Static and Dynamic Abstract Formal Models for 3D Sensor Images

Ioan Jivet, Alin Brindusescu, Ivan Bogdanov
Applied Electronics Department ,
University "Politechnica" Timisoara
V Parvan 2, Timisoara, ROMANIA
ioan.jivet@etc.upt.ro http://www.etc.upt.ro

Abstract: -The paper presents a perception oriented linguistic formal model for 3D sensors depth images. Both the static object extraction and short term dynamic evolution in the scene are analyzed. The target applications are subsystems in action involved in independent environment exploration and learning be it human or machine. Field of view depth images obtained with recently developed CMOS 3D sensors are analyzed in their capacity to provide immediate action oriented data. For the 3D scene images a selective segmented method is proposed in terms of salient objects in the depth image. The model as proposed uses a representation of the scene depth image in terms of object area and mean center location. An original abstract formal language representation is proposed. The extension of the context free grammar with attributes ads structure to the model. It is also shown that the generated language translates directly depth labeling into action planning on the environment. The performance of the proposed abstract representation method is analyzed in terms of estimated computation time and direct semantic relevance for a sample application. For applications of object motion detection and tracking the formal model was extended with attributes for direction and speed. The object position drift based on segment correspondence for speed determination is shown to be compatible to the formal model as proposed. Further development of the model for multi layered representations for more complex applications areas is also outlined.

Key-Words: - 3D sensor depth image, segmentation, formal language perception model, real time perception.

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 known to be redundant for the execution of immediate tasks in a complex system with more then one actor [1].

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

Interaction with the environment perceived in the scene in its static or dynamic behavior is the ultimate criterion of performance of the system.

The contextual semantic information is necessary from immediate actions following perception. Recently, system architectures have been proposed that models separately the task execution and scene perception in the context of general system management scheme [1].

The cases where task execution performance continues to be problem are the ones with a close coupling of perception and action. The relationship of 3D sensor data and its user in motion and in interaction as part of the environment is an illustrative example of a tight coupled system.

Recent literature reports intense interest in visual perception studies on the use of stereo cameras and 3D sensors with depth output in mobile robot navigation [3], [13]. The research on intelligent car visual sensor support for improved safety in automotive take advantage of the new 3D depth sensors technology [7].

The key requirement of fast real time operation requires a simple method for object detection [6][8].

The approach presented in this paper is based on a selective use of the information contained in a 3D sensor depth image in order to obtain a compact representation of the image in terms of significant salient objects and their positions.

The interaction with the objects of interest in the scene must be extracted from the scene representation as a compact model. To simplify this task we propose a abstract representation of the image that is formally context free by design. In this way the subsystem is relatively independent of the general system management and the context of other sensor subsystems present in the environment.

We also demonstrate that our model in relation to perception can be linked naturally with an a priory specified set of action like grasping or avoiding objects in the sensed 3D environment.

Even if scene objects are fixed they change relative position due to observer motion in its actions of avoiding obstacles or reaching to grasp objects. The sequence of images obtained in such a case raises the problem of time consistent representations from one to the other. The term *range flow* is used to characterize the motion of objects in the images of scene [20].

One practical area of application considered in the present work was real time depth image to sound representation for perception of the environment by visually impaired persons.

We shown that the speed and direction of motion are simple to incorporate in the same abstract model with object positions. This is due to the fact that both static and motion characteristics of objects are attributes naturally associated in their perception.

The model as proposed can also be used in an autonomous digital systems executing an approach or avoidance task using 3D depth images like the activity tasks of mobile robots.

Part 2 the paper presents a new 3D depth sensor and the characteristics of the depth image data obtained.

A efficient segmentation using histogram thresholding with an original extension for detection of oblique surfaces is presented in Part 3.

Part 4 introduces the formal language model for the 3D depth image based on context free attributes grammar. It is shown to be very useful and direct for human perception and supporting a natural translation to action planing.

The extension of the attribute grammar with attributes for the short term dynamics of objects is presented in part 5 of the paper. The natural semantic interpretation of motion is demonstrated to be similar to the case of static attributes. It also shown that range flow attributes validate static attributes and help action planning with time consistency.

The conclusions and an outline of further development of the proposed model is given in the last part 6 of the paper.

2 Depth Assessment by 3D Sensor

The CCD camera in mono or stereo format dominated until recently being perceived as the vision sensor technology of choice. Stereo CCD cameras had a technological advantage edge over depth sensors and have been used as the preferred vision sensor [15].

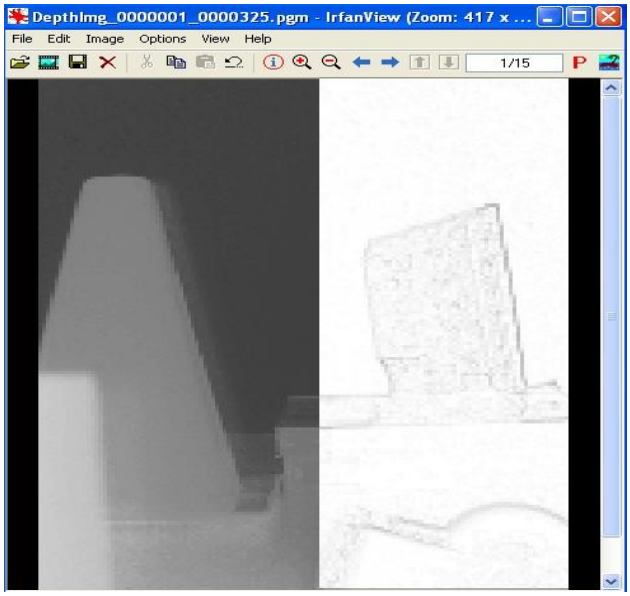
A well known problem with CCD stereo vision is the high sensitivity of the depth data to errors in locating corresponding features in each image. Small errors in the contrast areas limit in the two image results in significant depth measurement error.

The newly developed CMOS 3D sensors provide depth data in addition to reflected light intensity. They have become an important contender as the most frequently used sensor type.

Recent advances in CMOS Time of Flight (TOF) 3D infrared sensor increased interest in their use in real time visual perception applications. Fabricated using the standard CMOS process integrated circuits 3D sensors are now available commercially at a very low cost [9] [10], [11].

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

The operation of the sensor is based on an amplitude-modulated infrared light source and an array of CMOS transistors that determine the field depth 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.



a) b)

Fig. 1 Depth image sample of objects on a table [http://www.canesta.com]: a) original depth image; b) right half sample image following edge detection.

The camera module contains a light source constructed from a bank of infrared LEDs at 870nm wavelengths, a lens system for the detector chip incorporating 176 × 132 phase-sensitive pixels.

The whole chip is fabricated on standard CMOS process and the chip also includes an embedded processing unit for pixel depth value calculation.

The time-of-flight (TOF) sensor measures distances by determining the time delay between emission and detection of light pulses. The pulses are emitted by the light source switched on and off with a 50% duty cycle at a frequency on the order of 50 MHz. A simple calculus shows that the distance traveled by light in an impulse period is about 3 meters. At a digital signal resolution of 8 bits the depth resolution is on the order of one centimeter.

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

The light beam bounces off surfaces in the scene and returns to the 3D sensor. The phase of the reflected beam and geometrical position of light source to each pixel are used to determine the depth information in the 3D scene viewed.

The core of the sensor design is a pixel matrix consisting of special CMOS transistors with two gates. The differential structure accumulates photon generated charges in the two collecting gates. The gate modulation signals are synchronized in quadrature with the light source, and hence depending on the phase of incoming reflected light, one node collects more charges than the other. An integration in time over many cycles is performed to increase sensitivity. At the end of the integration cycle the voltage difference between the two nodes is read out as a measure of the phase of the reflected light.

The principle of operation can be described by the following set of equations. The light source is commonly best modeled by a sinusoidal form :

$$Source = cos(2\pi f_m t) \quad (1)$$

where f_m denotes the frequency of the pulsed light source.

The reflected signal of amplitude *Reflected* is of the same form but delayed by a phase φ due to the travel time.

$$Reflected = sin(2\pi f_m t - \varphi) \quad (2)$$

In the phase delay detection process the reflected signal is mixed with a quadrature pair of sine and cosine signals of the same frequency of magnitude $Mixed_{90/0}$.

$$Mixed_{90} = 1/2[sin(-\varphi)+sin(4\pi fmt-\varphi)] \quad (4)$$
$$Mixed_{0} = 1/2[cos(-\varphi)-cos(4\pi fmt-\varphi)] \quad (5)$$

After a low pass filtering, the low pass terms only depending of phase, are used to calculate (using a look-up table) the phase delay and the from it the corresponding distance *D* to the object:

$$\varphi = \arctan(V_{90}/V_{0}) \quad (6)$$
$$D = c\varphi / 4\pi f_m t \quad (7)$$

where V_{90}, V_{0} are the low pass terms of the reflected and mixed signals and *c* is the speed of light.

In practice all the above indicated operations are obtained by suitable electronic and detector matrix spatial arrangements [6].

For each pixel the sensor transistor pair is gated by a square wave with 90 phase shift implements the mixture operation. The difference of the mixed signals is low passed by integrating the light generated charges over a large number of pulses. Finally the distance is calculated using a look-up table.

Due to the finite time of charge accumulation in each pixel there are motion artifacts if fast movement occurs in the scene. The motion artifacts are observed mostly around the edges of a moving object. Correction techniques are used at each level of the processing of the depth image to reduce these artifacts.

Depth 3D images use in applications encounters all the issues known in 2D classical image processing. Methods used in 2D image processing are in general usable accepting the costs involved.

Pixel based edge detection and region growing segmentation as methods for object localization in the image most often used. These methods although very good in performance are known to be very intensive in the cost of processing time.

From efficient extraction of the information contained in a 3D sensor depth image we propose an approach that efficiently selects salient objects in the scene. The scene content in terms of the main objects is represented in a compact form that captures the geometrical relationships.

Also considering the perception oriented goal of the processing only significant detected objects are extracted and considered as possible candidates for a related action on them.

3 Segmentation by depth threshhold

A 3D sensor image is more explicit than a 2D image due to the presence of geometrical information inherited from the three dimensional space it projects.

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

In today's 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 represent objects is also very often used [12].

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 [2].

The scene segmentation algorithm was selected mainly due to its robustness and simplicity. It puts emphasis on a small number of components selected as dominant and sufficient for the environment perception process as the main objective of the present work.

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. is presented a sample 3D depth histogram segmentation based on thresholding. For the labeling of the objects the peaks of the histograms where first determined. In the second pass for each the threshold was set on the trailing slope at the value of 10% of the depth difference between the peak and the following minimum.

The simplicity and performance of the segmentation method used and tits robustness have been determined from experiments on a set of images with a limited number of salient objects.

The depth related scale of the different objects in the scene was not considered explicitly in the present study. The accuracy of area estimation was not considered as essential for object detection and localization. Further subsequent scene analysis for more demanding tasks like motion estimation would need to estimate this parameter with better accuracy.

The 10% margin was determined as sufficient to accommodate inter objects relative depth variance as well as other parameters like position and angle to the viewing direction. The value was found consistent with general image feature resolution for the purpose of object location as reported in recent literature [16].

The major problem of the global histogram segmentation is poor performance of the thresholding on lateral surfaces and floor. Due to large spread of depth values for such regions the image segmentation by thresholding is not effective. The depth values do not cluster in a peak on the histogram but distribute over plateau histogram regions. The same problem occurs in the detection of objects presenting 'side' surfaces at an angle to the line of sight.

In order to determine a point of reference other segmentation algorithms where used for the same set of 3D depth images for comparison.

As it can be seen from Fig. 4 the same problem occurs when segmentation is carried out using region growing algorithm. The depth values for adjacent pixels do not cluster in a narrow range for surfaces oblique to sight direction.

The solution proposed to discover the location of lateral and floor or ceiling surfaces is to partition the scene into slices. The slicing needs to be done in a direction perpendicular to the sensor direction of view. It was found that vertical and horizontal slicing will cover most natural scene surfaces like walls, floor and indoor ceiling.

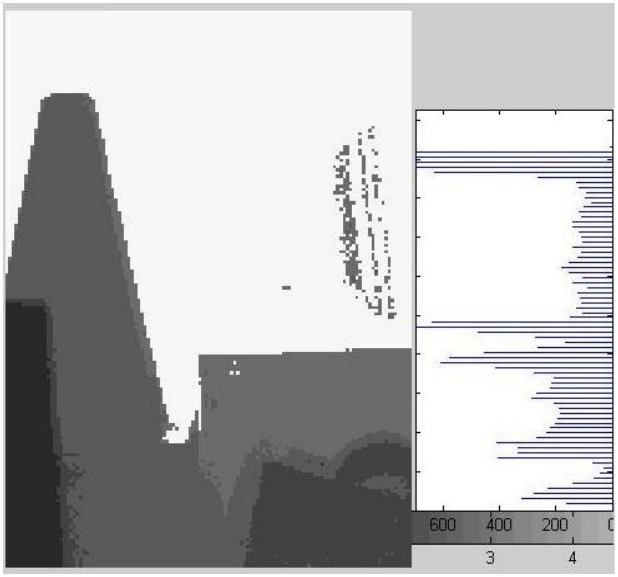


Fig. 3 Segmentation example by thresholding shows it is working for most objects in a scene. The exception are lateral running surfaces to the sensing direction.

The segmentation result will be an object represented by a set of areas with successive values depth labels. In the formal representation of the scene proposed in this paper this is not a problem. The object becomes a compound object generated with the same set of rules. The successive depth label constraint in fact aggregates the depth slices into a single 'super-object' easy to handle in associated action tasks as a solid unit.

Each slice of the image is processed separately. The slice histograms exhibit peaks for the depth on oblique surfaces when the slice width is small compared with the object running length.

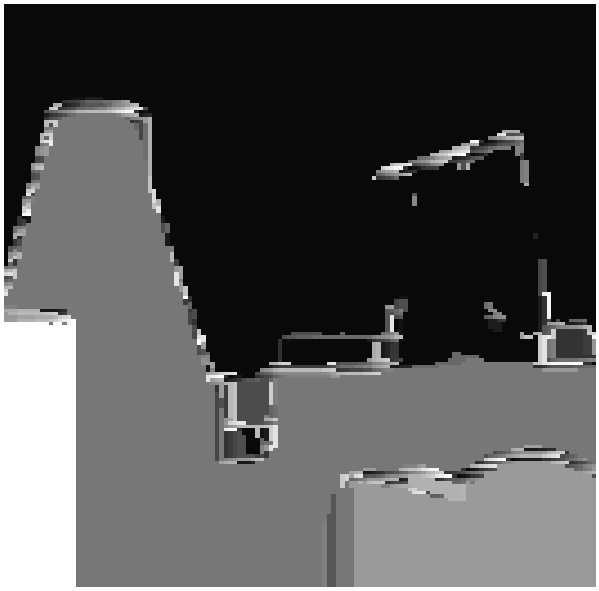


Fig. 4 Region growing segmentation of sample 3D depth image for comparison.

The pixels on the oblique surface cluster in the histogram if the slice was chosen perpendicular to the surface perspective direction. Vertical slices are effective for segmentation of lateral surfaces like walls and horizontal slices are necessary for the detection of floor and ceiling.

The processing time increases and the sliced object becomes multiple part object. The advantage of the method is that segmentation using histogram thresholding is in this way a unique robust and efficient method for all objects in the scene.

A simple parametrization of an object appearant area can be used like the rule of thump to determine the density of slices in a region.

It is also used to selects the salient semantic features essential for the determination of actions on the environment.

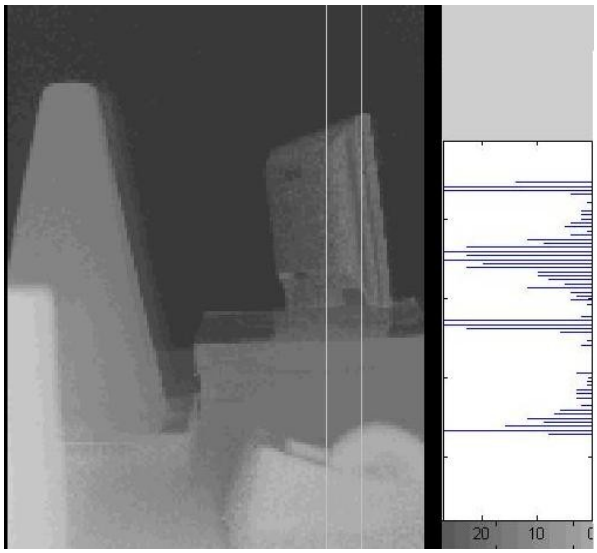


Fig. 5 Histogram based segmantation for a vertical slices of an image with lateral surfaces. The slice histogram has three peaks for corresponding objects as shown on the right .

The method of segmentation as proposed uses a 10% lateral threshold for the object separation from background. For consistency the same error margin should be preserved for each slice. Given an average dimension of an object the slice aperture must be in the order of 10% of the respective dimension.

In order to increase the detection probability several directions and slice aperture sizes must be tried and results compared for similar or different outcome.

One important observation must be made:

The slicing approach in compatible with bulk histogram segmentation in the sense that both produce the same result for objects facing the sensor.

The only inconvenience is a slight increase in the image decomposition time. A compromise among the precision, robustness and speed must be determined experimentally for each application.

To validate the model practical performance a limited study was made regarding the computation speed gain obtained. The results are summarized in Table 1.

The computation time becomes significant for larger images. At the present resolution of the 3D sensors the result is very good. The importance of computation time will certainly increase in the near future as the technology of 3D TOF sensor matures and resolution increases.

Table 1 Segmentation performance results and the estimated processing time.

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

To increase the precision of area parameter of the model more elaborate segmentation can be used but at the expense of increased processing time.

4 3D Depth Scene Formal Model

Perception and addaptivity in actions constitue the next most frequent higher level processing on the scene data aquired by sensors.

For the perception of a scene the representation of visual images by objects depth label, projected area and center location can be organized as a tree structure of objects and background.

The higher order perception logic responsible for the global decision process must subsequently make the appropriate contextual adjustments of scale.

A push down automaton algorithm of image registration can be considered. The description of the field of view as a list of objects defined by their center depth location and associated cross section area model is only sufficient for scene static description. A *regular grammar* formal model results using this approach.

In a dynamic scenario with the sensor moving in the field of view new objects will be entering the scene from obstructed areas by present visible objects.

The model for a dynamic scene as a depth oriented list of objects can not accommodate higher levels of actions in a layered form.

A more general abstract virtual model of the scene of context independent type and is therefore necessary. C*ontext-free grammars* (CFG) do the best compromise between computability and expressively. A sub class of CFG is the attribute grammars that have proven very useful in image representations [18],[19].

Although it is easy to describe the syntax a language using a context free grammar, accurate description of the language's semantics is notoriously difficult.

Attribute grammars enrich a language semantics by supplementing a context free grammar with node attributes. In this way attribute grammars add semantics to a basic context free grammar [17],[18].

Given an input image, our objective was to organize it in a representation of a hierarchical parsing graph where each non-terminal node corresponds to a production rule.

The CFG grammar we propose to represent a 3D depth image is a grammar with two classes of primitives as terminal nodes.

Each node represents a 3D objects projected on sensor 3D image of the scene. Following detection by segmentation each object is fully represented by its center location in image, its mean area and real mean depth as a label.

In the parsing graph, the vertical links represent the decomposition of scene into objects on the supporting background and the horizontal links indicate the spatial relations between components with relations among the attributes.

The *generative grammar with attributes* is a formal construct that can be defined as follows:

Definition 1: An attribute graph grammar is given by a 5-tuple:

$$\mathbf{CFG = (VN, VT , \textit{Bgd, Rules, Attributes})} \quad (8)$$

where VN and VT are the of non-terminal and terminal nodes respectively and *Bgd* (background) is the start node of the scene model.

Rules are a set of production describing the spatial relationship of the objects in the scene.

Attributes are a set of constraints that can be attached to each node in the scene.

The non-terminal nodes are denoted by capital letters P1, P2,... and terminal nodes are denoted by lower case o1, o2, ...

The CFG generative grammar has a start object denoted *Bgd* in the case presented in this paper. This is the object support of the scene (floor, table, wall etc.). The proposed CFG attribute grammar has the following definition:

$$\mathbf{S = \{ Bkg \}} \quad (9)$$

$$\mathbf{VN = \{ P\_object\}} \quad (10)$$

$$\mathbf{VT = \{ o\_[Depth\_label]\}} \quad (11)$$

$$\mathbf{Attributes = \{o\_[A, (x, y)] K , k = 1,2...N\}} \quad (12)$$

During the segmentation a 3D sensor scene is filled naturally with attributes by processing location, area parameters and depth labeling.

The usefulness of the abstract model proposed for the application taken as objective for the present paper is given by the following arguments.

The set of visual scenes perceivable by a observer is obtained by learning the semantics from the inherited attributes of the abstract attribute grammar.

Assessing adaptivity of the system through perception by a user of the sensor in this universe is direct using this model. In its simplest form a direct translation of the attribute tree into actions like approaching or avoidance objects in environment navigation is given by the graph of attributes.

The attainable performance for the model is expressed in the following proposition:

Proposition 1. The abstract attribute grammar model of the 3D sensor defines the *universe of visual scenes perceivable* semantically by the a sensor user.

This property is very useful in determining the all the scenes that will e successfully represented from the larger class of scenes including the ones with some degree of ambiguity.

In place of a formal demonstration of the proposition an immediate conjecture demonstration can be given applying formal language theory basics.

A grammar actually generates the language it models. Thus any phrase in the language can be parsed by the grammar, hence the user can perceive it by extracting the associated semantics.

The learning of the semantics inherited from the grammar in its most simple form is the projection of the grammar rules augmented with attributes into real environment actions using the determined attributes: coordinates, area and depth labeling.

No supplementary processing and supplementary semantic discovery is necessary making the model a self contained representation ready for use in higher order representation structure.

It is easy to observe that the formal generative model proposed with the attribute extension translates naturally to an independent action plan on the objects detected in the environment. The actions of *grasping* or *avoiding* of any of the objects are guided directly by the object attributes.

The object center indicates the action vector while the depth label outlines the cost of the action. Partial obstruction in the line of view translates to the clearance channel in navigation.

In the case of an application with human perception the abstract model as proposed as is the case of a visually impaired person the parsing tree as description of the scene translates naturally to sounds .

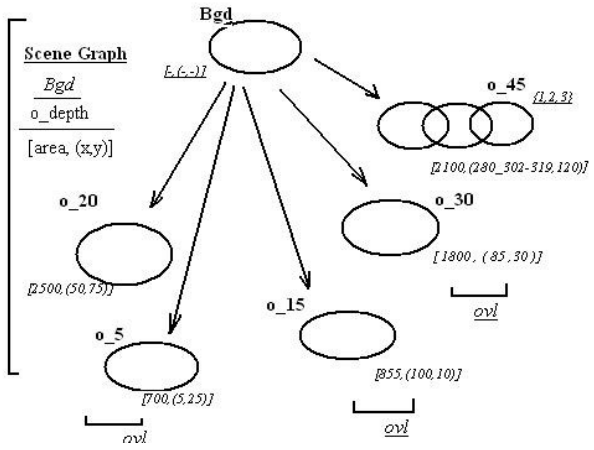


Fig. 6 The parsing tree for the scene in fig. 4 representing depth labeled objects and its attribute grammar graph structure.

The objects depth, area and center individualize the object in the scene. The case of objects with oblique surfaces there will be a set of successive objects that must be translated to a continuously variable sound to represent the angle of its surface in view in the scene.

5 Range Flow in 3D Depth Images

Motion estimation in 2D video images, a widely researched area is covered at present by remarkably good results demonstrated with computational efficiency and robustness. The motion estimation for 3D range images despite intense work in the last years is still in the search for a efficient and robust solution especially in real time applications.

Many recent papers are following the 2D image success methods of neighborhood correspondence in successive frames for range flow determination [20]. The method produces a dense field but is computationally very intense.

The approach of the present paper is focused on salient objects in the scene and is real time oriented. As such our algorithm is based on attributes of objects calculated over a sequence of frames in the past and present. As such the approach can be

considered a global type [21]. The work reported is tuned to the very recent extensive research for using 3D TOF sensors in automotive applications.

The principal contribution as proposed in the present paper is an extension of the object parametrization by area and center of cross section to a sequence of frames. The sensor radial range flow for each represented object is calculated from the latest few frames object depth parameters data stored in queues as presented in Fig 7.

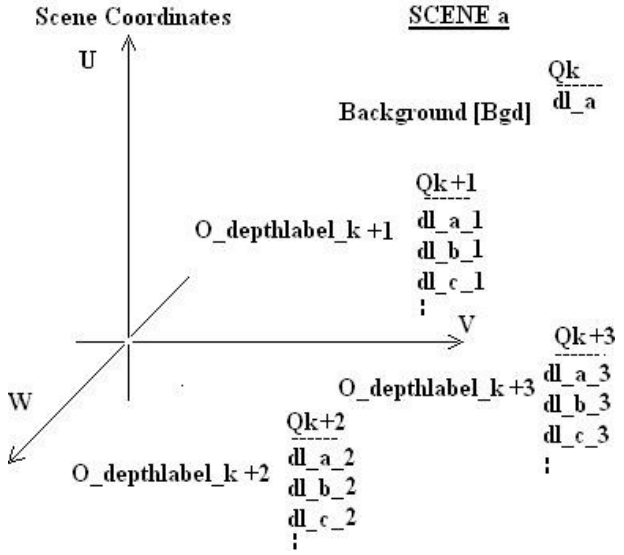


Fig. 7 Determination of radial range flow from object parameter queues.

Each queue will hold the depth_label of the object for the last frames. The number of frames is minimum two but using three or more will increase the robustness of the model with ought a major impact on the hardware resources required

The radial flow in the sensor reference can be calculated directly from the depth data. It is of central interest in applications of reaching for grasping or avoiding objects from the scene:

$$v_{RADIAL} = delta_{DEPTH} / inter\text{-}frame_{TIME} \qquad (13)$$

The tangential component is of lesser interest and its determination needs translation from sensor matrix coordinates into scene coordinates.

It can be shown that most of the traditional intermediary operations on the frame pixel values are preserved [21].

Individual pixel induced noise is filtered in the area determination stage from the histogram. A second stage that disperses local noise is the calculation of the mean depth for the object.

The robustness of the radial velocity calculation can be further improved by averaging over more then the last two calculated inter-frame mean depth values to three or more frames. The penalty paid in this case is a increase of the aperture of time for event localization.

The proposed model is able to cope pretty well with certain situations that are difficult to cope with in a dynamic environment.

The following two special situations have been incorporated into the model:

a) - objects entering and exiting the sensor field of view;

b) - partial or complete occlusion of some objects by the other;

In the first case a) the area of the object as it enters and departs from the sensor field of view is changing in time. It is a phenomenon that is not easy to cope with in 2D image processing. In 3D depth images the model as proposed presents neither scene spatial structure nor semantic interpretation as fading away on the sides of center of attention (area reduction) is natural especially if the interpreter is human.

The partial occlusion of case b) is similar to exiting the scene. The presence in the model of the queues will resolve the problem in many cases. Object retains identity due to queue and can do so for a longer period if the queue is set longer to begin with.

A special value of the *area parameter* avoids its use resulting in calculation errors. Object relative velocity is calculated based on the existing valid depth data in the queue only.

The sensor relative radial velocity of each object is calculated given the inter-frame period. If this is data not available or is not desired to be involved the advancing or receding speed can be calculated in depth units per frame.

Any one of the two possible situations support however the same semantic interpretation of progress in the grasping or avoiding task.

The model is very useful in the general case with objects at medium distances. In the final stage before the moment of grasping a supplementary detail processing may be necessary to increase the accuracy and avoid instability.

The abstract model of *context free generative grammar with attributes* is naturally extended to describe short period scene dynamic. Due to the fact

that the relative radial velocity v_{RADIAL} is calculated directly from object data parameters it qualifies as a *context free attribute*.

In the following the abstract model extension with the supplementary attribute is presented.

The Attributes rule extension from static case to short term dynamics form becomes:

$$\mathbf{S} = \{ \mathbf{Bkg} \} \quad (14)$$

$$\mathbf{VN} = \{ \mathbf{P\_object} \} \quad (15)$$

$$\mathbf{VT} = \{ \mathbf{o\_[Depth\_label]} \} \quad (16)$$

$$\mathbf{Attributs} = \{o\_[A, (x, y), v_{RADIAL}]K, k = 1,2...N\} \quad (17)$$

where the radial velocity v_{RADIAL} has been added as a supplementary attribute.

There are other candidates for the same procedure in order to have all data stored and do no translation or processing at the semantic interpretation stage.

One such parameter is the *time to collision* term used in robotics that can be naturally incorporated in the presentation for semantic interpretation.

In Fig 8. presents a sample 3D typical scene used for method validation after the decomposition in static objects with static attributes.

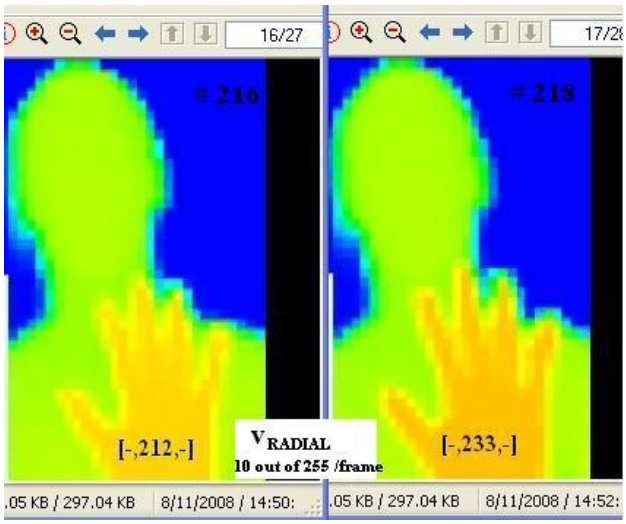


Fig. 8 Dynamic scene radial velocity attribute example: receding hand at a radial velocity of 10 out of 255 units/frame. [http://www.pmdtec.com]

The two frames differ only in the depth scene W true coordinate used for velocity determination. The middle frame was omitted to verify the hypothesis that the area changes are very small.

In this paper the grammar has a single level of objects derived directly from the start node. This is true for the simple sample problem exemplified for the illustration of the concept as well as for more complex scenes with many objects.

The model can be easily extended to multiple levels in the graph tree for scenes with clusters of object supported by a basic object like a table.

Further development effort is needed to extend the model for more demanding complex applications.

6 Conclusions and further work outline

A perception oriented formal linguistic model for 3D sensors depth images is proposed.

The 3D sensor used in the present work is a new CMOS TOF device generating both luminosity and depth images.

For object extraction from the depth image a segmentation based on histogram thresholding is proposed and extended to oblique surfaces by dividing the image in slices. Each slice is the segmented based on histogram thresholding.

The segmentation robustness and minimal necessary computation cost where the central constraints considered. The choice of global segmentation as proposed was determined by a minimal computational cost for real time applications.

For each object a model of the depth image in terms of object area and center location is proposed. The object mean depth parameters is used as a label for object identification in the scene.

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

For scene content structuring an original formal language representation is proposed. A context free grammar is shown to be natural to 3D sensor data. The extension of the context free grammar with attributes adds structure to the model.

It is also shown that the generated language translates directly the depth labeling into action planning on the environment a case of interest in mobile robots applications.

Human perception of scenes in terms of detected objects was proved to be also naturally derived from the attributes that add spatial semantics to objects. This is true for the simple sample problem exemplified for the illustration of the concept as well as for more complex scenes with many objects.

The paper also proves that the model can be extended to selectively include short term dynamic

attributes that can be mixed with the static one. The objects radial velocity named here range flow is exemplified as such a case.

The model can be easily extended multiple levels in the graph tree for scenes with clusters of object supported by a basic object like a table.

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

References:

[1] T. Winograd, Interaction Spaces for 21st Century Computing , *Human Computer Interaction in the New Millennium,* John Carroll, Ed., Addison Wesley, 2001, ISBN - 0201704471

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

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

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

[5] 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 Applications Berlin*, March 11-12, 2006

[6] I. Jivet and A. Brindusescu, Real Time Representation of 3D Sensor Depth Images, *WSEAS Transactions on Electronics*, Issue 3, Vol 5, March 2008 pp:65-71

[7] M Otesteanu, V Gui, 3D Image Sensors, an Overview, *WSEAS Transactions on Electronics*, Issue 3, Vol 5, March 2008 pp:53-56

[8] K Boehnke Fast Object Localization with Real Time 3D Laser Range Sensor Simulation, *WSEAS Transactions on Electronics*, Issue 3, Vol 5, March 2008 pp:83-92

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

[10] C. Bamji, E. Charbon, "Systems for CMOS-compatible three-dimensional image sensing using quantum efficiency modulation," US Patent 6,580,496, granted in 2003.

[11] 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

[12] 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, 2001

[13] 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 .

[14] 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

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

[16] T. Kadir, M. Brady, Scale, Saliency and Image Description. Timor Kadir and Michael Brady. I*nternational Journal of Computer Vision*. 45 (2), November 2001, pp. 83-105.

[17] M.S.Ryoo, J.K.Aggarwal, Recognition of Composite Human Activities through Context-Free Grammar based Representation, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,* New York, NY, 2006, pp. 1709-1718.

[18] Feng Han, Song-Chun Zhu Bottom-up/top-down image parsing by attribute graph grammar, C*omputer Vision, 2005. ICCV 2005. Tenth IEEE International Conference* Vol. 2, 2005 pp. 1778 - 1785.

[19] Y. Sakakibara, Learning context-free grammars using tabular representations, *Pattern Recognition,* Volume 38, Issue 9, September 2005, pp: 1372-1383.

[20] S Matzka, Y Petillot, A Wallace Fast Motion Estimation on Range Image Sequences acquired with a 3-D Camera *British Machine Vision Conference 2007*, University of Warwick, UK, Sept. 10-13, 2007

[21] X. Jiang, S. Hofer, T. Stahs, I. Ahrns, and H. Bunke. Extraction and tracking of surfaces in range image sequences, *Proceedings of the 2nd International Conference on 3-D Digital Imaging and Modeling*, pp: 252–260, 1999.