Automatic Image Collection of Objects with Similar Function by Learning Human Grasping Forms

Shinya Morioka, Tadashi Matsuo, Yasuhiro Hiramoto, Nobutaka Shimada, Yoshiaki Shirai

Ritsumeikan University, Shiga, Japan

Abstract. This paper proposes an automatic functional object segmentation method based on modeling the relationship between grasping hand form and the object appearance. First the relationship among a representative grasping pattern and a position and pose of a object relative to the hand is learned based on a few typical functional objects. By learning local features from the hand grasping various tools with various way to hold them, the proposed method can estimate 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.

1 Introduction

Generic object detection from an image classifies image subregions into object "categories", which is more difficult task than specific object detection. Since the object shape depends on the object function (thus on the grasping type) (as known as "affordance"[2]), the image patterns of the grasping hand can be a strong cues for the detection of a corresponding object category. Once a typical grasping pattern in an image is detected by some visual ques (SIFT, SURF or edgelets), a unique hand-object coordinate type can be fixed based on the relation between hand and object scale, position and orientation. Then according to the alignment with the estimated hand-object coordinate, the normalized object images can be automatically collected from live videos and its appearance or shape characteristic common to the category can be analyzed. Once the object-grasping-function relationship model[3, 5] is built for each grasping pattern from the aligned images, the object function can be inferred by estimating the grasping pattern through the model.

This paper proposes a machine learning based automatic object image collector for object grasping scene. First the relationship among a representative grasping pattern and a position and pose of a object relative to the hand is learned based on a few typical functional objects. By learning local features from the hand grasping various tools with various way to hold them, the proposed method can estimate 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.

2 Detecting wrist position with randomized trees

2.1 Training of Randomized Trees model providing the probability distribution of wrist positions

Randomized Trees (RTs)[4] is a multi-class classifier that accepts multi-dimensional features and it provides probability distributions over the class indexes. Here we construct RTs that can generate a probability distribution of a wrist position from Speeded-Up Robust Features (SURF)[1] features.

First, we specify a wrist position for each training image with a simple background by hand as shown in Fig.1(a). To learn relation between a wrist position and a set of SURF features as shown in Fig.1(b), we introduce a local coordinate system for representing a wrist position relatively. It is defined by the position, scale and orientation of the corresponding SURF feature as shown in (Fig. 1(c)). By using such a local coordinate system, a relative position of a wrist can be trained without depending on a position, scale or orientation of a hand in a training image. Since RTs generate a probability distribution of a discrete variable, a local coordinate space is segmented into finite number of blocks by a grid and a wrist position is represented by a block including the position as shown in Fig. 1(c). We assign a label for each block and assign a special label the condition that a wrist exists on the outside of all blocks. A position of a wrist can be represented by a pair of a label and a SURF feature. To estimate a label from a SURF descriptor, we train RTs so that they can calculate a probability distribution of such a label from a 128 dimensional SURF descriptor.

A label j is an index that means a block or background (the outside of all blocks). A local region of a SURF feature is divided into some square blocks. Each block is denoted as C_j. When a wrist position is $\mathbf{x}_{\mathrm{wrist}} = (x_{\mathrm{wrist}}, y_{\mathrm{wrist}})^t$ and a SURF feature f is detected on a hand, a label j of the SURF feature f is defined as a block index j such that $T_f(\mathbf{x}_{\mathrm{wrist}}) \in C_j$, where T_f denotes a normalization into the local coordinate based on the position, scale and direction of the SURF feature f.

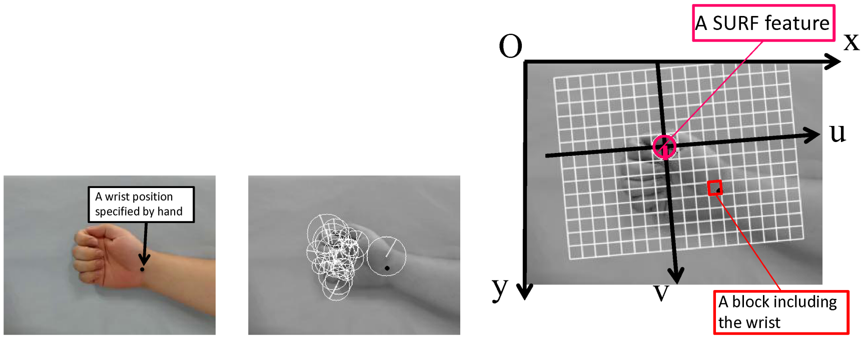
A label of a SURF feature detected on background is defined as a special index j_{back} and it is distinguished from block indexes.

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

2.2 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 $P(j|f)$ learned with the Randomized Trees. The votes function $V_{\mathrm{wrist}}(x, y)$ defined as $\sum_{f \text{ for all SURF features}} V_f(x, y)$, where

$$V_f(x,y) = \begin{cases} P(j = \tilde{j}|f) & \text{if } {}^{\exists}C_{\tilde{j}} \text{ s.t. } T_f(x,y) \in C_{\tilde{j}}, \\ 0 & (\text{otherwise}), \end{cases} \tag{1}$$


A wrist position specified by hand
A SURF feature
O
x
u
y
v
A block including the wrist

(a) A wrist position given by hand (b) SURF features (c) A local grid for a SURF feature

Fig. 1: A local grid for representing a wrist position relatively

and $P(j|f)$ is a probability distribution calculated by the trained Randomized Trees.

The position with the maximum votes, $\arg\max_{x,y} V_{\text{wrist}}(x, y)$, 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)).

If a background region is larger than a hand region, the above voting is disturbed by SURF features on the background. To overcome this, we roughly distinguish SURF features on a hand from those on the other region by a Support Vector Machine (SVM). The SVM classifies a SURF feature into two classes, which are "hand" and "background". Teacher samples for the "hand" class are extracted from images where a hand is placed on a simple background. Those for the "background" class are extracted from images consisting of a complex background. The above voting algorithm is performed with SURF features classified as the "hand" class.

An example of finding candidates is shown in Fig.3. First, SURF features are extracted from an input image (Fig.3(a)). By using a SVM, they are roughly classified and SURF features apparently originated from non-hand regions are excluded (Fig.3(b),(c)). For each SURF feature f, local coordinate is defined

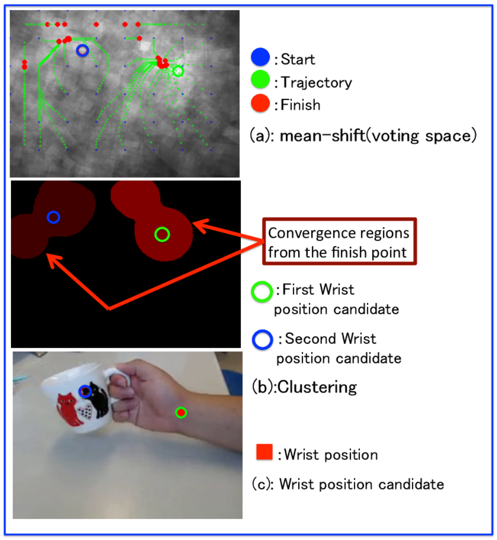

● : Start
● : Trajectory
● : Finish
(a): mean-shift(voting space)
Convergence regions from the finish point
○ : First Wrist position candidate
○ : Second Wrist position candidate
(b):Clustering
■ : Wrist position
(c): Wrist position candidate

Fig. 2: Flow of wrist candidate detection

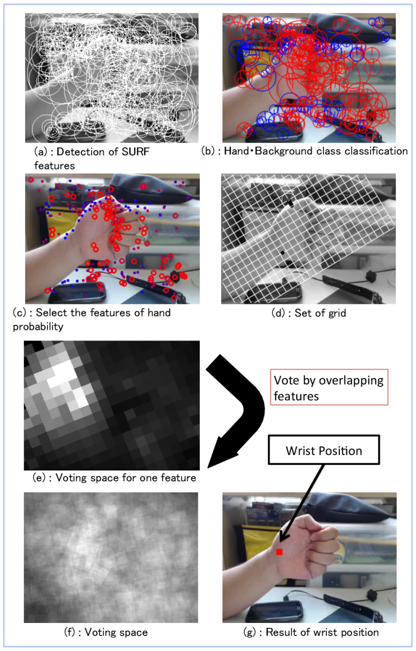

(a) : Detection of SURF features
(b) : Hand・Background class classification
(c) : Select the features of hand probability
(d) : Set of grid
Vote by overlapping features
(e) : Voting space for one feature
Wrist Position
(f) : Voting space
(g) : Result of wrist position

Fig. 3: Flow of wrist position detection

and a conditional probability distribution $P(j|f)$ is calculated by a Randomized Trees(Fig.3(d),(e)). By voting based on the probability distribution, candidates of wrist positions are determined (Fig.3(f),(g)).

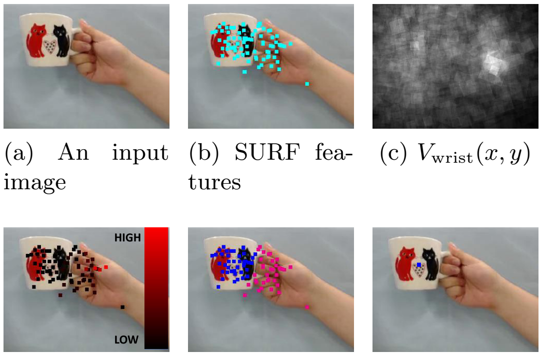
3 Extraction of hand and object regions

We extract hand and object regions by using relation with local features. Its rough process of training is following;

1. Generate training samples of pairs of an object center and a set of SURF features on a hand (the section 3.1).
2. Train the Randomized Trees so that outputs a probability distribution of an object center from a pair of a wrist position and a SURF feature on a hand (the section 3.2).
3. Train a one-class SVM[6] for finding an object region and the other one-class SVM for distinguishing whether a SURF feature is on a hand region or not (the section 3.3).

The rough process of detection is following;

1. Estimate a wrist position by the method in the section 2.



(a) An input image (b) SURF features (c) $V_{\mathrm{wrist}}(x, y)$

(d) Likelihood as a hand part (e) Classification (K-means) (f) An object center

Fig. 4: Estimation of an object center for generating training samples

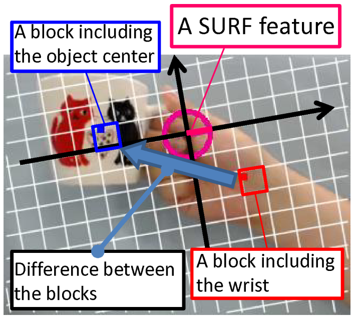

A block including the object center
A SURF feature
Difference between the blocks
A block including the wrist

Fig. 5: Positional relation between an object center and a wrist position

2. Estimate an object center by voting probabilities generated from the RTs trained at the training step 2. All SURF features take part in the voting.
3. Find an object region by the one-class SVM trained at the step 3. Distinguish SURF features on a hand from those on the other regions by another one-class SVM trained at the step 3.

3.1 Estimating an object center by coarse classification

A wrist position can be found by the algorithm described in the section 2. In addition to the wrist position, we use a center of gravity of an object, which makes a coordinate system suitable for learning positional relation between a hand and an object. For learning relation between the object center and a set of features, we generate training samples by coarse classification of SURF features extracted from images with simple backgrounds.

In an image of a hand grasping an object with simple background such as Fig. 4(a), a SURF feature belongs to a hand class or an object class. We classify such features (Fig. 4(b)) into the two classes by K-means clustering. On the clustering, each feature is represented by a triplet consisting of a coordinate value (x, y) of the feature and its "likelihood as a hand part" h. As a measure of the likelihood, we take how much the feature contributed to $V_{\mathrm{wrist}}(x, y)$ (Fig. 4(c)) used when determining the wrist position. The likelihood h of a SURF feature f is defined as $h = V_f(x_{\mathrm{wrist}}, y_{\mathrm{wrist}})$, where $(x_{\mathrm{wrist}}, y_{\mathrm{wrist}})$ denotes the estimated wrist position. An example of the likelihood are shown in Fig. 4(d). By classifying triplets, we can extract a set of SURF features on a hand as shown in Fig. 4(e). An object center is estimated as the average position of SURF features on an object as shown in Fig. 4(f).

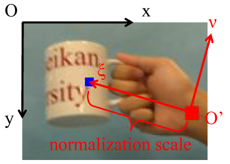


Fig. 6: A wrist-object coordinate system

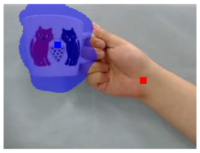


Fig. 7: An object region from M_{obj}

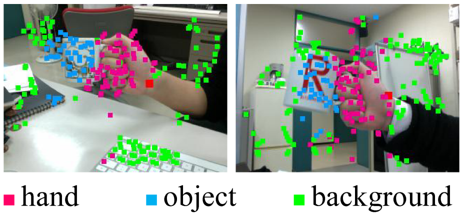


Fig. 8: A result of feature classification

3.2 Learning relation between an object center and a wrist position

By the method in the section 3.1, we have training sample images where an object center, a wrist position and a hand region are known. To represent a positional relation between an object center and a wrist position, we take a grid defined by a SURF feature, which is introduced in the section 2.1. On the grid, the relation can be represented by the positional difference between two blocks (Fig. 5). We train RTs with the differences so that it can calculate a probability distribution of a relative position of an object center from a pair of a wrist position and a SURF feature. By using the RTs similarly as the section 2, an object center can be estimated.

3.3 Learning one-class SVMs for finding an object region and features on a hand

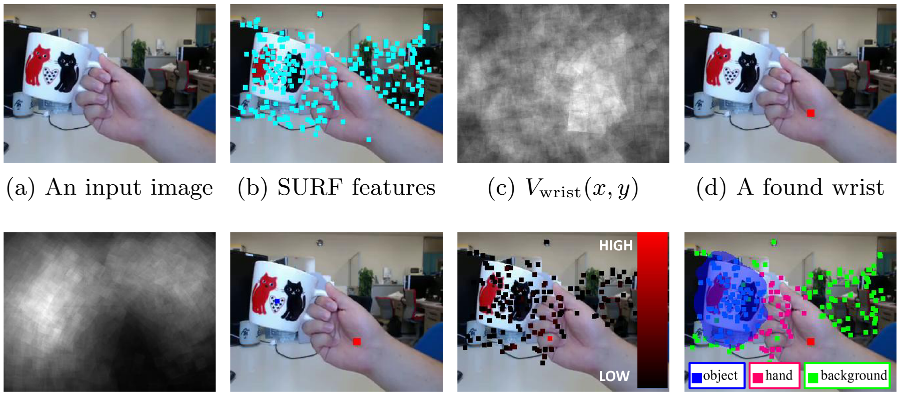
By using an object center and a wrist position, we can introduce a wrist-object coordinate system (ξ, ν), where the origin is the wrist position, one axis ξ extends to the object center and the distance between them is normalized as 1 (Fig.6). It is suitable for learning positional relation between a hand and an object.

We generate a one-class SVM M_{obj} that receives a coordinate value (ξ, ν) and outputs true if the position is included in an object region. Since "likelihood as an object part" cannot be estimated beforehand, we take a relative coordinate value (ξ, ν) of a SURF feature on an object region as a positive sample. Such a feature can be collected by the method in Sec. 3.1. An example of a object region derived from trained M_{obj} is shown as the blue region in Fig. 7, where the red point means the wrist position.

We also generate another one-class SVM M_{hand} for distinguishing SURF features on a hand from those on other regions. The SVM M_{hand} is trained with a set of a triplet (ξ, ν, h), where h means "likelihood as a hand part" defined in Sec. 3.1.

We classify each SURF feature f as follows;

1. If the SVM M_{hand} returns positive for the triplet (ξ, ν, h) of the feature f, it is classified as a hand feature.
2. If it is not classified as a hand feature and the SVM M_{obj} returns positive for the pair (ξ, ν) of the feature, it is classified as an object feature.



(a) An input image (b) SURF features (c) $V_{\mathrm{wrist}}(x, y)$ (d) A found wrist

(e) Votes for an object center (f) A found object center (g) Likelihood as a hand part (h) An estimated result

Fig. 9: A step-wise result of the proposed method

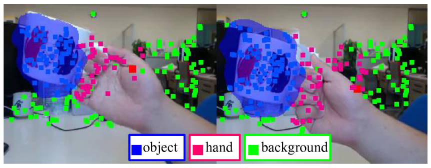


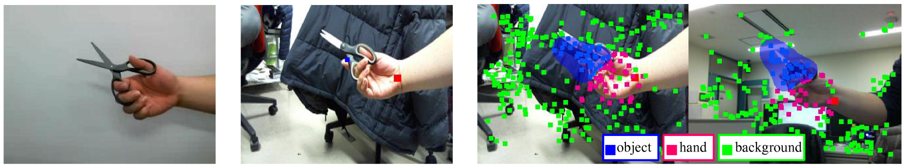
Fig. 10: Results for other images of a hand grasping a cup

3. Otherwise, it is classified as a background feature.

A result of feature classification is shown in Fig.8 In addition, an image can be divided into an object region and the other region because the SVM M_{obj} requires only a position.

4 Experimental result

We apply the proposed method to an image with complex background. The result images of each step of the method are shown in Fig. 9. In the input image (Fig. 9(a)), SURF features are extracted as Fig. 9(b). They generate $V_{\mathrm{wrist}}(x, y)$ as Fig. 9(c) and a wrist position can be estimated as (Fig. 9(d)). Although the estimated wrist position is a little off the true wrist, an object center is found correctly (Fig. 9(f)). By using the object center, the wrist position and the likelihood as a hand part (Fig. 9(g)), we can detect an object region and class of each SURF feature as Fig. 9(h). Results for other images of a hand grasping a cup are shown in Fig. 10. They show that object regions are extracted well if the ways of grasping a cup are different.



(a) One of images for training models (b) Estimated positions (c) An object region and feature classes

Fig. 11: Results for images of a hand grasping scissors

We also show results of a hand grasping scissors in Fig. 11. RTs and SVMs are trained with images such as Fig. 11(a). As shown in Fig. 11(b), an object center and a wrist position are correctly estimated, even though the grasped scissors differ from that in the training images. Object regions are also correctly estimated as shown in Fig. 11(c).

5 Conclusion

By integrating local features, a position of a hand can be estimated even if its background is complex and the hand is partially hidden. With Randomized Trees, a wrist can be found and a gravity center of an object can be estimated from a set of the wrist and local features. The wrist and the object center make a wrist-object coordinate system suitable for learning a shape of an object which depends on how the object is grasped. In the future, we will try object recognition by learning the relation between an object and a posture of a hand grasping it.

References

1. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding 110(3), 346 – 359 (2008), http://www.sciencedirect.com/science/article/pii/S1077314207001555
2. Gibson, J.J.: The Ecological approach to visual perception. Lawrence Erlbaum Associates, new edition edn. (Sep 1986), http://www.worldcat.org/isbn/0898599598
3. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31(10), 1775–1789 (Oct 2009), http://dx.doi.org/10.1109/TPAMI.2009.83
4. Lepetit, V., Fua, P.: Keypoint recognition using randomized trees. Pattern Analysis and Machine Intelligence, IEEE Transactions on 28(9), 1465–1479 (2006)
5. Pieropan, A., Ek, C.H., Kjellström, H.: Functional object descriptors for human activity modeling. In: Proc. of Int'l Conf. on Robotics and Automation 2013. p. 12821289 (2013)
6. Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. 13(7), 1443–1471 (Jul 2001), http://dx.doi.org/10.1162/089976601750264965