ML-Fusion based Multi-Model Human Detection and Tracking for Robust Human-Robot Interfaces

Liyuan Li[1], Jerry Kah Eng Hoe[1], Shuicheng Yan[2], and Xinguo Yu[1]
[1]Institute for Infocomm Research, Singapore
[2]National University of Singapore, Singapore
{lyli,kehoe,xinguo}@i2r.a-star.edu.sg, eleyans@nus.edu.sg

Abstract

A novel stereo vision system for real-time human detection and tracking on a mobile service robot is presented in this paper. The system integrates the individually enhanced stereo-based human detection, HOG-based human detection, color-based tracking, and motion estimation for the robust detection and tracking of humans with large appearance and scale variations in real-world environments. A new framework of maximum likelihood based multi-model fusion is proposed to fuse these four human detection and tracking models according to the detection-track associations in 3D space, which is robust to the possible missed detections, false detections, and duplicated responses from the individual models. Multi-person tracking is implemented in a sequential near-to-far way, which well alleviates the difficulties caused by human-over-human occlusions. Extensive experimental results demonstrate the robustness of the proposed system under real-world scenarios with large variations in lighting conditions, cluttered backgrounds, human clothes and postures, and complex occlusion situations. Significant improvements in human detection and tracking have been achieved. The system has been deployed on six robot butlers to serve drinks, and showed encouraging performance in open ceremony events.

1. Introduction

One of the essential functionalities for a mobile social robot is to continuously detect and track humans in its view. Recently, many social robots are emerging where the ability to interact with humans is an important component. Most existing systems employ a 1D laser sensor to detect human legs and apply simple image models (*e.g.*, skin color, face, head-shoulder contour, and motion) to find humans in images [2, 12]. The mean-shift, Kalman filters, or particle filters are usually used to track faces or skin blobs in image sequence [1]. In [13], disparity and color measurements are merged for human tracking. The employed algorithms are very efficient but too simple to be robust under varying real-world environments and crowded scenarios.

Recently, great progress in detecting human bodies within images has been achieved [3, 5, 9, 15, 16]. The state-of-the-art methods can capture the visual features of human appearances from a large training set by popular machine learning techniques. To detect humans of different sizes within an image, multi-scale window scanning is required, which results in high computational cost. Moreover, these algorithms are easy to produce duplicated detections around a true human under crowded scenarios.

With the availability of low price stereo cameras, stereo-based human detection becomes an interesting topic in computer vision. Most existing methods are based on the bottom-up depth segmentation of human bodies for human detection (*e.g.* [20]). As the disparity data on human bodies are often incomplete and/or inaccurate, and even merged together under cluttered or crowded scenes, these methods are unreliable under real-world scenarios. A top-down method to detect humans from a disparity image based on scale-adaptive filtering has been proposed in [11]. The top-down method is robust to the imperfection of disparity data, but the computational cost is high.

On the other hand, there have been tremendous literatures on visual object tracking through image sequence [18]. Typically, in a tracking approach, a visual target (color blob, feature point set, or object contour) is located in a new image frame by Kalman filters, particle filters, or mean-shift algorithm (see [9, 18]). To handle the interactions of multiple objects, a more complicated optimization process has to be introduced [8]. Recent multi-object tracking approaches (*e.g.* [17]) achieved more satisfactory performance in cases with occlusions, their computational cost however would increase significantly when many objects exist in crowded scenarios. With the impressive progress of human detection, several recent approaches have attempted to combine detection and tracking [8, 14, 17, 19], which are particularly promising for mobile platform. In the latest systems [4, 6],

978-1-4244-5498-3/09/$25.00 ©2009 IEEE

stereo cue was further used to suppress false detections from image-based detectors. However, the detection rate would reduce if the human evidence is poor from stereo measures.

In this paper, we present a novel system to integrate the individually enhanced stereo-based human detection, image-based human detection, color-based tracking, and motion estimation for robust human detection and tracking on a mobile service robot. It is motivated from the following observations:

1. Stereo-based human detection is robust to lighting changes and partial occlusions, but may fail if the disparity measures are poor, e.g., due to the textureless clothes, and also it may generate false detections in cluttered environments.
2. Image-based human detection is insensitive to the clothes colors, it may however produce duplicated responses around a real human instance in crowds.
3. Color-based tracking can connect the tracked humans in consecutive frames, but may degrade significantly in cluttered scenes or when encountering occlusions.
4. Motion estimation provides cues about partially or fully occluded humans when visual features are less reliable, but cannot well adapt to irregular movements.

By integrating all these models into one system, it is possible to compensate for each other and obtain satisfactory performance even through one or more models fail frequently. However, there exist two major difficulties in integrating them on a service robot. One is that the state-of-the-art human detection and tracking approaches require much computational resource, and the other is that the detection and tracking results from individual models may be inconsistent due to the possible false detections, missed detections, and duplicated detections. In this paper, first, the implementation of these two human detection approaches is substantially speeded up. Then, a Maximum Likelihood (ML) based multi-model fusion approach is proposed to integrate the outputs from all the four models for human detection and tracking, and implemented under a sequential near-to-far way for handling complex occlusions. The novelties of this paper can be summarized as follows:

1. Substantial improvements of individual components in both effectiveness and efficiency.
2. A framework of ML-based fusion to integrate possibly inconsistent results from different models.
3. An efficient system of integrating stereo- and HOG-based human detections, color-based tracking, and motion estimation for robust human detection and tracking in real-world environments.

Significant performance improvements in both human detection and tracking have been achieved from this ML-based fusion process. The system has been successfully deployed on six real robots for real-world events.

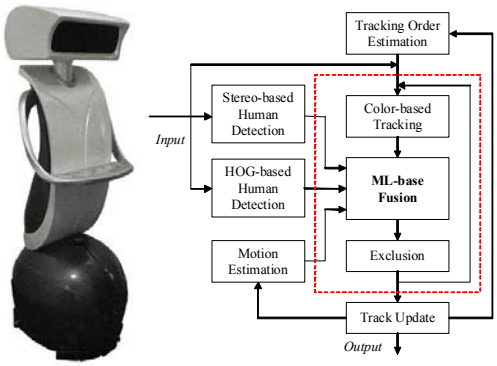

Figure 1. Illustration of the robot butler equipped with our vision system and its block diagram.

The rest of this paper is organized as follows. Section 2 gives a brief overview of our system. The details of four individually enhanced models and the ML-fusion framework are described in Section 3-7. The extensive experiment results are shown in Section 8, and the conclusive remarks are drawn in Section 9.

2. System Overview

The diagram of our proposed vision system and an example robot butler equipped with this system are shown in Figure 1. The input to this system is a pair of color and disparity images from a stereo camera on the robot head. For each frame, the blocks in the diagram are executed as follows. First, the humans in the view are detected from the disparity and color images. Meanwhile, the positions of occluded humans are predicted based on their temporal histories. To handle possible complex occlusions, multi-person tracking is performed sequentially from the closest one to the farthest (including the fully occluded ones) to approximate a globally optimal tracking process [10]. The order is determined by the predicted 3D positions of all the tracked humans. Then, the blocks within the dotted box of the diagram in Figure 1 are executed to track humans one by one. For each human, a mean-shift tracking is performed first, and then the new position is located in the image by the ML-based fusion. Finally, the exclusion step is performed to suppress the visual features of the tracked human in both color and disparity images. This operation is to avoid other humans being trapped in the positions of those tracked humans. When the sequential multi-person tracking is completed, the system updates the 3D positions and appearance models of the tracked humans, as well as the initializations and terminations of the tracks.

To our knowledge, it is the first system to integrate stereo-based and HOG (Histogram of Oriented Gradients) based human detections, color-based human tracking and motion estimation for a mobile platform and deployed on real robot for successful real-world demonstrations.

3. Stereo-based Human Detection

The disparity image from a stereo camera provides the 2.5D information of the observed objects in the view. Since each human object standing on the ground surface occupies certain physical volume in the 3D space, a top-down method for stereo-based human detection [11] is adopted in this system. To achieve better performance in both accuracy and efficiency, three improvements are made in this work: (a) introducing the logarithm model for the distance-disparity relation; (b) ML-based multi-person segmentation; and (c) multi-scale processing for better efficiency.

Theoretically, the relation between the depth distance (z) and the disparity (d) is described as $z = bf/d = K_1/d$, where b is the base-line and f is the focal length. In practice, K_1 is not a constant, resulting in small effective depth range of detection. By examining the curves of K_1 with respect to d from six stereo cameras, we propose the use of the logarithm model for describing the K_1 coefficient, *i.e.*

$$z = K_1(d)/d, \text{ with } K_1(d) = K_d \log d + B_d, \quad (1)$$

where K_d and B_d are the model parameters. For each stereo camera, the model can be obtained from offline calibration. The new model can extend the effect depth range for human detection more than 2 times, *e.g.*, for a 2.8mm lens 9cm baseline STOC camera from Videre Design, the effect depth range is extended from 1.2m-3.2m to 0.8m-5.5m.

Significant evidence of the standing humans can be observed from the projection of disparity values on the ground plane. Let $d(x, y)$ be the disparity image and $h(x, d)$ be the projection of it on x-d plane. Human evidence can be enhanced by scale-adaptive filtering [11]. In the filtered 2D histogram $\tilde{h}(x, d)$, disparity clouds of humans are clustered as significant, smooth, and separated peaks.

In [11], disparity pixels are segmented into a human region according to heuristic rules. Its performance degrades when the disparity measurements are poor or humans are close to each another. In this paper, human segmentation is performed under the ML framework.

The human candidates in the image are located at the significant peaks in $\tilde{h}(x, d)$ from the close to the far depth positions to the camera. Let $\mathbf{c}_i = (x_i, d_i)$ be the position of the ith significant peak. Suppose W_b and D_b are the average width and thickness of human bodies. The width and depth range of the ith human in the image can be estimated as $w_b(d_i) = fW_b d_i/K_1(d_i)$ and $d_\mp = \frac{K_1(d_i)d_i}{K_1(d_i)\pm d_i D_b/2}$. Then the cloud of the disparity data for the human would be within the bounding box $R_i = [x_-, d_-, x_+, d_+]$ in the x-d plane, where $x_\mp = x_i \mp w_b(d_i)/2$. A Gaussian distribution $\mathcal{N}(\mathbf{m}_i, \Sigma_i)$ can be obtained from the data $h(x, d)$ within the bounding box R_i, where $\mathbf{m}_i$ is the mean vector and Σ_i is the covariance matrix. In segmentation, for a pixel $d(x, y)$ in the disparity image with the feature vector $\mathbf{v} = (x, d(x, y))$, it is assigned to a human region according to the maximum likelihood probability

$$l_\mathbf{v} = \arg\max_i P_i(\mathbf{v}) \quad \& \quad P_i(\mathbf{v}) > T_1, \quad (2)$$

where $P_i(\mathbf{v}) = (2\pi)^{-1}|\Sigma_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{v}-\mathbf{m}_i)\Sigma_i^{-1}(\mathbf{v}-\mathbf{m}_i)^T}$.

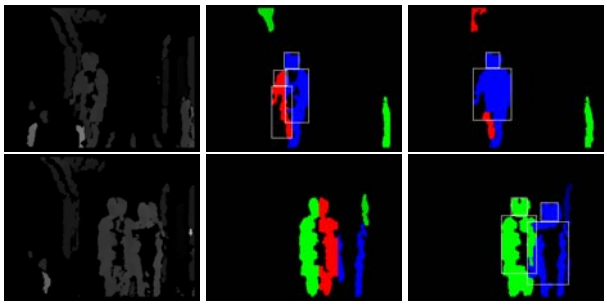

Figure 2. Two examples on the improvement in human segmentation performance from our approach compared with the approach in [11]. For each row, the images from the left to right are 1) the disparity image, 2) the segmentation result by [11], and 3) segmentation result by our approach.

Two examples to show the improvement in human segmentation performance from our proposed approach over the approach [11] are shown in Figure 2, from which we can see that our proposed approach is more robust to the poor disparity measurements for both isolated human and peoples in a group. The segmented humans are further verified according to the head-shoulder contour by deformable template matching as in [11].

As it is unnecessary to perform human detection in high resolution image [3], the close humans are detected in a smaller size image while the far away humans are detected in the original image. In such way, the detection approach is speeded up by three times.

4. HOG-based Human Detection

To compensate the errors of stereo-based human detection from disparity image, HOG-based human detection [3] from color image is adopted in this system. It is implemented with three characteristics: 1) three HOG models being used; 2) efficient detection under perspective constraint; and 3) clustering of the detections on depth measurements.

Three Models. When a human is very close to the robot for near face-to-face interaction, only the head and upper body are in the view, but when a human is quite far away from the robot, the full body is in the image. To be adaptive to such large variations of human appearances in the view, three HOG models of human appearances are used, *i.e.*, the *upper body*, *2/3 body*, and *full body* models. They are trained offline with 124, 350, and 746 ground truth samples, respectively, where the positive and negative samples are almost balanced.

Efficient Implementation. To be efficient in multi-model multi-scale HOG-based human detection, ground plane constraints [6, 7] are exploited. The 3D space in front of the robot is divided into a set of discrete layers. The humans in the closest layer appear with w_b^0 as average body width in the image. The next layers are selected according to $w_b^{k+1} = \gamma w_b^k$ with $\gamma < 1$ ($\gamma = 0.7$ in this system) until the smallest scale reached. With $w_b(d) = fW_b d/K_1(d)$ and (1), the disparity value (depth distance) for each layer can be computed using Newton's method. Then, for the kth layer, by using the perspective constraints, we can select the HOG model and the size of detection window, and locate the head top in the image. To detect humans in this layer, the detection window scans the image from the left to the right with $\triangle x = w_b^k/3$ as the shift gap, and at each horizontal position, the window slides vertically at 3 positions for human heights of about 1.9m, 1.75m, and 1.6m.

Let N be image width, w_b^0 and w_b^K be the maximum and minimum scales. The number of windows for the kth layer is $n_W^k = 3[N/(w_b^k/3)]$ and the number of layers is $K = 1 + \log(w_b^K/w_b^1)/\log\gamma$. Hence, the number of total detection windows is $n_W = \sum_{k=1}^{K} n_W^k$. For an image of 320×240 pixels, if $w_b^0 = 180$ and $w_b^K = 24$, the number of detection windows is 287, much less than conventional method (*e.g.* [15]) which requires over 200,000 detection windows for multi-scale scanning.

Clustering of Detections. HOG detector may produce duplicated detections for one target. The overlapping windows are grouped as follows. For the window cluster of the same scale, the one with maximum response is selected. For the overlapping windows of different scales, the one with minimum distance between the layer depth to the average depth within the window is select. Detections with too low disparity evidence are deleted. In this way, many redundant and false detections are suppressed.

5. Color-based Tracking

It was shown in [10] that the DCH (Dominant Color Histograms) performs better than conventional color histograms in differentiating the same target object from other objects under varying lighting conditions, which can benefit the human tracking task in the mobile platform. Hence, the DCH-based mean-shift tracking approach [10] is adopted in this system for the tradeoff of accuracy and efficiency. Since background subtraction is not applicable in mobile setting, we may not obtain explicit human segmentation in each frame. Therefore, DCH representation of the torso or head of a human is extracted from the core ellipse of the bounding box. The color blobs of torso and head are tracked by the mean-shift algorithm and coupled once they are separated too far away. The depth information, *i.e.*, the average disparity of body d_n^T, and disparity range $d_{n\pm}^T$ of a tracked human (*e.g.* the nth human) are roughly estimated from the new position with measurements of the occluding humans having been excluded.

Table 1. A summary of the outputs from four models.

Stereo	$d_i^S, d_{i\pm}^S, \mathbf{s}_{ih}^S, \mathbf{s}_{it}^S, B_{ih}^S, B_{it}^S, B_{ib}^S, i \in [1, N_S]$
HOG	$d_j^I, d_{j\pm}^I, \mathbf{s}_{jh}^I, \mathbf{s}_{jt}^I, B_{jh}^I, B_{jt}^I, B_{jb}^I, j \in [1, N_I]$
Tracking	$d_n^T, d_{n\pm}^T, \mathbf{s}_{nh}^T, \mathbf{s}_{nt}^T, B_{nh}^T, B_{nt}^T, B_{nb}^T, n \in [1, N_T]$
Motion	$d_n^M, d_{n\pm}^M, \mathbf{s}_{nh}^M, \mathbf{s}_{nt}^M, B_{nh}^M, B_{nt}^M, B_{nb}^M, n \in [1, N_T]$

6. Motion Estimation

Since the stereo camera is used in this system, humans are tracked in a 3D space. For a human in the view, we can label the perspective rays from his left and right sides to the camera as l_l and l_r. The angle between the line of l_l or l_r to the camera optical axis can be denoted as α_l or α_r. Suppose the humans H_A and H_B are in the view and $Z_A < Z_B$, the probability of human H_B being occluded by H_A can be computed as

$$P_o(B/A) = \begin{cases} \frac{\alpha_l^A - \alpha_r^B}{\alpha_l^B - \alpha_r^B}, & \text{if } \alpha_r^B < \alpha_l^A < \alpha_l^B; \\ \frac{\alpha_l^B - \alpha_r^A}{\alpha_l^B - \alpha_r^B}, & \text{if } \alpha_r^B < \alpha_r^A < \alpha_l^B; \\ 1, & \text{if } \alpha_l^B < \alpha_l^A \ \& \ \alpha_r^B > \alpha_r^A; \\ 0, & \text{others.} \end{cases}$$

Here, $P_o = 1$ means fully occluded, $0 < P_o < 1$ means partially occluded, and $P_o = 0$ means no occlusion. If more than one third body of a tracked human is occluded, *i.e.* $P_o > 0.33$, his/her position is predicted only according to the historical positions before occluded with a linear model.

7. ML-Fusion based Detection and Tracking

The above models locate humans based on different measurements, *i.e.* (a) stereo-based approach detects humans based on the disparity measures; (b) HOG-based approach detects humans based on the edge measures in color image; (c) color-based tracking is based on the color regions; and (d) motion estimation is based on the track's temporal history in the 3D space. They provide compensational and redundant information on the humans in the view. The object representations in different models are different, but they are transformed to the same format for each human. It consists of the average disparity value, the range of disparities, the centers of head and torso, the bounding boxes of head, torso, and full body in the 2D image, as shown in Table 1. The output of stereo-based human detection is from the explicit human segmentation, while the others are estimated from the locations according to average human metrics or previous tracking records.

Due to the false detections, missed detections, and duplicated detections from the human detectors, position bias from mean-shift tracking, and inaccuracy from motion prediction, the outputs of the four models would often be in-

consistent to each another. One-to-one assignment of detection to track is often impossible and inaccurate. An ML (Maximum Likelihood) based fusion approach is derived to integrate such possibly inconsistent measurements.

7.1. ML-based Fusion

Suppose we obtain R human candidates $(\mathbf{x}_1, \cdots, \mathbf{x}_R)$ from the above four models. Here, R may not be exactly four times of the true human number in the view due to the possible false detections, missed detections, and duplicated detections. With such an observation, the likelihood of the kth human's position, $\mathbf{x}_k^*$, in the image can be expressed as $P(\mathbf{x}_1, \cdots, \mathbf{x}_R|\mathbf{x}_k^*)$. Because the human candidates $(\mathbf{x}_1, \cdots, \mathbf{x}_R)$ are generated by different models from different measurements, it can be assumed that they are conditionally independent. The joint likelihood probability can be written as $P(\mathbf{x}_1, \cdots, \mathbf{x}_R|\mathbf{x}_k^*) = \prod_{i=1}^{R} P(\mathbf{x}_i|\mathbf{x}_k^*)$. Ideally, if $\mathbf{x}_k^*$ is the true position of the kth human in the image, the likelihood reaches its maximum at the position with $\partial P(\mathbf{x}_1, \cdots, \mathbf{x}_R|\mathbf{x}_k^*)/\partial \mathbf{x}_k^* = 0$. Since

$$\frac{\partial \prod_{i=1}^{R} P(\mathbf{x}_i|\mathbf{x}_k^*)}{\partial \mathbf{x}_k^*} = \left(\prod_{j=1}^{R} P(\mathbf{x}_j|\mathbf{x}_k^*)\right) \sum_{i=1}^{R} \frac{1}{P(\mathbf{x}_i|\mathbf{x}_k^*)} \frac{\partial P(\mathbf{x}_i|\mathbf{x}_k^*)}{\partial \mathbf{x}_k^*},$$

we can obtain

$$\sum_{i=1}^{R} \frac{1}{P(\mathbf{x}_i|\mathbf{x}_k^*)} \frac{\partial P(\mathbf{x}_i|\mathbf{x}_k^*)}{\partial \mathbf{x}_k^*} = 0. \tag{3}$$

Without lose of the generality, suppose $P(\mathbf{x}_i|\mathbf{x}_k^*)$ follows Gibbs distribution, specifically as

$$P(\mathbf{x}_i|\mathbf{x}_k^*) = \frac{1}{Z_{ik}} e^{-\beta_{ik} E(\mathbf{x}_i, \mathbf{x}_k^*)} = \frac{1}{Z_{ik}} e^{-\beta_{ik} \|\mathbf{x}_i - \mathbf{x}_k^*\|^2}, \tag{4}$$

where Z_{ik} is a constant for normalization, β_{ik} indicates the association of detection $\mathbf{x}_i$ to the kth tracked human, and $E(\mathbf{x}_i, \mathbf{x}_k^*) = \|\mathbf{x}_i - \mathbf{x}_k^*\|^2$ is the energy term. Note that here β_{ik} is not a constant, and instead dependent on $\mathbf{x}_i$ as introduced afterward. Then, there is

$$\frac{\partial P(\mathbf{x}_i|\mathbf{x}_k^*)}{\partial \mathbf{x}_k^*} = 2\beta_{ik}(\mathbf{x}_i - \mathbf{x}_k^*) P(\mathbf{x}_i|\mathbf{x}_k^*). \tag{5}$$

Substituting (5) into (3), we can obtain

$$\mathbf{x}_k^* = (\sum_{i=1}^{R} \beta_{ik} \mathbf{x}_i) / \sum_{i=1}^{R} \beta_{ik}. \tag{6}$$

This result indicates that, under the Maximum Likelihood (ML) formulation, the estimated true position is the weighted average of the detected and estimated positions according to their associations to the tracked human.

7.2. Sequential Implementation

To achieve near-optimal multi-object tracking performance with low computational cost, the sequential tracking strategy [10] is employed in this system. The basic philosophy behind the strategy is that if we track the humans in the image sequentially from the closest one to the furthest one, we could approximate the optimal result since the appearance of the closer human in the image is not affected by those occluded by him/her.

In this system, the order for tracking is determined by the depth distances of the humans and the visible clues from human detection and tracking. A measure of the priority for tracking is defined as

$$O_k = O_Z^k + (1 - P_o^k) P_c^k + A_S^k + A_I^k, \tag{7}$$

where $O_Z^k = 1 - Z_k/Z_{max}$ describes the priority on the depth distance ($Z_{max} = 6m$ in this system), P_o^k from (3) is the proportion of the occluded body part, P_c^k is the likelihood of observing the human from his/her visible part [10], A_S^k and A_I^k are the maximum associations to the stereo- and HOG-based detections. These measures come from the tracking results in the previous time step. Obviously, a closer human with less occluded part and frequently detected in the recent frames has a higher chance to be tracked correctly and reliably in the current frame. Humans are tracked sequentially according to the priority values $\{O_k\}_{k=1}^{N_T}$. In each iteration, three operations are performed, namely the color-based tracking, multi-model fusion, and exclusion. For the kth tracked human, the color-based mean-shift tracking is first performed.

Now we have the outputs from four models as shown in Table 1. Since we do not know the true position $\mathbf{x}_k^*$, we cannot compute the association β_{ik} directly. Here, we estimate β_{ik} from the correlation between the detected position and the position from tracking, since this measure is robust even if both detectors fail. The correlation of the ith human detected by stereo-based human detection with the kth tracked human can be defined as

$$a_{ik}^S = \alpha_s (w_h R_{oh}^{ik} + w_t R_{ot}^{ik}) + (1 - \alpha_s) R_{od}^{ik}, \tag{8}$$

where α_s is a weight to balance 2D spatial and depth matches ($\alpha_s = 0.75$ in this system as the disparity resolution is low), w_h and w_t are the weights for the head and torso ($w_h = |B_{kh}^T|/(|B_{kh}^T| + |B_{kt}^T|)$ and $w_t = 1 - w_h$), R_{oh}^{ik} and R_{ot}^{ik} are the overlapping rates of head and torso bounding boxes between the detected and tracked humans, and R_{od}^{ik} is the overlapping rate of disparity ranges. The overlapping rate is defined as the ratio of the intersection to the union of two boxes or range segments. To avoid the partially and fully occluded human being locked on the occluding human, the association of the detection with the tracked human is finally defined as

$$\beta_{ik}^S = (1 - P_o^k)(1 - P_{apt}^i) P_{hp}^{ik} a_{ik}^S, \tag{9}$$

where P_o^k from (3) represents the proportion of occluded body, $P_{apt}^i = \sum_{l=1}^{k-1} \beta_{il}^S$ is the accumulated association of

the detection with previous tracked humans who might occlude the kth tracked human, and P_{hp}^{ik} is the likelihood of that the heights of the detected and tracked human are close. P_{apt}^{i} is used for excluding duplicate association. P_{hp}^{ik} is used to prevent the tracked human from jumping to a close false detection in cluttered background when the true detection is missed. It is defined as $P_{hp}^{ik} = \exp\{-\frac{(H_i - \bar{H}_k)^2}{\sigma_H^2}\}$, where H_i and $\bar{H}_k$ are the heights of heads of the detected and tracked humans in 3D space and σ_H is the variance according to average human metrics.

Since the HOG-based human detector only returns a bounding box of the detected human in the image, the correlation of the jth detected human by HOG-based human detection with the kth tracked human is defined as

$$a_{jk}^{I} = \alpha_s R_{ob}^{jk} + (1 - \alpha_s) R_{od}^{jk}, \tag{10}$$

and the final association is defined as

$$\beta_{jk}^{I} = (1 - P_o^k)(1 - P_{apt}^{j}) P_{hp}^{jk} a_{jk}^{I}. \tag{11}$$

If the kth tracked human was occluded in the previous frames, the output from the motion prediction model should be a good estimation of the new position since less visual clue is available in recent frames. The association of the motion estimation with the true new position can be defined as $\beta_{kk}^{M} = P_o^k$. On the other hand, if the strong visual evidence can be observed from the position located by color-based tracking, we should have high confidence on the output of color-based tracking. The association of the tracked position with the true new position can be defined as $\beta_{kk}^{T} = (1 - P_o^k) P_c^k$, where P_c^k is the likelihood of being observed obtained from the color-based tracking model.

With the associations of the outputs from four models with the tracked humans, the ML-based fusion (6) can be applied as

$$\hat{\mathbf{x}}_k^* = \frac{\beta_{kk}^{M}\mathbf{x}_k^{M} + \beta_{kk}^{T}\mathbf{x}_k^{T} + \sum_{i=1}^{N_S}\beta_{ik}^{S}\mathbf{x}_i^{S} + \sum_{j=1}^{N_I}\beta_{jk}^{I}\mathbf{x}_j^{I}}{\beta_{kk}^{M} + \beta_{kk}^{T} + \sum_{i=1}^{N_S}\beta_{ik}^{S} + \sum_{j=1}^{N_I}\beta_{jk}^{I}}. \tag{12}$$

It is used to estimate the new positions of the head center ($\mathbf{x}_k = \mathbf{s}_{kh}$) and torso center ($\mathbf{x}_k = \mathbf{s}_{kt}$), the average depth position ($\mathbf{x}_k = d_k$) and depth range ($\mathbf{x}_k = d_{k\pm}$), and the 2D sizes of head and torso.

The last operation in one iteration of the sequential tracking is the *exclusion* which suppresses the measures of the kth tracked human from its new position in input images.

Once all the humans in the list have been tracked, their states in 3D world space and 2D image space are updated. The color model of a tracked human is updated when it has high associations with detections from both detectors, such that the model would be adaptive to human appearance and lighting variations but not easy to be distorted in cluttered and crowded scenes. If a human without occlusions does not have enough associated detections in 1-3 seconds, the track will be terminated. If a new human is detected, it is first compared with the occluded humans according to the color similarity and expected positions. If it is not matched to any occluded human, a new track is initialized; otherwise, the track of the occluded human is resumed.

8. Experiments

Robot Vision System

The vision system integrating all aforementioned models was implemented on six robot butlers. Each unit is composed of an onboard PC104 nano (C2D 2.0GHz 2GB RAM) and a STOC stereo camera from Videre Design. The camera has a chip for stereo computation so that it provides disparity images at 25-30fps. Our system runs at 11fps on average. For an empty scene, the frame rate can be over 15fps, and for a crowded scene, the frame rate can still reach 7-8fps. The parameters of detection and tracking models are well evaluated in [11, 3, 10], those in fusion part are chosen as [11, 10]. The system's performance is less sensitive to the slight changes of them.

The robots have been tested in various environments and scenarios before open to public. Then, in a series of ceremony events, they successfully performed tasks of serving drink to guests waving hands to them. One robot can serve up to five humans one-by-one each time before coming back to station for reloading. Some results and evaluations in complex real-world scenarios from the test runs are presented in this section. The system has been evaluated on 11 difficult sequences containing 3250 frames. The sequences were recorded through USB port during the tests of the robots. Hence, the frame rate (at 5-8fps) is lower than that of running without recording. All the benchmark data and results will be publicly available on our website.

Human Detection Results

The most important task for Human-Robot Interfaces (HRI) is to perform human detection and know how many and where the humans are in the view. Due to the variations of lighting conditions, backgrounds, human shapes and poses, clothes, occlusions, and distances to the robots, the individual human detection and tracking models often generate errors of missed detections, false detections, and duplicated detections. However, when integrated under the ML-fusion framework, such errors can be reduced significantly. Four examples from different complex crowded scenarios are shown in Figure 3. From these examples, we can see that although no individual human detector can always generate perfect detections, the final detections are correct in both 2D image space and 3D world space to the robot. Similar to [4], we evaluated the human detection on all the test sequences in every fourth frame, *i.e.* on totally 736 frames containing 1420 ground truths. The summary of the performance and comparison results are shown in Table 2, where DR

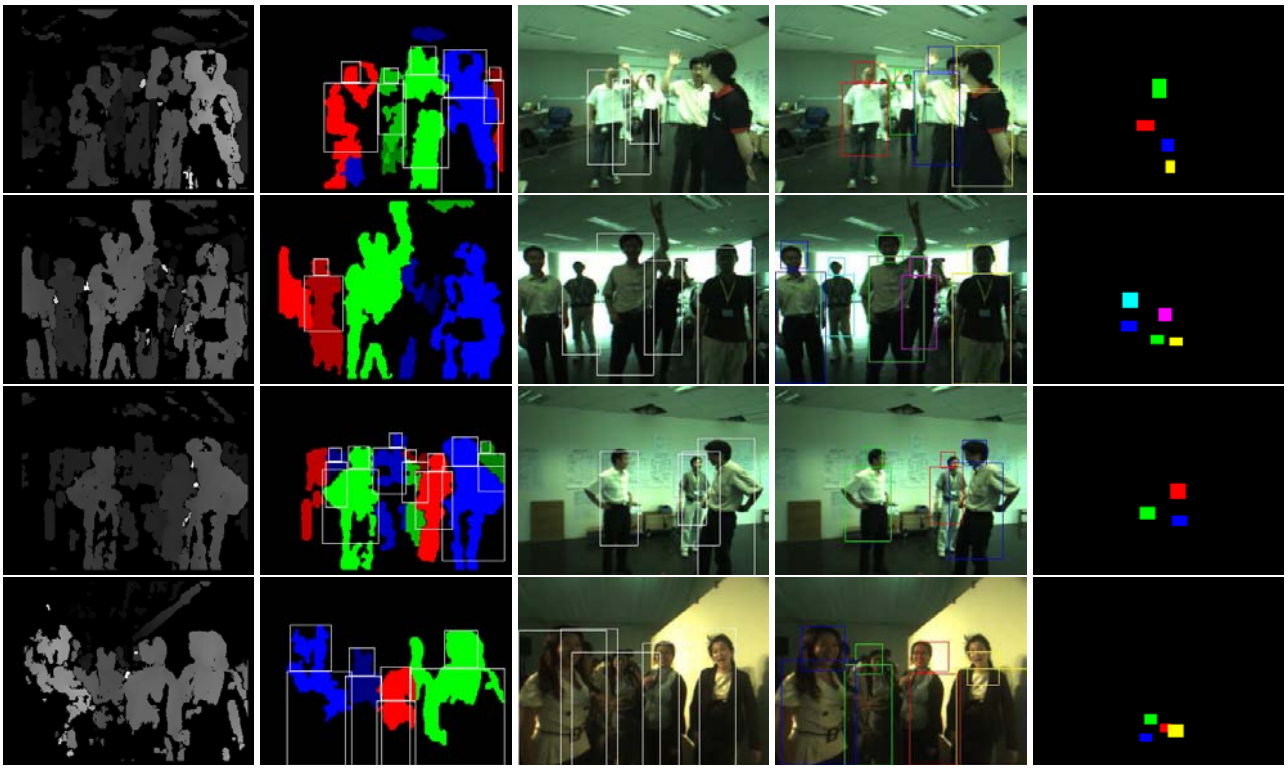

Figure 3. Examples of single-frame human detection from crowded scenarios. In each row, the images from the left to the right are: 1) the disparity image, 2) stereo-based human detection, 3) HOG-based human detection, 4) final human detection, and 5) the human locations on the ground plane viewed upright from the middle of the bottom line. For better viewing, please see the color pdf file. Note that the third row is to show the algorithmic performance under the scenario with cluttered background.

Table 2. The comparison results on human detection accuracies for stereo-based human detection (SHD), HOG-based human detection, and our proposed ML-fusing approach.

Method	DR	MDR	FPPF
SHD	71.5%	1.00	0.82
HOG	85.9%	1.07	0.08
Ours	94.1%	1	0.17

is detection rate (#detected truths/#ground truths), MDR is multi-detection rate (#detections on truths/#detected truths), and FPPF is false positives per frame. The low FPPF from HOG detector is due to that the disparity cue is exploited to suppress the false positives in our system.

Human Tracking Results

Human tracking in HRI is a necessary component for maintaining the interaction with the right human. For the robot butlers, it is required to track and approach the guests waving hands to them till within 1m for serving drinks. Four challenging scenarios are shown in Figure 4, including tracking four humans moving in high density, especially the yellow human moving between the red and blue humans (1st row), following a human walking around and partially occluded by the trolley (2nd row), tracking a human overlapping with two others, even fully occluded by the blue one (3rd row), and approaching a quite faraway human (pink one with red shirt) in crowds (4th row). Very high success rate is achieved in these real-world tests.

Table 3. The comparison results on human tracking performance.

	Tracks=86				Occlu. Events	
Method	MH	MM	AoC	Errors	q/p	n/m
HOG-Off	40	14	82.2%	34	41/48	18/22
SHD-Off	54	7	82.7%	36	32/48	12/21
Fusion	70	2	92.8%	9	58/65	23/29

The tracking performance was evaluated on all test sequences with two well-known criteria: the coverage of true trajectories [17, 4] and success rate through occlusion events [17]. The number of true trajectories (*including occluded durations*) and the numbers of mostly hit (MH, *i.e.* >80% covered) and mostly missed (MM, *i.e.* <20% covered) trajectories were counted. Overall, 92.8% of ground truth trajectories are covered by the tracker and the average life time of false tracks is 0.2 per frame. The numbers of occlusion events and humans involved (m and p), the numbers of success events and corrected tracked humans through the occlusion events (n and q) were recorded. The compari-

Figure 4. Example frames from tracking scenarios. 1st row: tracking four humans moving in high density; 2nd row: following a human walking around and partially occluded by the trolley; 3rd row: tracking a humans overlapping with two others; and 4th row: approaching a quite faraway human (pink one with red shirt) in crowds. For better viewing, please see the color pdf file.

son with tracking results obtained by switching off one of the human detectors was also made. The final statistics are shown in Table 3, where *Errors* include those caused by occlusions. The fewer of occlusion events recorded when stereo-based or HOG-based detector is switched off is due to the missed detections. It can be seen that significant improvement in performance has been achieved.

9. Conclusions and Future Work

In this paper, we presented a real-time vision system which integrates the stereo-based human detection, HOG-based human detection, color-based tracking, and motion estimation for robust human detection and tracking on a mobile service robot. An ML-fusion framework was proposed and implemented in a sequential near-to-far way to take the occlusions into consideration for human detection and tracking. This multi-model fusing method is robust to the missed detections, false detections, and duplicated detections from individual models, and hence the system performance has been significantly improved. This vision system has been deployed in six robot butlers and performed successfully for serving drinks in public events. Our current system may fail when the humans are too close or faraway to the robot, and one of our future research directions is to exploit the possibility of utilizing Pan-Tilt-Zoom cameras for further extending the sensible range of the robot.

References

[1] H.Bohme, *et al*. An approach to multi-modal human-machine interaction for intelligent service robots. *Robotis&Autonomous Systems*, 2003. 1

[2] T. Braun, K. Szentpetery, and K. Berns. Detecting and following humans with a mobile robot. *EOS Conf. Industrial Imaging and Machine Vision*, 2005. 1

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005. 1, 3, 6

[4] A. Ess, B. Leibe, K. Schindler, and L. Gool. A mobile vision system for robust multi-person tracking. *CVPR*, 2008. 1, 6, 7

[5] P. Felzenszwalb, D. McAllester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. *CVPR*, 2008. 1

[6] D. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73(1):41–59, 2007. 1, 4

[7] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *CVPR*, 2006. 4

[8] B. Leibe, K. Schindler, N. Cornelis, and L. Gool. Coupled object detection and tracking from static cameras and moving vehicles. *TPAMI*, 2008. 1

[9] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. *CVPR*, 2005. 1

[10] L. Li, W. Huang, I. Gu, R. Luo, and Q. Tian. An efficient sequential approach to tracking multiple objects through crowds for real-time intelligent cctv systems. *IEEE T-SMC-B*, 38(5):1254–1269, 2008. 2, 4, 5, 6

[11] L. Li, Y. Koh, S. Ge, and W. Huang. Stereo-based human detection for mobile service robots. *ICARCV*, 2004. 1, 3, 6

[12] M. Michalowski and R. Simmons. Multimodal person tracking and attention classification. *ACM HRI*, 2006. 1

[13] R. Munoz-Salinas, E. Aguirre, and M. Garcia-Silvente. People detection and tracking using stereo vision and color. *IVC*, 2007. 1

[14] K. Okuma, A. Taleghani, N. Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. *ECCV*, 2004. 1

[15] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63(2):153–161, 2005. 1, 4

[16] O. Tuzel, F. Poriki, and P. Meer. Human Detection via Classification on Riemannian Manifolds. *CVPR*, 2007. 1

[17] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *IJCV*, 75(2):247–266, 2007. 1, 7

[18] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):1–45, 2006. 1

[19] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. *CVPR*, 2008. 1

[20] L. Zhao and C. Thorpe. Stereo- and neural network-based pedestrian detection. *IEEE T-ITS*, 1(3):148–154, 2000. 1