Stereo-Assist: Top-down Stereo for Driver Assistance Systems

Gideon P. Stein, Yoram Gdalyahu and Amnon Shashua

Abstract—This paper presents a top-down approach to stereo for use in driver assistance systems. We introduce an asymmetric configuration where monocular object detection and range estimation is performed in the primary camera and then that image patch is aligned and matched in the secondary camera. The stereo distance measure from the matching assists in target verification and improved distance measurements. This approach, Stereo-Assist, shows significant advantages over the classical bottom-up stereo approach which relies on first computing a dense depth map and then using the depth map for object detection. The new approach can provide increased object detection range, reduced computational load, greater flexibility in camera configurations (we are no longer limited to side-by-side stereo configurations), greater robustness to obstructions in part of the image and mixed camera modalities FIR/VIS can be used. We show results with two novel configurations and illustrate how monocular object detection allows for simple online calibration of the stereo rig.

I. INTRODUCTION

Camera based driver assistance systems can be divided into two main types: monocular systems and stereo systems. The latter typically employ two symmetric cameras mounted side by side where epipolar lines are aligned with the horizontal image scan lines.

Due to the compact nature, simplicity in hardware and lower cost, driving assistance applications using a monocular camera system are gaining traction. Recent years have witnessed product launches of Lane Departure Warning (LDW), Adaptive High-Beam Assist (AHB) and Traffic Sign Recognition (TSR) as a bundle driven by a monocular camera (available in BMW 7 Series, Mercedes E-class and Audi A8 among others). LDW and Monocular vehicle detection for Forward Collision Warning (FCW) [1], [4], [5] are bundled in the Mobileye AWS. Fusion between radar and a monocular camera is used in applications such as LDW and Collision Mitigation by Braking (CMBB) on licensed vehicles (Volvo “Collision Avoidance Package”) and on Pedestrians [2]. The NHTSA safety ratings for model year 2011 include LDW and FCW [3] and underscore the need to integrate both customer functions into a single monocular camera system in order to save costs.

Taken together, one can identify three important trends: (i) monocular camera systems are in high demand and entering into the domain of (model-based) object detection (Vehicles for certain with possibility for Pedestrians as well), (ii) as a result, there is a strong drive to push the performance envelope of a monocular camera system in order to maximize the return on investment, (iii) monocular camera systems are quite mature. However, monocular systems have no direct method for measuring distance so information about the object class and/or context, such as the road plane, must be used.

Depth can be computed directly using multiple cameras. In the classic stereo approach, a dense disparity map is used to create a 3D map of the environment. This 3D representation is then used for foreground/background segmentation (e.g., for finding candidate regions for further processing), for triggering object detection processing and for estimating range and range-rate to detected objects. The classical stereo approach is a low-level pixel-based process and, in principle, can handle arbitrary shapes without modeling the object class beforehand. By contrast, monocular systems use pattern recognition for detection of a specific object class prior to monocular range estimation.

The classical stereo approach works well for close targets and in good weather conditions. However, the utility drops rapidly for longer distances and inclement weather. Furthermore, the effective range increases only by the square root of the camera resolution (see Sec. III-C for details) and thus, will not increase significantly with the introduction of the newer 1M pixel automotive sensors entering the market. Finally, it is not clear how truly useful the depth map is in cluttered scenes (see Fig. 1 and Fig. 2). The images themselves hold much more information than the depth map for segmenting out objects (and object parts) in the presence of clutter. Moreover, in an automotive setting, packaging a stereo rig with light-shield and windscreen wiper protection is not a simple feat. It adds to the cost of introducing a stereo system, and thereby is a hindrance for high-volume adoption.

Instead of having depth estimated at an early stage throughout the image (classical stereo), we propose to invert the process on its head. Monocular camera systems are ma-
Stereo-Assist and pedestrians. In Stereo-Assist, targets (and parts thereof) are detected in the monocular system and target distance is computed. This is the primary camera responsible also for all non-ranging applications (like LDW, IHC, TSR). The targets are then matched with a secondary camera image adding stereo-depth on the selected targets. (For an example see Fig. 2.)

This asymmetric configuration, which we call Stereo Assist, combines stereo-depth and monocular depth together thus maximizing the strengths from each modality. It can be viewed as a way of bolstering the performance envelope of a monocular system (whatever that envelope limit may be). In particular, more accurate distance information can be obtained and the stereo distance can also be used as an extra cue for target verification. Stereo Assist, on the other hand, can be viewed as an extension of classical stereo to handle high-level primitives (rather than pixel level measurements) covering the spectrum from complete model-based objects to object parts common to man-made objects in general. The move towards high-level primitives, thereby a stronger emphasis on “mono” abilities, endows the stereo system with increased robustness against clutter.

Following a review of current stereo and monocular systems in Section II, we discuss in more detail the rationale for Stereo-Assist in Section III. Section IV shows how the results of the monocular object detection can be used for online calibration of the stereo rig. Stereo-Assist can work well with a classic side-by-side stereo configuration, but to demonstrate the flexibility of the concept, Section VI describes experiments and results from two novel configurations. Section VII will discuss the results and suggest other possible camera configurations.

II. BACKGROUND

An example of the classic stereo approach is given in [8] which presents both sparse and dense disparity maps, and their use in both foreground/background segmentation and to trigger object detection. Effective results are presented for vehicle and pedestrian targets up to 20m using a 32deg FOV lens and a 30cm baseline. A simple epipolar geometry is used where the epipolar lines are parallel to the image rows. This is done for computational efficiency and is common to virtually all stereo systems. Since it is difficult to mount the cameras accurately enough, image rectification is performed. Schick [9] describes using cameras with different optics but still the images are warped to the simple epipolar geometry.

Detection range was improved using higher resolution, narrower lenses and wider baselines (see [10]). [11] pushes the limits using 35mm lenses and a baseline of 1.0m to detect targets with heights as small as 20cm at 100m range. Calibration becomes critical at these ranges and the solution was found in alignment of the road surface and then determining points that are above the road — an approach that has become known as v-disparity [10]. Suganuma [12] uses v-disparity and clustering to impressively segment a vehicle from a row of trees targets, one lane to the side, at a range of 30m using a only narrow 12cm baseline and a 30deg FOV. Vehicle targets at 70m however get blended together.

The disparity map is a computational challenge. Mark and Gavrila [13] give a review of the faster methods which are based on SSD and SAD computation and then investigate the various methods for choosing window size, smoothing and edge preserving. Their performance analysis is focused on intelligent vehicle (IV) applications including analysis of the effects of imager noise and alignment error. They do not, however, explore the effects of rain and lighting conditions. Hirschmuller [14] surveys more accurate methods which are more suitable to offline computation and presents a mutual information based cost function and an approximation to global matching (SGM). Technology has progressed and [15] reports a real-time SGM implementation on special hardware.
Subpixel disparity maps can be obtained by fitting the cost function to a parabola over local disparity perturbations and finding the maximum/minimum analytically. This topic is surveyed by [16] and results are shown for 3D reconstruction of buildings and for road scenes. A 26deg FOV and a 55cm baseline is used and gives improved shape at 40m in good weather. Subpixel disparity accuracy has not been demonstrated for inclement weather conditions.

Munder et.al. [17] use the depth from a dense stereo depth map as one input to the detection system. Keller et.al [15] use the dense stereo depth map to determine road geometry and in particular improve the ROI detection for possible pedestrian candidates.

Chiu et.al. [18] break from tradition and explicitly look for vehicle candidates in the grayscale images using vertical and horizontal edges maps before computing the disparity for the candidates patches. This significantly reduces the computational load. They limit themselves to simple epipolar geometry.

Monocular detection of specific object classes for driving assistance has been demonstrated successfully. Traffic recognition is already in serial production. Vehicle detection was presented in [1], [19] with serial production performance [4], [5] and pedestrian detection in [20], [21] also in serial production [2]. Monocular camera based vehicle and pedestrian detection technology is available as a black box. The experiments in this paper use such such a system for the primary camera.

III. WHY USE STEREO-ASSIST?

The advantages of the Stereo-Assist concept over classic stereo, can be grouped into several categories: (i) computational efficiency, (ii) increased robustness, (iii) increased effective range, (iv) flexibility in camera configurations, and (v) modularity. We expand on some of those categories in more details below.

A. Computational efficiency:

In classic stereo, a dense depth map is computed which requires significant computational resources. In order to make the computation more efficient, simple similarity scores such as SSD and SAD are used in most real-time systems and the epipolar geometry is reduced to a simple configuration where disparity is calculated by matching along image rows.

B. Robust Matching and Disparity Estimation

In Stereo-Assist, matching is performed only on a limited number of patches thereby allowing for matching techniques which are both more advanced and more computationally expensive, per image patch, than those affordable when computing a dense depth map. In this paper we use normalized correlation. This measure is robust to lighting variations and to partial occlusions of the target [22]. Alternatively one could consider measures such as Hausdorff distance [23] and mutual information [24].

One example of the increased robustness is shown in figure 3. Figure 3a and figure 3b show the images from the primary and secondary cameras respectively. In this frame, some moisture blocks part of the secondary camera image while the primary image is clear. This situation is quite common unless the windshield wipers are designed to clear the windshield in front of both cameras simultaneously. Figure 3c shows the three vehicle targets detected in the primary camera and aligned in the secondary camera. The alignment is good even though much of the targets in the secondary image is distorted by water.

This highlights a significant advantage of the Stereo-Assist system. It is quite tolerant to image quality issues in the secondary camera. In figure 2 the secondary camera is mounted exposed and away from the glass. Significant reflections appear from the dashboard. These reflections however, did not affect the overall matching of the objects.

C. Greater detection range

Consider the example of a classic stereo camera pair where each camera is of VGA resolution and with horizontal FOV
A. Calibration using targets tracked to infinity

We use targets that are detected when they are close and then tracked till they are under 8 pixels in the image. The calibration has four steps:

1) Calibration of relative rotation around optical axis Z: The targets, cropped from the primary image, are aligned with the targets in the secondary image allowing for planar translation, rotation and scaling. The larger (closer) targets are used to determine the relative rotation around the optical axis (Z axis) - see figure 5.

2) Calibration of the relative rotation around the X and Y axes: The distant targets are used as ‘points at infinity’ (Fig. 5c). This targets are vehicles (at least 1.5m wide) and thus must be over 150m in distance. The angles are small and we can simply use the x and y alignment values of the distant targets as an offset and not compute a true image rotation².

E. Flexible configurations and tolerance to calibration errors

Since no dense depth map is required, there is no need to perform image rectification. Not only does this reduce the computational need for image warping, it also reduces the adverse effects of rectification in the following cases: (i) wide angle lenses with significant distortion, and (ii) in the case where one camera is significantly in front of the other thereby having the rectified images no longer facing straight ahead. These complex geometries are made possible by allowing image translation, rotation and scale in the alignment of the small number of patches.

Moreover, as we match entire objects, which inherently have edges and corners in a variety of orientations, we are no longer limited, by the aperture problem, to search precisely along epipolar lines. The result is a system which to a large extent is tolerant to errors in epipolar geometry estimation.

IV. CALIBRATION

Calibration of a stereo rig involves calibration of internal and external camera parameters. In our case the focal length is known (within 1%). Furthermore, the position and orientation of the primary camera relative to the vehicle is known. What is left is to determine is the relative position and orientation of the two cameras. This is known as the epipolar geometry. Various methods have been proposed using calibration targets or point correspondences [25].

The online calibration method presented does not require specialized calibration targets. Instead, is bootstraps off the object detection available from the monocular vision system. Vehicle targets are used for calibration (not pedestrians or motorcycles) since they are wide and have strong texture edges. Two variants of this concept are described. We will assume that the relative orientation angles are a few degrees at most and thus small angle approximations can be used.

D. Detection in cluttered scenes

In "cluttered" scenes the foreground and background are at the same depth and the depth map, which is very expensive to compute, is not really useful in detecting the objects of interest. Some examples are shown in Fig. 1: pedestrians near rows of parked cars and people stepping out of a car or truck.

Figure 4b shows an image captured of a pedestrian in front of a store. The distance to the pedestrian is approximately 33m. If we assume a disparity map accuracy of 1 pixel[1], eqn. 2 implies that the pedestrian cannot be detected by stereo unless he/she is at least 3m from the wall.

Monocular pedestrian range in daylight is 50m [2] with a VGA sensor (f = 950 pixels) and this range increases linearly with increased resolution (either by narrowing the FOV or by moving to a 1M pixel sensor). For target verification we would like a minimum disparity of 4 pixels and that translates to a verification range of 72m. At those ranges, the distance estimate from stereo is only about 25% accurate and no better than the monocular distance. Thus the pedestrian in figure 4b can be detected by the monocular system and then verified by the Stereo-Assist algorithm for higher confidence.

1Although sub-pixel accuracies have been reported, those were limited to good weather conditions.

2This approximation can lead to a few pixels of error at the edges of the image. For accurate depth measurements further calibration refinement is required but this is beyond the scope of this paper.
3) Determining the epipolar lines: The \( x \) and \( y \) alignment values of the targets at closer range are used to determine the relative displacement in \( X \) and \( Y \) of the secondary camera. The scale gives an estimate of the displacement in \( Z \) as described in section IV-B.

4) The baseline: Often the absolute value of the baseline is known by design. If not, the relative baseline can be estimated from the monocular distance estimates as shown in the next section.

### B. Calibration using monocular distance estimates

In an urban environment distant targets might be rare. For such a setting, a different calibration method was developed, which does not require targets at infinity. This method uses the range measurements from the monocular system and the stereo matches to determine the epipolar geometry and stereo baseline.

Determine the relative rotation around the \( Z \) axis and correct the secondary image:

1) Targets, detected by the primary camera are matched to targets in the secondary camera using a coarse-to-fine search over a wide range of translations in \( x \) and \( y \), rotations and scales. To speed up the search, best matches are first found using only translation and only for even disparities over a wide disparity range (± 40 pixels). This was followed by a refinement stage which included a search in ±4 pixels in each direction and a rotation over a ±0.1 rad range.

2) As in the previous method, the final \( Z \) rotation estimate is the median rotation of all the targets that are larger than 20 pixels.

3) The secondary images are all corrected for \( Z \) rotation and then the alignment repeated using only image translation and scale.

In practice this last step is a refinement step with translation in a range of ±4 pixels in each direction and scale ranging from 0.95 to 1.05 in 0.01 increments. These translation and scale values are used to calibrate for the relative rotation around the \( X \) and \( Y \) camera axes and the relative translation in \( X \), \( Y \) and \( Z \).

We make the following assumptions and approximations:

1) The relative rotation around the \( Z \) axis (roll) has been computed and the secondary image has been corrected.

2) The relative rotation is small and can be approximated by an image shift. It can then be coupled in with the unknown principal points of the two cameras.

3) The effects of scale \( S \) due to the \( \delta Z \) relative translation can be decoupled from the image \( \delta x, \delta y \) displacement (disparity) due to the camera \( \delta X \) and \( \delta Y \) relative translations.

These leads to the following relationships:

\[
\delta x = \frac{f \delta X}{Z} + x_0 \tag{3}
\]

\[
\delta y = \frac{f \delta Y}{Z} + y_0 \tag{4}
\]

where \( f \) is the known camera focal length and the values \( x_0 \) and \( y_0 \) combine the relative orientation and the unknown differences in the principal points. The pairs of values of \( (\delta X, x_0) \) and \( (\delta Y, y_0) \) can be determined using a best line fit to (3) and (4) respectively.

If there is also relative camera displacement in the \( Z \) direction, there will some scaling required of the corresponding image patch where the scale factor depends on the distance to the object:

\[
S = \frac{Z}{Z - \delta Z} \tag{5}
\]

The scale \( S \) is in most cases too small a signal to be used for measuring the distance \( Z \). However, it cannot be neglected in the model as it can add a few pixels error to the disparity values \( \delta x \) and \( \delta y \). Equation 5 can also be written as:

\[
1 - S^{-1} = \frac{\delta Z}{Z} \tag{6}
\]

Dividing both sides by the dominant \( x \) or \( y \) disparity (for example \( \delta y \)) gives:

\[
\frac{1 - S^{-1}}{\delta y} = \frac{\delta Z}{Z} = \frac{\delta Z}{f \delta Y} \tag{7}
\]

where we have already corrected for the calibrated offset value \( y_0 \). The ratio \( \frac{1 - S^{-1}}{\delta y} \) is thus fixed for a particular relative camera geometry. This ratio is computed for all the targets and the median value is taken as the calibration value. Using this ratio one can perform a 1D stereo search over a range of disparities \( \delta y \) and for each disparity compute the appropriate scale factor.

### V. TOWARDS GENERAL OBJECT DETECTION

We have described the Stereo-Assist in the context model-based detections where the Primary camera is responsible for detecting generic object classes of vehicles, traffic signs and pedestrians. It is possible to extend the envelope and include object parts in the vocabulary of high-level entities
extracted by the Primary camera. Most objects of interest, including partially occluded pedestrians and man-made objects in general are rich with structured visual cues where the most notable ones are vertical line segments. Given texture analysis processes for extracting roadway segments (cf. [7]), it is possible to cluster vertical line segments based on lateral proximity, height from the roadway, motion and monocular depth cues into candidates of "meaningful structure parts" which are fed into the Stereo-Assist direct depth measurement process for purpose of target validation.

The possibility of obtaining direct depth measurements using stereopsis is crucial for making the process robust because otherwise accidental line segments, such as those arising from spurious shadows, can "collude" to form false positives. Those false positives can be removed using direct depth measurements or through model-based classification. Further details are outside the scope of this paper but Fig. 6 illustrates the main steps: (i) the roadway extraction, (ii) vertical line segments, and (iii) detected objects which include a partially occluded pedestrian and a sign-post further ahead.

Finally, the concept can be further extended to focus on all non-road texture clusters (not necessarily generated by vertical line segments). This case is not far from classical bottom-up stereo with some regions masked off, thus underscoring a fundamental principle: Stereo-Assist and classical bottom-up stereo lie on a spectrum defined by the degree of generality of the candidates sent to the Secondary camera for depth measurement. In classical stereo those candidates are single pixels whereas with Stereo-Assist those are high-level clusters defined by model-based classification and general texture extracts common to the type of objects of interest found in the driver assistance domain.

VI. EXPERIMENTS WITH TWO NOVEL CONFIGURATIONS

Stereo-Assist can work well with a standard symmetric stereo camera configuration. However this paper presents more novel camera arrangements in order to highlight the flexibility of the approach. The test vehicle sued (Ford Focus, MY 2010) already had a monocular object detection system for vehicles and pedestrians [2], [4] located just below the rear-view mirror. This was used as the primary camera. We chose two locations for the Secondary camera (see Fig. 7).

1) A diagonal configuration where the secondary camera head is mounted on the left tip of the rear-view mirror. The lateral displacement of the secondary camera (i.e. the stereo baseline) is half the rear-view mirror width or in this case 14cm. The camera is a few centimeters back from the windshield and there is no protection from reflections from the dashboard.

2) A vertical configuration where the Secondary camera was placed below the Primary camera approximately 13cm down and 26cm along the windshield (with a rake angle of 30deg). This configuration is attractive from a production standpoint as both cameras can fit into conventional standard mono packages.

The cameras provided with the units used a Micron MT9V024 CMOS sensor (VGA resolution) and a 5.7mm lens providing a horizontal FOV of 40deg. Camera mounting height of the primary camera was 1.2m.

A. Data collection

Data capture was performed using the object detection system. The system performs real-time vehicle and pedestrian detection and provides as output target data (image location bounding box and range estimate) and the images themselves. The bounding box and range data from the primary camera was used as targets to be matched between the stereo images.

Weather conditions during experiments included daytime winter conditions, i.e., overcast skies, on and off precipitation of light snow/rain mixture, wet roads after rain.

B. Calibration results using monocular distance estimates

The second data sequence, the vertical camera configuration, involved driving around a suburban shopping and parking area. In such a setting, no target vehicles could be tracked to beyond 100m and we used the second calibration method which does not require targets at infinity.
median value was $-0.0015$. Using (7) this gives:

$$
\delta Z = -0.0015 \times f \delta y = 0.22m
$$

### C. Range results with Vertical Configuration

Fig. 11 shows the range estimates for 20 targets (sample is shown in Fig. 8). One can see the robustness of the stereo range even in the presence of rain on the windshield. Targets (9,10), (11,12) and (13,14) are of two parked vehicles at almost the same distance. The range estimates for both vehicles is almost the same even though the dark vehicle on the left is obscured by rain in the secondary image.

After calibration, the search space is reduced to a 1D search along $\delta y$ disparity values with appropriate scaling used for each for $\delta y$ value. With pedestrians there is a strong vertical edge component. So that the vertical edges did not have any influence on the $\delta y$ search we allowed for a small ($\pm 1$ pixel) perturbation in the $\delta x$ positions. Figure 12 shows the results of Stereo-Assist on some pedestrian examples. Table I shows the distance results estimated (1) Manually assuming pedestrian height of 1.75m (2) Monocular System (c) Stereo-Assist.

<table>
<thead>
<tr>
<th></th>
<th>(b)</th>
<th>(c)</th>
<th>(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>20</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>Monocular</td>
<td>23</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Stereo Assist</td>
<td>24</td>
<td>26</td>
<td>26</td>
</tr>
</tbody>
</table>

### TABLE I

Distance estimates for the pedestrian targets in figure 12 calculated using three methods.
VII. SUMMARY

We have presented Stereo-Assist, a new approach to stereo for driver assistance systems. In the StereAssist concept the system fuses information from both Monocular and Stereo by having a Primary camera responsible for target attention and selection and range estimation and a Secondary camera providing stereo-range on selected targets for purposes of target verification and range-fusion from both mono and stereo sources. The system combines the strengths from both sources and improves the performance envelop of of both monocular systems and classic, bottom-up stereo systems. Furthermore, it is to a large extent immune to errors in stereo depth, and image quality issues that arise from a Secondary camera unprotected from reflections and rain. The Stereo-Assist supports packaging configurations with a higher degree of flexibility than classical stereo, as for example the novel configurations addressed in Sec. VI.

As production monocular camera systems proliferate in volume and customer functions, more innovation and resources are invested in extending the envelope of performance and variety of customer functions. The Stereo-Assist concept allows disparity to be added as a modular element to an existing (monocular) system, thereby, providing an evolutionary growth of computer-vision capabilities in the automotive market — as opposed to a ”zero sum game” between monocular and stereo as to which technology will gain dominance.

REFERENCES

[10] R. Labayrade, D. Aubert and J.P. Tarel, “Real time obstacle detection in stereovision on non flat road geometry through v-disparity representation”, in Intelligent Vehicle Symposium, June 2002

Fig. 12. Example of pedestrian targets detected and alignment results. Only the Primary camera image is shown. The secondary camera is mounted at the bottom of the windshield, below and in front of the primary camera so vertical alignment of the targets gives the range.