2004－CVIM－146　(6)
2004／11／11

ロボットの見まねによる運動技能獲得の試み

星　野　聖

筑波大学／JST　〒305-8573 茨城県つくば市天王台 1-1-1

概要　ヒトのような運動技能を，見まね（非接触的方法）によってロボットが獲得できるようにするためには，少なくとも次のような課題が解決されなくてはならない：第一に，とくにヒトの手や指の３次元動作の高速・高精度の運動推定．第二に，推定された軌道情報を，自由度数や構造，ダイナミクスなどが異なる自分のアクチュエータ動作に翻訳するアルゴリズム．第三に，動作のゴールが何であって，そのために軌道，力制御，時間的タイミングといった如何なる制御戦略を学習すべきかが理解できる能力，である．本報告では，同課題を達成するために著者が手掛け，最近，良好な成果が得られつつある研究内容，とくに高速・高精度の３次元ヒト手形状推定システムと，指先での紙つまみやペンを持っての書字動作生成が可能になりつつあるヒト型ロボットハンドについて紹介する．

Implementation of learning ability of human motion skills by watching in humanoid robots

Kiyoshi HOSHINO

University of Tsukuba / Japan Science & Technology Agency　1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

Abstract　To implement learning ability of human motion skills by watching, some problems at least should be solved as follows: Firstly, human hand posture estimation needs to be accomplished with high accuracy and high-speed processing. Secondly, algorithms to translate the estimated information into robot's own motor commands are necessary, where there are differences of, for example, dynamics, structures, and numbers of degree of freedom between humans as actors and teachers and robots as observers. Thirdly, ability to understand what is the goal of the motion should be implemented, then, the robot needs to know which trajectory should be selected, and how and when the force should be controlled. In this paper, the author introduces some results on robot's “learning by watching” project obtained recently, focusing mainly on human hand posture estimation systems and humanoid robot hands which can stably pinch at the fingertips.

1. Introduction

High-grade and complicated motions and information processing have come to be required of the robots, however, with the conventional level of engineering and technology, it was necessary for the user or designer to give consecutive and detailed instructions for motions to the robots. It is desirable for the robot itself to become capable of automatically acquiring and generating motions to be performed, by observing human actions. It is necessary in the first place to realize an imitating function of the “motions of hand and fingers” which are the most complicated of all human motions. To that purpose, it is essential that the three-dimensional shape of human hands and fingers can be estimated rapidly and with high accuracy.

A human hand, which has a multi-joint structure and a large degree of freedom, changes its shape in a complicated way. Moreover, since self-occlusions are produced in the hand due to its own palm or fingers, it was very difficult to estimate the shape. No satisfactory artificial system has yet been realized, in respect of both estimation accuracy and processing speed. In the past, a variety of attempts have been made, with a view to solving problems such as complexity of shape or self-occlusion in

the estimation of hand posture. For example, three-dimensional shape of hand was estimate by using a multi-camera and a set of small cubes called voxels [1]. However, no real-time processing is achieved, because it takes much time for the simulation of the model. Another group [2] realized real-time processing, but it required calculations of distance by a plural number of computers and infrared cameras. The hand shape was also estimated by using monocular camera in the group. However, because they perform matching processing by utilizing foresighted information about the object and that the protruding area on the image is a finger tip, the range of application is rather limited. There is also a study made in an attempt to estimate the hand shape, from the covered area of a three-dimensional shape model constructed inside the computer and the silhouette of the image [3]. However, a problem is that it takes much time for the processing, because of the difficulty of measures to be taken for adaptation to the complexity of shape and self-occlusion. A technical feature common to these series of studies is the processing for matching either a hand and finger model or a model of feature points constructed in advance in the computer with the human hand in the input image. In the matching processing, however, it seems difficult to achieve sufficient accuracy and processing speed, in the estimation of hand shape complicated in shape and having a lot of self-occluded areas.

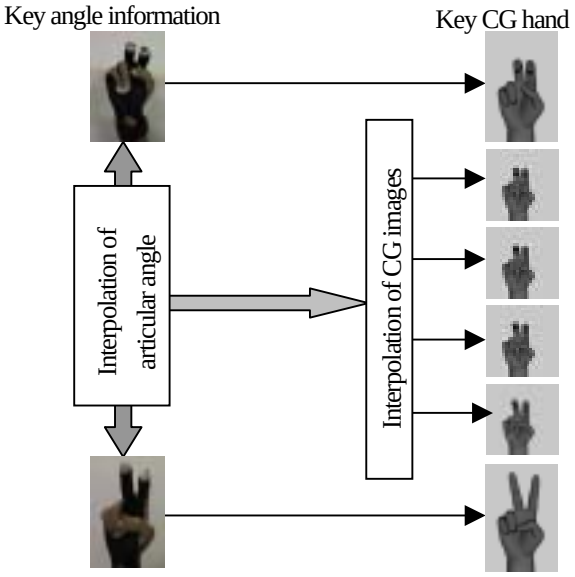

Fig.1 Interpolation of articular angle data and CG images of hand.

On the other hand, various types of robot hands have been developed [4]-[6] aiming at realization of cooperative and coexistence with men. One of important functions expected of a humanoid robot hand is a function of gently picking up something small, thin or fragile. However, putting this function into practice is more difficult than putting into practice the power-grasping function, and is delayed, because of existence of the following two problems: You have to accept to sacrifice the degree of freedom of motion due to the size and shape, and small motors for delicate control of fingertip may deteriorate the force control or increase friction loss.

For those reasons, in the present study, the author firstly introduces a system which performs estimation of shape by storing hand and finger images of as many different shapes as possible in advance in a database and searching similar images quickly and with high accuracy, against input image of hand shape to be estimated, regardless if the object shape is complicated or not and if there exists any self-occlusion or not. And then, the author proposes a proper and new mechanism of humanoid robot hands, by allowing to add some motional functions realizable only with a machine, without destroying the general harmony as a form of humanoid robot.

2. Hand posture estimation system

2.1 Database

In the first place, the author conducted measurement of articular angle data (hereinafter referred to as “key angle

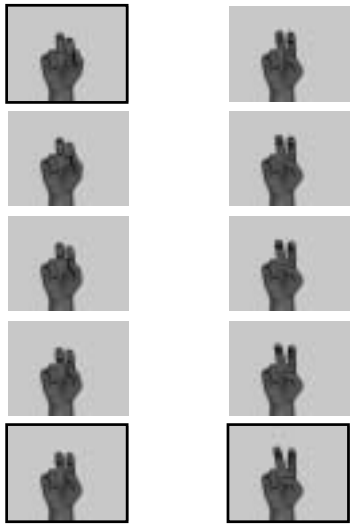

Fig. 2 Examples of interpolated CG images of hand.

data"). In the present study, the author determined articular angle at 24 points per hand in the form of oiler angle, by using data glove (cyberglove, Virtual Technologies). For the articular data in the shape of hand which requires a high detecting accuracy especially in the search for similar images, the measurement was made in a somewhat large number. And, we generated CG images of hand, based on the key articular angle data. CG editing software Poser 5 (Curious Labs Incorporated) was used, for the generation of images.

Secondly, from two key angle data, we interpolated a plurality of articular angle data in optional proportions. The interpolation of articular is a linear interpolation. Moreover, we also generated corresponding CG images of hand, based on the interpolated data. This operation makes it possible for the experimenter equipped with data glove to obtain CG images of hand in various shapes, with desired fineness, without taking the trouble of measuring the hand in all shapes. Fig.1 indicates a schematic chart of interpolation of articular angle data and CG images of hand. Furthermore, Fig.2 shows an example of interpolated CG images of hand. This figure represents an example of a case where articular angle was measured at 3 different points in time of actions changing from "guh (rock)" to"choki (scissors)" in "janken (rock-scissors-paper game)", and direct generation of CG and generation of CG by interpolation were made from adjoining 2 data. In both figures, the 3 images surround by a square represent the former, and the others show the latter.

Thirdly, we added data on differences among individuals. A wide variety of data are required for a database intended for searching similar images, because of existence of differences among individuals as shown in Fig.3. For example, in the hand shape "guh" of "janken", a great difference among individuals is liable to appear in (1) the curvature of the coxa position of the four fingers other than the thumb, and (2) the way of protrusion of the thumb coxa. Moreover, differences are liable to appear in (3) the way of opening of the index and the middle finger, and (4) the standing angle of the reference finger in the "choki" shape, but in (5) the way of opening, and (6) the way of warping, etc. of the thumb, in the "pah (paper)" shape. To express such differences among individuals in the form of CG hand, all you have to do is to adjust the parameters of the length of finger bone and the movable articular angle and, for that reason, we generated CG images of hand having differences among individuals on the basis of articular angle data obtained by the procedure described above. Fig.4 indicates an example of additional generation of CG hand in different shapes. In the figure, the X axis shows CG hands disposed in the order starting from those with larger projection of thumb coxa, while the Y axis presents from those with larger curvature formed by the coxa of the four fingers other than the thumb, respectively.

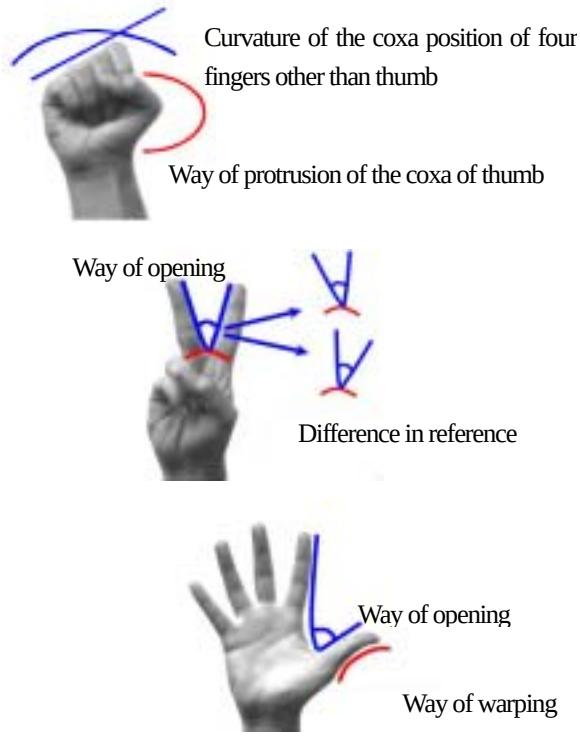

Fig. 3 Examples of differences among individuals.

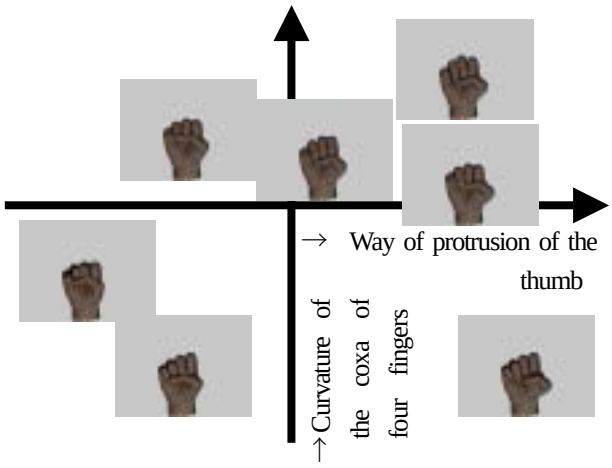

Fig. 4 Examples of supplemented data of differences among individuals.

By the procedure of the steps 1 to 3, we generated a total number of 15,000 CG hand images with this system.

In the fourth place, we calculated the amount of characteristics of the respective CG hands prepared. First, the respective CG hands were converted into images with 3 different kinds of resolution. Namely, 3 different kinds of image which are an image in which one picture element represents a single pixel (original image), an image in which one picture element represents 4 pixels (vertical) x 4 pixels (horizontal), and an image in which one picture element represents 8 pixels (vertical) x 8 pixels (horizontal). Next, after turning the images into black and white images with binary operation, we determined the center of the hand in the respective CG images.

$$x_g = \frac{1}{k}\sum_{i=1}^{160} x_{i*2} \qquad (1)$$
$$y_g = \frac{1}{k}\sum_{i=1}^{120} y_{i*2}$$

Here, x_g, y_g are coordinates of gravity center in the direction of X axis and the direction of Y axis, x_i, y_i are X, Y coordinates of white pixels, and k is the number of white pixels. In this system, we made a single processing per 4 pixels (2 pixels in vertical direction x 2 pixels in horizontal direction), to reduce the processing load. And, we split the respective CG hand images into 8 x 8 sections, based on the center point obtained. At present, one picture is constituted with 320 x 240 pixels. Lastly, we calculated the high-order local autocorrelational pattern at each level of resolution and for each split image. High-order local autocorrelational function is defined as the following formula:

$$x^N(a_1, a_2, \cdots, a_N) = \int f(r) f(r+a_1) \cdots f(r+a_N) dr \qquad (2)$$

Here, X^N is the correlational function near the point r in dimension N, (a_1, a_2, ---, a_N) is the direction of displacement, f(r) is the brightness value at the pixel position r, and N is the dimension number (N = 2 in the present study). Except for equivalent patterns due to parallel shifting, the high-order local autocorrelational patterns can be expressed in 25 different kinds [7]. However, since the patterns M1 to M5 become smaller in value compared with other patterns, we squared the number of pixels at the reference point for M1, and further multiplied it with the number of pixels at the reference point for M2 to M25. Finally, we normalized the concentration values of M2 to M25, by dividing them with the concentration value of M1.

In the fifth place, we reduced the amount of characteristics. Namely, by the above-described procedure, the total number of the amount of characteristics comes to 4,800 dimensions per CG image (= resolution 3 x split pictures 64 x high-order local autocorrelational patterns 25), and this number of dimensions is too large for searching CG images similar to an unknown image input at high speed. Therefore, we attempted to reduce the amount of characteristics, by using principal component analysis, as shown by the following formula:

$$z_{kp} = \sum_{l=1}^{div}\sum_{m=1}^{pnum}\sum_{n=1}^{25} a_{klmn} x_{plmn} \qquad (3)$$

Here, Z_{kp} is the marks of principal component of the data p in the **k**-th principal component, X_{plmn} is the **n**-th amount of characteristics in the **m**-th picture of the first resolution of the data p, a_{klmn} is the factor loading of the **n**-th amount of characteristics in the **m**-th picture of the first resolution of the **k**-th principal component, div is the number of resolutions, and pnum is the number of split pictures.

Next, we calculated the contribution ratio of the respective principal components. Here, contribution ratio is a coefficient which expresses to what extent the respective principal components explain the original information, and can be expressed as the following formula:

$$C_k = \sum_{l=1}^{div}\sum_{m=1}^{pnum}\sum_{n=1}^{25} b^2_{klmn} \Big/ (div * pnum * 25) \qquad (4)$$

Here, C_k is the contribution ratio of the **k**-th principal component, and b_{klmn} is the correlational coefficient of the

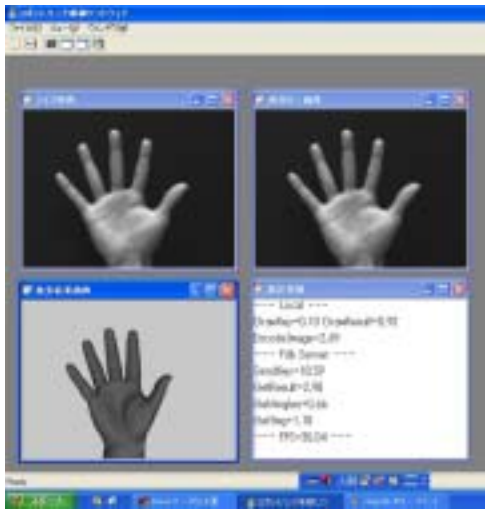

Fig 5 An example of results of estimation.

marks of principal component Z_{kp} and x_{plmn}, and they are defined as shown by the following formula:

$$b_{klmn} = a_{klmn}\sqrt{\lambda_k} \tag{5}$$

Here, a_{klmn} is the factor loading of the **n**-th amount of characteristics in the **m**-th picture of the first resolution of the **k**-th principal component, and λ_k is the characteristic value of the **k**-th largest correlational matrix. Lastly, we determined the number of principal components based on the cumulative contribution ratio. The following relational equation is established for the contribution ratio $C_1 \geqq C_2 \geqq C_3 \geqq \cdots \geqq C_k \geqq C_{k+1} \geqq \cdots \geqq C_{div*pnum*25}$. In the present study, we determined the number of principal components used for reduction of the amount of characteristics, with a cumulative contribution ratio of 95% or so as target.

$$\sum_{k=1}^{div*pnum*25} C_k \geq 0.95 \tag{6}$$

In this system, we decided to use the first principal component to the tenth principal component with which a cumulative contribution ratio of 97% is obtained.

In the sixth place, a table was prepared in which all the data are rearranged according to the magnitude of the marks of principal component, for each principal component from the first principal component to the tenth principal component. This makes it possible to perform collation of an input unknown image with CG images having similar amounts of characteristics.

In the seventh place, we determined in advance the number of objects to be searched in limited area, for efficient search of similar images. To be concrete, we selected data having the marks of principal component Z_{kp} either the same with the unknown image or closest to it in the respective principal components, and the data before and after the number corresponding to the contribution ratio of the respective principal components, as object of search. The number of candidates of the respective principal components is as shown by the following formula:

$$dc_p = Cc * \frac{\lambda_p}{\sum_{i=1}^{10} \lambda_i} \tag{7}$$

Here, dc_p is an estimated number of candidates of the **p**-th principal component (**p** = 1, 2, ---, 10), and Cc is the sum of the estimated candidates, which is a number determined in advance. In this system, *Cc* was set for *Cc* = 300. $\lambda p/\Sigma\lambda i$ is the contribution ratio of the component **p** among 10 principal components.

2.2 Search of similar images

Firstly, the motions of a human hand placed in front of the background screen are photographed with a single unit of monochrome high-speed camera (Megaplus, ES310/T). The sampling frequency was set for either 60 fps or 125 fps. No particular treatment was made for removing the background image, because a unicolor screen in black or white was used for the background.

Secondly, calculation is made of the amount of characteristics of the input image. In the first place, images of hand varied in 3 different levels of resolution are generated. They are an image in which one picture element represents a single pixel (original image), an image in which one picture element represents 4 pixels (vertical) x 4 pixels (horizontal), and an image in which one picture element represents 8 pixels (vertical) x 8 pixels (horizontal). In the case where the shape of hand is comparatively simple as when the palm faces the front side, etc., the extraction of characteristics becomes easier with a lower level of resolution. On the contrary, we may instinctively foresee that it will become difficult to estimate the way of bending of the respective fingers if the resolution is lowered, in the case where the hand faces a diagonal direction with turning of the wrist.

Next, in the same way as at the time of construction determine the center point of the hand according to the formula (1). And, split the input image into 8 x 8 sections

Fig. 6 Examples of behaviour of human hand and estimated joint angles.

of database, after giving a binary expression to the image, on the basis of the center point obtained. Although different ways of splitting were attempted (without splitting, splitting into 2 x 2 sections, 4 x 4 sections, 8 x 8 sections, 16 x 16 sections, 32 x 32 sections) in the preliminary tests regarding splitting, it has already been confirmed that, in small splitting into 16 x 16 sections or over, the accuracy of search does not make any marked improvement even with an increase in the amount of characteristics.

Lastly, calculate the high-order local auto-correlational pattern at each level of resolution and for each split image, according to the formula (2).

In the third place, calculate the marks of principal component for each principal component, according to the formula (3). Here, **p** = 1 applies in the case of input image.

In the fourth place, collate with the database, and select candidate CG images of hand. Here, the number of candidates of the respective principal components is determined in advance according to the formula (8). Therefore, the data having the closest marks of principal component Z_{kp} in the respective principal components and a number of candidate CG images of hand corresponding to the contribution ratio of the respective principal components are selected.

Next, calculate the degree of similarity, by the following formula, between the input image and the candidate CG images:

$$E_r = \sum_{i=1}^{10} (f_i(x_r) - f_i(x_t))^2 \qquad (8)$$

Here, $f_i(x)$ is the marks of principal component of the first principal component calculated from the amount of characteristics, x_r is the amount of characteristics by high-order local autocorrelational function of the candidate **r**, and x_t is the amount of characteristics by high-order local autocorrelational function at the hour t. This equation means that the smaller the value of Euclidean distance E_r the higher the similitude between the two images. Data p which minimizes E_r was set as the image to be searched, and the articular angle data possessed by the data p was taken as estimated angle.

Lastly, remove errors in the results of search, according to the formula given below. Namely, if the results of search at the hour t and the hour t-1 are found in the range of allowable articular angles, the search at the time **t** is terminated. On the contrary, in case an articular angle greatly different from that of the time **t**-1 is selected, select another candidate with the second smallest Euclidean distance E_r, and calculate to see if it is in the allowable range or not.

$$A_p = \sum_{i=1}^{24} (ang_{i(t)} - ang_{i(t-1)})^2 \qquad (9)$$

Here, A_p is an allowable value, **i** is the joint number of data glove (**i** = 24 in the present study), and $ang_{i(t)}$ is the **i**-th articular angle at the time **t**.

2.3 Experiments

To study effectiveness of this system, the author conducted search of similar CG images on moving images of human hand motions. By using a monochrome high-speed camera, the tester freely moved his hand fingers in front of a white screen, in the state in which his right palm faces the front side, but limited his wrist slewing motions. Fig.5 indicates an output image during the test captured on the screen, as example of estimated result. Of the 4 windows in the drawing, the window at the top left shows the captured image, the top right window shows the image forming the subject of search at the current time, the bottom left is the CG image as result of search, and the bottom right is the monitored picture. The sampling frequency of the high-speed camera is 125 fps. From this drawing, you can see that a processing speed of approximately 30 fps is obtained, even in the case where a personal computer of comparatively low functions having a clock frequency of CPU of 1 GHz (Pentium III, main memory 256 MB) is used. In the case where a personal computer of comparatively high functions (Pentium IV 2.8 GHz, main memory 1 GB) is used, the processing speed improved to 40 to 50 fps. In addition, by comparing the hand shape of the image forming the subject of search at the top right with that of the image as result of search at the bottom left, we can instinctively understand that similar images are searched with a fairly good accuracy.

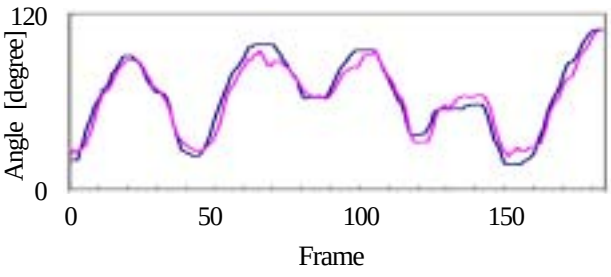

Fig. 7 Estimated accuracy of articular angle.

A weak point of this system is that there is no way to quantitatively evaluate the searching accuracy of the results in their original form. For that reason, to turn the articular angle of the hand fingers input as unknown image into a known value, the tester moved his hand fingers in front of a black screen, by wearing a thin white glove over the data glove. This enables to quantitatively study the difference between the two, because the CG images of hand are generated essentially from the data glove angle data in the database, and that the unknown image handled in the experiment in section 3.2 also has angle data.

Fig.6 indicates examples of behaviour of human hand and estimated joint angles superimposed with CG frame works. Fig.7 indicates an example in which articular angle data obtained with data glove and articular angle data possessed by the CG images of hand estimated from the image of its motions are drawn one upon another. The drawing shows, from top to bottom, the middle finger, the index and the thumb. It can be read quantitatively that estimation of hand shape is made with high accuracy with each finger.

By the way, with this method, a CG image of a shape most similar to the human hand shown on the unknown input image is selected. As described earlier, the CG images of hand stored in the database are prepared from articular angle information measured by data glove. On the other hand, the authors already reported, on the technologies of “robot hand mechanism capable of generating skillful motions” and “controlling robot hand from articular angle information of data glove” [8]. Therefore, at the point in time when the most similar CG hand is selected from the database, it becomes possible to control a robot hand in the same shape as that of a human hand without hardly any delay.

3. Humanoid robot hand

3.1 Mechanism of DIP joint

The mechanism of PIP (middle joint of finger) and MP (base joint of finger) require a large torque capacity because of a large mass of the finger mechanism to be supported. But there is a limit to the resolution of the produced torque, as general characteristic of a drive mechanism, and it is rather difficult to generate for the drive mechanism of PIP and MP to stably produce a fine fingertip force. On the other hand, the mechanism of DIP joint, which supports only a small and light mass, requires a small torque capacity, and can therefore produce a fine fingertip force comparatively easily. It also has an advantage that the transmission loss of force from the mechanism of DIP to the fingertip is zero, and is therefore an optimal mechanism to be provided with a function of producing a fine fingertip force.

In the present study, the author proposes, as a new system different from the conventional one, a system not imposing the function of strongly grasping an object on the DIP mechanism but expecting only production of a fine fingertip force required for delicately pinching an object. To be concrete, we add a small DIP-drive mechanism built up by giving priority to possibility of incorporation between DIP and PIP joints, and realize independent motions and fine control of fingertip force of the DIP joint.

A new grasping method of this robot hand will be explained below. To put it briefly, this method consists in transmitting a powerful grasping force to the object at portions on PIP and MP joints, and transmitting only a weak holding force to the object at the fingertip. For that purpose, provide a sufficient margin in the direction in which the fingertip moves away from the object in the movable range of DIP. At the same time, the author

Fig. 8 Stable pinching of the object.

Fig. 9 Twisting of the thumb.

Fig. 10 Writing characters with pen.

arrange the shape of the inside portion of PIP and MIP joints to facilitate their contact with the object.

The main advances of this new method are the following: As the acting point of the grasping force moves from the fingertip to portions on PIP and MP joints, it becomes possible for the mechanism of PIP and MP to generate a stronger grasping force with the same torque. Since a flexible fingertip force by DIP joint mechanism is added to this grasping force, a slip-free hold becomes easier. Moreover, because the movable range of rotation of DIP is widened here, it also provides an effect of stably pinching the object by putting the finger cushion widely in contact with the object, as shown in Fig. 8.

3.2 Twisting mechanism in thumb

When the tip of the thumb and the tip of other fingers being touch each other face to face in the human hand, the contact portion between the two is the cushion at fingertip on the thumb, but it is often a position off the fingertip cushion on the part of other fingers. This phenomenon is produced because the thumb does not have any twisting function. A human being has soft skin and flesh at the fingertip and a high control performance of motion and force at the respective fingertips, and can therefore realize a stable pinching function even if the thumb and the finger do not face each other exactly at the cushion part. However, a general robot hand has only a much lower control performance of motion and force compared with a human being, as mentioned before. DIP joint mechanism introduced previously to solve this problem demonstrates a force control function in one direction only. To fully utilize this capacity, it is desirable that the fingertip force produced by DIP joint mechanism at the tip of the two finger groups face each other justly, namely the two fingertips oppose each other exactly at the cushion. For that reason, the author adds a twisting function of the thumb, which does not exist on the human thumb, to realize the degree of freedom of motion necessary for stable grasping, as shown in Fig. 9.

3.3 Experiments

The experiment was carried out to confirmed the force control characteristics of the joint of the fingertip. The methods were as follows: the sin curve for the position control was given to the motor which linkingly drives the base and middle joints, which forces the fingertip to the mount. Simultaneously, the force control of the joint of the tip was carried out so that it may not exceed the limit of contact force between mount and fingertip. The film force sensor was attached on the mount, and the contact force at the fingertip was measured.

The results of generating force and angle displacement of the fingertip joint at the limitations of 90gf and 140gf respectively showed that the fingertip collided with the mount at first 1 second, and since then, the contact was maintained. In both cases, It was confirmed that the force control as contact force does not exceed the threshold and that the joint of the fingertip flexibly moves. These results suggest that adding the joints full of controllability makes the delicate force control effective at the fingertip, even if the generating force is weak.

Then, the experiments were carried out to confirm whether the robot hand can generate the motion of writing characters with a pen at the fingertips. The results of experiments confirmed not only that the motion of writing letters with a pen is possible but also that the maintenance of the pen with different contact points of fingertip of the thumb is possible, as shown in Fig. 10.

References

[1] E. Ueda, Y. Matsumoto, M. Imai and T. Ogasawara: "Hand pose estimation using multi-viewpoint silhouette images," *Proc. 2001 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS'2001)*, pp.1989-1996, 2001.

[2] M. H. Jeong, Y. Kuno, N. Shimada and Y. Shirai: "Recognition of shape-changing hand gestures," *IEICE Transactions Division D*, E85-D, 10, pp.1678-1687, 2002.

[3] N. Shimada, K. Kimura, Y. Kuno and Y. Shirai: "3-D hand posture estimation by indexing monocular silhouette images," *Proc. 6th Workshop on Frontier of Computer Vision*, pp.150-155, 2000.

[4] S. C. Jacobsen, J. E. Wood, D. F. Knutti and K. B. Biggers: "The UTAH/M.I.T Dextrous Hand: Work in Progress," *Intl. J. of Robotics Research*, 3, 4, pp.21-50, 1984.

[5] Gifu Hand: http://www.kk-dainichi.co.jp/

[6] Shadow Dextrous Hand: http://www.shadow.org.uk/products/newhand.shtml

[7] S.Odo and K.Hoshino: "Hand shape identification using higher-order local autocorrelation features of log polar coordinate space," *Journal of Robotics and Mechatronics*, 15, 3, pp.534-540, 2003.

[8] K.Hoshino: "Control of dexterous robot hand by data glove," *Proc. Intl. Tech. Conf. on Circuits/Systems, Computers and Communications*, 6F3L-5, pp.1-4, 2004.