Real-Time Stereo-Based Head Detection using Size, Shape and Disparity Constraints

Stephen J. Krotosky, Shinko Y. Cheng and Mohan M. Trivedi
University of California, San Diego
La Jolla, California 92093–0434
Email: {krotosky, sycheng, mtrivedi}@ucsd.edu

***Abstract*— A real-time stereo-based head detection algorithm is presented and evaluated as an occupant posture analysis system for "smart" airbag deployment. The algorithm uses several constraints to limit the number of head ellipse "candidates" found in the image. These constraints are based on shape, size and depth of the occupant's head. Results of ground truth experiments show that the head detection can accurately estimate the three-dimensional location of the occupant head. Extended experiments illustrate the robustness of the algorithm to poor lighting conditions, to occlusion, and to the presence of other competing head-like objects in the scene.**

I. Introduction

Head detection is a fundamental component in many algorithms designed to detect and track human activity in a scene. The shape and size of this body part is relatively universal across race and age. A robust, real-time head detection algorithm would be an invaluable input to many human-centered applications, including surveillance, recognition, and occupant posture analysis for "smart" airbag deployment.

This paper presents a stereo-based head detection approach that robustly and reliably detects three-dimensional head location information. This is accomplished by constraining the head search space based on its relative size, shape and disparity in the stereo image. The algorithm is evaluated in the context of occupant posture analysis measured against ground-truth in a series of extended in-car experiments.

II. Related Work

Stereo-based occupant detection has been studied in great depth in recent years. Many past occupant detection algorithms relied on extensive training and learning algorithms to obtain accurate detection results. Others focused on the static suppression case of differentiating an adult occupant from children and child seats. The following is a brief summary of previous work in occupant posture detection for airbag safety.

In [1]–[5], extensive examples and training are used to generate detection results using image features for occupancy classification. They are primarily concerned static occupant classification. Method [6] uses stereo range data to determine occupancy, but does not provide a method for actively detecting the head movements throughout the scene. Method [7] is also primarily concerned with occupancy classification for static suppression, but presents a method of head detection using elliptical detection. However, results are only presented for simulated data. Method [8] uses a two-tiered approach to head detection: surface and edge-based features. Head candidate selection is again based on the statistics of a large representative training set.

Head size and appearance models have been used in [9] for locating standing people under indoor laboratory conditions. Our method also employs a size-depth constraint with similar success, but in a more challenging in-vehicle environment, including night driving, uncontrolled daylight illumination, and more drastic changes in occupant size and appearance from the closer camera position and perspective projection.

III. Motivation and Research Objectives

Our previous work [10] also investigated the use of stereo-based techniques for head detection. The challenge was to develop an accurate and robust head detection algorithm without the use of extensive training data. To accomplish this, elliptical templates of various sizes, orientations and eccentricities were applied to the edge map of the disparity image to generate head candidate locations. The best head candidate was selected that maximized a likelihood function that incorporated the quality of the elliptical template to the disparity data, the vertical position in the image, and the relative disparity within the elliptical region. The algorithm described performed well for the experiments that we initially conducted.

In these initial experiments, the 10 cm baseline stereo camera restricted the minimum discernable disparity to about 1 m from the camera. This necessitated that the camera be placed outside the vehicle, looking down from the sunroof, to achieve acceptable cabin coverage. External camera placement is impractical, so a smaller baseline camera that could be mounted inside the vehicle was obtained.

From the external camera location, the occupant is viewed in profile, shown in Fig. 1, and the imaged head size varies little as the occupant moves about the cabin. However, the in-vehicle camera location captures the occupant from a frontal perspective, as illustrated in Fig. 2. From this view the imaged occupant's head size can vary greatly depending on his position in the cabin. Under these conditions, the original proposed stereo head detection algorithm fails, as it has no mechanism to capture the variability of imaged head sizes.

The motivation for this work, then, is to extend the methods proposed in [10]. As before, we would like to perform real-time stereo-based head detection without the use of a large,

Fig. 1. Example of two extreme positions in profile camera perspective. Note how the perceived head size is approximately constant.

Fig. 2. Example of two extreme positions in frontal camera perspective. Note how the perceived head size varies greatly.

representative training set. Additionally, we would like the algorithm to be able to accommodate arbitrary camera placements. To this end, we propose incorporating a size-depth constraint in the form of a foreground disparity image remapping using the probability distribution of the imaged head size at a given disparity (depth).

IV. Stereo Disparity Computation

Stereo Disparity computation methods for in-car occupant pose have been investigated in [11] and in general in [12]. Various correlation methods are discussed, including normalized window correlation, Laplacian of Gaussian transformations, and non-parametric census based approaches. We rely on the algorithm presented in [12] that has the following properties:

- Laplacian of Gaussian transform with L1 norm correlation measure.
- Variable Disparity Search (16-128) pixels.
- Post filtering with a texture-based interest operator and a left/right check.
- Multi-scale capable for capturing larger features.
- $\frac{1}{16}$ pixel range interpolation.

V. Imaged Head Size vs. Disparity

As any object moves closer to the camera, its size in the image increases. We also observe that the imaged head height remains relatively constant to changes in head pose at a given distance away from the camera. Thus, it is desirable to utilize this phenomenon in order to constrain the selection of head candidates. This technique has also been exploited in [9]. The relationship, shown in Fig. 3, gives the appropriate head sizes to search for given the disparities in the image.

We model this observation probabilistically as

$$d(h) = mh + b + w \tag{1}$$

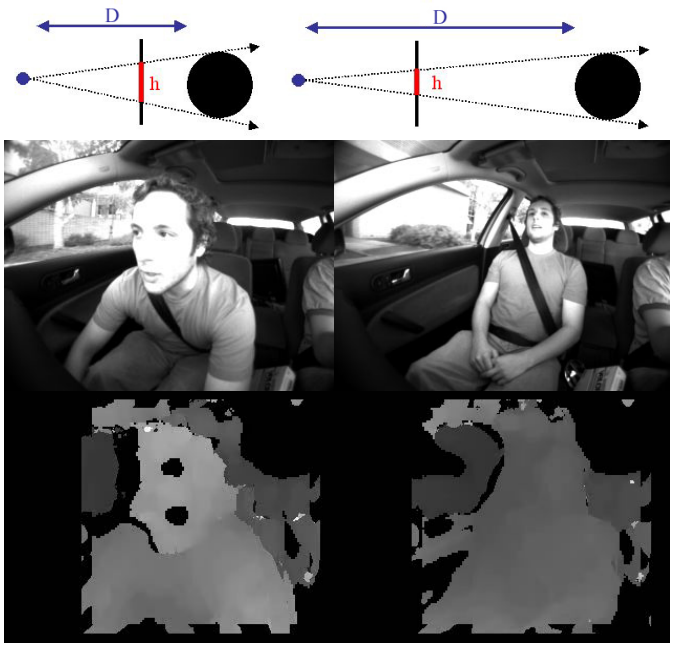

Fig. 3. Illustration of relative head size to disparity relationship: The vertical line is the image plane. The camera is denoted in small blue ball, and the head is denoted as the large black ball. The relative size on the head in the image plane is denoted in thick red bar along the image plane. Below each illustration is an example image and its corresponding disparity image.

where d is the disparity and h is the height of the head in pixels, and w is a random variable normally distributed with zero mean and σ standard deviation. The standard deviation is set at the distance between two disparities given the predefined number of head heights.

We estimate m and b using least squares by manually measuring the head size h and the average disparity d at four locations on the face (two eyes, mouth and forehead) of several medium-sized adults undergoing several head movements inside the cabin area. It is shown after extensive testing described below that this training on medium-sized adults is relatively universal and works well for a wide range of adult head sizes.

For our camera setup, the generalized best-fit line is

$$d(h) = \frac{255}{16p}(4.4366h - 73.825 + 16d_x) \tag{2}$$

where p is the maximum number of pixels over which the dense stereo algorithm will search, and the horopter d_x offsets the valid disparity (depth) range. The maximum number of disparities p ranges from 16 to 128 pixels, and the disparity $d(h)$ ranges from 0 to 255. In this case, the horopter was adjusted higher for closer depth range. The slope m and intercept b of the line is allowed to vary with the selection of appropriate p and d_x which define the valid head depths over which the stereo algorithm will actively search.

Note that (2) is specific for a given stereo pair but is invariant to camera location. Moving the occupant head with a fixed camera to change the perceived orientation can also be thought of as fixing the occupant's head position and changing the camera placement. If the algorithm can show robustness (and will below) to different head orientations and poses for

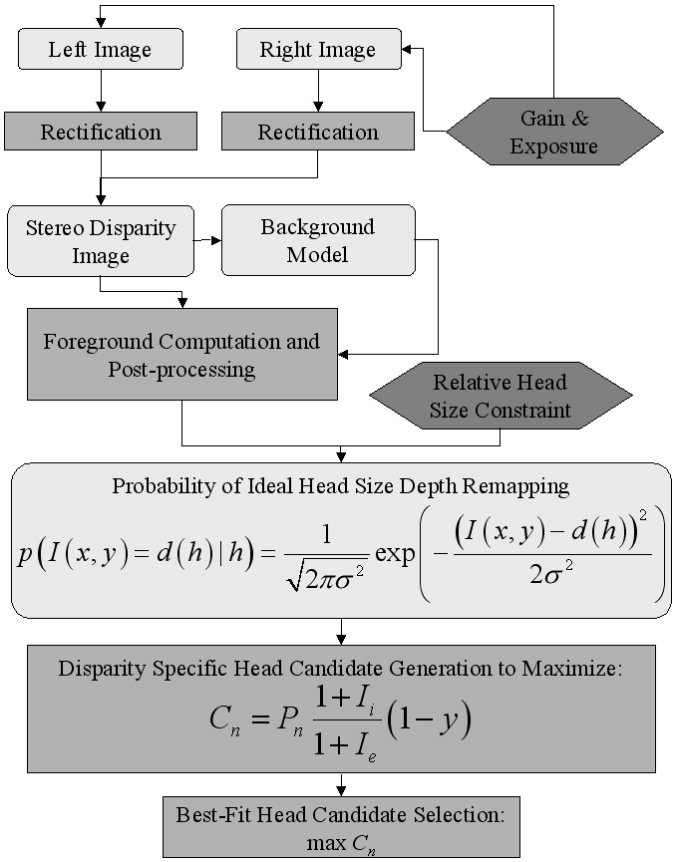

Fig. 4. Flowchart for Head Detection Algorithm.

a given camera position, it will also perform well for other camera positions as well.

VI. Head Detection Algorithm

The stereo-based head detection algorithm is outlined in the sections below. A flowchart for the algorithm is given in Fig. 4

A. Statistical Disparity Background Subtraction

The background model is computed from 300 frames of an empty passenger seat area. The frames are captured at various exposures and gain settings in order to expand the coverage of valid disparity regions in the scene. Based on work done in [13], the mean and variance are computed for each pixel in a recursive manner. Only frame pixels of non-zero disparities contribute to the background model.

Once the model is computed, the foreground disparity image can be found such that pixels in the current frame are considered foreground pixels if

$$|I(x,y) - \mu(x,y)| > k\sigma(x,y). \tag{3}$$

At the image location (x, y), $I(x, y)$ is the disparity of the current frame, $\mu(x, y)$ is the statistical mean of the background disparity image, and $\sigma(x, y)$ is the standard deviation of the background disparity image. The scale variable k allows us to vary the threshold. For a pixel to be considered part of the foreground, it must be k standard deviations away from the mean background value.

B. Stereo Computation and Post Processing

Given an input left and right image, lens distortion is compensated and the images are rectified. Area correlation is computed using the algorithm discussed in Section IV. The texture and left/right post-filtering are enabled and multi-scale correlation is performed to fill out larger objects in the scene.

Once the disparity image is computed, the current frame is subtracted from the background estimate and median filter, morphological opening and connected component analysis are all applied to remove smaller disparity blobs.

C. Probability of Imaged Head Size at Depth Re-mapping

A finite set of head sizes is chosen. For each head size, there is an ideal disparity value $d(h)$ found using (2). This disparity value is the mean of a Gaussian distribution used to re-map the intensities of the disparity image to values that correspond to the probability (proximity) that the surface projected onto the image is at the ideal disparity (depth) for a given head size. In other words, the resulting re-mapped image consist of depth slices with gradual depth boundaries. We refer to these slices as *probability of disparity images*.

$$P\big(I(x,y) = d(h)|h\big) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(I(x,y) - d(h))^2}{2\sigma^2}\right) \tag{4}$$

(4) is the distribution function used to create the re-mapped image for each head size. Example re-mappings are shown in Fig. 5. Notice how for a selected imaged head size, the head region has several high probability (close proximity) pixels forming a blob whose contour is in the shape of an oval, while the same head region has low probability (far proximity) pixels and less oval shaped contours at other depths. This will help ensure that the correct head size is found by reducing the weight of other head shaped objects at inconsistent depths.

D. Disparity Specific Head Candidate Generation

For each discrete head height and corresponding probability of disparity image, a best head candidate needs to be selected. To do so, we modify the approach presented in [14]. The input is the probability of disparity image, rather than the original disparity image. The intensity and probability of disparity images are down-sampled by 4, and edges are then found in the intensity image using Sobel operators. Edges that overlap valid probability of disparity image pixels are kept. Ellipses of varying eccentricities and angles are overlaid on the edge image anchored at the seed points. Only ellipses of the corresponding imaged head height are tried for each probability of disparity image. Unlike [10], this ensures that only ellipses of the right height and depth corresponding to head size are selected. The best candidate head location for each probability of disparity image is selected to maximize the likelihood function:

$$C_n = P_n \frac{1 + I_i}{1 + I_e}(1 - y) \tag{5}$$

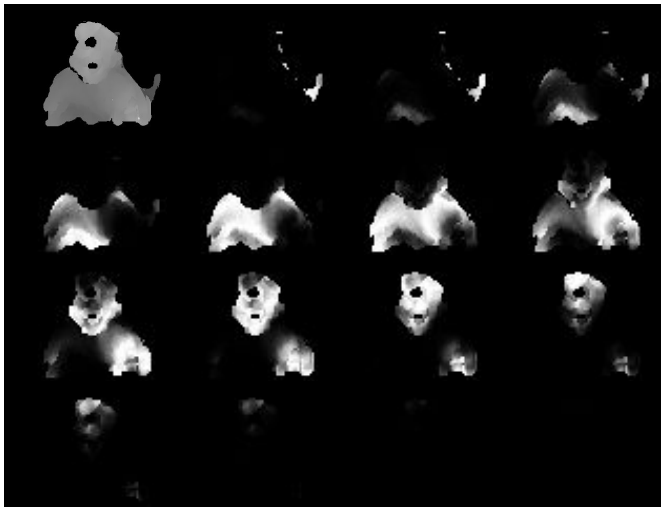

Fig. 5. Disparity re-mappings for increasing relative head sizes: The first image is the original foreground disparity image. The subsequent images show the probability of disparity images associated with various particular head sizes.

C_n is the likelihood of the ellipse at a given location, orientation and eccentricity in the n^{th} probability of disparity image. P_n is the average of all valid probability of disparity values within a 7×7 down-sampled region of the n^{th} probability of disparity image. I_i is the number of matches on the inner (main) ellipse template. I_e is the number of matches on the outer (surrounding the main) ellipse template. y is the normalized to 1 vertical position in the image. In general, the algorithm prefers elliptically shaped objects that have a high probability of having the correct disparity (of being at the correct depth) and are located higher in the imaged scene. For low P_n, no head candidate is selected.

The best overall head candidate from all the probability of disparity images is one that satisfies the equation

$$\hat{C} = \max_n C_n \tag{6}$$

VII. Performance Evaluation of Head Detection

Experiments were conducted in the LISA-P test-bed described in [10]. The LISA-P is a Volkswagen Passat equipped with a stereo camera pair and a Pentium 4 system to capture and process frames. The MEGA-DCS stereo camera system [15] is configured to capture stereo images at 30 fps at 320×240 resolution. The $\frac{1}{2}$ inch CMOS cameras are set with a 5 cm baseline and outfitted with 3.5 mm focal length lenses. Under these conditions, the nearest object that will produce valid disparities is approximately 15 cm. This minimum distance allows for a good amount of coverage of occupant movement inside the cabin. For these tests, the camera is affixed to the windshield below the rear-view mirror using a custom mount. Example captured images are shown in Fig. 3.

A. Experimental Setup

Four test subjects were selected so as to span the range of adult sizes. A 155 cm female and males 173, 178, and 185 cm were asked to perform a series of motions and tasks that were meant to test the robustness of the head detection system.

These tasks were as follows:

- Sit in normal position
- Lean halfway forward
- Lean completely forward
- Lean back against headrest
- Lean left
- Lean right
- Move clockwise about the cabin
- Move counterclockwise about the cabin
- Turn head to left and right
- Put left hand up next to head
- Put right hand up next to head
- Put both hands up next to head
- Lean forward with hands up by head
- Put left hand up in front of head
- Put right hand up in front of head
- Put both hands up in front of head
- Lean forward with hands in front of head
- Hold and read contents of box

The imaged head size to disparity constraint was trained using the middle two subjects. It is believed that head size is relatively constant across adult sizes and the constraint trained using the middle subjects should be sufficient for subjects at the extremes of the adult size ranges.

B. Ground-Truth Comparison

To show that the head detected location can be used to estimate the occupant's three-dimensional position, the results of the head detection algorithm is compared against ground-truth data acquired by detecting a visual light that the subject is asked to wear on their forehead. At alternating frames, the stereo pair captures images at normal exposure for the head detection algorithm and low exposure to isolate the light spot.

To determine the ground-truth location, the light spot is detected by searching for the brightest circular object in the scene. A threshold is applied to the low exposure image to eliminate illumination not originating from the light spot. The most round object is then found in the image by using the normalized covariances of the x and y blob pixel coordinates. Blobs with a large product of standard deviations $\sigma_x * \sigma_y$ tended to be circular blobs, and this criteria was used to filter out non-blobs. Once the light spot location is found, the 3D coordinates are determined by first computing the disparity on the light spot image using the stereo method described in Section VI. The X, Y, and Z values are then computed by averaging over a 7x7 window around the detected light spot point.

The head detected 3D point coordinates are also computed by averaging a 7x7 window around the center of the best-fit ellipse computed from (6).

These tests were conducted while the LISA-P was stationary to ensure that the light spot was easily and accurately detected. The subjects performed the test maneuvers listed in VII-A. The X, Y, and Z locations for the head detector and ground truth for each subject plotted in Fig. 6 - Fig. 9.

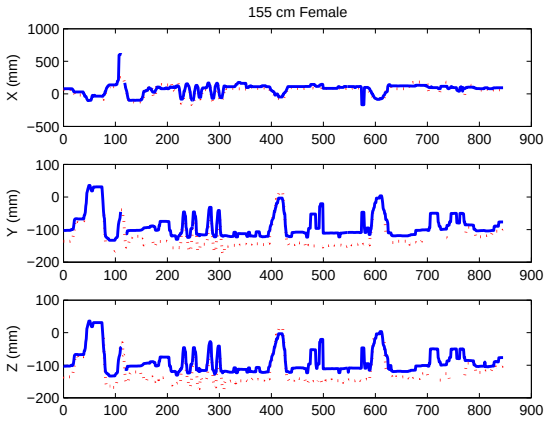

Fig. 6. Head Detection plotted with ground-truth for 155 cm female. For each subject, X, Y, and Z estimates are plotted. Ground-truth estimates are in dotted lines and Head detection estimates are in solid lines. Discontinuities in the ground-truth indicate frames when the light spot was not isolated.

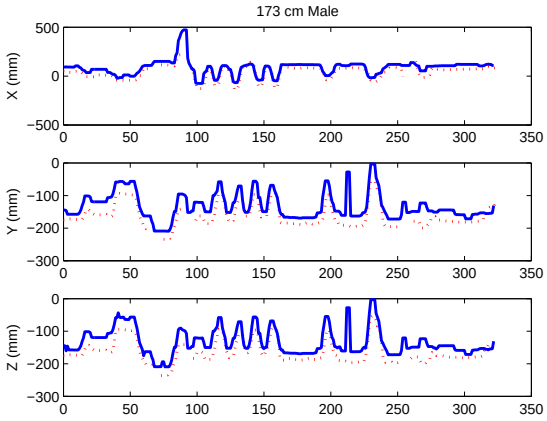

Fig. 7. Head Detection plotted with ground-truth for 173 cm male. For each subject, X, Y, and Z estimates are plotted. Ground-truth estimates are in dotted lines and Head detected estimates are in solid lines. Discontinuities in the ground-truth indicate frames when the light spot was not isolated.

Note that the head detection and ground truth follow each other very well. Large differences can be attributed to either head detection errors for a given frame, while smaller inaccuracies are the result of the light spot not being located at the center of the head. In general, the light spot will be located slightly higher than the center of an ellipse encapsulating the head. However, when the subject turns his head, the deviation from the detected center of the head is even greater.

C. Head Detection Experiment in a Moving Car

It is desired to also evaluate our head detection algorithm in the context of real driving situations. The same test detailed in VII-A was repeated with the same subjects in a moving car test. Throughout the test, the car was driven on local roads during a sunny afternoon. Head detection was considered correct if the center of the best ellipse point is located on the subject's head. The full results of the head detection algorithm are listed in Table I.

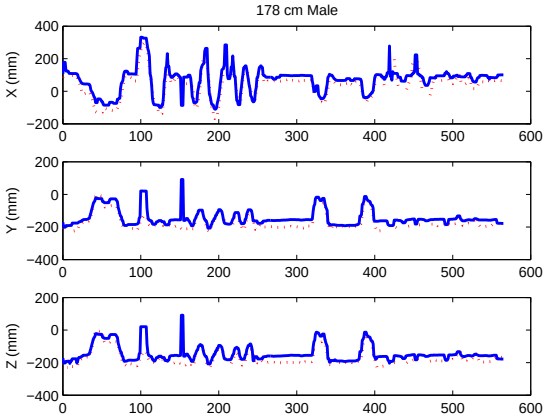

Fig. 8. Head Detection plotted with ground-truth for 178 cm male. For each subject, X, Y, and Z estimates are plotted. Ground-truth estimates are in dotted lines and Head detected estimates are in solid lines. Discontinuities in the ground-truth indicate frames when the light spot was not isolated.

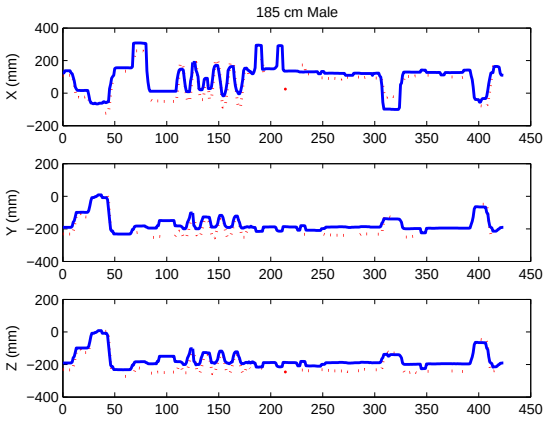

Fig. 9. Head Detection plotted with ground-truth for 185 cm male. For each subject, X, Y, and Z estimates are plotted. Ground-truth estimates are in dotted lines and Head detected estimates are in solid lines. Discontinuities in the ground-truth indicate frames when the light spot was not isolated.

TABLE I
HEAD DETECTION RESULTS

Subject	Number of Times Correctly Detected	Percent Correct
155 cm Female	1781 / 1910	93.3%
173 cm Male	1310 / 1394	94.0%
178 cm Male	1669 / 1836	91.0%
185 cm Male	1655 / 1753	94.4%

Fig. 10. Example of overexposed image due to harsh sunlight conditions is shown. The head detection algorithm will fail in cases like this.

TABLE II
HEAD DETECTION RESULTS OMITTING OVER-EXPOSURES

Subject	Number of Times Correctly Detected	Percent Correct
155 cm Female	1781 / 1812	98.3%
173 cm Male	1310 / 1328	98.6%
178 cm Male	1669 / 1722	96.9%
185 cm Male	1655 / 1703	97.2%

For these tests, 50-80% of the errors are due to situations where the sunlight was positioned directly in front of the camera, effectively washing out the entire scene. This effect happens so quickly and drastically that the camera auto gain does not properly adjust. Shown in Fig. 10, it is impossible to obtain much relevant information. The algorithmic failure in these cases is easy to explain, and it would be very difficult for many algorithms to perform in these conditions.

Since these errors are due to camera properties and position, it is valid to consider the head detection results with these frames omitted from consideration. If this is done, the results are improved as shown in Table II. Examples of successful head detection are shown in Fig. 11. Successful detections for more difficult examples including harsh lighting, partial occlusions and competing objects, such as hands, are shown in Fig. 12.

An experiment was also conducted in low-light night driving conditions. The 173 cm male subject performed the same poses outlined in Section VII-A under near-infrared illumination. The experimental results are listed in Table III. Although the overall illumination was small, the algorithm performed at the same level as during daylight conditions. Example results from the night experiment are shown in Fig. 13.

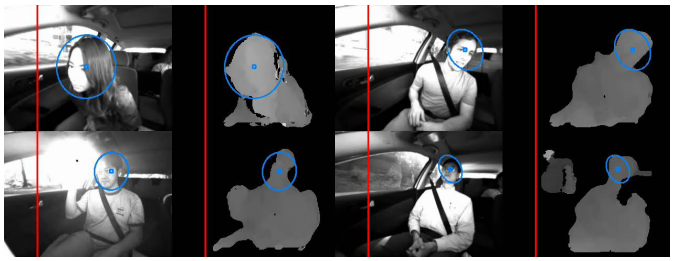

Fig. 11. Successful detection examples: The left stereo and the foreground disparity image are shown. The detected head location is shown with an ellipse.

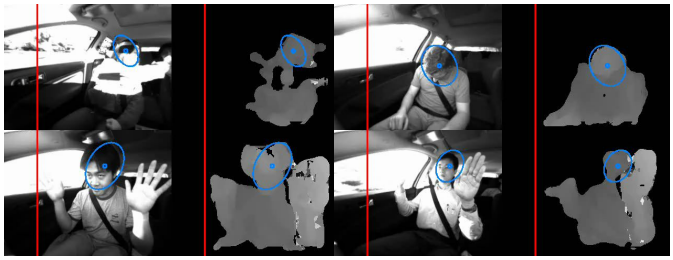

Fig. 12. Successful detection for difficult examples: The left stereo and the foreground disparity image are shown. The detected head location is shown with an ellipse. The examples show harsh lighting, partial occlusions as well as competing objects, such as hands.

TABLE III
HEAD DETECTION COMPARISON FOR DAY AND NIGHT TESTS

Subject	Number of Times Correctly Detected	Percent Correct
173 cm Male - Day	1310 / 1328	98.6%
173 cm Male - Night	625 / 632	98.9%

VIII. ERROR ANALYSIS AND DISCUSSION

A. Error Analysis

Head detection errors primarily occur for the following reasons: poor illumination, occlusion, and competing objects. An extreme example of poor illumination is shown in Fig. 10, but less harsh conditions can lead to head detection errors. Usually, when part of the scene where the face is located is overexposed, the face appears washed out. This reduces the texture on the face, invalidating the parts of the disparity image, making it difficult to find elliptical objects in that area using this method. Fig. 14 shows an example of this type of error.

Occlusion of the head area can also give incorrect results. In these situations, sometimes the subject's hands or an object would occlude the head. If the amount of occlusion was

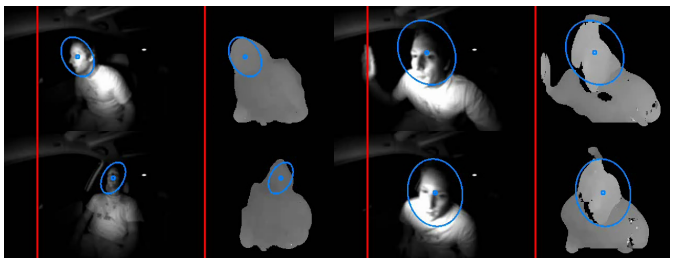

Fig. 13. Successful detection examples during night experiment with Near Infrared Illumination: The left stereo and the foreground disparity image are shown. The detected head location is shown with an ellipse.

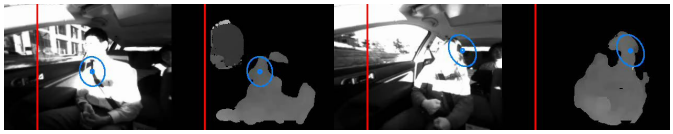

Fig. 14. Head detection errors due to harsh lighting: The left stereo and the foreground disparity image are shown. The detected head location is shown with an ellipse.

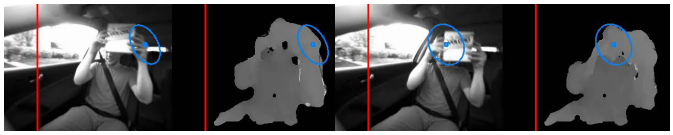

Fig. 15. Occlusion situations: The occlusion results in an incorrect detection in the first example, and a correct detection in the second one frame apart. The left stereo and the foreground disparity image are shown. The detected head location is shown with an ellipse.

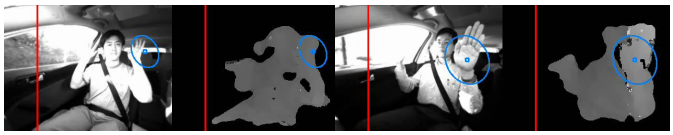

Fig. 16. Head detection errors due to competing objects: The left stereo and the foreground disparity image are shown. The detected head location is shown with an ellipse.

significant, the detected head location could be erroneous. However, the occlusions usually occur momentarily and head detection may still be correct for certain occlusions. An example of occlusion is shown in Fig. 15. The subsequently captured frames both show occlusion, but the head is detected incorrectly in the first and correctly again in the second.

Competing objects can also cause errors in head detection. For example, if the disparity associated with the occupant's hand for a given location and orientation has the same size, shape and disparity as an ideal head found in (2), it may be incorrectly selected as the best head location. This problem can be exacerbated by the fact that the hand may also partially occlude the desired face location. Such an example is shown in Fig. 16.

B. Discussion

In general, head detection errors occur in short bursts, at most 10 frames, normally over a 3 to 5 frames. For longer periods both preceding and following the error frames the head detection is correct. Clearly, tracking the head detection results could help alleviate these detection errors, since the short error bursts will not follow the same predicted track as the actual head location.

Tracking can also improve detection success and speed by predicting in which probability of disparity images the head is likely to be located. Currently, all images with some probability values are searched for head candidates. By tracking and predicting in which disparity range the head is likely to be, we can effectively reduce the head candidate search space. By incorporating this tracking constraint into the likelihood function, we can put less value on images that do not follow the predicted track. This may reduce the quality of candidates coming from competing objects and other incorrect locations.

IX. Conclusion

Development of a head detection algorithm that can perform robustly in outdoor and in-car situations requires that both the algorithmic input and head detection method be robust to lighting conditions and the uncontrolled environment associated with non-laboratory experiments. This stereo algorithm presented in this paper has shown robustness to both these conditions.

Modeling the relative head size and disparity allows a simple ellipse detector to accurately locate the head with very high accuracy. This method allows for a robust, real-time system that could be adapted to suit a variety of applications that require head detection.

Acknowledgment

This work was supported in part by a University of California Discovery Grant with Volkswagen Research Laboratory under the Digital Media Innovation Program. The authors gratefully acknowledge valuable inputs and assistance by Dr. Arne Stoschek and Dr. Cedric DuPont of VW Research Laboratory. We are also thankful for the assistance and support of our colleagues from the UCSD Computer Vision and Robotics Research Laboratory.

References

[1] Y. Owechko *et al.*, "Vision-based fusion system for smart airbag applications," in *IEEE Intelligent Vehicle Symposium*, vol. 1, 2002, pp. 245–250.

[2] J. Krumm and G. Kirk, "Video occupant detection for airbag deployment. Applications of computer vision," in *WACV'98*, 1998.

[3] R. Ryena *et al.*, "Head detection inside vehicles with a modified SVM for safer airbags," in *Proc. IEEE Intelligent Transportation System Conference*, 2001.

[4] M. Devy, A. Giralt, and A. Marin-Hernandez, "Detection and classification of passenger seat occupation using stereovision," in *Proc. IEEE Intelligent Vehicles Symposium*, 2000.

[5] Y. Luo, Y. L. Murphy, and F. Khairallah, "Human head detection using multi-modal object features," in *Proc. International Joint Conference on Neural Networks*, 2003.

[6] P. Faber, "Seat occupation detection inside vehicles. images analysis and interpretation," in *Proc.* 4^{th} *IEEE Southwest Symposium*, 2000.

[7] M. Fritzsche *et al.*, "Vehicle occupancy monitoring with optical range-sensors," in *Proc. IEEE Intelligent Vehicles Symposium*, 2004.

[8] B. Alefs *et al.*, "Robust occupancy detection from stereo images," in *Proc. IEEE Intelligent Transportation Systems Conference*, 2004.

[9] W. Huang and Y. Guo, "Real-time stereo tracking of multiple moving heads," in *Proc. IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, 2001.

[10] M. Trivedi, S. Y. Cheng, E. Childers, and S. Krotosky, "Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation," *IEEE Trans. Veh. Technol.*, vol. 53, no. 6, pp. 1968–1712, Nov. 2004.

[11] S. Guatama, S. Lacroix, and M.Devy, "Performance of stereo for occupant detection," in *Proc. of the International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, 1999.

[12] K. Konolige, "Small vision systems: hardware and implementation," in *Eighth International Symposium on Robotics Research*, 1997.

[13] T. Horprasert, D. Harwood, and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *Proc. IEEE ICCV Frame-Rate Workshop*, 1999.

[14] A. Eleftheriadis and A. Jacquin, "Face location detection for model-assisted rate control in H.261-compatible coding of video," *Signal Processing: Image Communication*, vol. 7, pp. 435–455, 1995.

[15] K. Konolige. (2004) STH-MDCS2-VAR/-C stereo head users manual. sthmdcs2-var.pdf. [Online]. Available: http://www.videredesign.com