrightarrow[\text{step 3}]{\mathcal{F}_t(\cdot)} [y_h, z_h]^T. \quad (45)$$

In this way, the 3D coordinate of this point, $[x_v, y_v, z_v]^T$, is reconstructed from, $[y_l, z_l]^T$ and $[y_r, z_r]^T$, as in step 1 and then transformed to the thermal reference frame as in step 2 via the rotation matrix, R, and translation vector, T. The transformed point is mapped to $[y_h, z_h]^T$ via the image formation process, $\mathcal{F}_t(\cdot)$ as in step 3. Since step 1 has been presented in Sec. 2, steps 2 and 3 will be described in this section. Similar to the linearization of $\mathcal{F}_l(\cdot)$ and $\mathcal{F}_r(\cdot)$, $\mathcal{F}_t(\cdot)$ is also assumed to be linearized as k_t. Then, the parameters, R, T and k_t can be calibrated simultaneously.

5.1. *Extrinsic calibration*

Given the configuration in Fig. 1, we have

$$P_t = RP_v + T, \quad (46)$$

where $P_t = [x_t, y_t, z_t]^T$ and $P_v = [x_v, y_v, z_v]^T$ are the coordinates of a 3D point in reference frames $O_tX_tY_tZ_t$ for the stereo rig and $O_vX_vY_vZ_v$ for the infrared thermal camera respectively, R is the rotation matrix representing the misalignment between these two reference frames, and T is the translation vector representing the offset. Assuming that step 1 in (45) has been done, the problem needed to be solved is to estimate R, T and k_t given $[x_v(i), y_v(i), z_v(i)]^T$ and $[y_h(i), z_h(i)]^T$, $i = 1, \ldots, n_p$ (n_p is the number of the observation points). Given a 3D point $p_t = [x_t, y_t, z_t]^T$, the image point in the infrared thermal image is obtained by

$$y_h = k_t \frac{y_t}{x_t}, \quad z_h = k_t \frac{z_t}{x_t}. \quad (47)$$

Eliminating the terms x_t and k_t in equation (47), we have

$$y_h z_t = z_h y_t. \quad (48)$$

In components, equation (48) can be written as

$$\begin{aligned} &y_h(r_{31}x_v + r_{32}y_v + r_{33}z_v + t_3) \\ &= z_h(r_{21}x_v + r_{22}y_v + r_{23}z_v + t_2), \end{aligned} \quad (49)$$

where r_{ij} is the component of the rotation matrix R, and t_i is the component of translation vector T. Equation (49) can be thought of as a linear equation for the 8 unknowns $h = [h_1, h_2, \ldots, h_8]^T = [r_{31}, r_{32}, r_{33}, t_3, r_{21}, r_{22}, r_{23}, t_2]^T$. Given n_p pairs of $[y_h(i), z_h(i)]^T$ and $[x_v(i), y_v(i), z_v(i)]^T$ ($i = 1, \ldots, n_p$), equation (49) can be rewritten as linear equations for the 8 unknowns [Trucco and Verri, 1998]:

$$Gh = 0, \quad (50)$$

where the components of the ith row of G are

$$\begin{aligned} &G(i,1) = y_h(i)x_v(i), \quad &&G(i,2) = y_h(i)y_v(i), \\ &G(i,3) = y_h(i)z_v(i), \quad &&G(i,4) = y_h(i), \\ &G(i,5) = -z_h(i)x_v(i), \quad &&G(i,6) = -z_h(i)y_v(i), \\ &G(i,7) = -z_h(i)z_v(i), \quad &&G(i,8) = -z_h(i). \end{aligned} \quad (51)$$

If $n_p \geq 7$ and these n_p points are not coplanar, the rank of G is 7. Due to the measurement noise, the rank of G is likely to be maximum. Thus, we rewrite equation (50) as

$$f(h) = h^T G^T G h \geq 0. \quad (52)$$

Since R is a normal matrix, we have

$$g(h) = \begin{bmatrix} h_1^2 + h_2^2 + h_3^2 - 1 \\ h_5^2 + h_6^2 + h_7^2 - 1 \end{bmatrix}. \quad (53)$$

Thus, the solution to R and T can now be formulated as an optimization problem as follows:

$$\min f(h) \ \ s.t.\ g(h) = 0. \quad (54)$$

We consider the objective function

$$\begin{aligned} C(h, \lambda_1, \lambda_2) = h^T G^T G h + \lambda_1(h_1^2 + h_2^2 + h_3^2 - 1) \\ + \lambda_2(h_5^2 + h_6^2 + h_7^2 - 1), \end{aligned} \quad (55)$$

where λ_1 and λ_2 are Lagrange multipliers. Since $C(h, \lambda_1, \lambda_2) \geq 0$, $C(h, \lambda_1, \lambda_2) = 0$ if and only if there is no measurement noise. In the presence of noise, $C(h^*, \lambda_1^*, \lambda_2^*) > 0$, where

$$[h^*, \lambda_1^*, \lambda_2^*] = \arg \min_{h \in \mathbb{R}^{8\times 1}; \lambda_1, \lambda_2 \in \mathbb{R}} C(h, \lambda_1, \lambda_2). \tag{56}$$

Taking the derivative of C over each individual component of h, λ_1 and λ_2, we have

$$\frac{\partial C}{\partial h_i} = \begin{cases} 2\sum_{j=1}^{8} a_{ij}h_j + 2\lambda_1 h_i, & \text{if } i \in [1,3] \\ 2\sum_{j=1}^{8} a_{ij}h_j, & \text{if } i = 4 \\ 2\sum_{j=1}^{8} a_{ij}h_j + 2\lambda_2 h_i, & \text{if } i \in [5,7] \\ 2\sum_{j=1}^{8} a_{ij}h_j, & \text{if } i = 8 \end{cases}, \tag{57}$$

where a_{ij} is a component of matrix $G^T G$ and

$$\frac{\partial C}{\partial \lambda_1} = h_1^2 + h_2^2 + h_3^2 - 1,$$

$$\frac{\partial C}{\partial \lambda_2} = h_5^2 + h_6^2 + h_7^2 - 1. \tag{58}$$

Let $\frac{\partial C}{\partial h_i} = 0$, $\frac{\partial C}{\partial \lambda_1} = 0$, $\frac{\partial C}{\partial \lambda_2} = 0$, and cancel λ_1 and λ_2, we have

$$h_2 \sum_{j=1}^{8} a_{1j}h_j = h_1 \sum_{j=1}^{8} a_{2j}h_j,$$

$$h_3 \sum_{j=1}^{8} a_{1j}h_j = h_1 \sum_{j=1}^{8} a_{3j}h_j$$

$$\sum_{j=1}^{8} a_{4j}h_j = 0,$$

$$h_6 \sum_{j=1}^{8} a_{5j}h_j = h_5 \sum_{j=1}^{8} a_{6j}h_j$$

$$h_7 \sum_{j=1}^{8} a_{5j}h_j = h_5 \sum_{j=1}^{8},$$

$$a_{7j}h_j \sum_{j=1}^{8} a_{8j}h_j = 0$$

$$h_1^2 + h_2^2 + h_3^2 = 1, \quad h_5^2 + h_6^2 + h_7^2 = 1. \tag{59}$$

The components in h^* is then calculated using the large-scale algorithm [Coleman and Li, 1996]. We observe that the last two rows of the rotation matrix, R, and the last two components of the translation vector T are now determined. The first row of the rotation matrix is then obtained by the vector product of the last two computed rows.

To compute the remaining unknowns t_1 and k_t, let us consider the first equation in Eq. (47). Given the observation $[y_h(i), z_h(i)]'$ $(i = 1, \ldots, n_p)$, we have

$$y_h(i)(r_{11}x_v(i) + r_{12}y_v(i) + r_{13}z_v(i) + t_1) = k_t(r_{21}x_v(i) + r_{22}y_v(i) + r_{23}z_v(i) + t_2), \tag{60}$$

and then

$$\hat{G} \begin{bmatrix} t_1 \\ k_t \end{bmatrix} = b, \tag{61}$$

where

$$\hat{G} = \begin{bmatrix} y_h(1) & -(r_{21}x_v(1) + r_{22}y_v(1) + r_{23}z_v(1) + t_2) \\ \vdots & \vdots \\ y_h(i) & -(r_{21}x_v(i) + r_{22}y_v(i) + r_{23}z_v(i) + t_2) \\ \vdots & \vdots \\ y_h(n_p) & -(r_{21}x_t v(n_p) + r_{22}y_v(n_p) + r_{23}z_v(n_p) + t_2) \end{bmatrix}$$

$$b = -\begin{bmatrix} y_h(1)(r_{11}x_v(1) + r_{12}y_v(1) + r_{13}z_v(1)) \\ \vdots \\ y_h(i)(r_{11}x_v(i) + r_{12}y_v(i) + r_{13}z_v(i)) \\ \vdots \\ y_h(n_p)(r_{11}x_v(n_p) + r_{12}y_v(n_p) + r_{13}z_v(n_p)) \end{bmatrix}. \tag{62}$$

The least square solution to (t_1, k_t) is

$$\begin{bmatrix} t_1 \\ k_t \end{bmatrix} = (\hat{G}^T\hat{G})^{-1}\hat{G}^T b. \tag{63}$$

The whole computation process is summarized in flowchart shown in Fig. 8. As R and T are solved, a 3D point can be projected onto an infrared thermal image. Figure 9 demonstrates the calibration performance where many candidate points are selected from the top image to cover the area of a rectangle disk. These points are projected to the infrared thermal image at the bottom through the process in (45). It is observed that the computed R and T provides the projection with satisfactory resolution. Each pixel of the kth human object, $O_k^*(y, z)$, can be projected to the corresponding infrared thermal image in the same manner as shown in Fig. 9. Let S_p be the area (in the infrared thermal image) covered by these projected points. Then, the kth human object is confirmed if

$$\frac{S_p}{S_t} \geq \eta_1, \tag{64}$$

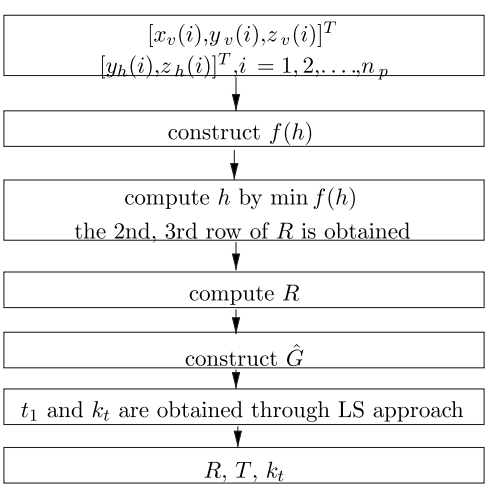


Fig. 8. Computation of R, T and k_t.

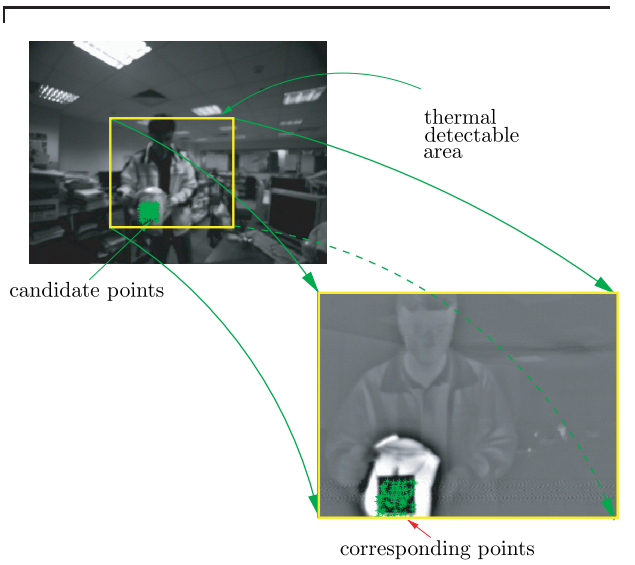


Fig. 9. Projection demonstration.

where η_1 is a threshold and S_t is the thermal image size for a person standing in front of the camera with distance x_t. The image size is computed by

$$\begin{aligned} S_t &= \int_{z_u}^{z_b}\int_{y_l(z_h)}^{y_r(z_h)} dy_h dz_h \\ &= \int_{z_t^u}^{z_t^b}\int_{y_t^l(z_t)}^{y_t^r(z_t)} \left(\frac{k_t}{x_t}\right) dy_t \left(\frac{k_t}{x_t}\right) dz_t, \\ &= \left(\frac{k_t}{x_t}\right)^2 \int_{z_t^u}^{z_t^b}\int_{y_t^l(z_t)}^{y_t^r(z_t)} dy_t dz_t \\ &= k_w \left(\frac{k_t}{x_t}\right)^2 S_b\,, \end{aligned} \tag{65}$$

where k_w is a constant, z_u and z_b are the upper and lower bounds of human image, while $y_l(z_h)$ and $y_r(z_h)$ are the left and right bound of the human image for a certain value of z_h. The bounds in 3D space are denoted by $[z_t^u, z_t^b]^T$ and

corresponding $[y_t^l(z_t), y_t^r(z_t)]^T$. To make use of thermal features, let S_f be the area covered by both the thermal feature and visual feature S_p. The human object is valid if

$$\frac{S_f}{S_p} \geq \eta_2, \tag{66}$$

where η_2 is set experimentally. These two simple conditions combine visual and thermal features, and profit the enhancement of human identification. This is successfully verified in our experiments.

6. Experimental Results

Three sets of experiments are conducted to verify the proposed approach. These include: (i) human detection using only stereo vision, (ii) the fusion of the stereo and infrared thermal images for the enhancement of human detection, and (iii) distinguishing human beings from a human-like object by using these two types of images. The parameters used in these experiments are shown in Table 2.

6.1. *Human detection using stereo vision alone*

Figure 10 shows single human detection using only stereo vision, wherein a person is surrounded by many background objects such as books, chars, tables and so on. The person moves in the front of the camera with the front facing the camera. It shows that the detection and identification of human candidates are achieved by using the proposed stereo based method.

Figure 11 shows the cases where a human candidate moves in the laboratory with side facing the camera and successful detections are achieved. Figure 12 demonstrates that the stereo based approach provides satisfactory results with the presence of multiple human candidates. Good performance is obtained even images of these two persons are overlapped as shown in the top right and bottom left graphs in this figure.

Table 2. Experimental parameters.

k_1	H_L	H_R	d_h	d_w
31887mm/pixel	2100mm	1375mm	180mm	460mm

6.2. *Human detection using both stereo rig and thermal camera*

Experiments using stereo based technique show that there are some false cases as shown in Fig. 13. The images in the left column of the figure show the false detection of a person. The shoulders of the person are identified as human candidates. The graphs in the right column of the figure show the $\hat{\Psi}(y, d)$ corresponding to the images in the left column. The peaks shown in these graphs indicate that shoulders of the person are able to provide significant detectable information. However, it can be solved by using the thermal feature provided by the infrared thermal camera as shown in Fig. 14. The first column shows images with the failure of human detection. The second column are filtered infrared thermal images, in which the hair, face and neck of human candidate are enhanced, while other objects such chairs, books, are removed. The contour of visual features are transformed and superimposed onto the infrared thermal images, which are shown by $*$ in the infrared thermal images. The last column demonstrates the fusion results of the stereo and infrared thermal images. The results indicate that the thermal features are capable of removing the failure detection and thus enhancing the robustness of the overall system.

Figure 15 shows some results where there are multiple persons with two different backgrounds. Among these results, there are several difficult cases where persons are very close to each other.

6.3. *Human detection with the presence of human-like object*

Generally, there are human-like objects in a variable environment. These objects may be an advertisement board with human shape, cube boxes and so on. To evaluate the detection

Fig. 10. Human detection with front facing the camera.

Fig. 11. Human detection with side facing the camera.

Fig. 12. Human detection with two human candidates.

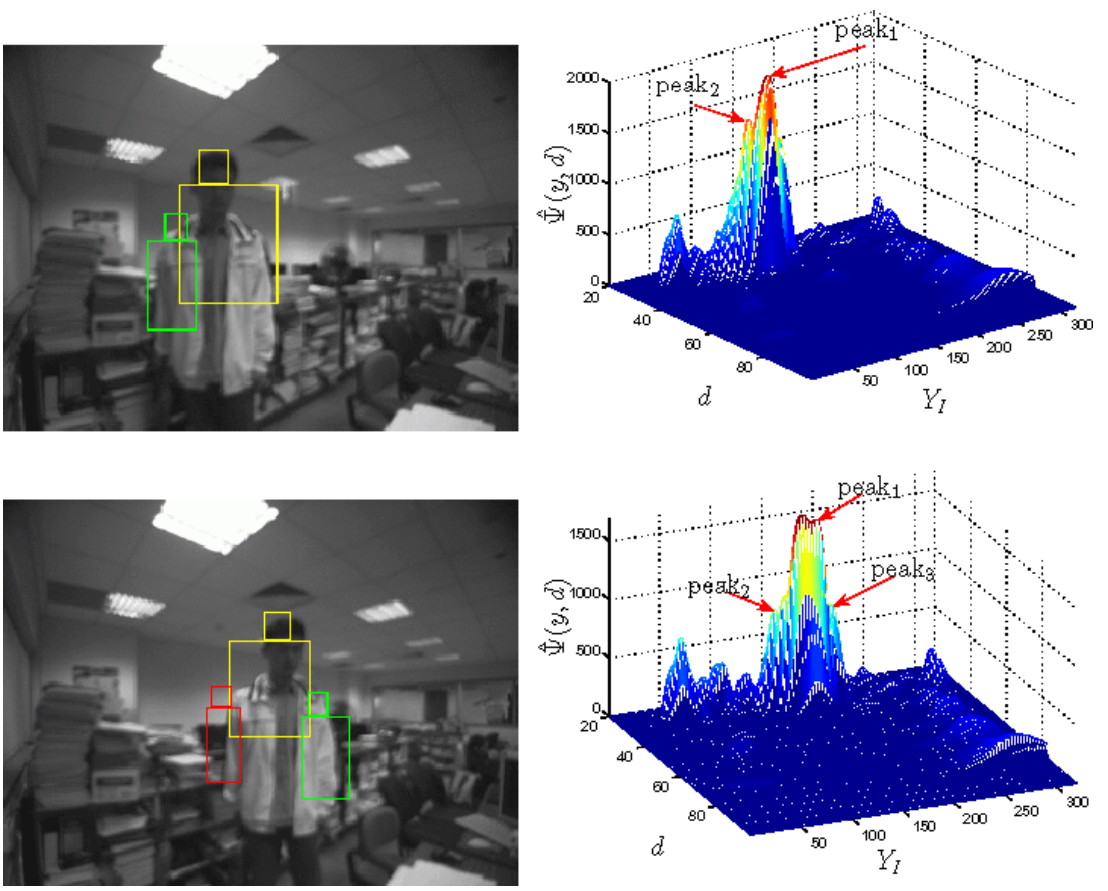


Fig. 13. Human detection with failure.

Fig. 14. Fusion based human detection.

Fig. 15. Multiple human detection with different background.

Fig. 16. Detection of object with human shape based on stereo approach.

Fig. 17. Fusion based human detection.

capability of both stereo-based and fusion based techniques, two cube boxes are used in our experiments as shown in Fig. 16. The small cube box is placed on the large cube box for the purpose of mimicking the head-shoulder shape of humans. It is also observed from this figure that stereo-based technique provides false detection. However, this failure can be avoided if thermal features are used. Figure 17 show the detection results using the fusion based techniques. It is obvious that the human-like object is ignored with the assist of thermal features.

Although the fusion based human detection technique provides satisfactory results, e.g. more than 90% success rate, there are several false cases. One example is shown in Fig. 18. The figure shows that the major portion of human object, e.g. the head, is not detected while other portions such as two shoulders are identified. Since the fusion based technique is based on the detection results provided by the stereo vision system, correct human identification cannot be achieved if major human features are not provided. Other false cases may include:

(1) Since disparity values are computed using cross-correlation method and the size of cross-correlation window is normally fixed, if a human candidate stands too close to the camera, the corresponding points (used to calculate disparity value) may not be in the same cross-correlation window. Thus, the disparity calculation may not be available in this case, which in turn provides no human detection results.
(2) If a human candidate is far from the vision system, the disparity value may be too small to be selected for segmentation purpose. In this case, the human detection may also fail.
(3) Due to the fact that the thermal camera has a limited field of view, no thermal feature can be used if a human candidate is outside its field of view.

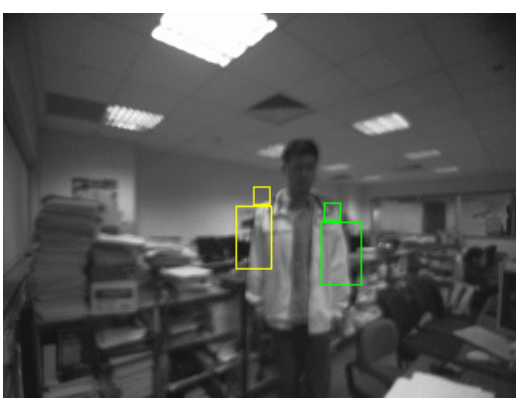


Fig. 18. Failure case using fusion based technique.

However these cases violate the assumptions made earlier in this paper.

7. Conclusion

In this paper, infrared thermal images have been incorporated with stereo images in an effort to develop a vision system to robustly detect and identify humans in 3D space for socially interactive robots with a friendly human robot interaction. Firstly, a scale-adaptive filter has been proposed for the stereo vision system to detect human candidates. Secondly, thermal features were incorporated to distinguish heat generating objects and non-heat generating objects. The fusion of these two types cameras thus enables the overall vision system effective and robust for human detection and identification, which can greatly enhance the interaction between intelligent social robots and humans.

Acknowledgment

The authors would like to thank Ms Yaozhang Pan, Beibei Ren and Mr Chenguang Yang for assistance in collecting image data. Special appreciation should be given to Mr Chee Siong Tan for assistance in designing the frame holder for the vision system.

References

Arrue, B. C., Ollero, A. and de Dios, J. R. M. [2000] "An intelligent system for false alarm reduction in infrared forest-fire detection," *IEEE Intelligent Systems* **15**(3), 64–73.

Breazeal, C. [2003]"Toward sociable robots," *Robotics and Autonomous Systems* **42**, 167–175.

Chan, W. L., So, A. T. P. and Lai, L. L. [2000] "Three-dimensional thermal imaging for power equipment monitoring," *IEE Proceedings — Generation, Transmission and Distribution* **147**(6), 355–360.

Coleman, T. and Li, Y. Y. [1996] "An interior trust region approach for nonlinear minimization subject to bounds," *SIAM Journal on Optimazation*, **6**(2), 418–445.

Gavrila, D. M. [1999] "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding* **73**(1), 82–98.

Ge, S. S. and Fua, C. H. [2005] "Queues and artificial potential trenches for multi-robot formations," *IEEE Transactions on Robotics* **21**(4), 646–656.

Ge, S. S., Guan, F., Loh, A. P. and Fua, C. H. [2006] "Feature representation based on intrinsic structure discovery in high dimensional space," *The 2006 IEEE International Conference on Robotics and Automation*, Orlando, Florida, May 15–19 2006, pp. 3399–3404.

Ge, S. S. [2007] "Social robotics: Integrating advances in engineering and computer science," *Proceedings of Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology International Conference*, Chiang Rai, Thailand, May 9–12 2007, pp. xvii–xxvi.

Haritaoglu, I.. Harwood, D. and Davis, L. S. [2000] "W^4: Real-time surveillance of people and their activities," *IEEE Transcations on Pattern Analysis and Machine Intelligence* **22**(8), 809–830.

Hsu, R. L., Abdei-Mottaleb, M. and Jain, A. K. [2002] "Face dectetion in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(5), 696–706.

Kakuta, N., Yokoyama, S. and Mabuchi, K. [2002] "Human thermal models for evaluating infrared images," *IEEE Engineering in Medicine and Biology Magazine* **21**(6), 65–72.

Li, L. Y., Huang, W. M., Gu, I. Y. H. and Tian, Q. [2004] "Statistical modeling of complex background for foreground object detection," *IEEE Transactions on Image Processing* **13**(11), 1459–1472.

Loh, A. P., Guan, F. and Ge, S. S. [2004] "Motion estimation using audio and video fusion," *Proceedings of the 8th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Kunming, China, pp. 1569–1574.

Luca, G., Marcenaro, L. and Regazzoni, C. S. [2002] "Automatic detection and indexing of video-event shots for surveillance applications," *IEEE Transactions on Multimedia* **4**(4), 459–471.

Maldague, X. P. V. [1994] *Infrared Methodlogy and Technology*, ser. Nondestructive Testing Monographs and Tracts, McGonnagle, W. J. (ed.), Switzerland: Gordon and Breach Sicence Publishers **7**.

Maragos, P., Schafer, R. W. and Butt, M. A. (eds.) [1996] *Mathematical Morphology and its Applications to Image and Signal Processing*, Boston, Kluwer Academic.

McKenna, S. J., Jabri, S., Duric, Z. and Rosenfeld, A. [2000] "Tracking groups of people," *Computer Vision and Image Understanding* **80**, 42–56.

Mohan, A., Papageorgiou, C. and Poggio, T. [2001] "Example-based object detection in images by components," *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, **23**(4), 349–361.

Nanadhakumar N. and Aggarwal, J. L. [1988] "Integrated analysis of thermal and visual images for scene interpretation," *IEEE Transcations on Pattern Analysis and Machine Intelligence* **10**(4), 469–480.

Socolinsky, D. A., Wolff, L. B., Neuheisel, J. D. and Eveeland, C. K. [2001] "Illumination invariant face recognition using thermal infrared imagery," *IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, December 11–13, pp. 527–534.

Trivedi, M. M., Cheng, S. Y., Childers, E. M. C. and Krotosky, S. J. [2004] "Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation," *IEEE Transations on Vehicular Technology* **53**(6), 1698–1712.

Trucco, E. and Verri, A. [1998] *Introductory Techniques for 3-D Computer Vision*, Upper Saddle River, New Jersey: Prentice Hall.

Tsuji, T., Hattori, H., Watanabe, M. and Nagaoka, N. [2002.] "Development of night-vision system," *IEEE Transactions on Intelligent Transportation Systems* **3**(3), 203–209.

Waxman, A. M., Fay, D. A., Gove, A. N., Seibert, M., Racamato, J. P., Carrick, J. E. and Savoye, E. D. [1995] "Color night vision: Fusion of intensified visible and thermal IR imagery," *Proc. SPIE Vol. 2463, p. 58–68, Synthetic Vision for Vehicle Guidance and Control*, Jacques G. Verly; (ed.), ser. Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, Verly, J. G. (ed.) **2463**, 58–68.

Wren, C. R., Azarbayejani, A., Darrell, T. and Pentland, A. P. [1997] "Pfinder: Real-time tracking of the human body," *IEEE Transcations on Pattern Analysis and Machine Intelligence* **19**(7), 780–785.

Zhang, B.-L., Zhang, H. and Ge, S. S. [2004] "Face recognition by applying wavelet subband representation and kernel associative memory," *IEEE Transactions on Neural Networks* **15**(1), 166–177.

Biography

Feng Guan received the BEng from Wuhan University of Hydraulics and Electrics, Wuhan, China, in 1997 and MEng from Shanghai Jiaotong University, Shanghai, China, in 2001, respectively. During 2001 to 2006, he is a research scholar working toward his PhD degree and then a research engineer in the Department of Electrical and Computer Engineering, National University of Singapore. His current research interests are in the fields of sound localization and sensor fusion.

Shuzhi Sam Ge, IEEE Fellow, P.Eng, is a professor at Department of Electrical and Computer Engineering, the National University of Singapore. He received his BSc degree from Beijing University of Aeronautics and Astronautics (BUAA), and the PhD degree and the Diploma of Imperial College (DIC) from Imperial College of Science, Technology and Medicine. He has (co)-authored three books: *Adaptive Neural Network Control of Robotic Manipulators* (World Scientific, 1998), *Stable Adaptive Neural Network Control* (Kluwer, 2001) and *Switched Linear Systems: Control and Design* (Springer-Verlag, 2005), and over 300 international journal and conference papers. He has served/been serving as an associate editor for a number of flagship journals including *IEEE Transactions on Automatic Control*, *IEEE Transactions on Control Systems Technology, IEEE Transactions on Neural Networks, and Automatica*. He also serves as an editor of the Taylor & Francis Automation and Control Engineering Series. His current research interests include social robotics, multimedia fusion, adaptive control, and intelligent systems.

Ai Poh Loh received the BEng (Electrical 1st class) from the University of Malaya, Kuala Lumpur, in 1983 and the DPhil degree in control from Oxford University in 1986. She is currently an associate professor and Deputy Head (Academic) at the Department of Electrical and Computer Engineering at the National University of Singapore. From 1994 to 1997, she was a visiting lecturer at MIT. Her research interests include auto-tuning, fault detection, signal processing and nonlinear adaptive control.