2004 IEEE Intelligent Vehicles Symposium
University of Parma
Parma, Italy • June 14-17, 2004

Human Posture Estimation Using Voxel Data for "Smart" Airbag Systems: Issues and Framework

Shinko Y. Cheng
sycheng@ucsd.edu

Mohan M. Trivedi
mtrivedi@ucsd.edu

Computer Vision and Robotics Research
University of California, San Diego
La Jolla CA 92032-0434
http://cvrr.ucsd.edu

Abstract

This paper examines the feasibility of a multi-camera voxel based occupant posture estimation system. Several new considerations are made to allow this tested human body modeling system to work reliably in the passenger seat of a vehicle, including camera position, segmentation, and body modeling with voxel reconstructions all from a constrained 4 camera setup. To describe occupant posture, a partial human body model consisting of a head and torso is proposed. The accuracy of the estimation of this model is compared against ground-truth.

1 Introduction

Estimating the posture of occupants in an automobile have wide ranging uses. One such use is in determining whether the occupant is of the right type and in the right position for airbag deployment. Several injuries and deaths have been attributed to the ill-timed deployment of the airbag. In this paper, we propose a shape-from-silhouette (SFS) based system that monitors the posture of the occupant [1, 2] in hopes to take a step towards prevention of such tragedies.

We present an occupant posture estimation system and explain the issues involved in using SFS voxel reconstructions of occupants for the purpose of estimating occupant pose. Several problems arise when taking the shape-from-silhouette technique from inside a controlled-lighting lab environment into the passenger seat of a car. Each stage of the system as it were requires additional considerations. These considerations are in the camera placement and calibration, image segmentation, voxel generation, and body modeling from voxel data.

SFS operates on silhouettes of the subject and calculates a voxel reconstruction of the subject's volumetric form. Silhouettes are generated using a statistical background subtraction and shadow detection technique [3]. Because only the upper body is visible in the seat, we consider a partial human body model consisting of ellipsoids and cylinders to represent the head and torso. Only voxel data is used in estimating the locations of the body parts. A spherical shell template that represents the largest and smallest expected head size is used as part of a template matching algorithm to locate the head. Because of the rough voxel reconstruction calculated from only 4 viewpoints, a Kalman filter is used to smooth and predict the tracks produced by the head estimates.

Several implementations of human posture have relied on image information in the form of skin color [4], appearance of a face or hand, contours of the subject [5]. Our approach relies solely on volumetric voxel data of the human subject to infer occupant posture information. The advantage of SFS is its completeness of coverage. Stereo vision provides depth measurements but provides only sparse measurements within the image, even for the "dense" versions of stereo [6]. In SFS, the entire image is utilized to form the volumetric representation of the subject, and is a reasonable upper bound of the shape of the human form. Nevertheless, this approach has its challenges when set inside a car. These issues largely exist in the need for calibrated cameras and a silhouette generating algorithm robust to widely varying lighting conditions.

2 System Considerations

In the following sections, new considerations in the occupant posture monitoring system in camera placement, calibration, segmentation, voxel reconstruction, and body modeling as the result of implementing the system inside a vehicle are described. The system flow diagram is shown in figure 1

0-7803-8310-9/04/$20.00 © 2004 IEEE

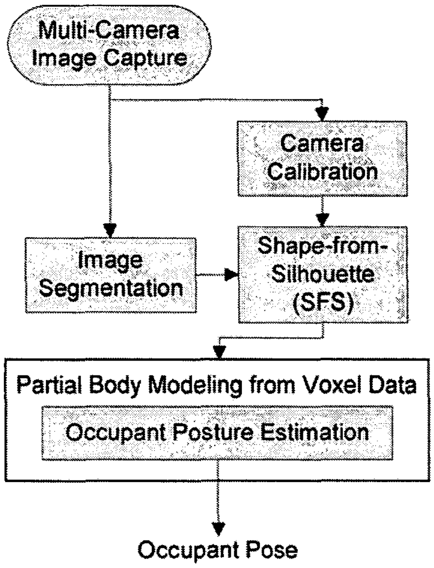
Multi-Camera Image Capture
Camera Calibration
Image Segmentation
Shape-from-Silhouette (SFS)
Partial Body Modeling from Voxel Data
Occupant Posture Estimation
Occupant Pose

Figure 1: System flow diagram.

2.1 Camera Placement and Calibration

The choices of camera placement is limited due to the confines of the front seat. Good locations of the camera are orthogonal with each other and as far away from the object as possible. Cameras placed orthogonal with each can carve away more of the volume not imaged by the other cameras. Placing cameras farther away leads to images that are nearly orthographic and the conic volumes projected from the camera are narrower, allowing the far side of the cone to be smaller and be carved more easily with the other cameras.

For the accurate capture of the visual hull of the object [7], increasing the number of viewpoints will result in better voxel reconstructions. Four has been used here and by the results appear to be adequate for the purpose of extracting the head and torso locations and orientations. Suitable places for these cameras are on the middle of the front dash, the passenger side-view mirror, the driver rear-view mirror, and the top of the passenger window.

The coordinates of a point $\bar{\mathbf{P}} = (X, Y, Z, 1)^{\mathrm{T}}$ in occupant space is related to the image coordinates $\bar{\mathbf{p}} = (u, v, 1)^{\mathrm{T}}$ and $\hat{\mathbf{p}} = (\hat{u}, \hat{v}, 1)^{\mathrm{T}}$ by the equation

$$\bar{\mathbf{p}} = \mathbf{KD}\hat{\mathbf{p}} = \mathbf{KD}\frac{1}{cZ}(\mathbf{R}\,\mathbf{t})\bar{\mathbf{P}}$$

where K is contains the set of intrinsic camera parameters and the extrinsic camera parameters are given by the rotation matrix $\mathbf{R} \in SO(3)$ and a translation vector $\mathbf{t}$. Radial distortion is modeled as the 3×3 matrix factor

$$D = diag(\lambda, \lambda, 1)$$

where

$$\lambda = 1 + \kappa_1 d^2 + \kappa_2 d^4$$
$$d = \hat{u}^2 + \hat{v}^2.$$

The Matlab Calibration Toolbox and Intel OpenCV Library was used to estimate this full set of camera extrinsic, intrinsic, radial distortion correction parameters necessary to perform voxel triangulation and voxel to pixel table calculation.

2.2 Segmentation

Silhouette generation by image segmentation is a critical component in shape-from-silhouette techniques. Lighting in a car is variable and highly volatile, posing a difficult problem for several color-based segmentation algorithms.

To partially address the issues of volatile changes in color, the system uses a statistical background subtraction technique that assumes that the background pixel values are gaussian distributed and only varies along the "chromaticity line" when a shadow is cast over or lifted from it. The "chromaticity line" is the line that connects the origin in RGB space and the mean value of the pixel when the pixel images the background. After collecting the background statistics, pixels in subsequent images, based on thresholds chosen for the amount of brightness and the chrominance distortion, are classified as background, shadowed background, highlighted background or foreground [3].

As is the case for all background subtraction techniques, the assumption is that the background does not change drastically, such as camera movement. And if the scene does change, it is due to the lighting and not object movement. Lighting changes can be addressed by allowing room for brightness distortion in subsequent images, and this occurs when the sun moves about the car when the car is in motion. Changes in the scenery outside the window however necessitate updating the background. The difficulty is in updating the background model without mistakenly incorporating foreground information into the background model, and in updating the background model quickly enough for the dynamic nature of scenery outside the car window.

2.3 Shape-From-Silhouette

The constituent volume element is the voxel. Several real-time voxelization algorithms are proposed [1, 2]. Generally, the procedure consists of considering each voxel and projecting its location onto each of image planes with the

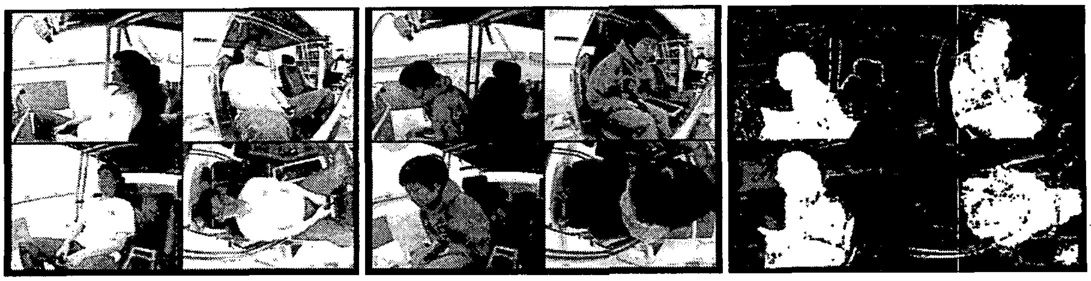

Figure 2: Input quad video sequence and segmented result using statistical background subtraction and linear color trajectory assumption for the behavior of shadows.

camera calibration information. The voxel is marked occupied if all pixels that the voxel projects onto are contained in the silhouette.

This system implements the optimizations proposed by [2] by limiting the number of voxels to check, considering only voxels indicated as occupied by the conic projection from the "dominant" camera. The occupancy of this subset of voxels is checked with the procedure of projecting the voxel onto the image as mentioned above. Another speedup is obtained by stopping the rest of the camera checks when a single camera indicates a voxel is unoccupied. A voxel-to-pixel and pixel-to-voxel table is also generated for each camera for a given set of extrinsic camera parameters to reduce the calculations during run-time to a simple table-lookup.

2.4 Body Modeling

In the indoor environment, it has been shown that a complete articulation of the human head, torso, upper arms, lower arms, upper legs, and lower legs from human voxel data is possible using twists formulation of these body parts and proceeding with a multi-stage fitting algorithm that estimates body part sizes and tracks body part orientation. For the car environment, the parts of the human body that can be found most consistently are the head and torso. The legs will most likely not be in view, unless the legs come in view when the occupant raises them onto the dash or other location. The occupant head and torso estimation flow diagram is shown in figure 3.

Estimating location and orientation of the various body parts begin with template matching a spherical shell kernel against the surface voxels of the voxel reconstruction of the occupant [9]. The spherical shell kernel is constructed using two radii. The larger and smaller radii represent the largest and smallest head one expects to find. The two radii were found to be 15cm and 5cm. The highest responses are frequently the head and less frequently the shoulders and hands. The tallest peak in the up direction is initially designated as the head.

A Kalman Filter tracks the location of the head. The state vector of the Kalman filter consists of six elements that describe its location and velocity. The state transition matrix assumes a damped velocity with half-life of 500ms. A valid head measurement is one that is the highest peak and within several centimeters away from the predicted location of the head. If no valid measurement existed, the predicted location of the head is used. The recovery mechanism for when the Kalman filter loses track of the feature is done by reinitializing the track when it is not updated with measurements for several consecutive frames.

The head location from detection step above is followed by a torso growing procedure. The neck is found by estimating the centroid of voxels contained in the spherical shell with radius 25% larger than the larger radius centered at the estimated head location. The neck serves as an anchor point for the torso fitting. An iterative fitting scheme is used to orient the predefined torso shape over the voxel data. A fixed cylinder with radius and length of 20cm and 30cm respectively is the torso model used. The iteration begins with an initial guess for torso orientation with the torso major axis pointing away from the head. The centroid of the voxels contained within the torso cylinder at this initial position is found. Then, the new torso cylinder is centered over the centroid. The algorithm iterates between these two steps until convergence. The algorithm has shown to always converge because there always exist neighboring voxels in the neck area provided the head was found.

The current configuration of cameras where placement is generally top-looking-down produces a voxel reconstruction where the bottom half of the occupant flares out into infinity. This flaring poses a challenge when determining the legs, and the arms when placed near the legs. However, as will be shown in the experimental section, accurate voxel reconstruction is helpful but not necessary for the detection of the head and torso locations.

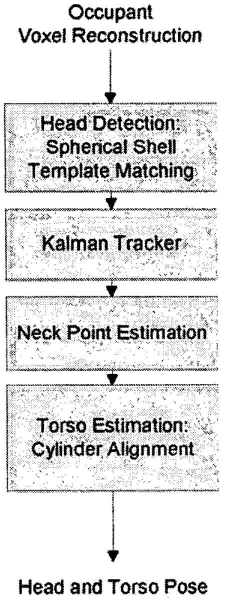

Figure 3: Flow Diagram of occupant posture estimation from voxel data.

An issue with body modeling from voxel data is the melding of limbs into the body. From voxel data alone, at times there is not enough information for even the human observer to decide a particular portion of a volume to be a limb or torso. These conditions usually occur when the arms are close to the torso and legs. In the car environment, this condition occurs more often than not. It thus appears necessary to use other cues to extract portions of the volume from the larger blob, such as appearance. To partially address the issue, the arms are detected when they are in view, and then tracked back into the volume blob. It can be shown to work well in momentary situations.

3 Experimental Results

Four NTSC images of the occupant are captured through a quad video combiner device. This device synchronizes the video to within 33ms of each other. After the background statistics are collected from a set of images of the seat unoccupied, the occupant is asked to perform a few poses summarized in table 1. Figure 2 shows the raw and segmented video of the occupant in the normal position. Figure 4 shows the result of the occupant pose estimation from voxel data.

As a ground-truth, the position of the head is tracked with the Polhemus FasTrak electromagnetic motion tracking device. Two receivers are placed on either side of the head of the occupant and their positions tracked. The average of the two measured receiver locations is taken to be the ground-truth location of the head at that instant in time. The difference between estimated head location and ground-truth are shown in figure 5. Three peaks in the normed difference plot that exceed 50cm difference indicate instances when the head was incorrectly detected to be elsewhere in the scene.

Table 1: List of test poses of the occupant. 50 frames are captured for each set of poses.

	Poses
1	Seated back. Move head back and forth.
2	Move head and body forwards and backwards.
3	Seated forward. Move head back and forth.
4	Seated back. Move head left and right.
5	Seated back. Move head counterclockwise.
6	Seated back. Move head clockwise.

The average error difference is 10.7cm. Omitting frames when the head is incorrectly detected, the average difference decreases to 7.55cm. This latter difference can be attributed to the approximate placement of the sensors on the subject's head during test to measure the centroid of the head. The alignment between the two tracks does not take into account head tilt. This can be observed in the y-axis plot of figure 5. The subject's head at 60cm is tilted back, at 40cm is upright, and 20cm is tilted forward, producing this deviation from ground-truth.

In 300 frames, 19 frames or 6.33% resulted in a head detection error. The primary cause is the a voxel reconstruction artifact in a region that appeared more spherical than the head. Figure 7 illustrates one such error.

The shape-from-silhouette algorithm runs at 7Hz with a voxel resolution of 60 × 50 × 50 using the C implementation on a 2.17GHz AMD Athlon PC. The body modeling runs offline in Matlab.

4 Conclusion

This paper proposes an occupant posture estimation system based on shape-from-silhouette voxel data. Given the voxel reconstruction of the occupant, using this voxel data alone is shown to provide fairly reliable location estimates of the head and torso. Experiments demonstrate an estimation accuracy of 7.55cm from ground-truth for head positions.

Special considerations arise when placing this tested system into the interior of a car with regards to camera placement, camera calibration, voxel reconstruction and body modeling, and most notably image segmentation for silhouette generation. The last issue remains only partially addressed.

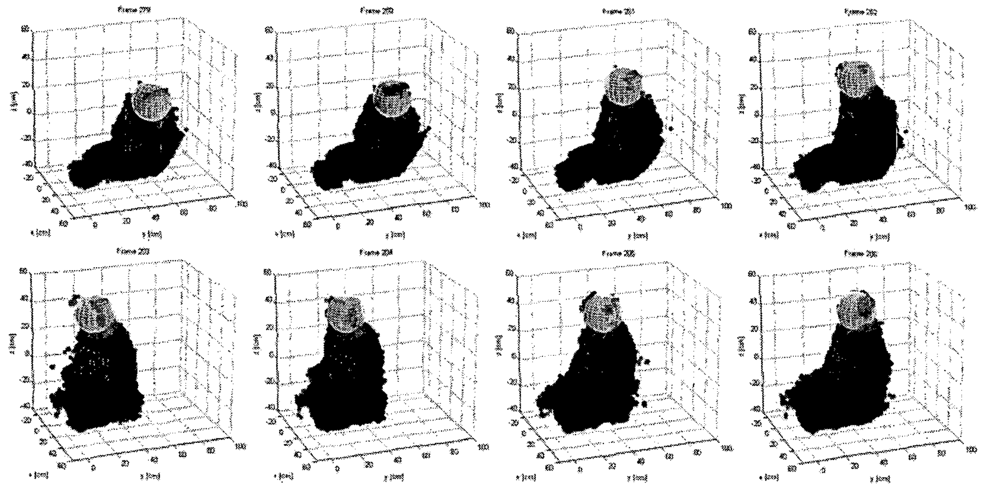

Figure 4: Result of spherical shell template matching and torso fitting. The top and bottom rows illustrate the tracking of the head and torso as the head moves left to right and front to back.

For more detailed posture information, cues such as appearance of arms and legs, and depth measurements within the silhouettes or a more elaborate model of the human subject may be required. Despite this, partial body modeling using voxel reconstructions show promising results.

In future work, we intend to incorporate knowledge of the boundaries of occupant in-position, out-of-position, and critically-out-of-position regions from the new airbag firing requirements set forth by the NHTSA in the FMVSS 208 towards the design of a classification system. Figure 6 illustrates the locations of these boundaries set in context with the occupant voxel reconstruction. Voxels that are contained within each region are shaded. The boundary planes are placed vertically 0mm, 100mm, 300mm and 700mm away and extends 100mm down below the level of the airbag.

Acknowledgements

This research is sponsored by a UC Discovery Grant in partnership with Volkswagen Research Laboratory. The authors gratefully acknowledge valuable inputs and assistance by Dr. Klaus Schaff and Dr. Arne Stoschek of VW Research Laboratory. We are also thankful for the assistance and support of our colleagues from the UCSD Computer Vision and Research Laboratory, especially those of Dr. Ivana Mikic, Kohsia Huang, Remy Chang, Jun-Wen Wu, and Pew Putthividhya.

References

[1] G.K.M. Cheung, T. Kanade, A Real-Time System for Robust 3D Voxel Reconstruction of Human Motions, In *IEEE Proceedings CVPR* , pp714-720, 2000.

[2] D.E. Small, L.R. Williams, Real-Time Shape-from-Silhouette, *Master's Thesis,* University of Maryland, 2001.

[3] T.Horprasert, D. Harwood, and L.S. Davis, A Statistical Approach for Real-Time Robust Background Subtraction and Shadow Detection, *Proc. IEEE ICCV'99 FRAME-RATE Workshop*, Kerkyra, Greece, September 1999

[4] R. Hoshino, D. Arita, S. Yonemoto, R. Taniguchi, Real-Time Human Motion Analysis Based on Analysis of Silhouette Contour and Color Blob, *Springer AMDO*, 2002.

[5] D.M. Gavrila, The Visual Analysis of Human Movement: A Survey, *CVIU*, Vol. 73, pp.82-98, 1999.

[6] M. Devy, A. Giralt, A. Marin-Hernandez, Detection and Classification of Passenger Seat Occupancy using Stereovision, *IEEE Proceedings Intelligent Vehicles Symposium*, 2000.

[7] A. Laurentini, How Many 2D Silhouettes Does It Take to Reconstruct a 3D Object?, *CVIU*, Vol. 67, No. 1, pp.81-89, July 1997.

[8] G.K.M. Cheung, S. Baker, T. Kanade, Shape-From-Silhouette of Articulated Objects and its Use for Human Body Kinematics Estimation and Motion Capture, In *Proceedings CVPR'03.*

[9] I. Mikic, M.M. Trivedi, E. Hunter, P. Cosman, Human Body Model Acquisitiona and Tracking Using Voxel Data, *IJCV*, 53(3), pp. 199-223, 2003.

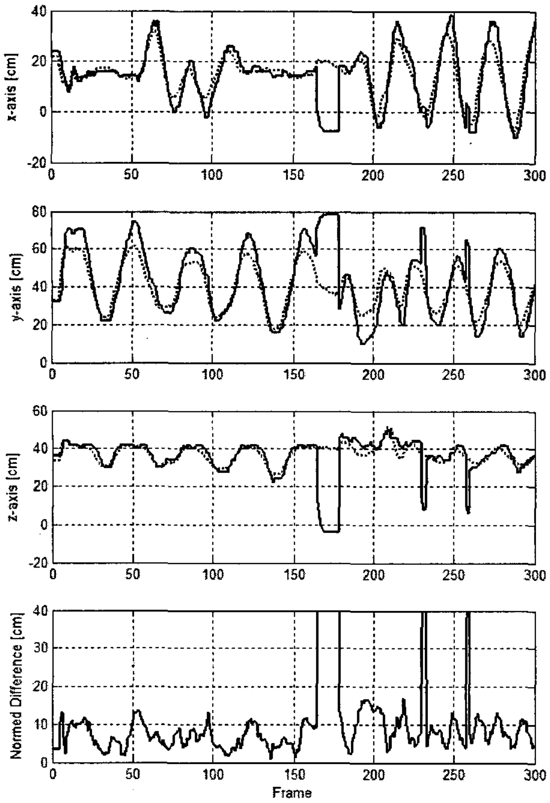

Figure 5: Head tracking results shown against ground-truth data. Head tracking from the occupant posture estimation is represented with a solid line and the ground-truth with a dotted line.

[10] I. Mikic, M.M. Trivedi, Vehicle Occupant Posture Analysis Using Voxel Data, *Ninth World Congress on Intelligent Transport Systems*, Oct. 2002.

[11] M.M.·Trivedi, S.Y. Cheng, E.C. Childers, S.Krotosky, Occupant Posture Analysis with Stereo and Thermal Infrared Video: Algorithms and Experimental Evaluation, to appear IEEE Trans. on Vehicular Technology, 2004.

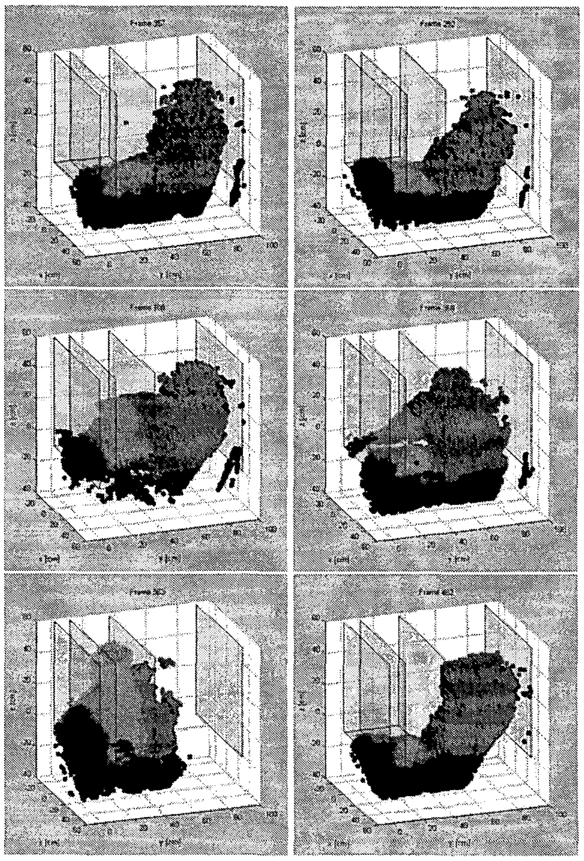

Figure 6: From the left most region demarcated by the boundary planes, they are the critically-out-of-position, out-of-position, and in-position, regions. In row-major order, the poses of the occupant shown are seated normally, knees on dash, feet on dash, reaching forward, on edge of the seat, and hands over head.

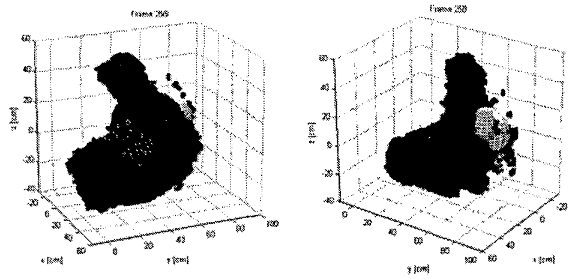

Figure 7: Error in head tracking from voxel reconstruction artifacts. Frame 258 shown.