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

Shinya Morioka, Tadashi Matsuo(✉), Yasuhiro Hiramoto, Nobutaka Shimada, and Yoshiaki Shirai

Ritsumeikan University, Shiga, Japan
matsuo@i.ci.ritsumei.ac.jp

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 image classifies the image subregions into object "categories", which is more difficult task than specific object detection. Since the target objects in one category have large variations of appearance in many cases, the framework which can detect targets even in clutter backgrounds and partial occlusions with other objects is required. Recently many machine-learning-based detection methods using structured local image features like Bag-of-features or graphical model are proposed. As a survey paper of image-based object detection by Zhang [22] such structural feature models are divided into "window-based" model like HoG (histogram of gradient) feature and "part-based" model like Implicit Shape Model [9]. While the former tends to be weak to large occlusion the latter gives comparatively good performance for it since it models loose connections between local features as parts of the object: boosting detector using edgelet feature [19], voting-based-detector using partial contourlet [17], constellation model based on SIFT local feature [21], 3-D shape constellation model using RGB-D cam [11] and conditional random field of dense label map of object region [18].

While the above methods detect objects using only image feature or appearance information, literature points out that the object category is defined by not simply appearance or shape but also its "function" [14]. From that viewpoint it is decisively important what kind of dynamical actions the 3-D shape of each

F. Schwenker et al. (Eds.): MPRSS 2014, LNAI 8869, pp. 3–14, 2015.
DOI: 10.1007/978-3-319-14899-1_1

part generates. Base on the assumption that the dynamical action of a common partial shape of one object category gives the object function unique for the category, [16] builds a graphical model of the 3-D partial shape and infers the object category.

This paradigm is reduced to "affordance" [4] since most of artificial objects assume human manipulations and are designed as tools with specific functions. Affordance means that the object shape reminds *the usage* of the object, that is how human physically uses the objects. The usage model requires the pairwise descriptions of the relative poses and motions of both human and the object. Gupta [6] shows an representative example of discrimination of PET bottle and spray can which have quite similar shape. It points out that they can be discriminated by considering the usage: drinking pose and motion for PET bottle and targeting pose and button pushing for spray can. This framework can be applied to object categories that include large shape or appearance variations. There are some researches on this framework: recognition of function as "chair" by 3-D human action simulation with the object shape [5], estimation of used objects in cooking scene by considering human motion [20], and object recognition by learning the relationship between object arrangement in the living room environment and human actions [7].

While these researches consider the macro-size poses and motions of human body, Gupta [6] refers an interesting suggestion reported in psychological field [2], which points out that when human recognizes the function of an object he/she often reminds the hand gesture grasping the object to be recognized. This means that the grasping details is necessary to categorize handy-size objects. Pieropan et al. [13] models the typical hand motion and position by clustering and identifies the functional object categories (tool, ingredient, support area, container). This research mainly considers only upper body and hand motions, not grasping details.

The literature revealed that human carefully selects the grasping patterns when using an object by considering the function to be invoked [8,12], for example, "Lat" type (intermediate grip: lateral grip) for a mug cap, and "PoD" type (power grip: distal type) for scissors. A recent research employs this grasping description for categorizing objects by building a graphical model describing the relationship between object-hand contacting point, object's appearance and the human's motion [3]. In the research the appearance of the specific object and its grasping hand are directly bound. In general, it requires too high costs to individually collect such combinations between object and grasping for various objects. The object appears in various size, orientation and position in an image. If the object shape depends on the object function (thus on the grasping type), the image patterns of the grasping hand can be a strong cues for the detection of a corresponding object category. After the detection of the typical grasping patterns in the image, the hand-object coordinate unique for each grasping type can be estimated and the scale-position-orientation normalized object regions can be automatically collected from the image database or live videos. In addition once the normalized object appearance or shape model is built for each

grasping pattern, the object function can be inferred by estimating the grasping pattern through the object-grasping-function relationship model. In this paper we focus on such an approach for hand-object-function model building.

In this paper, first the relationship between representative grasping patterns and object position and pose relative to hand is learned based on a few typical functional objects. Then based on the obtained object-grasping model the registered grasping pattern is automatically detected in the still image with cluttered background, the hand-object coordinate is attached onto the image region, the normalized object region is segmented and collected. The detail algorithm and the experimental results for this framework are shown.

2 Detecting Wrist Position with Randomized Trees

2.1 Training of Randomized Trees Model Providing the Probability Distribution of Wrist Positions

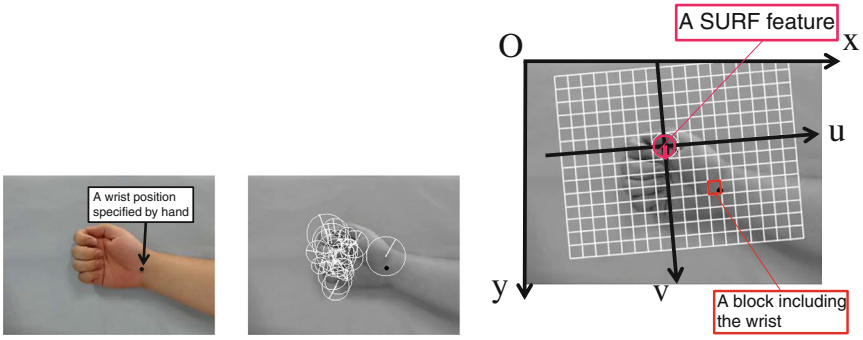
Randomized Trees (RTs) [10] is a multi-class classifier that accepts multi-dimensional features and it provides probability distributions over the class indexes. Here we construct RTs that can generate a probability distribution of a wrist position from Speeded-Up Robust Features (SURF) [1] features.

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

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

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



(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

When a wrist position is $\mathbf{x}_{\text{wrist}} = (x_{\text{wrist}}, y_{\text{wrist}})^t$ 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 $T_f(\mathbf{x}_{\text{wrist}}) \in C_j$, 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 $V_{\text{wrist}}(x, y)$ defined as $\sum_{f \text{ for all SURF features}} V_f(x, y)$, where

$$V_f(x,y) = \begin{cases} P(j=\tilde{j}|f) & \text{if } {}^{\exists}C_{\tilde{j}} \text{ s.t. } T_f(x,y) \in C_{\tilde{j}}, \\ 0 & (\text{otherwise}), \end{cases} \tag{2}$$

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.

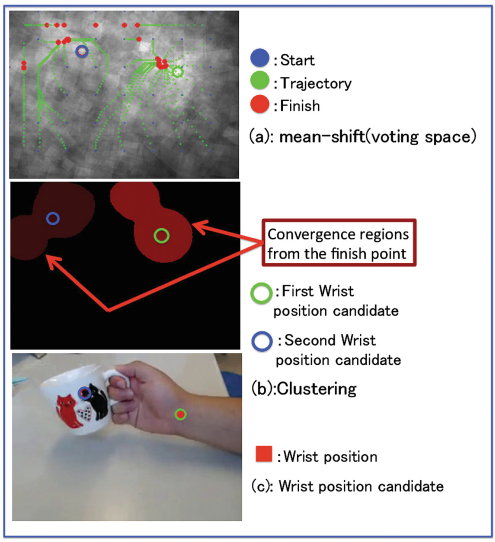


Fig. 2. Flow of wrist candidate detection (Color figure online)

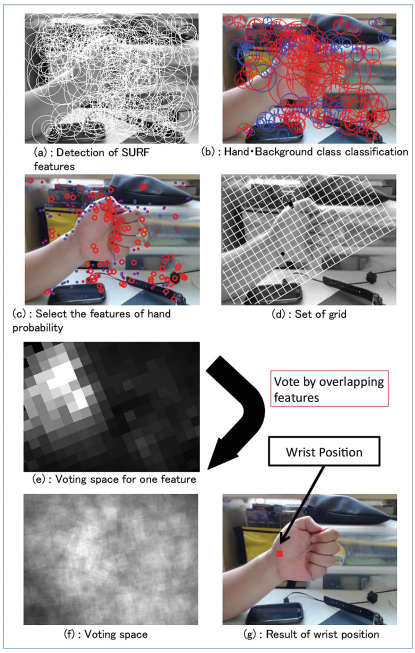


Fig. 3. Flow of wrist position detection

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 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 Sect. 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 Sect. 3.2).
3. Train a one-class SVM [15] 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 Sect. 3.3).

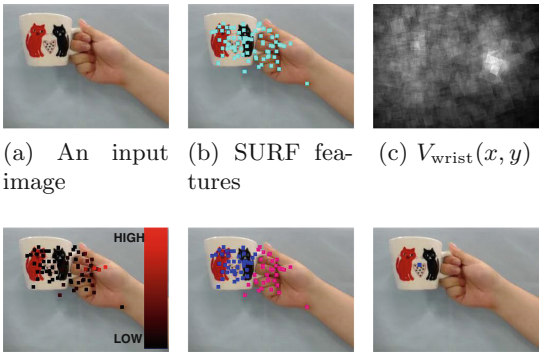
The rough process of detection is following;

1. Estimate a wrist position by the method in the Sect. 2.
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 Sect. 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 $V_{\text{wrist}}(x, y)$ (Fig. 4(c)) used when determining the wrist position. The likelihood h of a SURF feature f is defined as $h = V_f(x_{\text{wrist}}, y_{\text{wrist}})$, where $(x_{\text{wrist}}, y_{\text{wrist}})$ 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).



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

(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

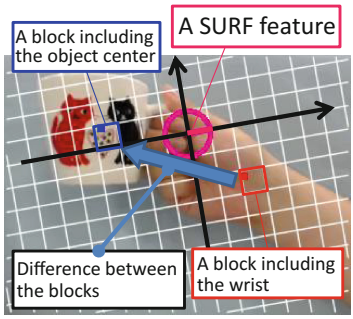


Fig. 5. Positional relation between an object center and a wrist position

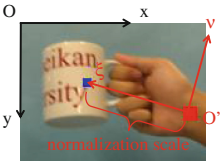


Fig. 6. A wrist-object coordinate system

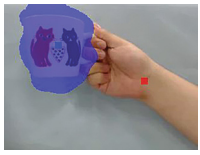


Fig. 7. An object region from M_{obj} (Color figure online)

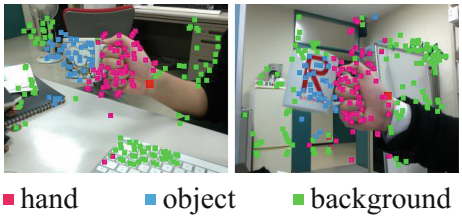


Fig. 8. A result of feature classification

3.2 Learning Relation Between an Object Center and a Wrist Position

By the method in the Sect. 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 Sect. 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 Sect. 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.

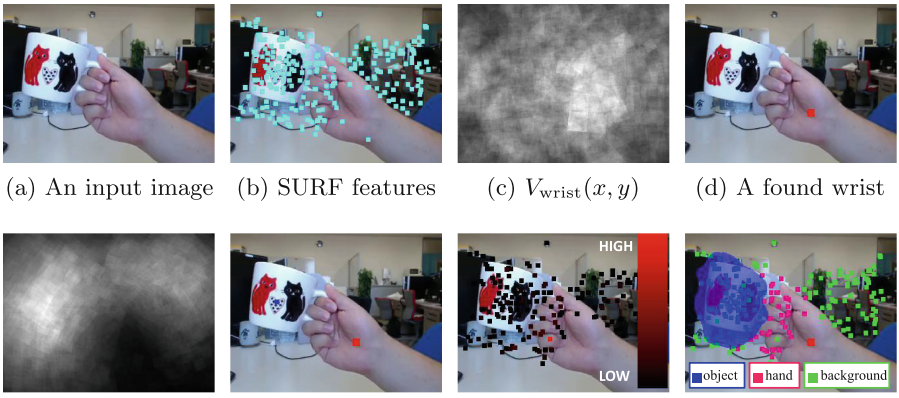
We generate a one-class SVM M_{obj} that receives a coordinate value (ξ, ν) and outputs true if the position is included in an object region. Since "likelihood as an object part" cannot be estimated beforehand, we take a relative coordinate value (ξ, ν) of a SURF feature on an object region as a positive sample. Such a feature can be collected by the method in Sect. 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 Sect. 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.
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.



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

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.

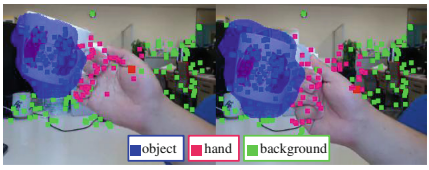
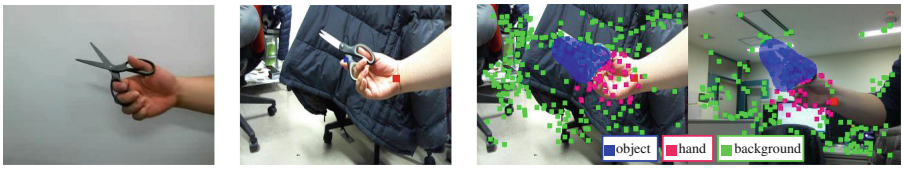


Fig. 10. Results for other images of a hand grasping a cup



(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

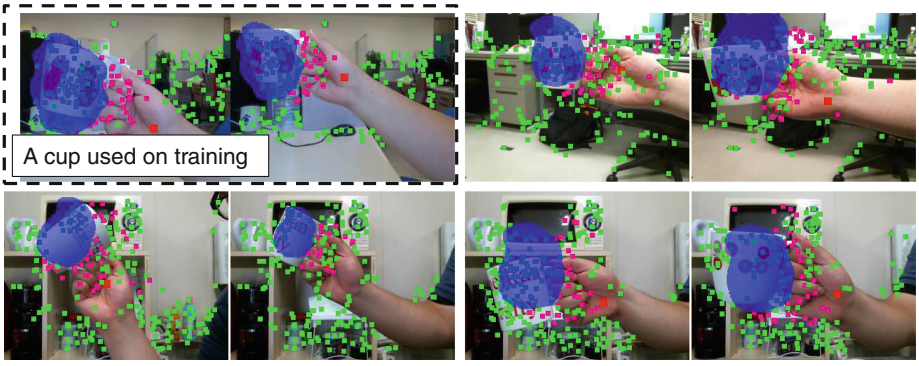


Fig. 12. Results for a hand grasping a cup not used on training

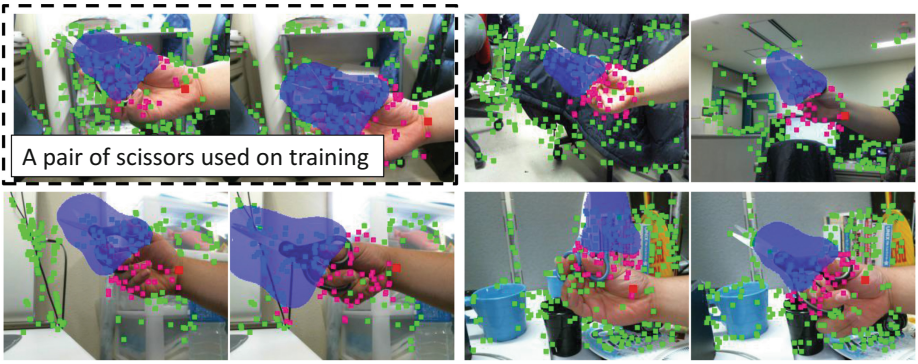


Fig. 13. Results for a hand grasping a scissors not used on training

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).

In Figs. 12 and 13, we show results of a hand grasping an object which is not used on training. The results show that the proposed method works well for unknown objects by focusing on how they are grasped.

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). Comput. Vis. Image Underst. **110**(3), 346–359 (2008). http://www.sciencedirect.com/science/article/pii/S1077314207001555
2. Bub, D., Masson, M.: Gestural knowledge evoked by objects as part of conceptual representations. Aphasiology **20**(9), 1112–1124 (2006). http://www.tandfonline.com/doi/abs/10.1080/02687030600741667
3. Filipovych, R., Ribeiro, E.: Recognizing primitive interactions by exploring actor-object states. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–7 (2008)

4. Gibson, J.J.: The Ecological Approach to Visual Perception. Lawrence Erlbaum Associates, Hillsdale (1986). http://www.worldcat.org/isbn/0898599598
5. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1529–1536 (2011)
6. 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 (2009). http://dx.doi.org/10.1109/TPAMI.2009.83
7. Jiang, Y., Lim, M., Saxena, A.: Learning object arrangements in 3D scenes using human context. In: Proceedings of the 29th International Conference on Machine Learning (ICML 2012), pp. 1543–1550 (2012)
8. Kamakura, N., Matsuo, M., Ishii, H., Mitsuboshi, F., Miura, Y.: Patterns of static prehension in normal hands. Am. J. Occup. Ther. **34**(7), 437–445 (1980). http://ajot.aotapress.net/content/34/7/437.abstract
9. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: ECCV Workshop on Statistical Learning in Computer Vision, pp. 17–32 (2004). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.6272
10. Lepetit, V., Fua, P.: Keypoint recognition using randomized trees. IEEE Trans. Pattern Anal. Mach. Intell. **28**(9), 1465–1479 (2006)
11. Madry, M., Afkham, H.M., Ek, C.H., Carlsson, S., Kragic, D.: Extracting essential local object characteristics for 3D object categorization. In: 2013 Proceedings of IEEE International Conference on Intelligent Robots and Systems, TuAT4.5 2013 (2013)
12. Napier, J.R.: The prehensile movements of the human hand. J. Bone Joint Surg. **38**(4), 902–913 (1956)
13. Pieropan, A., Ek, C.H., Kjellström, H.: Functional object descriptors for human activity modeling. In: 2013 Proceedings of International Conference on Robotics and Automation, pp. 1282–1289 (2013)
14. Rivlin, E., Dickinson, S., Rosenfeld, A.: Recognition by functional parts [function-based object recognition]. In: 1994 Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1994, pp. 267–274 (1994)
15. 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 (2001). http://dx.doi.org/10.1162/089976601750264965
16. Sgorbissa, A., Verda, D.: Structure-based object representation and classification in mobile robotics through a microsoft kinect. Robot. Auton. Syst. **61**(12), 1665–1679 (2013)
17. Shotton, J., Blake, A., Cipolla, R.: Contour-based learning for object detection. In: 2005 Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 1, pp. 503–510. IEEE (2005)
18. Winn, J., Shotton, J.: The layout consistent random field for recognizing and segmenting partially occluded objects. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 37–44. IEEE (2006)
19. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV 2005), vol. 1, pp. 90–97. IEEE Computer Society, Washington, DC (2005). http://dx.doi.org/10.1109/ICCV.2005.74

20. Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.: A scalable approach to activity recognition based on object use. In: 2007 IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8 (2007)
21. Zhang, H., Bai, X., Zhou, J., Cheng, J., Zhao, H.: Object detection via structural feature selection and shape model. IEEE Trans. Image Process. **22**(12), 4984–4995 (2013)
22. Zhang, X., Yang, Y.H., Han, Z., Wang, H., Gao, C.: Object class detection: a survey. ACM Comput. Surv. **46**(1), 10:1–10:53 (2013). http://doi.acm.org/10.1145/2522968.2522978