Multiple 3D Sensor Views Object Models Correspondence

I JIVET, V GUI, A BRINDUSESCU
Faculty of Electronics and Telecommunications
University Politehnica Timisoara
V Parvan No 2, Timisoara
ROMANIA

ioan.jivet@etc.upt.ro, vasile.gui@etc.upt.ro, alin.brindusescu@etc.upt.ro, http://www.etc.upt.ro

Abstract: - The paper presents the results of an investigation of the correspondence problem of scene object models obtained from multiple 3D sensors depth images. The recently developed 3D TOF (time of flight) sensor characteristics are analyzed relative to content of depth images produced. A model of the scene space for a pair of sensors is given in an adjusted epipolar geometry setting. The scene decomposition into objects according to a 3D specific recent segmentation is proposed. The determination of scene content and joint correlation of models by object correspondence is outlined. Visual hull method usefulness for object volume determination from occluded space is analysed. For a concrete experimental case the scene correspondence results of two 3D sensors is presented. Further work is also outlined.

Key-Words: - 3D time of flight sensor, multiple view, view space model, object correspondence

1 Introduction
Real scenes observed with sensors are populated with objects that are fixed in some applications as they can also change relative position due to observer motion or a different view from a different sensor.

The actions most frequently needed in applications are of avoiding obstacles or reaching to grasp objects. The scene content correspondence needs to be resolved. The reward for the effort is the corrections of positions in the scene.

The newly developed 3D sensors provide aside from a light intensity common image of the scene a distinct depth image [2], [3], [11].

Multiple view geometry and scene views consistency are problems intensely studied in the last period by researchers in the field of image processing.

The multiple views from 3D depth sensor is still an open filed with great potential due to the spread of use 3D sensors in many applications [4],[12].

The work reported in this paper made no assumption were made on the localization of the sensor or on its fixed or mobile statute. The general case of a randomly placed sensor pair in communication with the others is considered [14].

2 Characteristics of 3D TOF Sensors
Recent literature reports intense interest in the use 3D sensors in various applications from mobile vehicle navigation to automotive applications [6].

For the last period of time CCD or CMOS stereo cameras have been the sensor of choice for most applications using visual sensors.

A well known problem with CCD or CMOS 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 local depth measurement error.

Laser based 3D sensors have been in use for some time but their cost and sensitivity excluded them as candidates from wide use.

The latest developments in CMOS Time of Flight (TOF) 3D sensor increased the interest for their use in real time applications. The sensor is fabricated using the standard CMOS process integrated circuits and therefore cost effective [6], [7].

The CMOS 3D TOF sensors outputs for each pixel depth data as well as intensity of the reflected light. The new sensors about to be mass produced are likely to become a very common sensor due to their simplicity in use and cost.

The 3D TOF sensors do have advantages in comparison with other sensors used for depth estimation. The resolution of the depth image obtained with ought processing permits light miniaturized complete systems.

Stereo cameras require intensive post processing to construct depth images since data it is obtained by triangulation. TOF sensor post processing usually involves a simple table-lookup to map the individual pixel sensor reading to real range data.
The results reported in this paper where obtained using a SR 3000 3D TOF sensor.

The sensor is using 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.

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 x 144 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 calibration.

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 meters.

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.

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.

3  Space Model of 3D TOF Sensors

The sensor model and the visible part of the space through it are the prerequisites for its exploration from acquired data. Similar to all imaging sensors the 3D TOD sensor model is a projective space of a specified scale (focal distance) from the 3D scene on a plane.

The sensor performs two operations:
- derivation of a projective 2D optical image of the 3D space in the viewing angle;
- in the depth image each pixel outputs a value proportional to the projective distance;

Calibrated sensors adjust the proportionality constant to the optical system focal distance.

From the projective characteristics of the sensor a model is immediate. The real dimensions of objects in the projected depth scene correspond inverse proportionally to the value of depth in the image:

\[ X_{dim.3D} = X_{dim.2D} \times k \times Depth_{-object} \tag{1} \]

where \( k \) is a factor dependent on sensor optics.

Given a typical 3D sensor range of 10 m the precision of positioning in the near field (typical error order of a few milimetres) will have an effective viewing area a factor of 10 larger then the objects in the far field of the scene.

Observation of the same scene objects from more then one viewing point with multiple sensors can be used to compensate this degradation in position accuracy for different depth in the scene.

The case of a two sensor analyzed in this paper is similar in many ways to the stereo camera geometry for small angle separation. The epipolar geometry model can be used but supplemented with the depth information inherent to 3D depth sensors in addition to the projective image.
There are two issues to be resolved in order to determine the depth of the Object POINT from the two images:
- Object POINT – correspondence in the two images on the basis of its properties and those of its neighborhood;
- Position of the two cameras by their projection center;

The stereo camera problem statements above is reversed in the case of two sensors observing the scene. The depth of the Object POINT to each sensor is furnished by the sensor.

Figure 1. The experimental set up of the two 3D sensors in a real life scene with regular geometric background.

The positions of the two projective centers are considered be known in epipolar geometry. This is not at all the case with a pair of sensors that can be both in relative motion.

The relative position of the projective centers of the two sensors is the unknown of importance in applications with multiple sensors.

The determination of the projective center position from the depth information becomes the problem to be resolved as needed in the application.

In epipolar geometry the relation of the two cameras is given in terms of the fundamental matrix. It is the cross product of a translation and a rotation. From the knowledge of the depth information only the translation part can be determined. The rotation part needs additional constraints to be resolved.

Additional constraints can be obtained for the same pair of sensors if more objects in the scene are considered. From the background a reference wall can be selected. Relative orientation of the camera with respect to a normal to the reference plane contains the rotation information.

The correspondence problem however is more difficult to solve if the normal light image furnished by the sensor is not used. All the visible points of an object are characterized by their depth value only.

A characteristic point of the object in view uniquely determined is a better choice for its representation. The simplicity and repeatability of its determination is of great importance in applications.

In accordance with the approach proposed in the present paper of scene object representation by mean of depth and cross section area the median area center of the object is the characteristic point of choice.

One problem with the above choice is the dependence of the cross sectional area of the object on the position in the scene of the second sensor.

For a second sensor with a small rotation and hence also a small translation relative to the position of the first sensor this approximation may be negligible or acceptable.

The area changes with depth and symmetric deformations due to rotation have no influence on the mean cross area center.

A application envisioned for a pair of 3D sensor arrangement is the quick determination of the pose and orientation of a tested object in the scene. One of the sensors acting as reference will be set with an orthogonal orientation to the reference wall.

The mobile sensor aligned with the object to be observed will provide the position by depth determination and orientation by rotation of sensor coordinates to the normal to the reference wall.

4 Scene Content Determination

Analysis of image content and determination of objects in it is the central most difficult part of image processing and the 3D depth sensor images is not an exception.

Techniques form the long term experience in image processing can be used with depth images after careful review.

The 3D sensor depth images are explicit due to the presence of geometrical information. Every pixel of the depth image is not an abstract set of values but a geometrical measure of distance from the sensor to the surfaces of objects in the scene.

Many image segmentation methods are used in image processing. The region growing segmentation method is one of the more robust to noise methods. Object edge detection and subsequent aggregation to
represents objects is also very often used but is more sensitive to noise.

For 3D depth images a recently proposed segmentation method was found simple to implement and robust [4].

The algorithm is well suited for scenes with a small number of components selected as dominant. The algorithm consists of a two step process. In the first step global depth histogram is segmented by thresholding. In the second step the floor and lateral field of view enclosures are determined.

In Fig 2. is presented a sample 3D depth segmentation based on histogram thresholding. The peaks of the global histogram where determined as a sum of slice histograms.

Sliced histogram with a factor of 10 –15 per frame have been used to select the depth regions in the scene. A median filtering was applied to the image prior to thresholding for noise reduction.

Floor came out approximated by a set of steps. The artifact can be turned into a feature if properly used.

Areas with depth values in the selected range receive a labeled with the average value. The mean center of the area is taken as the representative point of the object.

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.

For tighter error margins a higher resolutions sensor is needed. Better resolution sensors have been announced for the future as the technology matures.

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.

The same problem occurs when segmentation is carried out using the region growing algorithm. The depth values for adjacent pixels do not cluster in a narrow range for surfaces oblique to sensor 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.

The slice histograms exhibit peaks for the depth on oblique surfaces when the slice width is small compared with the object running length. It was found that vertical and horizontal slicing will cover most natural scene surfaces like walls, floor and indoor ceiling.

The segmentation result will be an object represented by a set of areas with successive values depth labels. The object will result as 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.

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 for all objects in the scene.

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. The computation time

Figure 2. Real scene object detection using a SR 3000 3D sensor by histogram thresholding.
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 will increases.

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

5 Multiple Views Model Correlation
A 3D depth image in principle contains information of localization for all objects present in the scene. There are however reasons and situations that impose multiple view fusion or at a minimum communications on object out of viewing angle.

The work reported in this paper made no assumption were made on the localization of the sensor or on its fixed or mobile stature.

For the purpose of communication the description of one scene from any sensor in the network of the relative position in its coordinate system is the only information available.

Based on this information the correct correspondence is necessary to be carried out. Thus each sensor can add the positions of the other sensors in his model and possible the localization of the occluded objects in its view [19].

The problem of object correspondence once resolved only tracking the objects in the scene must be resolved in subsequent frames.

Representing the scene as an abstract space with objects as points in 3D coordinates each sensor view is an affine space with no common reference.

In the present paper the object correspondence is analyzed from two points of view:

- correctly identifying of objects detected for a coherent communication in between sensors;
- solutions for partially (or totally) occluded object sharing and object volume approximation;

Object correspondence from multiple views is an inverse problem to Depth from stereo. It debuts by locating the corresponding objects in the two depth images.

The features to be matched, in the correspondence of multiple sensor is of two basic types: methods using feature or parts (of area) with depth values of the objects;

Feature-based methods are using for the match process certain object features like contour edges, colors or form. The features are extracted form the object in a image analysis process.

Some of the features are better suited for the use in the matching. The features best candidates are the ones that do not change much from one view angle to the other.

The feature based matching methods advantage is that they can be processed in a short period of time. The criteria used in matching used in determining the correspondence use a measure of the norm type over the metrics of the feature space.

These methods increase the robustness of the matching process at the cost of processing time. [17] The matching feature for multiple 3D depth sensors the object attributes determined during scene analysis: cross sectional area, its center coordinates and depth.

One pair of sensors at a time with two different viewing angles perform a systematic attribute random check until all correspondence are found. The area needs to be scaled to correct for the variations of area (dimension) detected in the projective image by equation (1).

Object volumes are not known due to frontal nature of the image from the 3D sensor. Visual hulls can be used in multiple 3D sensor for object localization in the scene being a method to limit the hidden range occluded by scene objects.

Visual hulls are obtained by the intersection of the object contour (silhouette). Cones are constructed from each viewing center and their intersection determines a geometrical volume with minimal additions to the real 3D object.

Figure 3. Object localization from the two views in the experimental set up view using a visual hull.
In the process of multiple sensor depth image acquisition the sensors image capture is obtained using a simultaneous trigger signal. Sensor synchronization over time in dynamic environments can create correspondence problems. The depth images sequence synchronization guarantees the images from different views belong to the same time instant. Accurate time stamps on images is one method to assure the use of same time images for 3D localization of quality.

The results of the scene analysis for sensor B is presented in Fig. 4. A graph with objects extracted from depth image is labeled with attributes: depth, area and view coordinates. The depth is color coded in gray level as obtained form the sensor, the area is represented in pixel counts as is the center of the cross section area.

The graph for sensor B was used as a reference for the resolving of correspondence in the scene view of sensor A. Its displacement and rotation around the vertical axis was determined by the distance to center and angle to its normal.

The subject volume for the case studied came out to be very close in form to a cylinder due to similar cross sectional profile. The visual hull for more complex scene objects shapes needs additional work.

The case studied on the basis of available 3D sensor data did not addressed the case of a scene with objects at extreme depth.

Further work is also necessary for the multiple sensors positioned on opposing sides of the objects. The process of translation and rotation is ambiguous to the distinction of left and right oriented system of coordinates.

In all the work done on the 3D sensor depth images the normal light image was never used. The parallel processing of the two images furnished by the 3D sensor is due to resolve many problems.

6 Conclusions

The paper presents revised solutions to problems specific to newly developed 3D sensors in applications using more then one view.

The results of 3D sensor specific correspondence problem of scene object models obtained from multiple 3D sensors depth images are presented. The 3D TOF sensor operation is analyzed relative to depth images produced and their information content. A model of the scene space for a pair of sensors is given in an adjusted epipolar geometry setting.

The scene decomposition into objects according to a 3D specific recent segmentation is exemplified. The scene correlation of content form multiple views when multiple sensors are needed is defined.

Visual hull method is proposed and exemplified for object volume determination. For a concrete experimental case the scene correspondence results of two 3D sensors is presented. The proposed model weak parts are indicated and further work is also outlined.

References:


[5] Z. Lin, J. Jin, H. Talbot, Unseeded region growing for 3D image segmentation, Conferences in Research and Practice in


