Low Cost Real-Time Stereo Vision System

A Simple Stereo Block Matching Enhancement

Matthew Luke Kokshoorn

Electrical and Computer Engineering,

University of Canterbury,

Christchurch, New Zealand

[Matt.Kokshoorn@ieee.org](mailto:Matt.Kokshoorn@ieee.org)

Richard Green

Computer Science,

University of Canterbury,

Christchurch, New Zealand

[richard.green@canterbury.ac.nz](\\\\Maxwell\\mlk28$\\Universtity Projects and Documents\\Computer Vision 2013\\richard.green@canterbury.ac.nz)

***Abstract*— This paper proposes an approach to retaining depth information from a stereo block matching correspondence algorithm in order to produce a more stable disparity map. The new disparity map should contain more information than a single block matched disparity map. Camera intrinsic and extrinsic calibration is carried out using a planar calibration algorithm and the global Levenberg-Marquardt optimization algorithm respectively. Manual correspondence of the stereo webcams obtains depth information through the range of 50cm – 250cm with a maximum average error of 4.11%. The proposed memory approach achieves a much more stable disparity map with little tuning required and decreases the average amount of attempts required to obtain a valid depth value by as much 15 times.**

1. INTRODUCTION

There are currently a number of techniques that can be used to acquire depth. These methods can be both optical and audio however many of these require dedicated/specialized hardware and components that can become quite costly. While accuracy is important, it may be more desirable for an approach to sacrifice accuracy for cost. With ongoing advances in the field of electronics, commercial depth web cameras have become more and more financially viable. An example of this can be seen with Microsoft’s Kinect [1].

A large portion of these webcams, including the Kinect, are active sensors, both emitting and detecting light. This light is normally in the infrared range, so as not to visually alter the scene. From this light the scene’s depth information can be derived using structured-light triangulation or time of flight measurements [2]. These methods have been proven to be quite accurate with the Kinect, achieving errors within a 1-2mm to 3cm, over the range of 0.5m through to   
5.0m [3].

In general active cameras has been proven to be more reliable at acquiring accurate depth information from given scene when compared with devices lacking active illumination. These approaches are more affordable but commonly fail in “texture less regions along occluding edges with low intensity variation” [4].

An alternative to these active devices is low cost stereo webcams. This approach consists of a pair of forward facing web cams to acquire two similar but translated images. The distance to a point that is distinguishable in both images is then proportional to the horizontal pixel difference of the two images. By considering the intrinsic and extrinsic parameters of the camera, in conjunction with the pixel separation, the distance from the web cams to the feature can be calculated. This approach can be significantly more affordable allowing them to be used for a more diverse range of applications including; high frame rate tasks [5], embedded stereo vision design [6] and distance calculation in autonomous vehicles [7].

In order for low cost stereo webcam approaches to be a suitable alternative to active methods, the reliability and accuracy must be explored. In particular issues that arise in finding the similar features in the corresponding left and right images. This problem has been come to known as the correspondence problem [8] and has had a great deal of research carried out regarding it. Current methods often require a great amount of tuning for particular scenes, and have algorithms that struggle to consistently find matches between the two images, resulting in a lack of data to carry out depth calculations.

This remainder of this paper aims to investigate the performance of such a low cost approach and relative changes in accuracy due to stereo correspondence algorithms compared to manual correspondence identification. This paper also proposes a method to improve the amount of data retention in the disparity images produced by correspondence methods and thus produce a more stable result.

1. BACKGROUND

In order to improve upon current techniques we must first understand the theory required to gather depth information form a calibrated stereo vision pair. For this paper, the pinhole model will be assumed. This model can be seen in Fig. 2 and leads to the mapping of a 2D point on an image to a point in 3D space represented by. This mapping is expressed mathematically as,

(1)

Where Q is a 3x3 Intrinsic calibration matrix, R is 3x3 orthonormal matrix representing the cameras orientation and , t, is a column vector representing the cameras position [9].

**Depth Calculation**

Consider a point in 3D space, P, projected onto left and right 2D images at points and . From Fig. 1 below it can be deduced that,

(2)

(3)

Combining (1) and (2) we get,

[10] (4)

Where refers to the focal length and b represents the separation between lenses known as the baseline. represents the pixel separation between corresponding points in the two images and is known as the disparity. Therefore to get depth value, Z, we need to find these parameters through camera calibration.

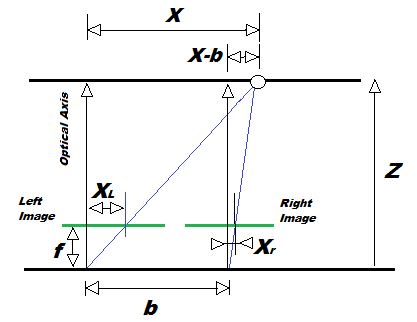


Figure 1 - Geometry Relation of Stereo Vision System [8].

**Stereo Calibration**

In order to retrieve depth information from a pair of stereo web cam images, calibration must be made to relate camera properties to the outside world. There are two parameters that need to be established to improve the capability of computer vision algorithms. The first of these is known as the internal parameters or intrinsic parameters. They describe the internal properties of the camera such as the defects associated with the camera lens [11]. The internal parameters are limited to the physical properties of the camera themselves and are not associative with the cameras external surroundings and resultantly need only be calculated once [12].

To calculate these focal length parameters, the upper triangular intrinsic camera matrix Q must now be considered [13]. This can be expressed as,

(5)

Where, & represent focal length with units of horizontal and vertical pixels. ‘S’ is a skew parameter typically considered as zero [14]. The matrix Q provides a mapping of the received image so that a point in 2D space can be matched to a ‘ray’ in three spaces. By using an object such as a chessboard with easily distinguishable points and known distances between them, the system of linear equations can be solved to achieve the calibration parameters and thus form the calibration matrix Q.



Figure 2 - Pinhole Camera Model Illustration Extrinsic Transformations [8].

The second set of parameters known as the external or extrinsic parameters *do* however take into account the external scene of the camera. These external parameters consider things like the positioning and orientation of the cameras [15]. For the case of stereo vision, extrinsic parameter acquisition is very important as it has the greatest effect on the depth accuracy. The extrinsic parameters effectively act as a transformation (Rotation and Translation) on each image to relate the camera co-ordinates to the real world co-ordinates [16].

This is illustrated in Fig. 2 through the translation and rotation of the blue world co-ordinate system to the black camera system. Mathematically this transformation is described by,

[9] (6)

Where, and represent a points in the camera and world co-ordinate systems respectively and R and t represent the rotational and translational mapping between them. It can now be seen that the pinhole model described by (1) is the result of (4) mapped by the intrinsic matrix from (3) as shown below in (7).

(7)

Calibration techniques typically require the use of the board with a known spacing of easy to identify features. For this paper a chessboard, with a known number of squares and square sizes, is used to solve the system of equations for the calibration matrices. This done using planar calibration algorithms and a number of image points to best estimate the transformation between the pair of cameras to create a stereo webcam pair. This effectively computes the intrinsic camera parameters that are required to relate the world co-ordinate system to the camera co-ordinate system previously discussed. [15].

In a similar manner once the external parameters can be calculated using the global Levenberg-Marquardt optimization algorithm in order to reduce total reprojection error and compute the rectification of each camera to create a virtual mapping so that each image (left and right) have features in the same epipolar lines. The finer details of the calibration algorithms and processes are not within the scope of this paper [15].

**Stereo Correspondence Algorithms**

In order to get the disparity used in (3) there are a number of options. The first, and most simple, of these required the manual detection of the pixel corresponding to a point of interest in each image plane. While this method works for individual points of interest, it is not a viable option for continuous data information of moving objects or the automatic processing of a large number of images. In order for this process to become automated, a stereo correspondence algorithm must be used to generate a disparity map. A disparity, or depth map, is a matrix of equal dimension to the original images whose elements contain the pixel separation between the feature at that pixel and its corresponding point in the second image. There are two main algorithms descried below that can generate this disparity map.

**Stereo Block Matching**

Stereo Block matching is commonly used to solve the stereo correspondence problem due to its simplicity and low overheads. The basic concept of stereo block matching is first to define a reference block that surrounds the point of interest in the left image. The goal is then to find the closest matching block in the corresponding right image within a pre-defined search area. From this the relative disparity between the points in the two images can be computed by considering the best found match [17].

Some common block matching search algorithms can be computationally taxing. These algorithms use matching criteria to check each pixel at all possible locations and therefore generate the best match with the best disparity image. By restricting the search region the execution time can be reduced.

Once the matching criteria has been established, the correspondence problem becomes a search problem with common criteria such as the sum of squared differences (SDD), the normalized sum of squared difference (SDD) and, the sum of absolute difference (SAD) [17].



Figure 3 – Illustration of stereo block matching using SAD on the left image (Top) and right (Bottom) [15].

Fig. 3**,** shown on the, on the previous page illustrates the SAD approach. This begins by selecting a pixel on left image and the SAD windows of size N x N. As the image on the right has been stereo rectified corresponding image features are in the same epipolar line and therefore the search region can be define by the right red rectangle in the lower image in Fig. 3 [18].

A drawback to the stereo block matching algorithm is the necessity of tuning the pre-defined search criteria such as window sizing and image filtering. These parameters can be difficult to establish and can vary for scenes, textures and lighting conditions [19].

An example of a disparity map produced by a block matching stereo correspondence algorithm and its corresponding scene can be seen below in Fig. 4. This particular disparity map is from a static pair of images, giving the user time to tune parameters. Real time disparity maps tend to show more instability as tuned parameters for one frame may not be consistent with the next.

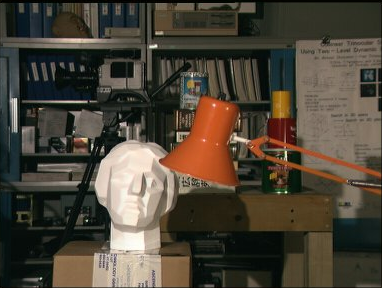




Figure 4 – Stereo Block matching Disparity Map (Bottom) and its Corresponding Scene (Top) [17].

**Stereo Graph Cutting**

Another common stereo correspondence method used is graph cutting. This is a bi-dimensional method that approaches the correspondence problem by considering the constraints of compatibility between a numbers of matching are on each epipolar line [20]. Graph cutting comes with a significant increase in algorithm intensity and therefore processing time. This renders the method ineffective for real time applications on most standard processors.

**3D Projection to get Depth information**

Once a disparity map has been obtained using a stereo correspondence algorithm, a 3D re-projection of the original image can be done. This requires the original 2D image along with the disparity map, carrying depth information, and the camera calibration matrix, Q.

As depth is inversely proportional to disparity, at smaller disparity values (larger depth values), the resolution of the approach greatly increases. That is for a small change in disparity, the corresponding change in depth greatly increases. It is important to note this relationship when selecting a baseline.

Fig. 5 below shows the relationship for a typical stereo vision pair. Accuracy of the method is largely dependent on the selected baseline size.



Figure 5 –Pixel Resolution H and Depth Z, at varying baseline values of a typical stereo vision system [15].

As this paper aims to investigate and improve on some of limitations seen in prior research. In particular this refers to stereo correspondence algorithms, their performance and stability under real time constraints and their dependence on scene specific tuning. In particular it seems that without prior tuning in terms of filters and window sizing that block matching methods produce rather poor results.

1. METHOD/ SOLUTION

**Hardware and Configuration**

To achieve an effective low cost stereo vision configuration the appropriate hardware had to be configured. To avoid any undesirable hardware complications, webcams of the same model were used. These were Cyber Computing’s CMOS sensor based 16MP (effective static pixels) webcams. The software was developed using Microsoft’s Visual Studio 2012 using C++ and openCV on a machine using running 64-bit Windows 7 with 8.0GB of RAM and an Intel 3.4GHz i7-2600 processor. To minimize the effects of the external camera misalignment, a frame was fabricated in order to hold the webcams both forward facing. This can be seen below in Fig. 6. The piece has the ability for the cameras to be fastened with baselines ranging from 3.4cm to 14cm.

**Software**



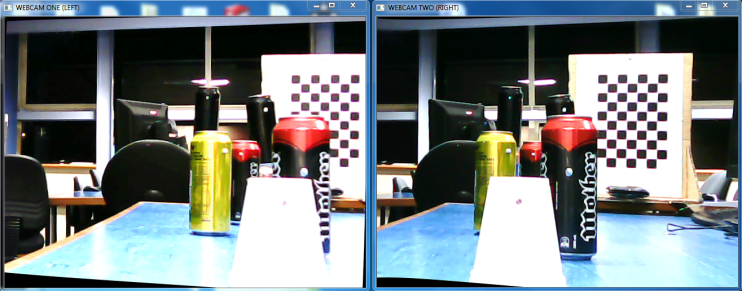


Figure 6 – Webcam Configuration (Top) and resulting left and right image (Bottom).

Once the hardware had been established in a way that would optimize performance the cameras had to be calibrated. This is an important part of stereo vision as it describes other relationship between image disparity and the desired depth information. The camera calibration software developed for this paper is an extended version of the open source code provided by [21].

This approach uses a chessboard image (6 x 9 Squares of 2.8cm Width) and a planar calibration algorithm. This requires two lists of the chessboard corner points from the left and right webcams and computes the extrinsic parameters required to rectify each webcam image. The code also uses epipolar lines to compute the average error of the computed matrices.

The next stage in the calibration procedure computes the intrinsic parameters required for relating the camera co-ordinate system to the world system. This done using and the global Levenberg-Marquardt optimization algorithm [15] which uses the intrinsic parameters previously calculated and the set of distance co-efficient for both images to minimize the reprojection errors. The completed matrix Q is then computed, allowing depth calculations to be achieved from disparity in the two images.

In order to achieve depth information, an interface was formed for the user to first be able to manually identify the disparity between the two images for a selected feature. The interface shows the two rectified images and allows the user to select a pixel in the left image and then the corresponding pixel for that feature in the right image. This allows the validity of the camera calibration matrices to be tested without the added variance of the correspondence algorithms. This was also tested over a range of baseline distances to establish a good testing setup for correspondence algorithms. This worked well and established that the calibration process was running correctly. Results can be seen in the subsequent section.

Next the stereo correspondence procedure was implemented to so that disparity could be calculated between the two images without manual interaction. To do this a stereo block matching algorithm was used. This used the rectified left and right images and produces the resulting disparity image. For interface purposes a separate image of this was then shifted to grey scale and displayed to the user.

The next step, in order to compute the depth information, was to use the camera calibration matrix, Q along with the disparity value for each pixel in order to estimate the corresponding 3D spatial co-ordinate for each pixel. This process was able to be carried using the equations described in, (4) (8) and (9). They use the previously computed disparity map and calibration matrix, Q. The computed 3D matrix is the same size as the original image but elements contain vectors of length three. Each vector corresponds to that pixels x, y and co-ordinate in the calibrated world co-ordinate system.

With this advancement a rectified left camera image was then able to have a single pixel selected and the depth associated with that pixel calculated through the above process.  
  
**Proposed Improvement on Prior Research**

Though the disparity image is now able to be calculated in real time through stereo block matching, the result was very unstable. In order to improve on the known issues with the block matching approach and still retain real time computation requirements, a simple ‘weighted memory’ approach was proposed to aid stability in the image.

The new approach was to create a second disparity image. On each new stereo BM computation, the weighted average of each element of the disparity image and the previous memory image is copied to the new memory image, if and only if the following two criterions were satisfied of the new disparity image element;

* A match has been found for that pixel.
* The disparity value produces a non-zero Depth.

This produced an image that retained previously calculated information. By incorporating (4), this can be expressed mathematically as,

(8)

(9)

(10)

0 (11)

Where, is the computed disparity matrix at the iteration, Represents the memory matrix and M[k+1] is the next iteration of memory matrix after being compared with the newly computed disparity matrix,. and represent the input weighting of the memory and new disparity elements respectively.

The memory image is then used to project 3D information as done with the original BM image. The resultant disparity maps can be seen in Fig. 7.



Figure 7 – The final interface showing; the left image (Top Left), the right image (Top Right), the memory disparity image (Bottom Left) and the raw disparity image (Bottom Right)

1. RESULTS

**Manual Correspondence Testing**

The resulting approach was tested first for accuracy with the by manual correspondence with a baseline of 12.5cm over a range of 100cm-250cm.The test was repeated three times. Fig. 8 below shows the results of this test.

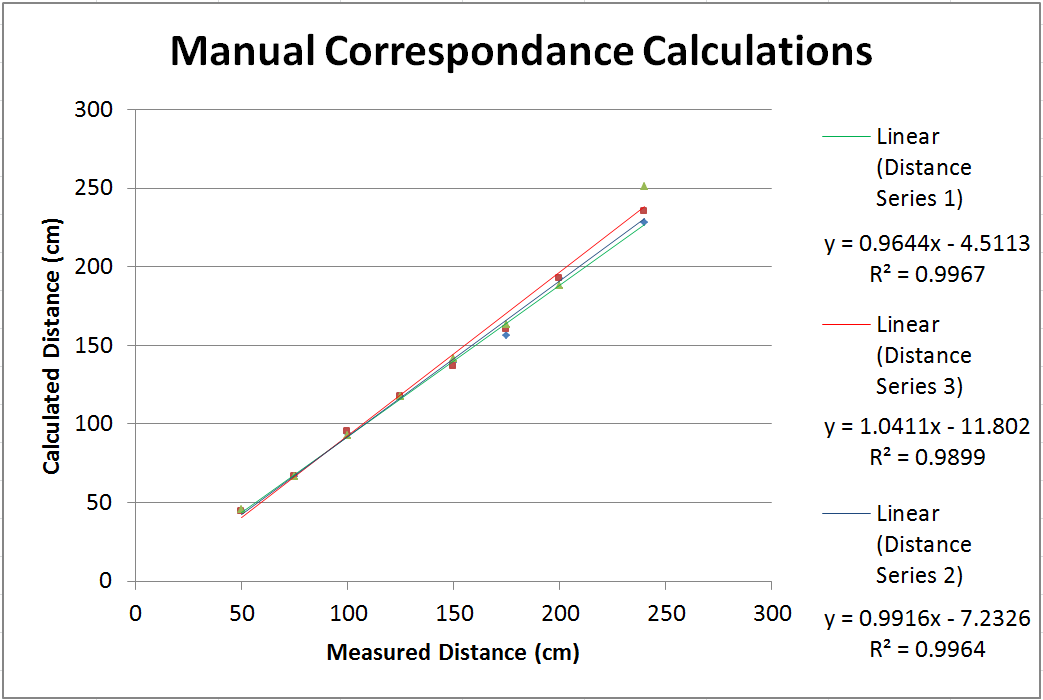


Figure 8 – Results of the Manual Correspondence Testing

As can be seen above, the results for this test show that the approach works well with on average for manual correspondence with a maximum average error produced of 4.11% by series three. These depth calculations will be used as the benchmark ‘ideal results’ for correspondence algorithm testing as to eliminate error due to calibration.

**Block Matching Correspondence Testing**

Using previous baseline and calibration data, the depth is then calculated over the same range using a stereo correspondence block matching algorithm alongside the proposed weighted memory approach with varying weighting. The results can be seen on the following page in Fig. 9.

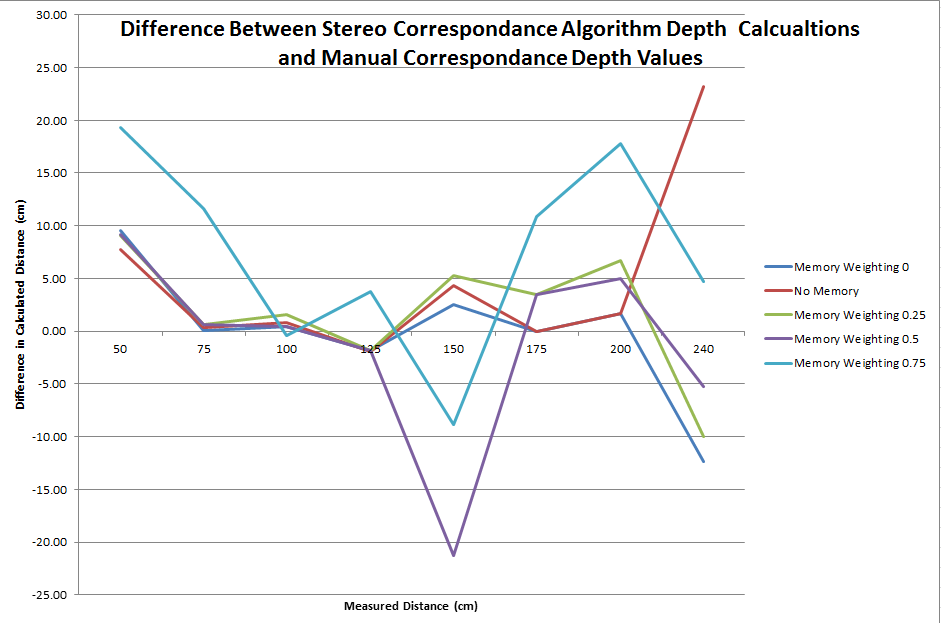


Figure 9 – Results of Stereo Correspondence Algorithm Testing.

Fig. 9 shows that, with the exception of highly weighted memory series (0.75 and 0.5), the memory approach produces similar results to that of no memory. Series with weighting 0.5 and 0.75 tend to produce an amplified error for measurements that were only off by a small margin in lower weightings and no memory.

Of particular interest is the number of attempts required to get a Non-zero reading for the same set of above. This data is below in Fig. 10.

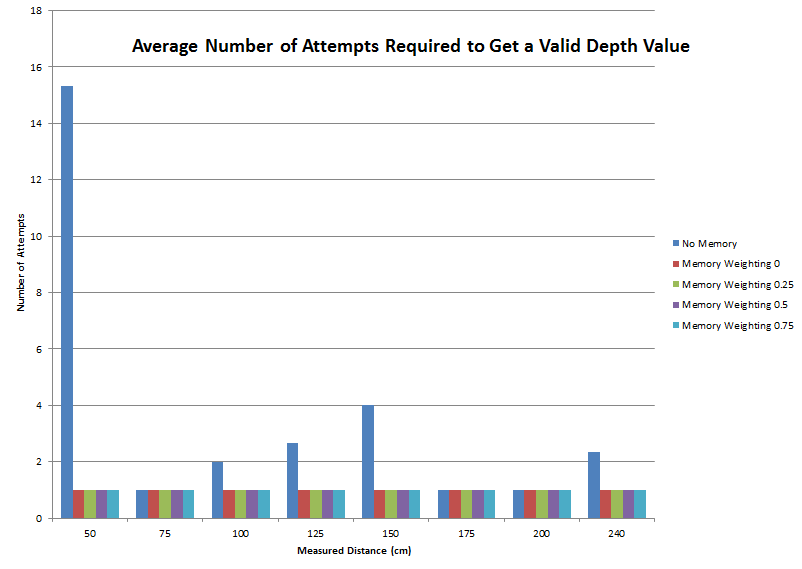


Figure 10 – Average Number of Attempts Required to Obtain a Valid Depth Value.

As can be seen, without the memory approach some areas require numerous attempts to achieve a matched depth value, in particular, are as close to camera or areas with a minimal features. Conversely, in all cases with the memory approach, a valid depth value was obtained first attempt.

**Visual Image Improvements**

The resultant disparity images from each test set and the corresponding scene can be seen below in Fig. 11. As can be seen the disparity maps increase in information significantly with the memory approach.







Figure 11 – From Top Left to Bottom Right:   
Left Image, No Memory, 0, 0.25, 0.5 and 0.75 Weighted Memory Disparities.

**Limitations**

Limitations to the proposed approach arise in captures with high motion. In these cases the object in motion’s depth can still be computed to a high accuracy however areas that have previously been occupied by the object in motion still retain information of the object until they are replaced by new valid data, causing a ‘ghosting’ effect. Similarly this occurs when the cameras themselves are in motion, but on a much larger scale therefore requiring the webcams to remain fixed in one place.

Another limitation is that memory is retained indefinitely for non-matched sections, so that if a single false disparity is obtained for an area that otherwise cannot be matched, it will remain there indefinitely as a false value.

1. CONCLUSION

The resulting solution overcomes limitations of prior research by providing a simple approach to achieving a stable disparity image in real time through stereo block matching. This differs from other methods as it requires little tuning and is not heavily dependent on the scene conditions such as texture [17] [21] .There are known limitations in the memory approach such as ghosting and undefined durations of memory retention. This approach could therefore be improved by filtering of some form, to remove ghosting effects and introducing a maximum duration that the disparity for a given pixel will be retained.

1. REFERENCES

|  |  |
| --- | --- |
| [1] | Microsoft, "Xbox 360 + Kinect," 2013. [Online]. Available: http://www.xbox.com/en-us/kinect/. [Accessed 22 April 2013]. |
| [2] | J. Wang, C. Zhang, W. Zhu, Z. Zhang, Z. Xiong and P. A. Chou, "3D scene reconstruction by multiple structured-light based commodity depth cameras," in *2012 IEEE International Conference*, 2012. |
| [3] | K. Khoshelham, "Accuracy analysis of kinect depth data," *In ISPRS workshop laser scanning,* vol. 38, p. 1, 2011. |
| [4] | J. rk, C. Kim, J. Na, J. Yi and M. Turk, "Using structured light for efficient depth edge detection," *Image and Vision Computing,* vol. 26, no. 11, pp. 1449-1465, 2008. |
| [5] | C. Murphy, D. Lindquist, A. Rynning, T. Cecil, S. Leavitt and M. Chang, "Low-Cost Stereo Vision on an FPGA," in *Field-Programmable Custom Computing Machines*, 2007. |
| [6] | B. Khaleghi, S. Ahuja and Q. J. Wu, "A New Miniaturized Embedded Stereo-Vision System (MESVS-I)," in *In Computer and Robot Vision, 2008*, Canadian Conference, 2008. |
| [7] | J. Roberts, P. Corke and G. Buskey, "Low-cost flight control system for a small autonomous helicopter," *Robotics and Automation,* vol. 1, pp. 546-551, 2003. |
| [8] | L. Torresani, V. Kolmogorov and C. Rother, "A Dual Decomposition Approach to Feature Correspondence," *Pattern Analysis and Machine Intelligence,* vol. 35, no. 2, pp. 259-271, 2013. |
| [9] | Q. Zhang and R. Pless, "Extrinsic calibration of a camera and laser range finder (improves camera calibration)," *Intelligent Robots and Systems,* vol. 3, pp. 2301-2306, 2004. |
| [10] | K.-H. Hsia, S.-F. Lien and J.-P. Su, "Height estimation via stereo vision system for unmanned helicopter autonomous landing," *Computer Communication Control and Automation,* vol. 2, pp. 257,260, 2010. |
| [11] | R. Horaud, G. Csurka and D. Demirdijian, "Stereo calibration from rigid motions," *Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1446-1452, 2000. |
| [12] | R. Illowsky and L. Huet, "Self-Calibration of a Pair of Webcams for Stereo Vision," Stanford University, Stanford, 2007. |
| [13] | I. V. Mikhelson, P. G. Lee, A. V. Sahakian, Y. Wu and A. K. Katsaggelos, "Automatic, fast, online calibration between depth and color camera," 27 March 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320313000503. [Accessed 22 April 2013]. |
| [14] | M. Wilczkowiak, E. Boyer and P. Sturm, "Camera calibration and 3D reconstruction from single images using parallelepipeds," in *Eighth IEEE International Conference on (Vol. 1, pp. 142-148)*, Chicago, 2001. |
| [15] | openCV, "Camera Calibration and 3D Reconstruction," 2009. [Online]. Available: http://opencv.willowgarage.com/documentation/camera\_calibration\_and\_3d\_reconstruction.html.. [Accessed 5 May 2013]. |
| [16] | P.-F. Luo and J. Wu, "Easy calibration technique for stereo vision using a circle grid," *Optical Engineering,* vol. 47, no. 3, 2008. |
| [17] | T. Tao, J. C. Koo and H. R. Choi, "A fast block matching algorthim for stereo correspondence," in *2008 IEEE Conference*, 2008. |
| [18] | "Stereo Matching Method and Height Estimation for Unmanned Helicopter," in *Machine Vision -Applications and Systems*, InTech, 2012, p. 272. |
| [19] | M. Michael, J. Salmen, J. Stallkamp and M. Schlipsing, "Real-time Stereo Vision: Optimizing Semi-Global Matching," 2013. [Online]. Available: http://www.ini.ruhr-uni-bochum.de/data/documents/michaeletal\_improvesgm\_iv2013.pdf. [Accessed 8 May 2013]. |
| [20] | A. Zureiki, M. Devy and R. Chatila, "Stereo Matching and Graph Cuts," in *Stereo Vision*, Viena, I-Tech, 2008, pp. 349-372. |
| [21] | Martin Peris, "OvenCV: Stereo Matching," 21 August 2011. [Online]. Available: http://blog.martinperis.com/2011/08/opencv-stereo-matching.html. [Accessed 6 May 2013]. |