RJ 10274 (A0211-101) November 25, 2002
Computer Science

IBM Research Report

Video-CRM: Understanding Customer Behaviors in Stores

Ismail Haritaoglu, David Beymer, Myron Flickner
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Video-CRM: Understanding Customer Behaviors in Stores *

Ismail Haritaoglu, David Beymer, Myron Flickner
IBM Almaden Research Center. San Jose, CA 95120
ismailh@almaden.ibm.com

Abstract

This paper describes a real-time computer vision system to detect and track people in stores to understand retail customer behavior while shopping. We propose an approach to detect and track people and their body posture without using an explicit 3D human model using overhead narrow-baseline stereo cameras. The proposed method is based on a 3D silhouette of people that is constructed from a 2D silhouette. The 2D silhouette is detected by color and disparity background subtraction. Once the 3D silhouette is generated, people are identified with iterative segmentation of a 3D silhouette based on the topological structure of human body. Once people are detected, their appearance model based on color and shape are generated and tracked over multiple camera using their 2D trajectory continuity and their appearance models. A shape histogram, the distribution of relative positions of points on a 3D silhouette, is introduced to estimate the posture and body parts to understand the people and object interactions in stores, such as "customer picking a object from a shelf". The initial pilot studies show the real-time performance and accuracy of the system to understand customer behavior.

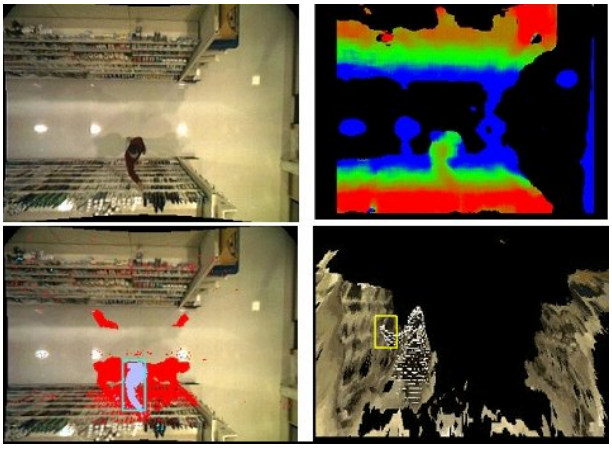

Figure 1: An example of understanding customer behavior in store: the color (top-left) and disparity image(top-right), the detection of 2D silhouette of people (in blue) and shadow removal (in red) (bottom-left), detection of "Pick Event" where 3D scene are constructed by disparity and the 3D silhouette of detected person (in white pixel).

1 Introduction

Leading edge retailers are using cameras to understand customer behavior to improve customer relationship management (CRM). CRM allows companies to understand their customers behavior for better customer understanding, such as, how customer interact with their brands in stores, product purchase decision, product promotion, and better customer satisfaction, such as, customer activity monitoring, improved operation efficiency and labor productivity, better store layout, self-service efficiency.

One challenging problem for retailers is to understand the interaction between customers and merchandise in the store to explore the shopping behavior. For example, a customer looks at an item on a shelf, picks up the item, looks for a price or information, then places it back on the shelf or in a shopping cart. In order to extract this information, we need to detect and track shoppers in the store, their body orientation, posture. In this paper, we describe a real-time system to detect and track people to understand their behavior in retail stores while shopping using overhead narrow baseline stereo cameras as shown in Figure 1(**movie1**).

Several challenging vision problem arose during our investigation of human shopping behavior. First, we need to detect each customer as soon as they enter the store. In many cases, people appear as a small groups and they move together in store where individual people may not visually isolated. We propose a method based on iterative segmentation and of a 3D silhouette of people to segment each individual people and determine their body orientation and posture. Second, we need to detect the head, shoulder, arms, and hands to detect and understand the "picking an object from shelf" event and other interactions. The computational models proposed in the system are based on observations obtained by analyzing human body structure. The human body has topological constrains on the relative location of body parts with respect to each other, e.g. head supported

* this paper includes color illustrations, please print this paper with color printer

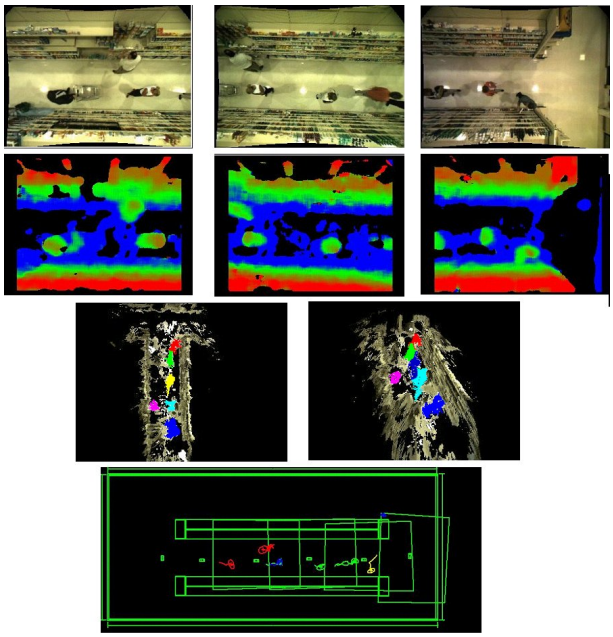

Figure 2: an instance of VCRM system with three camera monitoring an aisle in a retail store: color images (top-row), disparity images (second row), 3D reconstructed aisle and people segmentation in two different virtual view (third-row), and customer trajectories mapped floor plan (bottom-row)

by shoulder and torso, arms connected to shoulder. Third, we need to track them trough out entire store to recover the full trajectory to understand their moving pattern. That requires tracking each people over multiple camera. We combined appearance based tracking method with trajectory-continuity constraints in an hierarchical way to recover the full trajectory of each people over multiple camera. Appearance model is generated for each detected person using their color and shape of their 3D silhouette. The trajectory continuity constraints and appearance similarity allows us to resolve the tracking ambiguities while people leaving one camera's field of views and enter other's.

Our system consists of multiple narrow-baseline overhead digital stereo cameras [7] mounted on the 13 feet ceiling looking down to the floor as shown in Figure 2. The camera captures two images (left color and right monochrome) and a disparity image is computed using area based stereo matching [8]. The cameras are fully calibrated using an automatic calibration procedure developed for this project. Both extrinsic and intrinsic parameters are recovered, including radial and tangential lens distortion parameters. This enables us to construct a 3D scene from an overhead camera and compute the transformation from camera coordinate system to a common world coordinate system. The system first detects the 2D silhouette of people using background subtraction on color and disparity [5] images along with "volume of interest" thresholding [1]. After that, we reconstruct a 3D silhouette from 2D silhouette using the disparity information. Using iterative segmentation of the 3D silhouette based on body topological structure constraints, each person in group is segmented. A further body posture analysis for each person, which does not require a explicit 3D human model, are applied to determine body orientation and body posture to understand whether people interacting object on the shelf. In addition, a hierarchical multi-camera tracking algorithm based on appearance and trajectory-continuity constraints are employed to recover the trajectory of each person in the store as shown in Figure 2.

There are four main contributions of our system: (a)dot-coded checker pattern target used in multi-camera calibration which does not require a full view of target, (b) using 3D silhouettes to detect and segment each people, (c) analysis of body posture and parts without using explicit 3D human body model, (d) combining shape and color appearance of people with trajectory-continuity constraints to solve disambiguate while tracking people during hand-off. A large number of people single camera real-time tracking systems have been developed over the last several years, most of them using color [15], background subtraction [5, 2], contour modeling, stereo [1, 9, 10] to detect and track people and understand activities [11, 4]. Our system takes advantage of both using color and stereo for better detection and tracking which allows us to use 2D and 3D information. Recently, the detecting and tracking of people and body parts of people from video has been explored in surveillance, HCI, and animation to understand high level people activities. The majority of the researches rely on explicit 3D human body models where potential dynamic constraints and appearance of parts are used as main features to detect and track body parts [12]. Most of the models rely on edges and contours of body parts to estimate the location of joint angles. However, the method we are describing in this paper takes a different approach and leverages the method and results of our previous work using 2D silhouettes [5]. Our approach to estimating body configuration is similar to shape context, the distribution of relative positions of a 2D shape for shape matching, introduced by Belongie and Later [14] used the shape context to estimate human body configuration in 2D.

The remainder of this paper is organized as follows. Section 2 explains multi-camera calibration, synchronization, and color calibration used in our pilot-system. Section 3 describes the 2D foreground silhouette detection method we employ and the construction of the 3D silhouettes. Section 4 explains the computational model used to segment each people from 3D silhouette. Section 5 describes the

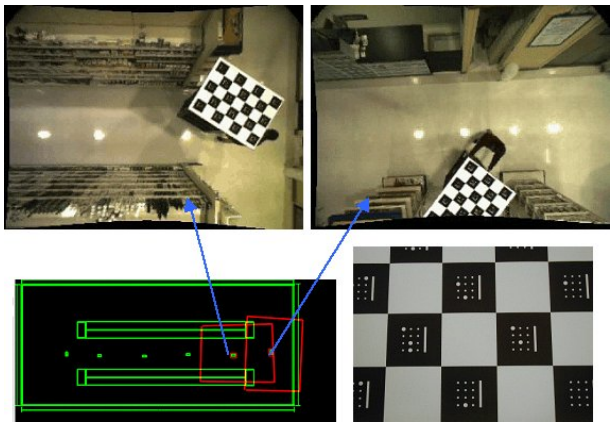

Figure 3: an instance of multi-camera camera calibration and dot-coded checker calibration target used in the system calibration procedure

multi camera tracking methods to recover full trajectories of people. Section 6 introduces the view-invariant shape model for body posture and determining the body posture and part to understand ”pick event”. Section contains the experimental results and discussion of future extensions.

2 Multi Camera Calibration

Our system consists of multiple narrow-baseline overhead digital stereo cameras [7]. The cameras connected to PC-based servers. Servers are synchronized using Network time protocol (NTP). The cameras are fully calibrated using an automatic calibration procedure developed for this project. Both extrinsic and intrinsic parameters are recovered, including radial and tangential lens distortion parameters. As our Video-CRM system is going to be installed in thousands stores, the multi camera calibration should not take considerable long time in such a broad installation. Therefore, we developed fast, easy of use, and automatic calibration procedure for this project. We developed a dot-coded checker pattern calibration target that allows us to multi-camera calibration even when the calibration target seen partially as shown in Figure 3. The each checker has special dot codes which indicated its relative location in the target. The calibration produce is simply moving the calibration target in the store, where the target is automatically detected by each camera system as soon as it enters the camera’s field of view partially. Each camera collects the measurements while the target in it’s field of view and those measurement are used to compute the camera transformation. This procedure enables us to construct a 3D scene from an overhead camera and compute the transformation from camera coordinate system to a common world coordinate system.

As the color mapping of each camera is not identical as shown in Figure 2, we need to calibrate the color setting in order to use appearance based methods across the cameras. We used a checker colored color calibration target and move the target across the camera to understand the color transformation from camera to camera using first camera as reference camera. In our system, experimental results shows that the color transformation is linear all color channel. Each camera normalize the color before processing and appearance information during multi-camera handoff.

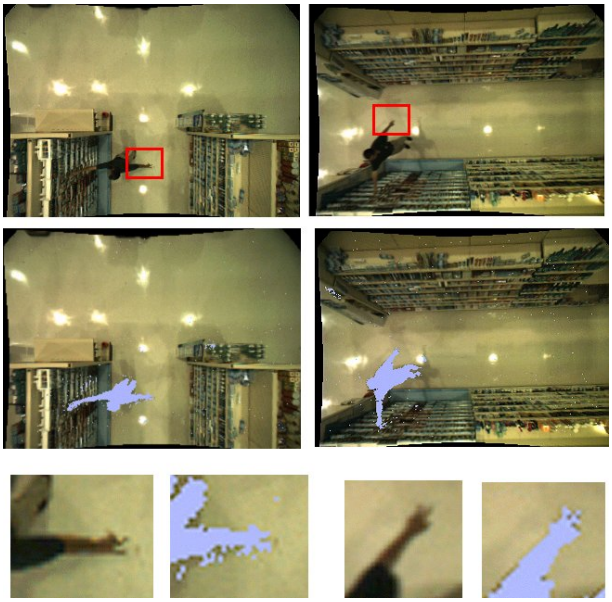

Figure 4: Example of 2D silhouette detection of a person in two neighbor camera: color images, detected silhouette in blue (middle row) and the detail (zoomed) of detected silhouette of area of person hand (bottom row). Note that, the shadows are extracted from silhouette and two fingers can be detected precisely from 13 feet distance.

3 Silhouette Detection

The silhouette of people is detected using a combination of color and disparity based background subtraction followed by a volume of interest filter. In our system, both color and disparity differences are computed separately, and combined together in a robust way that each method compensate the weakness of the other one. We used a robust and efficient color background subtraction algorithm that copes with shadows [6]. In Figure 4, examples of foreground detection and shadow elimination are shown where the detected silhouette is in blue color.Note that, the detection can achieve to distinguish two finger separation from 13 feet and good shadow separation. The background image is learned by computing the color chromaticity and brightness of each pixel over a period of time. During silhouette detection, the

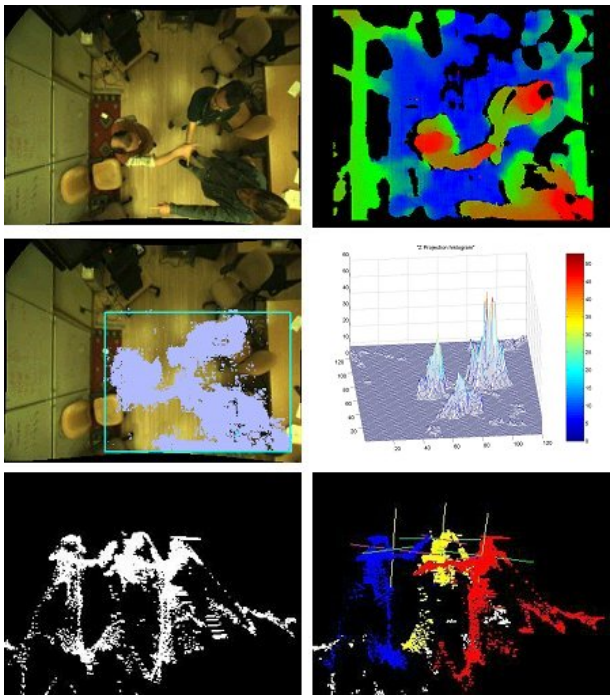

Figure 5: An example of Silhouette detection: (a) Color image, (b) disparity image (c) detected 2D silhouette, (d) occupancy map (e) constructed 3D silhouette (side view), (f) People segmentation results

chromaticity distortion and brightness distortion are computed for each pixel, and each pixel is classified as foreground, background, or shadow. The background model in range is modeled as a single Gaussian computed using over hundreds background disparity image. In order to reduce false positive errors caused by compression, camera jitter and digitization, we combine consecutive foreground detection results to obtain a 2D foreground silhouette.

We define a 3D silhouette a cloud of points in 3D, in which each point in 3D is constructed from a point in 2D silhouette using it's disparity value. Since the stereo camera has fully calibrated extrinsic and intrinsic parameters, we can easily defined a transformation which maps a 2D silhouette pixel $p^2 = [p^u, p^v, p^d]$ to 3D silhouette pixel $p^3 = [p^x, p^y, p^z]$. Here, u,v,d represents horizontal, vertical and disparity value respectively , and x, y, z represents its 3D location. 3D Silhouette is not a complete 3D construction of human body since an overhead camera has limited visibility of human body. However most of the upper body can be reconstructed and be identified easily in 3D. One of our main motivation is to be able use 3D silhouette for body posture and part analysis and person detection (**movie3**). An example of 2D silhouette detected by background subtraction and 3D silhouette generated by 2D detection and range information is shown in Figure 5.

4 People Detection

3D silhouettes allow us to locate people in the scene, however they do not locate an individual person within the silhouette boundary as a 3D silhouette may contain multiple people in its boundary. Person segmentation addresses the problem of how each person be segmented from given a 3D silhouette. In our system, the 3D silhouette are analyzed by scanning from near to far in z iteratively to detect any one of a head/shoulder region. In each iteration, a 3D silhouette of a person is segmented, remaining 3D silhouette goes under the same procedure to detect any other people until no more 3D silhouette points left. In each iteration, the head location of a person is detected first, since head regions are relatively simple to identify. Then, we compute a relative positions of each 3D silhouette point to the head location. We do this in a manner that exploits the topological constraints of body parts that enables us to identify the connection between body parts. Once a head/shoulder is detected, a normalized distance map based on 3D path distances between detected head/shoulder and any other 3D silhouette points are computed. This normalized distance map allows us to assign each 3D silhouette point to an individual people with an *owner likelihood*. All 3D points which have high owner-likelihood value, are segmented from 3D silhouette and assigned to one individual person. We applied same method to all remaining 3D points which has low or zero owner-probability recursively until there is no more head/shoulder regions is detected (**movie5**).

4.1 Head/shoulder detection

The system employs a global shape constraint derived from the requirement that the head be aligned with the axis of the torso. In particular, by projecting 3D silhouette points to the floor plane-projection histograms [5]-, the projection peak occurs near the head since the majority of points in the silhouette come from the head, torso, and shoulder regions. We vertically project silhouette points into a floor map representation H we call an *occupancy map* [1]. Consider a division of the floor plane into an $n \times m$ grid of vertical bins. We define a function $\gamma : \{x, y, z\} \rightarrow \{n, m\}$ which uniquely maps a 3D location to an index $\{n, m\}$ of H. In our current implementation, each bin size represents a $20 \times 20mm^2$ area in the floor plane. The occupancy map H is computed as:

$$H(n, m) = \sum_{p \in S^3} a(p)\, \delta[\gamma(p) - (n, m)] \qquad (1)$$

where δ is the Kronecker delta function, p is a 3D point in S^3, and $a(p)$ is a measure of the area covered by the point p. Figure 5(d) shows the occupancy map of 3D silhouettes,

where potential heads are retained only where there are significant peaks. The area measure $a(p)$ boosts the contribution of pixels that are further away relative to closer pixels. In our person tracking application, it helps to equalize the appearance of tall and short people in the occupancy map. To compute the area measure $a(p)$, we compute a surface approximation for each point in 3D. For a 2D silhouette pixel, $p^2 = [p^u, p^v, p^d]$, and its east and south neighboring pixels $p_e^2 = [p^{u+1}, p^v, p^d]$ and $p_s^2 = [p^u, p^{v+1}, p^d]$ and their corresponding 3D points $\Re(p^2), \Re(p_e^2)$ and $\Re(p_s^2)$ the area is computed as

$$a(p) = |\Re(p^2) - \Re(p_e^2)|_x |\Re(p^2) - \Re(p_s^2)|_y.$$

Here, the notation $|\cdot|_x$ or $|\cdot|_y$ means use the x or y coordinate in the computation.

Head and shoulder locations are estimated using a near-to-far scanning process in z. We observe that the total surface area of head-shoulder region in the 3D silhouette should be similar to the surface area of the head-shoulder region of a typical human body. Therefore, in an offline step, we estimate a total surface area A^{hs} of head-shoulder region of a typical human body. Then the system computes area $a(p)$ as discussed in the previous paragraph. As a 3D silhouette may contain multiple head/shoulder regions, we try to detect one head/shoulder region at a time. In order to find the group of 3D points in which represents a head-shoulder region, we analyze the 3D silhouette from top to bottom. While we are scanning the 3D-silhouette from top to bottom, we are grouping 3D points based on their location. We continue scanning until we scan enough 3D points where the total area of 3D points in any one group is larger than the expected area of a head-shoulder region A^{hs}. Depending on the body posture and configuration as well as the number of people in 3D silhouette, you can have more than one group of 3D points for each head/shoulder region. We pick only one group of 3D points in each scan that satisfies the total visible surface area constraints to determine the corresponding head/shoulder region. Once we find group of 3D points which satisfies the total visible surface area constraints, we use only those 3D points with their occupancy map values together to estimate a head/shoulder point which represents the head/shoulder region. In Figure 6(f) the weighted area of each 3D points are shown, the size of the each rectangle is directly related to the occupancy map value of that point: One observes that the rectangles around shoulder and head regions are bigger than the hands since the occupancy map tends to peak at head regions.

4.2 Segmenting rest of the body

Once a head/shoulder point is identified on 3D Silhouette, we analyze the connectivity and distance between head/shoulder point and other 3D points on 3D silhouette. The relative distances of each body parts to other parts is

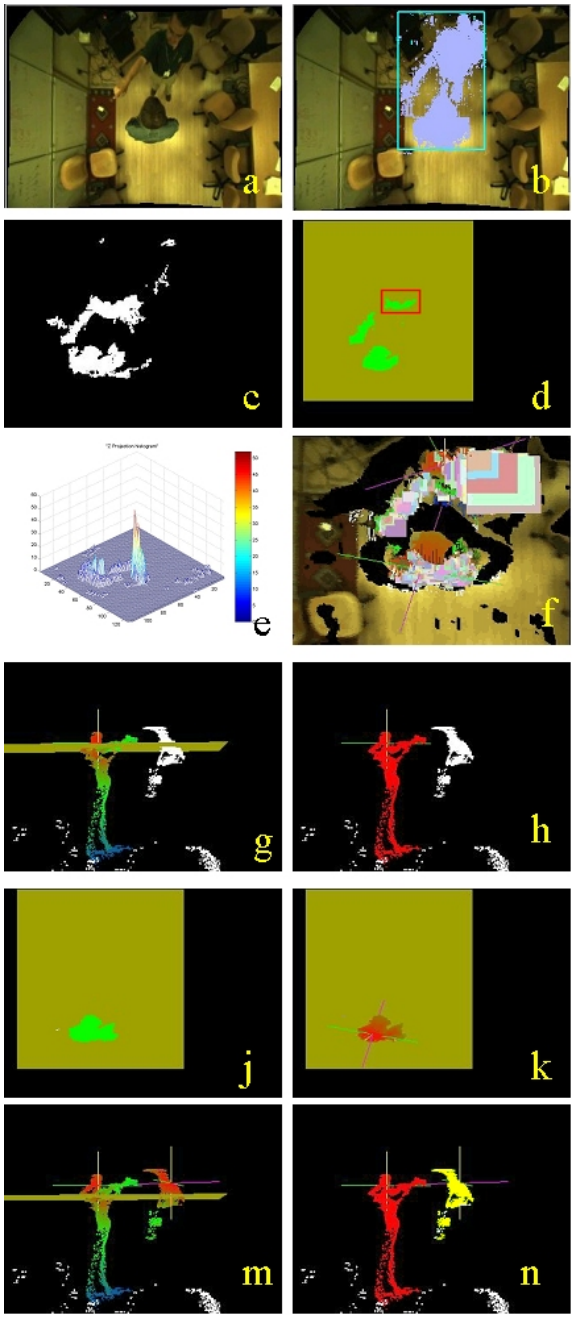

Figure 6: An example of person segmentation using top-down scanning: (a) color image, (b) detected 2D silhouette, (c) top-view of constructed 3D silhouette (d) scanned 3D points (in green) in first iteration of top-down scanning and detected head/shoulder region at this iteration (in red box) (e) occupancy maps (f) 3D rectangular surface for 3D points, (g) distance transform computation for detected head to other 3D point silhouette (h) segmented silhouette region (in red) in first iteration using distance maps (j) remaining scanned 3D points are shown in green in second iteration of top-down scanning(k) detected head/shoulder region on remaining 3D silhouette in second iteration (m) path distances computation for detected head in second iteration (n) segmented silhouette region (in yellow) in second iteration

an important feature to distinguish the body configuration. As human body is an articulated shape, the Euclidean distance may not be true to represent the topological structure of the body part. Rather, a *path* distance is more suitable for computing topological distance as it takes the connectivity properties of body parts into consideration. In our system, we employ a method which computes a distance transform from head location to each 3D point. Finding the neighbor points in 3D is a computational expensive method. Instead using 3D neighbor-connectivity, we describe a method using 2D neighbor-connectivity in 2D silhouette to find the neighbor pixels, then their distance is computed in 3D. So the total path distance from a detected head/point to another point is just a distance transform over 2D silhouette using 3D distances of neighbor pixels. Let $R(p^3) = [p_0^3, p_1^3, ..., p_k^3$ be a path (ordered point list in 2D) from a detected head point $p_0^3 = p_{head}^3$ to a point $p_k^3 = p^3$ in 3D silhouette. The distance d between any two consecutive point p_i^3 and p_{i+1}^3 in R is computed using Euclidean distance in 3D as

$$d(p_i^3, p_{i+1}^3) = |p_i^3, p_{i+1}^3| \tag{2}$$

where p_{i+1}^3 is the one of 8 neighbor pixel of p_i^3 in 2D silhouette and with a minimum 3D distance with p_i^3. The total path distance $D(p^3)$ is computed as

$$D(p^3) = \sum_{j=0}^{k} d(p_j^3.p_{j+1}^3) \tag{3}$$

If a neighbor pixel has more than 200mm in 3D, then we consider those points are not neighbor in 3D even two pixel are neighbor pixel in 2D.

Once a head/shoulder region detected and a path distances from head point to any other 3D points are computed, normalized distances $N(p^3)$ are computed to segment 3D points from silhouette using path distance $D(p^3)$ and the size (sz) of the head/shoulder region, where $N(p^3) = 2 * sz/D(p^3)$. Normalized distance $N(p^3)$ indicates a likelihood (owner-probability) of a 3D point belong a particular person. Therefore, we segment the 3D points which has high likelihood $N(p^3) > N^{thr}$ (N^{thr} is 0.8m in current implementation) and remaining silhouette with $N(p^3) < N^{thr}$ are analyzed with the same method to detect any other head and shoulder region to segment the remaining people. This recursive method continues until no more head/shoulder detected. Remaining 3D points which has very low owner-probability is considered as false-detected pixels in 2D points. Figure 6 illustrated an example of people segmentation for two people.

5 People Tracking

Tracking a single person in entire stores requires multiple camera as single camera's field of view is not enough to cover the entire surveillance area in entire stores. We proposed a hierarchical approach for people tracking using multiple camera to recover each person's trajectory in entire stores. This approach combines each camera's local tracking results in hierarchical way to obtain the full trajectories in system's common coordinate system. Each individual camera system tracks people while they are in camera's field of view using an appearance based tracking methods. Each camera system reports their local tracks to it's **track manager**. Track managers are responsible to convert local tracks to common coordinate system, combine them, and resolve disambiguates in overlap areas. The overhead cameras in our pilot-system are located and oriented as the overlapped area between each cameras field of view is minimal enough that allow us to handle hand-off tracking from. In Figure 7, an example of overlapped areas between three cameras are shown.

Each camera system employs an appearance based tracking algorithm based on mean-shift tracking to track people as long as they are in camera's field of view. This tracking allows us to recover *local* (camera-centric) trajectories of each people. The people appearance (color of clothing and shape) is an important cue for visual tracking. We developed a real-time tracker based on the visual appearance of people. The goals of person tracking stage are to initialize an appearance model based on color and edge density when a person appears on the camera's field of view, and compute the correspondence between person detected by the person segmentation and the people currently being tracked and recover the trajectories of each person in the cameras field of view.

An appearance model is constructed for a person as soon as they are detected in the person segmentation stage. The appearance model consists of the red, green, blue color and edge (gradient magnitude) densities of a person silhouette. For computational simplicity, we approximate the appearance distribution of a person using an n-bin histogram, θ_i. Once the appearance model for person is computed, the location of the person in the next frame is estimated by computing the similarity between its appearance θ_i and the appearance of each candidate location α_y in the next frame. The two distributions are most similar when their correlation is maximum. To find the location, in the next frame requires an exhaustive search in the neighborhood. Instead, we used a mean shift approach to achieve real-time performance. The details about the appearance based tracking used in our system can be found in [5].

Track managers handle tracking people in overlap areas (hand-offs) and resolve ambiguities using both appearance based methods (color and shape of people) and trajectory-continuity constraints as shown in Figure 7. Track managers also determine when a new person enters in the scene, gives an unique person ID and track them using their unique IDs.

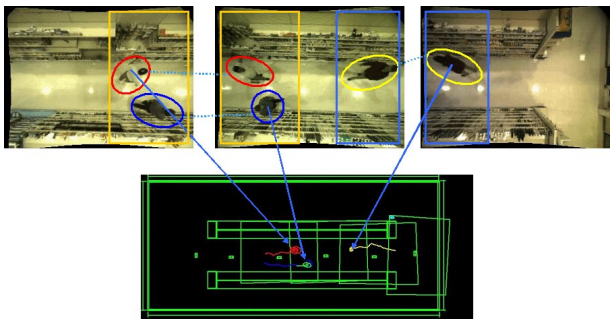

Figure 7: People in overlapped areas are matched using appearance information-color and shape in tracking, and their trajectory-continuity

Track Manager uses time-stamped local tracks to combine multiple local tracks to a single common coordinate trajectory. If a person in an overlap area, two or more local tracks for the same person are reported by camera systems. Track Manager combines these multiple tracks, which belongs two same person, in to one common coordinate track by using two constraints based on observations of how people move: (a) **Trajectory-Continuity:** when combined, their trajectories should show a continuation trends in common coordinate system, (b) **Appearance-Continuity:** The color and shape of the person should be similar in all the local camera system.

6 Analyzing "Pick Event"

We employed a further body posture analysis to understand the interaction between customers and merchandise in the store. For example, a customer looks at an item on a shelf, picks up the item, looks for a price or information, then places it back on the shelf or in a shopping cart. In order to extract this information, we need to recover their body orientation, posture and head orientation. We prefer an approach to detecting body posture without using an explicit 3D human model. We define a shape descriptor that expresses the configuration of an entire 3D silhouette of a person g as a distribution of path distances ($D(g)$) and relative orientations ($\alpha(g)$ and $\beta(g)$) between a point on 3D silhouette and the head point where the relative orientation is defined as the angle between a vector originating from head point to a point and the body orientation vector. We define the body orientation vector c as the minor axis of the 3D head-shoulder region that can be computed by applying a principal component analysis (PCA) to the head-shoulder pixels x and y locations in 3D. After we compute the body orientation vector c, we compute the angle between c and a vector v originated from head point to a point on 3D Silhouette. We compute two angles α and β as relative orientation descriptor. α is the angle on x axis between c_x and a_x, and β is the angle on z axis between c_z and a_z. We believe that the distribution of relative distance and orientation over 3D silhouettes is a robust, compact discriminative description that is invariant to orientation and translation and not effected by partial occlusions. We compute a *shape histogram* $W(g)$ of $D(g)$, $\alpha(g)$ and $\beta(g)$ to identify the shape of 3D silhouette. The shape histogram has large variations in four common upper body postures seen in a store: customer using no arm, using left arm, using right arm, and using both arms. Other postures are a small variation of one of the main body postures. The relative distances and orientations do not change significantly while a person is in one of the main postures. However, the relative distances and orientations do change when they change their main posture. Our system classifies the observed human upper body posture in a "hierarchical" manner: any posture is classified as one of the main postures in the first stage using shape histograms of D and α, and the local variation and arm orientation variation are computed in the second stage using shape histogram on β. We use an exemplar based shape similarity approach to estimate the body configuration. We experimentally generated an average of normalized shape histogram for each of the four main upper body postures. Each normalized shape histogram of a 3D silhouette is compared with the shape histogram of those main postures. Let $W(g)$ be normalized shape histogram of a 3D silhouette g and $W(M_i)$ be the normalized shape histogram of the ith main posture. The similarity $t(g, M_i)$ of those shape histogram is computed as $t(G, M_i) = \sum_{u=1}^{n} W_u(g)W_u(M_I)$ where n is the number of bins in the shape histograms. The two shape histogram are most similar when $t(g, M_i)$ is maximum. From the highest scoring main posture, $\beta(g)$ is used to estimate the relative orientation of arms and hand in the Z axis. Using both main and secondary posture estimation allows us to estimate rough location of body parts where we can detect "pick" events using topological property of body parts once the posture estimated (**movie7**).

Once the pick event has been detected, we try to determine on which shelf level and which item the customer is interacting. As we have already reconstructed 3D model of aisles and shelves and we estimated the arm/hand location in common world coordinate system using 3D silhouette, we can determine which shelf the customer is interacting. Currently, we do not attempt to recognize the merchandise that customer pick, instead, we used merchandise/layout/ information (which provide "where the merchandise is stacked in the shelf" information). In Figure 8 showing some instances of 3D reconstructed and silhouettes that the customer is picking object from different shelf levels.

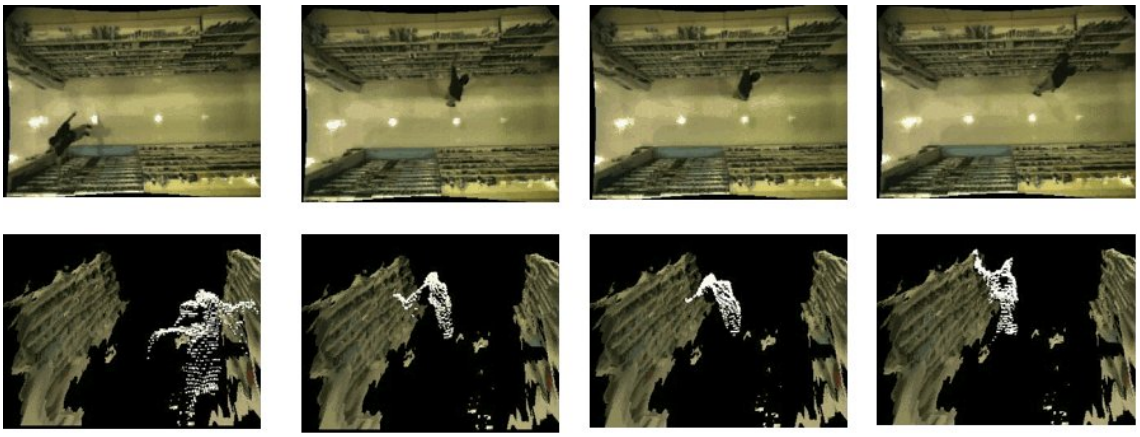

Figure 8: Examples of ”Pick Event Detection”: the 3D silhouettes and scene are shown in bottom row

7 Discussion

We preferred stereo camera solution over monocular color camera even though stereo is aq computational expensive approach for better detection and posture analysis as we can use 3D information to overcome some of the occlusion problems. We preferred an approach to detect people without using explicit 3D model of human, instead we use shape and appearance constraints on topology of human body.

We evaluated the performance of our system as qualitatively using two 15 minutes sequence captured at 10 fps. There are 17 people (5 of them moving as group) while shopping in one aisle that we were monitoring. The system correctly detected and labeled every person in the scene except one person detected as two people for first 1 minutes due to segmentation error and it is recovered later. During shopping, people were in mostly one camera's field of view and at most two camera's field of view at a time, however, they moved trough the aisle so each camera system can see them for some time interval. There are 29 hand-off during tracking and the systems handle hand-off 27 times correctly using both trajectory-discontinuity and appearance information where trajectory-discontinuity alone handles 18 hand-off correctly, appearance based hand-off alone handle hand-off 21 cases correctly. There are 83 ”pick-event” that the customer taking an object from shelves. The system correctly detected the 76 pick-event with 91% accuracy. For correctly detected pick event, the system correctly classify 69 pick-event with correct shelf determination. Currently in pilot system, there are 6 stereo head, two stereo head connected to a dual with processor 1200 MHz Pentium III computer runs at 5-7 fps using 320x240 resolution video, this includes stereo computation, background subtraction, 3D silhouette generation, body segmentation. The overall system has been calibrated once it is still good for 6 months.

Current system uses 3D silhouette segmentation by single camera. We are working on extending capabilities of 3D silhouette segmentation by combining 3D silhouettes generated by multiple camera and apply the 3D silhouette segmentation method onto these merged 3D silhouette. This yields better segmentation as some of the missing body part which can not be seen by one camera, can be seen by other camera.

References

[1] anonymous

[2] Boult, T.E., et al. Frame -Rate Omnidirectional Surveillance and Tracking of Camoflaged and Occluded Targets. In *IEEE International Workshop on Visual Surveillance.* 1999.

[3] Darrell, T., et. al. Integrated person tracking using stereo, color, and pattern detection. In *CVPR*, 601-608, 1998

[4] Grimson, W.E.L., et al. Using Adaptive Tracking to Classify and Monitor Activities in a Site. In *CVPR*, 1998

[5] anonymous

[6] T. Horprasert, D. Harwood, and L.S. Davis. A Robust Background Subtraction and Shadow Detection. In *Proc. Asian Conference on Computer Vision* , January 2000.

[7] Videre Design, MEGA-D Stereo Camera, *www.videredesign.com*

[8] Konolige, K. Small Vision Systems: Hardware and Implementation. Eighth International Symposium on Robotics Research, Japan, October 1997

[9] Krumm, J., et al. Multi-Camera Multi-Person Tracking for EasyLiving. In *IEEE International Workshop on Visual Surveillance*. 2000.

[10] Rehg, J.M., et.al. Vision for a Smart Kiosk. In *CVPR*, 690-696, 1997

[11] Rosales, R. et.al. 3D Trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions. In *CVPR*, 1999

[12] H. Sidenbladh, M. Black, D. Fleet Stochastic tracking of 3D human figures using 2d image motion In Proc. ECCV, 702-718, June, 2000

[13] Vivek Katra, Aaron Bobick, Amos Y. Johnson Temporal integration of multiple silhouette-based body-part hypotheses In Proc, CVPR, 2001

[14] G. Mori and J. Malik Estimating Human Body Configuration using Shape Context Matching In Proc. ECCV, 2002

[15] Wren, C., et al., Pfinder: Real-Time Tracking of the Human Body. In, PAMI, 1997. 19(7): p. 780-785.