A Comparison of Low-Cost Monocular Vision Techniques for Pothole Distance Estimation

S Nienaber*, RS Kroon**, MJ Booysen*
*Dept. of Electrical and Electronic Engineering, Stellenbosch University, Stellenbosch, South Africa.
**Computer Science Division, Stellenbosch University, Stellenbosch, South Africa. Centre for AI Research, CSIR Meraka
Email:{16082508, kroon, mjbooysen}@sun.ac.za

***Abstract*—Consider a single camera mounted on the inside of a vehicle's windscreen used for detecting potholes and other obstacles on the road surface. This paper outlines three approaches to the depth estimation problem of determining the distance to these obstacles in the range of 5 m to 30 m. We provide an empirical evaluation of the accuracy of these approaches under various conditions, and make recommendations for when each approach is most suitable. The approaches are based on the pinhole camera model: the simplest approach is based on the geometry of similar triangles, another employs the cross-ratio of a set of collinear points, and the final approach relies on calibration of the camera matrix. We recommend the use of the cross ratio approach for a fixed camera setup and depth estimation almost directly ahead, and an approach using similar triangles when predicting distances at wide angles or adjusting the camera height may be necessary.**

Keywords—depth estimation; monocular vision; projective geometry; cross ratio

I. INTRODUCTION

Depth estimation is not a new problem and many approaches have been developed for this task. These techniques employ various assumptions, so the appropriate depth estimation technique depends on the application domain. Major factors determining the applicability of an approach for a domain are the financial and computational cost of implementing it. This paper considers the problem of rapidly determining the distance from a vehicle to a pothole by depth estimation in high-resolution images. These images are generated by a windshield-mounted camera in a vehicle travelling at roughly 40 km/h - an example is given in Figure 1. Ideally, the solution should be suitable for deployment with the camera for real-time depth estimation of potholes and other obstacles.

In previous research, the authors developed a classifier for detecting potholes in such images [1]. As none existed, the authors have compiled and made publicly available a set of annotated images with potholes taken from a driver's vantage point [2]. Detecting distant potholes is considerably more challenging than detecting nearby potholes. The initial motivation for developing the depth estimation techniques described here was attempting to gain a better understanding of this aspect: by determining the distance of each pothole in set of images, it is possible to refine the results based on these distances. This allows one to quantify the performance degradation of the classifier as the depth increases, and to verify whether the degradation is monotone. In addition, this depth information can be used to construct classifiers for potholes at different ranges - it may be possible to combine these to improve overall pothole detection performance.

Fig. 1. Example of potholes in a road

For the work presented in this paper, the following factors were relevant. Potholes occur in countries and states that are subject to heavy rainfall, poor drainage, harsh winter weather and frequent freezing and thawing of road surfaces. However, in less affluent countries, potholes are usually not repaired in a timely fashion and therefore it was important to identify a cost-effective approach for depth estimation. This limitation prevents the use of multiple cameras or expensive laser mapping technology. Therefore, stereo vision techniques fall outside the scope of this paper. Additionally, the lack of laser mapping for establishing ground truth depths for objects necessitated the development of a low-cost but accurate approach to validate our depth estimation results. Finally, it was important to find a solution that could potentially be used in a real-time environment allowing a full real-time pothole detection and depth estimation solution. Approaches that enable real-time depth estimation within a road are also applicable to other vehicle automation tasks, such as incorporating rear-end collision avoidance into cruise control applications.

A. Contribution

Although there are existing approaches that perform distance estimation with monocular vision for various applications, there is limited work done to compare the

978-1-4799-7560-0/15/$31

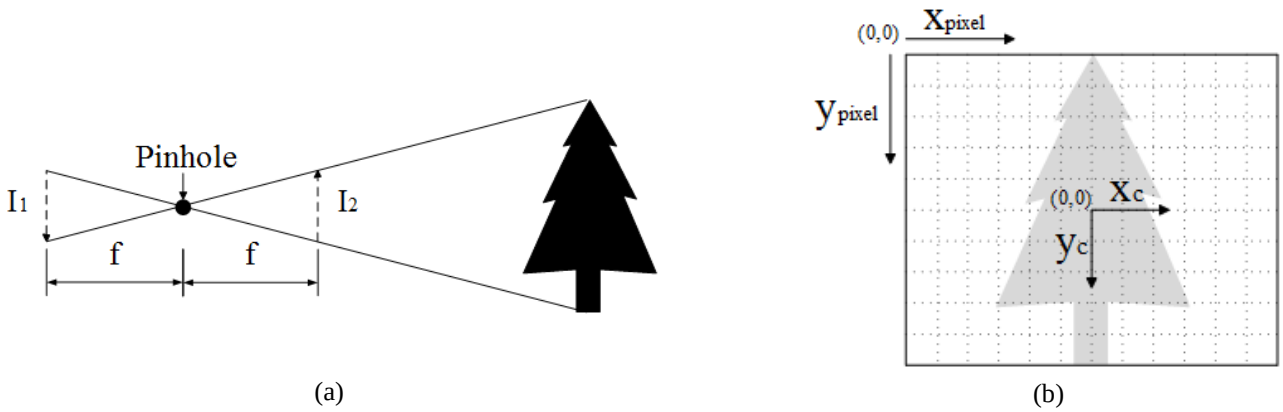

Fig. 2 Illustration of (a) the pinhole camera model and (b) the corresponding camera plane and image co-ordinate systems. I_2 represents the image plane, and the focal length f is the distance between the pinhole and the image length.

accuracy of the approaches. This is especially the case for the specific application considered in this paper, namely estimating the distance of objects in the road. The research contribution of this paper is the empirical comparison of the accuracy of three fundamental computer vision approaches to estimating the distance of potholes using monocular vision. Note that by the nature of the experiments, the results apply equally to other tasks involving depth estimation of markings or objects on a ground plane, where the relevant camera setup conditions are specified, and the distances involved are comparable (i.e. 5-30m).

II. Related Work

The depth estimation domain presented in this paper is novel, but depth estimation itself is widely applicable in practice. In robotics, it is crucial for a robot to determine its distance from obstacles in its path [2]. Another application pertains to forensics where only a single stationary camera is available and it is necessary to determine at what distance a person of interest was at a particular time or the height of a particular person in the image is required [3]. Motion capture has also become part of many computer games and films, and by using triangulation it is possible to determine the distance of actors and their limbs relative to other actors [4].

Depth estimation approaches are generally either active or passive. Active techniques generally analyse reflections of sound or light waves emitted into the scene to estimate depth [5]. An example of this is given in [2] where a laser light and camera are combined in such a manner that the camera can detect the laser light reflecting from obstacles. The camera was used to capture images at regular intervals and determine the centre point of the laser light pointing at an obstacle. The perceived shift of the laser light relative to the movement of the robot was then used to estimate depth with good accuracy. However, active approaches rely on the availability of additional equipment, which we aim to avoid.

Passive techniques perform depth estimation by analysing one or more captured images of a scene. Techniques have been developed for estimating depth from video feeds, estimating positional information from multiple images of the same scene (most commonly for stereo vision applications), and monocular depth estimation, where only a single image is used for depth estimation. The classical approach in all these cases involve trigonometric and (typically projective) geometric computations, incorporating whatever additional information is available from the particular domain. For the task we consider, only one camera is used, and using video feeds would only be feasible if an object tracking system were available for the potholes. Thus, our focus is on monocular vision approaches.

Two papers are of particular interest with respect to depth estimation of objects in a road using monocular techniques. In [6], a single camera was used to estimate the depth of a vehicle on an (assumed) planar road surface. The approach relies on the assumption that the camera's optical axis is parallel to the road's surface and edges. A geometric approach is applied that combines the pinhole model and ad-hoc error corrections presumably obtained from empirical measurements. An accuracy of 96% was reported over a distance of 70 m, but the methodology for obtaining the corrections is not presented, so that it is unclear what must be done to obtain this accuracy.

In [7], monocular depth estimation of vehicles on the road was also tackled using a pinhole camera model. In particular, the depth of the point where the vehicle meets the road was estimated. As in [6,8], a planar road surface and a parallel installation of the camera's optical axis to the road surface and edges are assumed. The work also included a method for determining the range rate that was based on scale change.

As an alternative to the classical geometry-based depth estimation techniques discussed above, it is worth noting that work has also been done on using probabilistic modelling of scenes and the structure of images for extracting depth estimation from single images. The most notable approaches in this direction make use of Markov random fields; e.g. [8,9,10]. These techniques are very versatile, generating range images for scenes appropriate to the situations they model without any major assumptions about the camera setup. These techniques are less appropriate for our use for two reasons. First, a suitable probabilistic model of a road and its contents is needed to apply the model; second, even with a suitable model, the accuracy of these is generally not as accurate as the geometric techniques discussed previously, so the latter should generally

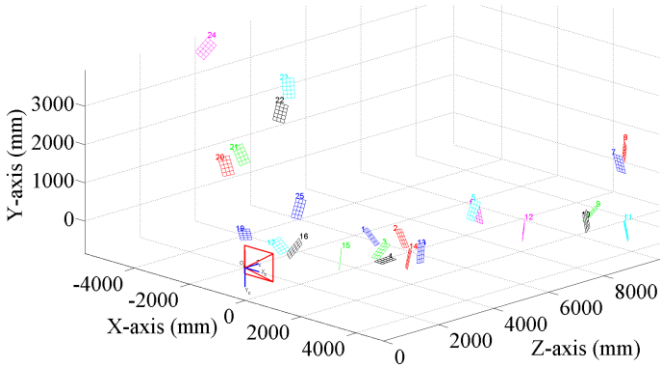

Fig. 3 Approximate board locations used for camera calibration. The camera is at the origin.

be applied when the relevant information for applying the geometric techniques is not available.

III. TECHNICAL BACKGROUND

This section discusses aspects of the pinhole camera model and camera calibration process relevant to the rest of our work.

A. *Pinhole camera model*

For the approaches used in this paper, the camera is modelled using the pinhole camera model [11] as illustrated in Figure 2.a, and only rays of light coming through the pinhole, are considered. The ray of light leaving the top of the tree in the figure projects onto a pixel on the top of the image plane I_2 and a ray of light from the bottom of the tree projects onto a pixel at the bottom of the image plane I_2. Similarly, other rays of light between these two rays project onto the image plane, and if the camera is aligned to the tree such that the middle of the lens is aligned with and pointing directly toward the middle of the tree, the ray of light from the middle of the tree will travel through the pinhole parallel to the ground. The distance between the image plane and the pinhole is known as the focal length f. The plane I_1 falls precisely on the internal CMOS sensor of the camera, and detects an inverted (upside-down) smaller image of a tree in this case. I_2 is symmetric to I_1 about the pinhole, so that the camera captures the image on the image plane by inverting (in software) the image captured on the CMOS sensor at I_1.

A different view of the image plane is given in Figure 2.b, which clearly indicates the relationship between the pixel coordinates (x_{pixel} and y_{pixel}) and the x- and y-axes of a camera coordinates (x_c and y_c). The z-axis of this system is then orthogonal to the image, and the z-coordinate of an object in this coordinate system thus represents its depth in the scene. Thus, estimating depth using this method involves directly estimating the z-coordinate in the camera world corresponding to a pixel location.

Since the origin in the camera coordinate system is at the pinhole, if a depth of 5 meters is calculated, it will refer to a distance of 5 meters in front of the camera. This is suited for the application, where the camera is mounted inside a vehicle and the camera reference point needs to be dynamic in relation to the movement of the vehicle on the road.

Fig. 4. Fisheye distortion before (left) and after (right) correction.

In general, the mapping from coordinates used for specifying a location in the real world (called world coordinates) to camera coordinates is a projective transformation that can be described by a projection matrix **P**. **P** depends on the camera's intrinsic and extrinsic properties as Equation 1. Our assumptions about the camera installation (see Section IV) simplify the situation by fixing the extrinsic parameters: the assumptions set the rotation matrix **R** to the 3 x 3 identity matrix and the translation vector **t** to the zero vector, i.e. we assume the camera and world coordinate systems are identical. The intrinsic parameters of the camera are represented by the calibration matrix **K** - these depend on the camera and its configuration. The next section discusses camera calibration, a process for estimating the calibration matrix.

$$\boldsymbol{P} = \boldsymbol{K}[\boldsymbol{R}|\boldsymbol{t}] \tag{1}$$

B. *Camera calibration*

Various forms of the calibration matrix can be found in [11]. One of the most common representations is given in Equation 2, where f_x and f_y represent the focal point in the x and y-axis respectively, and the centre point of the CMOS sensor of the camera (also referred to as the principal point) is represented by u_x and u_y.

$$\boldsymbol{K} = \begin{bmatrix} f_x & 0 & u_x \\ 0 & f_y & u_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

Setting $\mathbf{K}_{0,1} = 0$ indicates that the skew of the sensor is presumed to be zero i.e. the image is not distorted in a diagonal manner. In [9] the camera is not calibrated and is chosen as (u_x, u_y) = (W/2, H/2) where W represents the width of the image in pixels and H represents its height. This is usually an acceptable assumption but the origin can vary slightly depending on the manufacturing quality of the camera. For the purposes of this paper, the camera was calibrated, which yields the most accurate calibration matrix for the camera when needed.

The camera was calibrated with the Camera Calibration Toolbox for Matlab [12], using a board with 100 mm x 100 mm squares to enable more accurate calibration at the larger distances relevant to this task. The (estimated) board positions used for calibration are displayed in Figure 3. Note that the plane described by y = 0 in this space is not the road

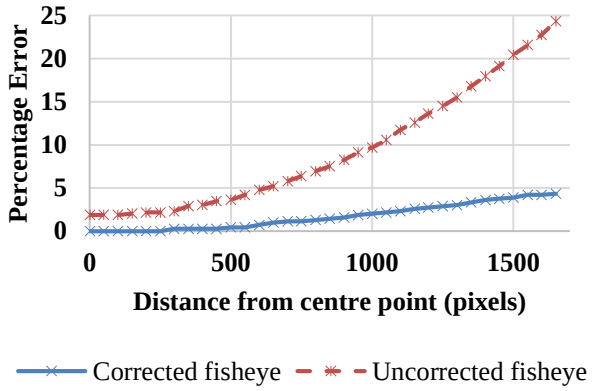

Fig 5. Fisheye correction measurements. Measurements with respect to the horizontal centre of the image

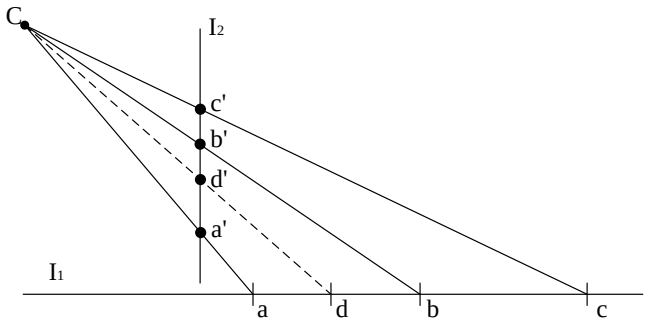

Fig. 6. The mapping from I_1 to I_2 projecting the points a, b, c and d to a', b', c' and d' respectively is an homography through C, and the cross ratio of sets of collinear points is preserved by homographies.

surface itself, because the camera (which is at the origin) is mounted within a vehicle.

Since cameras in practice make use of lenses that adjust the direction of light travel nonuniformly, the pinhole camera model is not entire accurate. It is thus generally desirable to remove the fisheye effect (or other perspective distortion) from images in which depth needs to be estimated, in order to maximize the accuracy of the depth estimation results.

In order to obtain accurate distortion parameters, it is necessary to place the board at various points near the edges of the field of view of the camera, where the fisheye effect is most noticeable. The left hand image in Figure 4 illustrates the fisheye effect of the camera; after calibration the distortion could be mostly eliminated, as seen in the right hand image in Figure 4.

It can be seen that the fisheye effect has not been eliminated; however the distortion has hopefully been reduced enough for its remaining effects to be negligible in future calculations. We also quantified the effects of not performing fisheye correction before depth estimation; see Section IV.B.

To measure the fisheye distortion in the horizontal plane, one of the horizontal lines of the venue where the images were taken was used. The line is indicated by the green crosses in Figure 4. This line is a straight line, however, due to the fisheye distortion introduced in the frame, it can be observed that it is not. The expected straight line was used as a reference and from its centre point, the deviation from this line was measured and a relative percentage error was calculated and is indicated in Figure 5. As can be seen, the corrected fisheye image's graph indicates an error of less than 5% at the edge of the image.

IV. Methodology

The approaches considered in this paper make use of two assumptions that simplify depth estimation. These assumptions are based on the fact that the camera will be mounted in a vehicle and will be facing the road when relevant images are taken. The first assumption is that the road surface is presumed to be flat relative to the vehicle. This should generally be the case below the vehicle, but the assumption may be incorrect for the road section ahead of the vehicle if the road slope changes. In this case, the depth estimation techniques will provide less accurate results. The second important assumption is that the mounted camera's lens is directly facing the road and that it faces the road in such a manner that a light ray through the centre point of the camera lens is parallel to the road surface and edges, thereby implying that there is zero yaw, pitch and roll when the camera faces the road. This is the assumption simplifying the form of the projection matrix by fixing the extrinsic parameters to convenient values, described in Section III.B. Note that there are methods for determining the ground plane, such as the work done in [13], if these assumptions are invalid. These techniques fall beyond the scope of this work, which focuses on contrasting the accuracy of depth estimation techniques.

Section IV.A describes a simple approach based on the geometry of similar triangles. An approach employing the cross ratio of collinear points is then described in Section IV.B. Section IV.C finally describes a versatile depth estimation which makes use of camera calibration, and only requires one other reference point in the image. This approach essentially determines the location of an object in the ground plane by solving equations involving the projection matrix **P**. We refer to this approach as the pinhole (camera) model approach. The low-cost experimental setup used to compare the accuracy of these methods is given in Section IV.D.

A. *Similar triangles*

A common geometric approach to estimating depth of an object directly ahead of the camera relies on the geometry of similar triangles [14] and can be deduced from a figure similar to Figure 2.a. The equation one arrives at is given in Equation 3 where z is the estimated depth in the image if the focal length (f), height of the camera (H) and y pixel value (y) at the point on the image plane are known.

$$z = \frac{fH}{y} \tag{3}$$

We can employ this formula to estimate the depth of an object on the road in our domain once we have the focal length of the camera: since the height of the camera is known, this formula implicitly maps pixel heights to estimated distances.

We consider two approaches to obtaining the focal length in this work.

B. Cross ratio formula

The second approach to depth estimation that we consider makes use of reference pixels corresponding to known distances within the image. The approach, presented in [3], employs the cross ratio projective transformation defined Equation 4 [13]. Here, points a-d refer to homogenous points in a one-dimensional plane $\mathbb{P}^1$, so that the right hand side is a quotient of products of matrix determinants.

$$Cross(a,b,c,d) = \frac{|a\ b||c\ d|}{|a\ c||b\ d|} \tag{4}$$

Consider four rays in a plane passing through a common point C. The projection of these rays on any other plane yields four collinear points, and the cross ratio of these points is invariant to the choice of the plane [15]. The use of this technique for our domain is illustrated in Figure 6, where C is the pinhole from the pinhole camera model. In this figure, the one-dimensional data on line l_1 represent depths in the road directly ahead of the camera and the one-dimensional data on line l_2 are the y-values of the corresponding pixels in the image, with x-coordinate equal to that of the principal point. We apply this technique by determining pixel locations for a number of known distances in the road in advance, and then selecting three equally-spaced distance for use as (a-c) together with an unknown distance d corresponding to a particular pixel height y. It is not required that the reference points be equally spaced, but it simplifies the exposition below. Thus the cross ratio for the data in the one dimensional road plane can be expressed as cross(a,b,c,d) = cross(0, L, 2L, d) and similarly that for the one-dimensional pixel y-values can be expressed as cross(a',b',c',d') = cross(y_0, y_1, y_2, y), where y_0, y_1, and y_2 have been determined in advance. Equating these cross ratios to determine d yields Equation 5. Note that this approach implicitly measures depths relative to the depth of reference point a.

$$d = \frac{(y_0-y_1)(y_2-y)-(y_0-y_2)(y_1-y)}{(y_0-y_1)(y_2-y)-\frac{1}{2}(y_0-y_2)(y_1-y)} L \tag{5}$$

For our experiments, we determined pixel values for reference points at 5 m intervals (from 5 m to 30 m), so we could always apply Equation 5 with L = 5. We selected our points a-c to include the reference points directly above and below the pixel of interest: for depths below 15 m, the reference points are 5 m, 10 m and 15 m. Between 15 m and 20 m, the reference points are 15 m, 20 m and 25 m. Lastly, for depths between 20 m and 30 m, the three reference points are 20 m, 25 m and 30 m.

The approaches described above (using similar triangles or the cross ratio) can only predict distances for points directly ahead of the camera. In order to perform depth estimation for other pixels, we simply predict the depth using the y-coordinate of the pixel. If the camera setup is correct and the perspective distortion has been eliminated, this will give the correct depth - essentially, we predict the depth of a point in the road next to the pothole.

C. Pinhole camera model formula

The pinhole camera model specifies that the projection of a ray through the pinhole onto the image plane is determined by the projection matrix **P** according to Equation 6, where $\mathbf{X_w}$ specifies the homogeneous world coordinates of the ray and the pixel coordinates are denoted by $\boldsymbol{x}$.

$$\boldsymbol{x} = \boldsymbol{P}\boldsymbol{X_w} \tag{6}$$

Using the pseudoinverse **P*** of **P** (note that **P** is not square), we can determine the homogeneous coordinates of the ray passing through a pixel of interest and the pinhole by

$$\boldsymbol{X_w} = \boldsymbol{P}^*\boldsymbol{x} \tag{7}$$

To determine the world coordinates corresponding to the pixel in question, it is necessary to determine the point where the ray travelling through this pixel and the plane of the road intersects. After computing $\mathbf{X_w}$, it must thus be dehomogenized to yield the desired world coordinates by finding a suitable scalar factor λ in Equation 8.

$$\boldsymbol{X_w} = \lambda \boldsymbol{P}^*\boldsymbol{x} \ \ where\ \lambda \in \mathbb{R} > 0 \tag{8}$$

The derivation to determine λ uses the normal form representation of a plane (Equation 9) which characterizes the points on a plane in terms of a reference point V in the plane and the normal **n** of the plane.

$$\boldsymbol{n} \cdot (P_1 - V) = 0 \tag{9}$$

The normal of the road plane is always the same in our application due to our assumption about the camera setup, and it can be determined by placing the calibration board flat on the road surface, and recording the translation vector and rotation matrix describing the position and orientation of the board. The location we place the calibration board at serves as our reference point V (Note that the thickness of the calibration board must be taken into account).

Next, setting $P_1 = \mathbf{X_w}$ in Equation 9 and solving for λ yields:

$$\lambda = \frac{\boldsymbol{n}\cdot V}{\boldsymbol{n}\cdot \boldsymbol{P}^*\boldsymbol{x}} \tag{10}$$

Finally, using this value of λ in Equation 8 yields the estimated world coordinates of the pothole in the road.

$$\boldsymbol{X_w} = \frac{\boldsymbol{n}\cdot V}{\boldsymbol{n}\cdot \boldsymbol{P}^*\boldsymbol{x}} \boldsymbol{P}^*\boldsymbol{x} \tag{11}$$

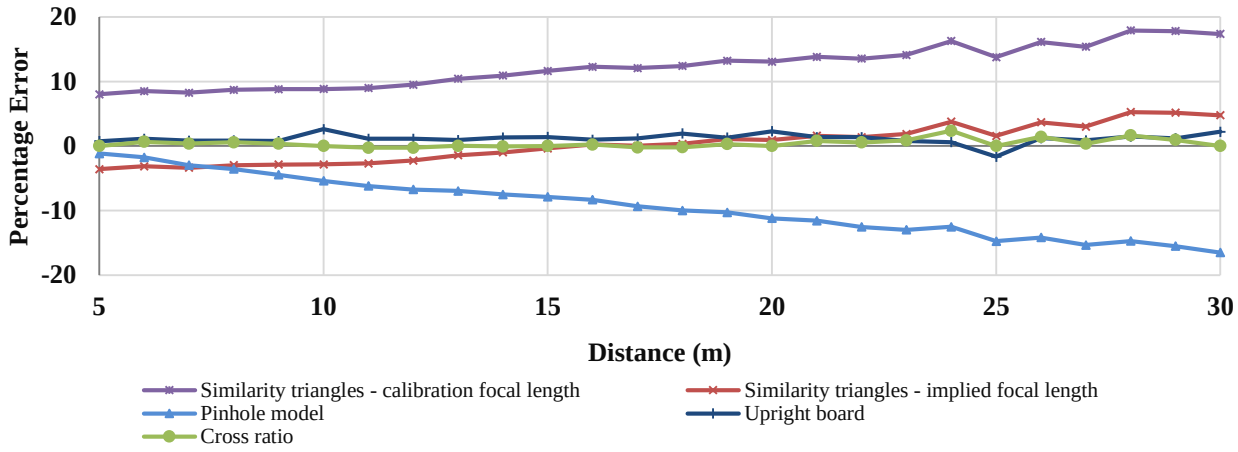

Fig. 7. Error percentages of the approaches considered at various distances directly ahead of the camera. A negative (positive) percentage error indicates underestimation (overestimation) of the depth.

An advantage of this method is that it explicitly provides the world coordinates corresponding to a pixel and therefore contains information of the relative position of the pothole for all three axes, not just the depth.

D. Experimental setup

To determine the true depth within the images, a measuring tape was placed on a flat surface in front of the camera such that the tape appears to run vertically in the camera image. Measurements were taken at 1 m intervals between 5 m and 30 m. At each of the one meter intervals between 5 m and 30 m the calibration board was held upright and flush with the surface and photographed. These results are recorded to demonstrate how accurate the pinhole model could be if more scale and orientation information were available at the point of interest. This process was performed using the same setup used for the camera calibration. In order to further quantify the accuracy of the techniques, an additional set of measurements were taken parallel to these, but at a distance of 4.93 m to the left of the camera. The camera was set up again, but not recalibrated, for this process.

In order to minimize alignment problem between the camera mounted within the vehicle and the centre point of the scene, the vehicle setup was eliminated in the final measurement setup. Consequently, the camera was placed on a tripod at the exact height it would have been in the vehicle.

The images used for this study were captured by a GoPro Hero 3+ camera with the resolution set to 3680 x 2760. The high resolution is necessary for the pothole detection system presented in [1] which uses the same images.

In a real-world scenario, a calibration procedure would be required to mount the camera to the vehicle to ensure the most accurate results. If the pothole detector becomes a commercial product, the manufacturer could develop a calibration device/procedure that would ensure the camera is mounted correctly.

V. RESULTS

The principal findings of the study are shown in Figure 7. This figure displays the error percentage in predicted distance at various distances from the camera, when the object is directly ahead of the camera. The values are plotted for each technique. For the similar triangles approach, results are given using the focal length estimated from camera calibration, as well as that obtained by averaging the focal length implied by the reference points used by the cross ratio approach. The pinhole model approach still has good performance for short distance, but the error percentage worsens approximately linearly as the distance increases, resulting in very poor performance for large distances. The similar triangles approach using the focal length from the camera calibration process also fares poorly over the entire range of distance, while the version using the averaged implied focal lengths performed much better. The disparity in these results is surprising, since other aspects (including the “Upright board” measurements discussed below) indicated that the values in the calibration matrix were fairly accurate. The cross ratio approach performed best among the approaches (with an error of less that 2% across the entire range), performing even better than the pinhole model approach augmented with additional information from the upright board used during the measurement process. (These results, labelled "Upright board" in the figure, are a dramatic improvement over the baseline performance of the pinhole model.) Omitting the fisheye removal step had a negligible effect on all these curves.

These results indicate that for the problem at hand, where we are estimating distances between 5-30 m directly ahead of us, and have reference pixels of known distance available, the cross ratio formula is the technique of choice. However, objects may not be directly ahead of the camera, which the cross ratio and similar triangles approach expect - our proposal of only considering the y-coordinate may introduce too much inaccuracy. Furthermore, in this situation, if camera calibration is not performed, fisheye removal may not be possible. Finally, one may not have as many reference points in practice as used above. We consider the effect of these limitations next.

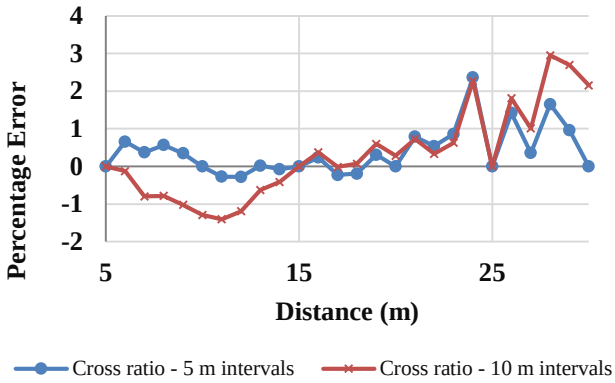

Fig. 8. Cross ratio comparison between 5 m intervals and 10 m intervals

Figure 9 provides results analogous to those in Figure 7, except that these errors are those obtained for the measurements 4.93 m to the left of the camera. At close distances, this corresponds to a wide angle, so the fisheye effect is much more noticeable for the location of the measurement on the image. Thus the figure also includes results when fisheye correction is not performed, since the difference is no longer negligible. Finally, at these angles, it is not always possible to reliably estimate the distance using the information from the upright board due to limitations in the camera calibration toolbox, so the corresponding curves for that method are omitted.

It is not surprising that the results are considerably worse in this scenario. Consider first the results where fisheye correction has been performed. Once again, the estimated distance for the pinhole model increase too slowly, resulting in a linear decrease on the graph; however, since the model initially overestimates the distance, there is a short window where it performs well. The cross ratio approach begins slightly worse than the pinhole model at 5 m, but becomes consistently more accurate as the distance increases, with less than 5% error for all distances beyond 13 m. Again, the similarity triangles approach using the calibrated focal length performed dismally. However, when this approach used the average implied focal lengths, it performed considerably better than the other approaches for shorter distances, and is only slightly worse than the cross ratio technique beyond 15 m.

Failure to perform fisheye removal in this setting has a notable effect, leading to increases in the estimated distances. The difference in accuracy resulting from the failure to perform fisheye correction reduces as the distance increases (from about 20% at 5 m, to a negligible effect at 20 m). This is because the angle from vertical decreases and the pixels under consideration thus move closer to the centre of the image, where the fisheye effect is weaker. These results indicate that if one's system should be able to estimate distances to objects at wide angles, it may be preferable to use the similar triangles approach (with implied average focal length) rather than the cross ratio approach, since the loss of accuracy directly ahead and at further distances may be countered by improved performance at wide angles.

In order to investigate the effect of the number of reference points used as anchors for the cross-ratio formula, Figure 8 compares the cross ratio results in Figure 7 to the performance of the cross ratio approach when only three reference points (the minimum needed for the cross ratio approach) were used. In particular, we used a 10 m interval, with reference points at 5 m, 15 m, and 25 m. Since the error of the techniques at a reference point for this technique is zero, there is more freedom for the technique to introduce error with the wider-spaced reference points, and we do see a reduction in performance. However, the performance is still much better than the pinhole model, and comparable to the best other approaches. (Note that we would also expect the similar triangles approach using implied focal lengths to worsen if only given three reference points.)

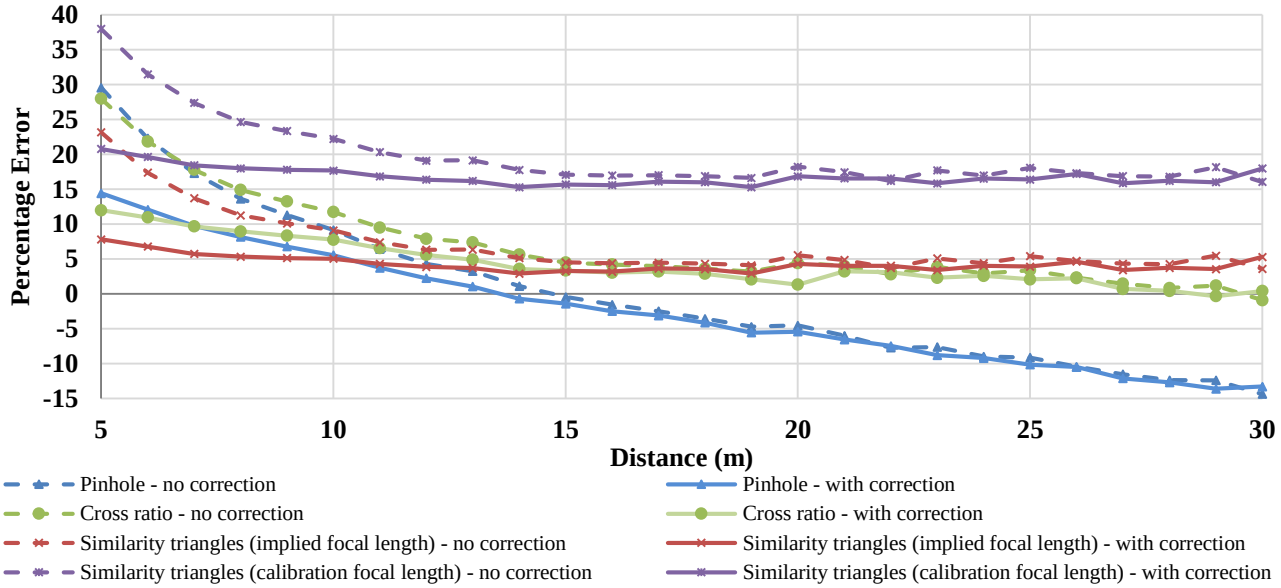

Fig. 9. Error percentages of the approaches considered for measurements 4.93 m to the left of the camera. Solid lines indicate error with fisheye correction performed, dotted lines indicate error with fisheye correction omitted. A negative (positive) percentage error indicates underestimation (overestimation) of the depth.

VI. Conclusion

The work presented in this paper assessed three different approaches to determining the depth estimation problem of determining the distance to an object, such as a pothole on a road. The pinhole model requires camera calibration to be performed, while the cross ratio approach requires reference pixels corresponding to known distances to be determined. The approach using similar triangles requires the focal length of the camera to be known. This information might be available for some camera models, but might need to be calculated (either as part of a camera calibration process, or from reference pixels as per the cross ratio approach). Furthermore fisheye distortion can impact distance estimation, so it is desirable to remove it if possible; however, this also requires camera calibration.

We found that for depth estimation directly ahead in the range 5-30 m, the fisheye correction has negligible effect, and the techniques based on reference points were far superior to the pinhole model approach, with the cross ratio approach performing the best directly ahead of the camera. At wide angles (and thus short distances) in this range, the fisheye effect is stronger and the similar triangles approach using implied average focal lengths performs better. If depth estimation in these cases are needed, we recommend performing camera calibration to remove the fisheye effect, and considering a hybrid approach, using the cross ratio approach to get good accuracy directly ahead and at larger distances, and using the similar triangles approach for depth estimates at wider angles.

A limitation of this study is that the results for the pinhole model depend on the calibration matrix, as does the similar triangles approach using its focal length. While care was taken to calibrate the camera accurately, there are always errors in such a process, and it may be that recalibration yields different results. While we believe our results qualitatively capture the nature of the problem with the pinhole model approach for this task - the distance estimation error percentage increases linearly to unacceptable levels at larger distances - it would be worthwhile to verify this,

Finally, we note that if the camera height is changed (for example, by deploying the camera in a different vehicle), the pinhole model and similar triangle approaches can be employed directly once the change in height has been established. On the other hand, the cross ratio approach requires that the location of the reference pixels be re-established. This is an argument in favour of using the similar triangles approach throughout.

Acknowledgment

The authors would like to thank MTN for their support via the MTN Mobile Intelligence Lab, as well as Dr W Brink at Stellenbosch University for his assistance.

References

[1] S. Nienaber, M. Booysen and R. Kroon, "Detecting potholes using simple image processing techniques and real world footage," in *South African transport conference*, Pretoria, 2015.

[2] S. Nienaber, M. Booysen and R. Kroon, "Electrical & Electronic Engineering, Stellenbosch University," October 2015. [Online]. Available: http://staff.ee.sun.ac.za/mjbooysen/Potholes/.

[3] V. Singh, A. Rai, N. Hemanth, D. Rohit and A. Mukherjee, "Algorithms for real time detection and depth calculation of obstacles by autonomous robots," in *IEEE Conference on Robotics, Automation and Mechatronics*, Chengdu, 2008.

[4] A. Criminisi, I. Reid and A. Zisserman, "Single view metrology," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123-148, 2000.

[5] I. Rapp, "Motion capture actors: body movement tells the story," [Online]. Available: http://www.nycastings.com/dmxreadyv2/blogmanager/v3_blogmanager.asp?post=motioncaptureactors. [Accessed 20 May 2015].

[6] A. Bhatti, Current advancements in stereo vision, InTech, 2012.

[7] A. Joglekar, D. Joshi, R. Khemani, S. Nair and S. Sahare, "Depth estimation using monocular camera," *International journal of computer science and information technologies*, vol. 2, no. 4, pp. 1758-1763, 2011.

[8] G. Stein, O. Mano and A. Shashua, "Vision-based ACC with a single camera: bounds on range and range rate accuracy," in *Intelligent vehicles symposium*, Columbus, 2003.

[9] A. Saxena, S. Chung and A. Y. Ng, "3D depth reconstruction from a single still image," *International Journal of Computer Vision*, vol. 76, no. 1, pp. 53-69, 2008.

[10] B. Lui, S. Gould and D. Koller, "Single image depth estimation from predicted semantic labels," in *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, 2010.

[11] A. Rahimi, H. Moradi and R. A. Zoroofi, "Single image ground plane estimation," in *20th IEEE International Conference on Image Processing*, Melbourne, 2013.

[12] R. Szeliski, Computer Vision - Algorithms and Applications, London: Springer, 2011.

[13] J. Bouguet, "Camera calibration toolbox for Matlab," [Online]. Available: http://www.vision.caltech.edu/bougeutj/calib_doc/. [Accessed 21 April 2015].

[14] R. Dragon and L. V. Gool, "Ground plane estimation using a hidden markov model," in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 2014.

[15] J. Bird, Basic engineering mathematics 4th Edition, Elsevier, 2005.

[16] A. Zisserman and R. Hartley, Multiple view geometry in computer vision, second edition, Cambridge, 2000.