On Evaluation of 6D Object Pose Estimation

Tomáš Hodaň, Jiří Matas, Štěpán Obdržálek

Center for Machine Perception, Czech Technical University in Prague

Abstract. A pose of a rigid object has 6 degrees of freedom and its full knowledge is required in many robotic and scene understanding applications. Evaluation of 6D object pose estimates is not straightforward. Object pose may be ambiguous due to object symmetries and occlusions, *i.e.* there can be multiple object poses that are indistinguishable in the given image and should be therefore treated as equivalent. The paper defines 6D object pose estimation problems, proposes an evaluation methodology and introduces three new pose error functions that deal with pose ambiguity. The new error functions are compared with functions commonly used in the literature and shown to remove certain types of non-intuitive outcomes. Evaluation tools are provided at: `https://github.com/thodan/obj_pose_eval`

1 Introduction

Object localization and detection are among the core problems of computer vision. Traditional methods work with 2D images and typically describe pose of the detected object by a bounding box, which encodes 2D translation and scale [1,2]. There is no information about the object orientation and only a rough notion of the object distance. A pose of a rigid object has 6 degrees of freedom, 3 in translation and 3 in rotation, and its full knowledge is required in many robotic and scene understanding applications.

Although methods trying to extract a richer pose description from 2D images exist [3,4], the task can be simplified if depth images are used as additional input data. RGB-D – aligned color and depth – images which concurrently capture appearance and geometry of the scene can be obtained by *e.g.* Kinect-like sensors that are common in robotic applications.

Evaluation of 6D object pose estimates is not straightforward. Object pose can be ambiguous due to object symmetries and occlusions, *i.e.* there can be multiple object poses that are indistinguishable in the given image and should be therefore treated as equivalent (Fig. 1). This issue has been out of focus in the work on 6D object pose estimation. In evaluation of pose estimates described by 2D bounding boxes, the indistinguishable poses are treated as equivalent implicitly since all are described by the same bounding box.

(a) (b) (c) (d) (e)

Fig. 1. Different poses of the cup (a-b) cannot be distinguished if the handle is not visible due to self-occlusion (c). Pose of the pen (d) is ambiguous if its discriminative ends are occluded by another objects (e).

The main contribution of this paper are three new functions to measure error of an estimated 6D object pose w.r.t. the ground truth 6D object pose. All three are invariant under pose ambiguity, *i.e.* they treat the indistinguishable poses as approximately equivalent. The Visible Surface Discrepancy (e_{VSD}) measures misalignment over the visible surface of the object model and is thus inherently invariant under pose ambiguity. The Average and the Maximum Corresponding Point Distance (e_{ACPD}, e_{MCPD}) measure misalignment over the entire model surface when considering all indistinguishable poses, which are assumed known.

We define two 6D object pose estimation problems in Sec. 2, propose an evaluation methodology in Sec. 3, review the commonly used pose error functions and introduce the new functions in Sec. 4, present experimental comparison of the pose error functions in Sec. 5, and conclude in Sec. 6.

2 6D Object Pose Estimation Problems

A 6D object pose estimator is assumed to report its predictions on the basis of two sources of information. First, at training time, it is provided with a training set $T = \{T_1, T_2, \ldots, T_n\}$ for a set of rigid objects represented by identifiers $O = \{1, 2, \ldots, n\}$. The training data T_i may have different forms, *e.g.* a 3D object model or a set of RGB or RGB-D images, where each image shows one object instance in a known 6D pose. Second, at test time, it is provided with a single test RGB or RGB-D image I, which might be accompanied with information about objects that are visible in the image. The goal is to estimate a single 6D pose for each visible object instance.

Prior information about the object presence in I distinguishes two problems:

6D Localization Problem

Training input: A training set T, as described above.

Test input: An image I and a multiset $L_I = \{o_1, o_2, \ldots, o_k\}$, where $o_i \in O$ are identifiers of the objects present in I. Note: Multiple instances of an object may be present in I, *i.e.* the same identifier

may be multiple times in L_I.

Test output: A sequence $E_I = ((o_1, \hat{\mathbf{P}}_1, s_1), (o_2, \hat{\mathbf{P}}_2, s_2), \ldots, (o_k, \hat{\mathbf{P}}_k, s_k))$, where $\hat{\mathbf{P}}_i$ is an estimated 6D pose of an instance of object $o_i \in O$ with confidence $s_i \in (0, 1]$. Note: $|E_I| = |L_I|$, the size of the output is fixed by the input.

6D Detection Problem

Training input: A training set T, as described above.

Test input: An image I. No prior information about the object presence is provided, there may be $j \geq 0$ instances of each object $o \in O$.

Test output: A sequence $E_I = ((o_1, \hat{\mathbf{P}}_1, s_1), (o_2, \hat{\mathbf{P}}_2, s_2), \ldots, (o_m, \hat{\mathbf{P}}_m, s_m))$, where $\hat{\mathbf{P}}_i$ is an estimated 6D pose of an instance of object $o_i \in O$ with confidence $s_i \in (0, 1]$. Note: The size of the output $|E_I|$ depends on the estimator.

The 6D localization problem is a generalization of the problem defined by Hinterstoisser et al. [5], where the goal is to detect a single instance of a given object per image, *i.e.* $|L_I| = 1$.

In evaluation of the 6D localization problem, if there are for some object more estimated poses than the specified number j of instances, which is given by L_I, only j estimated poses with the highest confidence s are considered.

3 Evaluation Methodology

We propose the following methodology to evaluate performance of a 6D object pose estimator in the problems defined above. It includes an algorithm that determines the estimated poses that are considered correct (Sec. 3.1), a definition of pose error functions (described later in Sec. 4), and a definition of performance scores (Sec. 3.2).

In this paper, a pose of a rigid 3D object is represented by a 4×4 matrix $\mathbf{P} = [\mathbf{R}, \mathbf{t}; \mathbf{0}, 1]$, where $\mathbf{R}$ is a 3×3 rotation matrix, and $\mathbf{t}$ is a 3×1 translation vector. An object is represented by a model $\mathcal{M}$, which is typically a mesh given by a set of points in $\mathbb{R}^3$ and a set of triangles. Matrix $\mathbf{P}$ transforms a 3D point $\mathbf{x}_m$ in the model coordinate system to a 3D point $\mathbf{x}_c$ in the camera coordinate system: $\mathbf{x}_c = \mathbf{P}\mathbf{x}_m$. The 3D points are represented in homogeneous coordinates.

3.1 Determination of Pose Correctness

For each test image I, there is a ground truth set $G_I = \{(o_1, \bar{\mathbf{P}}_1), (o_2, \bar{\mathbf{P}}_2), \ldots, (o_k, \bar{\mathbf{P}}_k)\}$, where $\bar{\mathbf{P}}_i$ is the ground truth pose of an instance of object $o_i \in O$. Determination of estimated poses that are considered correct is formulated as finding a maximal matching in a bipartite graph $B = ((E_I, G_I), F)$, where F is

a set of edges that connect the ground truth poses G_I with matchable estimated poses E_I.

An estimated pose $(o, \hat{\mathbf{P}}, s) \in E_I$ is considered matchable with a ground truth pose $(o', \bar{\mathbf{P}}) \in G_I$, if it satisfies the necessary matching condition: $o = o' \wedge e(\hat{\mathbf{P}}, \bar{\mathbf{P}}; \mathcal{M}, I) < t$, where t is a threshold of a pose error function e (Sec. 4). As in the PASCAL VOC challenge [6], estimated poses E_I are greedily assigned to the matchable ground truth poses G_I in the order of decreasing confidence s. This results in the maximal matching $M = \{(\hat{x}_1, \bar{x}_1), (\hat{x}_2, \bar{x}_2), \ldots, (\hat{x}_l, \bar{x}_l)\} \subseteq F$, where $\hat{x}_i \in E_I$ and $\bar{x}_i \in G_I$, and no two edges share an endpoint. The set of correct poses is defined as $E_I^c = \{\hat{x} \in E_I : \exists \bar{x} \in G_I : (\hat{x}, \bar{x}) \in M\}$, *i.e.* an estimated pose is considered correct if it is matched to some ground truth pose.

Alternatively, one may not prioritize matchable pairs based on the confidence s, since they all satisfy the necessary matching condition, and maximize the number of matches instead. This would correspond to finding a maximum cardinality matching in the bipartite graph B, which can be done using *e.g.* the Hopcroft-Karp algorithm [7]. However, if the threshold t is set judiciously, the two matching approaches lead to nearly identical results and thus we prefer the simpler greedy approach.

3.2 Performance Score

Following Hinterstoisser et al. [5], we suggest to measure performance in the 6D localization problem by the Mean Recall (MR), calculated as the mean of the per-object recall rates:

$$\mathrm{MR} = \operatorname*{avg}_{o \in O} \frac{\sum_I \left|\left\{(o', \hat{\mathbf{P}}, s) \in E_I^c : o' = o\right\}\right|}{\sum_I \left|\left\{(o', \bar{\mathbf{P}}) \in G_I : o' = o\right\}\right|}, \tag{1}$$

where $I \in \mathcal{I}$ and $\mathcal{I}$ is a set of test images.

In the 6D detection problem, we suggest to measure performance by the Mean Average Precision (MAP), calculated as the mean of the per-object Average Precision (AP) rates:

$$\mathrm{MAP} = \operatorname*{avg}_{o \in O} \operatorname*{avg}_{r \in S_o} \frac{\sum_I \left|\left\{(o', \hat{\mathbf{P}}, s) \in E_I^c : o' = o, s \geq r\right\}\right|}{\sum_I \left|\left\{(o', \hat{\mathbf{P}}, s) \in E_I : o' = o, s \geq r\right\}\right|}, \tag{2}$$

where $S_o = \bigcup_I \{s : (o, \hat{\mathbf{P}}, s) \in E_I^c\}$ is a set of confidence values of estimated poses that are considered correct. The AP rate effectively summarizes the shape of the Precision-Recall curve and we suggest to calculate it as in the PASCAL VOC challenge from 2010 onwards [1] – by the average of the precision observed each time a new positive sample is recalled, *i.e.* a correct pose is estimated.

Both scores, MR and MAP, depend on parameters of the necessary matching condition, which include the threshold t, the pose error function e and parameters

of e. The scores can be calculated for several interesting parameter settings or integrated over a reasonable range of settings. Sec. 4.5 discusses the parameters in more detail.

4 Measuring Error of Estimated Pose

This section introduces the notion of indistinguishable poses (Sec. 4.1) and the requirement on the pose error functions to be invariant under pose ambiguity (Sec. 4.2). It reviews the common pose error functions (Sec. 4.3), proposes new functions that are invariant under pose ambiguity (Sec. 4.4), and discusses the condition for matching of an estimated pose to the ground truth pose (Sec. 4.5).

4.1 Indistinguishable Poses

The set of poses of model $\mathcal{M}$ that are ε-indistinguishable from pose $\mathbf{P}$ in image I is defined as: $[\mathbf{P}]_{\mathcal{M},I,\varepsilon} = \{\mathbf{P}' : d(v_I[\mathbf{P}\mathcal{M}], v_I[\mathbf{P}'\mathcal{M}]) \leq \varepsilon\}$, where $v_I[\mathcal{M}] \subseteq \mathcal{M}$ is the part of model surface that is visible in I (*i.e.* the part that is not self-occluded or occluded by some other object), d is a distance between surfaces, and ε is a tolerance that controls the level of detail to be distinguished. A possible choice for d is the Hausdorff distance [8], which measures distance of surface shapes (appearance could be also considered if $\mathcal{M}$ is colored).

When object pose is ambiguous due to object symmetries or occlusions, the set of ε-indistinguishable poses $[\mathbf{P}]_{\mathcal{M},I,\varepsilon}$ contains various object poses, not only the poses that are nearly identical to $\mathbf{P}$. Note that $[\mathbf{P}]_{\mathcal{M},I,\varepsilon}$ is an equivalence class of $\mathbf{P}$ iff $\varepsilon = 0$ (for $\varepsilon > 0$, the binary relation defining the set is not transitive).

An object pose $\mathbf{P}' \in [\mathbf{P}]_{\mathcal{M},I,\varepsilon}$ is related to $\mathbf{P}$ by a transformation $\mathbf{T} \in T_{\mathbf{P},\mathcal{M},I,\varepsilon} : \mathbf{P}' = \mathbf{TP}$, which consists of a translation and a rotation. The set $T_{\mathbf{P},\mathcal{M},I,\varepsilon}$ represents partial ε-symmetries [8], which describe repetitions of the visible surface part $v_I[\mathbf{P}\mathcal{M}]$ on the entire surface of $\mathbf{P}\mathcal{M}$. It is allowed that $v_I[\mathbf{P}\mathcal{M}] \cap v_I[\mathbf{P}'\mathcal{M}] \neq \emptyset$, *i.e.* the matching surface patches can overlap. The partial ε-symmetries can be found by *e.g.* the method of Mitra et al. [9].

4.2 Invariance to Pose Ambiguity

The error $e(\hat{\mathbf{P}}, \bar{\mathbf{P}}; \mathcal{M}, I) \in \mathbb{R}_0^+$ of an estimated 6D object pose $\hat{\mathbf{P}}$ w.r.t. the ground truth pose $\bar{\mathbf{P}}$ of object model $\mathcal{M}$ in image I is required to be invariant under pose ambiguity, *i.e.* $\forall\hat{\mathbf{P}}' \in [\hat{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}, \forall\bar{\mathbf{P}}' \in [\bar{\mathbf{P}}]_{\mathcal{M},I,\varepsilon} : e(\hat{\mathbf{P}}', \bar{\mathbf{P}}') \approx e(\hat{\mathbf{P}}, \bar{\mathbf{P}})$, where the equality is approximate due to the tolerance ε. A pose error function e that satisfies this property is said to be *ambiguity-invariant*. Note: This property is required because a 6D object pose estimator makes predictions only from a single input image. There is no tracking or any other source of information which the estimator could use to remove the pose ambiguity.

4.3 Common Pose Error Functions

This section reviews the common pose error functions and discusses their properties. None of these functions that operate in 3D space are ambiguity-invariant.

Average Distance of Model Points The most widely used pose error function is the one proposed by Hinterstoisser et al. [5]. It is used for evaluation in *e.g.* [10,11,12,13,14,15,16]. The error of the estimated pose $\hat{\mathbf{P}}$ w.r.t. the ground truth pose $\bar{\mathbf{P}}$ of object model $\mathcal{M}$ that has no indistinguishable views is calculated as the average distance to the corresponding model point:

$$e_{\text{ADD}}(\hat{\mathbf{P}}, \bar{\mathbf{P}}; \mathcal{M}) = \underset{\mathbf{x} \in \mathcal{M}}{\text{avg}} \left\| \bar{\mathbf{P}}\mathbf{x} - \hat{\mathbf{P}}\mathbf{x} \right\|_2 . \tag{3}$$

If the model $\mathcal{M}$ has indistinguishable views, the error is calculated as the average distance to the closest model point:

$$e_{\text{ADI}}(\hat{\mathbf{P}}, \bar{\mathbf{P}}; \mathcal{M}) = \underset{\mathbf{x}_1 \in \mathcal{M}}{\text{avg}} \min_{\mathbf{x}_2 \in \mathcal{M}} \left\| \bar{\mathbf{P}}\mathbf{x}_1 - \hat{\mathbf{P}}\mathbf{x}_2 \right\|_2 . \tag{4}$$

Object model $\mathcal{M}$ is considered to have indistinguishable views if $\exists \mathbf{P}, \exists \mathbf{P}'$, $\exists C : d(v_C[\mathbf{P}\mathcal{M}],\ v_C[\mathbf{P}'\mathcal{M}]) \leq \varepsilon \wedge f(\mathbf{P}, \mathbf{P}') \geq \rho$, where $v_C[\mathcal{M}] \subseteq \mathcal{M}$ is the part of model surface that is visible from camera C (*i.e.* the part that is not self-occluded), the function d measures a distance between two surfaces (as in Sec. 4.1), and ρ is the minimum required distance f between the poses (this is required because there are many nearly identical poses for which the surface distance is below ε).

Although it became a common practise, values of e_{ADD} and e_{ADI} should not be directly compared. This is because e_{ADI} yields relatively small errors even for views that are distinguishable, and is thus more permissive than e_{ADD} (e_{ADI} is in fact the lower bound of e_{ADD}). The objects evaluated with e_{ADI} are therefore advantaged. Moreover, neither e_{ADD} or e_{ADI} is ambiguity-invariant (see Sec. 5).

Translational and Rotational Error Model-independent pose error functions are used in [16,17,18]. The error of the estimated pose $\hat{\mathbf{P}} = (\hat{\mathbf{R}}, \hat{\mathbf{t}})$ w.r.t. the ground truth pose $\bar{\mathbf{P}} = (\bar{\mathbf{R}}, \bar{\mathbf{t}})$ is measured by the translational (e_{TE}) and the rotational error (e_{RE}):

$$e_{\text{TE}}(\hat{\mathbf{t}}, \bar{\mathbf{t}}) = \left\| \bar{\mathbf{t}} - \hat{\mathbf{t}} \right\|_2 , \tag{5}$$

$$e_{\text{RE}}(\hat{\mathbf{R}}, \bar{\mathbf{R}}) = \arccos\left(\left(\text{Tr}(\hat{\mathbf{R}}\bar{\mathbf{R}}^{-1}) - 1 \right) / 2 \right) . \tag{6}$$

The error e_{RE} is given by the angle from the axis–angle representation of rotation [19]($p23$). Neither e_{TE} nor e_{RE} is ambiguity-invariant. As discussed in

Sec. 4.5, fitness of object surface alignment is the main indicator of object pose quality, model-dependent pose error functions should be therefore preferred.

Complement over Union A popular way how to measure accuracy of detection and segmentation methods in 2D domain is to calculate the Intersection over Union score [1]:

$$s_{\text{IOU}}(\hat{B}, \bar{B}) = area(\hat{B} \cap \bar{B}) \,/\, area(\hat{B} \cup \bar{B}), \tag{7}$$

where $\hat{B}$ and $\bar{B}$ is the estimated and the ground truth 2D region respectively. The related cost function is the Complement over Union:

$$e_{\text{COU}}(\hat{B}, \bar{B}) = 1 - area(\hat{B} \cap \bar{B}) \,/\, area(\hat{B} \cup \bar{B}). \tag{8}$$

Depending on the task, $\hat{B}$ and $\bar{B}$ can be rectangular regions (given by bounding boxes) or segmentation masks. For evaluation of 6D object pose estimates, the 2D regions can be obtained by projection of the object model $\mathcal{M}$ in the estimated pose $\hat{\mathbf{P}}$ and the ground truth pose $\bar{\mathbf{P}}$. Such pose error function is ambiguity-invariant, but since it operates in the projective space, it provides only a weak information about fitness of the object surface alignment. Another possibility is to extend e_{COU} to work with 3D volumes. Such function can be made ambiguity-invariant (by *e.g.* taking the minimum over the sets of ε-indistinguishable poses), but requires well-defined 3D models with hole-free surfaces. We define a more practical extension of e_{COU} in Sec. 4.4.

4.4 Ambiguity-invariant Pose Error Functions

We propose three pose error functions that are ambiguity-invariant. The Visible Surface Discrepancy is of the highest practical relevance since it is inherently ambiguity-invariant.

Errors Based on Corresponding Point Distance If the sets $[\hat{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}$ and $[\bar{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}$ are available, we propose to calculate the average or the maximum of distances between corresponding points of model $\mathcal{M}$ for each pose pair $(\hat{\mathbf{P}}', \bar{\mathbf{P}}') \in Q = [\hat{\mathbf{P}}]_{\mathcal{M},I,\varepsilon} \times [\bar{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}$, and take the minimum as the pose error:

$$e_{\text{ACPD}}(\hat{\mathbf{P}}, \bar{\mathbf{P}}; \mathcal{M}, I, \varepsilon) = \min_{(\hat{\mathbf{P}}', \bar{\mathbf{P}}') \in Q} \operatorname*{avg}_{\mathbf{x} \in \mathcal{M}} \left\| \bar{\mathbf{P}}'\mathbf{x} - \hat{\mathbf{P}}'\mathbf{x} \right\|_2, \tag{9}$$

$$e_{\text{MCPD}}(\hat{\mathbf{P}}, \bar{\mathbf{P}}; \mathcal{M}, I, \varepsilon) = \min_{(\hat{\mathbf{P}}', \bar{\mathbf{P}}') \in Q} \max_{\mathbf{x} \in \mathcal{M}} \left\| \bar{\mathbf{P}}'\mathbf{x} - \hat{\mathbf{P}}'\mathbf{x} \right\|_2. \tag{10}$$

The pose error e_{ACPD} is an extension of e_{ADD}. It can be used to evaluate results for objects with or without indistinguishable views, and thus allows their

impartial comparison, which is not the case of e_{ADD} and e_{ADI} (Sec. 4.3). The pose error e_{MCPD} might be more relevant for robotic manipulation, in which the maximum surface deviation strongly indicates the chance of a successful grasp.

Determination of the sets $[\hat{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}$ and $[\bar{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}$ complicates the evaluation process, especially because $[\hat{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}$ needs to be determined during evaluation. Hence, we suggest to prefer the Visible Surface Discrepancy in general. However, e_{ACPD} and e_{MCPD} can be still useful when the sets are easy to obtain, *i.e.* when object symmetries can be enumerated and occlusions (including self-occlusions) do not cause any ambiguity.

Visible Surface Discrepancy To achieve the ambiguity-invariance while avoiding the need to determine the sets $[\hat{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}$ and $[\bar{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}$, we propose to calculate the error only over the visible part of the model surface. The Visible Surface Discrepancy is defined as follows:

$$e_{\text{VSD}}(\hat{\mathbf{P}}, \bar{\mathbf{P}}; \mathcal{M}, I, \delta, \tau) = \underset{p \in \hat{V} \cup \bar{V}}{\text{avg}} \; c(p, \hat{D}, \bar{D}, \tau), \tag{11}$$

where $\hat{V}$ and $\bar{V}$ is a 2D mask of the visible surface of $\hat{\mathcal{M}} = \hat{\mathbf{P}}\mathcal{M}$ and $\bar{\mathcal{M}} = \bar{\mathbf{P}}\mathcal{M}$ respectively (Fig. 2). $\hat{D}$ and $\bar{D}$ are distance images obtained by rendering of $\hat{\mathcal{M}}$ and $\bar{\mathcal{M}}$. A distance image stores at each pixel p the distance from the camera center to the closest 3D point $\mathbf{x}_p$ on the model surface that projects to p[1]. δ is a tolerance used for estimation of the visibility masks, and $c(p, \hat{D}, \bar{D}, \tau) \in [0, 1]$ is the matching cost at pixel p:

$$c(p, \hat{D}, \bar{D}, \tau) = \begin{cases} d \,/\, \tau & \text{if } p \in \hat{V} \cap \bar{V} \;\wedge\; d < \tau \\ 1 & \text{otherwise,} \end{cases} \tag{12}$$

where $d = |\hat{D}(p) - \bar{D}(p)|$ is the distance between the surfaces of $\hat{\mathcal{M}}$ and $\bar{\mathcal{M}}$ at pixel p, and τ is the misalignment tolerance that limits the allowed range of d. The cost c linearly increases from 0 to 1 as d increases to τ. This allows to distinguish well aligned surfaces from surfaces whose distance is close to the tolerance τ. For pixels with $d \geq \tau$ or pixels which are not in the intersection of the visibility masks, the matching cost is set to the maximum value of 1.

Since pixels from both visibility masks are considered, the estimated pose $\hat{\mathbf{P}}$ is penalized for the non-explained parts of the visible surface of $\bar{\mathcal{M}}$ and also for hallucinating its non-present parts. The function e_{VSD} can be seen as an extension of the Complement over Union (Sec. 4.3) calculated on the visibility masks, where pixels in the intersection of the masks can have a non-zero cost.

[1] The distance image can be readily computed from a depth image, which at each pixel stores the Z coordinate of the closest scene surface.

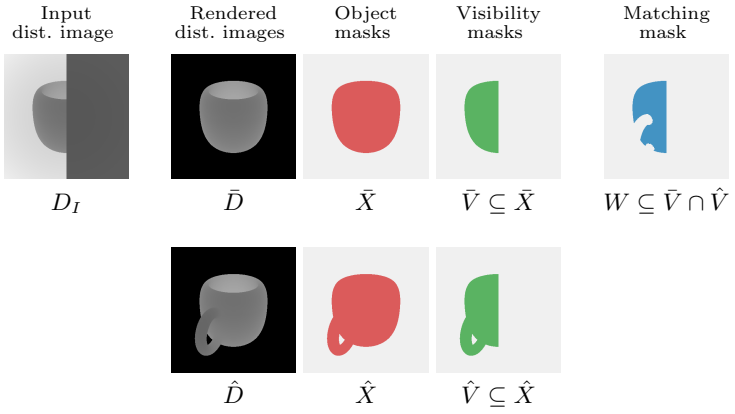

Fig. 2. Example of distance images and masks that are employed in calculation of the Visible Surface Discrepancy (e_{VSD}). The smaller the distance, the darker the pixel intensity in the distance image (pixels with unknown distances are black). Input distance image D_I captures a cup whose right part is occluded. The pose of the cup is ambiguous – from the given view it is impossible to determine the position of the handle. The matching mask W includes pixels at which the difference of the visible surface distance is smaller than τ.

Visibility Masks The visibility mask $\bar{V}$ is defined as a set of pixels where the surface of $\bar{\mathcal{M}}$ is in front of the scene surface, or at most by a tolerance δ behind:

$$\bar{V} = \{p : p \in X_I \cap \bar{X} \wedge \bar{D}(p) - D_I(p) \leq \delta\}, \tag{13}$$

where D_I is the distance image of the test scene, $X_I = \{p : D_I(p) > 0\}$ and $\bar{X} = \{p : \bar{D}(p) > 0\}$ is a set of valid scene pixels and a set of valid object pixels respectively. $D(p) = 0$ if the distance at pixel p in distance image D is unknown.

Similar visibility condition as in (13) is applied to obtain the visibility mask $\hat{V}$ of $\hat{\mathcal{M}}$. In addition to that, to ensure that the visible surface of the sought object captured in D_I does not occlude the surface of $\hat{\mathcal{M}}$, all object pixels $p \in \hat{X} = \{p : \hat{D}(p) > 0\}$ which are included in $\bar{V}$ are added to $\hat{V}$, regardless of the surface distance at these pixels. The visibility mask $\hat{V}$ is defined as follows:

$$\hat{V} = \{p : (p \in X_I \cap \hat{X} \wedge \hat{D}(p) - D_I(p) \leq \delta) \vee p \in \bar{V} \cap \hat{X}\}. \tag{14}$$

The tolerance δ should reflect accuracy of the ground truth poses and also the noise characteristics of the used depth sensor, *i.e.* it should increase with depth, as the measurement error typically does [20]. However, in our experiments we obtained satisfactory results even with δ fixed to 1.5 *cm*. Sample visibility masks are shown in Fig. 3.

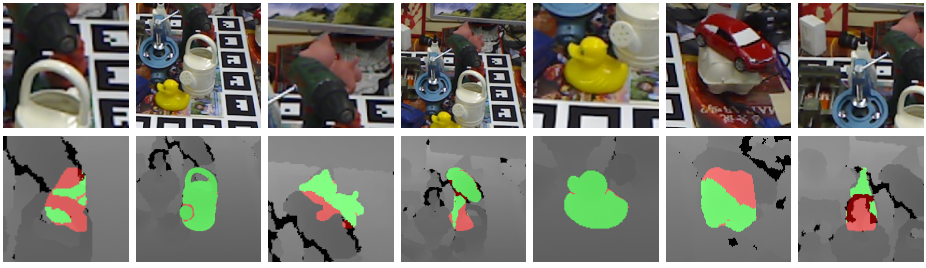

Fig. 3. Sample visibility masks $\bar{V}$ estimated with $\delta = 1.5\,cm$ in an RGB-D image from the dataset of Hinterstoisser et al. [5] using additional ground truth poses by Brachmann et al. [13]. The top row shows the color image, the bottom row shows masks overlaid on the depth image – the visibility mask $\bar{V}$ is shown in green, the occlusion mask $\bar{X} \setminus \bar{V}$ in red.

4.5 Discussion on the Necessary Matching Condition

An estimated 6D object pose $\hat{\mathbf{P}}$ is considered matchable with the ground truth pose $\bar{\mathbf{P}}$, if $e(\hat{\mathbf{P}}, \bar{\mathbf{P}}; \mathcal{M}, I) < t$ (Sec. 3.1). The choice of both the pose error function e and the threshold t largely depends on the target application. We discuss two areas in which the 6D object pose estimation is of great importance and which have different requirements on quality of the estimated pose – robotic manipulation and augmented reality.

In robotic manipulation, where a robotic arm operates in the 3D space, the absolute error of the estimated pose is important – especially in terms of misalignment of the object surface. The requirements are different for augmented reality applications, where the perceivable error is more relevant. This error depends on perspective projection and thus the closer the object to the camera, the more accurate the pose should be. Additionally, accuracy of the object position in the X and Y axis of the camera coordinate system is more important than accuracy in the Z axis, which represents the viewing direction of the camera.

Hinterstoisser et al. [5] adapt the threshold to the object size by requiring e_{ADD} or e_{ADI} to be below 10% of the object diameter. Others use fixed thresholds. Shotton et al. [18] require e_{TE} to be below $5\,cm$ and e_{RE} below 5°. Everingham et al. [1] require the e_{IOU} score to be above 0.5.

The adaptive threshold of Hinterstoisser et al. [5] makes a little sense. This is because the task is actually easier for larger objects since there are more pixels available to estimate the pose. It is more reasonable to adapt the threshold to the object distance from the camera (*e.g.* to the average distance of the visible surface of the model in the ground truth pose). This reflects the noise characteristics of the current RGB-D sensors (the depth measurement error increases quadratically with depth [20]), and also allows to control the perceivable error

which is important for the augmented reality applications. On the other hand, for robotic manipulation, it is more appropriate to keep the threshold fixed.

For e_{VSD} (Sec. 4.4), we propose to keep the threshold of the error fixed. Depending on the target application, the misalignment tolerance τ, which is used in calculation of e_{VSD}, can be either fixed or adapted to the object distance from the camera.

5 Comparison of Pose Error Functions

The discussed pose error functions were evaluated on a synthetic sequence ($\mathbf{P}_0$, $\mathbf{P}_1$, ..., $\mathbf{P}_{359}$) of 6D poses of a rotating cup (Fig. 4). Each pose $\mathbf{P}_i$ represents a rotation by $i°$ around axis perpendicular to the bottom of the cup. The poses were evaluated against the ground truth pose $\bar{\mathbf{P}}$, which was set to be the rotation by 90°. The handle of the cup is not visible in $\bar{\mathbf{P}}$ and thus its pose is ambiguous. $[\bar{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}$ was set to contain rotations from the range $[55°, 125°]$, which represent all poses from the sequence in which the handle is not visible. The set $[\mathbf{P}_i]_{\mathcal{M},I,\varepsilon}$ of the evaluated pose $\mathbf{P}_i$ was set to be the same as $[\bar{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}$ if $55 \leq i \leq 125$, and to $\{\mathbf{P}_i\}$ otherwise.

The calculated errors are shown in Fig. 5. Note that the error $e(\mathbf{P}_i, \bar{\mathbf{P}}; \mathcal{M}, I)$ calculated by the ambiguity-invariant pose error functions (e_{VSD}, e_{ACPD}, e_{MCPD}, e_{COU}) is close to zero for $\mathbf{P}_i \in [\bar{\mathbf{P}}]_{\mathcal{M},I,\varepsilon}$, which is the intuitive behavior.

Besides the synthetic sequence, we analyzed the pose error functions on the dataset of Tejani et al. [12]. The estimated poses produced by a method of the same authors were evaluated against the ground truth poses provided with the dataset. For e_{VSD}, the tolerances δ and τ were set to $1.5\,cm$ and $10\,cm$ respectively. The threshold t was set to 0.5 for both e_{VSD} and e_{COU}, and to 15% of the object diameter for e_{ADD} and e_{ADI}. Fig. 6 discusses several examples of the calculated errors.

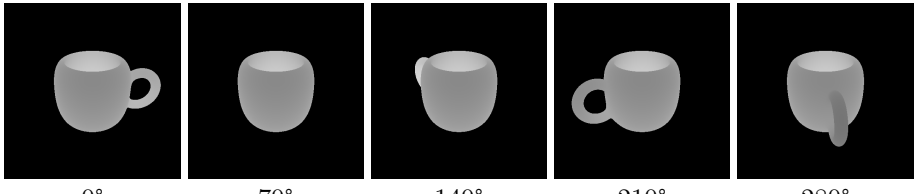

Fig. 4. Sample rendered depth images of a rotating cup (the rotation axis is perpendicular to the bottom of the cup).

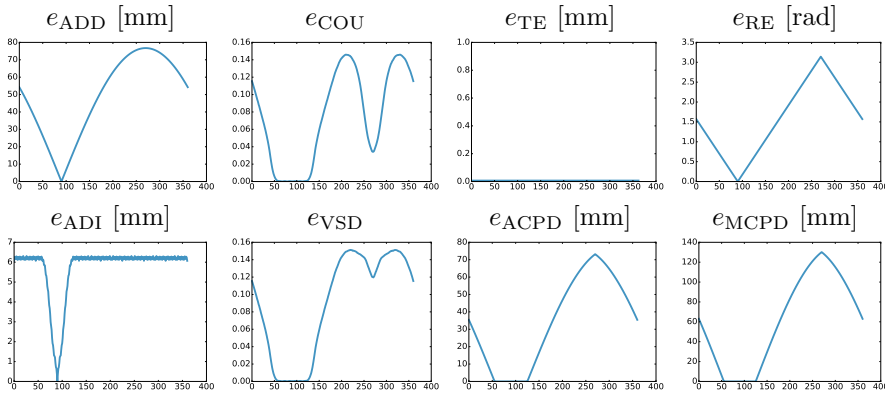

Fig. 5. Comparison of the pose error functions on the rotating cup. X axis shows rotation of the cup (from 0° to 359°). Y axis shows the calculated error.

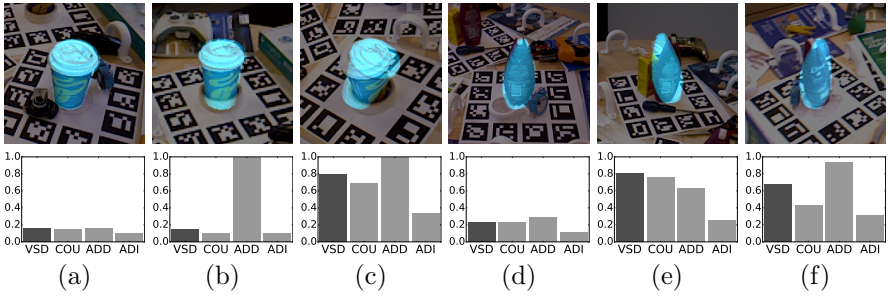

Fig. 6. Comparison of pose error functions on sample images from the dataset of Tejani et al. [12]. The visualized errors were normalized by the threshold t and thus all ranges from 0 to 1. The top row shows renderings of the object model in the estimated poses (blue), which are overlaid on the cropped color component of the input RGB-D image. (a,d) All errors are low for pose estimates that are close to the ground truth. (c,e) e_{VSD} and e_{COU} are sensitive to misalignment of object silhouettes, encouraging low perceivable error. (f) Unlike e_{COU}, which operates only in the 2D space, e_{VSD} penalizes also inconsistency in depth – the estimated pose is too close to the camera in this case. As expected, e_{ADD} produces non-intuitive values for these symmetric objects.

6 Conclusion

We defined two 6D object pose estimation problems – the 6D localization, in which prior information about presence of known objects in a test image is provided, and the 6D detection, in which no prior information is provided.

To measure error of an estimated 6D object pose w.r.t. the ground truth pose, we proposed to use the Visible Surface Discrepancy (e_{VSD}), which calculates the error over the visible surface of the object model. It is inherently ambiguity-

invariant, *i.e.* it treats the ε-indistinguishable poses as approximately equivalent. Alternatively, if the sets of ε-indistinguishable poses are available, we proposed to use the Average or the Maximum Corresponding Point Distance (e_{ACPD}, e_{MCPD}), which measure misalignment over the entire surface of the object model.

Determination of which estimated poses are correct is formulated as finding a maximal matching in a bipartite graph, where edges connect the ground truth poses with matchable estimated poses. The estimated poses are greedily assigned to the matchable ground truth poses in the order of decreasing confidence. An estimated pose is considered correct if the resulting matching includes an edge connecting the pose with some ground truth pose.

We proposed to apply a fixed threshold t on the value of e_{VSD} to decide if an estimated object pose is matchable with the ground truth pose. The misalignment tolerance τ, which is a parameter of e_{VSD}, can be either fixed or adapted to the object distance from the camera. For e_{ACPD} or e_{MCPD}, we proposed to keep the threshold t fixed or to adapt it to the object distance.

We suggested to measure performance of a 6D object pose estimator by the Mean Recall (MR) in the 6D localization problem, and by the Mean Average Precision (MAP) in the 6D detection problem.

The ongoing work is focused on a thorough validation of the proposed evaluation methodology, its application to data represented by sparse point clouds, and on extension of the Visible Surface Discrepancy to a multi-camera setting.

Implementation of the discussed pose error functions and the performance score functions is provided at: `https://github.com/thodan/obj_pose_eval`

Acknowledgements

We thank Caner Sahin, Rigas Kouskouridas and Tae-Kyun Kim from Imperial College London for providing results of the method by Tejani et al. [12], and Eric Brachmann from TU Dresden for discussion about the matching condition.

We thank the anonymous reviewer for pointing out the issue with data represented by sparse point clouds and the issue of extending the Visible Surface Discrepancy to a multi-camera setting. We are interested in a further discussion.

This work was supported by CTU student grant SGS15/155/OHK3/2T/13 and by the Technology Agency of the Czech Republic research program (V3C – Visual Computing Competence Center) TE01020415.

References

1. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1) (2015) 98–136

2. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3) (2015) 211–252
3. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: IEEE Winter Conference on Applications of Computer Vision, IEEE (2014) 75–82
4. Tulsiani, S., Malik, J.: Viewpoints and keypoints. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2015) 1510–1519
5. 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: ACCV. (2012)
6. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2) (2010) 303–338
7. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network flows. Technical report, DTIC Document (1988)
8. Mitra, N.J., Pauly, M., Wand, M., Ceylan, D.: Symmetry in 3d geometry: Extraction and applications. In: Computer Graphics Forum. Volume 32., Wiley Online Library (2013) 1–23
9. Mitra, N.J., Guibas, L.J., Pauly, M.: Partial and approximate symmetry detection for 3d geometry. ACM Transactions on Graphics (TOG) **25**(3) (2006) 560–568
10. Hodaň, T., Zabulis, X., Lourakis, M., Obdržálek, Š., Matas, J.: Detection and fine 3D pose estimation of texture-less objects in RGB-D images. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
11. Rios-Cabrera, R., Tuytelaars, T.: Discriminatively trained templates for 3d object detection: A real time scalable approach. In: ICCV. (2013) 2048–2055
12. Tejani, A., Tang, D., Kouskouridas, R., Kim, T.: Latent-class hough forests for 3d object detection and pose estimation. In: ECCV. (2014)
13. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J.: Learning 6d object pose estimation using 3d object coordinates. In: ECCV. (2014)
14. 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. arXiv preprint arXiv:1508.04546 (2015)
15. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3d pose estimation. arXiv preprint arXiv:1502.05908 (2015)
16. Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model globally, match locally: Efficient and robust 3d object recognition. In: CVPR. (2010) 998–1005
17. Choi, C., Christensen, H.: 3D Pose Estimation of Daily Objects Using an RGB-D Camera. In: IROS. (2012) 3342–3349
18. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in rgb-d images. In: CVPR. (2013) 2930–2937
19. Morawiec, A.: Orientations and Rotations: Computations in Crystallographic Textures. Springer Science & Business Media (2004)
20. Khoshelham, K.: Accuracy analysis of kinect depth data. In: ISPRS workshop laser scanning. Volume 38. (2011)