Computer Vision and Image Understanding 141 (2015) 38–51

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Real-time pose estimation of rigid objects in heavily cluttered environments☆

Blaž Bratanič[a,*], Franjo Pernuš[a,b], Boštjan Likar[a,b], Dejan Tomaževič[a,b]

[a] Sensum, Computer Vision Systems, Tehnološki park 21, Ljubljana 1000, Slovenia
[b] Laboratory of Imaging Technologies, Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, Ljubljana 1000, Slovenia

ARTICLE INFO

Article history:
Received 13 March 2015
Accepted 6 September 2015
Available online 11 September 2015

Keywords:
object pose estimation
textureless surfaces

ABSTRACT

In this paper, we present a method for real-time pose estimation of rigid objects in heavily cluttered environments. At its core, the method relies on the template matching method proposed by Hinterstoisser et al., which is used to generate pose hypotheses. We improved the method by introducing a compensation for bias toward simple shapes and by changing the way modalities such as edges and surface normals are combined. Additionally, we incorporated surface normals obtained with photometric stereo that can produce a dense normal field at a very high level of detail. An iterative algorithm was employed to select the best pose hypotheses among the possible candidates provided by template matching. An evaluation of the pose estimation reliability and a comparison with the current state-of-the-art was performed on several synthetic and several real datasets. The results indicate that the proposed improvements to the similarity measure and the incorporation of surface normals obtained with photometric stereo significantly improve the pose estimation reliability.

1. Introduction

Visual inspection systems are often crucial for the detection of defects in manufactured products and for diagnosing problems in a manufacturing process. Currently, one characteristic of automated visual inspection systems is their specialization. With few exceptions, nearly all of the existing automated visual inspection systems have been designed to inspect a single object or a part of one whose position is highly constrained [1]. Positioning is usually achieved by a mechanical manipulation of the object, which can be expensive, space and/or time consuming, or simply impossible in some scenarios. A visual inspection system that could inspect arbitrarily positioned objects would bypass the need for mechanical manipulation of the inspected parts, thus reducing the cost of the system and increasing its flexibility. It would also enable inspection in scenarios that were previously thought unfeasible.

For visual inspection in mechanically unconstrained environments, the system needs to detect and localize the objects, before the inspection can take place; this can be challenging because of variable object appearance, changes in the environment, mechanical properties, and background clutter. On the other hand, the fact that most industrial lines produce millions of similar products provides several opportunities to exploit. For example, 3D models are in most cases readily available, and even if they are not, the fact that the assembly lines handle millions of objects, makes an acquisition of such a model economically justifiable.

In this paper, we present a method for real-time pose estimation in heavily cluttered environments. The method estimates the poses of multiple objects of the same type that are arbitrarily positioned in an image acquired by a single camera. We present three ideas that allow us to achieve a fast, reliable, and accurate operation: an improvement of the template matching algorithm proposed by Hinterstoisser et al. [2], a combining of the orientation of depth edges with the surface normal orientation obtained by photometric stereo, and an iterative procedure for selecting pose hypotheses.

The primary goal of this research was to develop a robust and reliable system for real-time pose estimation in cluttered environments that could be used for on-line analysis of objects in mechanically unconstrained environments (Fig. 1).

The paper is organized as follows. Brief overview of the related literature is presented in Section 2. Template matching is described in Section 3, followed by a description of hypotheses generation and selection in Section 4. An evaluation and the comparison to the existing state-of-the-art are in Section 5. Section 6 contains a discussion of the results, followed by the paper's conclusion in Section 7.

☆ This paper has been recommended for acceptance by Nikos Paragios.

* Corresponding author.
E-mail address: blaz.bratanic@sensum.eu (B. Bratanič).

http://dx.doi.org/10.1016/j.cviu.2015.09.002
1077-3142/

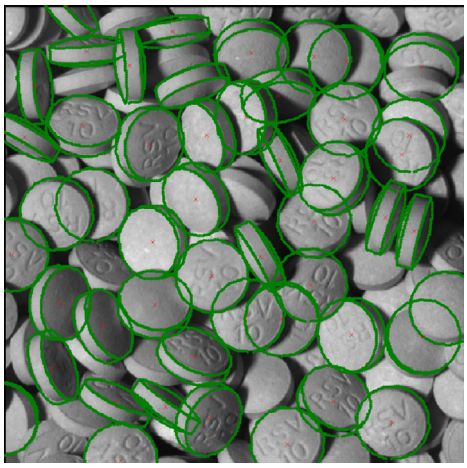

Fig. 1. Pose estimation in heavily cluttered environment.

2. Related work

A variety of approaches has been proposed for (textureless) pose estimation. Early attempts to estimate a 3D pose from a 2D image were made by Lowe [3], where a 3D wireframe model was matched to a 2D image, by establishing the correspondences between the distinctive lines on the model and an image. Correspondences were iteratively established by matching nearest lines on the model and an image. Similar approach with a modified optimization step was proposed by Dementhon and Davis [4] and subsequently by David and DeMenthon [5]. However, establishing correspondences by nearest neighbor search requires a starting pose close to the correct pose, therefore the applicability of the aforementioned methods is limited.

To circumvent this requirement, several descriptor-based methods have been proposed, which match the representative description [6,7] of the local regions on both the query and the reference image. For pose estimation, reference images are generated for all relevant views, thus the established correspondences determine both the position and the rotation of the object. Descriptor-based approaches demonstrate outstanding performance in several scenarios [8] but are generally restricted to textured objects. It is only recently that several attempts have been made to extract meaningful descriptors from textureless objects [9–12]. BOLD features [9] tackle textureless objects with a compact and distinct representation of groups of neighboring line segments aggregated over limited spatial supports. Damen et al. [12] combined an extraction of edgelet constellations with a library lookup based on rotation- and scale-invariant descriptors. Their distinguishing element was using path tracing for extracting edgelet constellations. Ferrari et al. [11] proposed a family of scale-invariant shape descriptors utilized in a shape-matching framework through a voting scheme in a Hough space.

Template matching methods, on the other hand, are usually based on exhaustive search and typically use *a priori* edge information for matching. Chamfer matching [13] was proposed decades ago for matching shape templates, and it remains the preferred method when simplicity is required. Nevertheless, it has a high computational complexity, which makes the naïve approach unfeasible for real-time applications. Borgefors [14] proposed a hierarchical Chamfer matching that significantly reduced the computational load. In Liu et al. [15,16], Chamfer matching was extended to include edge orientations and exhaustive search replaced with a 1D search along distinctive lines, leading to drastic improvements in both speed and accuracy. In Lampert et al. [17], exhaustive search was replaced by a branch and bound search strategy, significantly reducing the computational load. Choi and Christensen [18] combined the detection and tracking of textureless transparent objects within a particle-filtering framework, nonetheless still using Chamfer matching to find the set of starting states. Cai et al. [19] matched a compact representation of a shape in a sliding window with a library of reference representations. They extracted a compact representation at each position of a sliding window by finding the distance from uniformly positioned points in the window to their nearest edge point. All the aforementioned methods rely on binary edge images obtained by edge extraction methods [20–22].

Steger [23,24] proposed several similarity measures, for template matching, inherently robust against occlusion, clutter, and nonlinear illumination changes. One such measure was to sum the normalized dot product of the direction vectors of a transformed model and an image over all the points of the model; the measure returns a high score if all the direction vectors of the model and the query image align. To make the similarity measure truly illumination invariant, Steger discarded the magnitude of the orientation vectors and retained only the orientation difference. A comparable similarity measure was proposed by Hinterstoisser et al. [25] but with additional robustness to small translations and deformations. In addition, they extended the applicability of the method to arbitrary quantizable modalities and any combination thereof.

3. Template matching

The proposed method is based on the template matching proposed by Hinterstoisser et al. [25], which can match any modalities that can be quantized. Template matching is used to evaluate the similarity measure for all the reference templates and to generate probability maps for all the reference templates. We use edge orientation combined with surface normal orientation obtained with photometric stereo. In this section, we first briefly describe the method as proposed by Hinterstoisser et al. and then propose several improvements to it.

3.1. Similarity measure

Hinterstoisser et al. [25] proposed a similarity measure robust to small translations and deformations. The similarity measure $\mathcal{E}$ can be formalized as:

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

where $ori(\mathcal{O},r)$ is the orientation in radians on reference image $\mathcal{O}$ at location r and $\mathcal{P}$ is a list of locations r to be considered in $\mathcal{O}$. Template $\mathcal{T}$ is therefore defined as a pair $\mathcal{T}=(\mathcal{O},\mathcal{P})$ – a reference image $\mathcal{O}$ and a corresponding list of locations $\mathcal{P}$. Similarly, $ori(\mathcal{I},t)$ is the orientation on 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 r.

Hinterstoisser et al. describe a computationally efficient way for the evaluation of the aforementioned similarity measure. In short, the orientations on the query image are quantized and encoded in a binary representation, each bit representing an individual quantized orientation. This way each pixel can contain multiple (or no) orientations. Quantized orientations are "spread" around their original positions by shifting the bitwise representation of the orientation over the $\mathcal{R}$ neighborhood and merging all the shifted orientations with the bitwise OR operator. From the quantized and "spread" query image, n response maps are precomputed. Each response map contains an evaluated similarity measure (Eq. 1) where the orientation of template $ori(\mathcal{O},r)$ is replaced with an orientation corresponding to each of the n bins. With the precomputed response maps, a template $\mathcal{T}$ can be matched against the whole query image $\mathcal{I}$, simply by summing the (shifted) response maps.

We propose a modified similarity measure:

$$\mathcal{E}(\mathcal{I},\mathcal{T},c)=\sum_{r\in\mathcal{P}}\max_{t\in\mathcal{R}(c+r)}|\cos\left(ori(\mathcal{O},r)-ori(\mathcal{I},t)\right)|^{p}, \tag{2}$$