

Paper:

Gesture-World Environment Technology for Mobile Manipulation – Remote Control System of a Robot with Hand Pose Estimation –

Kiyoshi Hoshino*, Takuya Kasahara*, Motomasa Tomida**, and Takanobu Tanimoto*

*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

**Crescent, Inc.

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

[Received May 7, 2011; accepted August 31, 2011]

The purpose of this paper is to propose a remote-controlled robot system capable of accurate high-speed performance of the same operation strictly conforming to human operator movement without sensors or special control means. We specifically intend to implement high-precision high-speed 3D hand pose estimation enabling a remote-controlled robot to be operated using two cameras installed loosely orthogonally using one ordinary PC. The two cameras have their own database. Once sequential hand images are shot at high speed, the system starts selecting one database with bigger size of hand region in each recorded image. Coarse screening then proceeds based on proportional hand image information roughly corresponding to wrist rotation or thumb or finger extension. Finally, a detailed search is done for similarity among selected candidates. Experiments show that mean and standard deviation scores of errors in estimated angles at the proximal interphalangeal (PIP) index are 0.45 ± 14.57 and at the carpometacarpal (CM) thumb 4.7 ± 10.82 , respectively, indicating it as a high-precision 3D hand pose estimation. Remote control of a robot with the proposed vision system shows high performance as well.

Keywords: 3D hand pose estimation, two cameras installed at position of loosely orthogonal relationship, 3D shape reconstruction of a hand from a 2D image, remote control of a robot

1. Introduction

Robot research and development projects have not yet succeeded in incorporating a high level of intelligence in a robot. When an object having various poses, weights and centers of gravity is located in front, for example, it remains difficult to ensure that the robot hand holds the object in conformance to individual object features so that

the object can be manipulated. The level of intelligence built into a robot is currently that of a six-year-old child, at best. With countries such as Japan facing a declining birthrate and an aging population, robots are expected to be required to have an advanced level of intelligence especially in the fields of logistics and elder care meeting the needs and requirements of senior citizens.

A paradigm shift in thinking is needed. To be more specific, it is not easy to incorporate an advanced level of intelligence in a robot in such a way that the robot will take care of the work of assortment. Assume, for example, that a human operator in a room different from that of the assortment worksite monitors the area to confirm that items to be sorted travel as designated on a belt conveyor. In response to what is the operator's movements, a robot at the remote locations imitates the operator's movement. This would enable comparatively complex sorting without requiring that an advanced level of intelligence be built into the robot. This requires only that daily human action be done through a monitor.

Hand tracking is not the robot vision technology required in this case. What is needed is "hand pose estimation." Specifically, hand tracking in which images of hand movement direction and distance are analyzed and assigned to robot functions and information communication equipment. This is comparable to cases in which, for example, if the operator gestures "scissors" in a rock, paper, scissors game, the robot is made to do operation A. If the operator gestures "paper," the robot is made to do operation B. Hand tracking is enabled in a pointing device where hand direction and distance are detected and used to do the required work. The robot is not manipulated by daily operator action. Instead, in the technique of hand pose estimation, the "pose or posture of the hand" are associated with dynamic robot behavior. In hand pose estimation, the same movement as that of the operator is reconfigured by the robot. This does not require that the user to learn a specific action in advance to ensure that the robot does it. Once the user conducts a daily action, the robot will do the same.

Two approaches are used to roughly classify conventional hand pose estimation – 3-dimensional (3D) model-based and 2-dimensional (2D) appearance-based action. The 3D model-based approach [1–6] involves extracting local characteristics or silhouettes from images recorded by a camera and fitting a 3D hand model constructed beforehand on a computer. This approach estimates hand shapes highly accurately, but it processes self-occlusion poorly and requires long processing time. The 2D-appearance-based approach [7–9] involves directly comparing an input image to an image stored in a database, which cuts calculation time. If 3D changes in hand appearance – e.g., wrist and forearm movement – are involved, however, this approach requires a large reference database, and robot hand movement is difficult to be controlled 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 of all possible hand images, including complexity and self-occlusion. The feasibility of this approach thus depends on the search algorithm.

In 2D appearance-based approach, Hoshino et al. [8] proposed using computer graphics (CG) editing software and data gloves to create a large database containing personal hand pose attributes such as movable joint range and bone length. They developed a search algorithm that shortens search time in looking for unknown input images by using a multi-layer database based on a self-organization map accompanying self-multiplication and self-extinction so that similar hand images are brought closer and the search area is reduced to only that data near the search result during previous search time is inquired about [10].

In hand pose estimation using one camera, self-occlusion is fatal to manipulating an object by a remote-controlled robot. Assume, for instance, that an object captured by the camera from the back of the hand has almost the same the silhouette. This may involve at least two types of postures, such as power grasping and precision pinching. If the positional relationship between the finger and object to be grasped or pinched is inaccurate, the robot hand will easily lose the object. When an application example of hand pose estimation is considered, however, it is unrealistic to use a multiple-camera system to capture an object by surrounding it. If possible, requirements should be met by installing two cameras positioned loosely orthogonally, without camera installation being specifically or precisely positioned.

Given the above background, we propose a remote-controlled robot system capable of accurate high-speed performance of the same operation strictly conforming to human operator movement, but without sensors or special controllers. We are particularly interested in introducing a way to implement high-precision high-speed 3D hand pose estimation enabling real-time operation by a remote-controlled robot using two cameras, positioned loosely orthogonally, together with an ordinary PC.

2. System Configuration

2.1. Data Sets

Our previous system database was constructed using a single hand model, i.e., the operator's hand [11, 12]. The database stored individual hand images paired with finger and wrist angles synchronously acquired from a data glove and camera. Images were recorded using a camera with a resolution of 320×240 pixels, laterally and vertically viewing hands and fingers on an appropriately sized screen. Fingers and wrist angles were acquired using a data glove (Cyber Glove, Virtual Technologies Inc.) that simultaneously obtained 18 types of angle information on the hand.

The database must contain every possible hand pose for a hand model, without exception. Here, we therefore provide a system with two types of hand model pose patterns – called basic and additional – generated using 3D computer graphics [8] (Poser 5, Curious Labs). Basic pose patterns are created to cover all hand poses. We independently captured images on bending and extending the index, middle, ring, and little fingers in turn, the degree to which fingers spread or close toward one another in five stages, thumb motions with six stages, and wrist motion and forearm rotation with seven stages. We saved data sets combining these poses in the database. Individual stages were decided based on dynamic range and joint DOF (degree of freedom) number. For wrist motions, we only moved the wrist within the same plane, relative to the camera, for each rotation of the forearm.

We used additional pose patterns to add data sets for poses when the palm or back of the hand faced the camera. Whereas we had treated how much the fingers spread mutually as one degree of freedom, fingers are actually all capable of moving independently toward or away from each other, so appearance when the palm or back of the hand is facing the camera differs greatly. We added further hand pose data combining basic pose patterns for thumb and wrist motions with new patterns for finger bending and extension and how much fingers spread. In other words, hand CGs with various poses are systematically generated through the former “basic pose” procedure, and hand CGs with individual differences are generated through the latter “additional pose” procedure. **Fig. 1** shows examples of additional bending/extending and spreading of fingers. The resulting database contained 772,576 data sets from collecting large-scale data sets.

2.2. Calculation of Proportional Information on Hand Images

We first defined hand contours. Specifically, the outermost pixel becomes Labeling No.1 and the pixel internally adjacent to the outermost pixel Labeling No.2. Repeating this labeling yields the pixel position becoming the largest labeling found, i.e., the reference point. A hand range is defined and cut out. On the original image from the previous paragraph, the top, left and right ends of the hand image correspond to the top, left and right ends of