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

Sotaro TSUKIZAWA, Kazuhiko SUMI, and Takashi MATSUYAMA
tsucky@vision.kuee.kyoto-u.ac.jp
sumi@vision.kuee.kyoto-u.ac.jp
tm@i.kyoto-u.ac.jp

Graduate School of Informatics, Kyoto University
Sakyo Kyoto 606-8501, JAPAN

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 a human, should recognize the object and the intention of the human while the object is manipulated. 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 human manipulation for observation and related it 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 texture and silhouette information can be integrated in vacant space and the integration improves the efficiency of *dynamic space carving*.

1 Introduction

Human-computer symbiotic systems have been a popular subject of study in resent research. Symbiotic relationships between a person and a computer will require recognition both of the target object and the person. Ideally, the object should be recognized while it is in the person's hand. The feeling of the held object should be recognized too. Although, the person and the object can be observed by environmental sensors, we consider a wearable system because the computer system can share similar sight with the person. In other words, such a computer system can share the same experience with a human. Using a wearable vision system, it is easy to obtain similar visual information with the person including the object and the person's body. Fig.1 shows the wearable vision sensor we have developed. It is equipped with a gazing direction detector, which we refer to as the eye-mark recorder, and a stereo pair of cameras, each of which can control pan, tilt, and zoom. Manipulating an object in hand for observation is a kind of hand-held action. So far, various research has been conducted on holding [1][2]. Most of the research is concerned with hand-object relationships. Hand-object-view relationships are not studied yet. For recognizing the object

and understanding the person's 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 person's feelings, such as interest and intent concerning the object in hand. However, the main research target will be the 3D shape extraction both of the object and of the hand. In this paper, we analyze and classified the hand-object-view relationships into four typical cases and show what kind of visual information can be acquired from these cases.

Fig. 1. Wearable Vision Sensor

For 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 a time sequence. The synchronous approach can deal with a dynamic scene, however, the asynchronous approach assumes a static scene in which nothing changes along the time sequence and the images are equivalent to the synchronous approach. Because of this similarity, we can apply well studied 3D reconstruction methods like the factorization methods [4], volume intersection [5], and space carving [6]. However, a hand-manipulated object can also be obscured by the person's hands. Since each hand during manipulation changes its shape and location relative to the object dynamically, it is not treated by the asynchronous single camera approach. Although, the synchronous approach can manage this problem, it is not suitable for wearable vision. The 3D shape extraction of a hand-held object by wearable vision is a new computer vision problem.

In this paper, 3D shape extraction of a hand-held object is regarded as a new class of shape extraction problems 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 that is certain not to be occupied by any object. It can be derived both from silhouettes and from texture. Since the hand is a dynamic object occluding the static object inside, the *vacant space* will change from image to image, and extend its space until it reaches 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 analyze human manipulation for observation, and we show that a silhouette and a texture have a complementary relation.

2. We propose *dynamic space carving* for 3D shape extraction of a static object occluded by a dynamic object moving around the static object. We show that by using *vacant space*, the dynamic object will be eliminated along the carving.
3. We show that texture information and silhouette information 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

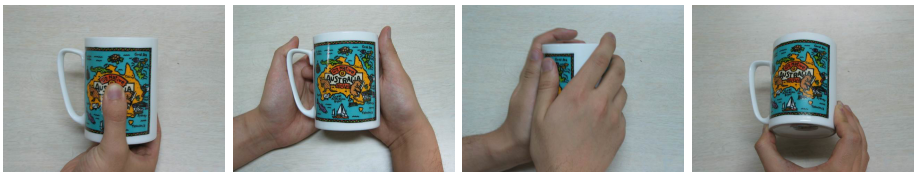
In this research, time series images captured with the wearable vision sensor are asynchronous multiple viewpoint images. In an object-centered coordinate system, this problem can be defined as a 3D shape extraction problem from asynchronous images in the situation where a dynamic occluding object exists.

It is not always easy to segment an object and a hand when the object is in the hand. Rusinkiewicz reconstructed a hand-held object in real time [11]. He made detection of a hand impossible by gloving a holder's hand. Since to glove and manipulate an object is unnatural for a human, the approach is effective in a laboratory, but is not suitable to use in everyday life. Therefore, we don't use a glove, and 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 [7] and space carving [6], because these techniques depend on correspondence on the stable object texture. Instead, we propose to detect *vacant space*. Since the hand moves but the object doesn't move in the object-centered coordinate system, if *vacant space* is carved, the space which a hand occupies will intersect that of the moving hand. The intersection will become zero if the hand moves in a sufficiently large area. On the other hand, the object space will never become *vacant space*. So, if *vacant space* is carved for a long time, only the object remains in space.

There is silhouette information and texture information available as information acquired from an image captured with the wearable vision sensor. In each viewpoint, because of the silhouette constraint [8], 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* obtained by silhouette. In addition, if we can obtain a depth map by template matching and stereo analysis, the space from the viewpoint to the foreground of the depth map is *vacant space* obtained by texture, 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 their hands and observes it, the person cannot acquire all of the information of the object simultaneously. Thus, the person manipulates the object to acquire necessary information. When the person manipulates it, the object's visible part changes depending on a type of holding, or the physical relationship of the object and the hand seen from the person's viewpoint. We classify the hand-object-view relationships into four classes according to the information which the holder of the object can acquire.



(a) Shape Acquisition (b) Visual Texture Acquisition (c) Haptic Texture Acquisition (d) Total Appearance Acquisition

Fig. 2. The Classification of the Relationships

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

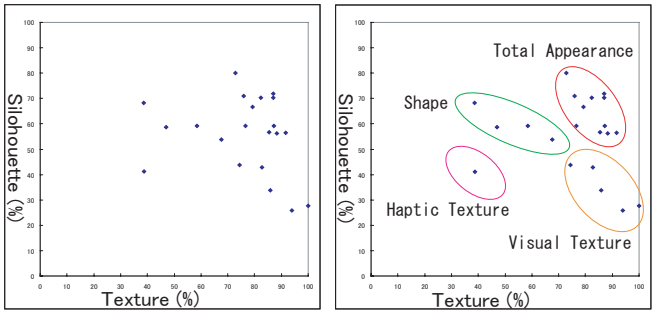
Visual Texture Acquisition :
When a person observes an object's visual texture, seen from the person's viewpoint, their hand is on the object's backside. In the captured image, mostly the object's texture is obtained, but small part of the object's silhouette is obtained as shown in Fig.2 (b).

Haptic Texture Acquisition :
When a person observes an object's haptic texture, their hand occludes most of the object. In the captured image, a small part of the object's 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 their viewpoint, their hand barely touches the object covering little important information. In the captured image, mostly the texture and the silhouette is obtained as shown in Fig.2 (d).

In order to confirm the classification, we developed the following experiment. We captured a scene that a person manipulated a cup to observe it. We captured 10 second and obtained 20 images. Fig.3 (a) shows that correlation of ratio of

the object's acquirable silhouette and ratio of the object's acquirable texture in one of the images. The ratio of its silhouette is the ratio of part which appeared as a contour of a silhouette to the object's contour in an image. And the ratio of its texture is the ratio of non-covered part of the object's texture to the object's part in one of the images. The correlation coefficient is -0.31 and there is a negative correlation between them. It means that a possibility that a silhouette can be acquired is high if it is hard to acquire texture, and a possibility that texture can be acquired is high if it is hard to acquire a silhouette. And this experimental result implied that hand-held action can be classified into the four classes as shown in Fig.3 (b). Since information obtained from a captured image



(a) Correlation Diagram (b) Classification

Fig. 3. Silhouette-Texture Correlation

changes with the class, ideally a computer system should change information used for reconstruction. 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 in hand-held action.

On the other hand, although it is impossible to obtain a silhouette of an object's concave part, the texture of the part can be obtained. Therefore, the part may be reconstructed by using texture information. And, although it is impossible to obtain an appropriate texture from a periodic texture or a monotonous texture, the silhouette of the part can be obtained. Therefore, the part may be reconstructed by using silhouette information.

Thus, silhouette and texture have a complementary relation. Therefore, using both is appropriate, and we use both to detect *vacant space*.

3 Dynamic Space Carving

3.1 Vacant Space Detection

Conventional 3D shape reconstruction approaches try to calculate the space of the object or the surface of the object. However in this problem under occlu-

sion by dynamic object, these approaches produce multiple inconsistent results caused by the dynamic object. Instead, we focus on the space, which is 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 information and from texture information. *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 from feature tracking of the object. See 4.1 for camera motion recovery. In this paper, we use the following notations as shown in Fig.4.

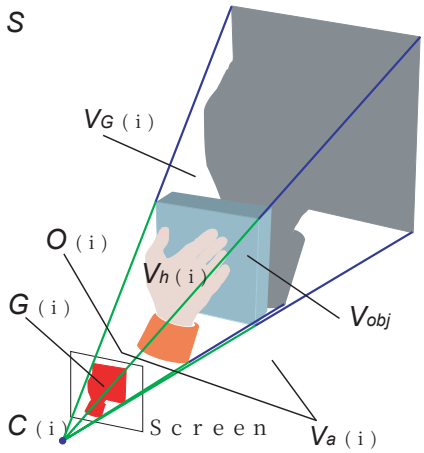


Fig. 4. The Parameters in Reconstruction Space

S : Space of an object-centered 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 that is back-projected $G(i)$ from viewpoint $C(i)$ (visual cone obtained by $G(i)$)
$O(i)$: Space from viewpoint $C(i)$ to the 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 in 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 voxel p is included in space V_x. In this research, the following conditions shall be fulfilled for S, V_{obj}, and $G(i)$.

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

3.3 Reconstruction by 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 voxels p ($p \in S$) where the proposition Eq.(2) is true.

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

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

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

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

Because Eq.(4) can be derived from our definitions, Eq.(5) can be concluded.

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

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

Therefore, since Eq.(6) and Eq.(7) are similarly derived from our definitions, Eq.(8) can be concluded.

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

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

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

The probability E that p ($p \in \overline{V_{obj}}$) is not included in $V_a(i)(i = 1, 2, \ldots, n)$ in all viewpoints is defined in 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 is captured for a long time, the probability becomes zero that the voxel 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 developed the following experiments. First, to check the fundamental effectiveness of the proposed technique and to evaluate errors, we reconstructed a simple shape object. Second, to check that the technique can apply to a common object, we reconstructed a toy figure of a monster.

4.1 The Flow of Processing

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

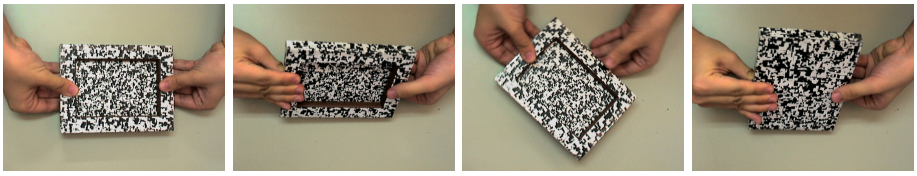
1. **Capture**
   We captured images in the situation of manipulating an object with a wearable vision sensor.
2. **Camera Motion Recovery**
   To recover the camera motion, we estimated the position of distance data. First, we extracted feature points with the Harris Corner Detector [9], and we obtained 3D feature points by stereo analysis in each image. Next, we estimated the camera position by using the advanced Iterative Closest Point algorithm [10], and we recovered the camera motion.
3. **Depth Map Acquisition**
   In each viewpoint, we match points between stereo pair images by coarse-to-fine template matching [12], and 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. **Dynamic Space Carving**
   We carved the vacant space by the technique shown in the previous section.

4.2 Evaluation with a Simple Shape Object

In the situation that a person manipulated a random patterned photo frame, we captured it with the wearable vision sensor at 1 frame per second as shown in Figs.5 (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 using the silhouette information is Fig.6 (a), and the result extracted by using the silhouette and texture information is Fig.6 (b). We compared those result models with the ground truth as shown in Fig.6 (c). The results are plotted in Fig.7 with graphs (a) and (b) showing respectively, the number of extra voxels, and missing voxels compared with the ground truth which is calculated from the geometrical facts. Extra voxels means voxels that the extracted model includes but ground truth does not, and missing voxels means voxels that the extracted model doesn't include but that the ground truth includes. The error ratios of final results to number

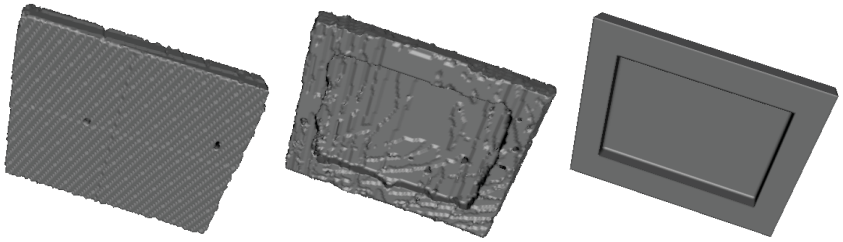
of ground truth voxels is as shown in Table 2. The time which this processing required is shown in Table 1.

In the experimental results, when we extract the object shape by using silhouette information with many viewpoint images, we can reconstruct the 3D object shape which is not included the hand. And, by using silhouette and texture information, we can reconstruct it with almost the same accuracy with fewer viewpoints, namely, by capturing for a short time. Moreover, we can reconstruct the concave part which is impossible to reconstruct only by using silhouette information. In the result by using silhouette and texture information, there are more missing voxels than the result by using only silhouette information. But compared with extra voxels, missing voxels are few. Therefore, it can be said that to use silhouette and texture information is effective for the 3D reconstruction problem of a hand-held object.



(a) 0 sec (b) 8 sec (c) 10 sec (d) 18 sec

Fig. 5. Captured Images of Photo Frame



(a) Extracted Shape by Silhouette Information

(b) Extracted Shape by Silhouette & Texture Information

(c) Ground Truth

Fig. 6. Photo Frame Data Set

4.3 Evaluation with a Complex Real Object

In the situation that a person manipulated a toy figure of a monster as shown in Fig.8 (a), we captured it with the wearable vision sensor. First, we obtained 13 asynchronous multiple viewpoint images. Next, we obtained vacant space, and

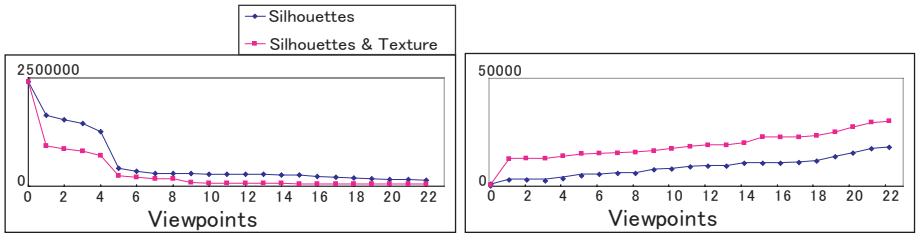
we extracted the figure shape by carving the vacant space. The extracted figure shape does not include the person's hand as shown in Fig.8 (b). Lastly, the toy figure's texture was mapped on the extracted shape as shown in Fig.8 (c).

Table 1. Processing Time

	Time (sec)	
	Silhouettes	Silhouettes & Texture
Capture	22	22
Camera Motion Recovery	43	43
Depth Map Acquisition	none	248
Silhouette Acquisition	4	4
Dynamic Space Carving	48	42
Total	117	359

Table 2. Error Ratios to Number of Ground Truth Voxels

	Silhouettes	Silhouettes & Texture
Extra Voxels	86.8 %	31.0 %
Missing Voxels	11.0 %	18.6 %



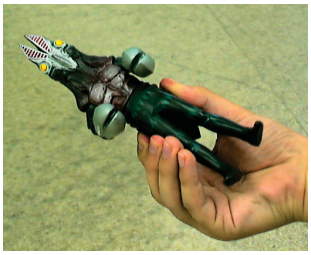
(a) Extra Voxels (b) Missing Voxels

Fig. 7. Graph of Errors

5 Conclusion and Future Work

In this paper, we showed the following results obtained with the wearable vision sensor. First, we analyzed and classified 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 experiment.

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



(a) Captured Image

(b) Extracted Shape

(c) Texture Mapped Shape

Fig. 8. Toy Figure Data Set

Acknowledgments

This paper is supported in part by the Ministry of Education, Culture, Sports, Science and Technology Grant No. 13224051.

References

1. I, Napier.: The prehensile movements of 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. K, N, Kutulakos., S, M, Seitz.: A theory of shape by space carving. IEEE International Conference on Computer VisionC(1999) 307–314.
7. H, Hoppe,CT, DeRose.CT, Duchamp.CJ, McDonald.CW, Stuetzle.: Surface reconstruction from unorganized points. Computer Graphics (SIGGRAPH '92 Proceedings)Cvolume **26**C(July 1992) 71–78.
8. 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.
9. C, J, Harris,, M, Stephens.: A combined corner and edge detector. In Proc. 4th Alvey Vision Conf.CManchesterC(1988) 147–151.
10. P, J, Besl.CN, D, McKey.: A method for registration of 3-D shapes. IEEE Trans. Patt. Anal. Machine Intell.Cvol. **14**(2)C(1992) 239–256.
11. Szymon, Rusinkiewicz.COlaf, Hall-Holt.C Marc, Levoy.: Real-Time 3D Model Acquisition. Transactions on Graphics (SIGGRAPH proceedings), (2002), 438–446.
12. S, T, Barnard.: Stochastic approach to stereo vision. International Journal of Computer Vision, (1989) 17–32.