Fast 6D Pose Estimation from a Monocular Image Using Hierarchical Pose Trees

Yoshinori Konishi[1](✉), Yuki Hanzawa[1], Masato Kawade[1], and Manabu Hashimoto[2]

[1] OMRON Corporation, Kyoto, Japan
{ykoni,hanzawa,kawade}@ari.ncl.omron.co.jp
[2] Chukyo University, Nagoya, Japan
mana@isl.sist.chukyo-u.ac.jp

Abstract. It has been shown that the template based approaches could quickly estimate 6D pose of texture-less objects from a monocular image. However, they tend to be slow when the number of templates amounts to tens of thousands for handling a wider range of 3D object pose. To alleviate this problem, we propose a novel image feature and a tree-structured model. Our proposed perspectively cumulated orientation feature (PCOF) is based on the orientation histograms extracted from randomly generated 2D projection images using 3D CAD data, and the template using PCOF explicitly handle a certain range of 3D object pose. The hierarchical pose trees (HPT) is built by clustering 3D object pose and reducing the resolutions of templates, and HPT accelerates 6D pose estimation based on a coarse-to-fine strategy with an image pyramid. In the experimental evaluation on our texture-less object dataset, the combination of PCOF and HPT showed higher accuracy and faster speed in comparison with state-of-the-art techniques.

Keywords: 6D pose estimation · Texture-less objects · Template matching

1 Introduction

Fast and accurate 6D pose estimation of object instances is one of the most important computer vision technologies for various robotic applications both for industrial and consumer robots. In recent years, low-cost 3D sensors such as Microsoft Kinect became popular and they have often been used for object detection and recognition in academic research. However, much more reliability and durability are required for sensors in industrial applications than in consumer applications. Thus the 3D sensors for industry are often far more expensive,

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46448-0_24) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part I, LNCS 9905, pp. 398–413, 2016.
DOI: 10.1007/978-3-319-46448-0_24

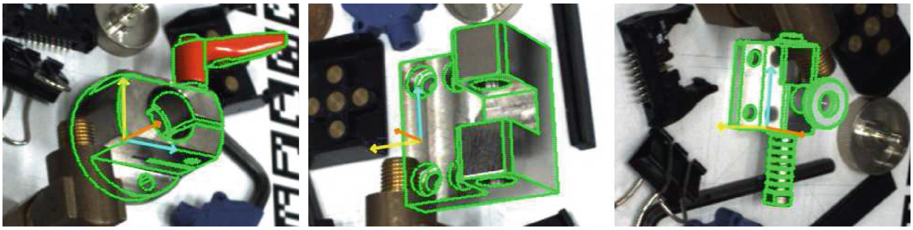


Fig. 1. Our new template based algorithm can estimate 6D pose of texture-less and shiny objects from a monocular image which contains cluttered backgrounds and partial occlusions. It takes an average of approximately 150 ms on a single CPU core.

larger in size and heavier than the consumer ones. Additionaly, most of 3D sensors even for industry cannot handle objects with specular surfaces, are sensitive to illumination conditions and require cumbersome 3D calibrations. For those reasons, monocular cameras are mainly used in the current industrial applications, and fast and accurate 6D pose estimation from a monocular image is still an important technique.

Many of industrial parts and products have little texture on their surfaces, and they are so-called texture-less objects. Object detection methods based on keypoints and local descriptors such as SIFT [1] and SURF [2] cannot handle texture-less objects because they require rich textures on the regions of target objects. It has been shown that template based approaches [3–9] which use whole 2D projection images from various viewpoints as their model templates successfully dealt with texture-less objects. However, they suffer from the speed degradation when the numbers of templates are increased for covering a wider range of 3D object pose.

We propose a novel image feature and a tree-structured model for fast template based 6D pose estimation (Fig. 1). Our main contributions are as follows:

- We introduce perspectively cumulated orientation feature (PCOF) extracted using 3D CAD data of target objects. PCOF is robust to the appearance changes caused by the changes in 3D object pose, and the number of templates are greatly reduced without loss of pose estimation accuracy.
- Hierarchical pose trees (HPT) is also introduced for efficient 6D pose search. HPT consists of hierarchically clustered templates whose resolutions are different at each level, and it accelerates the subwindow search by a coarse-to-fine strategy with an image pyramid.
- We make available a dataset of nine texture-less objects (some of them have specular surfaces) with the ground truth of 6D object pose. The dataset includes approximately 500 images per object taken from various viewpoints, and contains cluttered backgrounds and partial occlusions. 3D CAD data for training are also included.[1]

[1] http://isl.sist.chukyo-u.ac.jp/Archives/archives.html.

The remaining contents of the paper are organized as follows: Sect. 2 presented related work on 6D pose estimation, image features for texture-less objects and search data structures. Section 3 introduces our proposed PCOF, HPT and 6D pose estimation algorithm based on them. Section 4 evaluates the proposed method and compare it with state-of-the-art methods. Section 5 concludes the paper.

2 Related Work

6D Pose Estimation. 6D pose estimation has been extensively studied since 1980s and in the early days the template based approaches using a monocular image [3–5] were the mainstream. Since the early 2000s, keypoint detections and descriptor matchings became popular for detection and pose estimation of 2D/3D objects due to their scalability to the increasing search space and robustness to the changes in object pose. Though they can handle texture-less objects when using line features as the descriptors for matching [10,11], they were fragile to cluttered backgrounds because the line features were too simple to suffer from many false correspondences in the backgrounds.

Voting based approaches as well as template based approaches have a long history, and they have also been applied to detection and pose estimation of 2D/3D objects. Various voting based approaches were proposed for 6D pose estimation such as voting by dense point pair features [12], random ferns [13], Hough forests [14], and coordinate regressions [15]. Though they are scalable to increasing image resolutions and the number of object classes, the dimensionaliy of search space is too high to estimate precise object pose (excessive quantizations of 3D pose space are required). Thus they need post-processings for pose refinements, which spend additional time.

CNN based approaches [16–18] recently showed impressive results on 6D pose estimations. However, they take a few seconds even when using GPU and they are not suitable for robotic applications where near real-time processing is required on poor computational resources.

Template based approaches have been shown to be practical both in accuracy and speed for 6D pose estimation of texture-less objects [6,7,19]. Hinterstoisser et al. [8,9] showed their LINE-2D/LINE-MOD which is based on the quantized orientations and the optimally arranged memory quickly estimated 6D pose of texture-less objects against cluttered backgrounds. LINE-2D/LINE-MOD was further improved by discriminative training [20] and by hashing [21,22]. However, the discriminative trainig required additional negative samples and the hashing led to suboptimal performance in the estimation accuracy.

Image Features for Handling Texture-less Objects. Image features used in template matching heavily influence the performance of pose estimation from a monocular image. Though edges based template matchings have been applied to detection and pose estimation of texture-less objects, they often required the additional algorithm such as segmentation [19] or the additional hardware like a multi-flash camera [6] to suppress cluttered edges in the backgrounds.

It has been shown that the gradient direction vectors [23] and the quantized gradient orientations [24] were robust to cluttered backgrounds and illumination changes. However, it was pointed out that the similarity scores based on these features rapidly declined even if only slight changes in object pose occurred. To overcome this problem, dominant orientations within a grid of pixels (DOT) [25] and spread orientation which allowed some shifting in matching [8] were proposed. DOT and spread orientation are robust to the pose changes and slight deformations of target objects. However, they relax matching conditions both in foregrounds and backgrounds, and this possibly degrade the robustness to cluttered backgrounds. Konishi et al. [26] introduced cumulative orientation feature (COF) which was robust to the apperance changes caused by the changes in 2D object pose. However, COF did not explicitly handle appearance changes caused by the changes in 3D object pose.

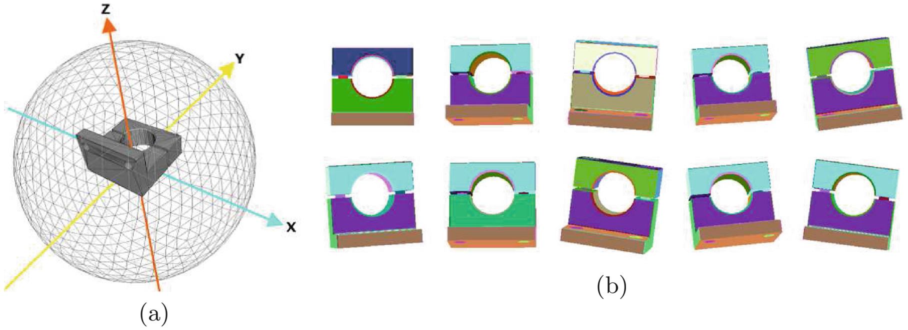
Tree-Structured Models for Efficient Search. Search strategies and data structures are also important for template based approaches. The tree-structured models are popular in the nearest neighbor search for image classification [27–29] and for joint object class and pose recognition [30]. These tree-structured models were also used in joint 2D detection and 2D pose recognition [31] and joint 2D detection and 3D pose estimation [32]. Though they offered efficient search in 2D/3D object pose space but not in 2D image space (x-y translations). The well-known efficient search in 2D image space is the coarse-to-fine search [33]. Ulrich et al. [7] proposed the hierarchical model which combined the coarse-to-fine search and the viewpoint clustering based on similarity scores between templates. However, their model is not fully optimized for the search in 3D pose space when 2D projection images from separate viewpoints are similar, as is often the case with texture-less objects.

3 Proposed Method

Our proposed method consists of a image feature for dealing with the appearance changes caused by the changes in 3D object pose (Sect. 3.1) and a hierarchical model for the efficient search (Sect. 3.2). The template based 6D pose estimation algorithm using both PCOF and HPT is described in Sect. 3.3.

3.1 PCOF: Perspectively Cumulated Orientation Feature

In this subsection, the way how to extract PCOF is explained using L-Holder shown in Fig. 2(a) which is a typical texture-less object. Our PCOF is developed from COF [26] and the main difference is two-fold: One is that PCOF explicitly handle appearance changes caused by the changes in 3D object pose, whereas COF can handle appearance changes only by 2D pose changes. Another is that PCOF is based on a probabilistic representation of quantized orientations at each pixel, whereas COF uses all the orientations observed at each pixel.



(a)

(b)

Fig. 2. (a) 3D CAD data of L-Holder, its coordinate axes and a sphere for viewpoint sampling. (b) Examples of the generated projection images from randomized viewpoints around the viewpoint on z-axis (upper-left image). Surfaces of objects are drawn by randomly selected colors in order to extract distinct image gradients.

Firstly many 2D projection images are generated using 3D CAD data from randomized viewpoints (Fig. 2(a)). The viewpoints are determined by four parameters those are rotation angles around x-y axes, a distance from the center of the object and a rotation angle around a optical axis. The range of randomized parameters should be limited so as to a single template can handle the appearance changes caused by the randomized parameters. In our research, the range of randomization were experimentally determined and those were ±12 ° around x-y axes, ±40 mm in the distance and ±7.5 ° around the optical axis. Figure 2(b) shows examples of generated projection images. The upper-left image of Fig. 2(b) is the projection image from the viewpoint where all rotation angles are zero and the distance from the object is 680 mm, and this viewpoint is at the center of these randomized examples. In generation of projection images, the neighboring meshes where the angle between them is larger than a threshold value are drawn by different color in order to extract distinct image gradients. In this study the threshold was 30 °.

Secondly image gradients of all the generated images are computed using Sobel operators (the maximum gradients among RGB channels are used). We use only the gradient directions discarding gradient magnitudes because the magnitudes depend on the randomly selected mesh colors. The colored gradient directions of the central image (the upper-left in Fig. 2(b)) are shown in Fig. 3(a). Then the gradient direction is quantized into eight orientations disregarding its polarities (Fig. 3(b)), and the quantized orientation is used for voting to the orientation histogram at each pixel. The quantized orientations of all the generated images are voted to the orientation histograms at the corresponding pixels. Lastly the dominant orientations at each pixel are extracted by thresholding the histograms and they are represented by 8-bit binary strings [25]. The maximum frequencies of the histograms are used as weights in calculating a similarity score.

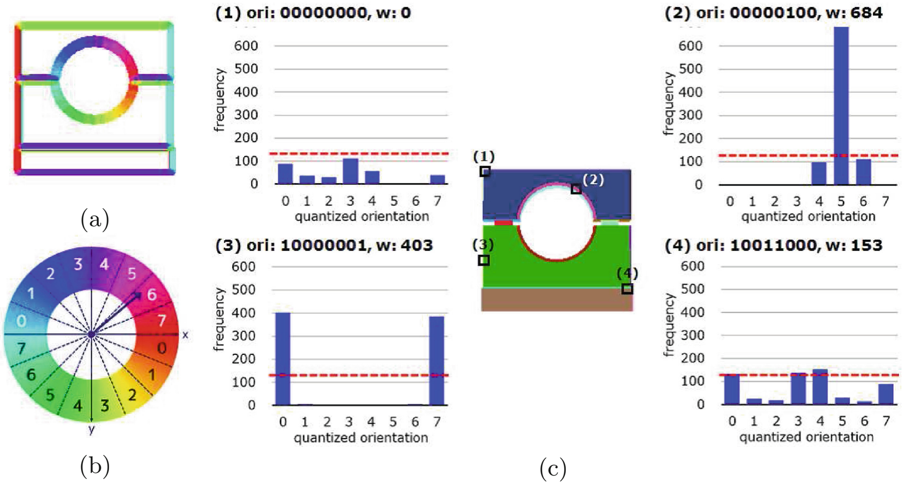


Fig. 3. (a) Colored gradient directions of the upper-left image in Fig. 2(b). (b) Quantization of gradient directions disregarding their polarities. (c) Examples of the orientation histograms, binary features (ori) and their weights (w) on arbitrarily selected four pixels. Red dotted lines show the threshold for feature extraction. (Color figure online)

The template T with n PCOF represented as follows:

$$T : \{x_i, y_i, ori_i, w_i | i = 1, ..., n\}\,, \tag{1}$$

and the similarity score is given by following equation,

$$score(x, y) = \frac{\sum_{i=1}^{n} \delta_k (ori^I_{(x+x_i, y+y_i)} \in ori^T_i)}{\sum_{i=1}^{n} w_i}. \tag{2}$$

If the quantized orientation of the test image (ori^I) is included in the PCOF template (ori^T), the weight (w) is added to the score. The delta function in Eq. (2) is calculated quickly by a bitwise AND operation (the symbol $\wedge$). Additionally, this calculation can be accelerated using SIMD instructions where multiple binary features are matched by a single instruction.

$$\delta_i(ori^I \in ori^T) = \begin{cases} w_i & \text{if } ori^I \wedge ori^T > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The orientation histograms, extracted binary features and their weights on arbitrarily selected four pixels are shown in Fig. 3(c). In our study, the number of generated images was 1,000 and the threshold value was 120. The votes were concentrated on a few orientations at the pixels along lines or arcs such as pixel (2) and (3). At these pixels the important features with large weights were extracted. On the contrary, the votes were scattered among many orientations at the pixels on corners and complicated structures such as pixel (1) and (4). At these pixels the features with small or zero weihts were extracted. Features with zero weights are not used for matching in pose estimation.

Algorithm 1. Building hierarchical pose trees

```
Input: a number of PCOF templates $T$ and their orientation histograms $H$
Output: hierarchical pose trees
  $T'_0 \leftarrow T$
  $H'_0 \leftarrow H$
  $i \leftarrow 1$
  loop
    $C_i \leftarrow$ cluster the templates in $T'_{i-1}$
    for each cluster $C_{ij}$ do
      $H_{ij} \leftarrow$ add histograms at each pixel of $H'_{i-1} \in C_{ij}$
      $H_{ij} \leftarrow$ normalize histograms $H_{ij}$
      $T_{ij} \leftarrow$ thresholding $H_{ij}$ and extract new binary features and weights
    end for
    for each $T_{ij}$ and $H_{ij}$ do
      $H'_{ij} \leftarrow$ add histograms of nearby $2 \times 2$ pixels
      $H'_{ij} \leftarrow$ normalize histograms $H'_{ij}$
      $T'_{ij} \leftarrow$ thresholding $H'_{ij}$ and extract new binary features and weights
    end for
    $N'_i \leftarrow$ minimum number of feature points in $T'_i$
    if $N'_i < N_{min}$ then
      break
    else
      $i \leftarrow i + 1$
    end if
  end loop
```

3.2 HPT: Hierarchical Pose Trees

A single PCOF template can handle the apparance changes caused by 3D pose changes generated in training (±12 ° around x-y axes, ±40 mm in the distance and ±7.5 ° around the optical axis). To cover a wider range of 3D object pose, additional templates are made at every vertices of the viewpoint sphere in Fig. 2(a) which contains 642 vertices as a whole and two adjacent vertices are approximately 8 ° apart. Additionally, the templates are made in every 30 mm in the distance to the object and in every 5 ° around the optical axes. These PCOF templates can redundantly cover the whole 3D pose space.

Our proposed hierarchical pose trees (HPT) are built in a bottom-up way starting from a lot of PCOF templates and their orientation histograms. The algorithm is shown in Algorithm 1 and it consists of three steps: clustering, integration and reduction of resolutions. Firstly all the templates are clustered based on the similarity scores (Eq. 2) between templates using X-means algorithms [34]. In X-means clustering, the optimum number of clusters are estimated based on Bayesian information criteria (BIC). Secondly the orientation histograms which belong to a same cluster are added and normalized at each pixel. Then the clustered templates are integrated to new templates by extracting the binary features and the weights from these integrated orientation histograms. Lastly the resolutions of the histograms are reduced to half by adding and normalizing

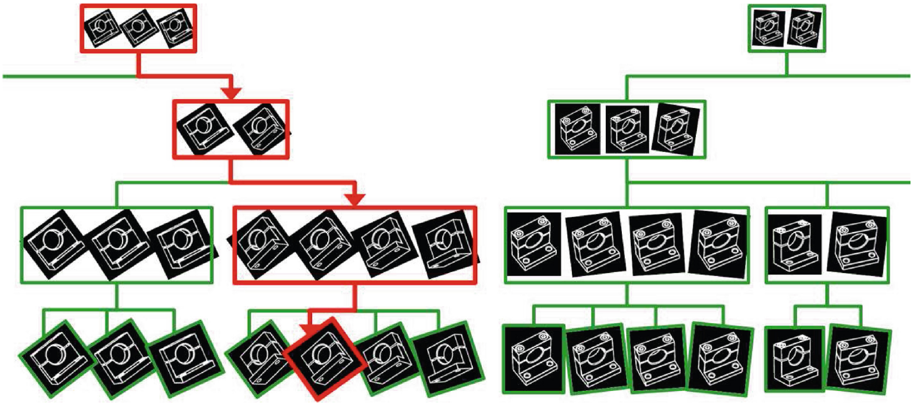


Fig. 4. Part of hierarchical pose trees are shown. Green and red rectangles represent templates used for matching. The bottom templates are originally created PCOF templates and the tree structures are built in a bottom-up way by clustering similar templates, integrating them into new templates and decreasing the resolutions of the templates. In estimation of object pose, HPT is traced from top to bottom along the red line, and the most promising template which contains the pose parameters is determined. (Color figure online)

histograms of neighboring 2 × 2 pixels. Then the low-resolution features and weights are extracted from these histograms. These procedures are iterated until the minimum number of feature points contained in low resolution templates is less than a threshold value (N_{min}). In our study N_{min} was 50.

Part of HPT are shown in Fig. 4. When the range of 3D pose was as same as the settings of experiment2 (±60° around x-y axes, 660 mm – 800 mm in the distance from the object and ±180° around the optical axis), the total number of PCOF templates amounted to 73,800 (205 viewpoints × 5 distances × 72 angles around the optical axis). These initial templates were clustered and integrated into 23,115 templates at the end of first round in Algorithm 1, and the number of templates was further reduced to 4,269 at second round and to 233 at third round. In this experimental settings, the iteration of hierarchization stopped at third round.

3.3 6D Pose Estimation

In 6D pose estimation, firstly the image pyramid of a test image is built and the quantized orientations are calculated on each pyramid level. Then the top level of the pyramid is scanned using the root nodes of HPT (e.g. the number of root nodes was 233 in experiment2). The similarity scores are calculated based on Eq. 2. The promising candidates whose scores are higher than a search threshold are matched with the templates at the lower levels, and they trace HPT down to the bottom. Finally the estimated results of 2D positions on a test image and the matched templates which have four pose parameters (three rotation angles and a distance) are obtained after non-maximum suppressions. 6D object pose of these results are calculated by solving PnP problems based on the correspondences between 2D feature points on the test image and 3D points of CAD data [35].

4 Experimental Results

We carried out two experiments. One is to evaluate the robustness of PCOF against cluttered backgrounds and the appearance changes caused by the changes in 3D object pose. Another is to evaluate the accuracy and the speed for our combined PCOF and HPT to estimate 6D pose of texture-less objects.

4.1 Experiment1: Evaluation of Orientation Features

Experimental Settings. In experiment1, we evaluated four kinds of orientation features on two test image sets ("vertical" and "perspective") shown in Fig. 5. A vertical image (a) was captured from the viewpoint on z-axis distanced by 680 mm from the center of the object. The upper-left image in Fig. 2(b) is the 2D projection image of 3D CAD from the same viewpoint. Perspective images (b) were captured from the same viewpoint as the vertical image with L-Holder slightly rotated around x-y axes (approximately 8 °, please see Fig. 2(b) as references). The number of the perspective images was eight (the combination of +/0/− rotation around x-y axes) and these images contain almost the same cluttered backgrounds. Our proposed PCOF was compared with three existing orientation features: normalized gradient vector [23], spread orientation [8] and cumulative orientation feature (COF) [26]. Existing methods used the upper-left image in Fig. 2(b) as a model image.

Similarity scores based on four kinds of orientation features were calculated at every pixel on the vertical and perspective images. We show the differences between the maximum scores at the target object (FG: foreground) and at the backgrounds (BG) in Table 1. This difference represents how discriminative each feature is against cluttered backgrounds on the vertical image and is both against cluttered backgrounds and changes in 3D object pose on the perspective images. The larger the score difference is, the more discriminative the feature is. Regarding the differences on the perspective images, mean values are presented in Table 1.

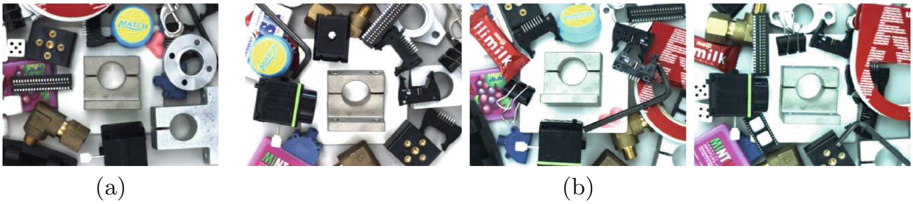


Fig. 5. (a) Vertical image and (b) three examples of perspective images for evaluation of the orientation features in experiment1. These images are almost identical except for the 3D pose of the target object (L-Holder at the center of the images).

Table 1. Differences between a maximum score at the target object and at the backgrounds on the vertical and perspective images in experiment1.

	Steger [23]	Spread [8]	COF [26]	**PCOF(Ours)**
Vertical	0.332	0.465	0.477	**0.485**
Perspective	0.214	0.421	0.403	**0.483**

Normalized Gradient Vector. Steger et al. [23] showed that the sum of inner products of normalized gradient vectors was occlusion, clutter and illumination invariant. Our experimental results in Table 1 showed that the differences between FG - BG scores both on the vertical and perspective images were much lower than other three features. This demonstrated that Steger's similarity score was fragile both to the background clutters and to the changes in 3D object pose.

Spread Orientation. Hinterstoisser et al. [8] introduced the spread orientation in order to make their similarity score robust to small shifts and deformations. They efficiently spread the quantized orientations of test images by shifting the orientation features over the range of ±4 × ±4 pixels and merging them with bitwise OR operations. In our experimental results in Table 1, the difference between FG - BG scores on the perspective images decreased from that on the vertical image. This indicated that the spread orientation was robust to cluttered backgrounds but not to the changes in 3D object pose.

Cumulative Orientation Feature (COF). Konishi et al. [26] introduced COF, which was robust both to cluttered backgrounds and the appearance changes caused by the changes in 2D object pose. Followoing their paper, we generated many images by transformimg the model image using randomized geometric transformation parameters (within the range of ±1 pixel in x-y translations, ±7.5 ° of in-plane rotation and ±5 % of scale). Then COF was calculated at each pixel by merging all the quantized orientations observed on generated images. The COF template was matched with the test images and the results were shown in Table 1. As with the spread orientation, the difference between FG - BG scores on the perspective images was decreased and COF was robust to cluttered backgrounds but not to the change in 3D object pose.

Perspectively Cumulated Orientation Feature (PCOF). PCOF was calculated as described in Sect. 3.1 and matched with the quantized orientations extracted on the test images. The difference between FG - BG scores in Table 1 were higher than other three features both on the vertical and perspective images, and the score difference was not decreased on the perspective images compared to that on the vertical image. This shows that PCOF was robust both to cluttered backgrounds and the changes in 3D object pose. Due to this robustness, the template which consist of PCOF can handle a certain range of

3D object pose (approximately 8 ° in out-of-plane rotation angles) without loss of the robustness to cluttered backgrounds. This advantage enables PCOF templates to handle a wider range of 3D object pose with fewer number of templates than other image features.

4.2 Experiment2: 6D Pose Estimation

Experimental Settings. In experiment2, we evaluated the accuracy and the speed of our 6D pose estimation algorithm on our texture-less object dataset. The dataset consists of nine mechanical parts which are texture-less and some of them have specular surfaces. These objects were captured from various viewpoints within the range of ±60 ° around x-y axes, ±180 ° around the optical axis and 660 mm − 800 mm in distance from the center of the object. The resolution of the camera was VGA (640×480) and approximately 500 images were taken per object where cluttered backgrounds and partial occlusions were contained. The ground truth of 6D object pose were estimated based on the surrounding AR

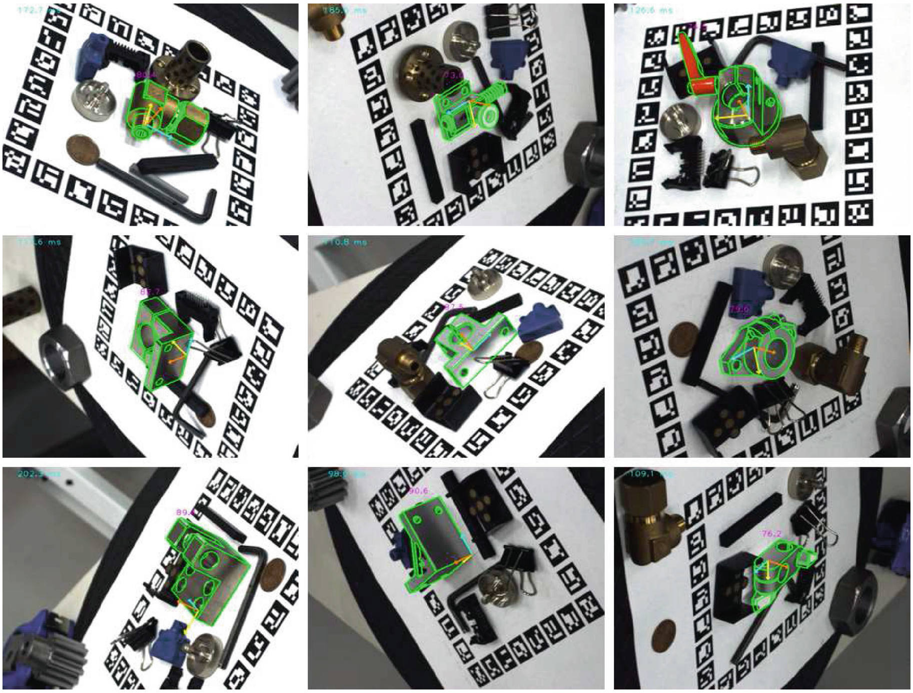


Fig. 6. The example images of our dataset are presented. The dataset consists of nine texture-less objects and contains cluttered backgrounds and partial occlusions. Top: Connector, SideClamp and Stopper. Middle: L-Holder, T-Holder and Flange. Bottom: HingeBase, Bracket and PoleClamp. The edges of the objects extracted from 3D CAD data (green lines) and the coordinate axes (three colored arrows) are drawn on the images based on the estimated 6D pose by our proposed method. (Color figure online)

markers printed on the board where the target objects were placed on. The AR markers were recognized using ArUco library [36]. We counted the estimated 6D pose as correct if the errors of the result were within 10 mm along x-y axes, 40 mm along z axis, 10 ° around x-y axes and 7.5 ° around z axis. The exmaple images of our dataset are shown in Fig. 6. The estimated results by our proposed method are drawn on the images.

The existing 6D pose estimation algorithms by Ulrich et al. [7], Hinterstoisser et al. (LINE-2D) [8] and Konishi et al. (COF) [26] were also evaluated on the dataset. We used the function "find_shape_model_3d" in the machine vision library "HALCON 11" (MvTEC in Germany) as an implementation of [7], LINE-2D implemented in OpenCV 2.4.11 and the source code of COF which was provided by the authors. We prepared 2D projection images from the same viewpoints as PCOF (total of 205 images per object) and used them for the training of LINE-2D and COF. All the programs were run on a PC (CPU: Core i7 3.4 GHz, OS: Windows7 64 bit) using a single CPU core.

Estimation Accuracy. Figure 7 shows the curves representing the relation between the success rate of correctly estimated 6D pose (vertical axis) and false positives per image (FPPI, horizontal axis). The estimation results with various search thresholds are plotted on the graphs. When the threshold is low and FPPI is high, the success rate for each object is less than 1. This is because 6D pose estimation requires not only correct positions but also correct rotation angles around x-y-z axes, and the estimated rotation angles do not depend on the search thresholds. All the graphs indicate that our proposed method achieves higher accuracy in comparison with other existing methods.

As shown in experiment1 (Sect. 4.1), COF and spread orientation of LINE-2D are not robust to the appearance changes caused by the out-of-plane rotations of the object. The numbers of viewpoints for making the templates are same in COF, LINE-2D and PCOF. Thus the differences in the success rate between these three methods are mainly due to the different image features.

In the algorithm of Ulrich et al. [7], the templates using normalized gradint vectors of Steger et al. [23] are made at the viewpoints sampled more densely than other three methods. Then the viewpoints are clustered to some aspects based on the similarity scores between the templates. Thus the viewpoint sphere is divided into some aspects which are optimized for a single template to keep its similarity score higher than a certain threshold. This viewpoint sampling is better than the regularly spaced sampling as in COF and LINE-2D, and the success rate of Ulrich et al. is higher than those of COF and LINE-2D. However, a single template represented each aspect and the similarity score should be degraded at the edges of the aspect. This is because our method surpass Ulrich et al. in the success rate of correctly estimated 6D object pose.

Processing Time. The processing times (ms) for 6D pose estimation when FPPI is 0.5 are shown in Table 2. Our proposed method achieved faster speed

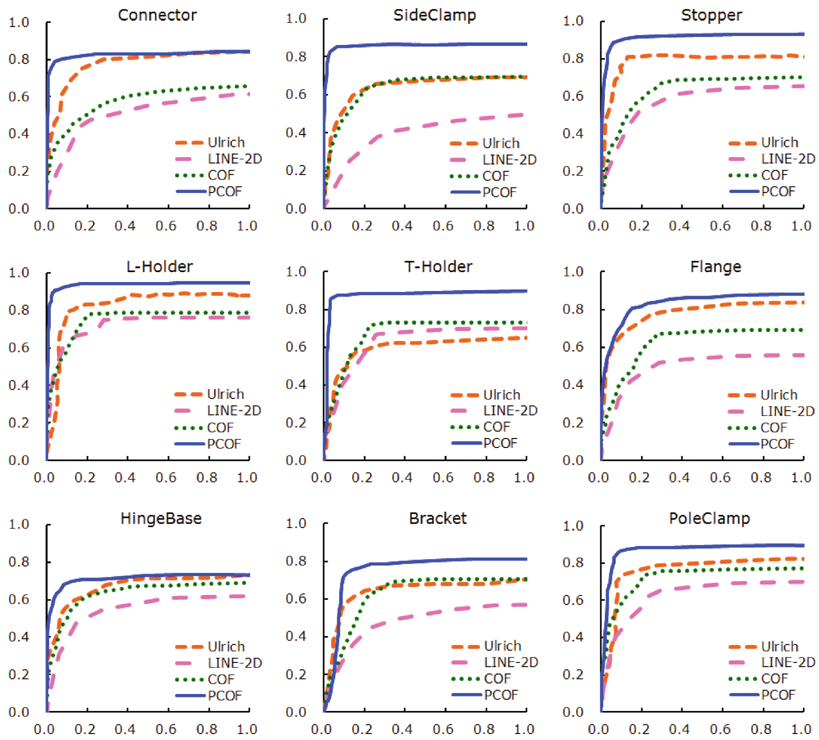


Fig. 7. The graphs showing the relation between the success rate of correctly estimated 6D pose (vertical axis) and false positives per image (FPPI, horizontal axis) are presented. There are nine graphs for each object in the dataset and the curves by four methods (Ulrich et al. [7], LINE-2D [8], COF [26] and PCOF (ours)) are drawn on each graph.

Table 2. The processing times (ms) for 6D pose estimation in experiment2 when FPPI is 0.5 are presented. The mean value is also shown at the bottom.

	Ulrich [7]	LINE-2D [8]	COF [26]	**PCOF (ours)**
Connector	964.1	375.8	1258.5	**167.1**
SideClamp	2724.4	383.2	1387.5	**220.4**
Stopper	2703.0	345.7	1149.9	**129.9**
L-Holder	963.8	357.1	1015.8	**122.6**
T-Holder	912.2	376.3	1140.1	**137.5**
Flange	973.0	390.5	1238.1	**137.4**
HingeBase	1137.1	348.9	1124.6	**226.1**
Bracket	792.4	358.5	961.4	**127.1**
PoleClamp	1439.0	375.9	1320.1	**137.4**
Mean	1401.0	368.0	1177.3	**156.2**

compared with the existing methods. PCOF and COF [26] use the same similarity scores calculated by bitwise ADD operations of binary features, and the main difference between them influencing the processing time is their search data structures. In COF the 2D object pose is estimated at each viewpoint independently, and the search strategy is optimized only in 2D pose space and not in 3D pose space. This is why the speed of COF was slower by approximately ten times than PCOF. The search model of LINE-2D [8] is also not efficient for search in 3D pose space. However, the similarity score of LINE-2D is calculated just by summing up the precomputed responce maps where the memory is linearized for reducing a cache miss, and this is much faster than the scores calculated by bitwise operations. Thus LINE-2D is much faster than COF.

Ulrich et al. [7] uses the normalized gradient vectors [23] which is not robust to the changes in 3D object pose, and their method requires more templates than PCOF in order to handle the same range of 3D object pose. Add to this, their search model is constructed by merging the neighboring viewpoints, and this is not fully efficient in the case that 2D views from separate viewpoints are similar, as is often the case with texture-less objects. Their similarity score which is based on floating-point arithmetic possibly lead to a slow matching of templates. From these reasons, 6D pose estimation of Ulrich et al. is slower by five to ten times than PCOF.

5 Conclusion

In this paper, we proposed PCOF and HPT for template based 6D pose estimation of texture-less objects from a monocular image. PCOF is extracted from randomly generated 2D projection images using 3D CAD data to explicitly handle a certain range of 3D object pose. HPT is built by clustering 3D object pose based on the similarities between 2D views and reducing the resolutions of PCOF features to accelerate 6D pose estimation using a coarse-to-fine search. The experimental evaluation demonstrated that PCOF was robust both to cluttered backgrounds and the appearance changes caused by the changes in 3D object pose. Another experimental result showed that our 6D pose estimation algorithm based on PCOF and HPT achieved higher success rate of correctly estimated 6D pose and faster speed in comparison with state-of-the-art methods on our challenging dataset.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
3. Grimson, W., Huttenlocher, D.: On the verification of hypothesized matches in model-based recognition. IEEE Trans. Pattern Anal. Mach. Intell. **13**(12), 1201–1213 (1991)

4. Lanser, S., Munkelt, O., Zierl, C.: Robust video-based object recognition using CAD models. In: Intelligent Autonomous Systems IAS-4, pp. 529–536 (1995)
5. Cyr, C.M., Kimia, B.B.: A similarity-based aspect-graph approach to 3D object recognition. Int. J. Comput. Vis. **57**(1), 5–22 (2004)
6. Liu, M.Y., Tuzel, O., Veeraraghavan, A., Taguchi, Y., Marks, T., Chellappa, R.: Fast object localization and pose estimation in heavy clutter for robotic bin picking. Int. J. Rob. Res. **31**(8), 951–973 (2012)
7. Ulrich, M., Wiedemann, C., Steger, C.: Combining scale-space and similarity-based aspect graphs for fast 3D object recognition. IEEE Trans. Pattern Anal. Mach. Intell. **34**(10), 1902–1914 (2012)
8. Hinterstoisser, S., Cagniart, C., Ilic, S., Sturm, P., Navab, N., Fua, P., Lepetit, V.: Gradient response maps for real-time detection of textureless objects. IEEE Trans. Pattern Anal. Mach. Intell. **34**(5), 876–888 (2012)
9. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 548–562. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37331-2_42
10. David, P., DeMenthon, D.: Object recognition in high clutter images using line features. In: CVPR, pp. 1581–1588 (2005)
11. Damen, D., Bunnun, P., Calway, A., Mayol-Cuevas, W.: Real-time learning and detection of 3D texture-less objects: a scalable approach. In: BMVC (2012)
12. Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model globally, match locally: efficient and robust 3D object recognition. In: CVPR, pp. 998–1005 (2010)
13. Rodrigues, J., Kim, J.S., Furukawa, M., Xavier, J., Aguiar, P., Kanade, T.: 6D pose estimation of textureless shiny objects using random ferns for bin-picking. In: IROS, pp. 3334–3341 (2012)
14. Tejani, A., Tang, D., Kouskouridas, R., Kim, T.-K.: Latent-class hough forests for 3D object detection and pose estimation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 462–477. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10599-4_30
15. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6D object pose estimation using 3D object coordinates. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 536–551. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10605-2_35
16. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3D pose estimation. In: CVPR, pp. 3109–3118 (2015)
17. Crivellaro, A., Rad, M., Verdie, Y., Yi, K.M., Fua, P., Lepetit, V.: A novel representation of parts for accurate 3D object detection and tracking in monocular images. In: ICCV, pp. 4391–4399 (2015)
18. Krull, A., Brachmann, E., Michel, F., Yang, M.Y., Gumhold, S., Rother, C.: Learning analysis-by-synthesis for 6D pose estimation in RGB-D images. In: ICCV, pp. 954–962 (2015)
19. Zhu, M., Derpanis, K., Yang, Y., Brahmbhatt, S., Zhang, M., Phillips, C., Lecce, M., Daniilidis, K.: Single image 3D object detection and pose estimation for grasping. In: ICRA, pp. 3936–3943 (2014)
20. Rios-Cabrera, R., Tuytelaars, T.: Discriminatively trained templates for 3D object detection: a real time scalable approach. In: ICCV, pp. 2048–2055 (2013)
21. Kehl, W., Tombari, F., Navab, N., Ilic, S., Lepetit, V.: Hashmod: a hashing method for scalable 3D object detection. In: BMVC (2015)

22. Hodan, T., Zabulis, X., Lourakis, M., Obdrzalek, S., Matas, J.: Detection and fine 3D pose estimation of texture-less objects in RGB-D images. In: IROS, pp. 4421–4428 (2015)
23. Steger, C.: Occlusion, clutter, and illumination invariant object recognition. In: International Archives of Photogrammetry and Remote Sensing, vol. XXXIV, Part 3A, pp. 345–350 (2002)
24. Ullah, F., Kaneko, S.: Using orientation codes for rotation-invariant template matching. Pattern Recogn. **37**(2), 201–209 (2004)
25. Hinterstoisser, S., Lepetit, V., Ilic, S., Fua, P., Navab, N.: Dominant orientation templates for real-time detection of texture-less objects. In: CVPR, pp. 2257–2264 (2010)
26. Konishi, Y., Ijiri, Y., Suwa, M., Kawade, M.: Textureless object detection using cumulative orientation feature. In: ICIP, pp. 1310–1313 (2015)
27. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR, pp. 2161–2168 (2006)
28. Silpa-Anan, C., Hartley, R.: Optimised KD-trees for fast image descriptor matching. In: CVPR, pp. 1–8 (2008)
29. Muja, M., Lowe, D.: Scalable nearest neighbor algorithms for high dimensional data. IEEE Trans. Pattern Anal. Mach. Intell. **36**(11), 2227–2240 (2014)
30. Lai, K., Bo, L., Ren, X., Fox, D.: A scalable tree-based approach for joint object and pose recognition. In: AAAI, pp. 1474–1480 (2011)
31. Gavrila, D.M.: A Bayesian, exemplar-based approach to hierarchical shape matching. IEEE Trans. Pattern Anal. Mach. Intell. **29**(8), 1408–1421 (2007)
32. Stenger, B., Thayananthan, A., Torr, P.H.S., Cipolla, R.: Hand pose estimation using hierarchical detection. In: Sebe, N., Lew, M., Huang, T.S. (eds.) CVHCI 2004. LNCS, vol. 3058, pp. 105–116. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24837-8_11
33. Borgefors, G.: Hierarchical chamfer matching: a parametric edge matching algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **10**(6), 849–865 (1988)
34. Pelleg, D., Moore, A.: X-means: extending k-means with efficient estimation of the number of clusters. In: ICML, pp. 727–734 (2000)
35. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004). ISBN: 0521540518
36. Garrido-Jurado, S., Muñoz Salinas, R., Madrid-Cuevas, F.J., Marín-Jiménez, M.J.: Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recogn. **47**(6), 2280–2292 (2014)