Detection of hand grasping an object from complex background based on machine learning co-occurrence of local image feature

Shinya Morioka*, Yasuhiro Hiramoto*, Nobutaka Shimada*, Tadashi Matsuo*, Yoshiaki Shirai*
*Ritsumeikan University, Shiga, Japan

***Abstract*— This paper proposes a method of detecting a hand grasping an object from one image based on machine learning. The learning process progresses as a boostrap scheme from hand-only detection to object-hand relationship modeling. In the first step, Randomized Trees that learns the positional and visual relationship between image features of hand region and wrist position and that provides the probability distribution of wrist positions is built. Then the wrist position is detected by voting the distributions for multiple features in newly captured images. In the second step, the suitability as hand is calculated for each feature by evaluating the voting result. Based on that, the object features specified by the grasping type are extracted and the positional and visual relationship between the object and hand are learned so that the object center can be predicted from hand shape. On the local coordinate constructed by the predicted object center and the wrist position, it is learned where the object, hand region and the background is located in the target grasping type by One class SVM. Finally, the grasping hand, the grasped object and the background are respectively extracted from the cluttered background.**

I. INTRODUCTION

There are researches of detecting unknown objects from image based on machine learning like [1]. In addition, another approach is proposed based on more image cues like human actions and the kind of objects. In order to recognize various actions of the human hands, such as holding or picking up something, detecting the hands and estimating their postures from images are to be solved. In some previous researches the object category is recognized by using functional relationship between an object and the hand action [2][3]. They simultaneously evaluate the image appearances of hand and object. However, when a hand is holding something, it is difficult to detect the hand and precisely classify its posture because some part of regions may be occluded by the grasped object. This paper proposes a method that detects a hand position holding an object based on local hand features in spite of the occlusion. By learning local features from the hand grasping various tools with various way to hold them, the proposed method estimates the position, scale, direction of the hand and the region of the grasped object. By some experiments, we demonstrate that the proposed method can detect them in cluttered backgrounds.

II. WRIST POSITION DETECTION BASED ON RANDOMIZED TREES LEARNING RELATIVE POSITIONS OF THE LOCAL IMAGE FEATURES

A. Training of Randomized Trees model providing the probability distribution of wrist positions

Randomized Trees (RT)[4] is a multi-class classifier that accepts multi-dimensional features and it provides probability distributions over the class indices. Here we construct a classifier that provides the probability distribution of wrist positions when SURF[5] image features are input into the classifier. It is necessary to estimate the distribution even if the hand is located at the image region, in the scale and orientation different from those of the training images. For the purpose, the training data of the wrist positions are once normalized based on the position, scale and orientation of each SURF feature of the grasping hands seen in the training images, and then the pairs of the normalized wrist positions and the features are input into the classifier for training. Since the probability distribution predicted by RT is discrete, however, the continuous distribution over the image space cannot be directly treat by RT. Therefore based on the normalized coordinate, the normalized image space is divided into discrete grid blocks each of that has a unique class label, the class label on which the wrist position lies is determined, and then an RT that outputs the probability distribution of the wrist position over the grids for the input of 128 dimensional SURF image feature is constructed.

A local coordinate space defined for each SURF vector is segmented into some blocks by a grid (Fig.1-(c)). The number of blocks is finite and they are indexed. A label j is an index that means a block or background.

The j-th block C_j is a region on a local coordinate space defined as

$$C_j = \left\{ \left[\begin{array}{c} u \\ v \end{array} \right] \middle| |u - u_c(j)| < S_{block}/2, |v - v_c(j)| < S_{block}/2 \right\}, \tag{1}$$

where $(u_c(j), v_c(j))^t$ denotes the center of the C_j and S_{block} denotes the size of a block.

When a wrist position is $\mathbf{x}_{wrist} = (x_{wrist}, y_{wrist})^t$ and a SURF vector v is detected on a hand, a label j of the SURF vector v is defined as a block index j such that $T_v(\mathbf{x}_{wrist}) \in C_j$, where T_v is defined the translation into the local coordinate

as

$$T_v(x,y) = \frac{1}{S_v}\begin{bmatrix} (x-x_v)\cos\theta_v + (y-y_v)\sin\theta \\ -(x-x_v)\sin\theta_v + (y-y_v)\cos\theta \end{bmatrix}, \quad (2)$$

and (x_v, y_v) denotes the position where the SURF v is detected, θ_v and S_v denote the direction and scale of the SURF v, respectively.

A label of a SURF vector detected on background is defined as a special index j_{back} and it is distinguished from block indexes.

To learn the relation between a SURF vector v and its label j, we collect many pairs of a SURF vector and its label from many teacher images where the true wrist position is known. Then, we train RT with the pairs so that we can calculate a probability distribution $P(j|v)$.

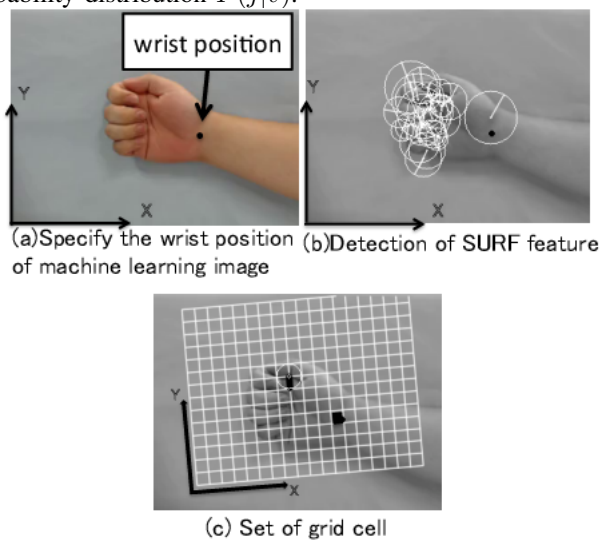

Fig. 1. Conversion of wrist position and positions of features to relative positional relationship class labels

B. Wrist position detection based on votes from the probability distribution of wrist positions

A wrist position is estimated by "voting" on the image space, based on the probability distribution $P(j|v)$ learned with the RT. The votes function $V_{wrist}(x,y)$ defined as

$$V_{wrist}(x,y) = \sum_{v \text{ for all SURF vectors}} V_v(x,y), \quad (3)$$

where

$$V_v(x,y) = \begin{cases} P(j=\tilde{j}|v) & \text{if } {}^{\exists}C_{\tilde{j}} \text{ s.t. } T_v(x,y) \in C_{\tilde{j}}, \\ 0 & (\text{otherwise}), \end{cases} \quad (4)$$

and $P(j|v)$ is a probability distribution calculated by the trained RT.

The position with the maximum votes, $\arg\max_{x,y} V_{wrist}(x,y)$, is considered as a position suitable for a wrist. However, the global maximum may not be the true position because the votes mean a local likelihood and global comparison of them makes no sense. Therefore, we allow multiple candidates of wrist positions that have locally maximum votes.

To find local maxima without explicitly defining local regions, we use mean-shift clustering. Seeds for the clustering are placed at regular intervals in the image space as shown in Fig.2(a), where a blue point denotes a seed. a green point denotes a trajectory, a red point denotes a limit of convergence and a green and blue circle denotes the first and second wrist candidate, respectively. The image space is clustered by limit positions of convergence (Fig.2(b)). For each cluster, the position with the maximum votes is taken as a candidate (Fig.2(c)).

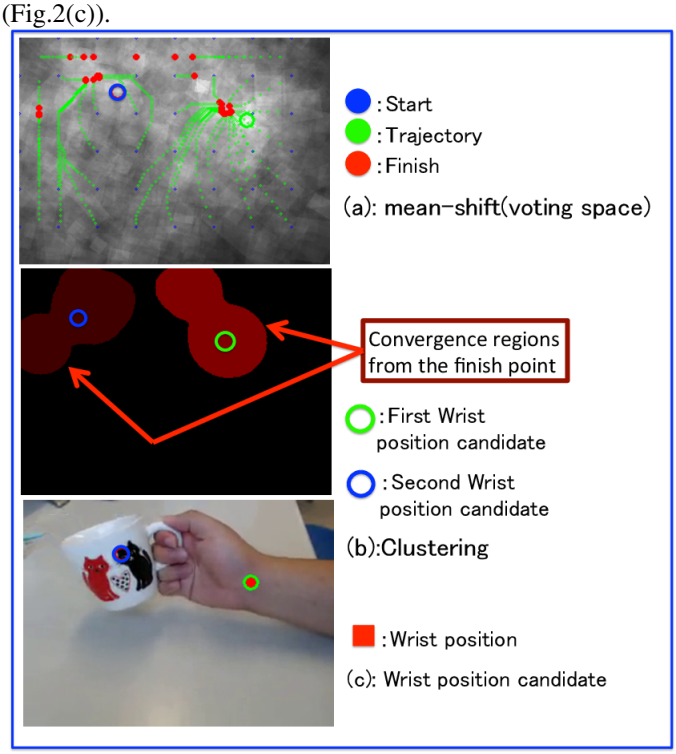

Fig. 2. Flow of wrist candidate detection

If a background region is larger than a hand region, the above voting is disturbed by SURFs on the background. To overcome this, we roughly distinguish a SURF seemingly originated from a hand and others by a SVM[6]. We perform the above voting algorithm with SURFs except those apparently originated from non-hand regions.

An example of finding candidate is shown in Fig. 3. First, SURFs are extracted from an input image(Fig. 3(a)). By using the SVM, they are roughly classified and SURFs apparently originated from non-hand regions are excluded (Fig. 3(b),(c)). For each of the rest SURF vectors v, as the above mentioned, the local coordinate is defined and the conditional probability distribution $P(j|v)$ is predicted by the RT classifier (Fig. 3(d),(e)). By voting the probability distribution, candidates of wrist positions are determined (Fig. 3(f),(g)). The red rectangle in the image Fig. 3(g) is the detected wrist position, and it is shown that the wrist is successfully detected.

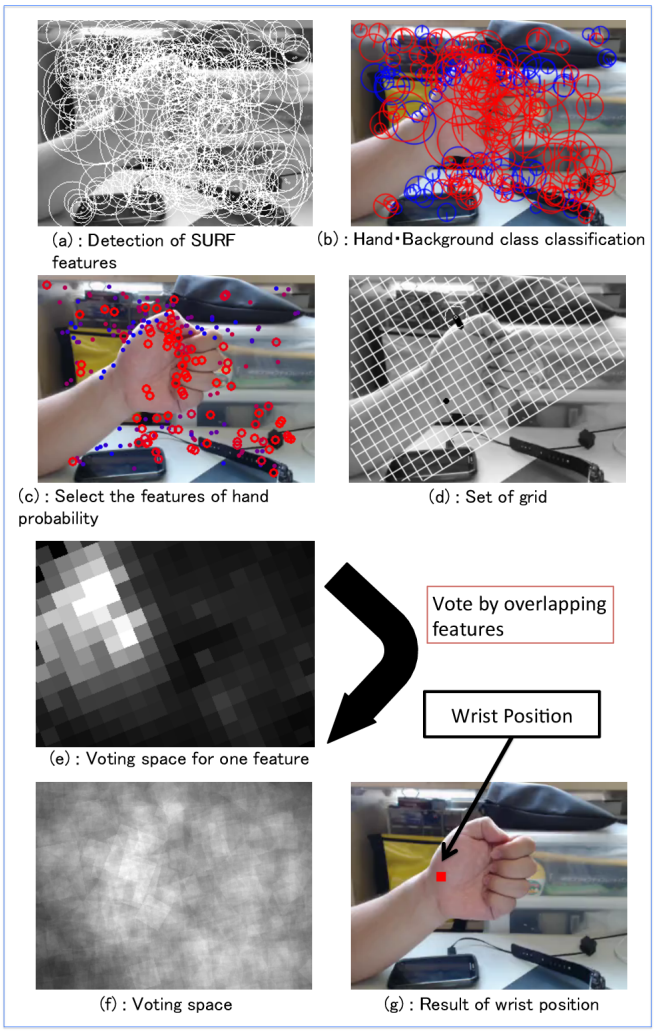

Fig. 3. Flow of based on votes from the probability distribution of wrist positions

III. Extraction of Hand and Object Region Based on Learning of Co-occurence of Hand and Object Features

To find a wrist position from features observed on a hand grasping an object, it is required to classify them into features on a hand and those on an object. We introduce a One Class SVM[7] for the classification. It is trained with a object relative position represented by a wrist-object coordinate system. First, we estimate the center of gravity of an object by RT and derive a wrist-object coordinate system from the estimated center. Then, we distinguish between feature points on an object and those on a hand by the One Class SVM.

A. Estimating gravity center of a hand by coarse classification

A wrist position can be found by the algorithm described the section 2. We can derive a wrist-object coordinate system by finding a center of gravity of an object. Here, we are looking at images showing object grasping states, and we need to classify these into the image features that belong to the hand and those that belong to the object. Thus, suitability as hand h represents how much the detected SURF features contributed to the voting for that position when the wrist position was being determined (i.e. the probability of votes) (Fig.4-(b)). The position (x, y) within the image is classified using h. The method of calculating suitability as hand h is to take the index of SURF features voted for v_{max} as i, and the values voted for by feature i in that place as P^i_{wrist}, and express this as in equation(5).

$$v_{max} = \sum_i P^i_{wrist} \tag{5}$$

The higher the value of P^i_{wrist} is greater, feature i has strongly contributed to determining the wrist position. In other words, the suitability as hand for each feature is shown in equation(6).

$$h_i = P^i_{wrist} \tag{6}$$

K-means clustering is used to segment the positions of features in the image and the groups representing their suitability as hand $x = (x, y, h)$ to determine whether they are hand class features or object class features. The h value of center $\mathbf{x}_{m1}$ and $\mathbf{x}_{m2}$ for each cluster is compared, with the larger one being judged to be the hand class L_{hand} and the other one to be the object class L_{obj}.

$$\begin{cases} h_{m1} > h_{m2} \rightarrow \begin{cases} m1 \rightarrow m_{L_{hand}} \\ m2 \rightarrow m_{L_{obj}} \end{cases} \\ otherwise \rightarrow \begin{cases} m1 \rightarrow m_{L_{obj}} \\ m2 \rightarrow m_{L_{hand}} \end{cases} \end{cases} \tag{7}$$

From this, we can use the key point x_i, and the Euclidean distance $d(x_i, x_m)$ of the center of gravity vector x_m to determine the class in which each key point belongs. The hand class L_{hand} and object class L_{obj} are calculated from equation(8).

$$\begin{cases} d(\mathbf{x}_i, \mathbf{x}_{\mathbf{m}_{\mathbf{L}_{\mathbf{hand}}}}) < d(\mathbf{x}_i, \mathbf{x}_{\mathbf{m}_{\mathbf{L}_{\mathbf{obj}}}}) \rightarrow i \in L_{hand} \\ otherwise \rightarrow i \in L_{obj} \end{cases} \tag{8}$$

This will generate the centers of gravity of the wrist position and object class features as well as the wrist-object coordinate system (Fig.4-(c)). Fig.4 shows the hand region based on the effective scope of SURF features classified into the hand class (dark color indicates greater suitability as hand).(Pink rectangle : hand feature class, Blue rectangle : object feature class)

B. Learning of an RT to output the probability distribution of object positions

Using the process in the previous section, it is possible to segment the respective areas for hand and object in the object grasping sample image for training use. By using the segmented image grasping the object for training purposes, it is possible to newly train it on the relationship between the wrist position and the center position of the object in these kinds of clasping patterns.

Thus, as in Section 2.A, RT is used to teach the center of gravity of objects in the sample training image of a grasping

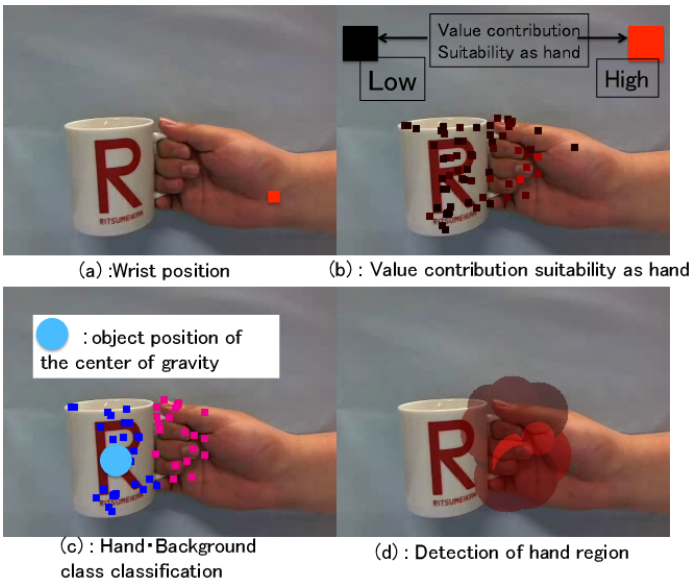

Fig. 4. Flow of Classification of hand and object features

hand extracted in the previous section and the relationship of hand SURF features in the same image and a system is built to predict the object's location relative to local features. After detecting the wrist position, the probability distribution output by the device for predicting the object's location is voted into the image space relative to SURF features that have high suitability as hand (h) and the object's center of gravity is detected. Fig.5 shows an example result of Detection.(Red rectangle : wrist position. Blue rectangle : center of gravity of objects)

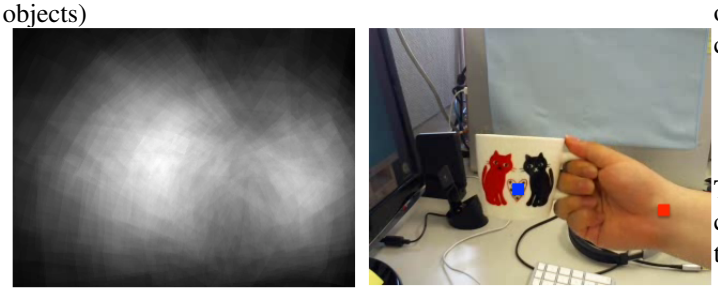

Fig. 5. Detection of center of gravity of objects.

C. Lerning of an One Class SVM to identify hand and object features based on the wrist-object coordinate system

By obtaining the classifiers in the previous section, it becomes possible, by inputting the hand grasping state, to predict what position within the image the center of the object is likely to appear. In this way, as it is possible to set the detected wrist position and wrist-object coordinate system estimated from the predicted central point of the object as previously explained, it is possible to train it from these standard coordinates where, relatively, the clasped object area will appear and make it possible to segment the object area and the background area.

In the wrist-object coordinate system defined in Section 3.A, in order to estimate the range of hand and object regions in the image, both a hand feature One Class SVM and object feature One Class SVM are built. One Class SVM is a classifier that learns using only positive instances to classify items into two classes. For hand features, it uses the input position and likeness to a hand (x, y, h) in the wrist-object coordinate system to learn classifiers. The conversion of image coordinates $\mathbf{x}^{'} = (x.y.h)$ to wrist-object coordinates $\mathbf{x}^{'}$ is expressed in equation(9).

$$\mathbf{x}^{'} = (x^{'}, y^{'}, h) \tag{9}$$

Specifically, the classifier $f_{hand}(\mathbf{x}^{'})$ of hand class feature becomes equation (10).

$$\begin{cases} f_{hand}(\mathbf{x}^{'}) > 0 \rightarrow \mathbf{x}^{'} \in L_{hand} \\ otherwise \rightarrow \mathbf{x}^{'} \notin L_{hand} \end{cases} \tag{10}$$

For object features, because the "suitability as object class" that is common to the appearance of all items that are normally grasped cannot be evaluated, it learns classifiers with only the two-dimensional position (x, y) in the wrist-object coordinate system of the image.
The object class feature classifier $f_{obj}(\mathbf{y}^{'})$, which uses $\mathbf{y}^{'} = (x', y')$ as the wrist-object coordinates converted from the image coordinates, is as expressed in equation(11).

$$\begin{cases} f_{cobj}(\mathbf{y}) > 0 \rightarrow \mathbf{y} \in L_{cobj} \\ otherwise \rightarrow \mathbf{y} \notin L_{cobj} \end{cases} \tag{11}$$

$\mathbf{x}^{'}$ and $\mathbf{y}^{'}$ are calculated from the input image. As countless objects exist each time there is the shape of a hand clasping an object, $f_{hand}(\mathbf{x}^{'})$ is trusted more than $f_{obj}(\mathbf{y}^{'})$. Thus, object class candidate recognition is expressed as in equation(12).

$$\begin{cases} f_{cobj}(\mathbf{y}) > 0 \cap f_{hand}(\mathbf{x}^{'}) \leq 0 \rightarrow \mathbf{y} \in L_{cobj} \\ otherwise \rightarrow \mathbf{y} \notin L_{cobj} \end{cases} \tag{12}$$

The classifier function $f_l(\mathbf{x}^{'}y^{'})$ for hand class L_{hand}, object candidate class L_{obj} and background class L_{back} from equation(10,11,12), is expressed in equation(13).

$$f_l(\mathbf{x}^{'}, \mathbf{y}^{'}) = \begin{cases} L_{hand} & f_{hand}(\mathbf{x}^{'}) > 0 \\ L_{cobj} & f_{cobj}(\mathbf{y}) > 0 \cap f_{hand}(\mathbf{x}^{'}) \leq 0 \\ L_{back} & otherwise \end{cases} \tag{13}$$

Fig.6 shows an example result of classification into hand, object, and background classes (features that are neither hand nor object) using these two classifiers.
(Pink rectangle : Hand feature , Blue rectangle : object feature , Green rectangle : background feature , Red rectangle : Wrist position)

IV. Conclusion

We constructed the RT that output the probability distribution of the wrist positions with the wrist having undergone the training in its positions as related to the SURF features of the hand class. Then we searched the wrist positions this complex

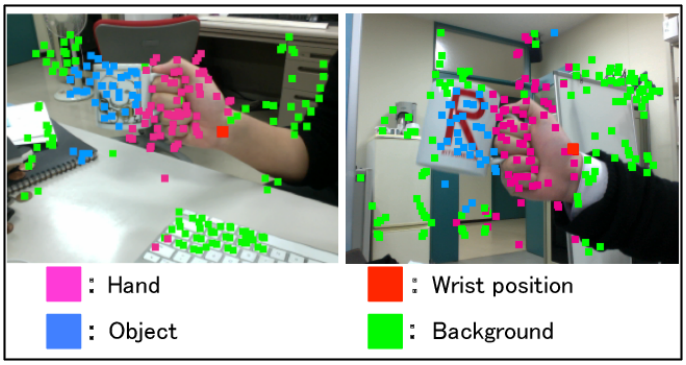

Fig. 6. Result of feature class classification

background. We calculated the suitability as hand of SURF features from the results thus detected on the wrist positions and classified the features into hand class and object class. From the results of our classification, we constructed the RT that output the wrist-object position probability distribution with the wrist having been trained in its relationship with the location of the center of gravity.Then we prepared the wrist-object coordinate system. To enable us to infer the hand and object area in the wrist-object coordinate system thus prepared, we constructed the One Class SVM of the hand features and the One Class SVM of the object features and proceeded to carry out the segmentation of the hand, object, background area under complex background from the two classifiers which we had constructed. In the future, how various objects change the grip of the hand will be studied in order to extract objects based on that relationship.

REFERENCES

[1] S. Kawabata, S. Hiura, and K.Sato, "A Rapid Anomalous Region Extraction Method by Iterative Projection onto Kernel Eigen Space," *The Institute of Electronics, Information and Communication Engineers* , D,J91-D(11), pp. 2673-2683, November, 2008

[2] N. Kamakura, *Shape of hand and Hand motion*. Ishiyaku Publishers, 1989.

[3] H.Kasahara, J.Matsuzaki, N.Shimada, H.Tanaka "Object Recognition by Observing Grasping Scene from Image Sequence," *Meeting on Image Recognition and Understanding 2008*, IS2-5, pp.623-628,2008.

[4] V. Lepetit and P. Fua: "Keypoint recognition using randomized trees", Transactions on Pattern Analysis and Machine Intelliggence, 28,9,pp.1465-1479

[5] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool: "SURF:Sppeded Up Robust Features", Computer Vision and Image Understanding, Vol.110,No.3,pp346-359(2008)

[6] V.Vapnik: "The Nature of Statistical Learning Theory", Springer, (1995).

[7] B.Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson: "Estimating the support of a high-dimensional distribution.", Neural Computation, 13(7):1443-1471,2001.