

Active Wearable Vision Sensor: Recognition of Human Activities and Environments

Kazuhiko Sumi[†]

[†]Graduate School of Informatics
Kyoto University
Kyoto 606-8501, Japan
sumi@vision.kuee.kyoto-u.ac.jp

Akihiro Sugimoto[‡]

[‡]National Institute of
Informatics
Tokyo 101-8430, Japan
sugimoto@nii.ac.jp

Takashi Matsuyama[†]

tm@i.kyoto-u.ac.jp

Masato Toda[†]

masatoda@vision.kuee.kyoto-u.ac.jp

Sotaro Tsukizawa[†]

tsucky@vision.kuee.kyoto-u.ac.jp

Abstract

To realize a symbiotic relationship between humans and computers, it is crucial to estimate the external and internal state of the human by observation. One promising approach is to acquire the same visual information as the human acquires. In this paper, we introduce our wearable vision sensor, which is equipped with a pair of active stereo cameras and a gaze-direction detector. From the visual information obtained by the wearable vision sensor, we present three basic three functionalities: a 3D gaze point detection and image retrieval, a 3D digitization of a hand-held object, and the measurement of a walking trajectory.

1. Introduction

Recently, human-machine symbiotic systems have been studied extensively. Their most important characteristic is an ability to interact with a human without manual invocation, conventional cooperative systems must wait for an invocation event to start interaction. Instead of waiting for an event, a symbiotic system observes the human physical state and estimates such emotional states as interest, intention, sympathy and feeling. Based on the estimation, a symbiotic system will start to interact with a human as well as synchronize itself with human action. Vision is a crucial tool to observe a human as well as his environment.

There are two approaches to human observation. One is the embedded ubiquitous vision in our environment, and the other is to wear vision. Although, recent progress in electronics has enhanced the feasibility of ubiquitous embedded vision, we believe that wearable vision is superior to the environmental vision, because it shares a similar sight with

the human wearing it. Since, environmental vision does not have manipulation capability, it is only able to observe human action from a viewpoint different than the human. On the other hand, wearable vision, which combines sensory motion with human vision, can share and experience almost exactly the same sight as humans. In addition to the realization of symbiotic systems, analyzing the sight of the wearable vision sensor will foster the study of humans. So, we have developed a wearable vision sensor[15].

One advance of the wearable vision sensor is that it corresponds to human sight. Human vision has both a very precise resolution at its center and a very wide field of view. In human vision, eyeball moves from point to point according to the interest of the human. To emulate human sight, our wearable vision sensor is equipped with a gaze-direction detector and a pair of active stereo cameras that have pan, tilt and zoom control features. We have researched the following three basic functionalities of the wearable vision sensor. One is to detect the three dimensional gaze point of humans and retrieve the image of interest. The second is to digitize the object shape and texture in 3D while it is manipulated by hand. The last is to recover the trajectory of human movement. In the following section, we present those functionalities one by one and conclude with directions for future research.

2. 3D gaze point detection and image retrieval

In this section, we present a method to retrieve the image of the object of interest, by 3D gaze point detection and camera control.

Wearable vision has been researched in various ways [7, 5, 10, 17]. The biggest advantage of wearable vision is that it shares its field of sight with the human who wears

it. However, balancing a wide field of view and resolution is a problem that must be solved. This research proposes to solve this problem by using gaze-direction and active cameras. It is also quite natural that a person's field of vision strongly reflects his/her interest or attention regardless of his/her consciousness [4].

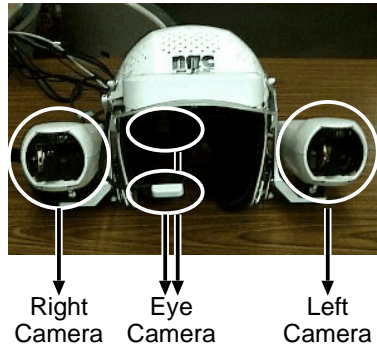


Figure 1. Appearance of active wearable vision sensor

A human's viewing line can be measured by a device, called a NAC Corporation EMR-8, that projects an infrared light beam onto an eye and detects its reflection from the retina (through the iris) and the cornea with an eye-direction sensor camera. Since the retinal and the cornea reflections reside on different spheres of the eye, the direction of the eye can be computed from those two reflections. Our wearable vision sensor, shown in Figure 1, uses this device to measure the gaze-direction. The sensor is equipped with a pair of SONY EVI-G20 active stereo cameras. Each of the camera is equipped with pan, tilt, and zoom mechanism, that are controlled by a computer. The camera is located on the side of the head. Those three devices are mounted on an bracket around a helmet to fix the geometrical relationship between the stereo camera and the gaze-direction detector. Their geometrical configuration is calibrated with the method shown in [15]. This configuration enables the sensor to share the same sight as the human.

In our proposed wearable vision sensor, the first problem is controlling the active cameras so that they share the same sight as the human who wears the sensor. The vision sensor should have the following capabilities:

- To acquire images from a wide field of view, if the human does not fix the sight.
- To track the human's gaze.
- To acquire images with high resolution, when the human fixes his sight on an object.

Since a wide field of view and a view with high resolution are mutually exclusive, we have developed a control algorithm based on gaze direction. To detect fixation of sight, temporal average of gaze direction is calculated. If the averaged gaze direction is fixed, the camera is zoomed up, but if it is changing, the camera is zoomed down. We have also developed a fast viewing-depth estimation algorithm that uses stereo cameras and a gaze-direction detector. In the following subsection, we will describe the detailed system and its algorithms.

2.1. Estimation of 3D gaze point

The 3D gaze-point is defined as the intersection of a viewing line and the viewed object. In Figure 2, cameras and the gaze-direction detector have been calibrated in advance. The viewing line is projected onto the screen of each camera. It is equivalent to the epipolar line of a viewing line. In the pair of stereo images in Figure 2, from the closest point to infinity on the viewing line, we assume the depth, and then crop the corresponding sub-image from the image on each screen. Those two images will match around the gaze point. Using this epipolar constraint, the number of trials required to calculate the best match is 10 to 100 times reduced, compared with standard stereo matching.

According to the 3D gaze-point measurement, we control the pan, tilt, and zoom parameters of the active camera as follows:

- If the gaze point is moving, the camera is zoomed out to retrieve a wide viewing field. Pan and tilt are controlled to follow the gaze direction.
- If the gaze point is fixed, the camera is zoomed in to retrieve a detailed image of the gazed object. The pan and tilt are controlled to converge toward the 3D gaze point.

We have implemented the above strategy with a dynamic memory architecture[13], whose control schematic diagram is shown in Figure 3. An example of the advantage of this system is shown in Figure 4. In Figure 4, both the left and right images are shown in the upper and lower rows, respectively. Each image is superimposed on the epipolar line with a white line, to represent the gaze direction. A yellow dot is marked on the 3D gaze point. The left image is taken without camera control, while the right image is taken with camera pan, tilt and zoom control. Based on the vantage point of the camera control, the gazed object (a computer box) can be observed in detail.

We have proposed a wearable vision sensor with a gaze-direction detector and a pair of stereo cameras. We showed that the 3D gaze point is quickly detected by finding the best image match along the epipolar line derived from gaze direction. Adding 3D gaze location to the camera control, we

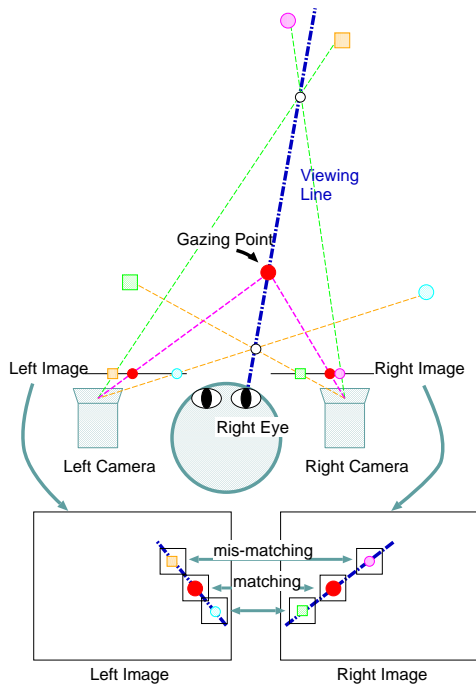


Figure 2. Optical configuration of wearable vision sensor and principal 3D gaze-point detection

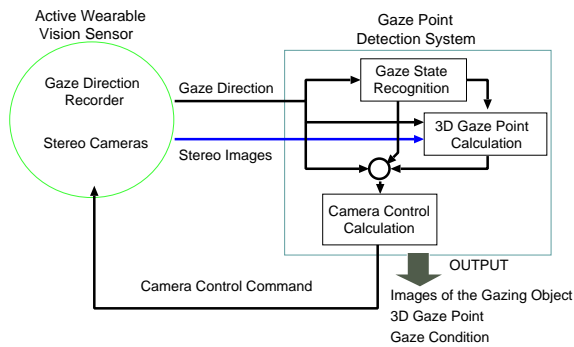


Figure 3. Control scheme of 3D gaze point detection and camera control

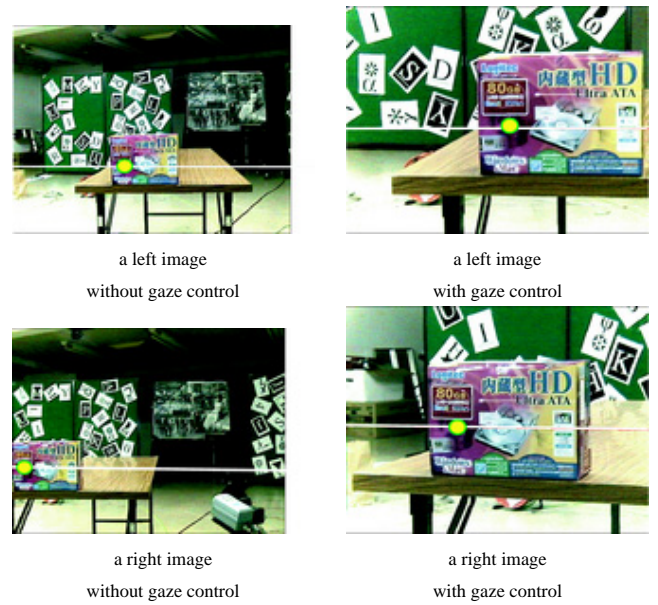


Figure 4. Results of 3D Gaze point detection and camera control

can resolve the exclusive difficulty for the viewing field and image resolution. Our future work will extend this proposal to include 3D object extraction.

3. 3D Digitization of a hand-held object

The second aspect of our research into the wearable vision sensor is the digitization of a hand-held object in 3D. A symbiotic system must recognize the action and feeling of the human living together. If the object of action is small enough, it can typically be held and manipulated by hand. Ideally, the object should be recognized while it is in his hand. The human feeling to the object in hand should also be recognized.

Most of the research into observation, based on hand-held manipulation of objects, has concentrated on the hand-object relationship[14][6]. But hand-object-view relationships have not been studied yet. Our biggest concern is the kind of view that can be acquired from the recognition of the object and the intention of the human manipulating it. Introducing the idea of “view” into hand-object relationships opens a new possibility for estimating human feelings, such as interest and intention, to the object in his hand. Therefore, our next research target will be 3D-shape extraction – both of the object and of the hand. In this paper, we separate the hand-object-view relationship into four typical cases and analyze what kind of visual information can be acquired from them.

There have been two major approaches to reconstruct the 3D shape. One is to use multiple synchronous images taken by multiple cameras [12]. The other is to use multiple asynchronous images taken along a time sequence. The synchronous approach can deal with dynamic scenes. However, the asynchronous approach assumes a static scene, in which nothing changes along the time sequence, so the images are equivalent to the synchronous approach. Once connected to the synchronous model, we can apply such well studied 3D reconstruction methods as the factorization method [16], volume intersection [1] [2], and space carving [11]. However, a hand-manipulated object is obscured by the hands. During manipulation, each hand changes its shape and location relative to the object dynamically, so it is not captured by the asynchronous single camera approach. Although, a synchronous approach can solve this problem, it is not suitable for our wearable vision. 3D shape extraction of the hand-held object by a wearable vision sensor presents a new vision problem.

In this research, 3D shape extraction of a hand-held object is regarded as a new class of shape extraction problem because of the asynchronous images which are captured when a dynamic occluding object exists.

Our approach is based on vacant space, which is defined as space that is certain to be vacant. It can be derived both from silhouettes and from stereo matching. Since the hand is a dynamic object occluding the static object, the vacant space will change from image to image and extend its space until it reaches the boundaries of the static object. Finally, we can get the 3D shape of the static object without the dynamic occlusive object. This paper proposes the following:

1. From the observation viewpoint, we analyzed human manipulation into four types: shape acquisition, visual texture acquisition, haptic texture acquisition, and total-appearance acquisition. We classified the relationships between the manipulation type and the visual information that we obtained.
2. We propose a dynamic space carving for 3D shape extraction of a static object occluded by a dynamic object moving around the static object. We showed that by using vacant space, the dynamic object will be eliminated along the carving.
3. We showed that the integration in vacant space of a stereo-depth map and a silhouette improves the efficiency of dynamic space carving.

3.1. Hand-Object-View Relationships

When a person takes an object in his hands and observes it, it is not possible to acquire all of its information simultaneously. When the object is manipulated by hand, its visible part changes depending on how it is held or the physical

relationship between the object and the hand. We classify hand-object-view relationships into four classes according to the information which the holder of the object can acquire.



Figure 5. Hand-object-view relationships

Shape Acquisition : The object obscures much of the hand, so its silhouette is visible. Figure 5(a),

Visual Texture Acquisition : The object is in front of the hand, so its texture is visible. Figure 5(b).

Haptic Texture Acquisition : The object is wrapped by the hands. Even though the object can be observed, little visual information can be acquired. Figure 5(c).

Total Appearance Acquisition : The object is turned by the hand, so that a total view, both of shape and texture, is acquired. Figure 5(d).

Since it is difficult for a computer to distinguish the captured image in the above examples, we propose to use both shape and texture for 3D shape extraction, and integrate them in our proposed vacant space.

3.2. Vacant Space

It is not always easy to distinguish between an object and a hand, especially if the object is being held in the hand. Therefore, the object and the hand are observed as one object, whose shape is changing in time. This makes it difficult to apply such conventional techniques as shape from

silhouettes [9] and space carving [11] because they depend on the correspondence of the texture on the stable object. Instead, we propose to detect vacant space, defined as space which is certain not to be occupied by any object. If vacant space is carved, the space occupied by a hand will intersect with the moving hand. The intersection will become zero if the hand moves in a sufficiently large area. On the other hand, since the object does not change, the object's space will never become vacant space. Therefore, if vacant space is carved using a long image sequence, only the object remains in space.

3.3. Processing Procedure

Based on vacant space detection, the 3D object extraction is implemented as follows:

1. **Capture** – A series of stereo images are captured by the wearable vision sensor.
2. **Camera Motion Recovery** – First, feature points are detected by a Harris Corner Detector [8]. The 3D location of the feature points are calculated by stereo analysis for each image. Next, the camera position is estimated by an advanced ICP algorithm [3].
3. **Depth Map Acquisition** – For each viewpoint, a depth map is computed by region based stereo analysis.
4. **Silhouette Acquisition** – For each viewpoint, a silhouette is computed by background subtraction.
5. **Carving Vacant Space** – The vacant space is updated by the above silhouette and the depth map.

3.4. Evaluation

First, we have evaluated our algorithm with an object whose exact dimensions were measured manually. From 22 viewpoints around the object, we extracted a 3D shape by dynamic space carving. The resultant shape extracted only by silhouette is shown in Figure 6(b), and the resultant shape by silhouette and texture is shown in Figure 6(c). These results are compared with the ideal shape shown in Figure 6(d). The number of the extra voxels and the missing voxels are shown in Figure 7(a) and (b), respectively. Figure 7 shows that even only by silhouette, dynamic space carving can extract the static 3D object shape without the hand, which is a dynamic occluding object. With silhouette and texture, the same accuracy can be achieved with fewer viewpoints. Extraction of a concave shape, which is impossible only by silhouette, is also possible by integrating silhouette and texture.

In the second experiment, we applied our algorithm to a more complex object as shown in Figure 8(a). The images were captured from 13 viewpoints around the object, and

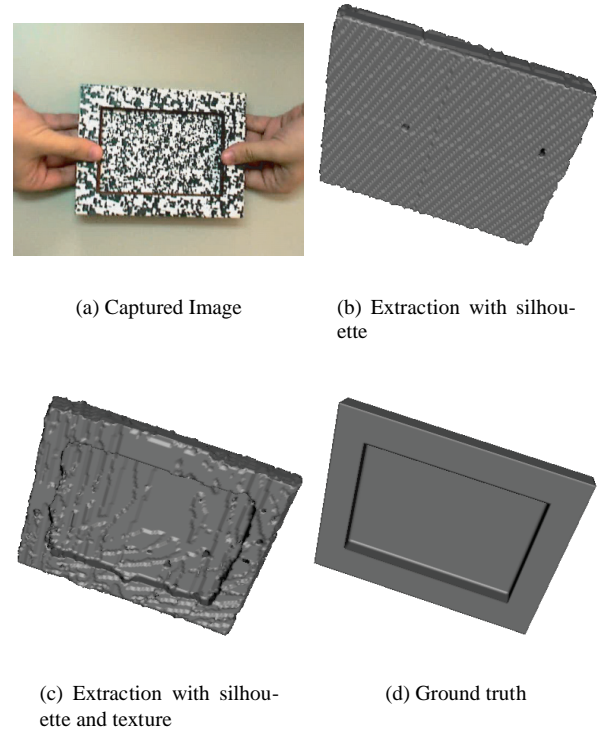


Figure 6. Photo frame data set

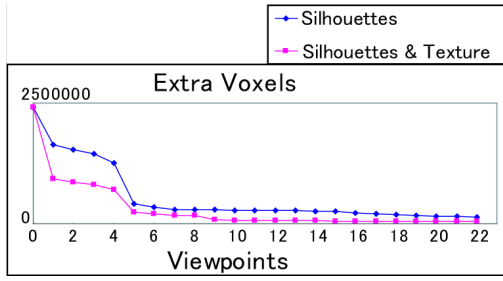
the obtained silhouettes are shown in Figure 8(b). The extracted shape does not include the hands as shown in Figure 8(c). After mapping the texture, we obtained the 3D digitized object shown in Figure 8(d).

3.5. Conclusion

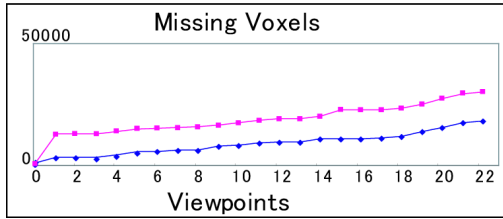
First, in this section, we showed that hand-object-view analysis, and their relationships can be classified into four types. Then we showed that shape from silhouette and shape from texture can be integrated in vacant space to extract a static object partially occluded by hand. Our experiments confirmed that dynamic space carving can extract the static object and eliminate dynamic objects around the static object. Finally, we showed that the integration of silhouette and texture enhances the extraction performance.

4. Estimation of human-motion trajectory by binocular-independent fixation control

In this section, we discuss human-motion trajectory estimation using an active wearable vision sensor. In previous sections, the sensor shares the view with the human; however, in this research, the sensor is used for acquiring objec-



(a) Extra Voxels



(b) Missing Voxels

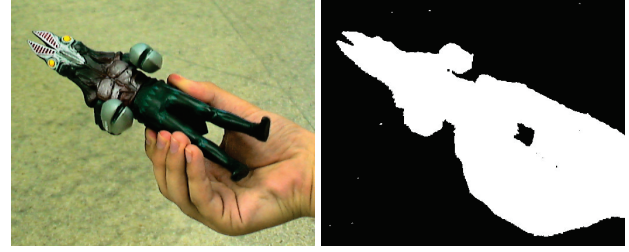
Figure 7. Error analysis of dynamic space carving

tive information: the field of view from the cameras is independent of the person's.

To estimate human-motion trajectory with two active wearable cameras, we introduce *fixation control*, i.e., camera control in which the camera automatically fixates its optical axis on a selected point (called the *fixation point*) in 3D and applies fixation control independently to each active camera. That is, while the person moves, we control the two active cameras independently so that each camera automatically fixates its optical axis to its own fixation point. We call this camera control the *binocular-independent fixation control* (Figure 9).

In binocular-independent fixation control, the two cameras need not share a common field of view because each camera fixates its optical axis on its own fixation point in 3D. We do not face the problem of feature correspondences between the images captured by the two cameras. Moreover, the estimation accuracy becomes independent of the baseline of two cameras.

To derive sufficient constraints to estimate human motion, we employ lines, which we refer to as focused lines, nearby the fixation point, because (i) we find many lines in an indoor scene, and (ii) we can easily and accurately detect lines with less computation by using the Hough transformation, and (iii) we can easily establish line correspondences



(a) Captured Image

(b) Silhouette Image



(c) 3D shape

(d) 3D shape with texture

Figure 8. Monster figure data set

over time-series frames due to their spatial extents. The constraints derived from line correspondences depend only on the rotation component of the human motion. We can thus divide the human motion estimation into two steps: the rotation estimation and the translation estimation.

The first step is the rotation estimation of the camera motion. We assume a correspondence of n focused lines over two time-series frames. Then, we have $n + 3$ unknowns (n is from the scale factors and 3 is from the rotation), whereas we have $3n$ constraints in this case. Therefore, we can estimate the rotation of the camera motion if we have more than two focused lines of correspondences.

When we finish estimating the rotation of the camera motion, translation are the only factors that need to be solved. The constraint derived from the fixation correspondence thus becomes homogeneously linear with respect to the unknowns. Hence, we can obtain the translation of the camera motion up to scale from two fixation correspondences with only linear computation.

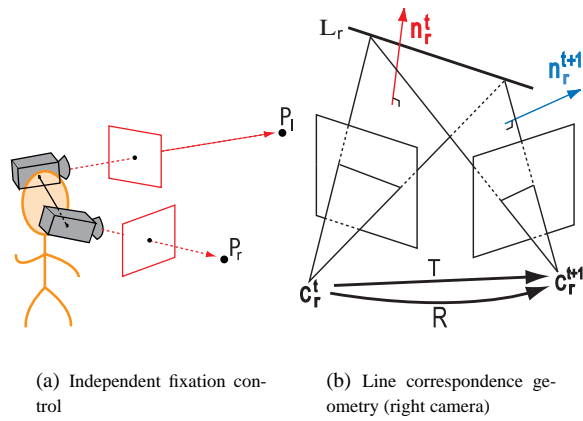


Figure 9. Binocular-independent fixation control and line correspondence.

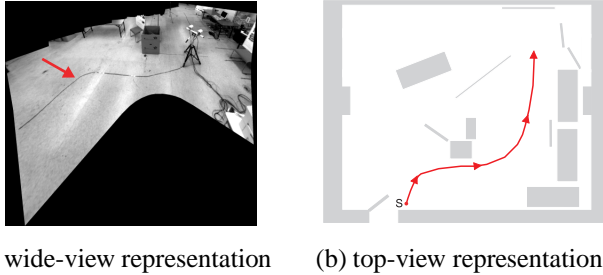


Figure 10. Camera motion trajectory.

4.1. Experiments

We moved a pair of stereo cameras indoors and created a simulation of the active wearable vision sensor. The trajectory of the right-camera motion is shown in Figure 10, and the length of the trajectory was about 6 m. We marked 35 points on the trajectory and regarded them as samples during the motion. We then applied the binocular-independent fixation control only to the samples to estimate the right-camera motion.

In the images captured by each camera at the starting point of the camera motion, we manually selected a point to serve as the fixation point. During the estimation, we manually updated fixation points 8 times. We used two focused lines for each camera, four focused lines in total. Edge detection followed by the Hough transformation is used for focused line detection. Figure 11 shows an example of image pairs captured at a marked point. In the image, the fixation point (the black circle) and two focused lines (the thick black lines) are overlaid.

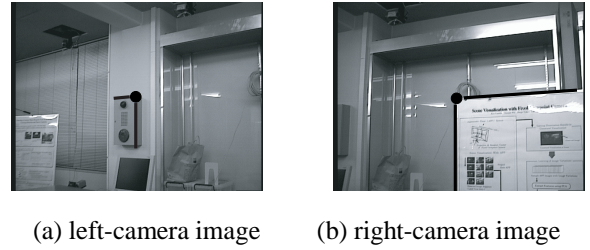


Figure 11. Example of images acquired by the two cameras during camera motion.

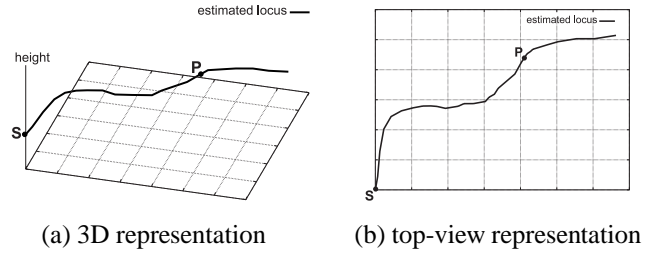


Figure 12. Estimated trajectory of the camera motion.

Under the above conditions, we estimated the right-camera motion at each marked point. Figure 12 shows the trajectory of the right-camera motion, obtained by concatenating the estimated motions at the marked points. In the figure, S is the starting point of the motion.

The height, which is almost constant, from the floor was almost accurately estimated from the trajectory. As for the component parallel to the floor, however, the former part (from S to P in the figure) of the estimated trajectory is fairly close to the actual trajectory; but the latter part (after P) deviates from the actual trajectory. We have a theory to explain this deviation. Perhaps the motion at P was incorrectly estimated, by mistakes in the fixation correspondence or in the line detection. Since the motion was incrementally estimated, an incorrect estimation at the marked point resulted in an aberration in the subsequent estimations.

4.2. Conclusion

We proposed a method for incrementally estimating the motion trajectory of a person wearing our sensor and by using binocular-independent fixation control, in which each of the cameras tracks a different point in a different field of view in the environment. We will address the problem of error accumulation, in future work.

5. Summary

An active wearable vision sensor and its three important functionalities are proposed in this paper. First, 3D gaze-point detection and image retrieval are proposed. They take advantage of the epipolar constraints given by the viewing direction and the stereo cameras and the achieved real time camera control which balances the viewing field and image resolution. Second, 3D digitization of a hand-held object is proposed. In this research, the hand-object relationship is classified by appearance variation. Then a new 3D shape extraction algorithm, which we refer to as dynamic-space carving, is proposed. The 3D shape of a partially occluded object is carved out of vacant space, which integrates both silhouette and texture. Third, a new technique to keep track of multiple features in a stable environment is proposed to recover human-motion trajectory. The camera control on a moving platform, which we refer to as binocular-independent fixation, is equivalent to the tracking of a moving object from wide-baseline stereo cameras embedded in the environment. This results in stable tracking and precise trajectory estimation. Our future work will extend gaze and manipulation analysis and focus on human observation and recognition to help a symbiosis between machines and humans in the near future.

Acknowledgements

Section 2 and Section 4 include work by Akihiro Nakayama and Wataru Nagatomo for their master's thesis, respectively.

This series of research is supported in part by the Informatics Research Center for Development of Knowledge Society Infrastructure, 21st. Century COE program and by contracts 13224051 and 14380161 of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

- [1] H. Baker. Three-dimensional modeling. In *Fifth int'l Joint Conf. on Artificial Intelligence*, pages 649–655, 1977.
- [2] B. G. Baumgart. Geometric modeling for computer vision. *Stanford University Technical Report*, AIM-249, 1974.
- [3] P. J. Besl and N. D. McKey. A method for registration of 3-d shapes. *IEEE Trans. PAMI*, 14(2):239–256, 1992.
- [4] R. Carpenter. *Movements of the Eyes*. Pion, London, 2nd edition, 1988.
- [5] B. Clarkson, K. Mase, and A. Pentland. Recognizing user's context from wearable sensors: Baseline system. *Vision and Modeling Technical Report, Media Lab. MIT*, (TR-519), 2000.
- [6] M. R. Cutkosky. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Trans. Robot. Automat.*, 5(3):269–279, 1989.
- [7] A. P. H. Aoki, B. Schiele. Realtime personal positioning system for wearable computers. *Vision and Modeling Technical Report, Media Lab., MIT*, (TR-520), 2000.
- [8] C. J. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conf.*, pages 147–151, 1988.
- [9] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle. Surface reconstruction from unorganized points. In *SIGGRAPH '92*, volume 26, pages 71–78, July 1992.
- [10] M. Kourogi, T. Kurata, and K. Sakaue. A panorama-based method of personal positioning and orientation and its real-time applications for wearable computers. In *Proc. of Int. Symposium on Wearable Computers*, pages 107–114, 2001.
- [11] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. In *IEEE Int'l Conf. on Computer Vision*, pages 307–314, 1999.
- [12] W. N. Martin and J. K. Aggarwal. Volumetric description of objects from multiple views. *IEEE Trans. PAMI*, 5(2):150–158, 1987.
- [13] T. Matsuyama, S. Hiura, T. Wada, K. Murase, and A. Yoshioika. Dynamic memory: Architecture for real time integration of visual perception, camera action, and network communication. In *Proc. Computer Vision and Pattern Recognition Conference*, pages 728–735, June 2000.
- [14] I. Napier. The prehensile movements of the human hand. *J. Bone and Joint Surgery*, 38B(4):902–913, 1956.
- [15] A. Sugimoto, A. Nakayama, and T. Matsuyama. Detecting a gazing region by visual direction and stereo cameras. In *Proc. of the 16th int'l Conf. Pattern Recognition*, volume III, pages 278–282, 2002.
- [16] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int'l Journal of Computer Vision*, 9(2):137–154, 1992.
- [17] B. T. W. W. Mayol and D. W. Murray. Wearable visual robots. In *Proc. of Int. Symposium on Wearable Computers*, pages 95–102, 2000.