Performance Characterization of a High-Speed Stereovision Sensor for Acquisition of Time-Varying 3D Shapes

Y. Xiao, R.B. Fisher*, M. Oscar

School of Informatics, University of Edinburgh

Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, Scotland, UK

Tel: +44 131 651 5645, Fax: +44 131 651 6899,

Corresponding author, E-mail: rbf@inf.ed.ac.uk

Abstract: Acquisition of dynamic dense 3D shape data is of increasing importance in computer vision with applications in various disciplines. In this paper we investigate the performance of a high-speed range sensor based on stereovision principle for 3D shape acquisition of animals. The investigation reveals some characteristics of the current version of the sensor with respect to its physical parameters, which suggest a more appropriate configuration of the sensor in real data acquisition. Due to the novelty of the sensor and the application, we believe that our evaluation of the sensor's performance will inspire new applications to follow using the dynamic 3D acquisition technology of similar type.

Keywords: 3D shape acquisition, stereo vision, range sensor, dynamic scene, performance evaluation

1. Introduction

The past decade has witnessed a significant advancement in 3D dynamic shape acquisition technology[1-5], however, 3D sensing of fast moving objects is still relatively unexplored area []. In the European Commission project "ChiRoPing: Developing Versatile and Robust perception using Sonar Systems that Integrate Active Sensing, Morphology and Behaviour"[6], an interest arises in studying 3D external morphology of bats in flight in relation to their echolocation behaviour. To this end, it is required to collect time-varying 3D shape data of bats flying at speeds 3-5m/s when they perform particular tasks such as capturing prey.

A custom-designed high-speed stereovision range sensor[1] is employed for the 3D shape acquisition. The sensor is capable of acquiring synchronized stereo image

[1] The high-speed stereo range sensor was built by a custom design by Dimensional Imaging Ltd [7] in which R.B. Fisher is a minor shareholder, but is now commercially available.

sequences at a speed up to 500 fps (frames per second). The stereo image sequences are processed off-line to produce time-varying range images and associated textures representing the dynamic 3D shape and appearance of the target. Since acquisition of 3D shapes from objects in high motion is still at very early stage in research, very little information are now known about how range sensors perform at high speeds. In order to deploy and utilise our stereovision range sensor properly, we carried out a study to investigate the capabilities and limitations of the sensor. We believe that the results from our investigation will stimulate new applications and studies of high-speed 3D/range sensors based on vision principles.

2. Related Work

2.1 3D Dynamic Shape Acquisition

Although extensively studied in the past three decades, 3D (surface) shape acquisition still maintains a strong interest of the computer vision community. Traditionally, 3D shape acquisition was achieved in a point-by-point or line-by-line fashion. In order to obtain data from the whole surface of the object to be measured, a mechanism is employed to manoeuvre the position or the scanline where the data was acquired. The employment of such a scanning mechanism imposes certain constraints in applications of 3D shape acquisition. For instance, the object to be scanned has to remain static during the scanning process, otherwise spatial consistency of data collected from the object cannot be guaranteed. With the recent improvement of digital imaging technology, it becomes increasingly common to apply 2D imaging sensors (e.g., CCD) in 3D surface shape acquisition. The benefit of using 2D imaging sensors is that a 2D array of data can be all acquired at a time, which significantly reduces the time of measuring a surface, making it possible to record dynamic 3D surface shapes.

Most 3D surface shape acquisition methods using 2D imaging sensors are based on triangulation principle. In a triangulation set-up, two devices (one of them must be a light-receiver, e.g. a camera, and the other could be a light-receiver or a light-emitter such as projector, laser emitter, etc.) are placed apart to form a baseline. The object to be measured is placed in front of the baseline. The depths of points on the

object surface can be calculated by intersecting the lights from the object surface projected onto the two light-receivers (stereo vision) or intersecting the lights from the object surface projected onto the light-receiver and the lights emitted to the object surface from the light-emitter (structured light). Compared with some other 3D surface shape acquisition methods using 2D imaging sensors such as photometric stereo [8], modulated light [9], triangulation-based methods provide a larger range of measuring depth [10], making it appropriate to dynamic shape acquisition since the moving object usually occupy a larger 3D space than its static counterpart.

The amount of research reported in the literature on stereo vision and structured light is huge. The core problem in stereo vision is to establish correspondence between pixels in the two images captured by the two cameras. Due to data deficiencies such as image noise, occlusion, surface discontinuities in the 3D scene, light reflections, repeated texture patterns, etc., establishing correspondence (or stereo matching) in stereo vision is a highly ill-posed problem. Despite an enormous amount of research published, a universally-accepted approach has not yet been established and the methods proposed in the literature all have their strengths and limitations [11]. Nevertheless, success has been reported to recover 3D shapes of moving objects in controlled environments. For instance, [12] reports a digital TV studio which is designed to acquire dynamic 3D data of a human actor performing a task. 24 video cameras are deployed to collect images from the subject in 8 different orientations, and dense stereo matching is applied to obtain range data from the 2D images. Finally the range data are merged using ICP registration [13] and re-organized into 3D mesh format using marching cube techniques [14]. This work demonstrated valid dynamic 3D surface acquisition at 25 fps, and inspired the application of dense stereo matching in 3D acquisition of dynamically free-formed shapes such as human bodies [7]. On the other hand, since stereo matching is an ill-posed problem, hybrid methods combining stereo vision with other methods such as voxel carving [15], morphable model fitting [16], template factorization [17] were proposed to improve the quality of 3D surface reconstruction.

Structured light methods establish correspondence by coding the spatial relationship of illumination from the light-emitter and decoding them on the imaging plane from

the light-receiver. In principle, the complexity of establishing correspondence in structured light methods is much smaller than that in stereo vision, resulting in more consistent and rapid 3D acquisition. However, the current optical technology is not sufficient to produce coded illumination at spatial and temporal resolutions comparable to those of digital imaging devices, creating a bottleneck for the application of structured light sensors. A very recent article in structured light [18] reported a 3D recording in resolution of 532X500 points per frame at 40 fps, which is still far below what has been achieved by stereovision sensors. Moreover, the use of extra illumination (structured light) sometimes distracts the subject (the object to be measured) limiting its applications.

The sensor to investigate in this paper is built on binocular stereovision principle. Due to the passive style of the sensor, it generates the least distraction to the subject in the particular application of bat study. With the recording speed of 500 fps and the spatial resolution 1280X1024, it is the state-of-the-art in dynamic 3D surface imaging to the author's best knowledge. This paper addresses the evaluation of the performance of the sensor.

2.2 Performance Evaluation of Stereovision Sensors

A stereovision sensor is a complex system comprising optics, electronics, mechanics and computing components. Such complexity makes it difficult to evaluate performance of a stereovision system as a whole. However, there are reports investigating performance of some particular components of a stereo vision system or analyse errors from some particular aspects in stereovision. [11] provides a comparative study of performance of dense stereo matching algorithms with quantitative evaluation results. [n1] discusses errors in 3D reconstruction of points from two views given known models of quantization errors on the image planes. A stochastic study of the 3D point reconstruction errors can be found in [n2], where a close-form solution to the error distribution is given with respect to baseline distance, focal length, and image sampling interval. Error analysis of 3D reconstruction of line segments [n3], quadratic curves [n4] has also been conducted using similar methods.

A comparative study of performance of dense stereo algorithms can be found in [11], and some preliminary quantitative evaluation results of the algorithms are provided. However, the evaluation results at algorithm level are hard to apply to a whole system for 3D acquisition which has to have some amount of engineering constraints.

In the medical field, accuracy [20-21], reproducibility [20] of 3D measurements and field of view [21] of stereo range sensors have been reported on acquisition of 3D data from human dummy heads. In the ChiRoPing project, the subjects are flying bats, which makes the method of applying dummies rather impractical due to the high cost in producing realistic bat dummies and simulating their motion in 3D. In a more affordable way instead, we propose to characterize performance of our stereo range sensor using a few artificial objects with representative shapes placed and manipulated with controlled motion in the capturing windows designed to acquire data from real bats.

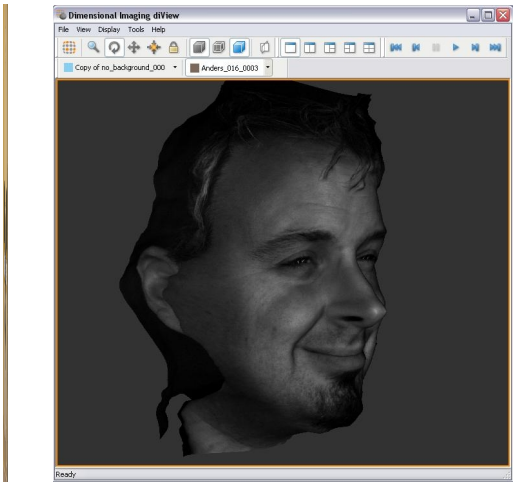
3. System Description and Configuration

3.1 System Description

The high-speed stereovision range sensor is manufactured by Dimensional Imaging Ltd [7]. The hardware of the sensor mainly comprises two Mikrotron^{TM} high-speed monochrome cameras, two infrared lights and two processing computers. The monochrome cameras are chosen to reduce data capacity therefore allowing higher frame rate in data acquisition given the same system bandwidth. The cameras are mounted on a rigid metal bar to form a stereo rig (see Fig 1a). The distance between the cameras can be adjusted to suit 3D capture of different scenes. Specially designed cables, along with frame grabbers, allow image capture up to 500 fps. The infrared lights are used to illuminate the acquisition scene. The infrared wavelength is carefully selected to overlap the visibility spectrum of the cameras and illuminate the acquisition scene without disturbing the bats. The computers are paired to receive, store and process raw intensity images captured by the stereo cameras. They share buffers so that the data can be processed in parallel. The frame grabbers

in the computers are synchronized externally through a synchronization cable. The synchronization is required when recording stereo images for 3D shape recovery.

The software of the stereo sensor consists of three major modules: image capture, 3D reconstruction, and 3D viewing. The image capture module allows users to trigger intensity image capture simultaneously for the two stereo cameras. Once the trigger button in the image capture module is activated, the cameras start to collect synchronized stereo images. The captured images are processed by the 3D reconstruction module. The outputs are 3D range images and associated texture maps, which can be visualized using the 3D viewing module (see Fig. 1b).



(a) Stereo rig (b) 3D viewer

Fig. 1. The high-speed 3D acquisition system

3.2 Acquisition Setup

Two groups of bats (insect gleaning and water trawling) are planned for study in the ChiRoPing project, for each of which an acquisition scenario has been considered. An insect gleaning bat usually hovers in front of a prey on a leaf for a few seconds before performing a capture. In this scenario we set up the stereo rig in a small bush behind the leaf on which prey is placed. When the bat is hovering within the working range of the stereo cameras, the cameras will be triggered to record stereo images of the bat. The distance between the bat and the stereo rig is expected to be 80cm. To suit this acquisition scenario, Fujinon CF50HA-1 50mm lenses are chosen. At the working distance of 80cm, a single CF50HA-1 lens allows a capture window of 20cm X 30cm which is about 2-3 times bigger than the insect gleaning

bat. The other acquisition scenario is for water trawling bats. The working distance is expected to be 2m in this scenario. Fujinon CF75HA-1 75mm lenses are chosen. At the working distance of 2m, a single CF75HA-1 lens allows a capture window of 30cm X 45cm which suits the bigger size of water trawling bats.

4. Performance Evaluation

It was reported in two previous studies [20-21] that errors of 3D measurement using stereo range sensors were less than 1mm on average for sparse landmark points in a working volume for static human face capture. However, the studies did not link the measurement accuracy to the system parameters of the sensors, which leads to a question whether or not the results of 3D measurement accuracy in these studies can be applicable to a more general context. Moreover, it is unknown if the accuracy obtained from a few sparse points is representative of the entire dense acquisition of 3D points on the object.

This paper investigates experimentally how the stereo sensor employed in ChiRoPing performs in various conditions. Compared with previous studies [20-21], the methodology in this research has three distinctive features. First, it varies system parameters of the stereo sensor. The aim is to find optimal configurations of the sensor for the two scenarios of capturing data from flying bats (see section 3.2). Second, the measurement errors are obtained directly from dense representation of the data instead of a few sparse points. We believe such errors are more representative of the object surface to be measured than those from a sparse sampling. Third, object motion is considered in the experiments. By varying the object motion, characteristics of the sensor related to dynamic acquisition can be revealed.

We categorize our experiments into two groups: static tests and dynamic tests. The static tests aim to reveal properties of the sensor when the test objects are still during data acquisition. System parameters of the sensor (including focal length, aperture, baseline length, and converging distance) and object appearance in terms of shape and texture are varied in these tests. On the other hand, the dynamic tests

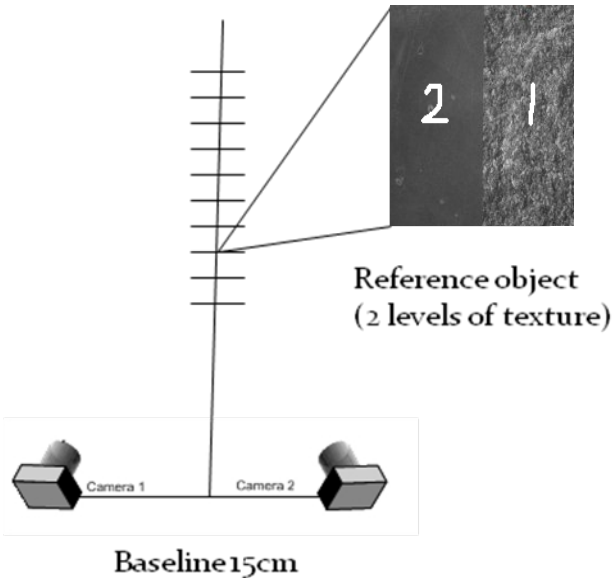
focus on characteristics of the sensor related to object motion. Various velocities of a test object will be applied in the experiments.

4.1 Static Tests

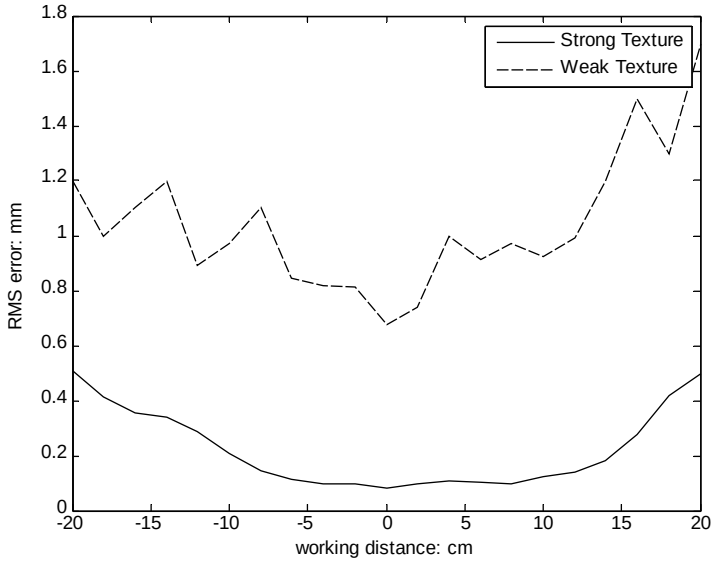
4.1.1 Working range

A stereovision sensor has a working volume that is observed by both cameras of the sensor. One of the major concerns in the study is how well the sensor would perform in terms of acquisition accuracy in the working volume. To answer this question, we tested the sensor with reference objects of known shapes placed at different locations within the working volume. The quality of the 3D range data acquired by the sensor at those locations can explicitly indicate the performance of the sensor in the working volume. By doing that, we could also identify a working range of the sensor in which sensible range data can be derived.

Our first reference object is a rigid planar surface (see Fig. 2(a)). The planar shape can be used as ground truth. When we acquire a range map of the object, we examine the variation of the range data against a plane, and then we know how well the sensor perform 3D acquisition of the planar surface.



(a) a planar reference object (2 levels of texture)



(b) RMS errors around working distance 120cm

Figure 2 Working range test. (a) shows the reference object in the test which has two levels of texture; (b) demonstrates RMS of residuals (in mm) after fitting a plane to 3D images of the reference object.

The original surface of the object has a texture of natural paint. Under the infrared light of the scanner, the original texture looks quite "weak" (which means that fine details in the texture are not in good contrast) in the intensity images captured by the cameras. To enhance the texture, we artificially imposed a "strong" texture on some parts of the object surface by attaching printed texture to the surface, thus we have two levels of texture (see Fig. 2(a), weak and strong texture are labeled as "1" and "2" respectively) and are able to get a comparison on how the scanner performs on different levels of texture.

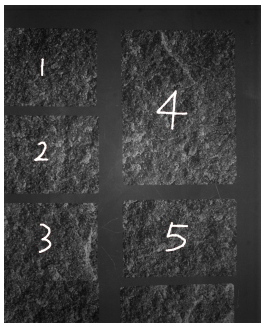
The first experiment was carried out at working distances between 100cm and 140cm, and the cameras were focused on and converged to a point at working distance 120cm. A working distance for the stereovision sensor is defined as the distance between the object and the baseline of the cameras. The reference object was placed to face the cameras with its normal perpendicular to the baseline and the camera imaging planes as closely as possible (Fig. 2(a)). In such a configuration, the planar surface at a testing position has only nearly a single depth to the sensor.

At each working distance of every 2cm in the range [100cm,140cm], a range image of the reference object was acquired. 2 regions (one with strong texture and one with weak texture) were manually selected from the range map for examination. The data points for each selected region in each range image were converted to a 3D point cloud and then fitted with a 3D plane. Standard deviation or RMS (Root Mean Square) of the fitting residuals was calculated. The RMS error characterizes the variation of the 3D data in the region, which avoids the bias towards a few points selected as landmarks in the previous studies [20-21].

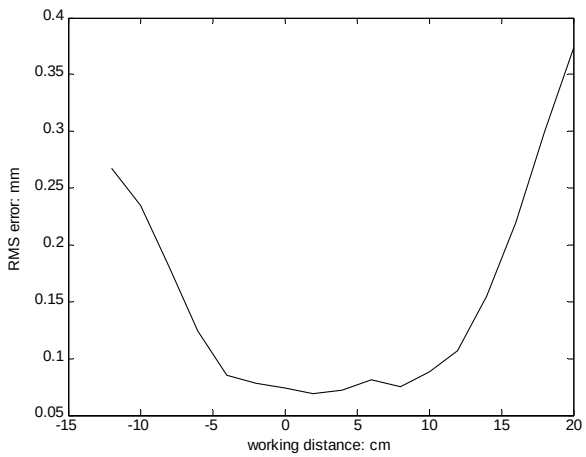
The results of the experiment are depicted in Fig. 2(b). The distance of 0cm corresponds to the working distance of 120cm, which is the center of the examined working distances. It can be seen that the RMS errors for the strongly textured region (region 1) exhibit a smooth basin shape with a flat bottom between -6cm and 8cm with a magnitude of 0.1mm. In comparison, the RMS errors for the weakly textured region (region 2) at the same distances are clearly larger with a magnitude of 0.9mm and also much more fluctuated. The RMS errors for both regions grow rapidly when working distance exceeds range [-8cm, 12cm]. The result suggests that the stereovision sensor in its current version is not able to capture quality 3D shapes for weakly textured surfaces.

The length of the baseline is 15cm and focal lengths of the camera lenses are both 50mm in the above test. With the same baseline length, we also tested the working distance range around 80cm and 200cm, since these are two most probable working distances for real acquisition of bats' 3D shapes. 50mm lenses were selected for working distance of 80cm and 75mm lenses were selected for working distance of 200cm. Only strongly textured regions were examined this time, since 3D acquisition for weakly textured regions was too noisy to characterize the sensor's performance (see Fig. 2(b)). The strongly textured regions are distributed in different parts of the object surface (Fig. 2(a)). RMS errors for all 5 regions were calculated and their medians were chosen to represent the sensor's performance. The results are depicted in Fig. 3. It can be seen that the RMS errors around working distance 80cm exhibit a fluctuating but fairly flat bottom in the range [-4cm,8cm] with a magnitude below 0.1mm. The flat bottom of the basin looks similar to that in Fig. 2(b), which reflects the influence of the depth of field of the

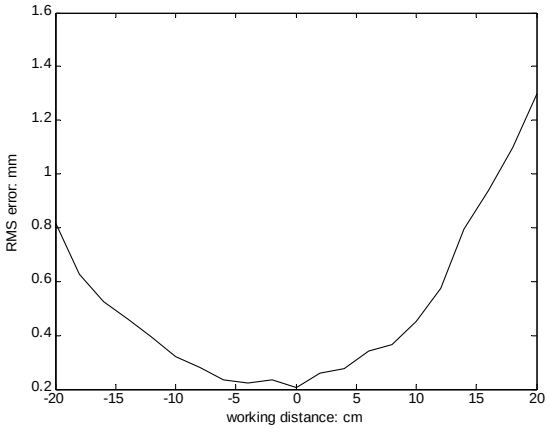
lenses. The RMS errors around 200cm (Fig. 3(c)) are at higher levels compared to those around 80cm with lowest value about 0.2mm, and the shape of the RMS errors looks more rounded at the bottom. The higher level of RMS errors around 200cm is due to larger error propagation in triangulation (more explanation in Section 4.1.4). Despite some small random fluctuation in the error curves in Fig. 3(b-c) due to manual intervention in the experiments, the curves are characteristic for the two capture scenarios in ChiRoPing. They suggest that we may have fairly good 3D acquisition in the depth range about [10-20cm] for the working distances 80cm and 200cm.



(a) Reference object (5 strongly textured regions)



(b) RMS errors around working distance 80cm (with 50mm lenses)



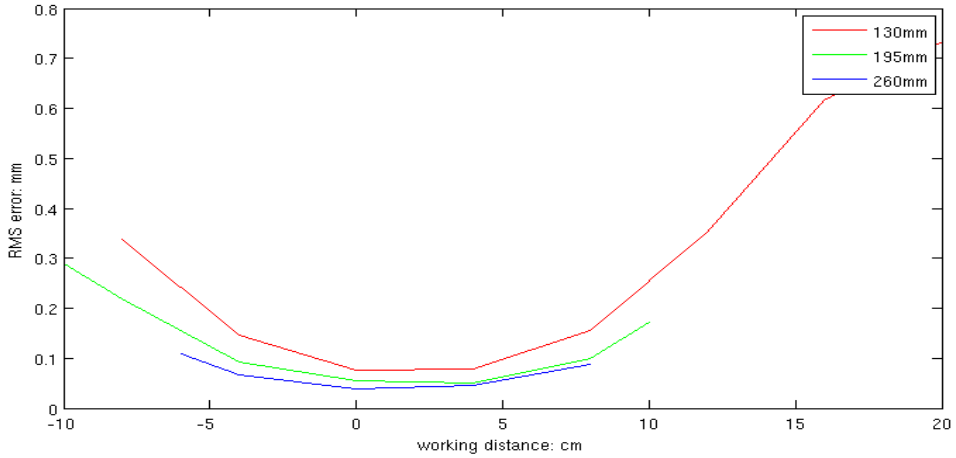
(c) RMS errors around working distance 200cm (with 75mm lenses)

Figure 3 RMS errors after fitting a plane to data points in selected regions 1-5 in 3D images of the reference surface

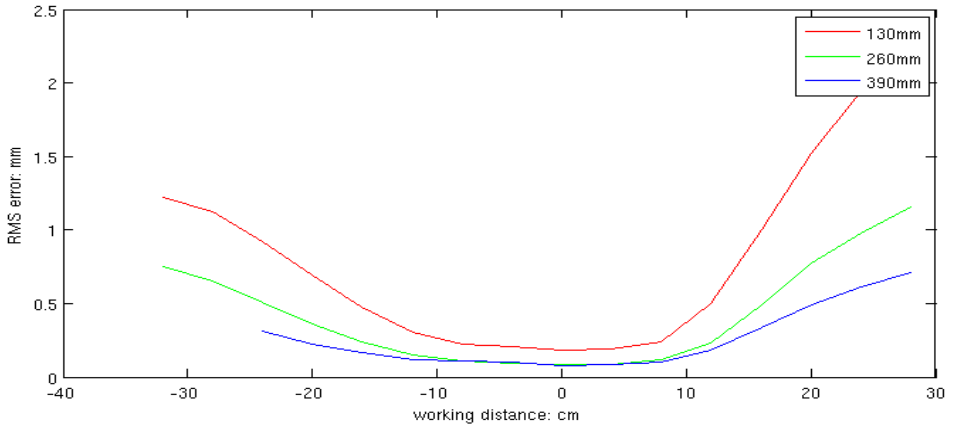
4.1.2 Baseline length

Another question about the stereovision sensor is how the length of the baseline of the sensor affects the accuracy of 3D measurements. To answer the question, we evaluated the sensor at 3 different baseline lengths for each working distance for the real data capture scenarios: baseline lengths 13cm, 19.5cm and 26cm for working distance 80cm and 13cm, 26cm and 39cm for working distance 200cm. The reference object and the evaluation method are the same as those in Fig. 3. The results with respect to baseline length are displayed in Fig. 4. It can be seen clearly that wider baselines generate lower RMS errors. However, the valid range of working distance for the sensor (in which the sensor is able to produce valid 3D measurements) may become shorter when the baseline is longer. For instance, for 50mm lenses with baseline length 260mm, the 3D measurements were observed with low RMS errors (below 0.1mm, see the blue curve in Fig. 3(a)) in the range [-6cm,8cm]. When the working distance exceeded the range, the scanner was not able to output valid 3D images representing the test object shape. In comparison, the 195mm baseline could allow valid 3D measurement in the range [-10cm, 10cm] with slightly increased RMS errors and the 130mm baseline can even achieve longer working range of [-8cm, 16cm] though the price is the even higher level of RMS errors. The experiment with the 75mm lenses confirms that wider baselines produce smaller RMS errors but result in shorter working ranges. The results indicate that the length of baseline is an important factor for the stereovision system

and it should be selected to reflect the balance of the 3D measurement accuracy and the valid working range. Our experiments show that 195mm baseline for 80cm working distance with 50mm lenses and 260mm baseline for 200cm working distance with 75mm lenses are good choices to have a good working range and yet still maintain a low level of measurement errors.



(a) RMS errors of different baseline lengths around working distance 80cm (with 50mm lenses)



(b)RMS errors of different baseline lengths around working distance 200cm (with 75mm lenses)

Figure 4 RMS errors measured in 3D images obtained at different working distances for different baselines.

4.1.3 Aperture

The aperture of a lens determines the amount of light that transmits to the camera image plane. In this experiment, we vary apertures of lenses of the stereovision sensor to examine how this affects 3D acquisition.

Fig. 5 illustrates error curves for three different apertures (denoted by F-stop numbers) of the 75mm lenses obtained in working range around 200cm with baseline 26cm. We tested only three apertures (F2.8, F4, F5.6) because the other F-stop numbers either render the images too dark or too bright to perform valid stereovision sensing. The same reference object and performance evaluation method were used as in Section 4.1.2.

It can be seen in those curves that F2.8 produces the most narrow RMS basin which has only about 15cm width, and nevertheless it achieves the lowest RMS error in the entire tests. F4 produces a wider bottom (a bit more fluctuating than that of F2.8 though) with slightly higher RMS value. However, the RMS errors of F4 go up at a rate much less than those of F2.8 when working distances exceed the range of the basin bottom. F5.6 generates a RMS error curve even less steep than that of F4 with RMS errors higher than those of F4 at basin bottom.

The explanation for the RMS curves in Fig. 5 is twofold. Firstly a larger aperture (lower F-stop number) generates a smaller depth of field, which results in a more narrow basin bottom in the RMS error curve, as illustrated in Fig.5. Secondly a larger aperture allows a large amount of light to enter the cameras and consequently enhance the contrast of the texture of the object, which improves the accuracy of stereovision sensing. That explains why F2.8 achieves the lowest RMS error. The quantitative analysis of the aperture effect will be discussed in Section 4.1.4.

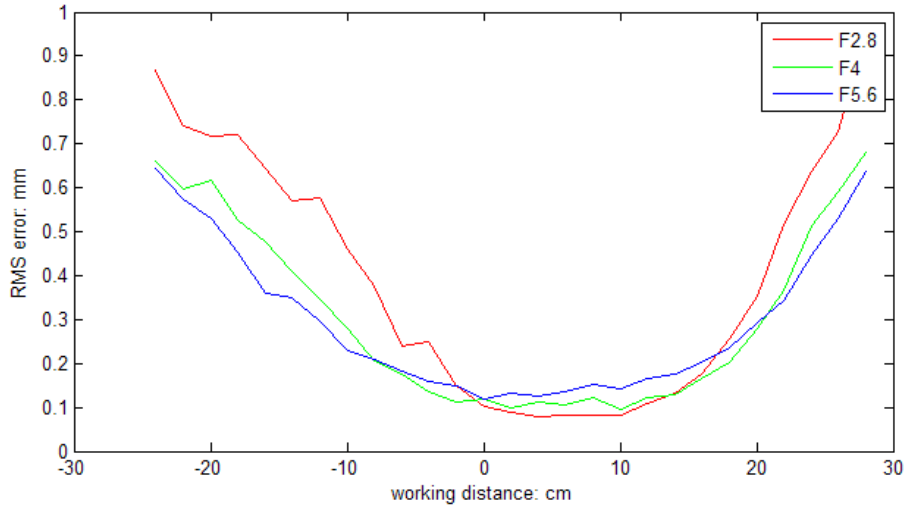


Figure 5 RMS errors with different apertures in working range around 200cm

4.1.4 Quantitative analysis

Errors of 3D measurement in stereovision sensing are the result of stereo matching errors and calibration errors, which propagate through the triangulation process. Ideally the distribution of 3D errors can be computed analytically given a known stereovison configuration and the distributions of stereo matching errors and calibration errors [n1-4]. However, for the stereovision sensor studied in this paper, it is extremely challenging to model the stereo matching errors sufficiently to allow valid calculation of the 3D error distribution, due to uncontrolled uncertainties in imaging condition, image formation and the underlying stereo matching algorithm. Furthermore the performance of stereo matching can be affected by system parameters of the stereovision sensor. For instance, a wider baseline will create more image distortion between the left and right view, consequently increasing stereo matching errors. While we accept that 3D errors from our stereovision sensor cannot be fully predicted, the evaluation results conducted in Section 4.1.1-3 at a few system parameters of the sensor can be interpreted using analytics at least partially.

From the results in the baseline tests in Fig. 4, it can be seen that the bottoms of the error curves in each test have nearly the same width despite the different baseline lengths for the sensor. Since the settings of the lenses and the working distance are the same in a baseline test, we hypothesize that the basin shapes in Fig. 4 are mainly caused by the out-of-focus blur of the lenses. Assuming a thin lens model (which suits the lenses for our stereovision sensor), the degree of out-of-focus blur can be calculated using the following formula []:

$$c = \frac{f^2 \left| z_{of} \right|}{F(z_f - f)(z_f + z_{of})} \tag{1}$$

where c stands for the diameter of the circle of confusion, F denotes F-stop number of the lens, f is the focal length, z_f represents the distance between the lens and its focus point , z_{of} denotes the distance from the object to the focus point. Given F , f , z_f, the relation between c and z_{of} typically exhibits a V-shape. Fig. 6 depicts such a relation for the lenses used in the test in Fig. 4(b). Note that c has been converted to pixels for the convenience of study.

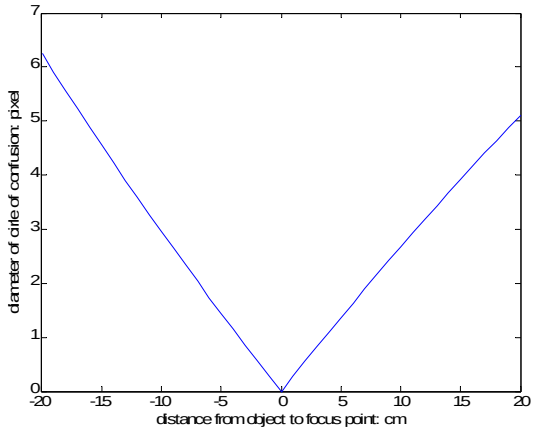


Figure 6 Relation between circle of confusion and object distance for lenses used in Fig. 4(b).

If we register the V-shape in Fig.6 to the U-shapes in Fig. 4(b), we can see that the bottoms of the U-shapes in Fig.4(b) are located around the working distance range [-10cm,+10cm], which corresponds to the object-to-distance range where diameter of circle of confusion is smaller than 3 pixels as shown in Fig.6. Since we know the disparity produced by the stereo matching algorithm employed for the sensor is measured by pixels, Fig.4(b) implies that 3 pixels is about the minimum size of texture the stereo matching algorithm can resolve. The 3-pixel ambiguity zone creates a flat bottom in the error curves in Fig.4(b) in which the 3D errors are averaged.

This hypothesis has been confirmed by the results in the aperture tests in Fig. 5. We calculated the 3-pixel ranges for the lenses at different apertures used in Fig.5 using Formula (1), and we measured the bottoms of the error curves in Fig. 5. A bottom of an error curve is defined as the part of the curve whose error values are less than 2 times of the minimum value in the whole error curve. The widths of the bottoms in Fig. 5 and those of the corresponding 3-pixel ranges are listed in Table 1. It can be seen that the corresponding error curve bottoms and 3-pixel ranges are highly correlated, which is a strong evidence to our hypothesis that the basin shape of 3D errors shown in Section 4.1.1-3 are mainly caused by the out-of-focus blur of the lenses used. Figures in Table 1 are also consistent with the observation that the stereovision sensor has texture resolvability of about 3 pixels. Given a new configuration of the sensor, we can calculate out-of-focus blur of the lenses using

formula (1) and then estimate the working range of the sensor based on the 3-pixel texture resolvability assumption.

Table 1 Comparison of widths of 3-pixel ranges for lenses in Fig.5 and widths of the corresponding RMS error curve bottoms.

	F2.8	F4	F5.6
Width of 3 pixel range (cm)	14.969	21.416	30.070
Width of error bottom (cm)	17.459	24.417	38.462

The above analysis correlated the out-of-focus blur and the 3D reconstruction errors, however, it should be noted that baseline and working distance have an effect on 3D errors too. In fact, in stereo vision, the following relation between disparity errors and errors in depth exists [n5]:

$$\sigma_z = \frac{z^2}{bf}\sigma_d \tag{2}$$

where σ_z stands for the standard deviation of depth errors, σ_d denotes the standard deviation of disparity errors, z represents the depth value of a 3D point, b, f are baseline and focal length of the lenses. Given the same σ_d , z, f, σ_z is inversely proportional to b. This explains the slopes of the sides of the U-shape curves in Fig. 4, where shorter baselines clearly generate steeper downward/upward slopes before/after the focus point. Therefore ideally the baseline should be as wide as possible. However, when the baseline gets wider, the geometric distortion between the left and right views increases as well, which results in shorter working range as shown in Fig. 4 and increased disparity errors. The increased disparity errors are probably the reason why the bottoms of the error curves for baseline 260mm and 390mm in Fig. 4(b) are at the same magnitude level.

It is also worth noting that the out-of-focus curve in Fig. 6 has a larger slope before the focus point and a smaller slop after the focus point, which is opposite to the U-shape curves in Fig. 4(b) where the slopes after the focus points are larger than before the focus point. The reason is due to the larger error propagation (expressed in Formula (2)) when the working distance gets longer. Let us assume the level of disparity errors is proportional to the diameter of circle of confusion, i.e., $c \sim \sigma_d$, then we use Formula (1) and (2) to compute the depth error for a particular working distance. The working distance is assumed to be the same as the depth value $z = z_{of}$

+ z_f. Since we placed our planar reference object perpendicular to the optical axes of the cameras, this assumption holds closely true. Fig. 7 illustrates the relation between depth error and working distance for the stereovision settings in Fig.4(b). It can be seen the slopes of the depth error curves are larger after the focus point than before the focus point.

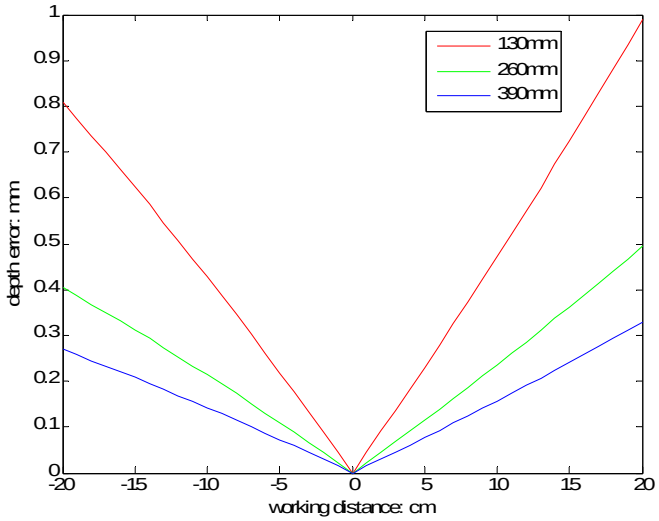


Figure 7 Relation between depth error and working distance for the stereovision settings in Fig.4(b), c=0.03σ_d is applied.

The above analysis links the relation between working distance and circle of confusion (Formula (1)), the relation between disparity error and depth error (Formula (2)), and the experiment results in Section 4.1.1-3. The uncertainty that prevents a complete analytical prediction of 3D errors lies in the relation between circle of confusion and disparity errors, i.e., c and σ_d , which depends heavily on the stereo matching algorithm and also on various factors such as illumination, texture, image noise, etc. The investigation how these factors affect the relation between c and σ_d is still an open question and beyond the scope of this paper.

4.1.5 Minimum Resolvable Features

The analysis in Section 4.1.4 revealed that the stereovision sensor is not able to resolve texture in 2D images smaller than 3 pixels. This section investigates further how well the sensor resolve 3D shape details. We tested the sensor with two types of 3D shapes: thread crosses (Fig.8(a)) and paper triangles (Fig.8(b)). Both types

have sharp shape features in the depth direction. A thread forms an impulse edge in the depth direction and a thread cross has a 3D saddle around the crossing point. A paper triangle generates three step edges in the depth direction and the edges intersect to form three corners. Our objective is to find out the minimum distances between the threads of a cross and between the sides of a paper triangle when they are still distinguishable in range data.

We set up the experiment at the working distance of 80cm with 13cm baseline length and 50mm lenses. To balance the experiment, two diameters of threads (0.5mm and 2.0mm) and two thicknesses of papers (0.4mm and 1.0mm) were used. The threads and the papers were stretched straight and placed tightly on top of their supporting planes (textured).

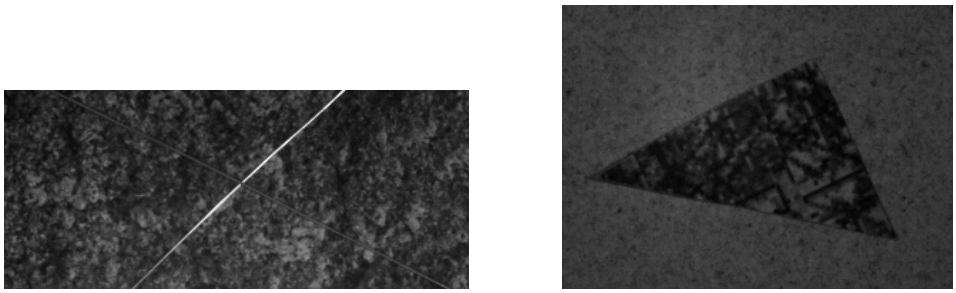
In the range data of the objects, it was found that the shapes of the threads and the paper edges are smoothed. In addition, the thread diameters and the paper thicknesses are too small compared to the working distance (80cm) so that the depth difference between the objects and the support planes is not easily noticeable in the range data. To visualize the 3D shape details of the objects, we applied a RANSAC method to find the supporting plane, then calculated the distances between the data points and the plane and normalized the distances to [-1,1] range. It can be seen that the threads and paper edges are clearly noticeable in the distance maps rendered in pseudo colors (left column in Fig.9 and 10).

We examined 1D profiles of the distance maps. A 1D profile is a slice (horizontal or vertical) of a distance map. It can be seen that impulse edges representing threads and step edges representing sides of paper triangles exist in the 1D profiles where the objects are present. In the critical 1D profiles near the thread crosses or the triangle corners, impulse edges or step edges start to merge together (Fig.9 (a) and (c), Fig.10 (a) and (c)). For each of the critical 1D profile, we measure the distance between the impulse edges or the step edges of the intensity image of the object. The distance is considered to be the minimum (horizontal or vertical) distance to distinguish the threads or triangle sides in range data.

Table 2 lists all the minimum distances we measured from data of the thread crosses and paper triangles in the experiment. The distances were measured initially in pixels from the intensity images. To give an idea about the scale of minimum shape details in 3D, we converted the distances to corresponding Euclidean distances in the world coordinates. The results suggest that the scale of minimum shape details to be distinguished in range data is about 10-15 pixels, which correspond to 2.0mm-3.0mm at working distance 80cm. The 10-15 pixels 3D shape resolvability and 3-pixel 2D texture resolvability have led us to think the stereo matching algorithm used in the sensor may have applied a smoothness constraint to produce disparity which results in smoothed 3D range data.

Table 2 Minimum distances between distinguishable shape details

	horizontal	vertical
Thin threads (diameter 0.5mm)	10 pixels/2.0mm	16 pixels/3.2mm
Thick threads (diameter 2.0mm)	27 pixels/5.4mm	23 pixels/4.6mm
Thin paper (thickness 0.4mm)	11 pixels/2.2mm	8 pixels/1.6mm
Thick paper (thickness 1.0mm)	12 pixels/2.4mm	6 pixels/1.2mm



(a) a thread cross (b) a paper triangle

Figure 8 Objects for minimum resolvable feature test

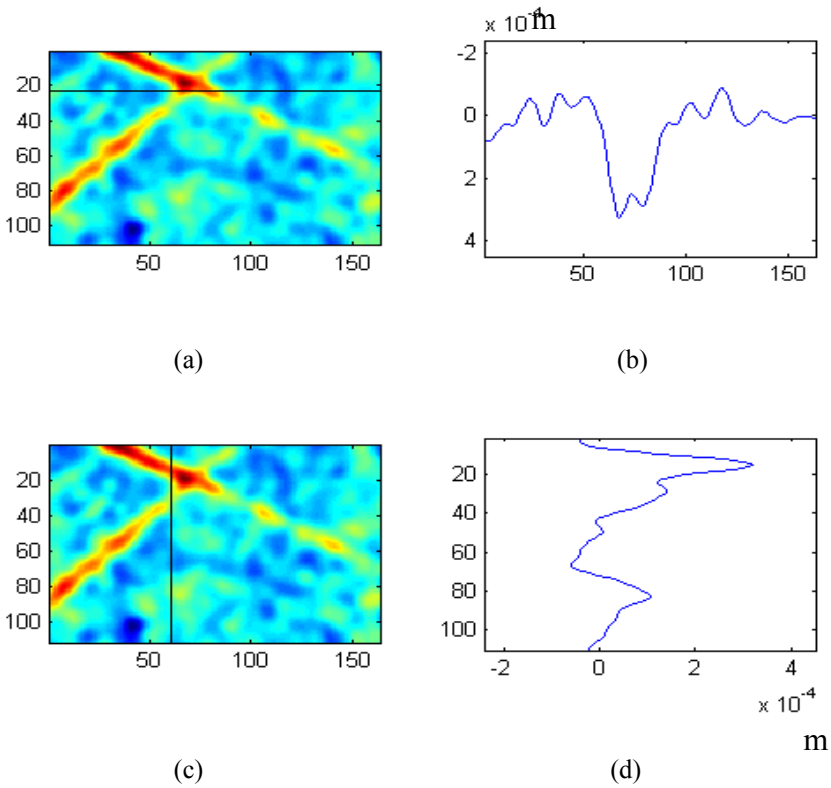


Figure 9 1D profiles of a thread cross (diameter 0.5mm): (a) distance map between the threads and the supporting plane; (b) profile of a critical horizontal slice of distance map (the location of the slice is depicted in (a)); (c) distance map between the threads and the supporting plane; (d) profile of a critical vertical slice of distance map (the location of the slice is depicted in (c));

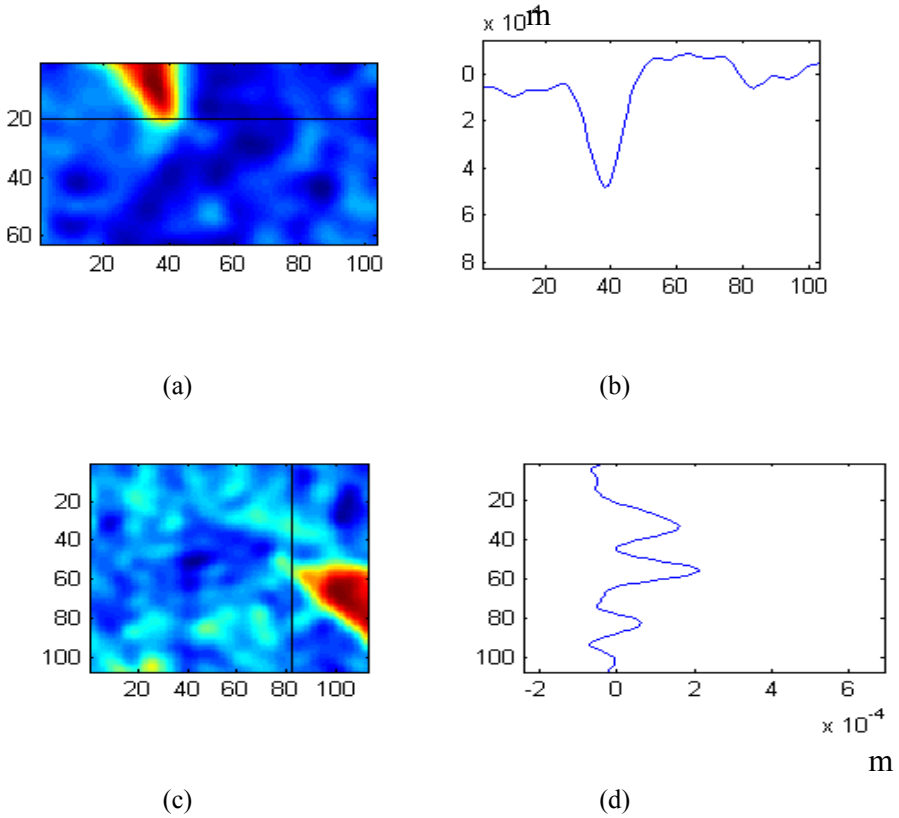
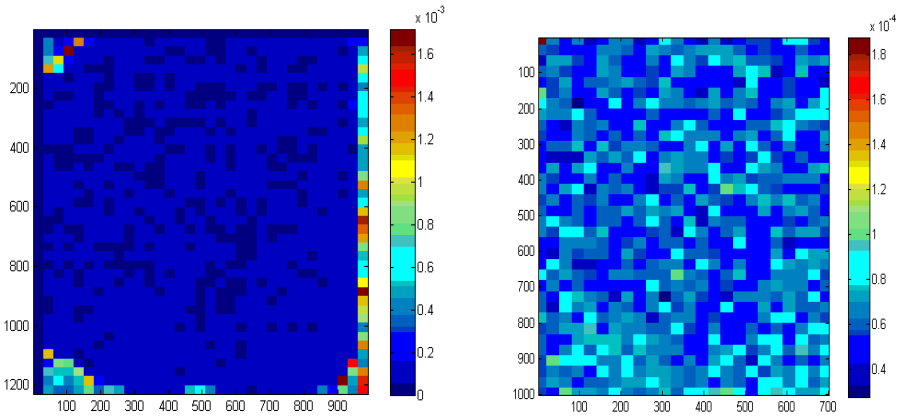


Figure 10 1D profiles of a paper triangle (thickness 0.4mm): (a) distance map between the triangle and the supporting plane; (b) profile of a critical horizontal slice of distance map (the location of the slice is depicted in (a)); (c) distance map between the triangle and the supporting plane; (d) profile of a critical vertical slice of distance map (the location of the slice is depicted in (c));

4.1.6 Homogeneity

One other question about the sensor is how the 3D acquisition noise varies spatially over the capture window. To answer this question, we tested the sensor with a bigger reference object that can cover the entire capture window. The object is well textured to minimize the noise level in 3D data. The range map obtained was partitioned to 30x30 patches and each patch was converted to a 3D point cloud and fitted with a 3D plane. RMS errors were calculated for each patch, and results are shown in Fig.11(a). It can be seen that the RMS errors remain stable over most part of the image plane apart from boundaries. We further examined the RMS errors in the central area of the image plane (where the RMS errors are stable) and it was found that the RMS errors fluctuate randomly in the central area and there is no obvious systematic pattern of the RMS errors (see Fig. 11(b)). The indication is that the 3D acquisition is homogenous in the capture window.



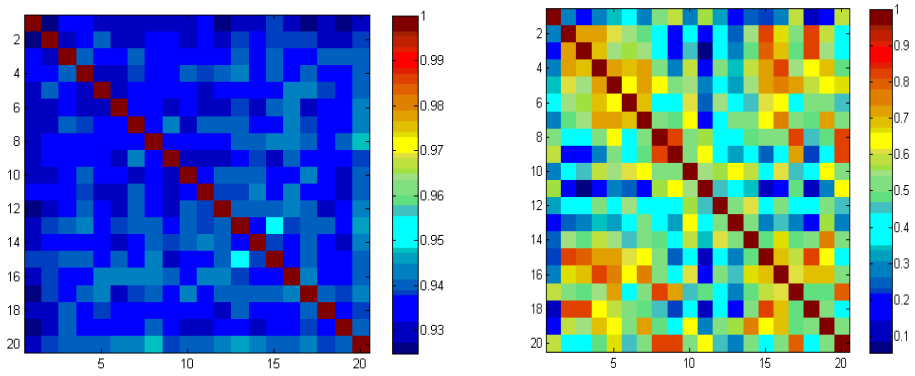
(a) noise distribution on the entire image plane (b) noise distribution in the central area of (a)

Figure 11 RMS errors of fitting a plane to 30x30 patches in a 3D image. (a) for the whole image; (b) for the central part of the image

4.1.7 Temporal Correlation

There is also a concern that 3D data captured in a sequence may be temporally correlated. To investigate into this question, we took a sequence of range maps of a static planar object. The object has two levels of texture (Fig.2(a)). Fitting a plane to 3D data in a frame of the sequence, we obtained a residual map. We fitted a plane to 3D data in the selected regions of the surface from both weak and strong textures respectively. We calculated the correlation matrix of the fitting residuals for the first 20 images of the sequence. An element C(i,j) in the matrix represents a cross-correlation factor between the fitting residuals in the i-th frame and j-th frame (Fig.12). It can be seen that the 3D data taken from strongly textured regions are highly correlated and those from weakly textured region are fairly random.

In the case of the strong texture, the RMS errors are quite low (around 0.1mm) in the 3D data, so we attribute the high correlations to the capability of the sensor to spot structural details on the object surface – these details being kept still during the acquisition. For the weak texture, there is larger magnitude of noise in the 3D data (RMS errors about 0.7mm) which override the surface details and cause correlation factors to get lower and more random.



(a) (b)

Figure 12 Correlation matrix computed from 20 consecutive frames of 3D images of a textured plane: (a) Correlation matrix for strong texture (non-diagonal elements within range [0.9249,0.9499]); (b) correlation for weak texture (non-diagonal elements within range [0.0547,0.8399]);

4.2 Dynamic Tests

The dynamic tests focus on the performance of the stereo range sensor related to object motion. Motion in 3D space can be categorized into translation and rotation. This paper only investigates the effect of translational motion, since the sensor is aimed to acquire 3D shapes from bats, which do not generate a large amount of rotational motion in the real world. We characterize translational motion in three directions: horizontal (x-direction), vertical (y-direction) and in depth (z-direction). The three directions are the implicit directions associated to range images output from the stereo sensor. x- and y- are the horizontal and vertical directions of the range data array and z- is the depth direction.

We employed a spherical reference object (Fig.13) in the dynamic tests. The reason for choosing a spherical shape is that the centre of the sphere can be estimated by using partial 3D shape data of the sphere. Knowing the centre positions of the sphere in 3D, the motion of the sphere can be calculated. The size of the sphere in Fig.13 is medium (of radius 17cm), which allows a good number of acquisitions when it is passing through the capture window and guarantees that the object appears big enough in the captured images to permit valid analysis of the object. To achieve the maximum number of acquisitions of the object in a sequence, the maximum capture speed of 500 fps is applied to the sensor.

4.2.1 Horizontal motion

To generate horizontal motion, the spherical object was swung across the capturing window. The object was attached to a fixed point by a string to form a pendulum. When the object reaches its lowest position, it generates the highest horizontal speed. The capturing window was placed to overlap the lowest position of the object in the pendulum so that the object could be captured at its maximum speed.

Range data of the swinging object was recorded and analyzed. RMS errors were calculated from the range data using the method similar to that for computing RMS errors of the planar object in the static tests. The range image of the object is first converted to a 3D point cloud and then fitted with a sphere. The standard deviation

of the fitting residuals (RMS errors) are used to evaluate the quality of the range data. A byproduct of computing the RMS errors is the estimation of centre positions of the spherical object. Knowing the centre positions of the spherical object, the motion of the object can be estimated. In this paper, we applied a backward difference operator to calculate the object speed.

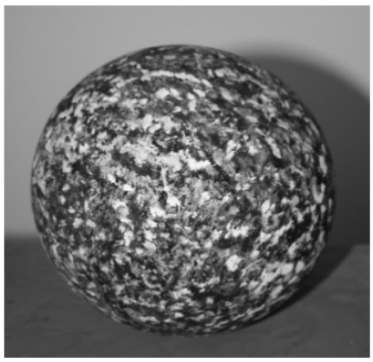


Fig. 13 Spherical reference object

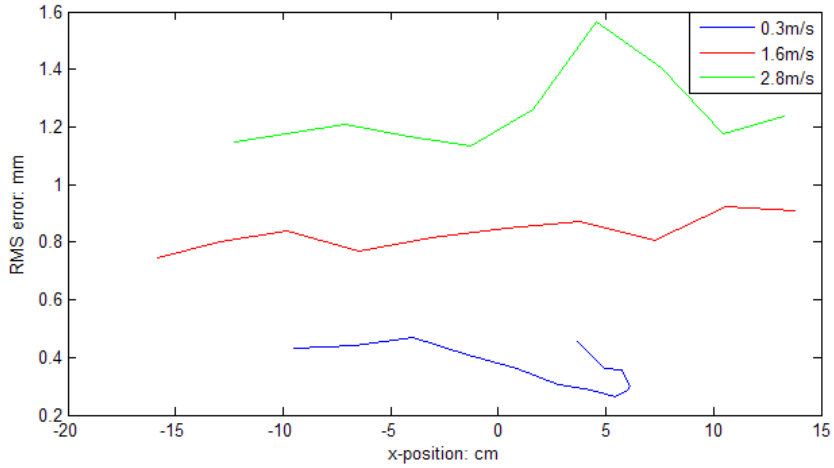


Fig. 14 RMS errors of fitting a sphere to 3D images of a swinging spherical object along x-direction at different speeds.

Fig. 14 illustrates the RMS error curves of the object swung at three different average speeds (0.27m/s, 1.64m/s, 2.83m/s) in the x-direction. It can be seen that the speeds influence the RMS errors significantly. The higher the speed is, the higher the RMS errors are. Also the error curves fluctuate more at higher speeds. This implies that horizontal object motion generates non-stationary noise in the recovery of 3D shapes by the stereo sensor.

4.2.2 Vertical motion

Vertical motion was generated by dropping the spherical object through the capture window of the sensor. Dropping the object at different heights produces different vertical speeds of the object. Fig. 15 illustrates RMS errors of the dropping object at three different average speeds (3.4m/s, 4.3m/s, 5.0m/s) in the y-direction. It can be seen that higher speeds generate larger RMS errors, similar to what is observed in the experiment with horizontal motion. However, the magnitude of the RMS errors related to vertical motion is much smaller than that with horizontal motion at the same velocity. For instance, the RMS errors of the reference object at vertical speed 3.4m/s (represented by the blue curve in Fig. 15) are about 0.5mm, which is much smaller than those of the same object at horizontal speed 2.83m/s (about 1.2mm). This observation suggests that the stereo sensor is more sensitive to horizontal motion than vertical motion. An explanation is that motion blur at horizontal direction generates larger disparity errors than motion blur at vertical direction at the same magnitude.

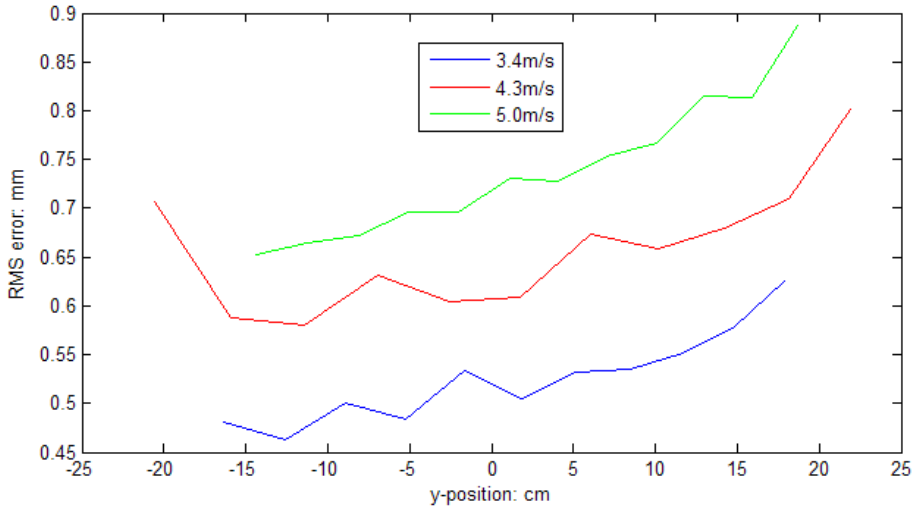


Fig. 15 RMS errors of fitting a sphere to 3D images of a falling spherical object in the y-direction at different speeds.

4.2.3 Motion in the depth direction

Motion in the depth direction was generated by swinging the spherical reference object toward/away from the stereo sensor. Using the same methods as in the experiments related to horizontal and vertical motion, RMS errors were calculated and object velocities were estimated from the dynamic range data of the object.

Three speeds were tested. At the lowest speed 0.9m/s, the RMS curve exhibits a smooth basin shape, which is similar to the RMS curves we obtained in the working range test (here the z-coordinate is very close to working distance). This implies that defocus is the main factor for RMS errors at this low speed. When speed increases, RMS errors rise and the RMS curve becomes less steep in its basin, which reflects the increasing influence of the motion in the depth direction. At the highest speed of 3.3m/s in the experiment, the RMS errors become almost a noisy flat curve, which means that the effect of z- motion has overtaken the defocus to become the main factor in the RMS errors.

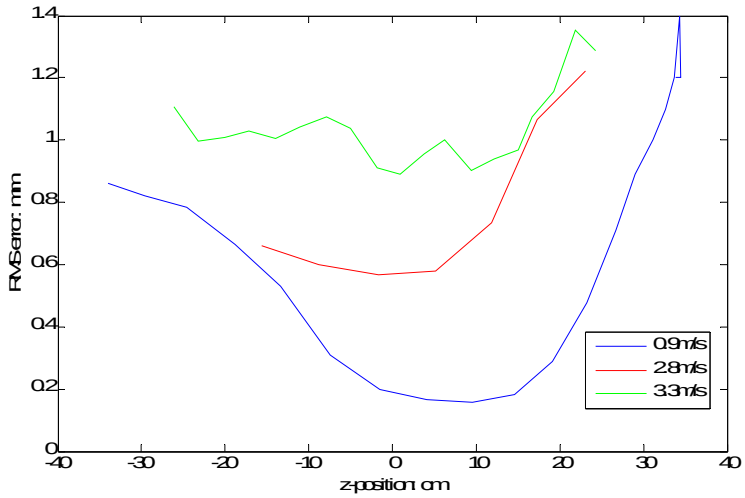


Fig. 16 RMS errors of fitting a sphere to 3D images of a swinging spherical object in the z-direction at different speeds.

Table 3 Characterization of error curves in x-, y-, z- motion

Motion	σ_z (mm)	σ_d (pixel)	3D speed (m/s)	Image speed (pixel/frame)
x	0.32	0.12	0.3	1.73
	0.81	0.30	1.6	9.23
	1.27	0.48	2.8	16.15
y	0.51	0.19	3.4	19.61
	0.63	0.24	4.3	24.80
	0.71	0.27	5.0	28.84
z	0.26	0.10	0.9	0.44 (x and y)
	0.61	0.23	2.8	1.37 (x and y)
	0.96	0.36	3.3	1.62 (x and y)

Table 3 lists some measurements made from the error curves in Fig.14-16. The 3D RMS errors are considered as depth deviations σ_z (since they must be very close), and disparity deviations are calculated using Formula (2). The speeds of the object in 3D have been converted to its speeds in the image plane (assuming the object is at the focus point), which give a hint of the degree of motion blur occurred in the images. Note that the motion in depth direction causes image motion in both x- and y- directions, and the directions of image motion of different points on the object surface are different depending on the locations of the point relative to the principal point of the image.

It can be seen in Table 3 that vertical motion (y-) does not generate a great amount of disparity errors and the same degree of motion in horizontal (x-) direction produces comparatively larger disparity errors. The depth (z-) motion does not generate a high degree of motion in the image plane, however, since the generated motion vector field is not uniform, a significant amount of disparity errors is caused. Therefore ideally the object motion should be constrained to vertical (y-) direction to get minimum disparity distortions.

5. Conclusions

Stereovision is a passive way of sensing the 3D world. The performance of a stereovision sensor is determined by factors related to the system (the sensor itself) and the scene. In this paper, we carried out an experimental study to investigate how the performance of a high speed stereovision sensor is related to its system parameters and some scene factors. Initially the experiments were designed to evaluate the sensor to find optimal configurations for acquisition of 3D shapes of flying bats in the EU project ChiRoPing. However, since the sensor is novel, we believe sharing our experience in evaluating the sensor will inspire applications using sensors of similar type.

The system parameters we investigated include aperture, baseline length and converging distance. It is found that the optimal baseline length is related to the converging distance of the cameras of the sensor. The wider baselines can produce more accurate 3D data but result in smaller working ranges, and the shorter

baselines can allow larger working range but may increase 3D measurement errors. For the two capture scenarios in the ChiRoPing project, optimal baselines are 195mm at the 80cm converging distance and 260mm at the 200cm converging distance. The change of aperture will affect working range. Higher F-stop numbers have longer working ranges, but lower F-stop numbers have better accuracy in 3D measurement when object is in focus.

We also conducted experiments to test the spatial and temporal relationships of the range images acquired by the sensor. It is found that measurement noises are distributed randomly in a range image, which indicates that the sensor performs homogenously in the space domain. The temporal correlations are found low for the weakly textured scene, which suggests that the noises in the range data are temporally independent. However, the correlations are quite high for the strongly textured scene, which we interpret as that the fine details of the scene content (which are captured by the sensor when the scene is strongly textured) have overtaken the noises to be the main contributor to the temporal correlations. The spatial homogeneity and temporal independence of the range images may well allow the application of some advanced techniques to enhance the image qualities, such as super resolution techniques [22].

The scene factors we considered in the study are texture, shape and motion of the object. It is found that proper exposure of texture is vitally important to achieve a good quality 3D acquisition. We experimented with planar and spherical shapes in the study, and the stereo sensor achieved similar level of errors for both shapes. For shapes with distinctive sharp features, the sensor is capable of discriminating the features when separated in about 10-15 pixels.

The tests related to object motion revealed that the sensor is most sensitive to horizontal motion, then to motion in the depth direction, and least sensitive to vertical motion. If we take a tolerance level of 0.8mm for RMS errors, the sensor allows velocities 1.6m/s, 5.0m/s, 2.8m/s in horizontal, vertical and depth directions respectively according to our experiments shown in Fig. 12, Fig. 13 and Fig. 14.

We designed our experiments in a finite space of the parameters (related to the system and the scene), since the number of experiments we could afford to conduct is limited. However, these parameters form an infinite space in practice, and some parameters may interact, which we have not explored in this study. We therefore expect the results presented here to be inspiring and reasonably representative but not thorough. We would hope more performance studies could be carried out to complement the findings in this paper.

Reference

[1] Ypsilos, I.A. , Hilton, A. , Rowe, S.. Video-rate capture of dynamic face shape and appearance. The 6th IEEE Conference on Face and Gesture Recognition, pp.117-122 (2004)

[2] Ypsilos, I.A. , Hilton, A. , Turkmani, A. , Jackson , P.,. Speech driven face synthesis from 3D video, IEEE Symposium on 3D Data Processing, Visualisation and Transmission, pp.58-65 (2004)

[3] Benton, L. , Nebel , J.-C., Study of the breathing pattern based on 4D data collected by a dynamic 3D body scanner**,** Proc. 7th Numérisation 3D/Scanning 2002 Congress, Paris, France, (2002)

[4] Deng, Z., Bailenson, J. , Lewis, J. P. , Neumann, U. , Perceiving visual emotions with speech, 6th International Conference on Intelligent Virtual Agents, Marina Del Rey, CA, USA, pp.107-120 (2006)

[5] Müller, P. , Kalberer, G. A. , Proesmans , M. , Luc Van Gool, Realistic speech animation based on observed 3d face dynamics, IEE Proceedings - Vision, Image & Signal Processing **152**(4):491-500 (2005)

[6] ChiRoPing, http://www.chiroping.org/, Accessed Jan 12, 2009

[7] Dimensional Imaging, http://www.di3d.com, Accessed Jan 12, 2009

[8] Wu, T.-P. , Tang, C.-K., Dense photometric stereo by expectation maximization, European Conference on Computer Vision, pp.159–172 (2006)

[9] Guidi, G. , Micoli, L., Russo, M., Frischer, B. , De Simone, M. , Spinetti, A. , Carosso, L. , 3D digitization of a large model of imperial Rome, International Conference on 3-D Digital Imaging and Modeling, pp.565-572 (2005)

[10] Hartley, R. I. , Sturm, P. , Triangulation, Proceedings of ARPA Image Understanding Workshop, pp.957-966 (1994)

[11] Scharstein, D. , Szeliski, R. , A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, International Journal of Computer Vision, **47**:7-42 (2002)

[12] Cockshott, W.P. , Hoff, S., Nebel, J.-C., An experimental 3D digital TV studio**,** IEE Proceedings - Vision, Image & Signal Processing, **150**(1):28-33 (2003)
[13] Besl, P. J. , Mckay, N.D. , A method for registration of 3D shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence, **14**(2):239-256 (1992)
[14] Lorensen, W. E. , Cline, H. E. , Marching cubes: A high resolution 3D surface construction algorithm, ACM SIGGRAPH Computer Graphics, **21**(4):163-169 (1987)
[15] Starck, J. , Hilton, A., Surface capture for performance based animation, IEEE Computer Graphics and Applications, **27**(3):21-31 (2007)
[16] Ramnath, K., On the multi-view fitting and construction of dense deformable face models, Technical Report CMU-RI-TR-07-10, Robotics Institute, Carnegie Mellon University, (2007)
[17] Del, A. , Agapito, L. , Non-rigid stereo factorization, International Journal of Computer Vision, **66**:193-207 (2006)
[18] Zhang, S., Huang, P.S., High-resolution, real-time three-dimensional shape measurement, Optical Engineering, **45**(12), (2006)
[19] Xiao, Y. , Segmentation and modelling of whole human body scan data, PhD Dissertation, Department of Computing Science, University of Glasgow, (2006)
[20] Khambay, B. , Narin, N. , Bell, A., Miller, J. , Bowman, A., Ayoub, A. ,Validation of reproducibility of a high-resolution three-dimensional facial imaging system, British Journal of Oral and Maxillofacial Surgery, **46**(1):27-32 (2008)
[21] Windera, R.J., Darvannb, T.A., McKnight, W., Mageed, J.D.M., Ramsay-Baggse, P., Technical validation of the Di3D stereophotogrammetry surface imaging system, British Journal of Oral and Maxillofacial Surgery, **46**(1):33-37 (2008)
[22] Caron, J.N. , Rapid supersampling of multiframe sequences by use of blind deconvolution, Optics Letters, **29**(17): 1986-1988, (2004)
[n1] Chan, K.L., Forrest, A.K. , An empirical study on the effects of spatial discretization error in a stereo vision system , BMVC, 1990
[n2] Balasubramanian, R., Das, S., Udayabaskaran, S., Swaminathan, K., Quantization error in stereo imaging systems, International Journal of Computer Mathematics, 79(6):671-691, 2002
[n3] Balasubramanian, R., Das, S., Udayabaskaran, S., Swaminathan, K., Error analysis in reconstruction of a line in 3-D from two arbitrary views, International Journal of Computer Mathematics, 78(2):191-212, 2001
[n4] Sukavanan, N., Balasubramanian, R., Kumar, S., Error estimation in reconstruction of quadratic curves in 3D space , International Journal of Computer Mathematics, 84(1):121-132, 2007
[n5] Chang, C. and Chatterjee, S., Quantization error analysis in stereo vision, 26th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, Calif., 26–28 October 1992.