MVA2015 IAPR International Conference on Machine Vision Applications, May 18-22, 2015, Tokyo, JAPAN

# Pose estimation of textureless objects in cluttered environments

Blaž Bratanič[1], Boštjan Likar[1,2], Franjo Pernuš[1,2], Dejan Tomaževič[1,2]
[1]Sensum, Computer Vision Systems, Tehnološki park 18, 1000 Ljubljana, Slovenia
[2]University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, 1000 Ljubljana Slovenia
blaz.bratanic@sensum.eu

## Abstract

*In this paper, we propose a method for pose estimation of multiple (textureless) objects of the same type in heavily cluttered environments. The method as such could be used as a component for building more flexible automated visual inspection systems, removing the need for precise mechanical manipulation and enable inspection in settings previously thought unfeasible. The method consists of three phases. In the first phase, template matching is used to calculate similarity measure maps for different object poses. Template matching combines both edge and surface normal information to improve the pose estimation accuracy. Second, a large number of pose hypotheses are generated with a non-parametric clustering of the similarity measure maps and finally, the best hypotheses are iteratively selected. Method was evaluated and compared with the current state-of-the-art on two synthetic and one real-world datasets. The results show that the proposed system performs better than the current state-of-the art for pose estimation in industrial environments.*

## 1 Introduction

Automated visual inspection (AVI) systems are often crucial for detection of defects in manufactured products, or for diagnosing problems in the manufacturing process. Currently, most of AVI systems are highly specialized. Nearly all of them, with a few exceptions, have been designed to perform inspection of a single object whose position and pose are highly constrained. Positioning is usually achieved by precise mechanical manipulation, which can be expensive, space/time consuming, or requires additional customization for each specific part [1].

Visual inspection systems that could inspect randomly positioned and rotated objects, would remove the need for precise mechanical manipulation. This would consequently reduce the cost of the visual inspection systems and enable inspection in scenarios that were previously thought unfeasible. One of the challenges to achieve this goal is to provide the visual inspection system with the ability to detect and estimate the poses of the inspected objects in an unconstrained environment (Figure 1).

In this paper, we address the pose estimation of multiple (textureless) objects of the same type for automated visual inspection. This is a challenging task, because objects are randomly positioned, rotated, and variously occluded. Several methods have been proposed for pose estimation; however, most of the proposed methods detect only the best matching object. Methods for pose estimation of (textureless) objects can be classified into model-based and shape-based methods. Model-based methods directly match the 3D CAD model to the 2D image by determining correspondences between the two [2, 3, 4]. Establishing correspondences between a 2D image and a 3D model is a difficult task; consequently, model-based methods are usually complemented with a 3D information.

Shape based methods match reference shapes with the shapes on the image. Object's position and its pose can be determined by matching the shapes of various views of the reference objects with the image. Shape-based methods could further be split into descriptor-based methods that allow utilizing a standard descriptor-based pipeline [5] and into methods that use exhaustive search to match the shape with the image (template-matching); usually utilizing a sliding window. Several meaningful descriptors for textureless objects have been proposed. BOLD features [6] tackle with textureless objects with a compact and distinct representation of groups of neighboring line segments aggregated over limited spatial supports. Damen et al. [7] proposed a pose estimation framework based on edge constellations combined with a library lookup, where edge constellations were extracted with a path-tracing algorithm. In [8], Ferrari et al. proposed a family of scale-invariant descriptors utilized in a shape-matching framework, through a voting scheme in Hough space.

Template matching is one of the more perspective approaches to (textureless) object detection and pose estimation in cluttered environments. Although proposed decades ago, Chamfer matching [9] remains the preferred method when simplicity is required; however, the naïve approach has a high computational complexity that makes it unfeasible for real-time applications. Liu et al. [10] extended Chamfer matching with edge orientations and replaced an exhaustive search, with a 1D search along distinctive lines, leading to drastic improvements in both pose estimation accuracy and speed. In [11], shape descriptors were utilized to avoid multiple sliding window passes over the query image. Instead, a shape-matching framework was used to match distinct representation of the shape in the sliding window. All the aforementioned methods operate on a binary-edge image obtained by one of the

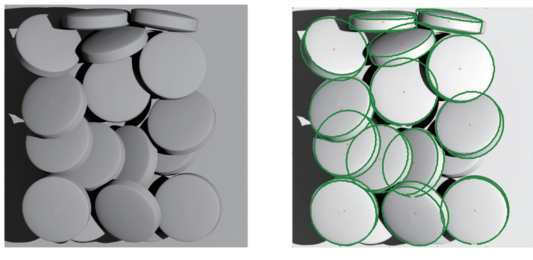


Figure 1: Pose estimation in heavily cluttered environment.

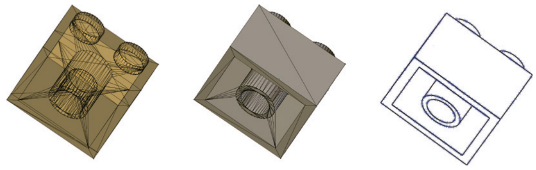


Figure 2: Extracted edge orientations from a 3D CAD model.

edge extraction methods. Steger [12] proposed several similarity measures inherently robust to illumination changes that do not require a binary-edge image. Hinterstoisser et al. [13] improved the similarity measure proposed by Steger. They proposed a similarity measure robust to small translations and deformations and presented a highly optimized procedure for its evaluation.

# 2 Method

The proposed method consists of three steps. In the first step, template matching [13] is used to calculate similarity measure maps for various object poses. Template matching combines both edge and surface normal information - obtained with photometric stereo - to improve the pose estimation accuracy. Second, a large number of pose hypotheses are generated with a nonparametric clustering of the similarity measure maps and finally, the best pose hypotheses are iteratively selected.

## 2.1 Template matching

Hinterstoisser et al. proposed a similarity measure robust to small translations and deformations:

$$\mathcal{E}(\mathcal{I},\mathcal{T},c) = \sum_{r\in\mathcal{P}}\Big(\max_{t\in\mathcal{R}(c+r)} |\cos(ori(\mathcal{O},r) - ori(\mathcal{I},t))|\Big), \tag{1}$$

where $ori(\mathcal{O},r)$ is the orientation on the reference image $\mathcal{O}$ at positions $r \in \mathcal{P}$ and $\mathcal{P}$ is a list of all positions to be considered on the reference image. The template $\mathcal{T}$ is then defined as a pair $(\mathcal{P},\mathcal{O})$. Similarly, $ori(\mathcal{I},r)$ is the orientation on the query image $\mathcal{I}$ at location $t \in \mathcal{R}(c+r)$ where $\mathcal{R}(c+r) = [c+r-\frac{T}{2}, c+r+\frac{T}{2}] \times [c+r-\frac{T}{2}, c+r+\frac{T}{2}]$ defines a neighborhood of size $T$ centered at the current position $c$ on the query image shifted by $r$.

Their proposed similarity measure is not limited to edge orientations and can be used with any modality. In the proposed method, we combine edge orientations with the surface normal orientations obtained with a photometric stereo. Similarity measure, matching surface normal orientations with the object model, is evaluated separately for orientations along x- and y-axis. Evaluated similarity measures for all modalities are in the end combined with a multiplication.

## 2.2 Template generation

To use template matching for pose estimation, a reference template is required for each pose. Usually, templates are generated for view directions sampled uniformly on a sphere. However, for some object shapes, e.g. for rotationally symmetric objects, the number of views can be significantly reduced.

Templates for edge orientations are generated directly from a 3D CAD model. Starting from a polygonal mesh, we first identify visible edges at selected view (Figure 2). Each visible edge is either:

- a sharp edge between two polygons with very different orientations or
- a boundary silhouette line of the object.

Template for edge orientations is then generated by sampling the points on the visible edges and extracting their orientations. Points should be sampled uniformly over the object to improve robustness to partial occlusions.

Templates for surface normal orientations are generated by identifying visible surfaces and then uniformly sampling the points and their orientations on the visible surfaces.

## 2.3 Query image preprocessing

**Gradient orientation maps** were computed from the captured images by first extracting the gradient orientation separately for each image and then for each location $\mathbf{c}$ select the orientation of the image with the highest gradient magnitude. Same as was done in [13], we mask the gradient orientations at locations, where the gradient magnitude is below a certain threshold.

**Surface orientation maps** were computed with photometric stereo [14] that can produce a dense normal field at a very high level of detail, given an assumption of Lambertian scene. Similarly as above, we mask the orientations that are impossible given the illumination setup.

## 2.4 Pose hypotheses generation

To generate a number of pose hypotheses, the evaluated similarity measure maps $\{\mathcal{E}(\mathcal{I},\mathcal{T}_i)\}$ for all the shape templates $\{\mathcal{T}_i\}, i = 1,\ldots,N_t$ are first combined together. The combined similarity measure map $\widetilde{\mathcal{E}}$ is calculated by:

$$\widetilde{\mathcal{E}}(\mathbf{c}) = \max_{i=1,\ldots,N_t} \mathcal{E}(\mathcal{I},\mathcal{T}_i,\mathbf{c}), \tag{2}$$

For each location $\mathbf{c}$ we also store the index of the template with the highest similarity measure.

On $\widetilde{\mathcal{E}}$ we detect clusters with a modified mean shift clustering algorithm proposed by Derganc et al. [15]. They modified the update step in the direction of the weighted mean of the density in the kernel with the step to the maximum value in the kernel:

$$\mu = \arg\max_{\mathbf{c}\in K} \widetilde{\mathcal{E}}(\mathbf{c}) - \mathbf{c} \tag{3}$$

where $\mu$ is the update step from the current position $\mathbf{c}$ and $K$ denotes the kernel. Each detected cluster is a pose hypothesis with a corresponding similarity score $s$, position and a template index (which also determines the object's rotation).

Pose hypotheses are then evaluated in an iterative fashion. At each iteration, the hypothesis with the highest similarity score $s_{max}$ is selected. Similarity

(a) (b) (c)

Figure 3: Example images of synthetic datasets (a, b) and an example image of real-world dataset (c).

score is again evaluated on the updated query image $\mathcal{I}$. If the re-evaluated similarity score equals $s_{max}$ the pose hypothesis is selected. In addition the query image $\mathcal{I}$ is updated, by masking the area of the selected pose hypothesis. On the other hand, if the re-evaluated similarity score differs from $s_{max}$ the iteration starts over. We observed that the non-occluded objects have a higher similarity score; therefore, will be selected first.

## 3 Experiments

We conducted evaluations on several synthetic datasets and one real-world dataset. The method was compared to the method proposed by Liu et al. [10] (FDCM), which is state-of-the-art for pose estimation in industrial environments. Since there is no simple way of obtaining ground truth data for heavily cluttered dynamic environments, we performed quantitative evaluation only on synthetic datasets. For real-world dataset, we performed a qualitative evaluation, where the correctly estimated poses were determined visually.

### 3.1 Datasets

Synthetic datasets were generated with Blender and rendered with its Cycles rendering engine. The first dataset (Figure 3a) represents shapes that are common in the pharmaceutical domain and the second synthetic dataset (Figure 3b) represents a more general shape. For capturing real images (Figure 3c) we used a setup consisting of a single industrial camera with telecentric lens and the lights directed $\sim 35°$ from the view direction.

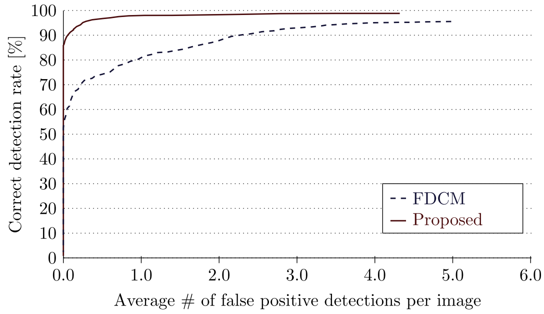
### 3.2 Implementation details

The method was implemented in C++ and optimized for CPU operation. Processing time for one $1000 \times 1000$ image was $\sim 100ms$, with the majority of time spent on the evaluation of the similarity measure. For the FDCM [10] method we used the publicly available implementation found on the author's website.

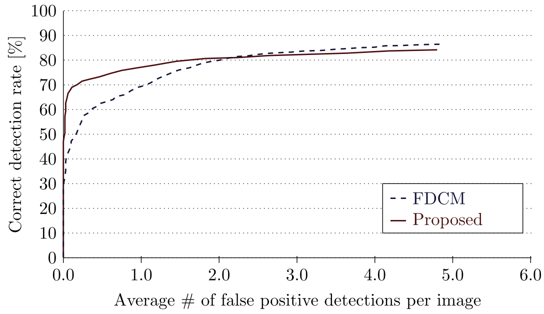
### 3.3 Evaluation on synthetic datasets

For quantitative evaluation, we use the criterion proposed in [13]. With the model $\mathcal{M}$, ground truth translation $T$, and rotation $R$ we can calculate the average error $e$ of the estimated rotation $\tilde{R}$ and translation $\tilde{T}$ as:

$$e = avg_{x \in \mathcal{M}} ||(Rx + T) - (\tilde{R}x + \tilde{T})||. \quad (4)$$




(a)




(b)

Figure 4: Comparison of the proposed approach with the method proposed by Liu et al. [10]. Figures (a) and (b) correspond to the first and second synthetic dataset, respectively.

For symmetric objects, the matching score was calculated as:

$$e = avg_{x_1 \in \mathcal{M}} \min_{x_2 \in \mathcal{M}} ||(Rx_1 + T) - (\tilde{R}x_2 + \tilde{T})||. \quad (5)$$

In our case, the detection was successful if the error measure $e < 0.05d$, where $d$ is a diameter of $\mathcal{M}$. Moreover, if less than 30% of the object surface is visible, it does not count towards false negative detections.

### 3.4 Evaluation on real-world datasets

On the real-world dataset, we performed a qualitative evaluation. For each estimated pose, a model with the corresponding pose was overlaid on the image. To obtain unbiased results, all poses were evaluated with a standalone application, where the method with which the pose was estimated was unknown.

### 3.5 Results

Results for quantitative evaluation and comparison using two synthetic datasets are presented in figure 4. Results for qualitative evaluation and comparison using the real-world dataset are presented in figure 5.

## 4 Discussion

Method was evaluated and compared with the current state-of-the-art on two synthetic and one real-world datasets. Pose detection in a heavily cluttered environment is a difficult task due to shadows, occlusions, self-occlusions and many ambiguities occurring due to neighboring objects. We show that a high pose

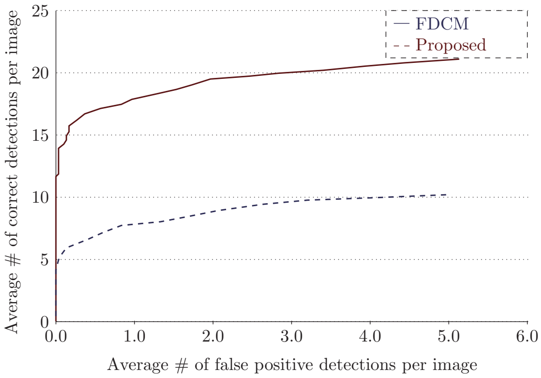



Figure 5: Comparison of the proposed approach with the method proposed by Liu et al. [10](FDCM) on the real-world dataset.

estimation accuracy is achievable, even in heavily cluttered environments, using only a single industrial camera. For (partly) diffuse surfaces, the precision can be improved by incorporating surface normal information. We use photometric stereo, which produces dense surface normal information unachievable by other triangulation methods, thus it is well suited for small detailed objects.

The experiments using several synthetic and one real dataset show superior detection rate comparing to the method proposed by Liu et al. [10]. The improvement is more prominent on simple objects where edge orientation alone is often not sufficient for pose disambiguation (first synthetic dataset and the real-world dataset). On synthetic datasets, we use surface normal information in both x- and y-axis, but for the real dataset, only the surface normals along a single axis are available, because we used a two-light setup.

The pose hypotheses were generated from a combined similarity measure map, due to the computational complexity of the clustering algorithm. Performing clustering separately on each evaluated similarity measure map would have several advantages; for example, detecting two objects with the same location but different poses. Therefore, finding an algorithm for more efficient non-parametric clustering will be a part of our future work.

Poses were selected by a simple iterative procedure, based on the assumption that non-occluded objects have a higher similarity score; therefore, will be selected first. Experiments have shown that iterative selection performs better than just selecting the best $N$ clusters. We assume this is, because neighboring objects can significantly affect the detected position and the pose.

## 5 Conclusion

In this paper, we propose a method for pose estimation of multiple (textureless) objects of the same type in heavily cluttered environments. The method as such could be used as a component for building more flexible automated visual inspection systems, removing the need for precise mechanical manipulation and enable inspection in settings previously thought unfeasible.

Method was evaluated and compared with the current state-of-the-art on two synthetic and one real-world dataset. The results show that the proposed method performs better than current state-of-the-art for pose estimation in industrial environments. However, by using photometric stereo we assumed that object surfaces are at least partly diffuse. Future work will be directed to extending the method to arbitrary materials, utilizing specular cues.

## References


[1] T. S. Newman and A. K. Jain, "A survey of automated visual inspection," *Computer vision and image understanding*, vol. 61, no. 2, pp. 231–262, 1995.

[2] P. David and D. DeMenthon, "Object recognition in high clutter images using line features," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, pp. 1581–1588, IEEE, 2005.

[3] D. G. Lowe, "Three-dimensional object recognition from single two-dimensional images," *Artificial intelligence*, vol. 31, no. 3, pp. 355–395, 1987.

[4] A. Chia, S. Rahardja, D. Rajan, and K. Leung, "Object recognition by discriminative combinations of line segments and ellipses," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2225–2232, IEEE, 2010.

[5] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1615–1630, 2005.

[6] F. Tombari, A. Franchi, D. Automation, and L. Di Stefano, "Bold features to detect texture-less objects,"

[7] D. Damen, P. Bunnun, A. Calway, and W. W. Mayol-Cuevas, "Real-time learning and detection of 3d textureless objects: A scalable approach.," in *BMVC*, pp. 1–12, 2012.

[8] V. Ferrari, F. Jurie, and C. Schmid, "From images to shape models for object detection," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 284–303, 2010.

[9] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," tech. rep., DTIC Document, 1977.

[10] M.-Y. Liu, O. Tuzel, A. Veeraraghavan, R. Chellappa, A. Agrawal, and H. Okuda, "Pose estimation in heavy clutter using a multi-flash camera," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 2028–2035, IEEE, 2010.

[11] H. Cai, T. Werner, and J. Matas, "Fast detection of multiple textureless 3-d objects," in *Computer Vision Systems*, pp. 103–112, Springer, 2013.

[12] C. Steger, "Similarity measures for occlusion, clutter, and illumination invariant object recognition," in *Pattern Recognition*, pp. 148–154, Springer, 2001.

[13] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit, "Gradient response maps for real-time detection of textureless objects," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 5, pp. 876–888, 2012.

[14] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical engineering*, vol. 19, no. 1, pp. 191139–191139, 1980.

[15] J. Derganc, B. Likar, and F. Pernuš, "Shading correction and segmentation of color images," in *Image and Signal Processing and Analysis, 2001. ISPA 2001. Proceedings of the 2nd International Symposium on*, pp. 345–350, IEEE, 2001.