https://doi.org/10.20965/jrm.2009.p0749

Paper:

3D Hand Pose Estimation Using a Single Camera for Unspecified Users

Kiyoshi Hoshino and Motomasa Tomida

Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
E-mail:hoshino@esys.tsukuba.ac.jp

[Received April 20, 2009; accepted October 26, 2009]

The three-dimensional hand pose estimation this paper proposes uses a single camera to search a large database for the hand image most similar to the data input. It starts with coarse screening of proportional information on hand images roughly corresponding to forearm or hand rotation, or thumb or finger bending. Next, a detailed search is made for similarity among selected candidates. No separate processes were used to estimate corresponding joint angles when describing wrist's rotation, flexion/extension, and abduction/adduction motions. By estimating sequential hand images this way, we estimated joint angle estimation error within several degrees – even when the wrist was freely rotating – within 80 fps using only a Notebook PC and high-speed camera, regardless of hand size and shape.

Keywords: 3D hand pose estimation, single camera, unspecified users, proportional information on the hand images

1. Introduction

Enabling computer vision to estimate human hand poses or postures at high speed and highly accurately regardless of hand shape while the wrist or shoulder is rotating would make possible two things:
(1) A robot could automatically acquire movement by observing human movement without robot designers' computer programs. To design a robot to communicate in sign language, for example, signing could be demonstrated and meanings taught to the robot, which could then predict hand or arm movement and automatically generate such movement. This "learning by observation" may be the human being's most intelligent and adaptive control. Remote control could be realized enabling users to control robots or machines by their hand or finger movement similar to movements in their daily living, as shown in **Fig. 1**.
(2) Information could be input regardless of the user's positioning, e.g., lying in bed. Multitouch screens (multipoint input devices) and holographic displays cannot yet replace the mouse or keyboard in functionality, e.g., data entry while lying down. If hand pose and positioning could be estimated at high speed highly accurately, however, using a single camera, images displayed to the user or a head-mounted display could be changed based on user hand movement, such as shown in **Fig. 2(a)**. For inputting three-dimensional (3D) models such as clay art, the user need only move a hand or arm as desired to form a shape, as shown in **Fig. 2(b)**.

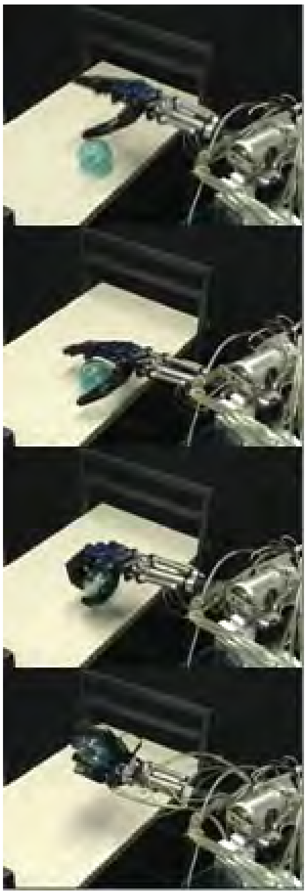
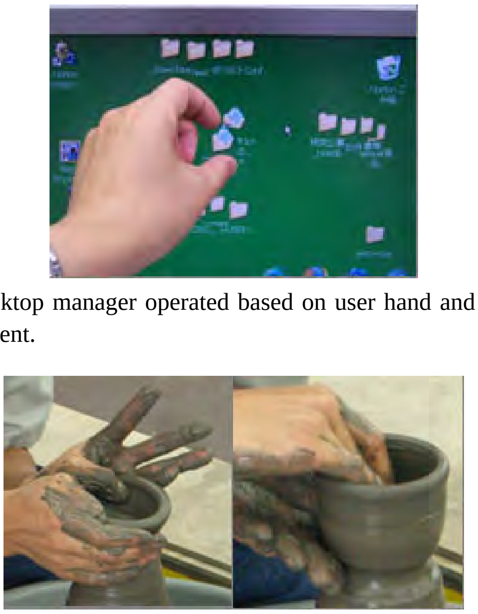


Fig. 1. Robot telecontrol by vision.

Two approaches are used to roughly classify conventional hand pose estimation – 3D-model-based and 2D-appearance-based. The 3D-model-based approach [1–6] involves extracting local characteristics, or silhouettes, in image recorded using a camera and fitting a 3D hand model constructed beforehand on a computer. While this approach estimates hand shapes highly accurately, it processes self-occlusion poorly and requires long processing



(a) Desktop manager operated based on user hand and finger movement.

(b) 3D models such as clay art input simply by their hand and arm movement.

Fig. 2. Examples of projected system use.

time. The 2D-appearance-based approach [7–9] involves directly comparing an input image to an image stored in a database. While this approach reduces calculation time, if 3D changes in hand appearance – including wrist and forearm movements – are not an issue, this approach requires a large reference database and robot hand movement is difficult to control using imitation. If basic difficulty in estimating hand poses lies in hand shape complexity and self-occlusion, high-accuracy poses become theoretically possible to estimate, but this requires an extensive database including all possible hand images, including complexity and self-occlusion. The feasibility of this approach therefore depends on the search algorithm.

Regarding the 2D-appearance-based approach, Hoshino et al. proposed using computer graphics (CG) editing software and data gloves to create a large database containing personal hand pose attributes such as joint movable range and bone length [8]. They developed a search algorithm that shortens search time in looking for unknown input images by using a multilayer database based on a self-organization map accompanying self-multiplication and self-extinction so that similar hand images are brought into closer proximity and by reducing the search area so that no data other than that near the search result during the previous search time will be inquired about [10]. A study using information on fingernail location with lower level image characteristics for estimating hand poses [11] has proposed using a multilayer database, but does not show image characteristics, including fingernail location information, applicable for individual layers defined by representative values.

One of the disadvantages of this in finding similar images in a multilayer database is that where the hand pose is changing quickly, candidate images may exit the search range, failing to find the best matching image. A second disadvantage is that even if the hand pose changes slowly, once a dissimilar image is found, the search range for the next time cannot be changed from that where the image was found. A third disadvantage is that a database created as multilayer using a priori facts instead of statistical information on the similarity of images raises problems in determining appropriate image characteristics, including joint angle and fingernail location, for layers as defined by a representative value.

We therefore are proposing 3D hand pose estimation using a single camera that retrieves the hand image most similar to data input from a large nonmultilayer database. We also studied whether this works for unspecified users. Specifically, coarse screening is used first as proportional information on hand images, e.g., height and right-bias roughly corresponding to forearm rotation, bending of the thumb or fingers. A detailed search is then made for similarity for selected candidates. To describe wrist rotation, flexion/extension, and abduction/adduction motions, no separate was used to estimate corresponding joint angles. Instead, hand images already containing these were prepared in a database to provide real 3D change in hand appearance.

2. System Configuration

2.1. Constructing a Database

Hand images, finger joint angles and forearm rotational angles

The database was prepared with storage by pairing individual hand images and finger and wrist angles synchronously acquired from a camera and data glove. Images were recorded using a universal serial bus (USB) camera at a resolution of 320 × 240 pixels laterally and vertically viewing hands and fingers on a big enough screen. Fingers and wrist angles were acquired using a data glove (Cyber Glove, Virtual Technologies Inc.), that obtained 18 types of angular information on the hand at a time. When the data glove was worn, however, image features specific to the data glove appeared, so a thin white glove was worn over the data glove. Although this makes color, texture, shape, and posture slightly different from those of the human hand, it could handle hand images better, as explained later. Wrist rotation was recorded using a small light-weight bar on the wrist, recording rotation with anther USB camera above. Although a database can be prepared even if background image already exist using skin-color extraction, we used a black screen as the background to minimize background influence on estimation.

The current database has 30,000 datasets. Some 100,000 were first prepared with experimenters' thumbs and fingers moved every 20° of wrist rotation within 0 to 180°. Each wrist rotation thus has 10,000 hand poses. Thumb and fingers were moved in the manner of a binary scale. The database has both bent and stretched fingers and various hand poses, because the data glove and

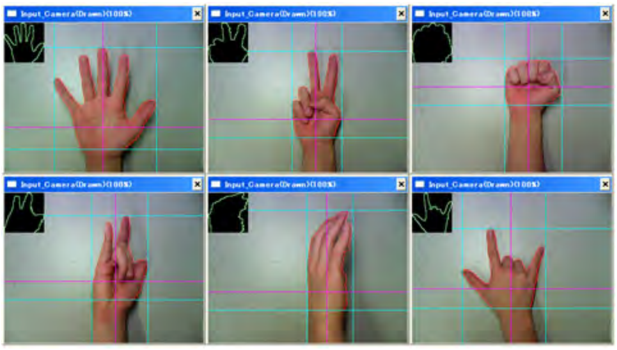


Fig. 3. Examples of hand extractions.

USB camera were recording based on a sampling frequency of 30 fps, then the number of datasets was decreased based on similarity, eliminating poses with almost the same angles. Based on preparatory investigation of different database sizes, 30,000 datasets was the minimum required for accurate estimation.

Extraction of hand images

Hand contours were defined first, meaning the outermost pixel is given Label No.1, the next pixel internally adjacent Label No.2, etc., until no more pixel locations are found. This is the reference point.

A hand range was then defined and extracted, i.e., the top, left, and right ends of the original hand image obtained above correspond to the top, left, and right ends of the hand's contour. The bottom end of the hand image is lower than the reference point by the distance to such a pixel on the outermost contour closest to the reference point – the distance is defined by the number (N) of pixels – as shown in **Fig. 3**.

Extraction of image characteristics

To characterize hand images, we used higher-order local autocorrelational [12]. The characteristics defined using the following expression were calculated for the reference point and its vicinity:

$$x^N(a_1,a_2,\cdots,a_N) = \int f(r)f(r+a_1)\cdots f(r+a_N)dr, \quad \ldots\ldots\ldots\ldots \quad (1)$$

where x^N is the correlational function near point r in dimension N. Because pixels around the point are important when a recorded image is generally used as a processing object, factor N was limited to the second order. When excluding equivalent terms due to parallel translation, x^N is expressed using 25 types of characteristic quantities, but patterns at $N=0$ and $N=1$ should be normalized because they have a smaller scale than characteristic quantities of patterns at $N=2$. By further multiplying pixel values of the reference point for patterns at $N=1$ and by multiplying the square of the pixel value of the reference point for pattern at $N=0$, we obtained good agreement with other characteristic quantities.

An image was divided into 64 sections of – 8 × 8 each vertically and laterally – divided images were represented by 25 types of characteristic quantities using higher-order local autocorrelational, so a single image is described using characteristic quantities of 25 patterns × 64 divided sections. Image characteristics of the hand and joint angle data of the finger and wrist were paired as a set for preparing the database.

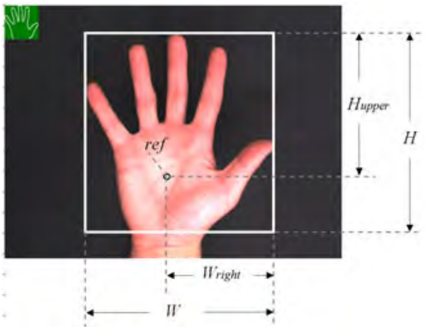


Fig. 4. Three hand proportions.

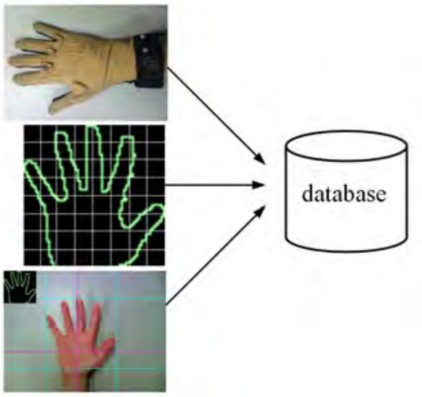


Fig. 5. Information in datasets.

Quick determination of hand poses using proportional information on images

For a hand image as found above, the following three different proportions are calculated as shown in **Fig. 4**:

(1) Tallness: $R_{tall}[i] = H[i]/(H[i]+W[i])$,
(2) Top-heaviness: $R_{topheavy}[i] = H_{upper}[i]/H[i]$,
(3) Right-bias: $R_{rightbiased}[i] = W_{right}[i]/W[i]$.

H is the number of pixels measured vertically within the extraction. W is the number of pixels measured horizontally within the extraction. H_{upper} is the number of pixels above the base point. W_{right} is the number of pixels in the region right of the base point. Suffix i is the dataset number in the database.

These three proportions correspond roughly to forearm rotation, nonthumb finger bending, and thumb bending. This image interpretation by proportional information is used for coarse screening, so each dataset in the database consists of hand and finger characteristics, finger joints angles with forearm rotation, and proportional information, as shown in **Fig. 5**.

2.2. Estimating Hand Poses

Color coordinate system

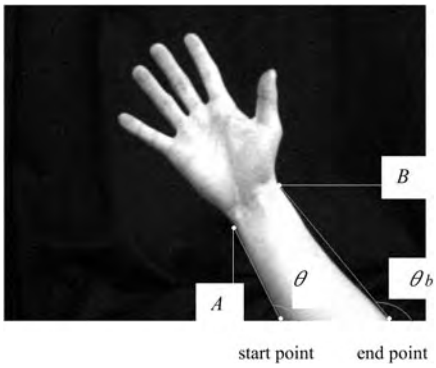


Fig. 6. Detection of forearm inclination.

Each captured image was converted from RGB to HSV color coordinate system as follows:

$$H = \begin{cases} 60 \times (\frac{G-B}{\max-\min}), & if \max = R, \\ 60 \times (2+\frac{B-R}{\max-\min}), & if \max = G, \\ 60 \times (4+\frac{R-G}{\max-\min}), & if \max = B, \end{cases} \qquad (2)$$

$$S = \frac{\max - \min}{\max}, \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (3)$$

where, max=max(R,G,B), min=min(R,G,B), H and S indicate hue and saturation. We do not use color value (V) in HSV color coordinate system.

To extract hand and forearm region, means, and standard deviations of H and S were calculated within several frames. If a pixel satisfies the following equation, the pixel was considered to belong to the hand and forearm region:

$$k\sigma_H < |\mu_H - f_H| \bigcap k\sigma_S < |\mu_S - f_S|, \quad . \; . \; . \quad (4)$$

where f_H, μ_H, σ_H are the value, average, and standard deviation of hue at $f(x,y)$. f_S, μ_S, σ_S are the value, average, and standard deviation of saturation at $f(x,y)$. k is a constant fixed based on the illumination environment.

Forearm inclination detection

User should be able to move their arms freely when the camera is taking pictures. This flexibility, in turn, allows the hand image to be oriented other than upright though images in the database are upright. We must correct image orientation so that it is upright. To do so, we use an algorithm to rotate the hand contour until forearm representation extending from the bottom of the frame of the image rotates to upright, as shown in **Fig. 6**.

Specifically, the forearm and hand contour are extracted from the image, then the pixel (start point) is found from which the contour image begins using horizontal and downward searches starting from the upper left corner of the frame. Whether a pixel is on the contour line is determined by checking the pixel value as luminance. Contour image pixels are tracked sequentially from the start to the end of the contour, calculating the contour line inclination in each pixel. Due to general characteristics of human arm and hand form, inclination is roughly constant throughout the forearm length, i.e., lines from start to point A and from point B to the end are almost straight, while inclination varies along the length of the hand. Standard deviation in the inclination of the contour for each of the specified ranges within the contour image is calculated. The region from the start to where standard deviation exceeds previous pixels and that from where standard deviation becomes smaller than previous pixels to the last point represent the forearm. The average of these two standard deviations is calculated, and the hand contour is rotated by the average angle thus calculated. This hand image is compared to data in the database.

Two-stage search

The first stage in a 2-stage search is coarse screening using proportional hand image information. The second stage is detailed screening for determining the image most similar among candidates selected in the first stage. The second stage uses similarity calculation based on specified image characteristic types.

The first screening uses the three parameters defined by proportional information. If all three parameters fall within the specified threshold, the dataset is chosen as a candidate for the second screening. These three parameters and their thresholds are shown below.

(1) Tallness threshold: $Th_{tall}[i] = |R_{tall} - R_{current-tall}|$,
(2) Top-heaviness threshold:
$$Th_{topheavy}[i] = |R_{topheavy} - R_{current-topheavy}|,$$
(3) Right-biased threshold:
$$Th_{rightbiased}[i] = |R_{rightbiased} - R_{current-rightbiased}|,$$

where R_{tall}, $R_{topheavy}$, and $R_{rightbiased}$ are proportions representing tallness, top-heaviness, and right-biased of the hand image in the dataset under inquiry. $R_{current-tall}$, $R_{current-topheavy}$, and $R_{current-rightbiased}$ are proportions representing tallness, top-heaviness and right-biased of the current input image. Suffix i is the dataset number.

The second screening uses a Euclidean-distance-based similarity search to determine the highest possible image similarity. Dataset joint angles having the shortest distance among candidates chosen represent the result to be determined as the image having the highest possible similarity to the input image.

Threshold determination

Tallness threshold Th_{tall}, top-heaviness threshold $Th_{topheavy}$ and right-biased threshold $Th_{rightbiased}$ are determined as follows: For any of the three thresholds, given a larger value, a search result having an acceptable level of accuracy is obtained via exponential convergence. For a smaller value, the first screening reduces the number of candidates so that search time is reduced. As in determining the optimum operating point from a set of economic demand and supply curves, these thresholds are chosen so that a tradeoff is possible between these two exponential functions. A preliminary experiment for checking estimation error variation as these three thresholds change from 0.001 to 0.061 in steps of 0.01 showed that the best combination of (Th_{tall}, $Th_{topheavy}$, $Th_{rightbiased}$) is (0.011, 0.011 and 0.011) or (0.011, 0.011 and 0.021).

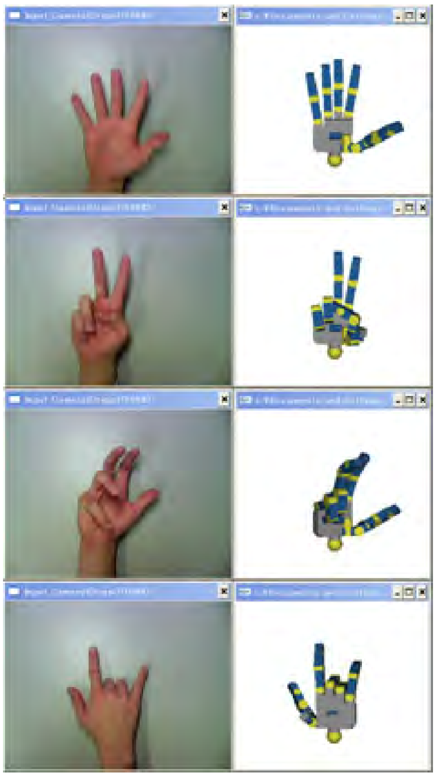


Fig. 7. Captured hand images and results of hand pose estimation.

3. Estimation experiment

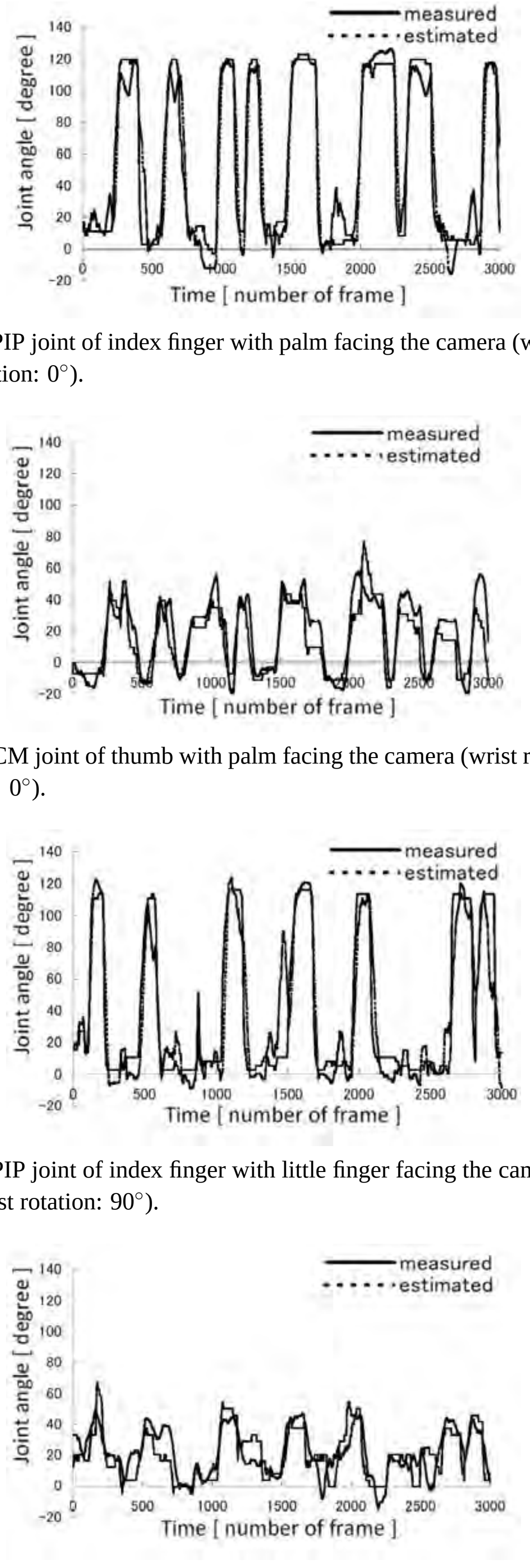
3.1. Procedures

To verify the effectiveness of our approach, actual images were experimentally estimated. A subject held up a hand 1 m in front of the high-speed camera and moved fingers and thumb, wrist, and arm freely. Hand movement was allowed in all directions as long as the hand stayed within camera range.

In the experiments, our database used had stored 30,000 pairs of characteristic quantities and angles for the finger and wrist. In experiments, we used a notebook PC (Dell Precision M4300), a CoreTM 2 Duo Processor T8300 (2.40 GHz, 800 MHz FSB), main memory of 4 GB, a high-speed camera (Dragonfly Express^{TM}, Point Grey Research Inc.) and data gloves (Cyber Glove, Virtual Technologies Inc.).

3.2. Results

Figure 7 shows examples of estimation in snapshots. Estimated results were drawn using the CG hand on the right side. Finger angles may have been estimated with high precision when the hand and fingers were continuously moved, even with wrist rotation. Estimation could be made provided that the hand image did not blend into the background even when the illumination environment changed.



(a) PIP joint of index finger with palm facing the camera (wrist rotation: 0°).

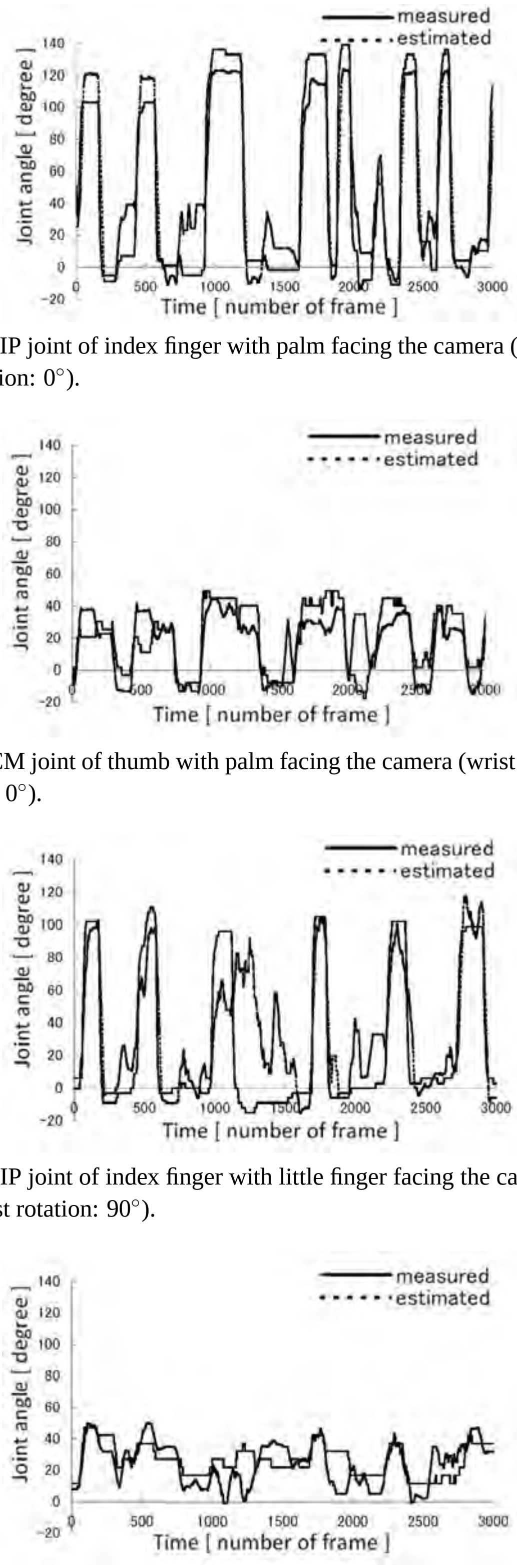
(b) CM joint of thumb with palm facing the camera (wrist rotation: 0°).

(c) PIP joint of index finger with little finger facing the camera (wrist rotation: 90°).

(d) CM joint of thumb with little finger facing the camera (wrist rotation: 90°).

Fig. 8. Examples of estimated results in a subject whose hand information is stored in the database.

For quantitative assessments, measured and estimated values must be compared, but in an ordinary environment using our approach, measured values of joint angle information from the hand and fingers moving in front of the camera cannot be obtained, so we conducted estimation experiments wearing the data glove and a white glove as stated earlier. Results in **Figs. 8** and **9** showing angular data measured using the data glove and estimated results.



(a) PIP joint of index finger with palm facing the camera (wrist rotation: 0°).

(b) CM joint of thumb with palm facing the camera (wrist rotation: 0°).

(c) PIP joint of index finger with little finger facing the camera (wrist rotation: 90°).

(d) CM joint of thumb with little finger facing the camera (wrist rotation: 90°).

Fig. 9. Examples of estimated results in another subject whose hand information is “not” stored in the database.

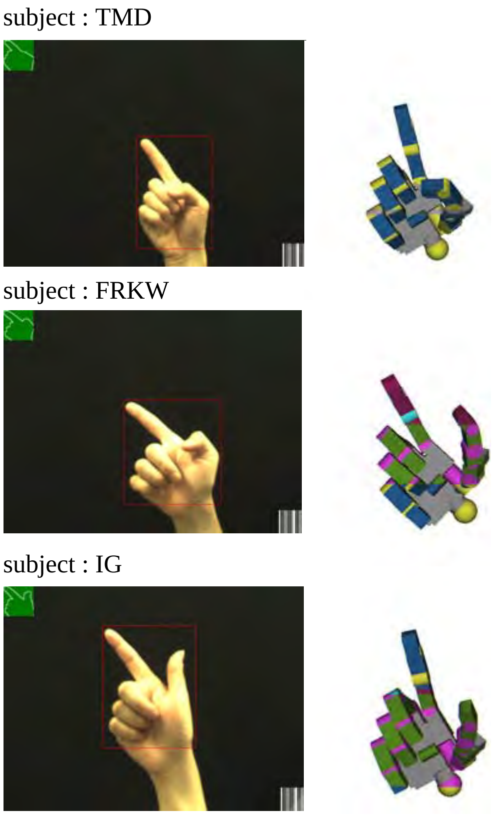


Fig. 10. Examples of individual differences in estimated results to a hand pose.

Subjects were instructed to move their hand and fingers freely in front of a high-speed camera. **Fig. 8** shows results for a subject whose hand information is stored in the database. **Fig. 9** shows results for another subject whose hand information is not stored.

In these figures, (a) and (c), and (b) and (d) show the PIP joint of the index finger and CM joint of the thumb, when the palm is facing the high-speed camera and the little finger is facing the camera respectively. The state with the joint extended was set to 180°. Mean and standard deviation scores of errors in estimated angles were 2.4±14.5 at index PIP and −5.3±44.4 at thumb CM, in **Fig. 8**. Scores in **Fig. 9** were −9.6±27.0 and −11.3±16.5. Standard deviations of errors seem to be bypassed, but mean error is small.

The system operates at 80 fps using a notebook PC with a single high-speed camera and enables real-time estimation. Coarse screening in the first stage, done with proportional information as low-order image features, selected 150 datasets as the average from 30,000 database, and only 0.5% of candidates are the target for precise calculation for similarity in the second stage which requires relatively long processing.

Figure 10 shows snapshots of estimated results for three subjects. Users having different hand size and shape in 3D hand pose estimation are handled. This may lead an interface for unspecified users for information input device or remote control of a robot by “human hand movements” without any sensors attached to the users.

Image and joint angle information are paired in our database. Once we output results for hand pose estimation to a robot hand, the robot reproduces the same movements as those of the fingers mimicked. **Fig. 11** shows a

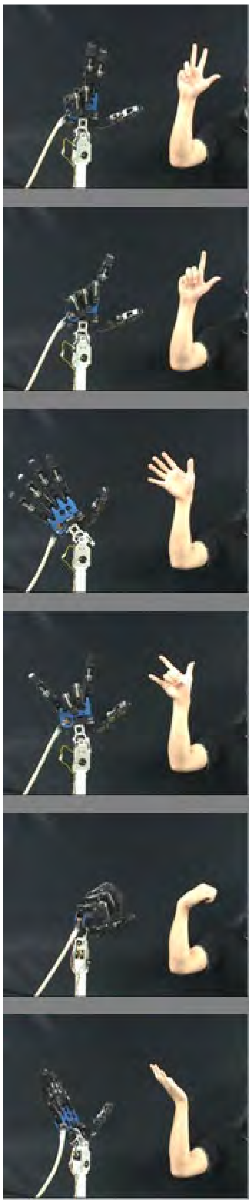


Fig. 11. Dexterous robotic hand imitating human movements without a time delay.

dexterous robot hand without sensors [13] imitating human hand movement. The simple moving average was used within 100 sampling points to avoid vibration of the robotic hand, which imitates the human hand because hand pose estimation calculates at high speed highly accurately.

4. Discussions

The technology proposed here involves finding similar images out of a massive database at high speed. As stated earlier, two types of approach can be used for hand pose estimation. The first – 3D-model-based [1–6] – involves extracting local characteristics, or silhouettes, from an image recorded using a camera and fitting a 3D hand model, constructed in advance on a computer. This approach requires long processing time, however – video rate at best, or some 10 fps. The second, a 2D-appearance-based approach [7–9, 14] involves directly comparing input images to the appearance of the image in the database. This reduces calculation time. If 3D changes in hand appearance are included, a large reference database is needed, making it difficult to control the robot hand without a time delay. The feasibility of this approach thus depends on the search algorithm used for rapidly finding similar images in an extensive database. Previous research based on a 2D-appearance-based approach accomplished 150 fps processing or more [10] while realizing high-accuracy estimations of hand poses almost the same as for the 3D-model-based approach.

Our approach has four advantages:

(1) Even if a larger database is used for higher accuracy and resolution in estimation, an image having the highest possible similarity is obtained at high speed.

(2) Even if the hand pose changes quickly, estimation accuracy remains unchanged so the image having the highest possible similarity is obtained.

(3) Search time immediately before the current time does not affect the search result, enabling the image having the highest possible similarity to be obtained. Our previous system [10] may fail to find the best matching image because it searches based on past estimation results, as stated, but the proposal in this paper has no such problem. Estimation is very stable because it uses no past results.

(4) Our approach requires no time-consuming construction of a multilayer database, and thus no need to consider how many classes and data must be defined or which types of image characteristics must be defined for each class. On the contrary, our approach here uses coarse screening of the proportional information on hand images such as tallness and right-biased corresponding to forearm rotation, thumb bending, and nonthumb finger bending. A detailed search is then made for similarity among selected candidates. To describe forearm rotation, internal and external wrist rotation (abduction/adduction), and bending and stretching (flexion/extension), no separate processes were used for estimating corresponding joint angles. Instead, such hand images already containing these movements were prepared in the database, which could give real, 3D changes in hand appearance.

A tree-based search to estimate the hand pose may have to be discussed for optimal solution searching, e.g., a cascade of classifiers was arranged in a tree to recognize multiple object classes [15]. Classifiers were obtained from a geometric 3D model or from training images. Although their performance is not as good as classifiers learned from image data, they have the advantage of being easy to generate and labeled with a known 3D pose. In research using a tree-based filter [16], alternative tree construction, including regular partitioning of the eigenspace and vector quantization, were considered and hand dynamics were captured by keeping a histogram of tree node transitions in training data. Although single camera pose estimation with unconstrained hand movement is ambiguous due to occlusion, a tree-based search may be able to initialize tracking without imposing too many constraints on the user [17]. A tree-based search may have the potential to provide more natural and effective solutions.

In our early study [8], for example, a lookup table was used whose dimensions were lowered by principal component analysis. It was not easy, however, to determine orders of principal components to be considered and processing time could not be effectively shortened. Concerning binary tree indices such as KD-tree or Quad-tree advanced, it was hard to determine the height of trees and the number of divisions. Our approach is effective where coarse screening is done with proportional information on hand images such as tallness and right-biased roughly corresponding to forearm rotation, thumb bending or finger bending, and then a detailed search is done for similarity among selected candidates. Candidates are limited to fewer than 0.5% by coarse screening.

The main features of our approach are as follows:

Our approach is unique in normalizing the hand image and that in the database to a fixed size, e.g., 64×64 pixels, by dividing the original image by the maximum width (top, left, and right ends) of the hand view or the distance (to the lower end) measured from the reference point determined by labeling where by the distance from the camera to the object (hand) need not be considered in images. User can move their arms and hands freely as usual to give a specified type of gesture for the robot or computer.

Note that data in the database consists of hand images only. Image taken by the camera are of a hand and part of the arm. To enable accurate estimation through comparison of lower-order image features such as a higher-order autocorrelational function, it is necessary to extract a “hand part” from the hand and arm image taken by the camera. This may be done “intuitively” by considering the constricted part the wrist and the part extended from the wrist as the hand. This may not, however, be effective in all cases. To estimate a hand pose rotated by the forearm upright as when the hand has been rotated so that the palm faces parallel to the camera view, with fingers stretched, the constricted part (horizontal location of the axis perpendicular to the forearm) may not exist at the same height or may be at other than the actual wrist position.

The system should, rather than precisely recognizing and extracting the hand from the image taken, is to extract the hand from the image and simultaneously extract a similar hand image from the database so appropriate datasets can be chosen by precise comparison of similar images using lower-order image features. The main purpose of extracting the hand is to find image features available for coarse screening. This is, however also effective for extracting a hand having the same pose, irrespective of the wrist, and can be done very quickly. This is a prerequisite for our high-accuracy 3D hand pose estimation here.

As shown in **Fig. 11**, our approach using one notebook PC estimates 3D hand poses of a person freely moving their arm in front of a high-speed camera, whose result is output to a humanoid robot to let it imitate hand movement. Our pose estimation is so accurate and quick that the robot can operate stably. This is, to our knowledge, the only system that can currently provide such capability.

The second feature of our approach is a dataset containing information on 3D wrist bending, stretching, and rotation included in the database. Because of this, even hand postures accompanied by 3D wrist movement are quickly estimated without needing additional algorithms.

For data entry or robot control using 3D hand poses, the following conditions must be met:

(1) Hand pose estimation must be sufficiently accurate with a joint angle estimation error of a maximum of 4 to 5°.

(2) Processing speed must be sufficiently high – at least 100 fps.

(3) All users must be processed, regardless of different hand size and shape.

Our approach meets these three conditions. Other conditions that should be additionally satisfied include:

(4) Relatively fast hand movement such as for sign language – possibly representing a user’s natural movements – must be accepted.

(5) Both hands must be used, if possible.

Using a data glove may currently be the only way of obtaining quantitative data for quantitative hand poses evaluation. The concern is that the data glove tends to have the same shape and size, regardless of who uses it, and this would introduce a bias in similarity of input and database images. The major factor possibly preventing accurate, stable estimation, however, is due to longer or shorter fingers in relation to palm length. As is seen from snapshots in **Fig. 10**, the first hand has long, thin fingers, while the second has short, fat fingers. Our system copes well with both types of subjects.

5. Conclusion

A 2D-appearance-based approach, which consists of directly comparing input images to images stored in a database, reduces calculation time, but if 3D changes in hand appearance are not an issue, which includes wrist and forearm movement, a large reference database is required, and it becomes difficult to control the robot hand using imitation. The 2D-appearance-based approach using a multilayer database is effective but has disadvantages such as candidate images exiting the search range, leading to a failure to find the best matching image, where the hand pose is changing quickly. The second disadvantage is that even if the hand pose changes slowly, once a dissimilar image is found, the search range for the next time cannot be changed from where the image was found. Third, such a database as created into a multilayer structure using a priori facts instead of statistical information on image similarity has difficulty in determining appropriate types of image characteristics, including joint angle and fingernail location for each of the layers as defined by representative values.

We have therefore proposed a 3D hand pose estimation using a single high-speed camera that retrieves the hand image most similar to data input from a large nonmultilayer database as a 2D-appearance-based approach.

By estimating sequential images of finger shape in this way, we realized an approach involving joint angle estimation error within several degrees even with the wrist freely rotating, and processing time of 80 fps using a notebook PC having a CoreTM 2 Duo Processor T8300 (2.40 GHz, 800 MHz FSB) and main memory of 4 GB. Our approach worked equally well for subjects regardless of different hand sizes and shapes.

Because image information and joint angle information are paired in the database and the wrist, our approach reproduced the same actions as those of the fingers and wrist of a person using a robot with a slight time delay outputting estimation results to the robot hand, using only one notebook PC.

References:

[1] 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.

[2] M. H. Jeong, Y. Kuno, N. Shimada, and Y. Shirai, "Recognition of Shape-Changing Hand Gestures, IEICE Trans. on Information and Systems," E85-D, pp. 1678-1687, 2002.

[3] S. Lu, D. Metaxas, D. Samaras, and J. Oliensis, "Using Multiple Cues for Hand Tracking and Model Refinement." Proc. CVPR2003, 2, pp. 443-450, 2003.

[4] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara, "Hand Pose Estimation for Vision-Based Human Interface," IEEE Trans. on Industrial Electronics, 50, 4, pp. 676-684, 2003.

[5] M. H. Jeong, Y. Kuno, N. Shimada, and Y. Shirai, "Recognition of Two-Hand Gestures Using Coupled Switching Linear Model," IEICE Trans. on Information and Systems, E86-D, pp. 1416-1425, 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, 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, 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, 21, pp. 1743-1761, 2007.

[11] R. Sano, M. Tomida, and K. Hoshino, "3D Hand Posture Estimation Using Relative Positions of Fingernails as First-Stage Screening," ITE technical report, 33, pp. 21-24, 2009 (in Japanese).

[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.

[13] K. Hoshino and I. Kawabuchi, "Pinching at Finger Tips for Humanoid Robot Hand," J. of Robotics and Mechatronics, 17, pp. 655-663, 2005.

[14] K. Hoshino and T. Tanimoto, "Realtime Hand Posture Estimation with Self-Organizing Map for Stable Robot Control," IEICE Trans. on Information and Systems, E89-D, pp. 1813-1819, 2006.

[15] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla, "Hand Pose Estimation Using Hierarchical Detection," Lecture Notes in Computer Science, 3058, pp. 105-116, 2004.

[16] A. Thayananthan, B. Stenger, P. H. S. Torr, and R. Cipolla, "Learning a Kinematic Prior for Tree-Based Filtering," British Machine Vision Conference, 2, pp. 589-598, 2003.

[17] C. Tomasi, S. Petrov, and A. Sastry, "3D Tracking = Classification + Interpolation," Ninth IEEE Int. Conf. on Computer Vision, 2, pp. 1441-1448, 2003.

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 at Tokyo Medical and Dental University
1995- Associate Professor at University of the Ryukyus
2002- Associate Professor at University of Tsukuba
2008- Associate Professor at 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:

- "Interpolation and extrapolation of repeated motions obtained with magnetic motion capture," IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, E87-A, 9, pp. 2401-2407, 2004.

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)

Name:
Motomasa Tomida

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

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

Brief Biographical History:

2008- B.E. degree from University of Tsukuba
2008- Master Candidate at University of Tsukuba