https://doi.org/10.20965/jrm.2015.p0167

Paper:

Wearable Device for High-Speed Hand Pose Estimation with a Ultrasmall Camera

Motomasa Tomida and Kiyoshi Hoshino

Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibakaki 305-0006, Japan
E-mail: motomasa.tomida@gmail.com, hoshino@esys.tsukuba.ac.jp
[Received September 10, 2014; accepted January 14, 2015]

Operating a robot intentionally by using various complex motions of the hands and fingers requires a system that accurately detects hand and finger motions at high speed. This study uses an ultrasmall camera and compact computer for development of a wearable device of hand pose estimation, also called a hand-capture device. The accurate estimations, however, require data matching with a large database. But a compact computer usually has only limited memory and low machine power. We avoided this problem by reducing frequently used image characteristics from 1,600 dimensions to 64 dimensions of characteristic quantities. This saved on memory and lowered computational cost while achieving high accuracy and speed. To enable an operator to wear the device comfortably, the camera was placed as close to the back of the hand as possible to enable hand pose estimation from hand images without fingertips. A prototype device with a compact computer used to evaluate performance indicated that the device achieved high-speed estimation. Estimation accuracy was $2.32° \pm 14.61°$ at the PIP joint of the index finger and $3.06° \pm 10.56°$ at the CM joint of the thumb – as accurate as obtained using previous methods. This indicated that dimensional compression of image-characteristic quantities is important for realizing a compact hand-capture device.

Keywords: wearable hand capture, hand pose estimation, dimension compression of image characteristic quantity, hand image without fingertip, remote robot control

1. Introduction

Among the ways that rescue robots have been developed in various disaster cases is a humanoid rescue robot competition called the DARPA Robotics Challenge. The challenge introduces robots performing a variety of tasks. Google has acquired robot venture companies, so it is expected that the humanoid robot industry will expand further in the future.

Humanoid robots remain extremely difficult to operate, however, due to their complex mechanisms. Humanoid robot operations are classified as self-operated using artificial intelligence, i.e., self-operation, or as human-operated. In self-operation, the complexity of human operation need not be studied because the robots operate themselves, however difficult it may be to define their response to an unknown situation. Even humanoid robots with artificial intelligence require functions operable by a human operator. Human operation for robots to respond to unknown situations requires an input system that enables the operator to control robots with as complicated a structure as desired by the operator. Specifically, fine control of both the arms and upper body and of the fingers is necessary for robots to perform various tasks by way of the input system.

One of the simplest ways to control humanoid robots is to use human motion as input [1, 2]. This reflects the operator's motion in robot motion so that the operator control the robot as desired without having to undergo special operation training.

Many existing system technologies detect large parts of the body, such as the head, arm, or hand, accurately but cannot simultaneously detect small complicated parts such as the fingers. Optical motion capture system, for example, captures body highly precisely. Even with this system, however, it is extremely difficult to detect fingers and other small parts because the motion range of small parts is limited compared to that of large parts and because the structure of small parts is usually quite complex. Detecting finger motion thus requires that other technology be combined with existing ones.

The data glove is a typical technique capturing motion of hands. When a user wearing the data glove moves the hand and fingers, it is measured how much a built-in strain gauge bends to detect the angle of finger joints. The glove must contact the hand closely, which may make a user uncomfortable. The system with data glove is also complicated, which makes the device expensive and difficult to use. A depth camera could be used in contact-free detection that would minimize the load on users, but the detection range is still small and users cannot move around much due to the depth camera's low resolution.

To solve these problems, we decided to develop a system that controlled a robot hand and fingers without disturbing operator motion. We developed a wearable device of estimating the full articulation of a hand, also called a hand-capture device that used an ultrasmall camera and

a compact computer to estimate the shape of the hand and fingers. Note that in this paper, we do not use the term "hand capture" as it is used in clustering technology, which classifies hand-finger orientation into several patterns. Instead, we use hand capture as it is used in detecting the angles of individual finger joints as is done for the data glove. The ultrasmall camera on the device is 10 mm^2. It applies minimal load on the user and does not disturb hand motion. Google Glass, for example, has an ultrasmall camera that users are not aware of when wearing the glasses. A hand pose estimation algorithm that realizes high-speed image recognition even on a low-performance compact computer was used for the system device.

Hand pose estimation based on image recognition is classified roughly to two types – 3D-based and 2D-based.

With 3D-based models [3–6], the 3D hand-finger model estimates hand and finger shape by comparing a hand-finger profile created from the model and the profile of the hand and fingers obtained from input images. This makes it necessary to create multiple hand-finger profiles from the 3D model for comparison in individual estimation processes. This requires high computation power, so high-speed detection is difficult.

With 2D-based methods [7–10], however, the system has large amounts of hand-finger profile information in its database and compares the hand and fingers with profile data and hand-finger profiles of input images. This requires that the database be corrected. Shape estimation could be easily accelerated, however, by speeding up the data search.

Our group developed a high-speed, high-precision hand pose estimation method [11] that was basically 2D-based. Although this worked on a high-performance PC at high speed, it did not work on a low-performance PC or a compact computer. One of the reasons for this was the high number of dimensions – 1,600 – of characteristic quantities in the database.

The authors of this study solved the problem by dimensionally compressing characteristic quantities to reduce computational cost. In previous studies, 25 patterns of higher-order local auto correlation (HLAC) [12] were used as characteristic quantities but only one of them was found to be the important pattern. This is why we used this pattern in studying how to reduce the 1,600 dimensions of characteristic quantities to 64 dimensions.

To make the device more comfortable for an operator to wear, the camera was attached as close as possible to the back of the hand so that it did not disturb the operator's movement. If the camera was set this way, however, the user's fingertips often went outside the camera's imaging area because of how greatly the operator's fingers moved. We are therefore proposing a way to estimate the hand-finger shape if the entire image of the hand and fingers was shot when the camera was initially adjusted. This enabled shape to be estimated even if the fingertips were not in the imaging area during the shape estimation process.

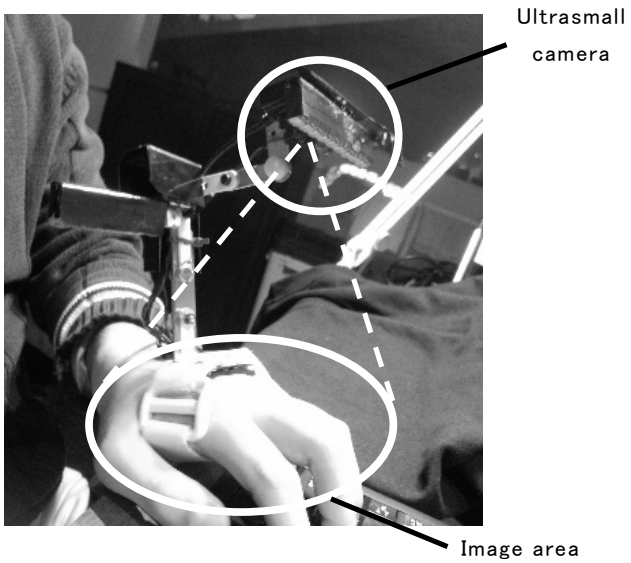


Fig. 1. Wearing the device.

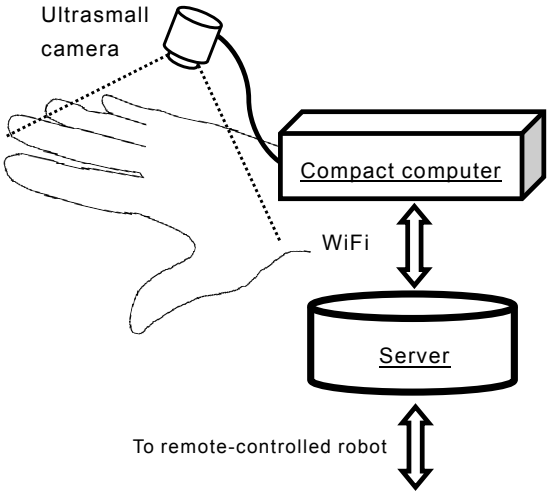


Fig. 2. Wearable hand-capture device configuration.

2. System Configuration

2.1. Hardware

Hardware consisted of the ultrasmall RGB camera and compact computer shown in **Figs. 1** and **2**. The user wears the camera on the hand to take images of the entire hand and fingers and puts the compact computer at the waist. A white LED makes the hand and fingers bright enough to separate the hand-finger profile from other image areas, as explained below. The device enables wireless communication by using a WiFi module to transmit estimation results wirelessly. Estimation results contain only joint angle data, so data volume is small enough that results are transmitted at high speed through WiFi.

2.2. Estimation Algorithm

(1) Database development

The system's database consists of hand pose data sets.

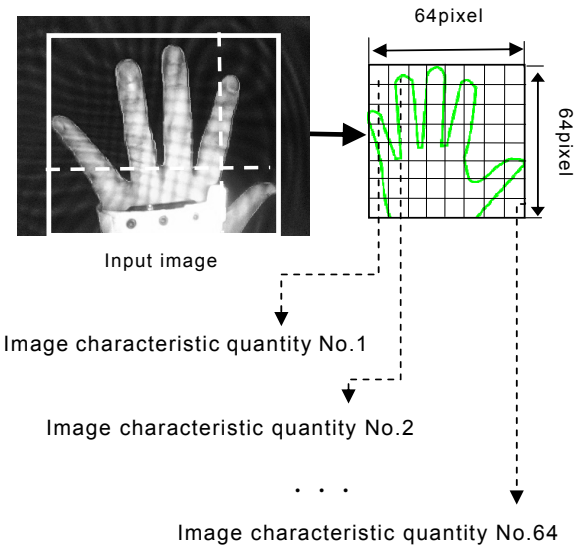


Fig. 3. Calculation of image shape ratio and image-characteristic quantities.

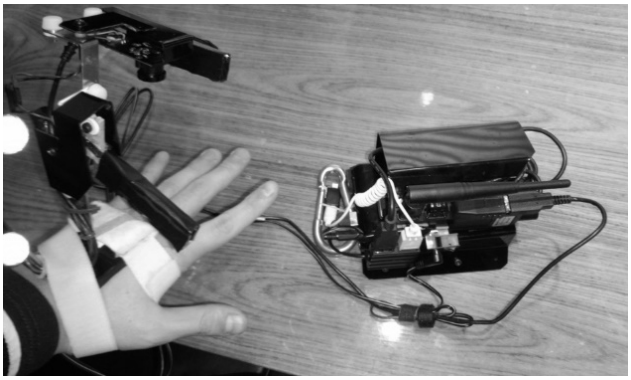


Fig. 4. Prototype used for evaluation.

Each data set contains the image-characteristic quantities acquired in the process explained below and the angle of each finger joint measured with data gloves.

To create the database, the operator wears the device on one hand and the data glove on the other, then moves both hands carefully synchronized to acquired data sets of image-characteristic quantities and finger joint angles.

The camera must first be calibrated to calculate image-characteristic quantities. The operator wears the device and adjusts the camera position and orientation. The hand is placed in the initial position shown in **Fig. 3** with fingers open and the camera placed to take images of the entire hand and fingers. The hand-finger area within the white frame in **Fig. 3** is detected automatically for image processing of the hand and fingers. The camera position is adjusted so that the horizontal dashed line in **Fig. 3** comes to the bases of all four fingers. The horizontal dashed line divides the area length from the top to the bottom of the image at a certain ratio. Camera orientation is adjusted in the same way so that the vertical dashed line comes to the base of the thumb. The vertical dashed line divides the frame into left and right at a certain ratio, as the horizontal line does. Image-characteristic quantities are detected from the hand-finger area obtained in this process and estimation is made by using quantities even if parts of fingertips lie outside the area.

The operator next moves the hand and fingers to various locations and image-characteristic quantities are calculated from a hand-finger image shot by the ultrasmall camera the operator is wearing. To calculate image-characteristic quantities, the profile of the hand and fingers is extracted from the image. The white-light LED of the device makes the hand and fingers brighter than the rest of the image area, so the hand-finger profile is extracted from the image by simply judging using a brightness threshold. The contour of the profile is then normalized to 64 pixel $\times$ 64 pixel and the normalized contour image is divided into 8 directions vertically and horizontally to create 64 divided images. The number of contour line pixels in each divided image is then counted and used as the 64-dimensional image-characteristic quantity.

(2) Estimation method

In estimation, the operator first puts on the device following the above procedures, and image-characteristic quantities are calculated.

Image-characteristic quantities are next compared to those in the database based on the Euclidian distance to find the data set which is the nearest from image-characteristic. The finger joint angle in the data set is then output as the estimation result.

3. Evaluation Experiments

3.1. Evaluation Method

In the first evaluation experiment, the estimation precision of the system using 64-dimensional image-characteristic quantities and using 1,600-dimensional image-characteristic quantities was compared. To create the system database, input data were obtained using the data glove and the number of data sets was found to be 30,000. Databases of 64-dimensional quantities and of 1,600-dimensional quantities are created in the same way, except that that with 64-dimensional quantities is created with the method presented in this paper and that with 64 with a conventional method [9] based on HLAC.

In the second evaluation experiment, present and conventional methods are compared for the same total data volume, i.e., the precision of the estimation with the conventional method using about 30,000 data sets is compared to the precision of the estimation with the method proposed in this paper using about 750,000 data sets – or 25 times more than in the conventional method. The prototype device shown in **Fig. 4** was used to evaluate the precision of hand pose estimation. A camera board of Buffalo BSW20KM11BK was used for the prototype device. Camera resolution was set to 640 $\times$ 480 pixel and image shooting speed to 30 fps. The lens used had a view

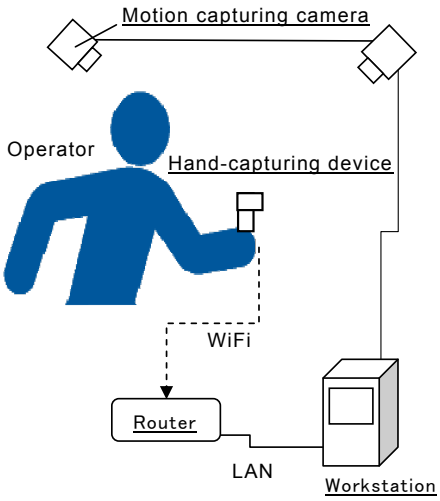


Fig. 5. System configuration for evaluation.

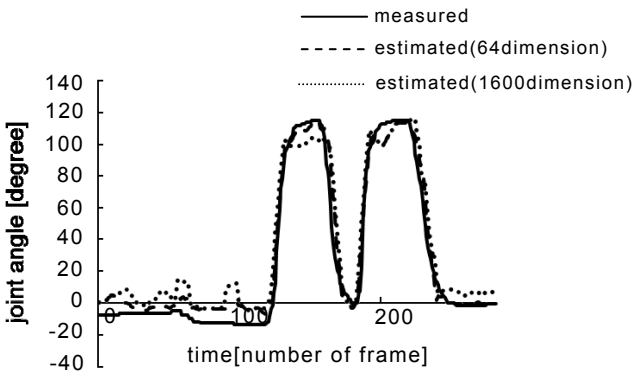
angle of 120°. A Beagle Bone Black A5C (CPU: 1 GHz, memory: 512 MB) was used. In the experiment, the operator wore the device on the left hand and a data glove (Cyber Glove Systems: CyberGloveII) on the right hand and carefully made the same motion with both hands. Data from the device were used as estimated data and data from the data glove as actual measurement data. Both types of data were then used to evaluate estimation precision.

The input system shown in **Fig. 5** combined the prototype device and an optical motion capture VICON was then assessed in experiments. The estimation result from the prototype was transmitted to the input system through WiFi. The prototype was powered by a portable battery. The motion capture process output used in this study was sent to a VR simulation system, not to a robot. Although the final purpose of the study was to control a humanoid robot, the VR system enabled users to see object motion that reflected their hand motion input. They could also see the VR object from the robot's viewpoint through HMD by making various operations which are such as grasping and pinching with their hands.

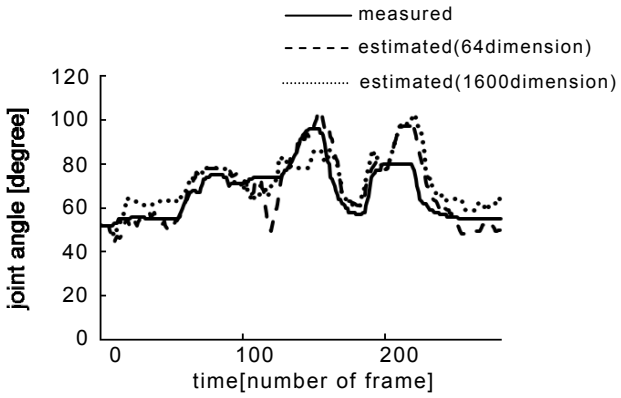
3.2. Experimental Results and Discussion

Estimation results for the first evaluation experiment are graphed in **Fig. 6**. The thick solid line represents actual measurement data, the thin solid line values estimated from the 64-dimensional image-characteristic quantities, and the dotted line represents the values estimated from the 1,600-dimensional image-characteristic quantities. **Fig. 6(a)** shows estimation results for the proximal interphalangeal (PIP) joint of the index finger and **Fig. 6(b)** that of the carpo metacaepal (CM) joint of the thumb. Results were obtained with the hand facing the camera. **Table 1** lists average errors and standard deviations in the estimation results for finger joint angles obtained in the 64-dimension system and in the 1,600-dimension system.

As the figure shows, there is almost no difference in estimation precision between 64 and 1,600 dimensions.



(a) PIP joint of index finger



(b) CM joint of thumb

Fig. 6. Estimation results for different dimensions of image characteristics.

Table 1. Estimation errors for different numbers of image-characteristic quantity dimensions.

		dimension	
		1,600	64
PIP joint of index finger [°]	average	−1.96	−3.92
	s.d.	8.42	10.88
CM joint of thumb [°]	average	−2.94	−4.81
	s.d.	7.02	7.55

Quantitative data on errors in **Table 1** show that average errors or standard deviation differed by only 2° at most between the system using 64-dimension image-characteristic quantities and that using 1,600-dimension image-characteristic quantities.

The fact that estimation precision was almost the same for 1,600-dimension data and 64-dimension data may be due to image division into 64 images. In the conventional system, 25 HLAC patterns were used to derive characteristic quantities from the length, tilt angle, and shape of the contour. In the 64-dimension method, only the contour length was used to reduce the dimensions. Estimation precision would ordinarily be low if contour line tilt information or shape information was provided. In the system proposed in this paper however, image-characteristic

Table 2. Estimation error for different numbers of data sets.

		dimension	
		1,600	64
PIP joint of index finger [°]	average	7.46	2.32
	s.d.	19.82	14.611
CM joint of thumb [°]	average	3.44	3.06
	s.d.	20.55	10.56

quantities are calculated for each of the divided images, and we estimated contour line tilt and the shape in each divided image from length information in adjacent images. The actual amount of information therefore remains almost the same as for the conventional method and the estimation precision of this method was at the same level as that of the conventional method.

Results of the second evaluation experiment are shown in **Fig. 7**. To enable estimation precision to be recognized intuitively, snapshots show estimation result computer graphics on the left and actual hand shape on the right.

Figure 8 shows estimation results and specifically compares actual data measured with the data glove and estimated data with the conventional and proposed methods. Although we could have used any joint from the index, middle, fourth, or fifth finger, we chose the PIP joint of the index finger for this study because it could be moved more than the others. The CM joint of thumb was chosen as a study target because its motion differed markedly from that of the four fingers.

Upper graphs in **(a)** and **(b)** show time variations in measured and estimated values. The solid line shows measured data, the dashed line estimated data with 64-dimensional image-characteristic quantities, and the dotted line estimated data with 1,600-dimensional image-characteristic quantities.

Lower graphs in **(a)** and **(b)** show estimation errors. The dashed line shows estimated data with 64-dimensional image-characteristic quantities, and the dotted line shows estimated data with 1,600-dimensional image-characteristic quantities. **(a)** is estimation results for the PIP joint of the index finger and **(b)** for the CM joint of the thumb. In experiments, the hand faces the camera.

As figures indicate, estimation with the conventional method was unstable and had large errors in many cases. In contrast, estimation with the proposed method was highly accurate in most hand-finger cases. This may be due to one of the following reason: the conventional method does not have many data sets that correspond to measured hand-finger postures, resulting in unstable estimation. The method, with 25 times more data sets than the conventional method, had data sets that approximated measured hand-finger postures, resulting in high precision estimation.

Table 2 lists average estimation errors and standard deviations for the PIP and CM joints. In estimation experiment results, the average and standard deviation of the estimation error for the PIP joint was $7.46° \pm 19.82°$ using the conventional method and $2.32° \pm 14.611°$ using the proposed method. Average and standard deviations for the estimation error for the CM joint was $3.44° \pm 20.55°$ us-

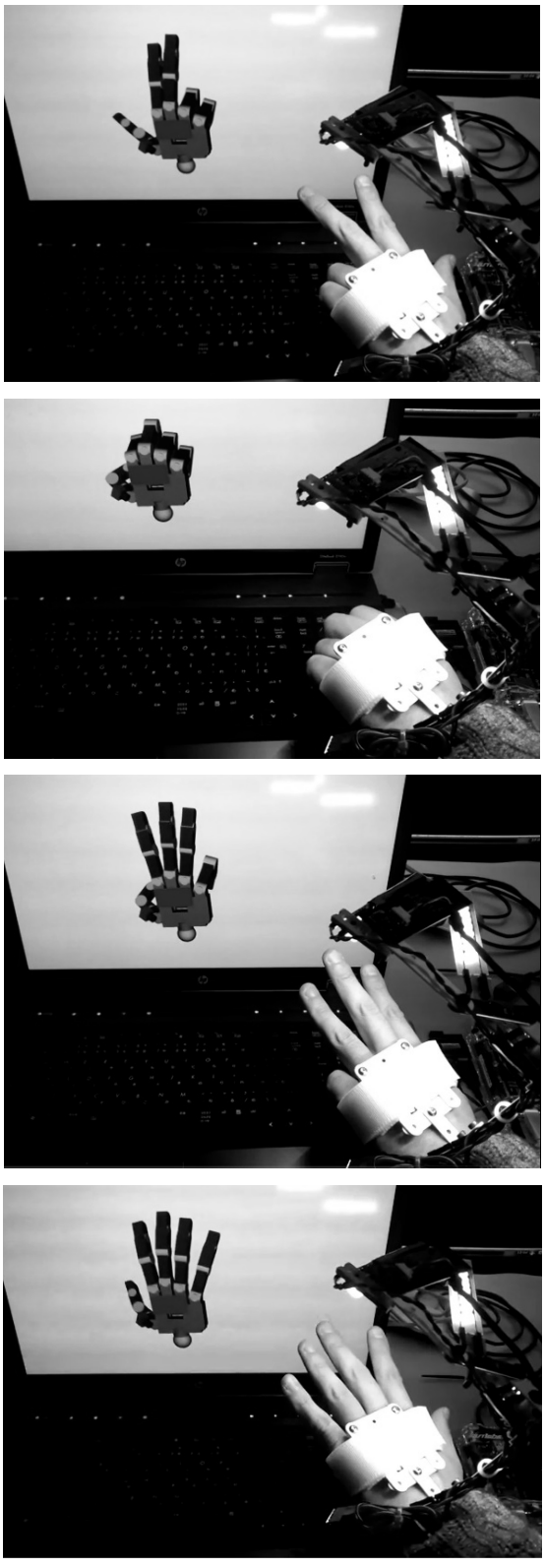
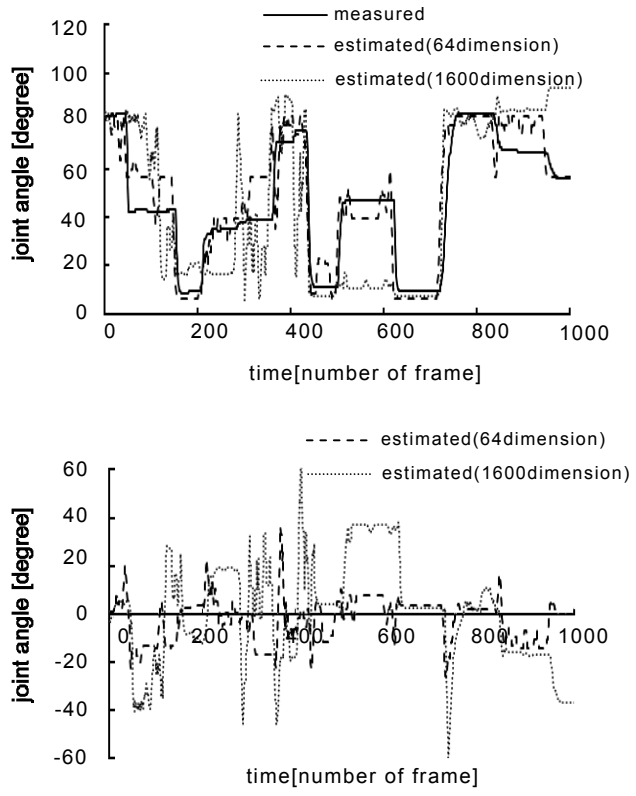
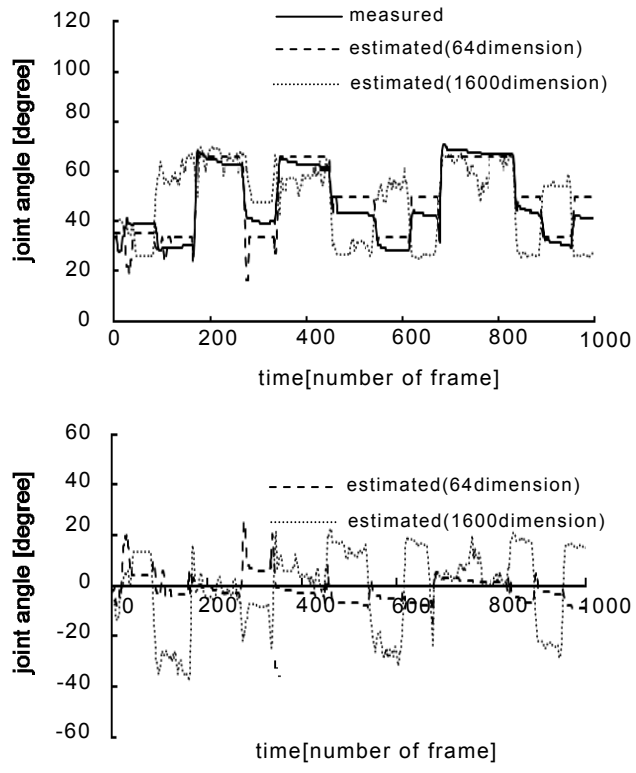


Fig. 7. Snapshots of estimation results.



(a) PIP joint of index finger



(b) CM joint of thumb

Fig. 8. Examples of estimation results.

ing the conventional method and $3.06^\circ \pm 10.56^\circ$ using the proposed method. In other words, average estimation error was almost the same for the two methods but standard deviation of error was decreased by about three fourths to one half using the proposed method.

The input system for evaluation shown in **Fig. 5** was demonstrated at various exhibitions, where users could experience through virtual reality how a remote-controlled humanoid robot disassembled an engine.

The motion of a user's head, arm, and fingers was detected by optical motion capture and was used to control the robot. Viewing the engine on a head-mounted display, users "grasped" engine parts by moving their hands and fingers to disassemble the engine. The device was used by many visitors to exhibitions and all operated the device successfully. Users would have needed training, however, if device operation had required a special skill such as that required by game controllers. The device we proposed only required ordinary motion of the hand and fingers and users were not puzzled by operation. These results indicate the possibility that a humanoid robot could be controlled as desired by the user if the device receiving input from the operator's motion were applied to the robot.

It is to be noted in demonstrations at exhibitions, however, that the weight of the prototype became an issue. The prototype was made with commercially available parts and hence was not as compact as it could have been. A dedicated machine developed could, of course, be made smaller. The smallest camera device for example would be 7–8 mm^2 and the compact computer just several centimeters in size if dedicated FPGA or SoC were employed – in short, as small as Google Glass. The weight would then no longer be an issue for a dedicated device.

4. Conclusion

We have realized hand pose estimation difficult with existing motion-capture systems by developing a wearable hand-capture device. In developing the device, we aimed to estimate the full articulation of a hand accurately – not to simply classify hand grasping or release – to reproduce actual hand pose. This technique is necessary to realize the robot function of grasping the appropriate part of an object in an appropriate way. The device also needed to be small enough to reduce the load on users and to enable their free motion.

With this target in mind, we have proposed a wearable hand-capture device consisting of a ultrasmall camera and compact computer. The device, worn on the user's hand, recognizes hand images and estimates hand pose. The largest problem was that the device required an algorithm for high-speed and high-precision estimation even on a compact computer with low performance. To solve this problem, our group realized high-speed high-precision hand pose estimation by using a two-step database search in which two kinds of image information, i.e., image shape ratios and image-characteristic quanti-

ties, were used. In this study, we focused on more important image-characteristic quantities to reduce 1,600 dimensions to 64 without lowering estimation precision. This dimensional reduction made the algorithm faster. To make the device more comfortable to wear, we put the camera close to the back of the hand. Fingertips sometimes moved outside of the area being used, however. This method estimated the hand pose from images even if fingertips were occluded, provided that an image of the entire hand and fingers was taken in advance in the camera adjustment process. This wearable hand-capture device realized high-speed stable estimation even with a low-performance computer and even with the ultrasmall camera placed close to the user's hand.

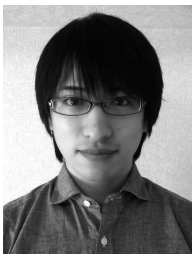
A prototype installed on equipment was evaluated to test the device. The joint angle of the PIP joint of the index finger was estimated to be $7.46^\circ \pm 19.82^\circ$ with the conventional method and $2.32^\circ \pm 14.611^\circ$ with the proposed method. For the CM joint of the thumb, the joint angle was estimated to be $3.44^\circ \pm 20.55^\circ$ with the conventional method and $3.06^\circ \pm 10.56^\circ$ with the proposed method. Although average estimation error remained the same, the standard deviation of the estimation error was reduced from about three fourth to one half. Users could also make complex operations simply by using the input system combined with an optical motion capture unit.

Acknowledgements

This study was supported in part by the Strategic Information and Communications R&D Promotion Programme (SCOPE), the KDDI Foundation, and the JST Adaptable & Seamless Technology Transfer Program through Target-driven R&D (A-STEP).

References:

[1] N. S. Pollard, J. K. Hodgins, M. J. Riley, and C. G. Atkeson, "Adapting human motion for the control of a humanoid robot," Proc. of IEEE Int. Conf. on Robotics and Automation, Vol.2, pp. 1390-1397, 2002,
[2] T. Asfour and R. Dillmann, "Human-like motion of a humanoid robot arm based on a closed form solution of the inverse kinematics problem," Proc. IEEE/RSJ Conf. Intelligent Robots and Systems, pp. 1407-1412, Las Vegas, Nevada, October 2003.
[3] J. M. Rehg and T. Kanade, "Visual tracking of high DOF articulated structures: an application to human hand tracking," European Conf. Computer Vision, pp. 35-46, 1994.
[4] Y. Kameda and M. Minoh, "A human motion estimation method using 3-successive video frames," Proc. Virtual Systems and Multimedia, pp. 135-140, 1996.
[5] S. Lu, D. Metaxas, D. Samaras, and J. Oliensis, "Using multiple cues for hand tracking and model refinement," Proc. CVPR2003, Vol.2, pp. 443-450, 2003.
[6] T. Gumpp, P. Azad, K. Welke, E. Oztop, R. Dillmann, and G. Cheng, "Unconstrained real-time markerless hand tracking for humanoid interaction," Proc. IEEE-RAS Int. Conf. on Humanoid Robots, CD-ROM, 2006.
[7] V. Athitos and S. Scarloff, "An appearance-based framework for 3D hand shape classification and camera viewpoint estimation," Proc. Automatic Face and Gesture Recognition, pp. 40-45, 2002.
[8] K. Hoshino and T. Tanimoto, "Real time search for similar hand images from database for robotic hand control," IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, Vol.E88-A, pp. 2514-2520, 2005.
[9] Y. Wu, J. Lin, and T. S. Huang, "Analyzing and capturing articulated hand motion in image sequences," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.27, pp. 1910-1922, 2005.
[10] K. Hoshino, E. Tamaki, and T. Tanimoto, "Copycat hand – Robot hand imitating human motions at high speed and with high accuracy," Advanced Robotics, Vol.21, No.15, pp. 1743-1761, 2007.
[11] K. Hoshino and M. Tomida, "3D hand pose estimation using a single camera for unspecified users," J. of Robotics and Mechatronics, Vol.21, No.6, pp. 749-757, 2009.
[12] N. Otsu and T. Kurita, "A new scheme for practical, flexible and intelligent vision systems," Proc. IAPR. Workshop on Computer Vision, pp. 431-435, 1998.



Name:
Motomasa Tomida

Affiliation:
Graduate School of Systems and Information Engineering, University of Tsukuba

Address:
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

Brief Biographical History:
2008- Graduate School of Systems and Information Engineering, University of Tsukuba
2011- Crscent, Inc.
2015- Ph.D. Candidate, Graduate School of Systems and Information Engineering, University of Tsukuba

Main Works:
• "Gesture-world environment technology for mobile manipulation – remote control system of a robot with hand pose estimation –," J. of Robotics and Mechatronics, Vol.24, No.1, pp. 180-190, 2012.
• "Optical tactile sensor assuming mathematical cubic distortion of elastic membrane," J. of Robotics and Mechatronics, Vol.21, No.6, pp. 780-788, 2009.
• computer vision
• virtual reality

Membership in Academic Societies:
• The Institute of Electronics, Information and Communication Engineers (IEICE)
• The Virtual Reality Society of Japan (VRSJ)

Name:
Kiyoshi Hoshino

Affiliation:
Professor, Graduate School of Systems and Information Engineering, University of Tsukuba

Address:
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

Brief Biographical History:
1993- Assistant Professor, Tokyo Medical and Dental University
1995- Associate Professor, University of the Ryukyus
2002- Associate Professor, University of Tsukuba
2008- Professor, University of Tsukuba
1998-2001 Senior Researcher of PRESTO project, Japan Science and Technology Agency (JST)
2002-2005 Project Leader of SORST project, JST

Main Works:
• "Copycat hand – Robot hand imitating human motions at high speed and with high accuracy," Advanced Robotics, Vol.21, No.15, pp. 1743-1761, 2007.

Membership in Academic Societies:
• The Robotics Society of Japan (RSJ)
• The Institute of Electronics, Information and Communication Engineers (IEICE)
• Japanese Society for Medical and Biological Engineering (JSMBE)