Real Time Representation of 3D Sensor Depth Images

I.JIVET
Applied Electronics Department ,
University "Politechnica" Timisoara
V Parvan 2, Timisoara, ROMANIA

A. BRINDUSESCU
Institute for Automation,
Universitat Bremen,
Otto-Hahn-Alle1, Bremen, DEUTSCHLAND

ioan.jivet@etc.upt.ro, http://www.etc.upt.ro, brindusescu@iat.uni-bremen.de, http://iat.uni-bremen.de

Abstract: - The paper presents a representation model for 3D sensors depth images for use in independent sub system real time action. Field of view depth images obtained with recently developed CMOS 3D sensors are analyzed on their usability in real time perception applications. Based on the characteristic features of the 3D depth image an abstract model with an associated segmented representation is proposed. The objective of the global segmentation method proposed is real time scene perception. The segmentation robustness and the computation cost are the central constraints considered. Classic segmentation methods in use are used as reference since the use pixel or edge criteria have been found insufficiently efficient due to their intensive processing needs and susceptibility to local noise. A model of the depth image in terms of object area and mean center location is proposed as a real time local depth image representation. The results of the use of the proposed model in independent action oriented applications is also presented. The representation method proposed is analyzed in its performance in terms of estimated computation time and semantic relevance for the sample applications.

(Key-Words: - 3D sensor depth image, segmentation, real time perception, action oriented sensor data model.
modelvi1vtemp
.

1 Introduction.

One of the challenges of using image sensors in applications is the high quantity of information conveyed by images. Most of this wealth of information is recognized as redundant for the immediate application in most cases [1].

Extracting the usefull information in a readily usable format is a key problem for many intensely studied area of applications from machine learning in robots to enactive perception in humans [2], [3].

The semantic perception requires a formal representation of the image representation for an a priory specified action objective of the application. Interaction with the environment perceived in the scene in its static or dynamic behavior is the ultimate performance criteria of the whole system.

Recent advances in CMOS Time of Flight (TOF) 3D infrared sensor increased interest in their use in real time visual perception applications [6], [7].

The approach presented in the paper is based on a selective use of the information contained in a 3D sensor depth image in order to obtain good processing performance in real time. The applications targeted are independent actions for top level system decision and coordination tasks.

The use of 3D sensor provides cost effective solutions in visual perception due to their compact size with adequate resolution.

The image content perceived at a primary level of detail is registered in the system as an abstract model based on action requirements of the application. The proposed model is characteristic to both human

behavior as well as deductive machine learning. The focus of the work presented was directed to a representation supporting the understanding of the image content as a process of learning.

A practical area of applications considered for the present work is an autonomous digital system providing strategic decision form visual images, like orientation of mobile robots. A second direction of application considered was the general real time image to sound environment representation for perception of the environment by visually impaired persons.

1.1 Previous work

In comparison to general 2D image sensing, the 3D depth image obtained by the recently developed sensor was found to a be just as complex and as demanding with the same problems well known and intensely studied in 2D vision [4][5].

Real time applications pose an even more severe constraint on the complexity and cost of the feature extraction form 3D sensor depth images. Nevertheless there have been reports of promising use of newly developed CMOS 3D depth sensors in commercial applications for strategic real time control decisions.

An example of recent work on visual perception as reported covers the use of stereo camera and 3D sensor with depth output as used in mobile robot navigation [9], [10]

A second direction of development opened up by the availability of cost efficient CMOS 3D sensor with depth images is intelligent car visual sensor support in automotive. Prominent automotive research teams conduct intensive research in using 3D depth sensors for intelligent vision based systems in car safety [4].

Stereo CCD cameras used to have a technological advantage edge until recently and considered as the best available vision sensor.

A well known problem associated with CCD stereo vision is the high sensitivity of the depth measurement to errors in feature localization in each camera image. Small errors in the position of the feature in the image result in significant depth measurement error.

Our approach to the problem presented in the paper is directed on a 3D depth vision representation to be used for perception of the sensed environment at a high order semantic level of perception.

2 Depth Assessment by 3D Sensor

The depth information in the field of view was recognized as very important since the early days of use of visual sensors in robotics [2].

The CCD camera in mono or stereo format dominated until recently as the vision sensor technology of choice [11].

New CMOS 3D sensors providing depth image in addition to reflected light intensity have become an important contender as the best sensor type. Commercially available CMOS 3D sensors, with high efficiency in hardware implementation provide pixel level direct depth data have made a real success recently [5] [6], [7].

The operation of the camera is based on an amplitude-modulated infrared light source and a CMOS sensor that provides information on the field depth of objects from the back scattered light. The ambient light is not affecting the sensor operation since it is not modulated. Several other methods are used to eliminate the effect of ambient noise.

The camera module contains a light source constructed from a bank of infrared LED 870nm wavelengths, a lens system for the detector chip, a detector chip with 176×132 phase-sensitive pixels. The whole chip is fabricated on standard CMOS process and also contains an embedded CPU for advanced image processing.

The time-of-flight (TOF) sensor measures distance by observing the time delay between emission and detection of light pulses. The pulses are emitted by an active light sources switched on and off with 50% duty cycle at a frequency on the order of 50 MHz. The light beam bounces off a surface in the scene and returns to the camera.

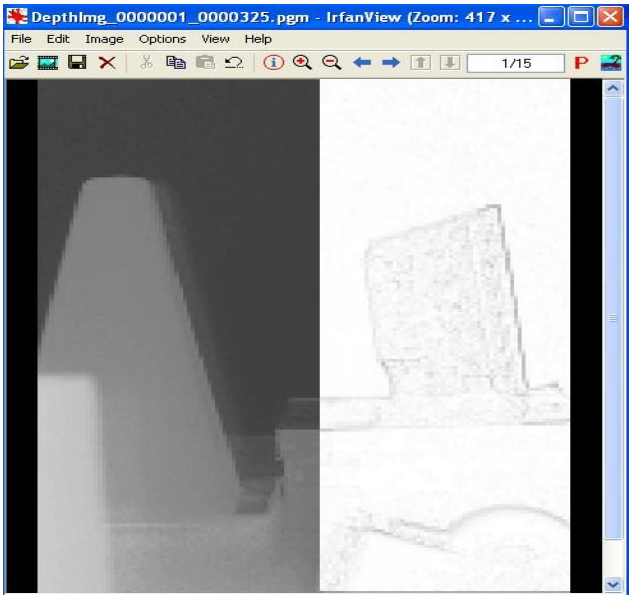
The phase of the pulse and geometrical relations of light source to each pixel are used to determine the depth information.

The key part in the sensor design is the pixel matrix structure using a special CMOS transistor with two gates. The differential structure accumulates photon generated charges in the two collecting nodes. The gate modulation signals are synchronized with the light source, and hence depending on the phase of incoming light, one node collects more charges than the other. At the end of integration, the voltage difference between the two nodes is read out as a measure of the phase of the reflected light beam.

The maximum unambiguous range for 50 MHz is around 3m with a resolution on the order of one cm.

The frequency of the light source defines the maximal depth of the field of view. There are methods using multiple frequencies to extend the range up to 100 m.

Motion artifacts exist when there are fast movements in the scene. The motion artifact is observed mostly around the edges of a moving object. Correction measures are used at each level of the processing of the depth image.



a) b)

Fig. 1 Depth image sample with objects on a table: a) original depth image; b) right half of the image sample following edge detection.

The 3D TOF sensors have numerous advantages over other depth sensing devices. Triangulation-based methods such as stereo CCD require intensive post processing to construct depth images. This is not necessary for the TOF sensor and the post processing usually involves a simple table-lookup to map the sensor reading to real range data.

Depth 3D images exhibit all the issues known in 2D classical image processing.

Pixel base edge detection and region growing segmentation methods of the image content are known to be very expensive in processing time.

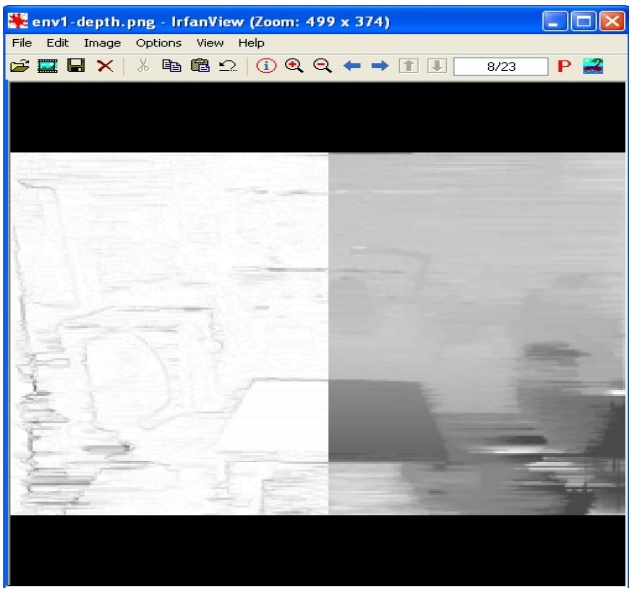
The main advantages of direct (TOF) 3D sensors are their direct assessment of depth with ought the need for a front end processing like in the case of CCD stereo image cameras.

3 Fast object extraction in depth images

3.1 Space representation in 3D depth image

A 3D sensor image is more efficient than a 2D image due to the geometrical properties inherited from the three dimensional space it is projecting.

Every pixel of the depth image is not an abstract value but a geometrical measure of distance to the surfaces of objects in the scene.



a) b)

Fig.2 Depth image of a laboratory view from stereo CCD camera: a) image following edge detection b) right half of original depth image.

Most geometrical relations among the scene objects are preserved in the depth image of the CMOS sensor. The scene is projected on a planar matrix structure at the level of the integrated circuit scaled in width by the associated optic lens system.

The 3D sensor image of the field of view in projection on the sensor detecting area includes also the 3D perspective relations of objects in the scene.

Two notable relations where found in the 3D depth image of the objects on the scene as reflected by the sensor and can be simply expressed as follows:

RI. Objects in the scene obstruct one another in the line of sight all the way to the background or ground level depending on their position, relative depth and cross sectional areas.

RII. The perceived objects following detection by the 3D sensor on the 2D sensing matrix plane have areas proportional to a cross section perpendicular to the exploring ray. Different objects in the scene have cross sectional area that are proportional to the real 3D object maximal cross section scaled by the inverse square of relative depth to the object.

A notable approximation to be considered in evaluation of the modeling method employed is due to the object shape. Different geometrical shapes result in different proportional mean area factor for each object in the scene.

Taking in consideration the geometrical relations described above it is natural to consider a model of the scene in terms of real objects projected cross section and the cross section mean center position. The depth difference between the shape projection plane and the first tangent plane to the object represents the localization approximation.

3.2 Segmentation by depth threshholding

To obtain the scene model comprising of objects and associated depths a segmentation of the depth image is the necessary first step.

In the present state-of-he-art image processing there are many image segmentation method in use. The pixel based region growing segmentation method is one of the most often used. Object edge detection and subsequent aggregation to construct objects is also very often used [8][9].

In the present paper a specific solution applicable to 3D depth images is proposed. The solution is similar to color code region segmentation successfully used for real time applications [1].

The scene segmentation algorithm selection was based on its robustness. It puts emphasis on a small number of components selected as dominant in the perception process [2].

The proposed object extraction basic principle is a two step process. The first step uses global depth histogram segmentation by thresholding to select objects from background. In the second step the detection and localization of floor and lateral field of view enclosures are executed.

In Fig 3. an example of the segmentation of an image using histogram thresholding is presented. For the peaks of the depth histogram the threshold value was set at the trailing 10% level of the peak with respect to the following minimum:

$$\textbf{Depth\_label} = 1/10\,(\textbf{Max}_N - \textbf{min}_{N+1}) \qquad (1)$$

where *Max* and *min* are the successive values of a peak and valley in the histogram.

The depth value determined is used to label the corresponding segmented object in the scene.

The objects also receive as attachments other parameters for image perception in applications. The area of the object in terms of pixel count at the given relative resolution of the image is calculated by counting the pixels with the same depth level.

The mean center of each object is also calculated as follows:

$$\mathbf{X_K} = \mathbf{Sum_i}\,[\,\mathbf{X_i}\,(\mathbf{depth} = \mathbf{k})] \qquad (2)$$
$$\mathbf{Y_K} = \mathbf{Sum_i}\,[\,\mathbf{Y_i}\,(\mathbf{depth} = \mathbf{k})] \qquad (3)$$

where X_K , Y_K are the coordinates of the center of the area of the object and X_i , Y_i are the coordinates of the pixels of the segmented object.

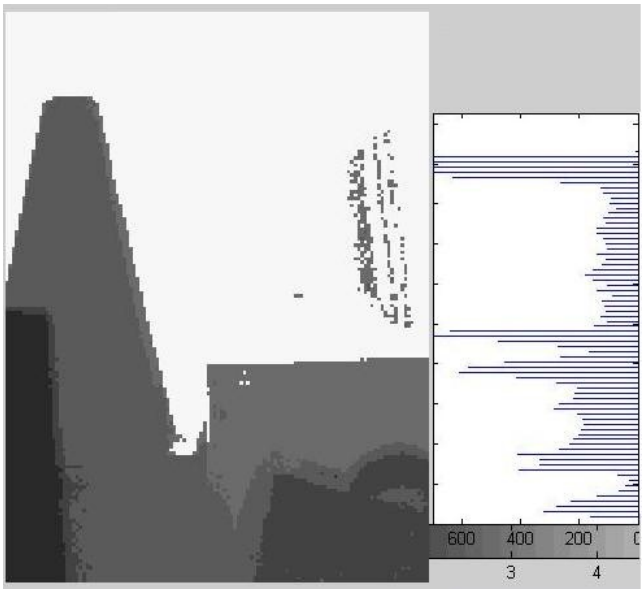


Fig. 3 Segmentation by thresholding example shows it working for most objects in a scene. The exception are lateral running surfaces to the central direction.

The major problem of the global histogram segmentation as a method is the poor performance of the thresholding on lateral surfaces and floor. Due to large spread of depth values for such regions of the image thresholding is not possible. The depth values

do not cluster but distribute over plateau histogram regions.

The same problem occurs in the detection of objects presenting 'side' oblique surfaces at an angle to the line of sight. As it can be seen from Fig. 4 the same problem occurs when using segmentation with region growing algorithm.

The solution proposed in the present paper is to determine the location of lateral and floor surfaces by partitioning them into slices. The slicing must be done in the direction perpendicular to the sensor view direction.

The number of slices per image was determined experimentally to be consistent with the segmentation threshold of 10% of the area margin in area determination.

D slice = D image /(10 x Nr. of objects) (4)

For the image of 132 x 176 resolution obtained form the CMOS TOF sensor the number of slices was determined at 12 for the width of the image and at 16 for its height.

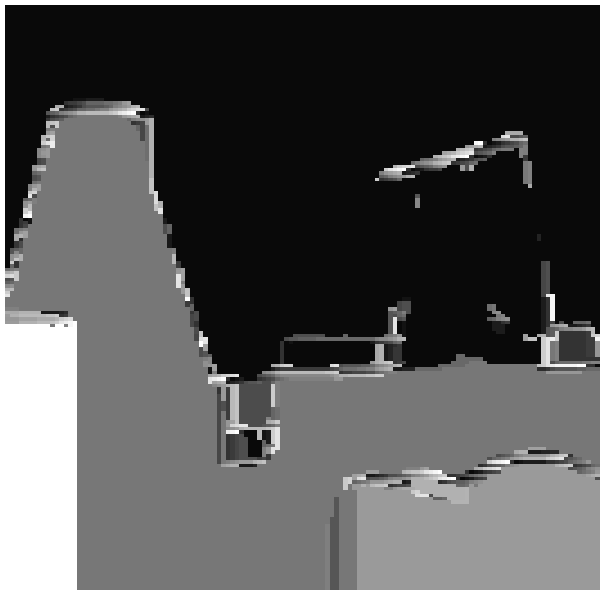


Fig. 4 Segmentation by region growing of the sample depth image for comparison.

Each slice is processed separately. The slice histograms do have peaks for the oblique surface due to the small slice width compared with the running height. The pixels along the height of the slice do cluster in the histogram if the slice was chosen perpendicular to the perspective direction. The slices result vertical on lateral surfaces and horizontal on the floor of the scene.

The processing time increases and the sliced object become multiple part objects. The advantage of the method is that segmentation using histogram thresholding is unique for all objects and easy o implement in real time.

The advantage of histogram thresholding segmentation techniques used in 3D sensor depth data processing is due to its robustness to noise in the captured image.

It is also demonstrated that when histogram thresholding is applied to 3D depth image the segmentation method needs extensions to address the specific constraints of the 3D sensor space.

The time necessary to segment the depth image and determine the object area and mean center location was estimated taking into consideration the depth image resolution of the sensor. For the 3D sensor analyzed in the present paper the dimensions of the sensor matrix where 176 x 132.

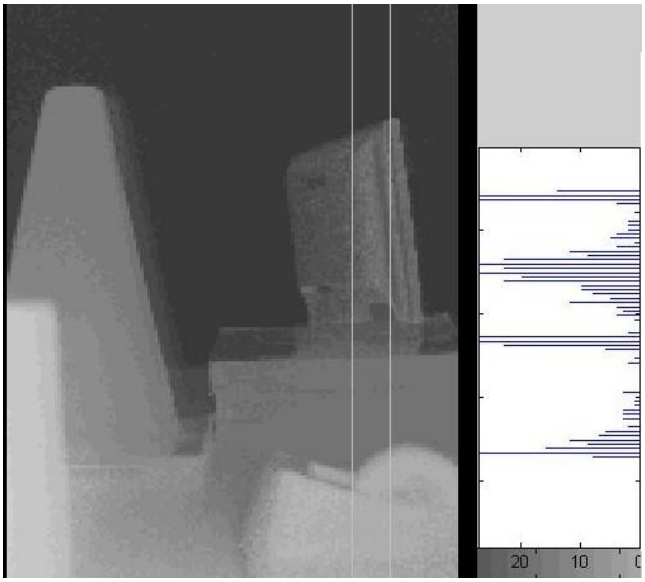


Fig. 4 Histogram of the vertical slices of the image contains peaks associated to lateral surfaces. Slice localization by thresholding is demonstrated.

No assumptions where made on the real hardware implementation of the algorithms since this is highly dependent on the technology used and the application.

Table 1 Segmentation performance results and the estimated processing time.

	Thresholding	Region Growing
Objects	Front Facing only	Front Facing only
Area %	91	96
Processing time	x1	x8

The results in Table 1 indicate comparable quality object extraction and very low computation complexity. The segmentation by slicing can be extended to a finer image granularity. A finer width slicing will improve the precision of the depth labels for each slice. The computation complexity will increase and the performance will degrade in this case.

The general performance criterion for the representation method proposed is the computation time average per pixel or object in the scene perceived. The model proposed relies on selective area object detection. The selection of salient features based on depth becomes cost efficient only as long as the selective process results in objects with good statistics.

All segmented object area needs to be no smaller then one order of magnitude of the image dimension. In the case of the image used in the example the five objects occupy from 5 % to 25 % for a total of about 60 % object area in the image. The 10% error margin accepted for an object area is thus justified given the 176 x 132 resolution of the image.

4 3D Depth Objects Representation use Real Time Actions

The performance of a system in real time using visual sensors relies on the local segmented representation in a simple format for independent action. It also satisfies the minimal computation requirement to meet the real time constraint.

The sensor module using the scene model with objects and their geometrical parameters can act as a independent entity.

The coordinates of the center and depth are vectors for approach actions as well as clearance data for avoidance of obstacle objects.

The registration of perspective in a higher order representation of the sensor image needs care full consideration at the higher level before further contextual use.

Perception and adaptivity in the environment of the sensor user in its actions on the environment will be subsequently amended by the higher level parser and planer of the scene activity when its actions reach out of bounds.

In the global system perspective at the perception level of the scene models by the object area must be preserved at the real perspective scale of the depth sensor image. There is no scale adjustment expected in the global perception model.

Objects at different depth in the image require scaling for correct representation of objects in the global real scene model. A classical situation is the registration of sensed objects in the abstract simultaneous localization and mapping environment model in mobile robots (SLAM).

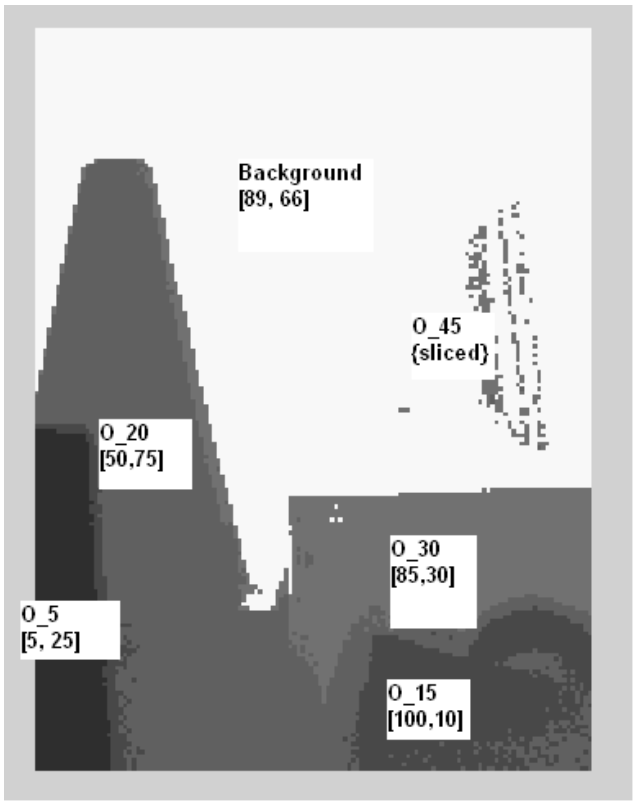


Fig 5. An annotated segmented image with the object list and center coordinates. The object name contains the depth label.

The higher order perception logic responsible for

the global decision process must subsequently make the appropriate contextual adjustments of scale.

The above distinction of object effective area marks the difference between abstract perception by machine sensors and human natural perception. Both are dependent on local or hierarchical model of the perception system.

Locally representing visual images by objects depth, mean center location and area permits environment modeling as an ordered list of objects with properties.

In a dynamic scenario of the environment the sensor motion in the field of view finds new objects entering the scene from obstructed areas by present visible objects. These need to be addressed with out changes to the topology of the local scene model.

Direct scene extensions of the list of objects and space structuring relations are usually required.

At the system level the sensor function is analogous to a push down automaton with an associated algorithm for sensor image registration.

Actions like targeting one object or avoidance of collision are natural for the lists of object representation.

A registered abstract virtual model of the scene is context independent and is therefore expandable.

Proving the sufficiency of the model as proposed in an action oriented system demonstrates the usefulness of the model as proposed.

5 Conclusions

A 3D depth sensor scene model based on histogram segmentation abstract representation is proposed. The segmentation robustness and computation cost where the central constraints considered.

The objective of the global segmentation method proposed is real time scene perception. The model of the depth image in terms of object area and mean center location is proposed for action oriented tasks. The final objective of the depth image representation proves self sufficient for real time perception.

The performance of the proposed representation method is analyzed in its efficiency in terms of estimated computation time and semantic relevance in applications.

Further work is necessary to include support for object surface texture to extend the system capabilities to object recognition.

References:

[1]D. Comaniciu, P. Meer, "Robust analysis of feature spaces: color image segmentation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1997* pp. 750-755.

[2] Bruce, J., Balch, T. and Veloso, M Fast and inexpensive color image segmentation for interactive robots, *Proceedings IROS 2000 Conference on Intelligent Robots and System*s, Takamatsu, Japan,2000

[3]A. Talukder, R. Manduchi, A. Rankin, and L. Matthies, Fast and Relaiable Obstacle Detection and Segmentation for Cross-country Navigation, *Intelligent Vehicle Symposium*, Versaille, France, June 2002.

[4]S. Hsu, S. Acharya, A.. Rafii and R. New, Performance of a Time-of-Flight Range Camera for Intelligent Vehicle Safety Applications, *12th International Forum on Advanced Microsystems for Automotive Application*s Berlin, March 11-12, 2006

[5]Gokturk, S.B. Yalcin, H. Bamji, C. A Time-Of-Flight Depth Sensor - System Description, Issues and Solutions, *Conference on Computer Vision and Pattern Recognition Workshop* June 2004

[6]CMOS-compatible three-dimensional image sensor IC. *United States Patent 6,323,942,* Canesta, Inc. (Palo Alto, CA), September 22, 1999

[7] Coded-array technique for obtaining depth and other position information of an observed object *United States Patent 7,212,663*, Canesta, Inc., Sunnyvale, CA, June 19, 2003

[8] Z. Lin, J. Jin, H. Talbot, Unseeded region growing for 3D image segmentation, *Conferences in Research and Practice in Information Technology*, Vol. 2. P. Eades and J. Jin, Eds. Australian Computer Society, Inc. 2001

[9] Weingarten, J.W.; Gruener, G.; Siegwart, R. A state-of-the-art 3D sensor for robot navigation (IROS 2004). *Proceedings of IEEE/RSJ International Conference* . 2004 Volume 3, 2004, pp. 2155 – 2160

[10] P. Dorninger , C. Nothegger 3D Segmentation of Unstructured Point Clouds for Building Modelling, *Stilla U et al (Eds) PIA07. International Archives of Photogrammetry,* Remote Sensing and Spatial Information Sciences, 2007, pp.191-196

[11] D. Litwiller, "CCD vs. CMOS: Maturing Technologies, Maturing Markets", *Photonics Spectra,* August 2005, Pages 54-58.