Measuring Distance with Mobile Phones Using Single-Camera Stereo Vision

Clemens Holzmann
University of Applied Sciences Upper Austria
Softwarepark 11, 4232 Hagenberg, Austria
clemens.holzmann@fh-hagenberg.at

Matthias Hochgatterer
University of Applied Sciences Upper Austria
Softwarepark 11, 4232 Hagenberg, Austria
matthias.hochgatterer@fh-hagenberg.at

***Abstract*—Computer stereo vision is an important technique for robotic navigation and other mobile scenarios where depth perception is needed, but it usually requires two cameras with a known horizontal displacement. In this paper, we present a solution for mobile devices with just one camera, which is a first step towards making computer stereo vision available to a wide range of devices that are not equipped with stereo cameras. We have built a prototype using a state-of-the-art mobile phone, which has to be manually displaced in order to record images from different lines of sight. Since the displacement between the two images is not known in advance, it is measured using the phone's inertial sensors. We evaluated the accuracy of our single-camera approach by performing distance calculations to everyday objects in different indoor and outdoor scenarios, and compared the results with that of a stereo camera phone. As a main advantage of a single moving camera is the possibility to vary its relative position between taking the two pictures, we investigated the effect of different camera displacements on the accuracy of distance measurements.**

***Keywords*-Depth perception, single-camera stereo vision, inertial sensing.**

I. Introduction

Visual distance estimation is a fundamental part of human perception. It is based on several depth cues including the size and contrast of objects as well as the exploitation of parallax, which refers to a displacement of the apparent position of an object viewed from different lines of sight. This parallax phenomenon is used in binocular or stereo vision for human depth perception, where the difference in the views from the left and right eye is analyzed. As stereo vision is a robust and reliable depth perception technique, it has been used in robotics and other computer systems to calculate the distance to objects. In computer stereo vision systems, two cameras observe a scene from different locations. The different camera locations result in different image locations of the objects. The difference in the image locations is called disparity and the distance between the two cameras is called baseline. Based on the disparity and baseline, the distance can be calculated.

However, stereo vision is not constrained to two separate cameras. Recording two images with a single camera from different locations results in the same image disparities. This approach is referred to as single-camera stereo vision. Compared to traditional stereo vision, the baseline between the cameras is not fixed but can vary instead. This leads to a potentially higher distance accuracy, which can be achieved by increasing the distance between the camera positions from where the two pictures are taken. Another obvious advantage of using a single camera are the lower hardware costs. On the other hand, the single-camera approach requires a measurement of the camera's translational and rotational movements to calculate its relative positions, which can be achieved with inertial sensors.

In this paper, we present a single-camera stereo vision system based on a mobile phone with integrated inertial sensors. The aim of our work was to explore the distance accuracy in comparison to stereo vision systems using two cameras. In particular, we wanted to find out whether bigger baselines lead to more accurate distance measurements or not. Thus, the contributions of this paper are as follows:

- We have developed a single-camera stereo vision prototype, which will be presented in Section III. The most challenging parts were the baseline measurement with inertial sensors as well as the disparity calculation from image features. The theoretical background on distance calculation with stereo vision will be given in Section II.
- Several experiments have been conducted to show whether bigger baselines actually lead to more accurate results on the one hand, and how different distances between camera and object affect the measurement accuracy on the other hand. The results of the experiments will be discussed in Section IV.

II. Depth Perception with Stereo Vision

In this section, stereo vision fundamentals will be described, which provide the basis for our single-camera approach presented in Section III. First, motion parallax and binocular vision will be explained, which are the two general principles of stereo vision. Second, the stereo triangulation technique for calculating the distance to an object will be explained and the achievable depth resolution will be discussed. Stereo triangulation is based on the difference of an object's position in two images due to a changed line of sight, which requires to match those image points from two images at a time that are projections of the same 3D point. This is referred to as the correspondence problem, and it will be explained at the end of this section.

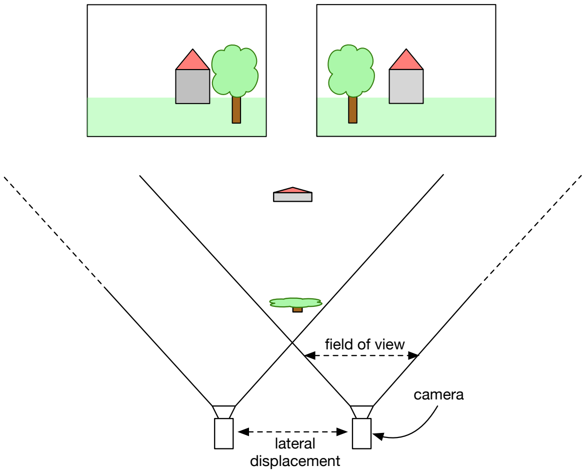


Figure 1. Two images recorded from a laterally displaced camera. The displacement of the objects provides relative depth information.

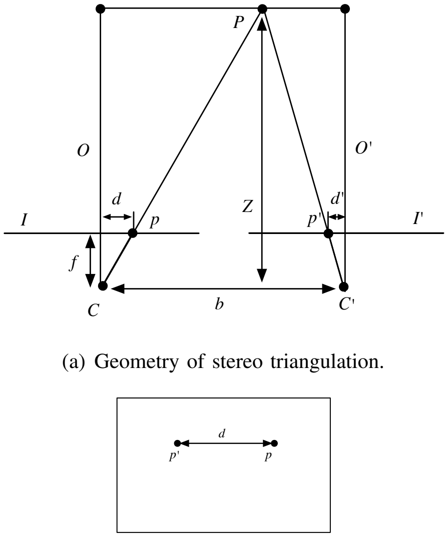
A. Motion Parallax and Binocular Vision

Motion parallax basically describes the displacement in the apparent position of objects viewed from different lines of sight. The displacement of an object is called disparity, which is a cue for its distance to the camera. An example of a motion parallax is shown in Figure 1, where two images depict a scene recorded from a laterally displaced camera. The different camera viewing angles to the objects result in displaced object locations on the images. In this example, the tree has a bigger disparity and therefore seems to be closer than the house.

Binocular vision, also known as stereo vision, allows humans to perceive the environment in three dimensions as a result of a combination of images from both eyes. In computer stereo vision the same technique is used, where two cameras – whose relative positions are known – capture a scene from two positions [1]. The disparity between the images is calculated and used for a reconstruction of the three-dimensional scene. The matching of an object's projection on the two images is a great challenge in computer stereo vision systems, and it is referred to as the correspondence problem [2] described further below.

In stereo vision systems, it is required that the distance between the two cameras is known. However, as will be described in Section III, binocular vision can also be used in a monocular vision system with one camera. Applying a lateral displacement with a known distance to a camera results in the same perspective projection and motion parallax than in a system with two fixed cameras.

B. Stereo Triangulation



(a) Geometry of stereo triangulation.

(b) Overlay of the images shows the difference between the image locations p and p'.

Figure 2. Stereo vision setup with two cameras with parallel optical axes. The cameras have equal focal length f and are laterally displaced by a baseline b.

A common technique to calculate the distance in a stereo vision system is called stereo triangulation, which requires that the two optical axes of the cameras are parallel. As shown in Figure 2, the distance of a point P in 3D space is defined by the intersection of the two rays from the optical centers C and C' through the respective image points p and p' on the image planes I and I' [3]. The triangulation is based on similar triangles, for which the ratios

$$\frac{b}{Z} = \frac{b - (d + d')}{Z - f} = \frac{d + d'}{f} \tag{1}$$

are equal. This leads to the following triangulation formula to compute the distance Z, where f is the focal length, D the disparity ($D = d + d'$) and b the baseline:

$$Z = \frac{f \cdot b}{D} \tag{2}$$

The focal length f of a lens is the distance from the lens to the digital camera sensor when a far away object is in focus, and it is predefined by the used camera. The baseline b is defined by the distance between the optical centers of two displaced cameras. For a constant distance, an increasing baseline results in a higher accuracy of the system due to limitations of the camera resolution.

The disparity D is the horizontal displacement of a stationary object on images which are captured from two laterally displaced camera positions. The object disparity from a camera translation is an important cue for depth perception, which is based on the motion parallax phenomenon described above. In contrast, rotation around the optical center, which also causes object disparity, doesn't include any depth information.

C. Depth Resolution

The depth resolution of a stereo vision system is limited. An obvious limitation is the camera resolution, which specifies how many pixels an image consists of. An image with lower resolution consists of fewer image points, and therefore changes in the scene may not be seen as well as on a high-resolution image. The focal length of the camera also influences the depth resolution as it is directly related to its field of view, but it is usually fixed. Another limiting factor is the baseline. As can be seen in Figure 3, an increasing baseline results in a better depth accuracy, where R denotes the size of uncertainty for a fixed baseline and different distances. For single-camera systems with varying baselines, the depth resolution could theoretically be better than in traditional stereo vision system.

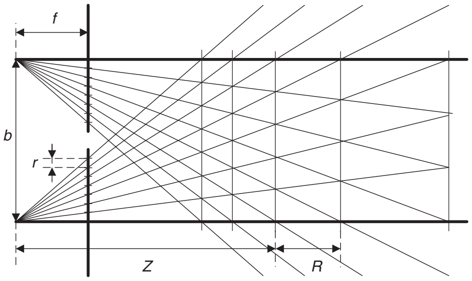


Figure 3. The limitations of stereo triangulation for a specific baseline [4].

The depth resolution R can be calculated with the triangulation equation 2 as

$$R = Z_1 - Z_2 \tag{3}$$

$$Z_1 = \frac{f \cdot b}{d_1} \tag{4}$$

$$Z_2 = \frac{f \cdot b}{d_1 + r_m} \tag{5}$$

where d_1 is the discretized disparity and r_m is the width of one pixel in metric units. A discretization of the disparity occurs when the camera chip rasterizes the light rays to form an image.

When designing a stereo vision system, the camera resolution and baseline have to be chosen properly to meet the given accuracy requirements.

D. The Correspondence Problem

The correspondence problem is the problem to match image points from two images which are projections of the same 3D point. It is the main problem in computer stereo vision when calculating the disparity for stereo triangulation. The change in perspective, lighting and image location as well as object occlusions make the matching process difficult. There are many different approaches available, an overview of the techniques and algorithms is given in [3] and [5].

In general, two correspondence methods can be distinguished. The first category are direct methods, which compare the pixel intensity values between the two images to match similar image regions. They can be fast to compute, but they much likely lead to matching errors. The reason is that the pixel intensity depends on a lot of external factors like lighting conditions, sensor noise and camera resolution, which makes pixel matching more difficult. The second category are indirect methods, which locate distinctive features in the images. These features are associated with image properties like edges, corners or other interesting regions. A main requirement of good features is their repeatability, which includes invariance against image transformations and robustness against deformations [6].

For our work, we have used the SURF (Speeded Up Robust Features) algorithm [7] to detect and describe image features. The features are scale and rotation-invariant, which makes them robust against affine image transformations, and they are fast to compute. Especially on mobile devices, which have limited computational resources, SURF features provide an acceptable tradeoff between repeatability and efficiency [6].

III. Single-Camera Stereo Vision Prototype

Stereo vision systems are usually equipped with two cameras which are displaced by a fixed baseline. This is an advantage on the one hand, as the spatial relationship of the cameras is known and does not have to be measured during operation, but it is a drawback on the other hand, as the depth resolution in stereo triangulation is limited by this baseline. However, there are also systems which use just one camera, and they are referred to as single-camera stereo vision systems. Due to the fact that only one camera is used, the baseline can be varied. Therefore, a single-camera stereo vision systems may achieve a much higher depth resolution since it is not tied to a fixed baseline.

There are two different approaches for single-camera stereo vision: axial and lateral motion stereo. In axial motion stereo, the camera is moved along the optical axis, which results in a disparity on the image [8]. The second approach for single-camera stereo vision is lateral motion stereo, which is equal to a common stereo vision system. The camera is displaced laterally which results in the same motion parallax phenomena than in a normal stereo vision system, and the distance can be calculated with stereo triangulation as presented in Section II.

We have implemented the lateral motion stereo approach, as we were interested in how a mobile single-camera system performs in comparison to a stereo vision system with two cameras. In the following, our prototypical implementation, and especially how the baseline and disparity have been determined, will be explained.

A. Prototype Implementation

The prototype has been implemented on an iPhone 4 mobile phone with iOS 4.3.3. The used sensors of the mobile phone were its five-megapixel camera, three-axis gyroscope and three-axis accelerometer. The prototype processes raw camera frames in a 32 Bit RGBA (Red Green Blue Alpha) format and with a resolution of 640 x 480 pixels. The translational movement of the device is recorded by the gyroscope and the accelerometer. Based on the acceleration values in combination with the gyroscope data, the baseline can be calculated.

The disparity calculation of the object is based on image features. As mentioned in Section II, we used SURF features to extract interesting image regions. An implementation of the SURF algorithm is included in the Open Source Computer Vision (OpenCV)[1] library, which is also available for the iOS platform. As the SURF algorithm in OpenCV uses a custom image structure, the image captured from the camera has to be converted into an OpenCV 8-bit (unsigned) grayscale image.

Afterwards, features are extracted from the first image and matched with similar features from the second image. In the first image however, just the features of a 70x70 pixels region in the middle of the image are considered, which allows the user to select an object of interest to which the distance should be calculated. The disparity is then calculated by measuring the horizontal displacement of matched features. The overall disparity, which is used in the stereo triangulation equation, is the median of the five matched features that are nearest to the center of the first image. Using the median helps to eliminate outliers due to incorrect feature matches.

In contrast to stereo vision systems with two cameras, we calculated the baseline by a double integration over the acceleration values in x-direction as described in [9]. The accelerometer output is accessed via the iOS Core Motion framework[2]. It tracks the device's motion using both the built-in gyroscope and the accelerometer, and can therewith differentiate between gravity and user acceleration. This differentiation is necessary, as just the user acceleration reflects the device movement over time and is used for the baseline calculation. Using only the accelerometer would not be sufficient, as it measures the sum of these two acceleration vectors.

Once the disparity and baseline are known, the distance can be calculated with the stereo triangulation formula presented in Section II. First, the feature locations in the second image have to be corrected to get the real disparity. This is done by multiplying the feature positions with a transformation matrix. Its rotation and translation components are determined by calculating the difference of the device's orientations at the two points in time when the images are recorded. This step is necessary, as the displacement of objects due to a rotation of the camera does not provide any depth information. Afterwards, the disparities are converted from pixel into metric units, and the distance is finally calculated with the triangulation formula.

IV. Experiments

In order to evaluate the accuracy of our proposed single-camera stereo vision approach, we have conducted several experiments which are described in this section. We will first give an overview of the research question and the design of the experiments. Afterwards, we will present the results in detail, and conclude the section with a summary and discussion of our findings.

A. Overview and Setup

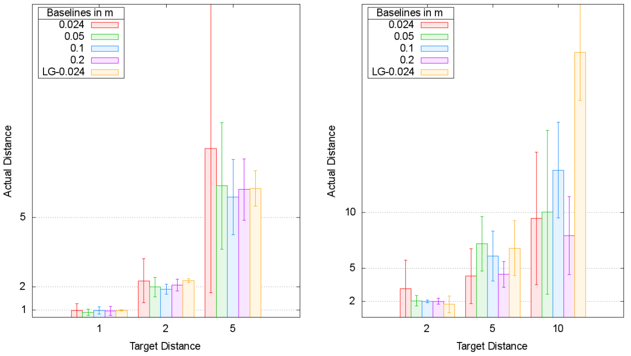
The goal of the experiments was to investigate the accuracy of the distance estimation of our prototype under real-world conditions. First, we were interested in how different distances of objects to the camera affect the accuracy of the distance estimation. Second, we wanted to find out whether bigger baselines lead to more accurate distance estimations or not. On the one hand, the distance accuracy should become higher with an increasing baseline, which is obvious from the stereo vision fundamentals presented in Section III. On the other hand, the baseline is measured with inertial sensors, which leads to higher measurement errors for longer baselines.

In order to answer these research questions, we designed two experiments. In the first experiment, distances to a coffee maker (indoor) and a parked car (outdoor) have been measured with our prototype under good lighting conditions. The measurement have been performed from different distances (1, 2 and $5m$ indoors and 2, 5 and $10m$ outdoors) and with four different baselines (2.4, 5, 10 and $20cm$) at a time. The prototype has been moved horizontally on a plate with markings for the four baselines. This should reduce errors from the inertial sensors, as the prototype could not be rotated but just moved in one direction. The two images have been taken manually before and after moving the prototype. We have performed 10 test runs for each combination of baseline and distance, which resulted in 240 test runs in total for the first experiment.

In the second experiment, the same measurements like in the first experiment have been performed, with the difference that the prototype device has been moved free-handedly. It has been grabbed with both hands, and moved horizontally without holding on – again by using the markings on the plate which has been placed below. With this second experiment, we wanted to see if our approach also worked for a free-handed use of the device, and how big the difference in

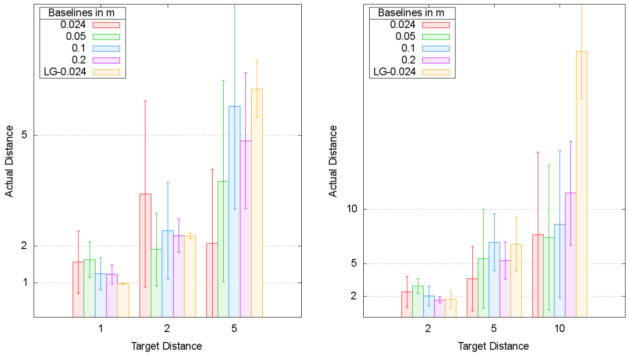
[1] http://opencv.willowgarage.com

[2] developer.apple.com/library/IOS/#documentation/CoreMotion/Reference/CoreMotion_Reference/



(a) Coffee Maker - Indoor (b) Car - Outdoor

Figure 4. Mean value and standard deviation of estimated distances in the first experiment, where the prototype has been moved on a plate.



(a) Coffee Maker - Indoor (b) Car - Outdoor

Figure 5. Mean value and standard deviation of estimated distances in the second experiment, where the prototype has been moved free-handedly.

terms of accuracy was compared to the controlled movement of the phone in the first experiment.

The results from both experiments have been compared with that of an LG Optimus 3D (P920) [3] mobile phone which has a built-in stereo camera. The cameras have a resolution of five megapixels and a focal length of $4.6mm$. The baseline of the LG smartphone is $2.4cm$, which is the reason for choosing this baseline also for the experiments with our prototype. The image resolution has been limited to 640x720 pixels, which is the same width than that of our prototype with 640x480 pixels; the heights can be disregarded as just the horizontal disparity has been taken into account. The distance calculations have been performed exactly the same way as for our prototype presented in Section III, with the only difference that the fixed baseline of $2.4cm$ was used.

B. *Results of the Experiments*

For every baseline and distance specified in the previous subsection, Figure 4 shows the mean values of the ten test runs together with their standard deviations. The experiments also included ten test runs for the LG smartphone, because even though the baseline was fixed, the image pairs recorded by the stereo camera of the phone did not result in the same distance every time. First, the images from the stereo camera are exported as JPEG files. JPEG compression adds compression artifacts, which may influence the feature matching and therefore also the disparity calculation. Second, since the LG smartphone calibrates the exposure settings of the stereo camera automatically, some areas in the images can be very bright or dark, which makes the disparity calculation more difficult. We observed this effect especially in outdoor scenes, where in some cases the exposure was much higher for the stereo image pairs from the LG smartphone compared to the iPhone 4.

With the first experiment, we found out that bigger baselines actually increased the accuracy, although the baseline has been measured with inertial sensors. Especially for the indoor scenario, a decrease of the standard deviation for bigger baselines can be clearly seen in Figure 4. Second, we can also see from the results that larger distances between camera and object lead to less accurate results. The reason is the decreasing disparity with an increasing distance between object and camera, which is obvious from the descriptions in Section III-A. Although the results of the outdoor scenario are not that obvious, the general findings concerning variations in baseline and distance are the same like in the first scenario. However, a main difference was that the LG smartphone performed worst in the outdoor scenario, which was caused by the image quality and exposure of the stereo camera explained before. This result also shows that small variations in the disparity may lead to huge errors if the baseline is small.

The results from the second experiment, where the prototype device was moved free-handedly, show that the standard deviation became bigger compared to the first experiment. The reason for that were unintentional rotations of the device. They influenced the measured baselines on the one hand, which have shown to become worse compared to a displacement on a plate, and the disparity calculation on the other hand. In both experiments, the standard deviation is the lowest for a large baseline of $20cm$ compared to smaller baselines, and it is the highest for long distances of $5m$ or $10m$ compared to shorter distances.

C. *Summary and Discussion*

To sum up, the experiments have shown that the distance accuracy becomes better with an increasing baseline, although it was not predefined as it is the case in stereo vision systems with two cameras. Instead, the baseline has been measured with built-in inertial sensors of our prototype. Since other factors besides the baseline influence the distance accuracy as well, especially unintentional rotations of the devices when moving it from one position to the other as well as bad lighting conditions which may influence the

[3] http://www.lg.com/uk/mobile-phones/all-lg-phones/LG-android-mobile-phone-P920.jsp

feature matching, a baseline of $20cm$ did not always lead to the best results. However, longer baselines improved the distance average and standard deviation in most cases.

For long distances, both the LG smartphone as well as our prototype were unable to accurately determine the distance to the objects. A main reason for that is the discretization of the disparity, which occurs due to a low horizontal image resolution of 640 pixels. Another important factor is the baseline, which is very short in the case of the LG smartphone and inaccurate in the case of our prototype where it has been measured with inertial sensors. For shorter distances, both systems perform better.

A surprising finding was that results of the LG smartphone were worse in our experiments, which we trace back to the lower image quality due to JPEG compression and exposure settings of the stereo camera. The compression artifacts from the JPEG export and the image exposure, where some regions in the image are lightened or darkened, influence the disparity calculation and furthermore the calculated distance. As our prototype device processed raw camera frames, no compression artifacts occurred.

V. Related Work

Single-camera vision systems which use geometric constraints are presented in [10] and [11]. By exploiting certain characteristics of the environment and the objects of interest, the distances to these objects can be estimated. In contrast to our approach, those systems are not applicable in unknown environments, but they are rather designed for specific scenarios.

In [12], a glass plane, which is mounted in front a camera, is used to calculate the distance to objects. Changing the pose of the planar plane changes the image position of the projected object because of the light ray's incident angle, the plane's thickness and refraction index. This leads to a parallax effect which is needed for stereo vision, even though the camera is not translated.

The idea of actually translating a single camera has already been presented in [13] and [14]. In both systems, the camera is displaced to get two images from different positions, which allows single-camera stereo triangulation as proposed in this paper. The main difference to our approach though is that the camera is fixated and translated mechanically.

VI. Conclusions and Future Work

In this paper, we presented our implementation of a single-camera stereo vision system using a state-of-the-art mobile phone which is able to measure the distance to everyday objects. It can be used with mobile phones which are equipped with just one camera by determining the distance between the camera positions, from where the two pictures are taken, with inertial sensors. We conducted indoor and outdoor experiments with different distances and baselines. The results showed that an increasing baseline leads to a higher distance accuracy in most cases, but we also observed that lighting conditions and unintentional device rotations influenced the results.

References

[1] D. Murray and J. J. Little, "Using real-time stereo vision for mobile robot navigation," *Autonomous Robots*, vol. 8, no. 2, pp. 161–171, 2000.

[2] A. S. Ogale and Y. Aloimonos, "Shape and the stereo correspondence problem," *Int. J. Computer Vision*, vol. 65, no. 3, pp. 147–162, Dec. 2005.

[3] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 993–1008, Aug. 2003.

[4] B. Cyganek and J. P. Siebert, *An Introduction to 3D Computer Vision Techniques and Algorithms*. Wiley, Mar. 2009.

[5] D. Scharstein and R. S. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Computer Vision*, vol. 47, no. 1-3, pp. 7–42, Apr. 2002.

[6] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2007.

[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[8] N. O'Brien and R. Jain, "Axial motion stereo," in *Proceedings of the IEEE Workshop on Computer Vision, Representation and Control*, 1984, pp. 88–92.

[9] K. Seifert and O. Camacho, *Implementing Positioning Algorithms Using Accelerometers. Application Note AN3397*, Freescale Semiconductor, Feb. 2007.

[10] J. Larson, M. Bruch, R. Halterman, and J. Rogers, "Advances in autonomous obstacle avoidance for unmanned surface vehicles," in *Proceedings of the AUVSI Unmanned Systems Conf.*, 2002.

[11] A. Wedel, U. Franke, J. Klappstein, T. Brox, and D. Cremers, "Realtime depth estimation and obstacle detection from monocular video," in *Proc. of the DAGM Symp. on Pattern Recognition*, 2006, pp. 475–484.

[12] C. Y. Gao and N. Ahuja, "Single camera stereo using planar parallel plate," in *Proc. of the Int. Conf. on Pattern Recognition (ICPR)*, 2004, pp. IV: 108–111.

[13] Y. L. Murphey, Y. L. Murphey, J. Chen, J. Crossman, and J. Zhang, "Depthfinder, a real-time depth detection system for aided driving," in *Proc. of the Intelligent Vehicles (IV) Symp.*, 2000, pp. 122–127.

[14] J. Yu, J. Song, and X. Sun, "Design and experiment of distance measuring system with single camera for picking robot," in *Proc. of the World Automation Congress (WAC)*, 2010, pp. 445–448.