Occipant Posture Analysis with Stereo and Thermal Infrared Video: Algorithms and Experimental Evaluation*

Mohan M. Trivedi, Shinko Y. Cheng, Edwin C. Childers, and Stephen Krotosky
Computer Vision and Robotics Research Laboratory (cvrr.ucsd.edu)
University of California at San Diego
La Jolla, CA 92093

Abstract

Dynamic analysis of vehicle occupant posture is a key requirement in designing "smart airbag" systems. Vision based technology could enable the use of precise information about the occupant's size, posture, and in particular, position in making airbag deployment decisions. Novel sensory systems and algorithms need to be developed for capture, analysis and classification of dynamic video based information for new generation of safe airbags. This paper presents a systematic investigation where stereo and thermal infrared video based real-time vision systems are developed and systematically evaluated. It also includes design of several test beds, including instrumented vehicles for systematic experimental studies for evaluation of independent and comparative evaluation in automobiles. Results of extensive experimental trials suggest basic feasibility of video based occupant position and posture analysis system.

I. Introduction

According to the National Highway Traffic Safety Administration (NHTSA) [14], in the past 10 years, airbags were deployed more than 3.3 million times in the United States. Airbags are credited with saving more than 6,000 lives and preventing a much greater number of serious injuries. These numbers clearly highlight the life saving attributes of airbag technology. Alas, there are other rather disheartening numbers that are presented in the same report. It states that since 1990, over 200 fatalities have been recorded as a direct result of an airbag deployment. The majority of these deaths have been children. The number of severe injuries is much higher. Obviously, these deaths and injuries must be prevented by exploring and adopting the most appropriate technologies, policies and practices. A new law, which will go into effect in the United States in 2004 requires that airbag systems be able to distinguish small persons and persons in unsafe positions from persons in safe positions for airbag deployment [1]. One of the main difficulties encountered by the decision logic systems used in airbag deployment deals with the critical assumption about the occupant size and position in the car at the time of a crash. Most airbag systems consider a single standard for the occupant's size and the nature of the crash. Vision based technology enables the use of precise information about the occupant's size, position, and posture.

The primary objective of this research is to describe the design, implementation and evaluation of vision based occupant posture analysis systems to control the deployment of an airbag in a safe manner. These efforts resulted in the development of, (1) a framework for sensing the most relevant visual information, (2) a set of robust and efficient algorithms for extracting features which characterize the size, position and posture of the occupant, and (3) a pattern recognition module to classify the visual cues into categories, which can trigger the safe deployment logic of the airbag system. We are

* This research is sponsored by a UC Discovery Grant in partnership with Volkswagen Research Laboratory under the Digital Media Innovation Program
devising vision-based technology that will allow our airbags to recognize the difference between “deploy to save” and “deploy to hurt”.

In this paper, we describe our experiments estimating the body position and pose of a vehicle occupant using thermal IR, stereo depth, and 3D voxel reconstruction data. We will show that these systems are capable of extracting the position of the head in the case of thermal infrared and stereo depth data, and the posture of the body, with 3D voxel data, with good accuracy.

A novel test bed was constructed to perform the comparison between these vision-based “smart airbag” systems. The test bed is built around a Volkswagen Passat outfitted with the necessary equipment to perform real-time side-by-side tests. This framework not only subjects the systems to realistic conditions, but also allows us to quickly accommodate new systems for comparison under identical conditions.

II. Research Objectives and Approach

The objective of the proposed research is the development of a highly reliable and real-time vision system for sensing passenger occupancy and body posture in vehicles, ensuring safe airbag deployment and helping prevent injuries. We propose using video cameras, which are very unobtrusive and provide information that in addition to allowing for “smart airbags”, could be used for other purposes.

We propose investigating multiple approaches in three aspects of system design: (1) real-time scene sensing, (2) feature (“cue”) selection and (3) body size, posture and movement analysis (Figure 1). For scene sensing, the use of single or multiple cameras, as well as stereo and a thermal systems, will be analyzed. For feature selection and analysis, approaches ranging from simple region occupancy features to detailed human body model pose estimation will be investigated. Table 1 shows comparative estimates of the levels of robustness, accuracy, speed and cost for different combinations of these approaches.

For both scene sensing and feature selection and analysis, the choice of the method will depend on its robustness and cost. However, using stereo or multi-camera systems with high-level human body modeling would provide information useful not only for optimal airbag deployment but for other applications with minimal extra effort. High quality input data and detailed analysis of body pose can also be used to enhance safety by analyzing driver alertness and could also be used to build intelligent interfaces to different in-car devices such as the mobile phone or radio [18].

To determine a passenger’s position in the seat, the area between the back of the seat and the dashboard can be divided into different sections. A diagram of the In-Position, Out-of-Position, and Critically Out-of-Position areas in the passenger seat is shown in Figure 2. Analyzing these regions, we can categorically examine the human body under the various positions a person can take on in the passenger seat, including sitting in position, leaning forward, reaching down, seated with the seat advanced, reclined, slouched, knees on the dashboard, on the edge of the seat, etc. The set of possible positions can be expanded to include infants, three-year-old children, six-year-old children, 5th-percentile female adults and 50th percentile male adults, which is a subdivision that this standard makes. Associated with each position is a desired airbag response, either deploy normally, with limited power, or completely suppressed. Associated with each desired operation is a cost of making the wrong decision. This cost is weighted relative to the other positions that the systems are to recognize. Figure 3 shows selected positions considered in our investigations.
Figure 1: Occupant Position and Posture Based Safe Airbag Deployment Approach

Figure 2: Schematic showing the In-Position, Out-of-Position, and Critically Out-of-Position regions of the passenger seat.
III. Related Studies

The focus of this research is to investigate vision-based systems that estimates occupant size, position and pose. Vision is attractive because it is passive and can provide a multitude of cues for determining how to deploy an airbag and as well as other uses from a single sensor. Alternative solutions for this purpose include a system of measuring the weight distribution of a seat. Technologies for measuring the presence and position of an occupant include ultrasound, capacitance, near-infrared spot. For detecting a rear-facing infant seat – a “must not deploy” occupant – resonating proximity sensor have been employed to detect the presence of that.

A number of experiments exist using active illumination [1], [2], [3], [4], [5] to capture the scene features. They range from unobtrusive near-infrared LEDs to projecting light patterns and emitting multiple flashes to light the scene. The benefits of active illumination is a system insensitive to different types of scene illumination changes, but the gain comes at the cost of being obtrusive to the environment. In contrast, our focus is purely on unobtrusive active illumination or passive scene sensing, and we show that obtrusive lighting schemes are not necessary for robust estimation of occupant posture information. Reference [6] presents an approach where four image features and a series of learning algorithms are used to classify conditions that are safe or unsafe for airbag deployment. These four features were 2-D and 2½-D features, edges, motion, shape and range. In [6] however, it does not consider emitted energy from the passenger nor volumetric representations of the passenger acquired from multiple points of view in the context of side-by-side system comparisons. Furthermore, these systems rely on the fact that a complete and varied reference image set of all passenger seat occupant types and positions, including the many types of infant seats, are known a priori for the training of the system to recognize occupant posture.

Faber [7] previously documented the approach that restrict the problem of describing the occupant space as occupied or unoccupied using sparse range data of the passenger seat back. Stereo-based range data was used to detect the presence of a backrest to determine occupancy. Although this system was able to acquire occupancy information from the seat well, this system lacked detailed occupant size, position, and pose information.

In [8], Krumm and Kirk augmented the class space to include the presence of rear-facing infant seat, which together with the unoccupied classification requires the airbag to be turned off. Krumm took both image intensity (2-D) and stereovision-based
(2½-D) range data and found for each class the principle components, with which nearest neighbor classification was performed. Both systems were found to be quite successful in classifying the passenger seat into these 3 classes. However, this approach required an even larger training data set to include the types of car interiors along with person type and position to achieve this accuracy. And like [7], the information this system provides lacks detailed occupant position information.

In better describing occupant position, the head appears to be the easier human part to detect that simultaneously provides rich implicative positional information on the rest of the body. Remaining in the 2-D category, Reyna, et al [9] uses a modified Support Vector Machines (SVM) method to estimate the location of the head, much in the same way face recognition is performed using SVM. They found an accuracy rate of 72% correct-detection and 33% false alarm. Drawbacks in such a system assume that a representative set of head images are a priori known, and hardware requirements for real-time execution of this algorithm are considerable. In contrast, we propose a method that gathers thermally emitted, stereo depth and volumetric information that requires no training and, with the exception of the third, operates in excess of 15fps on an off-the-shelf Windows PC.

In the 2½-D and 3-D category, the research community has made some effort. We can subdivide their approaches into two main classes: One based on region occupancy detection and the other based on object tracking. The first is an approach where the region in which the feature resides, regardless of whether it be a head, arm, limb or noise, determines the outcome of the classification. The other is an object tracking-based approach where a particular body part is tracked, providing unambiguous body pose information, limb orientation, angle with respect to other body parts, sizes, and lengths. The lack of a fit to a body model usually implies the existence of an object other than a human body, requiring other models to be used in detection of child seats, inanimate objects, etc.

Lequellec et al [3] approached the problem of modeling occupants by way of projecting a pattern of dots onto the occupant and detecting their position in 3-D space using epipolar geometry, a stereovision based technique (2½-D). Devy et al [21] gets rid of the active illumination requirement of [3], relying solely on stereovision and features in the scene to provide a dense stereo reconstruction of the occupant – 3000 to 5000 as opposed to 400 3D points in [3]. Both systems by Devy and Lequellec rely on the surface of the occupant. Our effort differs in that we fit the acquired data to a model which we then track from frame to frame.

IV. Design of Multimodal Video Capturing System, Testbeds, and Instrumented Vehicle

To provide an adaptable experimental test bed for evaluating the performance of various sensing modalities and their combination, two test environments, based upon a Daimler-Chrysler S-class test frame and a Volkswagen Passat (LISA-P), were outfitted with a computer and a multitude of cameras and acquisition systems. Of principal importance in the hardware specification and software architecture was the ability to capture and process data from all the sensor subsystems simultaneously and to provide facilities for algorithm development and offline testing. Figure 1 shows the laboratory based testbed. Various sensory and computing modules used in this laboratory test frame and the instrumented automobile LISA-P include the three types of camera modules (thermal infrared, trinocular stereo, and NTSC color cameras), high speed synchronized video stream capturing hardware, and power supply, and high volume storage. Details of these test beds are presented below.
The LISA-P instrumented automobile is equipped with a 120V alternating current power distribution system. This comprises a marine-type 1.0 kW true sine wave inverter, an auxiliary battery and appropriate isolation circuitry. The availability of 120 VAC allowed for expeditious system development using readily available cameras and power supplies that can be vetted in the laboratory. This approach requires less engineering than the alternative of adapting equipment to the vehicle’s 12 VDC supplies, and yields an easier transition from laboratory to vehicle.

The computing platform consists of a commercial Xenon PC with a high throughput disk subsystem for streaming video data. This subsystem consists of four 15 Krpm Ultra320 SCSI disk drives in a RAID0 array which achieves in excess of 220 MB/s sustained, formatted throughput. This data rate allows for the capture of several simultaneous high-resolution video streams. The computing platform allows for a good deal of processing to be done in real-time as well, but normally in the course of algorithm development speed is achieved after functionality, and data collection is expensive. Hence, the overriding requirement is to capture the data for efficient off-line development.

The video capture hardware currently consists of a dual CamerLink PCI-X capture board, an analog, color RS-170 PCI capture board, and an IEEE 1394 PCI host controller. Additionally, a quad splitter is available to allow four half resolution images to be acquired simultaneously using the analog capture board. The variety of acquisition devices allows for experimentation with a wide range of imaging sensors. Three video sensing systems are being used in the experiments described in this report.

The first sensor system is a trinocular stereo system from PtGrey Research, which provides 2½-D imagery. This consists of 3 black and white 640 X 480 element 1/3" CCD sensors mounted with 3.8 mm focal length lenses in a common enclosure as two stereo pairs in an isosceles right triangle. Integrated digitization and buffering circuitry and an IEEE 1394 interface allows transfer of video from all three cameras to the host. This system is supplied with host software that performs the correspondence calculation and provides a disparity image. The nearest object that can be seen by all cameras is roughly 30 cm from the face of the enclosure, which poses some restrictions on camera placement in the automobile.

The second sensor uses a miniature 2-D thermal infrared sensor, Raytheon model 2000AS. This device provides response in the TIR spectrum (7 - 14 microns) and an analog video output. The camera uses an 160 by 120 element amorphous silicon focal plane array and lens that produces a 35° x 50° field of view. It requires no cooling; the absence of active cooling provisions allows the sensor head/lens assembly to be quite small (~125 cm³) so that it can be unobtrusively mounted on the dashboard. The camera has no absolute calibration and is subject to considerable drift in both gain and offset with temperature. It does have a mechanism for correcting per pixel gain variation, which employs a shutter that momentarily drops in front of the focal plane.
array every few minutes. Miniaturization and cost reduction is moving at a rapid pace in thermal IR (TIR) cameras, with roughly a four-fold decrease in both size and price in the last two years; in selecting this device, we intend to test something representative of what may be reasonably added to a passenger car a few years hence.

The third sensing system provides 3-D imagery through voxelization. The hardware is comprised of 3 color 1/3” CCD cameras, each producing NTSC output with 2.8mm focal length lenses and a quad video signal combiner to merge the three video signals into one, such that each input video frame occupies a quadrant of the output video frame.

The placement of the cameras is shown in Figure 5. The output of all seven cameras capture synchronously and the stored images of one experiment with a male subject are shown in Figure 7.

![Figure 5: Multimodal video cameras, synchronized capture, computing, storage, power modules in the LISA-P instrumented vehicle testbed.](image)

The software architecture was designed to allow for efficient development of multiple video processing algorithms and their incorporation into an integrated framework. To this end the acquisition, storage, user interface, and display functionality are modular and separate from the image processing functionality. Two applications have been developed, one for processing and capturing live data, and one for processing the captured data for off-line algorithm development in the laboratory. Both applications use the same frame processor C++ object interface encapsulating video processing algorithms. The block diagram for this basis frame processor is shown in Figure 8. This standard interface ensures that algorithms developed separately in the laboratory can be painlessly integrated on the test bed.
Figure 6: Top Left: LISA-P Test bed with the power supply, PC-based capturing system. Right: Three NTSC cameras (front driver and passenger windows and front dash), thermal camera (front dash), and a stereo camera (roof rack) in view. Bottom: Test subject in LISA-P during an experiment.
Figure 7: Example acquired image data from LISA-P. Top Left, Top Middle, Bottom Left: Multiple perspective NTSC images. Bottom Middle: Thermal IR image. Top Right: One of three grayscale images for depth map calculation. Bottom Right: Resulting depth image from stereo correspondence.
The live data application is configurable at compile time to capture one or multiple streams using a digitizer-independent image acquisition framework derived from the Microsoft Vision SDK (VisSDK). This approach minimizes the effort required to run the live data system on different hardware, so laboratory machines fitted with various makes and models of digitizer may be utilized for algorithm development. The live data application acquires multiple video streams in a synchronous manner, so that for each time step, one frame from each stream is available. Therefore the results of the various sensor modalities may be compared on a frame-by-frame basis. The central and synchronous nature of the capture ensures that system resources are cooperatively shared among the acquisition subsystems. The processed and/or raw data is combined in a single AVI file to minimize disk seeks and written to the RAID0 array.

The video display functionality is also derivative of the VisSDK and independent of the processing algorithms. A tiled array of video pane windows displays the results of each frame processor object included at compile time. Each object can request multiple panes for output display upon initialization and may provide a pointer to a processor-specific GUI window that may be summoned from the main application for changing parameters, switching on and off functionality, etc. This allows the design of interfaces specific to each frame processor while ensuring compatibility with the system as a whole.

The captured data application allows raw video streams captured using the live system to be played back and routed to the same frame processor object used in the live system. Similar display functionality is provided and the identical processor specific GUI may be summoned, simulating the test bed environment precisely from the viewpoint of the frame processor object. In this way each frame-processing object may be developed separately off-line and later incorporated into the live data application with a minimum of integration issues.

![Figure 8: Software Architecture for Live Side-by-Side Processing](image)
V. Stereo Based Head Detection and Tracking

1. Stereo based Head Detection and Tracking Algorithm

The stereo-based approach consists of a correlation matching step, where the head is located in the background-removed disparity data. The template used is a zero-mean oval-shaped filter meant to generalize the various depth images of human heads. The coordinate of the window with the highest response indicates the location of the head. A Kalman filter is used to remove the noise associated with missed head detections. The depth pixels of the background are updated recursively pixel-wise with an AR(1) filter, while the depth pixels of the detected foreground are not updated. The system flow diagram is shown in Figure 9.

![Flowchart of the real-time Binocular stereo based head detection and tracking algorithm](image)

Depth maps are captured using the Triclops trinocular stereo camera, capturing at 160x120 pixel resolution at 15fps. Using the disparity map, the system then filters out any pixels that fall outside of the car. Once the in-car pixels have been determined, a background model of the unoccupied car is determined by averaging the first 10
frames of the disparity image. After the first 10 frames, the background is adaptively updated every frame using the relationship

$$B_{k+1} = \begin{cases} 
\alpha B_k + (1 - \alpha)F & F \in \text{Background} \\
B_k & F \in \text{Foreground} 
\end{cases}$$

where $B_{k+1}$ is the new background, $B_k$ is the old background, $F$ is the current frame, and $\alpha$ is a scalar between 0 and 1, often called the “forgetting factor” that controls the rate at which new frames are incorporated into the background. To ensure that a still passenger is not incorporated into the background, once the body of the passenger is found, that portion of the image is not included in the background model update.

A simple zero-mean oval filter, shown in Figure 11a, is then applied on the disparity difference image. The dimensions of the filter are 20x30 pixels. The gray portion of the filter represents negative values, and the white, positive values. The maximum filter response is found within a specified window and considered the location of the centroid of the head if that value exceeds a minimum set threshold.

The max-search is limited to within a rectangular search window around the last head location. Using the window allows the algorithm to perform the head filter operation on a much smaller portion of the image and thereby increase the speed of the algorithm. If the head is not found in the smaller window, then the algorithm attempts to locate the head by filtering the entire image, and searching for the maximum filter response in the larger area.

The approximate length of the torso can be determined by extending a line from the centroid of the head down to the waist, assumed to be at the intersection of the seat bottom and seat back. Currently this intersection is manually initialized. In Figure 11, the line drawn from the head to the seat intersection is used as the torso length estimate. The current system is able to report the average length of the torso, and a future implementation will use this information to constrain the head location.
2. Performance Evaluation

Initial tests were performed in non-moving car situations, both indoors and outdoors under various lighting conditions [15]. Indoor experiments were conducted using the Daimler-Chrysler experimental frame, while outdoor experiments occurred in the LISA-P test bed. The intent was to make the head detection algorithm robust to various lighting conditions, including indoor light, NIR illumination, sun, and shade.

Successful head detection in the initial tests lead to algorithm refinement and the development of the most recent test.

A series of experiments were conducted where data was captured for both the stereo and IR methods simultaneously. The desire is to have a direct comparison of the two head tracking methods on a frame-by-frame basis. The following describes the outcome of the stereo portion of the tests.

The tests were conducted from 8-11 am on a sunny morning using the LISA-P test bed. The stereo camera was placed on the driver’s side roof rack, looking down on the passenger's seating area. For each test run, 300 frames of video are captured frames at 15fps, during which time the background model is captured and the subject enters the vehicle. The car is then driven at road speeds until the entire 300 frames are captured. The resulting videos are approximately 20 seconds for each test run.

In the test sequences, the subject enters the vehicle:

- without anything on her head.
- wearing a white baseball cap with the brim facing forward
- wearing a white baseball cap with the brim facing backward
- wearing the hat backwards, resting her feet on the dashboard

The test run data is analyzed beginning when the subject has entered the car and the door has completely closed. Each frame is analyzed for both the head detector only, as well as with a head tracker used to smooth the result. For the head detector only, the results of each frame are categorized as either correct, incorrect or not found. The head detector is classified as correct if the tracked point lies anywhere on the subject’s head. For the head tracker, results are specified as either correct or incorrect. Taking advantage of the Kalman filter to estimate the location of the head in frames when the head detector does not find a valid head region, there is no need for a not found classification.

In the following examples, the tracked positions are represented by the ellipse and the dot marking the ellipse center. The instantaneous head detection positions are represented by the dot that is not in the center of the ellipse.
Figure 12: (a) The Head Detector and Tracker are correct. (b) The Head Detector does not find a valid region, but the Tracker is still correct. (c) Here the Head Detector gives an incorrect location, but the Tracker is still correct.

Figure 13: Here the Head Detector is correct, but the Tracker is incorrect due to errors in previous frames. In the next frame, the Tracker has corrected.

Table 1 – Summary of Stereo Test Results

<table>
<thead>
<tr>
<th>Stereo Tests</th>
<th>Head Detector Only</th>
<th>Head Detector with Tracker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Head Present</td>
<td>Head Present</td>
</tr>
<tr>
<td></td>
<td>Correct</td>
<td>Wrong location</td>
</tr>
<tr>
<td>No Hat 1</td>
<td>149</td>
<td>6</td>
</tr>
<tr>
<td>No Hat 2</td>
<td>164</td>
<td>5</td>
</tr>
<tr>
<td>Hat Forward 1</td>
<td>243</td>
<td>2</td>
</tr>
<tr>
<td>Hat Forward 2</td>
<td>110</td>
<td>52</td>
</tr>
<tr>
<td>Hat Forward 3</td>
<td>108</td>
<td>17</td>
</tr>
<tr>
<td>Hat Forward 4</td>
<td>206</td>
<td>6</td>
</tr>
<tr>
<td>Hat Backward, Feet on Dash</td>
<td>125</td>
<td>26</td>
</tr>
<tr>
<td>Total Result</td>
<td>1324</td>
<td>115</td>
</tr>
</tbody>
</table>
An average of 85.5% correct detection rate is obtained using stereo cameras in conjunction with the elliptical filtering method. It is clear that an immediate way to improve the results is to reduce the number of frames for which the head tracker gives no response. For example, the lowest two correct rates occur when the algorithm gives no response 38% of the time. As a comparison, this method obtains 99.6% accuracy when only 3-4% of the frames have no response. One way to increase the response would be to use more accurate cameras or cameras that are able to mount on the dash. This would increase the density of our depth map and would make it more likely that the head would be contained in a valid region of stereo correspondence. Improvements to the Tracker could be made by taking head velocity into account when tracking. Currently, when long sequences of frames with no response occur, the Tracker remains static. If the subject is moving across the field of view, the Tracker prediction will eventually be invalid. However, by taking the head’s velocity into account, the Tracker prediction will likely remain valid longer.

Incorrect head responses are fairly low in these tests, but can also be improved. The majority of these responses occur when the ellipse filter locks on to things far away from the head, like the back of the seat or the headrest. These areas can be eliminated through more sophisticated background modeling and more accurate cameras. By improving the background accuracy and reducing the noise, the algorithm should be able to more easily segment the background from the foreground as well as generate a more accurate and complete depth image.

In runs “Hat Forward 2” and “Hat Forward 3”, the head detection algorithm was less successful than in the other tests. These failures can be attributed to two major causes: incorrect head detection in a frame and frames where no valid head region is found.

In “Hat Forward 2”, the incorrect head detection rate was 22.3%, out of proportion with the other tests, which averaged 9.8%. Naturally, this high misdetection rate leads to the 57% accuracy rate for this run. In Figure 14, there are three examples of incorrect head detection. In the first two frames, the head is “detected” near the small of the back. In the third frame, the head is “detected” on the seat’s headrest.

These incorrect detections are due to both the limitations of our current stereo setup in conjunction with the nature of our ellipse filter. In using the Triclops camera system, it is clear that for certain lighting conditions and situations, the depth image produced by the Triclops is both noisier and more incomplete than in ideal conditions. This often leads to the head area being too large, too small, or missing altogether. Also, other areas in the field of view are accentuated in the depth map. For example, in the first image of Figure 14, the head rest area is not apparent in the depth image, while it clearly is in the last image, captured only a few seconds later.

The ellipse filter is based on the assumption that the head is the most elliptically shaped object in the scene. However, when noise and incomplete depth maps make this assumption false, it is clear that the maximal filter response is likely to be in an area that does not contain the subject’s head.

In “Hat Forward 3”, these limitations lead to a similar problem. For a large percentage of the run, the algorithm is unable to detect a valid head region. For run 6, this rate is 49.4% compared to an average rate of 22.0% for the other test runs. In Figure 15, notice that the detector does not find a valid region for this frame. Also, the Tracker position is incorrect. This is due to the long sequence of previous frames where the head was undetected. Since there is no new data on the head location, the Tracker could not be updated.
In the case where no valid head region is found, the limitations of the stereo setup and the ellipse filter are also to blame. However, instead of selecting an invalid region, the maximal filter response of the ellipse filter on the depth image is lower than the user-defined threshold, so no point is selected. In the above figure, this most likely occurs because the brim of the subject’s hat creates a head area that is not elliptical enough to register a high enough response with the elliptical filter. Also, no other region in the frame gives a high enough filter response to have an invalid region selected.

VI. Head Detection and Tracking using Thermal Infra Red Video

A. Head Detection and Tracking Based Upon Uniformity and Contrast of Human Thermal IR Signatures

Two algorithms have been developed and evaluated for the detection and tracking of heads in the Thermal IR spectrum. These each make use of a simplification of the concept of soft segmentation by human skin class membership probability from
TIR intensity values as described in Socolinsky, et al [11]. They may be broadly described as region-based and edge-based.

1. Face Detection and Tracking Based on Uniformity of Thermal IR Signature

The first algorithm is region-based, derived directly from the algorithm developed above for stereo imagery. It was modified in two ways:

- The image intensity values were remapped to approximate the per pixel probability of membership in the human skin class using a simple triangular probability distribution, with the peak value and slope manually set, as shown in Figure 16. This results in an enhancement of skin areas while suppressing those regions with much greater or lesser TIR emissions.
- The adaptive background subtraction was not used. Unlike the stereo data, where events outside the window of the moving car can be filtered out by range thresholds, the background for the IR images is constantly changing due to car movement, and therefore the background subtraction is inappropriate here. It may be possible to mark only those areas on the interior of the car for background subtraction, but no attempt was made here.

After remapping the intensity values, the image is convolved with the same zero-mean ellipse template as described in the stereo section and the head location is given by the maximal response.

![Figure 16: Triangular Intensity to Probability Mapping](image)

**Figure 16: Triangular Intensity to Probability Mapping**

![Figure 17: TIR Images: (a) before, and (b) after Intensity to Probability Mapping](image)

**Figure 17: TIR Images: (a) before, and (b) after Intensity to Probability Mapping**
After experimentation with the region-based algorithm (see results below), it became apparent that the zero-mean ellipse template was not sufficiently discriminating for the head as opposed to hands or other similarly sized (in image space) body parts. It also failed to provide any estimate of head size or pose without adding multiple convolution templates and considerably increasing the computational load. Head size and pose estimates may be useful in future depth from size extensions; therefore a new algorithm was developed to address these issues.

2. Face Detection and Tracking Based on Face/Background Contrast in Thermal IR Signatures

The edge-based algorithm uses the skin class probability from TIR intensity concepts from [11] and also modifications of Eleftheriadis, et al [12] to yield an edge-based head detection algorithm specifically for TIR imagery providing head pose and size estimates and which is easily implemented in real time (in excess of 30 fps on commercial PCs).

The Edge-based algorithm flowchart is shown Figure 19, and the various steps are described below:

1. The 8-bit per pixel intensity image is remapped to approximate the probability of membership in the human skin class using a simple Gaussian PDF with the mean and variance manually, empirically set. This is implemented using a pre-calculated look-up table, and yields an 8-bit output image with pixels valued 255 most likely to be skin and pixels valued 0 least likely. The probability look-up table, \( P(I) \), was populated with values given by

\[
P(I) = 255 \frac{(I - \mu)^2}{2\sigma^2}
\]

where intensity is \( I = 0 \ldots 255 \), and the mean, \( \mu \), and variance, \( \sigma \), are manually set on a per-run basis.

![Figure 18: Gaussian Probability Remapping](image)

2. The probability image obtained from step one is subjected to one iteration of grayscale erosion where each pixel is replaced by the minimum of its eight neighborhood. This eliminates the consideration of one pixel wide edges in the probability image caused by the boundaries of hot objects passing rapidly through the human skin emission region.

3. The resultant image is then smoothed with a Gaussian kernel and down-sampled by a factor of 4 from its original 240x320 pixels to 60x80 pixels similar to the algorithm described in [12].

4. The gradient magnitude image is calculated using 1x3 Sobel operators in vertical and horizontal orientations. This is stored as a floating point image.

5. A two-stage horizontal run detector as described in [12] is applied to the gradient magnitude image. This process consists of first applying a threshold to the gradient magnitude image to obtain a binary edge image. In our implementation this threshold was set to 0.6 standard deviations of the gradient magnitude above the mean gradient magnitude, which gave a binary edge image of fairly constant density. This image is then examined in two stages: a coarse scan and a fine scan as in [12]. In the
coarse scan the image is divided into 5x5 pixel blocks and those blocks that have at least one edge pixel are marked. In the fine scan, each run of consecutively marked horizontal blocks are examined. For each such run, the first scan-line with an edge pixel is found. On this scan-line, within this run of blocks, all pixels between the first and last edge pixels possibly correspond to the top of a head and are marked as possible seed points for ellipse contour template matching.

(6) A set of pairs of pre-calculated templates for ellipse contours and ellipse boundaries are overlaid on the gradient magnitude image from step 4. These template pairs consist of one template that represents the contour of the ellipse and one that represents the boundary just outside the ellipse contour and are described in [12]. Unlike [12] two slightly different figures of merit are used to filter ellipse candidates. In [12] the ellipse templates were applied against the binary edge image and the figures of merit were calculated by the number of edge pixels underlying the templates normalized for the size of the template, with the pixels at the top of the ellipse contour given 50% more weight. In this algorithm, the templates are applied to the gradient magnitude image itself and the average gradient magnitude underlying the template is used as a figure of merit (also with 50% extra weight given to those pixels at the top of the ellipse.) This approach yielded better fitting that justified its small impact on computation speed. Only ellipses that concurrently exceeded a threshold on average gradient underlying the contour and were below a threshold for average gradient underlying a boundary were given further consideration. Our implementation tested 231 template pairs per seed point ranging from 10 to 44 pixels high, 60 to 120 degree inclination and with eccentricities ranging from 1.10 to 1.55, more or fewer templates can be used for more precise fitting or greater speed, respectively.

(7) At this point in the processing, we have a list of ellipse candidates. Whereas in [12] this step was followed by the selection of the "best" ellipse, we found that for TIR images this approach would sometimes find elliptical objects that were not body parts or elliptical contours that occurred due to the chance arrangement of several non-elliptical objects. To eliminate most of these candidates, the average skin probability of the ellipses' bounding boxes are calculated by summing the pixel values in the image from step 2, we refer to this quantity as $P_{AVE}$. Only those candidates exceeding a threshold for this average probability are further considered.

(8) We now have a list of ellipse candidates that each have large mean gradient values on the contour, $I_i$, and small mean gradient values on the boundary surrounding the ellipse, $I_e$, and with an underlying skin class probability, $P_{AVE}$. From among these candidates, we select the "best" by choosing the candidate which maximizes for this expression:

$$P_{AVE} \cdot \frac{(1+I_i)}{(1+I_e)}$$

(9) To reduce the effects of measurement noise in our estimates, all five ellipse parameters are tracked using a Kalman filter. The tracked parameters are ellipse center coordinates, the axes dimensions, and the inclination angle.

B. Performance Evaluation

The relative performance of the two TIR algorithms was evaluated on a data set consisting of 2400 frames of video acquired in a moving automobile with 4 passenger subjects.
Results were deemed correct if the algorithm placed a head center point somewhere on the subject's head or neck if a head was present in the image or if the algorithm failed to find a head if none was present. It is not always possible to accurately determine the jaw-line in the TIR imagery, and for the purpose of head tracking in the context of intelligent airbag deployment, the neck/head position is a sufficiently accurate indicator. Failures by the algorithm were further classified into categories of the head not being found although present, the head being found in the wrong location or a head being found when no head was present.

As was mentioned earlier, experiments with the region-based algorithm sometimes gave poor results. This was principally due to the algorithm finding some other body part of approximately elliptical profile. This was usually a hand, but
Occasionally the algorithm would find a shoulder, or a knee. The zero-mean template used gave similar response for a hand held near the camera as a head in normal sitting position. This sort of failure accounted for the poor performance on the Adult Male subject as shown in Table 2. The poor performance with the Female 2 subject was due largely to the algorithm triggering on an area of the car’s upholstery that had been warmed in the sun in the absence of the subject. The algorithm triggered on this area even though the area was not even roughly elliptical.

The edge-based algorithm’s performance was much more robust, with typically 90% accuracy. It too suffered the occasional failure by misidentifying a hand as a head, but only when the hand’s aspect was quite elliptical (i.e. palm out, fingers not spread). It rarely found some other body part, usually only for a frame or two in sequence. It would sometimes fail to find the head when a head was present. This happens most often during rapid head movements. The nature of the thermal camera’s focal plane array is such that the pixels’ time constants are on the order of the frame rate. This causes an afterimage and therefore motion blur. Since the algorithm requires a sharp edge (high gradient magnitude) in its very first, coarse scanning, the appropriate seed points for the subsequent ellipse fitting are never found. On a couple of occasions, with a strong wind blowing in the open window at roughly 35 mph, the passenger’s skin temperature was lowered to the point where it was momentarily cooled and classified as background.

The fact that the failures of the edge-based algorithm were only occasional and usually consisted of a sequence of few wildly incorrect frames (for instance finding a hand) among a series of a hundred or more correct frames caused the tracker to degrade the performance of the system as a whole. Although it did improve the stability of the head location during those sequences where the head was found correctly, a single result far from the true head location would require a few frames for the Tracker to settle on the new position whereas the raw reading was correct in the very next frame. The Kalman filter used in the head tracker is designed to model a system with Gaussian process and measurement noise, whereas this noise is of an impulse flavor. It is probably appropriate to apply some non-linear temporal filter (i.e. a median filter) to the sequence of head position readings before they are incorporated into the tracker estimate.

C. Ongoing and Future Enhancements

The performance of the edge-based algorithm is quite good, but it does occasionally suffer from a preference to other body parts from the head. In these cases an ellipse candidate was usually found for the head, but scored lower than the hand when the final selection of the best ellipse was made. One possible solution to this is to use more computationally expensive algorithms on the small set of ellipse candidates that survive to this stage of the selection process. One such approach is to search the candidate ellipses for one or more sub-features that one would expect to find on a face, but not on a hand. Eyes, ears, and mouths show up quite distinctively in TIR imagery although they appear completely different than in visible spectrum images.

The algorithm may also benefit from calculating $P_{AVE}$ as the estimated probability underlying the ellipse as opposed to that underlying the ellipse’s bounding box, at the cost of slightly more computational complexity.

The size of the ellipse may be used to get a rough depth from size estimate, although currently the head is sometimes found from the top of the head to the chin, and sometimes found from the top of the head to the neck, yielding two distributions of head size. If we can robustly measure one or the other, then the head image size data might be combined with switches in the seatback to give a reference head image size for the passenger sitting fully back. It would then be possible to give an estimate of the passenger distance from the camera (and hence the dashboard) by comparing the head
image size with that in the reference position. The results of the process for inferring depth information from the ellipse size in thermal IR images is illustrated in Figure 20.

Table 2: TIR Algorithm Comparison Results (Numbers reflect count for when head is present)

<table>
<thead>
<tr>
<th>TIR Tests</th>
<th>Region-based Head Detector w/tracker</th>
<th>Edge-based Head Detector w/o tracker</th>
<th>Edge-based Head Detector w/tracker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Wrong location</td>
<td>Not Detected</td>
</tr>
<tr>
<td>Baby Region</td>
<td>443</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Male Region</td>
<td>302</td>
<td>298</td>
<td></td>
</tr>
<tr>
<td>Female 1 Region</td>
<td>403</td>
<td>39</td>
<td>8</td>
</tr>
<tr>
<td>Female 2 Region</td>
<td>483</td>
<td>99</td>
<td>107</td>
</tr>
<tr>
<td>Total Result</td>
<td>1631</td>
<td>443</td>
<td>115</td>
</tr>
</tbody>
</table>

VII. Stereo and Thermal IR Head Detection and Tracking Algorithms: Comparative Analysis

As mentioned above, a series of experiments were conducted where data was simultaneously captured for both the stereo and IR methods. The data was collected on a sunny morning with the subject with under the conditions listed above in section V.2.
The stereo camera was placed on the roof rack above the driver, looking down on the passenger’s seating area. The IR camera was placed on the dashboard, also directed at the passenger’s seating area. 300 frames of video at 15fps was captured of the subject entering the car, closing the door, and the car then driven at road speeds.

Ideally, a method of comparison would consist of answering the following question: how well does a system make decisions on the most critical occupant sizes, positions, and poses, at the same time addressing the less critical ones airbag deployment conditions. In the standard issued by the NHTSA [14], it states that airbag systems must be able to disable itself when occupants of five standardized sizes are in various positions. One way is to associate with each occupant size and position the cost for making the erroneous decision and having the airbag deploy or not deploy by mistake. For every system, there is an associated miss detect rate. False alarms are not considered since there is a comprehensive set of occupant sizes and positions. An acceptable cost function would be the sum of false alarm and miss detection rates weighted by the cost of making an erroneous classification for each occupant size and position. This “likelihood of safe deployment” would be the measure of goodness for the system under test.

In the stereo and TIR tests, the distance from the detected head to the dash is the only variable to be considered when classifying a person to be in position or out of position. Since only one subject was used in these tests, the other variable, occupant size, is constant. A correct head detection is equivalent to a valid distance estimate between the dash and the head, within some distance variance. Below is table of detection rates for the Stereo and TIR head detection and tracking algorithms.

<table>
<thead>
<tr>
<th>Test Run</th>
<th>Stereo Detector Only</th>
<th>Stereo Detector w/ tracker</th>
<th>TIR Detector Only</th>
<th>TIR Detector w/ tracker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Incorrect</td>
<td>No Response</td>
<td>Correct</td>
</tr>
<tr>
<td>No Hat 1</td>
<td>60%</td>
<td>2%</td>
<td>38%</td>
<td>85%</td>
</tr>
<tr>
<td>No Hat 2</td>
<td>68%</td>
<td>2%</td>
<td>30%</td>
<td>92%</td>
</tr>
<tr>
<td>Hat Forward 1</td>
<td>95%</td>
<td>1%</td>
<td>4%</td>
<td>100%</td>
</tr>
<tr>
<td>Hat Forward 2</td>
<td>47%</td>
<td>22%</td>
<td>30%</td>
<td>57%</td>
</tr>
<tr>
<td>Hat Forward 3</td>
<td>44%</td>
<td>7%</td>
<td>49%</td>
<td>73%</td>
</tr>
<tr>
<td>Hat Forward 4</td>
<td>87%</td>
<td>3%</td>
<td>11%</td>
<td>95%</td>
</tr>
<tr>
<td>Hat Backward Only</td>
<td>52%</td>
<td>11%</td>
<td>37%</td>
<td>83%</td>
</tr>
<tr>
<td>Hat Backward, Feet on Dash</td>
<td>96%</td>
<td>0%</td>
<td>3%</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td>69%</td>
<td>6%</td>
<td>26%</td>
<td>86%</td>
</tr>
</tbody>
</table>

The stereo method greatly benefited from the addition of the tracker. When sequences of frames illicit no response from the head detector, previous tracker estimates are often still valid. Naturally, this increases the correct detection rate. On the other hand, the TIR algorithm does not provide a no response detection. When the detection is incorrect, the head tracker is misled away from the head. The rate at which the head tracker is misled can be reduced by having it track slower, but at the cost of not being able to track quick head movements.

VIII. Discussion and Extensions

A. Stereo and Thermal IR Head Detection and Tracking Evaluation Summary

This test set up is particularly unique in that it allows for a direct comparison of the stereo and TIR head detection methods on a frame-by-frame basis. The test bed allows for the simultaneous capture of multiple inputs from differing modalities all...
synchronized to the same clock. The test bed is also set up to allow for easy modularization so that both new cameras and new algorithms can be introduced into the system and tested with minimal further development. As both camera and algorithm advancement are anticipated, the test bed gives an ideal circumstance to further and extend this research.

Both algorithms achieve a high average accuracy in detecting and tracking the head. At success rates of 86% and 93% respectively, it can be concluded that both the stereo and TIR systems can be used to robustly detect the location of the head with high reliability. The resulting head location information can give the necessary information to decide in the manner in which an airbag should be deployed.

Although the algorithms achieve a high success rate, it is still apparent that both algorithms suffer from certain drawbacks that should be resolved through further research.

The current placement of the stereo camera is non-ideal. Due to the Triclops camera’s enclosure size and focal length limitations, the only viable position for camera placement is looking down on the passenger from the driver’s side. If the camera were placed so the subject was viewed from a frontal perspective, the stereo algorithm could more easily exploit the depth map, which would be a direct indicator of the subject’s distance from the dashboard. A new stereo camera system is being designed to allow for this frontal camera placement.

The TIR camera also exhibits unwanted characteristics. Currently, the TIR camera sensors have a non-stationary skin temperature to intensity mapping. That is, over time the intensity mapping for skin often changes in the TIR image. This provides difficulty for the TIR algorithm, which relies on a skin intensity probability density function that becomes invalid if the intensity mapping changes. An adaptive recalibration of the intensity mapping would rectify this issue.

The ellipse filter in both the stereo and TIR algorithms had difficulty differentiating the head from other elliptical objects such as the subjects hand, elbow or the headrest. Improvements to the filter to invalidate these types of features would greatly improve detection accuracy. Also, allowing for more ellipse orientations and sizes could improve both the accuracy and robustness of the system.

Despite the success of this initial test, further testing is imperative. This test of the stereo and TIR systems included only one subject at a particular time of day in particular weather. Clearly, many different subjects need to be tested on the system. It is still untested how well the algorithms will perform when subject’s have features such as facial hair, large hats, objects in their lap, are eating or drinking, are very large or very small, or are sitting in unconventional positions. The algorithms are also untested in driving conditions other than a sunny day. Although an exhaustive test of the permutations of subject type and driving condition is impractical, a much larger and extensive test of these variations is necessary to deem the algorithms reliable enough for commercial use.

With head detection and tracking, we are able to acquire a great deal of information concerning the occupants posture and position. We have not yet taken into consideration another very informative cue, which are the limbs and torso of the passenger. For this, we developed a 3D voxel reconstruction system and human body model for the purpose of modeling a passenger inside a vehicle.

B. Multiple Camera 3D Voxel Reconstruction and Human Body Modeling

In [13], we have described a fully automated motion capture system that uses the input from multiple cameras to construct full human body models. The system computes the 3D voxel reconstruction of the human body from its silhouettes from each perspective. The system performs the automatic model acquisition and tracking.
Voxel reconstruction of the human body shape is computed by checking for each voxel in the volume of interest, it projects to a foreground body pixel in each of the image planes. This is shape-from-silhouette procedure is based on Small’s HybridVolume algorithm, which efficiently *voxelizes* multi-camera silhouette data [18].

The automatic model acquisition consists of two parts. First, initial estimates of body part sizes and locations are found using a heuristic procedure that uses the knowledge of the average shapes and sizes of the parts. Then, these estimates are refined using a Bayesian network that imposes the known proportions of the human body.

The body model used for the full-body model is shown in Figure 21. Each axis of rotation in different joints is modeled using the twists formulation. By imposing the limits on the joint angles, such a model is guaranteed to be in a physically valid configuration.

For measurements, we chose a set of points on the human body that are either centroids or endpoints of different body parts. The twists framework for describing the joint rotations leads to a simple formulation of the head tracker, which adjusts the model position and configuration to the measurements in each new frame. To find the locations of measurement points in each new frame, the voxel data is labeled using the model position and configuration prediction from the previous frame. The voxel-labeling algorithm takes advantage of the fact that the voxel reconstruction is of the dimensions of the real person, and uses the known sizes and shapes of body parts to locate them even for large frame-to-frame displacements.

We have applied the described algorithm to several sequences involving sitting people who perform head and torso tilting to simulate the body movements inside a car. Figure 22 shows sample frames for the sitting adult sequence. Figure 23 shows some key joint angles for the sequence with the adult.
The key features of the performed motion are clearly visible. The knees are kept bent at a constant angle, hip angles agree with the torso tilting forward and backward and the neck angle with the head tilting backward and forward. The coordinates of the torso centroid match the same torso tilting motion.

The analysis of the results reveals very good capture of the important quantities, such as the hip angle, torso centroid and neck angle. Body part sizes are given in Table 4.

In this paper, we have demonstrated that a voxel-based motion capture system can successfully extract the posture of a sitting person. Important details such as the head and torso position and orientation are extracted accurately.

The conditions under which these experiments were conducted were idealistic in several aspects compared to the real conditions inside a car. First there was little occlusion of the person’s body in the camera views. The only occlusion was caused by the chair in one of the camera views, but the legs were clearly visible by other cameras. Second, the current system takes approximately 20 seconds to compute the tracking result, which is far from the necessary real-time performance of the “smart airbag” system. However, in implementing this system, speed was not of great concern and many improvements could be made to improve it. Finally, to successfully deploy an airbag, a classification system is needed that makes a decision on the type of posture the person is in. This task requires detailed knowledge of the airbag operation parameters and with previously mentioned tasks is left for future research.

We have demonstrated that a voxel-based motion capture system can successfully extract the posture of a sitting person. Important details such as the head and torso position and orientation are extracted accurately. Furthermore, the uses of a full or even a partial body model are undeniably far reaching. The potential goes beyond occupant posture estimation for the purpose of “smart airbag” deployment, and enters...
the domain of driver fatigue analysis, driver attentiveness, human-machine interfaces inside the car.

Figure 23: Some joint angles as functions of time for the sequence containing a sitting person. $x$-axis: frame number, $y$-axis: radians

Table 4: Average volume and body part sizes extracted from the sequences containing a sitting adult and a child

<table>
<thead>
<tr>
<th></th>
<th>Average volume</th>
<th>Torso</th>
<th>Head</th>
<th>Upper arm</th>
<th>Lower arm</th>
<th>Thigh</th>
<th>Calf</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>voxels/cm³</td>
<td>dim</td>
<td>dim</td>
<td>dim</td>
<td>dim</td>
<td>dim</td>
<td>dim</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$x/y/z$ [mm]</td>
<td>$x/y/z$ [mm]</td>
<td>$x/y/z$ [mm]</td>
<td>$x/y/z$ [mm]</td>
<td>$x/y/z$ [mm]</td>
<td>$x/y/z$ [mm]</td>
</tr>
<tr>
<td>Adult</td>
<td>7740.8/120949.7</td>
<td>138.146/120.216</td>
<td>98.31/113.63</td>
<td>50.5998/50.5998</td>
<td>59.7998/120.216</td>
<td>78.1998/179.399</td>
<td>87.3997/262.199</td>
</tr>
<tr>
<td>Child</td>
<td>3891.5/60804.8</td>
<td>109.109/80.7406</td>
<td>68.7532/91.7084</td>
<td>49.3922/98.7844</td>
<td>38.4874/134.706</td>
<td>64.6589/161.647</td>
<td>51.3166/179.608</td>
</tr>
</tbody>
</table>
IX. Concluding Remarks

New generation of air bags will incorporate information about the position and posture of the occupant in making the “deploy”, “no-deploy” or “partial deploy” decisions. Development of vision based systems for human posture analysis within the challenging constraints of time and automobile interior, in an accurate and robust manner was the main research issue addressed in this paper. It presented a systematic investigation where stereo and thermal infrared video based real-time vision systems are developed and systematically evaluated. It also includes design of several test beds, including instrumented vehicles for systematic experimental studies for evaluation of independent and comparative evaluation in automobiles. Results of extensive experimental trials suggest basic feasibility of video based occupant position and posture analysis system.

X. 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, Dr. Arne Stoschek and Mr. Sascha Sertel 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, Tim Schoenmackers, Kohsia Huang, and Jun-Wen Wu.
XI. References


[14] Occupant Crash Protection Standard (FMVSS 208) NHTSA


