The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Head-Mounted 3D Multi Sensor System for Modeling in Daily-Life Environment

Hiroaki Yaguchi, Kei Okada and Masayuki Inaba

Abstract— **Model-based approaches in recognition and planning of robots work effectively, and these approaches can apply to model-less situation using autonomous model construction by an agent. There are problems about segmentation or shape fitting of various objects with different scales or shapes. In this paper, we construct a Head-mounted 3D multi sensor for 3D environment modeling and propose a method of 3D reconstruction for various objects using intentional behavior of human.**

I. INTRODUCTION

In recent years, the growth of robot technology is remarkable; many robots have humanlike manipulation and motion capabilities as like humanoids and can work both moving tasks and handling tasks. Robots working several different tasks need different levels or scales about recognition for each task. Especially in daily-life environment, moving tasks in room and handling tasks of hand-held objects are demanded and models are necessary for these tasks, but room environment and hand-held objects have different scales and tolerance, so there is different means of results in spite of same mathematical solution can be applied. So there are needs for modeling approach manageable different tasks comprehensively.

We define a Daily-life environment modeling; translation to inner representation of computers about various objects in daily-life environment using measurement by sensors. Modeling approach of real environment is useful for applications about recognition or planning, but traditionally those are based on manual measurement or transformation by human. In this research, we attempt to autonomous modeling by 3D measurements.

A big problem of automating in modeling is daily-life environment is constructed by various objects which have different scales or shapes. In this research, we propose an approach using physical intervention by human to separate object from environment and make attention to object, and construct Head-mounted 3D multi sensors that can measure both human behavior and 3D environment. In recent research Kojima et. al [1] proposed a method of estimation joints or movable parts of known furnitures supported by human manipulation, we extend this solution to create models thmeselves with shape information. We also propose a method of solid modeling using hypothesis from category of object, and present about some methods: pose estimation of sensor, measurement 3D objects, and shape fitting for hypothesis.

This paper is constructed as follows: related works are described in next section. Construction about head-mounted 3D multi sensor is described in section III. Daily-life environment modeling approach is explained in section IV. Results about modeling experimentations are shown in section V. Finally section VI concludes this paper.

II. RELATED WORKS

A. 3D MODELING

There are many approaches proposed about 3D object modeling. For example,factorization [2] is mathematical reconstruction methods in computer vision. Plane segment finder [3] is boundary fitting approach using stereo vision. There are also approaches about solid model fitting to generalized cylinder [4] and ACRONYM [5]. But identity or physical boundary of objects can not recognize only from shape information. Modeling approach for recognition about attention and model identification is needed.

Differences of modeling representation are caused by demand for ability of aplication. E.g. MonoSLAM[6], Photosynth[7]using Snavely's approach[8], and VideoEnhancement by Bhat et. al. [9] represent 3D information as 3D point cloud. In MonoSLAM aproach they attach importance to estimate continuous camera poses. Photosynhth and VideoEnhancement aim to image rendering from sparse 3D information. These approaches have no consideration about reuse memorized 3D information to recognize identity of objects. On the other hand, view-based approach [10] realized to reuse visual information by memorize images with no 3D information. This suggests images and visual features are essential to reuse object models. It is necessarynot only 3D shape information but visual features to models.

B. RECOGNITION OF HUMAN BEHAVIOR

Intentional sensing [11] is an approach to switch a precondition about sensing by recognize human intention. For recognize intention, sensors must be able to observe human behavior. Observation approach about human behavior, e.g. robotic room [12] embedded sensors in environment, is observe from third person viewing. Human behavior observation approaches from external sensors have advantage to recognize whole body behavior, but intention or attention of human can be observed indirect, it must be observed by behavior estimation. On the other hand, wearable sensors can observe directly intention by sharing attention with equipped human. Especially, head-mounted camera provides an equivalent viewing of humans, so it can observe automatically handling or noticed object by human.

Authors are with The Department of Creative Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Build. eng. 2 73B2, Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan, 133-8656 {h-yaguchi,k-okada,inaba}@jsk.t.u-tokyo.ac.jp

III. HEAD-MOUNTED 3D MULTI SENSOR SYSTEM

There are same concepts about head mounted cameras in wearable or ubiquitous computing. Steve Mann's personal imaging [13] is the epoch of wearable computing. In recent, Tsukizawa et al. [14] developed a 3-D reconstruction system for hand held objects using head-mounted stereo vision.

Recently head-mounted vision approach aimed to "capturing images from near human's view point". When using images similar to human's viewing, system can search from database or construct 3D model about noticed object by human. In this paper we aim to reconstruct loci of human walking using images in walking behavior. In recent works there is small attention area such as reachable range of human hand, we consider to movable range of human walking as attention area.

Fig. 1 shows a constructed sensor system. The system is composed of a stereo vision sensor (SVS STH-DCSG-C made by Videre Design Inc. : (A) in Fig. 1), two LRFs (URG-04LX made by HOKUYO Inc. : (B) in Fig. 1) and a motion sensor (A3U9S made by NEC Tokin Inc. : (C) in Fig. 1). Multi sensors are connected to backpacked PC (Mac Book Pro made by Apple Inc. , 2.53GHz Core2Duo CPU, 2GB Memory), and obtained information provide to human using HMD. System is also equipped Bluetooth remote controller for interaction with human to notify intention.

Stereo vision is a core of system. Stereo camera can obtain 3D information by 1 frame capturing, this is an advantage from monocular vision obtaining from plural images. Camera module is connected by IEE1394 to PC, frame rate is 15[fps](in VGA capturing) or30[fps](in QVGA capturing). Each camera has 81.2° view angle.

Laser range finder can obtain high accuracy range data on 2D plane. It is difficult to dense 3D information from single sensor because of limitation in measurement range, so we attempt to equip two sensors in different angle. These sensors can obtain a part of floor and walls which have sparse textures as 2D lines. Frame rate of these sensors are 10[fps].

Motion sensor is used to estimate sensor pose without measurement from external sensors, is composed by acceleration sensor , gyro sensor , and magneto sensor. Only this sensor is affected drift in equipped sensors, so estimated pose from this sensor is used as initial pose for estimation using visual odometory with stereo vision. Frame rate of this sensor is high speed against other sensors, about 125[fps], it is obtained synchronous with stereo vision.

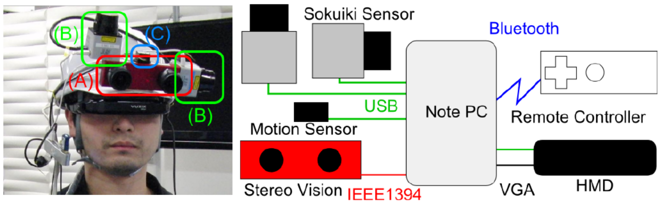


Fig. 1. Head-mounted 3D multi sensor system

IV. SOLID MODELING IN DAILY-LIFE ENVRIONMENT

In this section, we propose a method of solid modeling by primitive shape fitting with 3D point clouds for various objects with different scales and shape in daily-life environment. 3D point clouds are obtained by merging between measurements in different frames using visual odometry pose estimation. Also we construct a method of shape fitting using combination shape features of artificial objects: plane feature and symmetric feature.

A. VISUAL ODOMETRY

A continuous camera poses estimation method is constructed by 3D feature tracking using KLT tracker[15] with stereo depth estimation [16].

Feature points selected or tracked by KLT tracker are estimated 3-D position on current camera coordinate using stereo disparity. Relations between measured points and camera poses for each time t are:

$$ {}^{t}C = {}^{t}T_{t-1}{}^{t-1}C \qquad (1)$$

$$ {}^{t}p_n = {}^{t}T_{t-1}{}^{t-1}p_n \qquad (2)$$

where ${}^{t-1}T_t$ is a homogeneous transformation matrix of camera pose from $t-1$ to t, ${}^{t}p_n$ is a nth feature point on a camera coordinate in t, ${}^{t}C$ is homogeneous representation of camera coordinate. Eq. 2 is not equivalent because of measurement errors. An error minimization equation is:

$$\min \sum_{i=1}^{N_p} ||{}^{t}p_i - {}^{t}T_{t-1}{}^{t-1}p_i||^2 \qquad (3)$$

where N_p is a number of all feature points. This equation can be solved by closed form using Dual Number Quaternion[17]. However least mean square is sensitive for outlier, therefore we employ least median for robust estimation. Finally the minimization becomes:

$$\min \text{median}||{}^{t}p_i - {}^{t}T_{t-1}{}^{t-1}p_i||^2 \quad (i = 1 ... N_p) \qquad (4)$$

B. PLANE SEGMENT FINDER

We apply plane fitting to 3D point clouds using plane segment finder [3]. When a plane L_p is defined in 3D space, likelihood betwenn L_p and 3D points is defined as:

$$v_p = \sum_{i=1}^{N_n} \exp(-0.02 d_i^2) \qquad (5)$$

where N_n is a number of 3D points, and d_i is distance betwenn the ith 3D point and L_p. L_p which maximaize this likelihood is estimated by paricle filter.

C. SYMMETRICAL SEGMENT FINDER

We propose a method of symmetrical shape detection from 3D points with definition likelihood about planar symmetry based on a method of Podolak et. al. [18]. Fig. 2 shows an image of evaluation of likelihood. 3D points P_0' are created mirroring from point clouds at one side P_0 when point clouds P_a are separated to plane L_x. Then nearest neighbor search is applied to all points of P'_0 and P_1, and the matching point exists when distance between current point and nearest neighbor d_i is below constancy. This calculation apply to P_1 similarly. If assuming the number of points where the matching point exists to be a likelihood, it is possible to catch with a plane search that maximizes the likelihood Thus likelihood v_x of symmetric plane L_x with P_a is:

$$v_x = \sum_{P_a}^{p} \begin{cases} 1 & (di \approx 0) \\ 0 & (\text{otherwise}) \end{cases} . \quad (6)$$

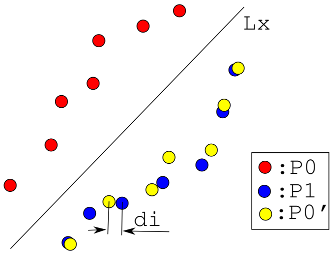


Fig. 2. Planar symmetric likelihood

D. RECTANGLE DETECTION

After deciding base coordinate using plane segment finder and symmetric segment finder, we apply rectangle detection to 3D points to detect a face composing cube. Fig. 3 shows an image of rectangle detection. Likelihood v_{rect} of rectangle and points P is defined as:

$$v_{rect} = \sum_{i=0}^{N} \exp(-0.02 d_i^2). \quad (7)$$

where N is a number of P, (x, y) is center of rectangle, w, h are width and height of rectangle. Rectangle fitting is realized using particle filter that maximize the likelihood v_{rect} with parameters (x, y, w, h).

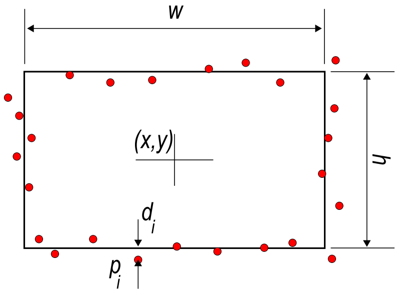


Fig. 3. image of rectangle detection

V. MODELING EXPERIMENTATION IN DAILY LIFE ENVIRONMENT

In this section, we present some experimentation and evaluation about proposed modeling methods for different objects in daily-life environment. Methods of modeling are switched from intention notified using remote controller.

A. ROOM AND DOOR ESTIMATION

We propose a method of model construction of room and door. Fig. 4 shows the algorithm of method. Images of experiment are shown in Fig. 5.

Firstly, we obtain an outward form of a room. A size of room is estimated by detecting walls on all sides when room is defined as a large cube. Human faces for each wall as sequence of obtaining all sides walls. Sensor direction is obtained from magnetic azimuth and gravity vector, and all range data are projected on 2D ground plane. If walls of another side are detected simultaneous, walls can registry. An accerelation vector and a horizontal laser range finder are used in wall detection. Range data considered slant of sensor system can be obtained when human is stopped and gravity vector is estimated using acceleration vector from motion sensor. Walls estimated by robust line fitting to laser range data. In this experiment human stands on near center of a room, and room size can be estimated by estimate walls on all sides sequentially. In this expreiment, estimated size of room is 9680 × 5040[mm], this includes about 1% errors from ground truth 9750 × 5000[mm].

Secondly, we construct a model of door. When door is closed, door is cannot detected because the door is buried in the wall. Thus we propose a method using opening door behavior of human. We apply door size estimation. A door should be opened for door estimation. In our method, door is estimated by measurement with facing to opened door. It consider to difficult about estimation using images because of door is constructed plate, has sparse texture, and is larger than human. So estimation method uses laser range finder similarly to room. Fig. 6 (a) shows constructed door model. Estimated size of door is 802 × 2350[mm], includes about 2% error from ground truth 800 × 2320[mm].

Thirdly we propose a method of estimation knob position and addition to constructed door. Knob's precondition is to handle by human hand, thus knob position considered from hand detection using measurement of handling behavior. At this time, there is problem about which side is fixed edge. Knob's position must be opposite side from fixed side, so fixed side is considered by knob positioning simultaneous. In our method, knob's position related with door is estimated by registration with fixed side and bottom edge after knob estimation using hand position detection. Fig. 6 (b) shows a result of knob positioning. In this experiment estimated height of knob is 1060[mm] includes over 100[mm] from ground truth 920[mm]. It considered a hand position shift to upper because a part of arm is included hand region in hand detection.

Finally relative positioning between door and wall is needed because door is fixed to wall. Fixed position of door is estimated using registration door with wall. When door is opened, space that is same size of door appears in wall, so door position is estimated by measuring space. However, we must consider about fixed side of door and opening direction of door. When door open to outside of the room there is no problem, but door open to inside of the room, wall detection

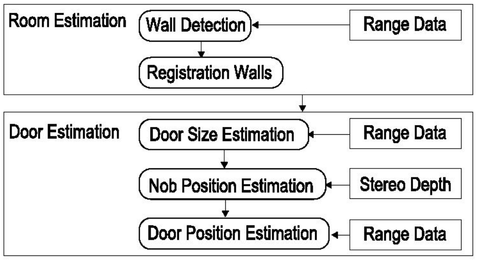


Fig. 4. Algorythm of room and door modeling

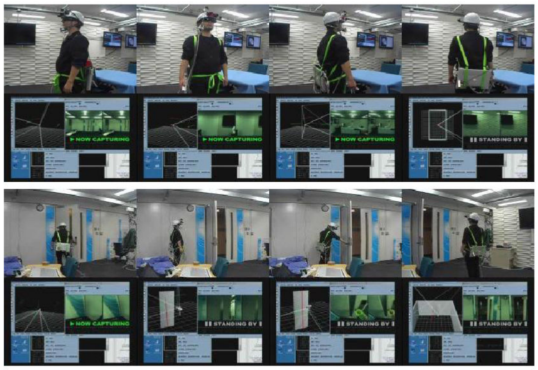


Fig. 5. Image of experiment: room and door estimation

is prevented by opened door. Thus door positioning uses edge of space in side of free side of door. Fig. 6 (c) shows a result of door positioning using our approach. In this case estimated position of door is 2420[mm] from side wall, includes under 1% from ground truth 2400[mm]. Especially result shows that estimated result matches the ground truth.

B. TABLE MODELING

We consider to a table modeling as important structure in room. Table has rectangle top plate and 700[mm] height from floor in general, and table is not measure whole size at once according to its size. Algorithm of a table modeling is shown in Fig. 7. In proposed method table is estimated online. Firstly base frame and end frame is decided by intention of human. Side edges of table must be visible in both frames. Camera locus from base to end frame is obtained by visual odometry, initial pose is decided from gravity vector. Online visual odometry runs about 5[fps]. Secondly top plate is detected form range data of laser range finder, by detected line segments using robust line fitting. When line segment is 700[mm] height, parallel by floor and visible then it is line segments composing table top. Fig. 8 shows image of experiment. Fig. 9 shows experimental result by table top estimation and registration. Estimation of table takes about 0.3[s]. In this case estimated table size is 1163 × 764[mm] includes about 10% errors from ground truth 1250 × 800[mm].

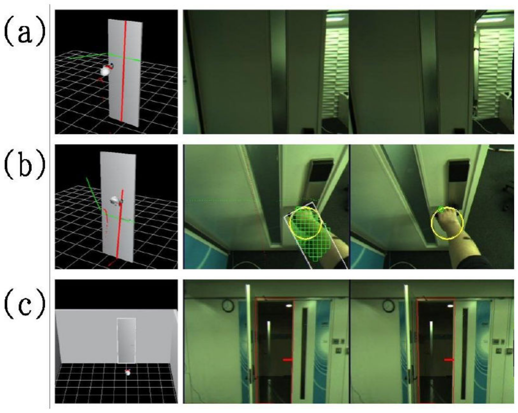


Fig. 6. Experimental result

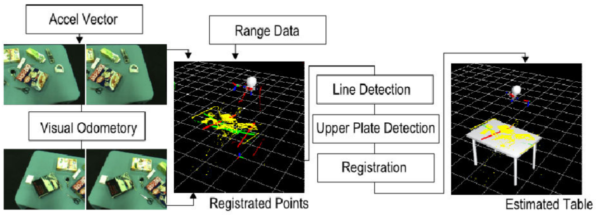


Fig. 7. Algorithm of table modeling

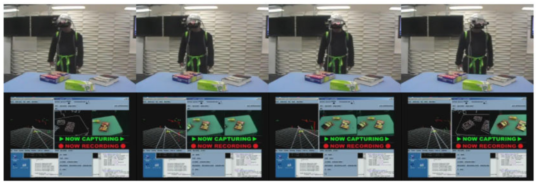


Fig. 8. Image of experiment: table estimation

C. CHAIR FITTING

Chair is a part of daily-life environment, has sparse texture similarly to other furniture, and is composed complex curves to fit human body. Complex curves are difficult to represent by solid model, but can approximate by constructive primitive solid model. We propose methods of chair detection and modeling using abstract chair model. We define abstract chair primitive model. Primitive model is composed by a sheet, a back rest and two arm rest, each parts are represented by cube. For chair detection, particle filter is composed using 3D edge points as input, and number of points inside primitive as likelihood. Abstract chair primitive is larger than real instance for apply various size of chair. The size of detected chair can obtain by 3D points inside each part, and which exist or not about arm rest using a number of points. Also color of each part can be decided by color of input images. Fig. 10 shows a image of experiment. In proposed method there is a precondition that chair measures from front, human help recognition by moving chair to proper pose. Obtained chair models are shown in Fig. 11. Fitting between 3D edge points and abstract chair primitive run about 10[fps] using

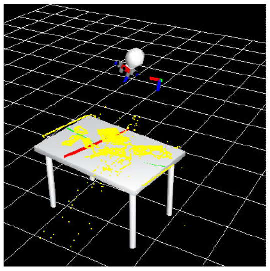


Fig. 9. Experimental result of Table fitting; table registration

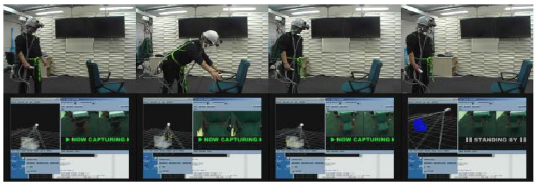


Fig. 10. Image of experiment: chair fitting

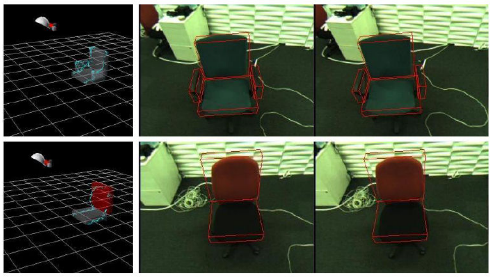


Fig. 11. Experimental result of chair fitting

about 2000 input points and 20 particles. Modeling worked well whether armrest was exist or not, and can decide that.

D. CUBE FITTING OF FURNITURE

Furniture with large scale or build-in to wall is difficult to measure except from front, but it can be managed as wall when it is fixed on wall. Especially frontal plane of furniture is important because of doors or consoles. We propose a furniture modeling method by cube fitting from measurement frontal plane. Algorithm of proposed method is shown in Fig. 12. Large furniture is considered to not estimate by one frame measurement. Measurements in plural frames are merged using visual odometry. Then plane segment finder is applied to 3D points to detect main plane of furniture, and rectangle finder is applied to contour points of plane segment. Fig. 13(b) shows input images and super impose of estimated result, Fig. 13(a) shows an estimated furniture object with texture mapping. The depth of furniture can not observe but we composed it as cube using precondition that depth equals to width of furniture. In this experiment estimated size is 512×1524[mm]. This result includes about 10% from ground truth 580 × 1730[mm], however in aspect ratio, estimated result is 2.98 nearly equals to ground truth 2.98. Errors in absolute size are considered to be caused by scale error by calibration error.

E. MODELING OF HAND-HELD OBJECTS

Cubic-shaped object model can construct by estimating each face composed object independently and can compose using estimated faces. Algorithm of our method is shown in Fig. 14. Firstly we obtain 3D edge points of each face of an object. When area of region close to camera and displacement of area is less than constancy over 2[s], then obtain 3D edge points. This check runs in about 20[fps]. Secondly we estimate symmetrical line on 2D plane from obtained 3D points. For 2D estimation, it is necessary to segment points on plane, thus plane segment finder is applied to 3D edge points. Then contour detection is applied for eliminate effects from texture. Face reconstruction using symmetrical line and texture mapping is applied after contour detection. Finally we construct an object model using obtained faces. Order of faces is decided as precondition, model is constructed when difference of width of parallel faces is less than constancy, about 10% of width. If human handles bottom side of object, bottom side of object is considered to be occluded. Thus height of object is estimated by maximum height of all faces, other faces are extended to bottom. Fig. 15 shows images of experiment, and Fig. 16(b) shows result of reconstruction. Face reconstruction took about 3[s] for each face. Object model was obtained in about 30[s].

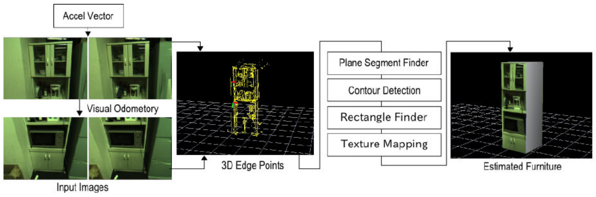
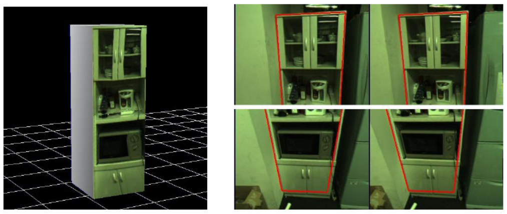


Fig. 12. Algorithm of large furniture modeling



(a) obtained model (b) Input images

Fig. 13. Experimental result of a kitchen shelf

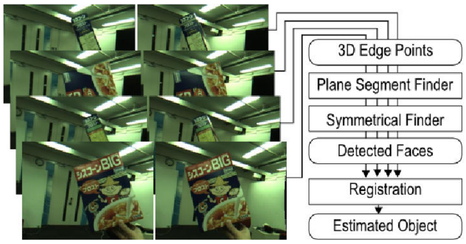


Fig. 14. Algorythm of hand-held object modeling

We also propose a method of modeling for cylindrical object. For cylindrical object, cylinder fitting is applied to 3D

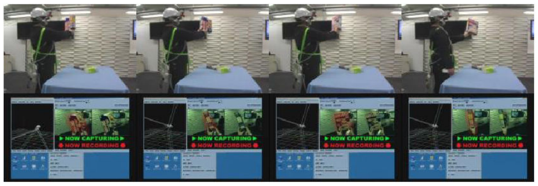
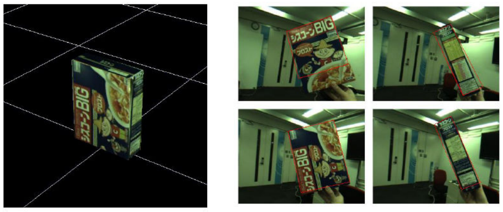


Fig. 15. Image of experiment: hand-held object estimation



(a) Obtained model (b) Input images

Fig. 16. Experimental result of serial box

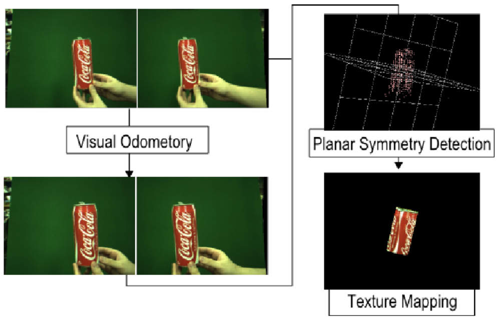


Fig. 17. Algorythm of cylinderical object modeling

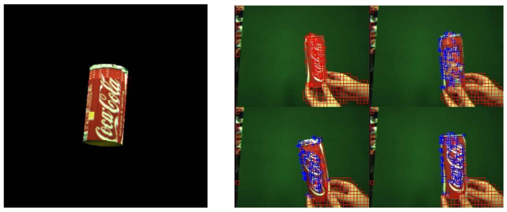

(a) Obtained model (b) Input images

Fig. 18. Experimental result of coke can

points obtained using visual odometry, by planar symmetrical detection. Algorithm of proposed method is shown in Fig. 17. Fig. 18(b) shows input images, and Fig. 18(a) shows an obtained cylindrical object. In this case estimated size of obtained model is $70 \times 75 \times 140$[mm], includes about 10[mm] for radial direction from ground truth $65 \times 65 \times 165$[mm].

F. CONSTRUCTED MODELS

Fig. 19 shows comparison between real objects and obtained models. Multi scale and shape modeling using single system was realized by switching different estimation methods for various kind of object using human intention.

VI. CONCLUSION AND FUTURE WORK

In this paper, we constructed a head-mounted multi sensor system to observe indicated objects and human behavior simultaneously, and composed a system to construct models of various kind of objects which has different scales and shapes by switching different modeling methods using human intention notified by remote controller. Auto modeling in daily-life environment is realized using proposed method. We showed some experimental results about modeling in real environment, constructed models are with about 1 to 10% errors, and provide online model creation.

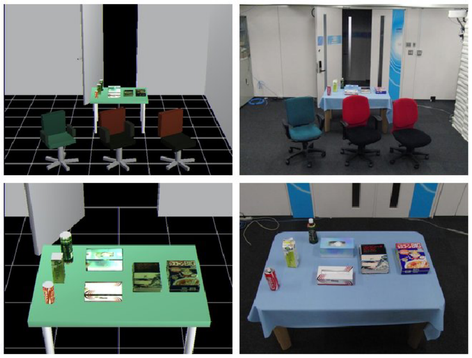


Fig. 19. Experimental result

Composed system cannot switch estimation methods atuomatically by recognize human behavior because of there are only lower level interection, so we plan to implement higher level interaction in our system. We also plan to use obtained models to manipulate objects or manage objects database system, with object matching and positioning methods. Especially objects having dense textures can be applied image search algorithm to its texture to object matching and positioning.

REFERENCES

[1] Mitsuharu Kojima, Kei Okada, and Masayuki Inaba. Manipulation and Recognition of Objects Incorporating Joints by a Humanoid Robot for Daily Assistive Tasks. In *Proceedings of The 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1564–1569, 9 2008.

[2] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, Vol. 9, No. 2, pp. 137–154, 1992.

[3] Kei Okada, Satoshi Kagami, Masayuki Inaba, and Hirochika Inoue. Plane segment finder : Algorithm implementation and applications. In *Proc. of International Conference on Robotics and Automation (ICRA'01)*, pp. 2120–2125, 2001.

[4] T. Lozano Perez, W.E.L. Grimson, and S.J. White. Finding cylinders in range data. In *CRA87*, pp. 202–207, 1987.

[5] Rodney A. Brooks. *Model-Based Computer Vision*. UMI Research Press, 1981.

[6] Andrew J. Davison, Ian Reid, Nicholas Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 6, pp. 1052–1067, 2007.

[7] Microsoft Live Labs. Photosynth. `http://photosynth.net/Default.aspx`.

[8] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *Int. J. Comput. Vision*, Vol. 80, No. 2, pp. 189–210, 2008.

[9] Pravin Bhat, C. Lawrence Zitnick, Noah Snavely, Aseem Agarwala, Maneesh Agrawala, Brian Curless, Michael Cohen, and Sing Bing Kang. Using photographs to enhance videos of a static scene. In Jan Kautz and Sumanta Pattanaik, editors, *Rendering Techniques 2007 (Proceedings Eurographics Symposium on Rendering)*, pp. 327–338. Eurographics, June 2007.

[10] Yoshio Matsumoto, Masayuki Inaba, and Hirochika Inoue. View-based approach to robot navigation. *Journal of the Robotics Society of Japan*, Vol. 26, No. 5, pp. 506–514, 2002.

[11] Masatoshi Ishikawa. Parallel processing architecture for sensory information. In *The 8th Int. Conf. on Solid-State Sensors and Actuators, and Eurosensors IX*, pp. 103–106, 1995.

[12] Tomomasa Sato, Yoshifumi Nishida, and Hiroshi Mizoguchi. Robotic room: Symbiosis with human through behavior media. *Robotics and Autonomous Systems*, Vol. 18, No. 1-2, pp. 185–194, 1996.

[13] Steve Mann. Wearable computing: A first step toward personal imaging. *Computer*, Vol. 30, No. 2, 1997.

[14] Sotaro Tsukizawa, Kazuhiko Sumi, and Takashi Matsuyama. 3d digitization of a hand-held object with a wearable vision sensor. In *ECCV Workshop on HCI*, pp. 129–141, 2004.

[15] Jiambo Shi and Carlo Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600, 1994.

[16] Hiroaki Yaguchi, Kei Okada, and Masayuki Inaba. A simultaneous estimation method of camera pose and environment model using head mounted stereo vision. *Journal of the Robotics Society of Japan*, Vol. 26, No. 6, pp. 470–475, sep 2008.

[17] M.W. Walker and L. Shao. Estimating 3-d location paramters using dual number quaternions. *CVGIP: Image Understanding*, Vol. 54, No. 3, pp. 358–367, 1991.

[18] Joshua Podolak, Philip Shilane, Aleksey Golovinskiy, Szymon Rusinkiewicz, and Thomas Funkhouser. A planar-reflective symmetry transform for 3D shapes. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, Vol. 25, No. 3, July 2006.