Measuring Error of 6D Object Pose in The SIXD Challenge 2017

Tomáš Hodaň, Jiří Matas, Štěpán Obdržálek

Center for Machine Perception, Czech Technical University in Prague

July 3, 2017

1 Introduction

Evaluation of 6D object pose estimates is not straightforward. Given an image, object pose can be ambiguous, *i.e.* there can be multiple poses that are indistinguishable. This is due to the existence of multiple fits of the visible part of the object surface to the entire object surface – the visible part is determined by self-occlusion and occlusion by other objects and the multiple surface fits are induced by global or partial object symmetries [1].

Sec. 2 of this document defines the Visible Surface Discrepancy (VSD), a pose error function that can deal with pose ambiguity and that is used for the evaluation in the SIXD Challenge 2017. Sec. 3 defines the criterion of pose correctness and Sec. 4 compares VSD with the Average Distance of Model Points by Hinterstoisser et al. [2].

Evaluation scripts are available at: https://github.com/thodan/sixd_toolkit.

2 Visible Surface Discrepancy

In the SIXD Challenge 2017, where only rigid objects are considered, the error of an estimated pose is measured by the Visible Surface Discrepancy (VSD) [1] that calculates the error over the visible part of the model surface. The visible part is the same in all indistinguishable poses which are thus treated as equivalent. This is a desired property that is not present in the case of the other pose error functions commonly used in the literature, *e.g.* both versions of the average distance of model points [2] or the translational and rotational error [3] (see [1] for a detailed analysis).

We use the following definition of VSD to calculate error $e \in \mathbb{R}_0^+$ of estimated 6D object pose $\hat{\mathbf{P}}$ w.r.t. ground truth pose $\bar{\mathbf{P}}$ of object model $\mathcal{M}$ in image I:

$$e_{\text{VSD}}(\hat{\mathbf{P}}, \bar{\mathbf{P}}; \mathcal{M}, I, \delta, \tau) = \underset{p \in \hat{V} \cup \bar{V}}{\text{avg}} c(p, \hat{D}, \bar{D}, \tau), \tag{1}$$

$$c(p, \hat{D}, \bar{D}, \tau) = \begin{cases} 0 & \text{if } p \in \hat{V} \cap \bar{V} \ \wedge \ |\hat{D}(p) - \bar{D}(p)| < \tau \\ 1 & \text{otherwise,} \end{cases} \tag{2}$$

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. 1). 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. Matrix $\mathbf{P}$ transforms 3D point $\mathbf{x}_m$ in the model coordinate system to 3D point $\mathbf{x}_c$ in the camera coordinate system: $\mathbf{x}_c = \mathbf{P}\mathbf{x}_m$ (the 3D points are in homogeneous coordinates). An object is represented by a mesh model $\mathcal{M}$ given by a set of points in $\mathbb{R}^3$ and a set of triangles. $\hat{D}$ 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 – it can be readily computed from a depth image which at each pixel p stores the Z coordinate of $\mathbf{x}_p$. δ is a tolerance used for estimation of the visibility masks, $c(p, \hat{D}, \bar{D}, \tau) \in [0, 1]$ is the matching cost at pixel p, and τ is a misalignment tolerance. It should be $\tau \geq \delta$. Typical values are $\delta = 15\,\text{mm}$ and $\tau = 20\,\text{mm}$.

Since pixels from both visibility masks are considered, the estimated pose is penalized for the non-explained parts of the visible surface and also for hallucinating its non-present parts. The cost c is 0 for pixels at which the surface of $\hat{\mathcal{M}}$ and the surface of $\bar{\mathcal{M}}$ are visible and the distance between the surfaces is below τ. Otherwise, the cost is 1. Note that this is different from the original definition of VSD [1], where the cost c linearly increases from 0 to 1 as the distance

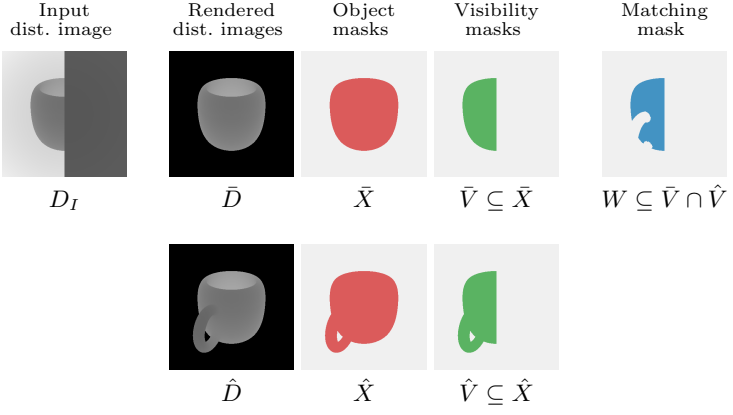

Fig. 1. Example of distance images and masks that are employed in the calculation of the Visible Surface Discrepancy (e_{VSD}). The smaller the distance, the darker the pixel intensity in the distance image (pixels with unknown distances are black). Input distance image D_I captures a cup whose right 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 error is calculated over the union of visibility masks $\hat{V}$ and $\bar{V}$. Matching mask W includes pixels $p \in \hat{V} \cap \bar{V}$ at which $|\hat{D}(p) - \bar{D}(p)| < \tau$ (the matching cost c is 0 at these pixels).

between the surfaces increases to τ. The new formulation does not distinguish well-aligned surfaces from surfaces whose distance is close to the tolerance τ, but is easier to interpret as it directly expresses the misaligned fraction of the union of visible surface parts. Moreover, the new definition does not penalize distance differences below δ that can be caused be imprecisions of depth measurements or the ground truth pose.

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:

$$\bar{V} = \{p : p \in X_I \cap \bar{X} \wedge \bar{D}(p) - D_I(p) \leq \delta\}, \tag{3}$$

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 (3) 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:

$$\hat{V} = \{p : (p \in X_I \cap \hat{X} \wedge \hat{D}(p) - D_I(p) \leq \delta) \vee p \in \bar{V} \cap \hat{X}\}. \tag{4}$$

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 [4]. However, in our experiments we obtained satisfactory results even with δ fixed to 15 mm. Sample visibility masks are shown in Fig. 2.

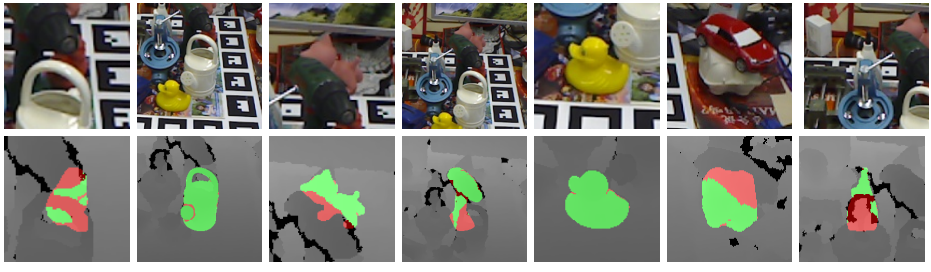

Fig. 2. Sample visibility masks $\bar{V}$ estimated with $\delta = 15$ mm in an RGB-D image from the dataset of Hinterstoisser et al. [2] using additional ground truth poses by Brachmann et al. [5]. The top row shows cropped color images, the bottom row shows corresponding depth images overlaid with the visibility mask $\bar{V}$ in green and the occlusion mask $\bar{X} \setminus \bar{V}$ in red.

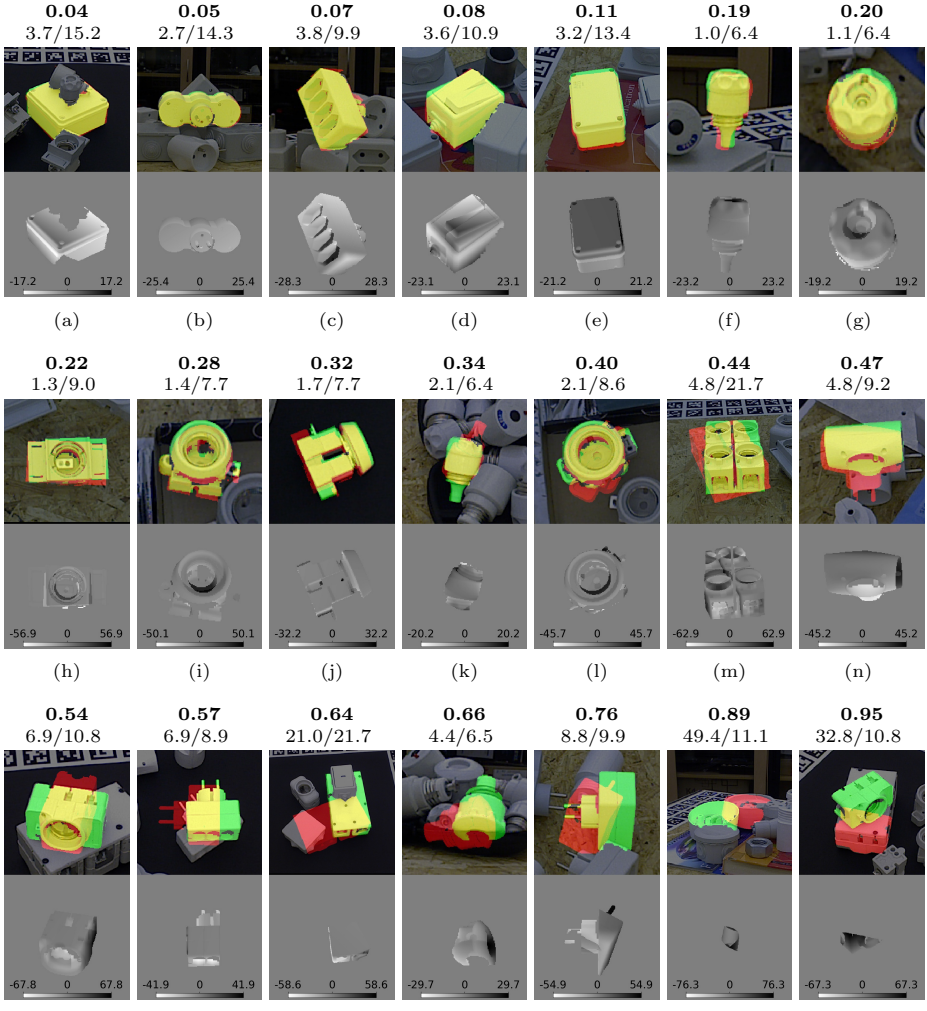

Fig. 3. Comparison of e_{VSD} (bold, $\delta = 15\,\text{mm}$, $\tau = 20\,\text{mm}$) and e_{ADI}/t_h (in mm) values on example pose estimates. $t_h = 0.1 \cdot d$, where d is the object diameter, is the threshold of correctness as defined by Hinterstoisser et al. [2]. The upper are cropped test images overlaid with visibility masks $\hat{V}$ and $\bar{V}$ in red and green respectively – they blend to yellow at $\hat{V} \cap \bar{V}$. The lower images show differences $\hat{D}(p) - \bar{D}(p)$ for pixels $p \in \hat{V} \cap \bar{V}$ that are used in calculation of the pixel-wise cost c in e_{VSD}. The examples are ordered by increasing e_{VSD} value.

3 Correctness Criterion

Estimated pose $\hat{\mathbf{P}}$ is considered correct w.r.t. ground truth $\bar{\mathbf{P}}$ if $e_{\text{VSD}}(\hat{\mathbf{P}}, \bar{\mathbf{P}}) < t$. If there are more estimated poses, at most one of them is considered correct w.r.t. a ground truth pose (Sec. 3.1 in [1] describes how the estimated poses are matched with the ground truth poses).

For the SIXD Challenge 2017, the value of threshold t has not been decided yet. However, a reasonable choice seems to be a value between 0.3 and 0.4 that would be fixed for all objects. Fig. 3 shows examples of e_{VSD} values for pose estimates produced by the method of Hodaň et al. [6] on the T-LESS dataset [7].

4 Comparison with Average Distance of Model Points

Probably the most widely used has been the pose error function by Hinterstoisser et al. [2]. It calculates the error of estimated pose $\hat{\mathbf{P}}$ w.r.t. ground truth pose $\bar{\mathbf{P}}$ of object model $\mathcal{M}$ that has no indistinguishable views as the average distance to the corresponding model point:

$$e_{\text{ADD}}(\hat{\mathbf{P}}, \bar{\mathbf{P}}; \mathcal{M}) = \operatorname*{avg}_{\mathbf{x} \in \mathcal{M}} \left\| \bar{\mathbf{P}}\mathbf{x} - \hat{\mathbf{P}}\mathbf{x} \right\|_2 . \tag{5}$$

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

$$e_{\text{ADI}}(\hat{\mathbf{P}}, \bar{\mathbf{P}}; \mathcal{M}) = \operatorname*{avg}_{\mathbf{x}_1 \in \mathcal{M}} \min_{\mathbf{x}_2 \in \mathcal{M}} \left\| \bar{\mathbf{P}}\mathbf{x}_1 - \hat{\mathbf{P}}\mathbf{x}_2 \right\|_2 . \tag{6}$$

Note, however, that even e_{ADI} is not invariant under pose ambiguity – see the example with a rotating cup in Sec. 5 of [1].

Fig. 3 compares e_{VSD} and e_{ADI} values on the T-LESS dataset. e_{ADI} was chosen because almost all objects in this dataset have some indistinguishable views. e_{ADI} is relatively low for pose estimates (o)-(s) which we do not consider correct although they satisfy the criterion of Hinterstoisser et al. [2]: a pose estimate is correct if $e_{\text{ADI}} \leq 0.1 \cdot d$, where d is the object diameter – the largest distance between any pair of model vertices.

See [1] for further discussion about commonly used pose error functions.

References

1. Hodaň, T., Matas, J., Obdržálek, Š.: On evaluation of 6d object pose estimation. European Conference on Computer Vision Workshops (ECCVW) (2016)
2. 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)

3. Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model globally, match locally: Efficient and robust 3d object recognition. In: CVPR. (2010) 998–1005
4. Khoshelham, K.: Accuracy analysis of kinect depth data. In: ISPRS workshop laser scanning. Volume 38. (2011)
5. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J.: Learning 6d object pose estimation using 3d object coordinates. In: ECCV. (2014)
6. 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)
7. Hodaň, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. IEEE Winter Conference on Applications of Computer Vision (WACV) (2017)