装着型視覚センサを用いた手持ち物体の3次元ディジタル化

築澤 宗太郎　鷲見 和彦　松山 隆司
{tsucky, sumi}@vision.kuee.kyoto-u.ac.jp　tm@i.kyoto-u.ac.jp
京都大学大学院 情報学研究科 知能情報学専攻

概要　人にとって, 興味ある小さな対象を手に持って, 取りまわして観察するという行為は, 日常的によくする行為の一つである. 人と共生する計算機システムは, 人が取りまわして観察している物体と, その物体に対する人の意図や意志を自律的に認識すべきである. 人が手に持った物体を自律的に認識し, また, 人の把持行為や把持物体に対する観察行為を理解するために, 我々は装着型視覚センサーを用いた研究を行ってきた. 装着型視覚センサーは, 装着した人とほぼ同じ視点と視界で物体を見ることができるという特徴がある. 本論文では装着型視覚センサーを用いた画像から得られた次のような研究結果を示す. まず第一に, 観察のための手繰り操作という視点で, ものの持ち方を４通りに分類できることを示した. 第二に, ダイナミックスペースカービングと言う, 変化する物体に隠蔽される剛体の三次元形状を抽出する手法を提案した. 最後に, 空白空間という概念で距離画像とシルエットとが統合できることを示し, この統合が三次元形状の抽出を効率化することを示した.

3D Digitization of Hand-held Object with a Wearable Vision Sensor

S. Tsukizawa　K. Sumi　T. Matsuyama
{tsucky, sumi}@vision.kuee.kyoto-u.ac.jp　tm@i.kyoto-u.ac.jp
Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University

Abstract. It is a common human behavior to hold a small object of interest and to manipulate it for observation. A computer system, symbiotic with human, should recognize the object and the intention of the human, while he is manipulating it. To realize autonomous recognition of a hand-held object as well as to study human behavior of manipulation and observation, we have developed a wearable vision sensor, which has similar sight to a human who wears the sensor. In this paper, we show the following results obtained with the wearable vision sensor. First, we analyzed the human manipulation for observation into four types and related them with acquirable visual information. Second, we proposed *dynamic space carving* for 3D shape extraction of a static object occluded by a dynamic object moving around the static object. Finally, we showed that stereo depth map and silhouette can be integrated in *vacant space* and the integration improves the efficiency of *dynamic space carving*.

1 Introduction

Recently, human-computer symbiotic systems have been studied extensively. Symbiotic relationships between a human and a computer will require recognition both of the target object and of the human manipulating it. Ideally, the object should be recognized while it is in his hand. Also, human feeling against the object in hand should be recognized too. Although, the human and the object can be observed by environmental sensors, we consider wearable system is better, because the computer system can share the similar sight with the human. In other words, the computer system can share the same experience with human. Using a wearable vision system, it is easy to obtain the similar vi-

sual information with the human including the object and his body. Fig.1 shows the wearable vision sensor we have developed. It equips with a gazing direction detector, which we refer to as eye-mark recorder, and a pair of stereo cameras, each of which can control pan, tilt, and zoom. Manipulating an object in hand for observation is a kind of hand hold action.So far, various researches have been studied on holding [1][2]. Most of them are concerned with hand-object relationships. Hand-object-view relationships are not studied yet. For recognizing the object and understanding his will, what kind of view can be acquired is the biggest concern. Introducing the idea of "view" into hand-object relationships will open a new possibility to estimate the human feelings, such as interest and intention, against the object in his hand. However, first order research target will be the 3D shape extraction both of the object and of the hand. In this paper, we analyze the hand-object-view relationships into four typical cases and shows what kind of visual information can be acquired from these cases.

Figure 1: Wearable Vision Sensor

For the 3D shape reconstruction, there have been two major approaches. One is to use multiple synchronous images taken by multiple cameras [3]. The other is to use multiple asynchronous images taken along time sequence. Synchronous approach can deal with dynamic scene, however, asynchronous approach assume static scene, in which nothing changes along the time sequence and the images are equivalent to the synchronous approach. Once concluded to synchronous model, we can apply well studied 3D reconstruction method like factorization method [4], volume intersection [5] [6], and space carving [7]. However, a hand-manipulated object is obscured by hands. Since each hand in manipulation changes its shape and location relative to the object dynamically, it is not treated by the asynchronous single camera approach. Although, synchronous approach can manage this problem, it is not suitable for wearable vision. The 3D shape extraction of hand-held object by a wearable vision is a new computer vision problem.

In this paper, the 3D shape extraction of a hand-held object is regarded as new class of shape extraction problem from asynchronous images which are captured in the situation where a dynamic occluding object exists.

The approach we propose is based on *vacant space*. *vacant space* is defined as the space, which is confident of vacancy. It can be derived both from silhouette and from stereo matching. Since the hand is a dynamic object occluding the static object inside, the *vacant space* will changes from image to image, and extend its space until the boundary of the static object. Finally, we can get the 3D shape of the static object without the dynamic occluding object. The contribution of this paper is as follows:

1. From the observation viewpoint, we analyzed the human manipulation for observation into four types, shape acquisition, visual texture acquisition, haptic texture acquisition, and total appearance acquisition. We classified relationships between the manipulation type and the visual information we can obtain.

2. We proposed *dynamic space carving* for 3D shape extraction of a static object occluded by a dynamic object moving around the static object. We showed that using *vacant space*, the dynamic object will be eliminated along the carving.

3. We showed that stereo depth map and silhouette can be integrated in *vacant space* and the integration improves the efficiency of *dynamic space carving*.

The composition of this paper is as follows. Section 2 describes our approach and the classification of relationships between the manipulation type and the visual information. Section

3 describes the technique of 3D shape extraction of a hand-held object by using silhouette information and texture information from asynchronous multiple viewpoint images. In Section 4, we evaluate our approach with real images. Finally, in Section 5, we summarize our approach.

2 Our Approach

2.1 vacant space

In this research, time series images captured with the wearable vision sensor are asynchronous multiple viewpoint images. In an object centered coordinate system, this problem can be defined a 3D shape extraction problem from asynchronous images in the situation where a dynamic occluding object exists.

It is not always easy to segment an object and a hand, if the object is in the hand. Therefore, the object and the hand are observed as a combined object. The shape of the combined object is changing along time. This makes it difficult to apply conventional techniques such as shape from silhouettes [8] and space carving [7], because these techniques depend on correspondence on the stable object texture. Instead, we propose to detect *vacant space*. *vacant space* is defined as space which is certain not to be occupied by any object. Since the hand moves, if *vacant space* is carved, the space which a hand occupies will be intersection of the moving hand. The intersection will become zero if the hand moves in sufficient large area. On the other hand, since the object does not change, the object space will never become *vacant space*. According to these considerations, if *vacant space* is carved using the images captured for a long time, only the object remains in space.

There are silhouette information and texture information as information acquired from an image captured with the wearable vision sensor. In each viewpoint, because of silhouette constraint [9], a hand and an object surely exist in a visual cone obtained from a silhouette. Therefore, the outside of the visual cone is *vacant space*. In addition, if we can obtain a depth map by stereo analysis, the space from the viewpoint to the foreground of the depth map is *vacant space*, too.

2.2 The Classification of Hand-Object-View Relationships

In this subsection, we discuss the silhouette information and texture information in the image of a hand-held object. When a person takes an object in his hands and observes it, he cannot acquire all the information of the object simultaneously. Thus, he manipulates it to acquire information which he needs. When he manipulates it, its visible part changes depending on a type of holding, or the physical relationship of the object and his hand seen from his viewpoint. Then, we classify hand-object-view relationships into four classes according to the information which the holder of the object can acquire.

Shape Acquisition :
When a person observes an object shape, his viewpoint, the object, and his hand are located in a line. In the captured image, because of occlusion, small part of the object texture is obtained, but most part of the object silhouette is obtained as shown in Fig.2 (a).

Visual Texture Acquisition :
When a person observes an object visual texture, seen from his viewpoint, his hand is in the object backside. In the captured image, most part of the object texture is obtained, but small part of the object silhouette is obtained as shown in Fig.2 (b).

Haptic Texture Acquisition :
When a person observes an object haptic texture, his hand occludes most part of the object. In the captured image, small part of the object texture and the silhouette is obtained as shown in Fig.2 (c).

Total Appearance Acquisition :
When a person pinches an object to observe total balance of the object, seen from his viewpoint, there is no part to which his hand almost contacts the object. In the captured image, most part of the object texture and the silhouette is obtained as shown in Fig.2 (d).

So, information obtained from a captured image changes with the class, but it is difficult for a computer to identify the class of the image.

Therefore, it is appropriate to use both silhouette information and texture information. On the other hand, although it is impossible to obtain a silhouette of an object concave part, the texture of the part can be obtained. And although it is impossible to obtain an appropriate texture in part of a periodic texture, and a monotonous texture, the silhouette of the part can be obtained. Thus, a silhouette and a texture have a complementary relation. Therefore, using both of them is appropriate, and we use both of them to detect *vacant space*.

(a) Shape Acquisition

(b) Visual Texture Acquisition

(c) Haptic Texture Acquisition

(d) Total Appearance Acquisition

Figure 2: The Classification of the Relationships

3 Integration of Volume Intersection and Stereo Analysis

3.1 Vacant Space Detection

For the 3D reconstruction using asynchronous images, conventional approaches, such as volume intersection and stereo analysis, try to calculate the space of the object or the surface of the object. However in this problem under occlusion by dynamic object, both of these approaches produce multiple inconsistent results caused by the dynamic object. Moreover, those two approaches provide different type of representation, space and surface. Instead, we focus on the space, which occupied neither by the static object nor by the dynamic occluding object. We refer to this space as *vacant space*. It can be calculated both from silhouette and from stereo analysis. *vacant space* grows as we add new viewpoints and as the dynamic object changes its shape, but does not extend beyond the surface of the static object. Finally, we get the shape of the static object.

3.2 Definition of Parameters

We assume that images are captured from time t_1 to time t_n. Before applying our approach, the camera motion is recovered form feature tracking of the object. See 4.1 for camera motion recovery. In this paper, we use the following notations as shown in Fig.3.

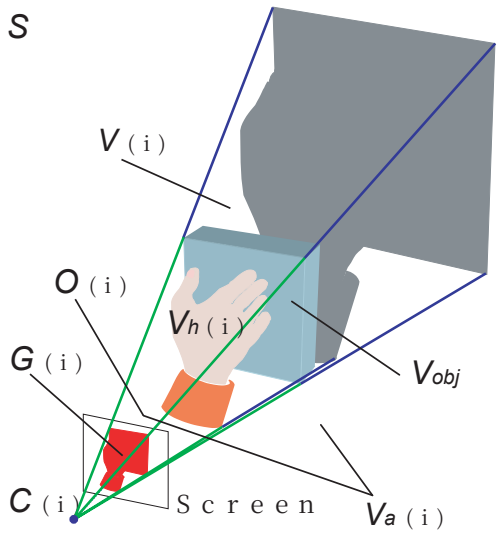


Figure 3: Space at Time t_i

S : Space of an object center coordinate system

V_{obj} : Space where the object occupies

$V_h(i)$: Space where a hand occupies at time t_i

$C(i)$: A viewpoint at time t_i

$G(i)$: A silhouette of V_{obj} and $V_h(i)$ obtained in viewpoint C_i

$V_G(i)$: Space where is back-projected $G(i)$ from viewpoint $C(i)$ (visual cone obtained by $G(i)$)

$O(i)$: Space from viewpoint $C(i)$ to foreground of a depth map obtained in $C(i)$

$V_a(i)$: *vacant space* at time t_i

At time t_i, *vacant space* $V_a(i)$ is defined Eq.(1)

$$V_a(i) = O(i) \cup \overline{V_G(i)} \tag{1}$$

And We define $P(p, V_x)$ as the probability that a point p is included in Space V_x. In this research, the following conditions shall be fulfilled about S, V_{obj}, and $G(i)$.

- S and V_{obj} are contained in the sight of the camera in all the viewpoints $C(1), C(2), \cdots$, and $C(n)$.
- Silhouette $G(i)$ is obtained exactly.

3.3 Dynamic Space Carving

We extract the hand-held object space by carving *vacant space*. We call this technique "*Dynamic Space Carving*".

The space where the object does not exist certainly is described as a set of point $p(p \in S)$ which the proposition Eq.(2) is truth.

$$\bigcup_{i=1}^{n} \{p \in V_a(i)\} \tag{2}$$

Therefore, to prove that a hand is not included in an extracted shape, we should just show that a probability that p $(p \in \overline{V_{obj}})$ is not included in $V_a(i)(i = 1, 2, \cdots n)$ in all viewpoints bocomes zero.

We suppose that a person manipulates an object for a long time $(n = \infty)$, and images are captured in sufficient many viewpoints. If it is supposed that a hand moves randomly the space except the object, in a viewpoint C_i, it can be known in probability theory whether a point p which is not included in V_{obj} fulfills Eq.(3) or not.

$$p \in \overline{V_a(i)} \tag{3}$$

Since Eq.(4) is concluded, Eq.(5) is concluded.

$$\overline{V_a(i)} \in S \tag{4}$$

$$P(p, \overline{V_a(i)}) \leq P(p, S) \tag{5}$$

Therefore, since Eq.(6) and Eq.(7) are concluded, Eq.(8) is concluded.

$$P(p, S) = 1 \tag{6}$$

$$S \cap \overline{V_a(i)} \neq \phi \tag{7}$$

$$P(p, \overline{V_a(i)}) < 1 \tag{8}$$

Probability E that p $(p \in \overline{V_{obj}})$ is not included in $V_a(i)(i = 1, 2, \cdots n)$ in all viewpoints is defined Eq.(9).

$$E = P(p, \overline{V_a(1)})\ P(p, \overline{V_a(2)}) \cdots P(p, \overline{V_a(n)}) \tag{9}$$

Eq.(10) is derived from Eq.(8) and Eq.(9).

$$\lim_{n \to \infty} E = 0 \tag{10}$$

Therefore, if a hand moves randomly and capturing for a long time, it becomes zero the probability that the point p which is not included in the object $\left(p \in \left(S \cap \overline{V_{obj}}\right)\right)$ is not included in $V_a(i)$. Thus, the object shape which does not include the hand can be extracted by *dynamic space carving*.

4 Experimental Results

In order to evaluate the proposed technique, we experimented.

4.1 The Flow of Processing

We extracted 3D shape of a hand-held object by the following procedures.

1. **Capture**
   We captured images in the situation of manipulating an object with the wearable vision sensor.

2. **Camera Motion Recovery**
   To recover the camera motion, we estimated position of a distance data. First, we extracted feature points with Harris Corner Detector [10], and we obtain 3D feature points by stereo analysis in each image. Next, we estimated the camera position by using the advanced Iterative Closest Point algorithm [11], and we recovered the camera motion.

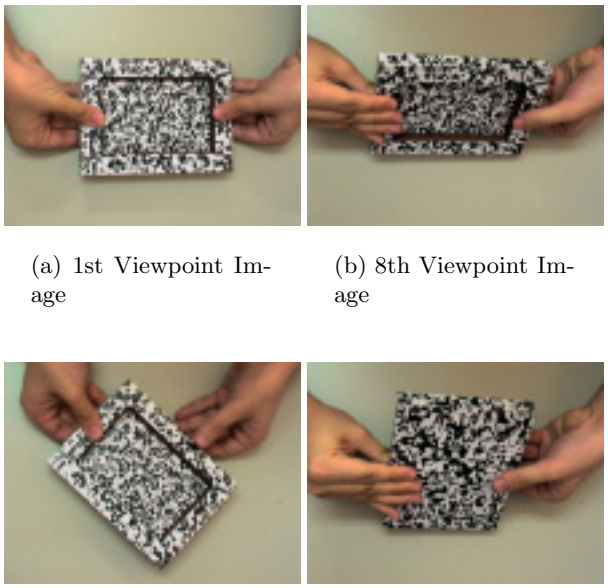
3. **Depth Map Acquisition**
   In each viewpoint, we obtained a depth map and detected *vacant space* by stereo analysis.

4. **Silhouette Acquisition**
   By background subtraction, we obtained a silhouette containing the object and the hand in each viewpoint. And we obtained *vacant space* from the silhouette.

5. **Carving Vacant Space**
   We carved the *vacant space* by the technique shown in last section.

4.2 Evaluation with Simple Shape Object

In order to evaluate errors of extracted models, we reconstructed a simple shape object.



(a) 1st Viewpoint Image

(b) 8th Viewpoint Image

(c) 10th Viewpoint Image

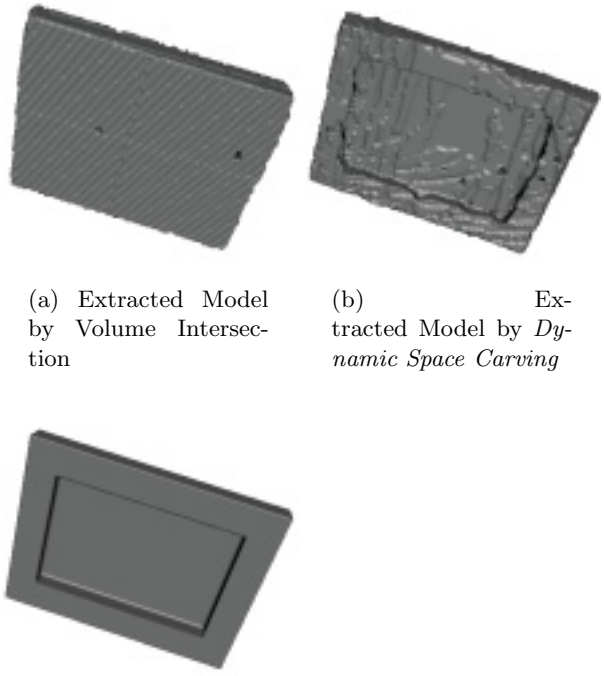
(d) 18th Viewpoint Image

Figure 4: Captured Images of Photo Frame

In the situation that a person manipulated a photo frame which stuck random pattern, we captured it with the wearable vision sensor as shown in Fig.4 (a), (b), (c), (d). We obtained asynchronous images with the wearable vision sensor in 22 viewpoints, and we extracted the 3D object shape from these asynchronous images.

The result extracted only by the volume intersection is Fig.5 (a), and the result extracted by the proposed technique is Fig.5 (b). We compared those result model with ideal model as shown in Fig.5 (c). The results are plotted in Fig.6 with graphs (a), (b) and (c) showing respectively the number of extra, missing , and total error (extra plus missing) voxels between the extracted models and ideal model against the number of viewpoints used.

In the experimental result, when we extract the object shape by volume intersection with many viewpoint images, we can reconstruct the 3D object shape which is not included the hand. And by *dynamic space carving*, we can reconstruct it with almost same accuracy with fewer viewpoints, namely, for a short time. Moreover, we can reconstruct the concave part which is impossible to reconstruct by volume intersection. In the result of *dynamic space carving*, there are more missing voxels than the result of volume intersection. But compared with extra voxels, missing voxels are quite few. Therefore, it can be said that the proposed technique is effective for the 3D reconstruction problem of a hand-held object.



(a) Extracted Model by Volume Intersection

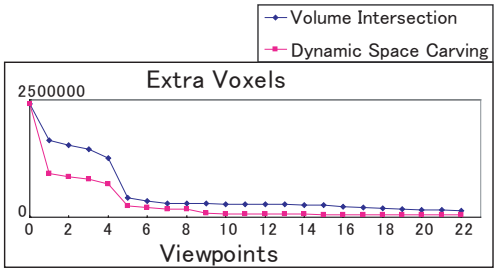
(b) Extracted Model by *Dynamic Space Carving*

(c) Ideal Model

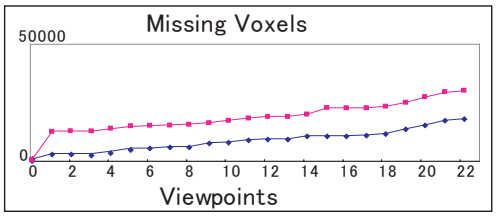
Figure 5: Photo Frame Data Set

4.3 Evaluation with Complex Real Object

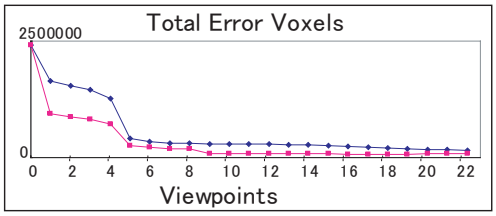
In the situation that a person manipulated a monster figure, we captured it with the wearable vision sensor as shown in Fig.7 (a), (b), (c), (d). First, we obtained 13 asynchronous multiple viewpoint images, and obtained their silhouettes as shown in Fig.8 (a), (b), (c), (d). Next, we obtained *vacant space* by silhouette constraint and stereo analysis, and we extract



(a) Extra Voxels



(b) Missing Voxels



(c) Total Error Voxels

Figure 6: Graph of Errors

(a) 1st Viewpoint Image

(b) 4th Viewpoint Image

(c) 7th Viewpoint Image

(d) 10th Viewpoint Image

Figure 7: Captured Images of Monster Figure

(a) 1st Viewpoint Silhouette

(b) 4th Viewpoint Silhouette

(c) 7th Viewpoint Silhouette

(d) 10th Viewpoint Silhouette

Figure 8: Silhouettes of Monster Figure

the figure shape by carving the *vacant space*. The extracted figure shape was not included his hand as shown in Fig.9 (a). Last, its texture was mapped on the extracted shape as shown in Fig.9 (b).

(a) Extracted Shape (b) Texture Mapped Shape

Figure 9: Monster Figure Data Set

5 Conclusion and Feature Work

We showed the following results obtained with the wearable vision sensor. First, we analyzed the human manipulation for observation into four types and related them with acquirable visual information. Second, we proposed *dynamic space carving* for 3D shape extraction of a static object occluded by a dynamic object moving around the static object. Finally, we showed that stereo depth map and silhouettes can be integrated in *vacant space*, and showed that our approach is effective by the experiment.

Now, we are studying detection of 3D gazing object position, and camera control for capturing the gazing object appropriately. Therefore, we will integrate them with our approach of this paper in the future.

Acknowledgments

The paper is supported in part by Ministry of Education, Culture, Sports, Science and Technology grants no. 13224051.

References

[1] I, Napier.: The prehensile movements if the human hand. J.Bone and Joint Surgery, **38B**(4), (1956) 902–913.

[2] M, R, Cutkosky.: On Grasp Choice, Grasp Models, and the Design of Hands for Manufacturing Tasks. IEEE Trans. Robot. Automat., **5**(3), (1989) 269–279.

[3] W, N, Martin., J, K, Aggarwal.: Volumetric description of objects from multiple views. IEEE Transactions on Pattern Analysis and Machine Intelligence, **5**(2) (1987) 150–158.

[4] C, Tomasi., T, Kanade.: Shape and motion from image streams under orthography: a factorization method. Int'l Journal of Computer Vision, Vol. **9**(2) (1992) 137–154.

[5] H, Baker.: Three-dimensional modelling. Fifth International Joint Conference on Artificial Intelligence, (1977) 649–655.

[6] B, G, Baumgart.: Geometric modeling for computer vision. Technical Report AIM-**249**, Artificial Intelligence Laboratory, Stanford University, (1974).

[7] K, N, Kutulakos., S, M, Seitz.: A theory of shape by space carving. IEEE International Conference on Computer Vision, (1999) 307–314.

[8] H, Hoppe,, T, DeRose., T, Duchamp., J, McDonald., W, Stuetzle.: Surface reconstruction from unorganized points. Computer Graphics (SIGGRAPH '92 Proceedings), volume **26**, (July 1992) 71–78.

[9] A, Laurentini.: How far 3d shapes can be understood from 2d silhouettes IEEE Transactions on Pattern Analysis and Machine Intelligence, **17**(2), (1995) 188–195.

[10] C, J, Harris,, M, Stephens.: A combined corner and edge detector. In Proc. 4th Alvey Vision Conf., Manchester, (1988) 147–151.

[11] P, J, Besl., N, D, McKey.: A method for registration of 3-D shapes. IEEE Trans. Patt. Anal. Machine Intell., vol. **14**(2), (1992) 239–256.