Direct 3D Pose Estimation of a Planar Target

Hung-Yu Tseng Po-Chen Wu
National Taiwan University
{hytseng,pcwu}@media.ee.ntu.edu.tw

Ming-Hsuan Yang
UC Merced
mhyang@ucmerced.edu

Shao-Yi Chien
National Taiwan University
sychien@ntu.edu.tw

Abstract

Estimating 3D pose of a known object from a given 2D image is an important problem with numerous studies for robotics and augmented reality applications. While the state-of-the-art Perspective-n-Point algorithms perform well in pose estimation, the success hinges on whether feature points can be extracted and matched correctly on targets with rich texture. In this work, we propose a robust direct method for 3D pose estimation with high accuracy that performs well on both textured and textureless planar targets. First, the pose of a planar target with respect to a calibrated camera is approximately estimated by posing it as a template matching problem. Next, the object pose is further refined and disambiguated with a gradient descent search scheme. Extensive experiments on both synthetic and real datasets demonstrate the proposed direct pose estimation algorithm performs favorably against state-of-the-art feature-based approaches in terms of robustness and accuracy under several varying conditions.

1. Introduction

Determining the 3D pose of a target object from a calibrated camera is a classical problem in computer vision that finds numerous applications such as robotics and augmented reality (AR). While much progress has been made in the past few decades, it remains a challenging task to develop a fast and accurate pose estimation algorithm, especially for planar target objects lacking a textured surface.

Existing pose estimation methods can be broadly categorized into two categories. The approaches in the first category are based on features extracted from target objects with rich texture. The core idea behind feature-based methods is to compute a set of n correspondences between 3D points and their 2D projections from which the relative position and orientation between the camera and target can be estimated. In recent years, numerous feature detection and tracking schemes [26, 5, 21, 33, 2] have been developed and applied to a wide range of AR and simultaneous localization and mapping applications [16, 24, 30] with demonstrated success. In order to match features more robustly, variants of RANSAC algorithms [11, 7] have been used to eliminate outliers before object pose is estimated from a set of feature correspondences. Typically the Perspective-n-Point (PnP) algorithms [34, 20, 38] are applied to the last reliable feature correspondences after using RANSAC algorithm for estimating the 3D object pose. We note that feature-based methods are less effective in pose estimation when the tilt angle between the camera and the planar target is large. While the Affine-SIFT (ASIFT) [37] approach matches feature points well when there are large changes in viewpoint, it is more computationally expensive than other algorithms. Since the performance of feature-based pose estimation methods hinges on whether or not point correspondence can be correctly matched, such approaches are less effective when the target image contains less texture or the camera image is blurry.

The second category consists of direct methods that do not depend on features. Since the seminal work by Lucas and Kanade [28], numerous algorithms for template matching based on global, iterative, nonlinear optimization have been proposed [13, 35, 4, 29]. As the pose estimation problem can be reduced to the template matching problem with reference frame, 2D or 3D poses can be estimated through optimizing the parameters to account for rigid transformations of observed target images [8, 10]. However, these methods rely on initial reference parameters and may be trapped in a local minimum. To alleviate the limitations of nonlinear optimization problems, non-iterative approaches [6, 18, 14] have recently been proposed. Nonetheless, these template matching approaches have the main shortcoming of misalignment between affine or homography transformation space and pose space. It would cause the additional pose error produced by transformation matrix decomposition while estimating the 3D pose.

In this paper, we propose a direct method to estimate the 3D poses of planar targets from a calibrated camera by measuring the similarity between the projected planar target and the 2D image based on appearance. As the proposed method is based on a planar object rather than a 3D model, the pose ambiguity problem as discussed in pri-

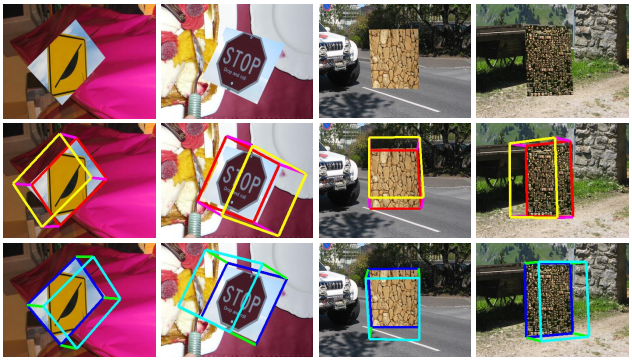


Figure 1. Pose estimation results on synthetic images. First row: original images. Second row: images rendered model with ambiguous pose obtained from proposed algorithm without refinement approach. Third row: pose estimation results from the proposed algorithm.

or art [31, 34, 22, 36], is inevitably bound to occur. Pose ambiguity is related to situations where the according error function has several local minima for a given configuration, which is the main cause of jumping pose estimation results in an image sequence. Based on image observations, one of the ambiguous poses with local minima, according to an error function is the correct pose. Therefore, after obtaining an initial rough pose using an approximated pose estimation scheme, we determine all ambiguous poses and refine the estimates until they converge to local minima. The final pose is chosen as the one with the lowest error among these refined ambiguous poses. A few results are shown in Figure 1. Extensive experiments are conducted to validate the proposed algorithm. In particular, we evaluate the proposed algorithm on different types of templates with different levels of degraded images caused by blur, intensity, tilt angle, and compression noise. Furthermore, we evaluate the proposed algorithm on the datasets by Jegou *et al*. [15] against the state-of-the-art pose estimation methods.

The main contributions of this work are summarized as follows. First, we propose an efficient non-feature based pose estimation algorithm for a planar target undergoing arbitrary 3D perspective transformations. Second, the proposed pose estimation algorithm performs favorably against the state-of-the-art feature-based approaches in terms of robustness and accuracy. Third, the proposed pose refinement method not only improves the accuracy of estimated results but also alleviates the pose ambiguity problem effectively.

2. Related Works

The template matching problem has been widely studied in the literatures, and one important issue is how to efficiently obtain accurate results with evaluating only a subset of the possible transformations. Since the appearance distances between a template and two sliding windows shifted by a few pixels (e.g., one or two pixels) are usually close due to the nature of image smoothness, Pele and Werman [32] exploit this fact to reduce the time complexity of pattern matching. Alexe *et al*. [3] derive an upper bound of the Euclidean distance (based on pixel values) according to the spatial overlap of two windows in an image, and use it for efficient pattern matching. In [18], Korman *et al*. show that the 2D affine transformations of a template can be approximated by samples of a density function based on smoothness of a given image, and propose a fast matching method.

The proposed refinement method is motivated by fast motion estimation methods. Liu and Feig [25] propose the Gradient Descent Search (GDS) algorithm that evaluates the values of a given objective function from a centralized search neighborhood for motion estimation. When the minimum within a neighborhood is found, it is used to determine the position for the next search until it converges. Compared with the full search method, the GDS algorithm achieves similar performance but with much lower computational complexity. Zhu and Ma [40] develop an algorithm for block-based motion estimation based on two designed diamond-shaped search patterns, and it further reduced the required number of search points. A motion estimation method that exploits more elaborated coarse-to-fine search patterns is subsequently developed by Zhu *et al*. [39].

The pose ambiguity problem occurs not only under orthography but also for perspective transformation, especially when the target plane is significantly tilted with respect to camera views. In [34], Schweighofer and Pinz show that two local minima exist for cases with images of planar targets viewed by a perspective camera, and develop a method to determine a unique solution based on an iterative pose estimation algorithm [27]. Zheng *et al*. [38] formulate the PnP problem in a functional minimization problem and retrieve all the stationary points by using the Gröbner basis method [19], and one of the two the stationary points with smallest objective values will be the correct pose in most cases.

3. Problem Formulation

Given a target image I_t and a camera image I_c with pixel values normalized in the range $[0, 1]$, the task is to determine the object pose of I_t in six degrees of freedom parameterized based on the orientation and position of the target with respect to a calibrated camera. With a set of 3D coordinates of reference points $\mathbf{x}_i = [x_i, y_i, 0]^\top, i = 1, \ldots, n, n \geq 3$ in object-space coordinate of I_t, and a set of camera-image coordinates $\mathbf{u}_i = [u_i, v_i]^\top$ in I_c, the transformation between them can be formulated as

$$\begin{bmatrix} hu_i \\ hv_i \\ h \end{bmatrix} = \begin{bmatrix} f_x & 0 & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} [\mathbf{R}|\mathbf{t}] \begin{bmatrix} x_i \\ y_i \\ 0 \\ 1 \end{bmatrix}, \qquad (1)$$

where

$$\mathbf{R} = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \in SO(3), \mathbf{t} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \in R(3), \tag{2}$$

are the rotation matrix and translation vector, respectively. In (1), (f_x, f_y) and (x_0, y_0) are focal length and principal point of the camera, respectively.

Given the observed camera-image points $\hat{\mathbf{u}}_i = [\hat{u}_i, \hat{v}_i]^\top$, the pose estimation algorithm needs to determine values for pose $\mathbf{p} = (\mathbf{R}, \mathbf{t})$ that minimize an appropriate error function. In principle, there are two possible error functions. One is the reprojection error, which is mostly used in the PnP algorithms,

$$E_r(\mathbf{p}) = \frac{1}{n} \sum_{i=1}^{n} \left[(\hat{u}_i - u_i)^2 + (\hat{v}_i - v_i)^2 \right]. \tag{3}$$

Another error function is based on the sum of absolute differences (also known as appearance distance) and is mostly used in direct methods and this work,

$$E_a(\mathbf{p}) = \frac{1}{n_t} \sum_{i=1}^{n_t} |I_c(\mathbf{u}_i) - I_t(\mathbf{x}_i)|, \tag{4}$$

where n_t represents the total number of pixels in I_t.

4. Proposed Algorithm

The proposed algorithm consists of two steps. First, the 3D pose of a planar target with respect to a calibrated camera is estimated. Second, the object pose is further refined and disambiguated. We describe these steps as follows.

4.1. Approximated Pose Estimation

Let $T_{\mathbf{p}}$ be the transformation at pose $\mathbf{p}$. Assume a reference point $\mathbf{x}_i$ in a target image is transformed separately to $\mathbf{u}_{i1}$ and $\mathbf{u}_{i2}$ in a camera image with two different poses $\mathbf{p}_1$ and $\mathbf{p}_2$. It is shown in [18] that if any distance between $\mathbf{u}_{i1}$ and $\mathbf{u}_{i2}$ is smaller than a positive value ε, with upper bound in the Big-O notation,

$$\forall \mathbf{x}_i \in I_t : d(T_{\mathbf{p}_1}(\mathbf{x}_i), T_{\mathbf{p}_2}(\mathbf{x}_i)) = O(\varepsilon), \tag{5}$$

then the following equation holds

$$|E_a(\mathbf{p}_1) - E_a(\mathbf{p}_2)| = O(\varepsilon \bar{\mathcal{V}}), \tag{6}$$

where $\bar{\mathcal{V}}$ denotes the mean variation of I_t, which represents the mean value over the entire target image of the maximal difference between each pixel and any of its neighbors. The mean variation $\bar{\mathcal{V}}$ can be constrained by filtering I_t. The main result is that the difference between $E_a(\mathbf{p}_1)$ and $E_a(\mathbf{p}_2)$ is bounded in terms of ε. In the proposed direct method, we only need to consider a limited number of poses by constructing a ε-covering pose set $\mathcal{S}$ based on (5) and (6).

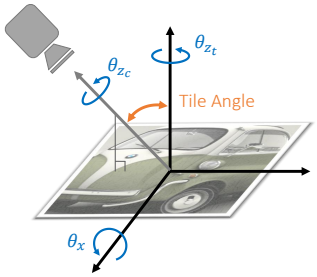


Figure 2. Illustration of rotation angle: θ_x indicates the tilt angle between camera and target image when the rotation is factored as $\mathbf{R} = \mathbf{R}_z(\theta_{z_c})\mathbf{R}_x(\theta_x)\mathbf{R}_z(\theta_{z_t})$.

Construct the ε-Covering Set. By factoring the rotation as $\mathbf{R} = \mathbf{R}_z(\theta_{z_c})\mathbf{R}_x(\theta_x)\mathbf{R}_z(\theta_{z_t})$ [9] as shown in Figure 2, the pose then can be parameterized as $\mathbf{p} = [\theta_{z_c}, \theta_x, \theta_{z_t}, t_x, t_y, t_z]^\top$. These Euler angles θ_{z_c}, θ_x, and θ_{z_t} are in the range $[-180°, 180°]$, $[0°, 90°]$, and $[-180°, 180°]$, respectively. A pose set $\mathcal{S}$ is constructed such that any two consecutive poses, $\mathbf{p}_k$ and $\mathbf{p}_k + \Delta\mathbf{p}_k$ on each dimension, satisfy (5) in $\mathcal{S}$. To construct the set favorably, the coordinates of $\mathbf{x}_i \in I_t$ are pre-normalized to the range $[-1, 1]$. Starting with t_z, we derive the equation below by using (1) for each $\mathbf{x}_i$,

$$\begin{aligned} &d(T_{\mathbf{p}_{t_z}}(\mathbf{x}_i), T_{\mathbf{p}_{t_z+\Delta t_z}}(\mathbf{x}_i)) \\ &= \sqrt{[(\frac{f_x x_i}{t_z}) - (\frac{f_x x_i}{t_z + \Delta t_z})]^2 + [(\frac{f_y y_i}{t_z}) - (\frac{f_y y_i}{t_z + \Delta t_z})]^2} \\ &= O(\frac{1}{t_z} - \frac{1}{t_z + \Delta t_z}). \end{aligned} \tag{7}$$

To make (7) satisfy the constraint in (5), we use the step size, with tight bound in Big-Theta notation,

$$\Delta t_z = \Theta(\frac{\varepsilon t_z^2}{1 - \varepsilon t_z}), \tag{8}$$

which means that (7) can be bounded if we construct $\mathcal{S}$ with the bounded step (8) on dimension t_z.

Since θ_x describes the tilt angle between camera and target image as shown in Figure 2, we obtain the following equation depending on the current t_z,

$$\begin{aligned} &d(T_{\mathbf{p}_{\theta_x}}(\mathbf{x}_i), T_{\mathbf{p}_{\theta_x+\Delta\theta_x}}(\mathbf{x}_i)) \\ &= \sqrt{d_{\mathbf{u}_i}^2 + d_{\mathbf{v}_i}^2} \\ &= O(\frac{1}{t_z - \sin(\theta_x + \Delta\theta_x)} - \frac{1}{t_z - \sin(\theta_x)}), \end{aligned} \tag{9}$$

for each $\mathbf{x}_i$, where

$$\begin{aligned} d_{\mathbf{u}_i} &= (\frac{f_x x_i}{y_i \sin\theta_x + t_z}) - (\frac{f_x x_i}{y_i \sin(\theta_x + \Delta\theta_x) + t_z}), \\ d_{\mathbf{v}_i} &= (\frac{f_y y_i \cos\theta_x}{y_i \sin\theta_x + t_z}) - (\frac{f_y y_i \cos(\theta_x + \Delta\theta_x)}{y_i \sin(\theta_x + \Delta\theta_x) + t_z}). \end{aligned} \tag{10}$$

In addition, to make (9) satisfy the constraint in (5), we set the step size,

$$\Delta\theta_x = \Theta(\sin^{-1}(t_z - \frac{1}{\varepsilon + \frac{1}{t_z - \sin(\theta_x)}}) - \theta_x). \tag{11}$$

Table 1. Bounded step size on each dimension for constructing the ε-covering pose set.

Dimension	Step Size
θ_{z_c}	$\Theta(\varepsilon t_z)$
θ_x	$\Theta(\sin^{-1}(t_z - \frac{1}{\varepsilon + \frac{1}{t_z - \sin(\theta_x)}}) - \theta_x)$
θ_{z_t}	$\Theta(\varepsilon t_z)$
t_x	$\Theta(\varepsilon(t_z - \sqrt{2}\sin(\theta_x)))$
t_y	$\Theta(\varepsilon(t_z - \sqrt{2}\sin(\theta_x)))$
t_z	$\Theta(\frac{\varepsilon t_z^2}{1-\varepsilon t_z})$

Similarly, we derive the bounded steps for the other dimension depending on the current t_z and θ_x. Table 1 summarizes the bounded step size on each dimension for the ε-covering set, and the derivation details are available in the supplementary document.

Coarse-to-Fine Estimation. Due to the large parameter space, the computational and memory costs are prohibitively high if the ε-covering pose set is used directly for pose estimation. For fast and accurate pose estimation, a coarse-to-fine approach is employed. The pose set $\mathcal{S}$ is first constructed with a coarse ε. After obtaining the best pose $\mathbf{p}_b$ and the associated error measure $E_a(\mathbf{p}_b)$, we select the poses within a threshold,

$$\mathcal{S}_L = \{\mathbf{p}_L \mid E_a(\mathbf{p}_L) < E_a(\mathbf{p}_b) + L\}, \tag{12}$$

to be considered in the next round. Here the constant L is a threshold set empirically. Based on E_a, we create sets with finer $\varepsilon^{'}$,

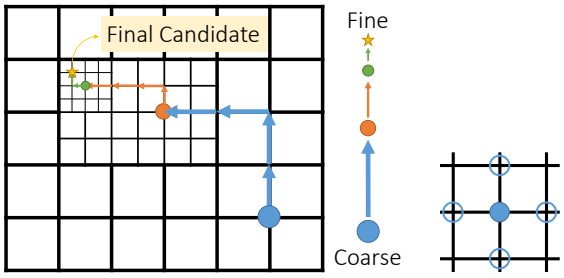
$$\mathcal{S}^{'} = \{\mathbf{p}^{'} \mid \exists \mathbf{p}_L \in \mathcal{S}_L : (5) \text{ holds for } \mathbf{p}^{'}, \mathbf{p}_L \text{ and } \varepsilon^{'}\}, \tag{13}$$

and repeat search until we obtain the desired precision parameter ε^*. The best pose in the last set is used as the approximated estimate.

Approximate the Error Measure. If we approximate the error measure $E_a^{'}$ with random sampling only a portion of pixels instead of computing E_a with sampling total pixels in I_t, according to Hoeffding's inequality [1], $E_a^{'}$ is probably close to E_a within a precision parameter δ if the number of sampling pixels m is large enough,

$$P(|E_a^{'} - E_a| > \delta) \leq 2e^{-2\delta^2 m}, \tag{14}$$

where $P(\cdot)$ represents the probability measure. This inequality suggests that if m is properly selected, the approximation error between $E_a^{'}$ and E_a can be bounded with high probability. In other words, $E_a^{'}$ is a close approximation of E_a within the probably approximately correct (PAC) framework [1]. With this approximation, the runtime of estimating the error measure can be dramatically reduced by inspecting only a small fraction of pixels in a target image.



(a) (b)

Figure 3. (a) 2D illustration of coarse-to-fine gradient descent search. We carry out the GDS on the coarse level in the beginning. After reaching the local minimum, we move to the finer level and repeat this coarse-to-fine GDS approach until we obtain minimum within the desired precision. (b) The checking pattern in 2D view, including 1 center checking point and its 4 neighnbors (2 neighnbors per dimension).

The proposed approach can be further improved to be invariant to different lighting conditions by normalizing the intensity term or adding the chroma term to the appearance distance measure.

4.2. Pose Refinement

We obtain $(\mathbf{R}^{'}, \mathbf{t}^{'})$ after the proposed approximated pose estimation scheme. However, this result is bounded based on the distance in the appearance space rather than the pose space. Therefore the estimated pose and actual pose may be significantly different even when the appearance distance is small, which is often the case when the tilt angle of a target image is large. In the meanwhile, the pose ambiguity problem is likely occur as illustrated in Figure 1. Consequently, a pose refinement scheme is proposed to further improve the accuracy and address the ambiguity problem.

Explore the Candidate Poses. In order to address the problem of pose ambiguity, we first transform four corner points $\mathbf{x}_{c1}$, $\mathbf{x}_{c2}$, $\mathbf{x}_{c3}$, and $\mathbf{x}_{c4}$ in the target image I_t to $\mathbf{u}_{c1}$, $\mathbf{u}_{c2}$, $\mathbf{u}_{c3}$, and $\mathbf{u}_{c4}$ in the camera image I_c with $(\mathbf{R}^{'}, \mathbf{t}^{'})$, respectively. Using the functional minimization method [38], we compute all its stationary points of the error function (3) based on the Gröbner basis method [19]. Finally, only the stationary points with the two smallest objective values in (4) are plausible poses, and these two ambiguous poses are both chosen as the candidate poses.

Refining Pose Estimation. After obtaining the two candidate poses, we can further improve the accuracy using a coarse-to-fine gradient descent search scheme. In contrast to the 2D motion estimation in video coding, we consider a 6D pose motion with infinity resolution in this work. A 2D view of the coarse-to-fine gradient descent search is shown in Figure 3(a). The largest blue circle denotes the approximate pose estimated in Section 4.1, and the smaller one (orange) represents the local minimum found by the search

pattern at the starting ε-precision. As the minimum under the current precision level is found, we diminish the precision parameter ε and perform gradient descent search again on the next level. This process is repeated until we obtain the local minimum under the desired precision parameter ε^*. Finally, the pose with smaller E_a is chosen from the two refined candidate poses.

The 2D view of the checking pattern in the coarse-to-fine GDS scheme [25] are shown in Figure 3(b). It is formed by 13 checking points, including the center point and its 12 neighbors. These 12 neighbors are ε-away from the center separately in the 6D pose space. Let $\mathbf{p}_c = [\theta_{z_c}, \theta_x, \theta_{z_t}, t_x, t_y, t_z]^\top$ be the center point of the checking pattern in the pose space and $\mathbf{P}_c$ be the 6×13 matrix with repeating $\mathbf{p}_c$ in a row. Also let $\mathbf{s}_\varepsilon = [s_{\theta_{z_c}}, s_{\theta_x}, s_{\theta_{z_t}}, s_{t_x}, s_{t_y}, s_{t_z}]^\top$ be the step size listed in Table 1 with precision parameter ε and $\mathbf{S}_\varepsilon$ be the 6×13 matrix with repeating $\mathbf{s}_\varepsilon$ in a row. The mathematical description of the checking pattern $\mathbf{M}$ can then be written as

$$\mathbf{M} = \mathbf{P}_c + \mathbf{D} \circ \mathbf{S}_\varepsilon, \tag{15}$$

where

$$\mathbf{D} = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix} \tag{16}$$

and $\circ$ represents the Hadamard product. Each column in $\mathbf{M}$ represents one of the checking points within the checking pattern. The main steps of the proposed pose estimation method are summarized in Algorithm 1.

Algorithm 1 Proposed Direct 3D Pose Estimation

Input: Target image I_t, camera image I_c, intrinsic parameters, and precision parameters $\varepsilon_c^*, \varepsilon_f^*$.
Output: Estimated pose result $\mathbf{p}^*$.

Create an ε-covering pose set $\mathcal{S}$.
Find $\mathbf{p}_b$ from $\mathcal{S}$ with $E_a^{'}$ according to (14).
while $\varepsilon > \varepsilon_c^*$ **do**
  Obtain the set $\mathcal{S}_L$ according to (12);
  Diminish ε;
  Replace $\mathcal{S}$ according to (13);
  Find $\mathbf{p}_b$ from $\mathcal{S}$ with $E_a^{'}$ according to (14);
end while
Explore the candidate poses $\mathbf{p}_1$ and $\mathbf{p}_2$ with $\mathbf{p}_b$.
for $i = 1 \rightarrow 2$ **do**
  Let $\mathbf{p}_c = \mathbf{p}_i$ and $\varepsilon_i = \varepsilon$
  while $\varepsilon_i > \varepsilon_f^*$ **do**
    Find $\mathbf{p}_b$ from (15) with $E_a^{'}$ according to (14).
    if $\mathbf{p}_c \neq \mathbf{p}_b$ **then**
      $\mathbf{p}_c = \mathbf{p}_b$
    else
      Diminish ε_i;
    end if
  end while
  Let $\mathbf{p}_i = \mathbf{p}_c$
end for
Return the pose $\mathbf{p}^*$ with smaller E_a from $\mathbf{p}_1$ and $\mathbf{p}_2$

5. Experimental Results

We experimentally evaluate the proposed algorithm for the 3D pose estimation problem using both synthetic and benchmark datasets, and compare it with the feature-based schemes. Through some preliminary experiments, we find the SIFT [26] method performs better than other alternative features in terms of repeatability and accuracy. Similar observations can also be found in [12]. As the ASIFT [37] method is considered the state-of-the-art affine-invariant method to find correspondences under large view change, we use both the SIFT and ASIFT methods in the compared feature-based schemes. The RANSAC-based method [11] is then used to eliminate outliers before object pose is estimated by the PnP algorithms. It has been shown that, among the PnP algorithms [34, 20, 38, 17], the OPnP [38] algorithm achieves the state-of-the-art results in terms of efficiency and precision. Therefore we use the OPnP algorithm as the pose estimator in the feature-based schemes.

We run all codes in MATLAB on a desktop computer with 3.4 GHz CPU and 16 GB RAM. Table 2 shows average runtimes for different algorithms. The source code and datasets will be made available to the public. Due to the space limit, we leave more results in the supplemental material.

Given the true rotation matrix $\hat{\mathbf{R}}$ and translation vector $\hat{\mathbf{t}}$, we compute the rotation error of the estimated rotation matrix $\mathbf{R}$ by $E_{\mathbf{R}}(degree) = acosd((\mathrm{Tr}(\mathbf{R}^\top \cdot \hat{\mathbf{R}}) - 1)/2)$, where $acosd(\cdot)$ represents the arc-cosine operation in degrees. The translation error of the estimated translation vector $\mathbf{t}$ is measured by the relative difference between $\hat{\mathbf{t}}$ and $\mathbf{t}$ defined as $E_{\mathbf{t}}(\%) = \|\hat{\mathbf{t}} - \mathbf{t}\|/\|\hat{\mathbf{t}}\| \times 100$. We define a pose to be successfully estimated if its both errors are under predefined thresholds. We use $\delta_{\mathbf{R}} = 20^\circ$ and $\delta_{\mathbf{t}} = 10\%$ as the threshold on rotation error and translation error empirically, as shown in Figure 4. The success rate (SR) is defined as the percentage of the successfully estimated poses within each test condition.

5.1. Synthetic Images

We use a set of synthetic images consisting of 8400 test images for experiments, including 21 different test conditions. Each test image is generated from a warping template image according to the randomly generated pose with

Table 2. Average runtimes for three approaches on synthetic and real test data. Numbers in parentheses represents the average steps of checking pattern in the refinement approach. Although SIFT-based Approach is the fastest method among these three different schemes, its performance is quite limited.

Data Type	SIFT-based Approach				ASIFT-based Approach				Proposed Direct Method		
	SIFT	RANSAC	OPnP	Total	ASIFT	RANSAC	OPnP	Total	Approximated	Refinement	Total
Synthetic	10.56 s.	0.08 s.	0.02 s.	10.67 s.	46.45 s.	0.07 s.	0.02 s.	46.58 s.	38.35 s.	(26.3) 2.16 s.	40.51 s.
Real	5.09 s.	0.08 s.	0.02 s.	5.19 s.	24.91 s.	0.09 s.	0.02 s.	25.08 s.	35.13 s.	(19.5) 1.29 s.	36.42 s.

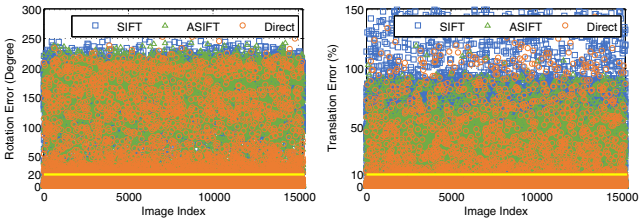


Figure 4. Distributions of rotation and translation errors over experiments. The horizontal lines correspond to the thresholds used to detect unsuccessfully estimated poses. There is a total of $15,289$ poses estimated by each pose estimation approach.

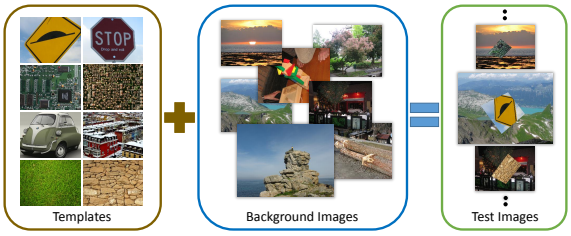


Figure 5. The test image was generated from a warping template image according to the randomly generated pose on randomly chosen background image.

tilt angle in the range $[0°, 75°]$ in a randomly chosen background image, as shown in Figure 5. The template image size is 640×480. These templates are classified into four different classes, namely "Low Texture", "Repetitive Texture", "Normal Texture", and "High Texture" [23] as shown from top to bottom in Figure 5. Each class is represented by two targets. The background images are acquired from the database [15] and resized to 800×600 pixels.

Normal Conditions. The pose estimation results of the SIFT-based, ASIFT-based, and the proposed direct methods using the undistorted test images are shown in Table 3. Each test condition contains the average rotation error $E_{\mathbf{R}}$, translation error $E_{\mathbf{t}}$, and success rate. The evaluation results show that although the proposed method is sometimes slightly less accurate than the feature-based approaches, it performs more robustly with different templates. Although the SIFT-based approach can detect and match the features accurately under small tilt angle, it frequently fails in the experiments when the template undergoes large pose change. In most cases, the feature-based approaches cannot correctly estimate the pose of textureless template images.

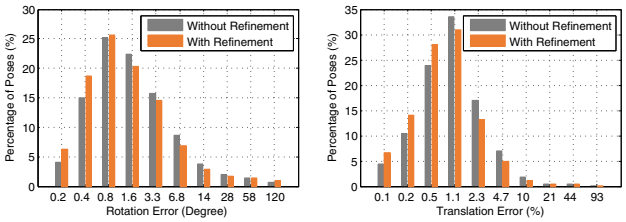


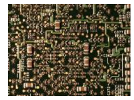
Figure 6. Pose estimation results with and without refinement approaches. The average value of rotation and translation error are both reduced by the refinement approach.

Varying Conditions. We further evaluate the proposed methods using all templates with five degradation levels: Gaussian blur with kernel width of $\{1, 2, 3, 4, 5\}$ pixels, JPEG compression with the quality parameter set to $\{90, 80, 70, 60, 50\}$, intensity change with pixel intensity scalar parameter set to $\{0.9,0.8,0.7,0.6,0.5\}$, and tilt angle in the range of $\{[0°15°), [15°30°), [30°45°), [45°60°)$, and $[60°75°)\}$. The results are shown in Figure 7. The proposed algorithm outperforms the other two feature-based methods with blurry images. All three approaches are able to deal with certain levels of distortion in intensity or JPEG compression noise. The SIFT-based approach performs well when the tilt angle is small since the marker images are not perspective distorted in the camera images. In the other conditions, however, the proposed algorithm and the ASFIT method are able to estimate 3D poses relatively well. In this synthetic image experiment, the proposed direct method achieves an overall success rate of 95.62%, while SIFT-based and ASIFT-based approaches achieve success rates of 47.62% and 74.74% respectively.

Refinement Analysis. To improve the accuracy of our pose estimation algorithm, we propose a refinement approach as described in Section 4.2. Pose estimation results (i.e., rotation and translation error) with and without the refinement approach are shown in Figure 6. The rotation and translation error can be reduced averagely by $-0.258°$ and -0.233% respectively with proposed refinement scheme.

To demonstrate the proposed algorithm is able to disambiguate among plausible poses, we design another experiment conducted as follows: For each test, we choose a test image from the synthetic images. The template image in this test image is warped according to pose $\mathbf{p}_t$. An ambiguous pose $\mathbf{p}_a$ is then determined from $\mathbf{p}_t$ using the functional minimization method [38]. One of the two plausi-

Table 3. Evaluation results for two feature-based approaches and the proposed direct method with undistorted test images in terms of average number of rotation error $E_{\mathbf{R}}$, translation error $E_{\mathbf{t}}$, and success rate in each test condition. The best values are highlighted in bold.

	Bump Sign			Stop Sign			Lucent			MacMini Board			Isetta			Philadelphia			Grass			Wall		
																								
	$E_{\mathbf{R}}$(°)	$E_{\mathbf{t}}$(%)	SR(%)	$E_{\mathbf{R}}$(°)	$E_{\mathbf{t}}$(%)	SR(%)	$E_{\mathbf{R}}$(°)	$E_{\mathbf{t}}$(%)	SR(%)	$E_{\mathbf{R}}$(°)	$E_{\mathbf{t}}$(%)	SR(%)	$E_{\mathbf{R}}$(°)	$E_{\mathbf{t}}$(%)	SR(%)	$E_{\mathbf{R}}$(°)	$E_{\mathbf{t}}$(%)	SR(%)	$E_{\mathbf{R}}$(°)	$E_{\mathbf{t}}$(%)	SR(%)	$E_{\mathbf{R}}$(°)	$E_{\mathbf{t}}$(%)	SR(%)
SIFT	100	54.8	10	69.2	35.3	38	30.1	16.5	72	28.0	13.4	78	20.6	16.2	82	13.5	4.57	90	97.3	212	24	17.1	27.3	86
ASIFT	72.1	24.3	22	5.07	**0.74**	**96**	**1.90**	**0.38**	**100**	6.37	2.59	**96**	2.08	**0.49**	98	**1.16**	**0.35**	**100**	51.2	16.7	52	**2.76**	**0.36**	**96**
Direct	**3.84**	**0.80**	**96**	**2.81**	0.91	**98**	2.62	1.06	98	**5.37**	**2.56**	**96**	**1.49**	0.87	**100**	1.95	0.87	**100**	**2.87**	**1.06**	**96**	6.68	1.64	94

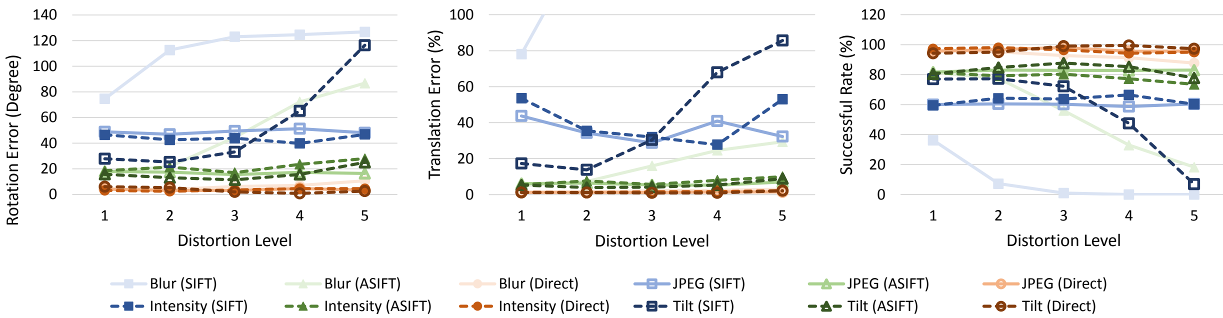


Figure 7. Experimental results on synthetic data under varying conditions.

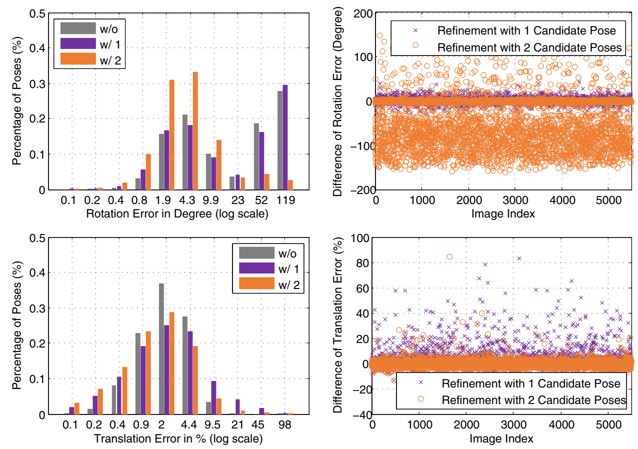
ble poses $\mathbf{p}_a^{'}$ is randomly chosen and added some Gaussian noise. Later the refinement approach is applied to $\mathbf{p}_a^{'}$ for estimating the pose of the warped template image. Finally, we compute $E_{\mathbf{R}}$ and $E_{\mathbf{t}}$ of both the initial noisy pose $\mathbf{p}_a^{'}$ and the refined pose $\mathbf{p}_r$ according to $\mathbf{p}_t$.

Thus, if the proposed refinement approach is able to disambiguate the plausible pose $\mathbf{p}_a^{'}$, the rotation error can be reduced significantly. We compare the proposed refinement method to the approach with only one candidate pose in Algorithm 1, and present the results in Figure 8. The success rate before refinement, refinement with one candidate pose, and refinement with two candidate poses are 51.26%, 51.19% and 90.34%, respectively. These results show that the proposed refinement method can help improve estimation accuracy and address pose ambiguity problem effectively. We also note that the pose estimates can be further improved by filtering the results from single images.

5.2. Real Images

In this experiment we investigate the performance of proposed method on a benchmark dataset by Gauglitz *et al*. [12], originally used to evaluate the tracking-related algorithms. This dataset consists of 96 videos with a total of 6889 frames including 6 different templates with 16 different conditions. The frame size in this data set is 640 × 480 pixels, and we resize the template to 570 × 420 pixels. It is definitely a challenging database for the pose estimation problem due to significant viewpoint change, drastic illumination difference, and noisy camera images.

The complete comparison results of two feature-based methods and the proposed direct algorithm are shown in Table 4. While OPnP performs well in pose estimation, the success hinges on whether feature can be well matched. The difficulty of the feature-based approaches to cope with motion blur is apparent. On the other hand, the proposed method can still estimate poses with low translation error and slightly higher rotation error under severe blur condi-



(a) Distribution of errors. (b) Difference of errors.

Figure 8. The proposed method without refinement (w/o), refinement with one candidate (w/ 1), and refinement with two candidates (w/ 2) are evaluated. (a) The rotation errors are reduced drastically in the ambiguous cases, but the translation errors are relatively not, because the translation terms of ambiguous poses are quite similar in most cases. (b) The difference of pose errors before and after applying two kinds of refinement approaches. While the proposed refinement approach can disambiguate the object pose effectively, approach with only one candidate pose suffers from the risk of getting trapped into local minimum.

Table 4. Experimental results of the visual tracking dataset [12] under different conditions. The SIFT-based (S), ASIFT-based (A), and the proposed direct (D) methods are evaluated under different conditions (uc: unconstrained; pn: panning; rt: rotation; pd: perspective distortion; zm: zoom; mX: motion blur level X, X = 1...9; ls: static lighting; ld: dynamic lighting). The best results in each condition are highlighted in bold.

		Bricks			Building			Mission			Paris			Sunset			Wood		
		$E_R(°)$	$E_t(\%)$	SR(%)	$E_R(°)$	$E_t(\%)$	SR(%)	$E_R(°)$	$E_t(\%)$	SR(%)	$E_R(°)$	$E_t(\%)$	SR(%)	$E_R(°)$	$E_t(\%)$	SR(%)	$E_R(°)$	$E_t(\%)$	SR(%)
uc	S	73.3	126	**41.6**	110	183	11.2	63.6	34.5	48.4	52.2	103	56.2	113	205	2.00	124	350	0.00
	A	75.1	19.2	35.8	82.6	25.5	29.4	49.3	15.2	57.0	11.1	3.80	91.0	63.1	19.5	40.4	80.0	**24.1**	28.0
	D	**59.9**	**15.1**	41.0	**16.8**	**10.5**	**83.8**	**17.0**	**8.33**	**84.4**	**1.30**	**1.18**	**99.4**	**6.34**	**10.3**	**57.2**	**73.1**	43.9	**28.2**
pn	S	14.8	3.96	**90.0**	124	83.4	0.00	33.2	26.2	70.0	15.1	13.2	74.0	116	87.2	0.00	124	138	0.00
	A	36.6	13.4	64.0	109	46.1	0.00	7.45	**1.06**	90.0	16.0	1.23	60.0	70.4	25.1	22.0	112	41.7	0.00
	D	**7.84**	**1.13**	84.0	**2.59**	**1.57**	**100.0**	**5.31**	1.92	**100**	**3.82**	**1.05**	**96.0**	**14.2**	**1.32**	**84.0**	**47.0**	**7.51**	**60.0**
rt	S	**1.08**	**0.28**	**100**	72.9	80.3	34.0	2.65	**0.39**	98.0	3.59	0.67	98.0	44.3	22.5	28.0	126	144	0.00
	A	3.76	0.61	98.0	29.6	10.0	52.0	**2.02**	0.44	**100**	**1.34**	**0.35**	**100**	15.1	1.58	58.0	**5.38**	**1.38**	**94.0**
	D	27.7	71.5	76.0	**10.6**	**4.90**	**90.0**	3.05	0.99	**100**	1.88	0.51	**100**	**4.08**	**1.26**	**100**	110	66.3	0.00
pd	S	41.1	138	66.0	82.0	104	34.0	37.3	15.4	70.0	32.1	30.8	74.0	102	86.2	2.00	120	428	0.00
	A	40.7	**13.1**	70.0	50.2	16.8	64.0	24.5	**7.21**	80.0	**24.6**	7.45	**84.0**	43.9	**14.5**	64.0	51.0	**13.5**	**62.0**
	D	**23.0**	17.5	**82.0**	**23.8**	**21.6**	**82.0**	**16.8**	12.3	**84.0**	26.5	**24.9**	82.0	**25.8**	17.4	**78.0**	**50.1**	64.0	54.0
zm	S	**1.18**	**0.30**	**100**	95.0	128	16.0	**5.30**	**0.56**	94.0	**2.57**	**0.42**	**100**	95.5	117	14.0	111	146	8.00
	A	27.9	8.56	58.0	50.0	15.7	50.0	8.73	0.85	78.0	7.51	0.43	74.0	21.7	4.11	54.0	**58.9**	**17.1**	**42.0**
	D	30.77	40.8	66.0	**11.1**	**7.42**	**94.0**	5.76	0.93	**98.0**	2.67	0.58	**100**	**5.95**	**1.23**	**100**	98.5	61.0	0.00
m1	S	**6.23**	**0.39**	**100**	127	100	1.12	**8.80**	**0.48**	90.9	16.1	1.29	69.0	118	75.6	0.00	119	80.0	0.00
	A	65.1	34.4	39.8	113	48.0	0.00	18.9	2.06	55.7	15.7	0.86	67.8	106	37.3	1.14	95.9	46.1	0.00
	D	10.6	1.52	90.9	**16.1**	**2.07**	**85.4**	11.8	1.69	**94.3**	**6.48**	**0.67**	**100**	**21.5**	**1.18**	**44.3**	**65.8**	**4.23**	0.00
m2	S	68.6	35.5	31.1	130	52.6	2.22	13.2	4.98	**95.6**	22.9	22.9	68.2	137	263	0.00	126	125	0.00
	A	106	46.8	6.67	104	43.2	0.00	18.6	2.62	57.8	17.1	1.33	63.6	125	47.9	0.00	102	43.7	0.00
	D	**16.0**	**2.72**	**84.4**	**14.7**	**2.05**	**56.7**	**12.7**	**1.03**	91.1	**7.69**	**0.98**	**100**	**20.0**	**1.23**	**51.1**	**73.1**	**5.03**	**2.22**
m3	S	123	88.1	9.38	141	67.0	0.00	89.9	46.2	18.8	93.2	429	16.1	128	87.0	0.00	130	221	0.00
	A	99.4	43.6	6.25	98.7	44.2	0.00	21.1	4.78	71.9	20.0	1.56	54.8	119	47.0	0.00	111	50.0	0.00
	D	**17.2**	**3.05**	**90.6**	**17.1**	**2.37**	**75.0**	**13.3**	**1.90**	**93.8**	**10.1**	**0.96**	**100**	**23.4**	**1.19**	**33.3**	**78.0**	**3.93**	0.00
m4	S	124	104	8.70	127	76.9	0.00	99.1	52.4	13.0	102	354	4.55	131	77.3	0.00	122	154	0.00
	A	106	42.1	4.35	111	43.3	0.00	96.7	38.8	8.70	37.5	9.72	54.5	112	58.1	0.00	100	42.3	0.00
	D	**27.4**	**3.43**	**60.9**	**30.0**	**5.17**	**56.5**	**15.8**	**1.84**	**78.3**	**9.31**	**0.63**	**100**	**25.7**	**1.12**	**17.4**	**68.4**	**4.85**	0.00
m5	S	115	104	15.8	146	87.5	0.00	91.2	537	15.8	111	216	11.1	139	74.9	0.00	140	104	0.00
	A	93.7	42.3	10.5	109	46.6	0.00	92.8	40.3	15.8	92.4	42.8	11.1	112	50.2	5.00	101	50.4	0.00
	D	**42.5**	**7.18**	**47.4**	**35.8**	**5.09**	**66.7**	**20.7**	**2.31**	**26.3**	**8.87**	**0.64**	**100**	**25.3**	**1.32**	**25.0**	**83.6**	**5.01**	0.00
m6	S	115	121	16.7	140	111	0.00	101	79.4	16.7	103	207	16.7	128	57.7	0.00	123	249	0.00
	A	105	51.2	0.00	102	43.9	0.00	90.3	36.8	11.1	80.7	31.9	11.1	126	42.9	0.00	128	42.2	0.00
	D	**71.8**	**13.7**	**27.8**	**59.9**	**11.9**	**38.9**	**20.8**	**1.32**	**38.9**	**12.4**	**1.44**	**94.4**	**28.9**	**1.55**	**5.56**	**81.0**	**8.22**	0.00
m7	S	105	85.0	18.8	131	120	0.00	102	107	18.8	111	157	18.8	122	148	0.00	119	163	0.00
	A	109	51.2	0.00	114	40.4	0.00	90.4	36.7	12.5	94.4	35.8	18.75	122	48.3	0.00	115	43.7	0.00
	D	**44.4**	**5.92**	**25.0**	**70.0**	**17.1**	**31.3**	**21.0**	**1.04**	**37.5**	**14.3**	**1.87**	**68.8**	**24.5**	**1.57**	**43.8**	**90.1**	**4.97**	0.00
m8	S	125	195	13.3	133	180	0.00	106	50.5	20.0	102	191	20.0	132	74.4	0.00	127	205	0.00
	A	104	35.0	6.67	98.4	45.2	0.00	70.8	33.8	13.3	71.9	36.5	20.0	119	54.8	0.00	102	46.4	0.00
	D	**72.8**	**8.27**	**20.0**	**83.1**	**23.6**	**28.6**	**23.7**	**2.16**	**40.0**	**16.5**	**1.90**	**60.0**	**28.9**	**1.76**	**13.3**	**75.3**	**3.58**	0.00
m9	S	108	70.9	14.3	130	98.2	0.00	99.6	36.5	15.4	93.6	183	14.3	135	160	0.00	122	91.6	0.00
	A	109	49.5	0.00	92.6	41.2	0.00	82.7	35.6	15.4	78.7	32.6	21.4	95.9	42.3	0.00	115	48.4	0.00
	D	**73.1**	**9.64**	**21.4**	**71.1**	**21.5**	**42.9**	**23.4**	**1.08**	**46.2**	**18.0**	**1.75**	**42.9**	**32.7**	**1.91**	**7.14**	**83.9**	**3.56**	0.00
ls	S	58.6	78.1	50.0	76.4	91.6	40.0	61.0	26.6	50.0	55.5	33.3	56.3	108	44.3	8.75	125	114	0.00
	A	**24.8**	**9.21**	**75.0**	69.5	21.4	37.5	**0.89**	**0.44**	**100**	**0.90**	**0.61**	**100**	39.7	10.4	61.25	52.5	18.3	51.3
	D	52.9	33.2	56.25	**0.96**	**0.66**	**100**	1.92	0.97	**100**	1.30	0.80	**100**	**6.38**	**6.89**	**77.5**	**5.37**	**3.67**	**92.25**
ld	S	86.3	317	29.0	111	122	14.0	94.4	40.4	26.0	84.3	76.8	27.0	125	78.6	2.00	126	182	0.00
	A	**45.9**	**14.1**	**58.0**	73.6	23.6	32.0	3.80	1.04	98.0	**0.88**	**0.40**	**100**	55.4	**19.2**	45.0	51.3	20.2	50.0
	D	94.2	69.0	23.0	**6.59**	**8.94**	**94.0**	**1.90**	**0.84**	**100**	17.6	16.0	84.0	**21.4**	25.8	**47.0**	**23.4**	**15.6**	**82.0**

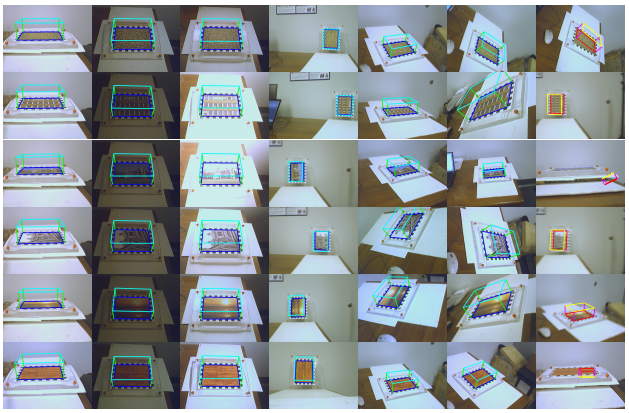


Figure 9. Estimation results by the proposed direct method on real images under different conditions. The success cases are represented with rendered cyan boxes, and the failure cases are represented with rendered magenta boxes.

tion. As motion blurs are likely occur in AR applications, the proposed algorithm can be better applied to estimate 3D pose than feature-based approaches. However, if the target image appears a exetreme flat color in the camera image, our proposed method still might fail because the appearance between the template and its local patches are almost undistinguishable. Sample images rendered model with pose obtained from proposed algorithm are shown in Figure 9. Overall, the proposed direct method outperforms the feature-based approaches within an success rate 68.08%. The success rate of the SIFT-based and ASIFT-based approaches are 29.34% and 46.10% respectively.

6. Conclusion

In this paper, we propose a robust direct method for 3D pose estimation based on two main steps. First, the pose of a planar target with respect to a calibrated camera is approximately estimated using a efficient coarse-to-fine scheme. Next, we use a gradient descent search method to further refine and disambiguate the object pose. Extensive experimental evaluations on both synthetic and real datasets demonstrate the proposed algorithm performs favorably against two state-of-the-art feature-based pose estimation approaches in terms of robustness and accuracy under several varying conditions. Our future work includes extensions of the proposed algorithm on a GPGPU platform as the algorithm is highly parallelizable.

References

[1] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning from data*. AMLBook, 2012. 4

[2] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *CVPR*, 2012. 1

[3] B. Alexe, V. Petrescu, and V. Ferrari. Exploiting spatial overlap to efficiently compute appearance distances between image windows. In *NIPS*, pages 2735–2743, 2011. 2

[4] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *CVPR*, 2001. 1

[5] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3), 2008. 1

[6] Y.-T. Chi, J. Ho, and M.-H. Yang. A direct method for estimating planar projective transform. In *ACCV*. 2011. 1

[7] O. Chum and J. Matas. Matching with prosac-progressive sample consensus. In *CVPR*, 2005. 1

[8] A. Crivellaro and V. Lepetit. Robust 3d tracking with descriptor fields. In *CVPR*, 2014. 1

[9] D. Eberly. Euler angle formulas. *Geometric Tools, LLC, Technical Report*, 2008. 3

[10] J. Engel, T. Schöps, and D. Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*. 2014. 1

[11] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 5

[12] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *IJCV*, 94(3):335–360, 2011. 5, 7, 8

[13] G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE TPAMI*, 20(10):1025–1039, 1998. 1

[14] J. F. Henriques, P. Martins, R. F. Caseiro, and J. Batista. Fast training of pose detectors in the fourier domain. In *NIPS*, 2014. 1

[15] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*. 2008. 2, 6

[16] G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. In *ISMAR*, 2007. 1

[17] L. Kneip, H. Li, and Y. Seo. Upnp: An optimal o (n) solution to the absolute pose problem with universal applicability. In *ECCV*. 2014. 5

[18] S. Korman, D. Reichman, G. Tsur, and S. Avidan. Fast-match: Fast affine template matching. In *CVPR*, 2013. 1, 2, 3

[19] Z. Kukelova, M. Bujnak, and T. Pajdla. Automatic generator of minimal problem solvers. In *ECCV*. 2008. 2, 4

[20] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate O(n) solution to the pnp problem. *IJCV*, 81(2), 2009. 1, 5

[21] S. Leutenegger, M. Chli, and R. Y. Siegwart. BRISK: Binary Robust Invariant Scalable Keypoints. In *ICCV*, 2011. 1

[22] S. Li and C. Xu. Efficient lookup table based camera pose estimation for augmented reality. *Computer Animation and Virtual Worlds*, 22(1):47–58, 2011. 2

[23] S. Lieberknecht, S. Benhimane, P. Meier, and N. Navab. A dataset and evaluation methodology for template-based tracking algorithms. In *ISMAR*, 2009. 6

[24] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele. Real-time image-based 6-dof localization in large-scale environments. In *CVPR*, 2012. 1

[25] L.-K. Liu and E. Feig. A block-based gradient descent search algorithm for block motion estimation in video coding. *IEEE TCSVT*, 6(4):419–422, 1996. 2, 5

[26] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004. 1, 5

[27] C. Lu, G. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *IEEE TPAMI*, 22(6):610–622, 2000. 2

[28] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981. 1

[29] E. Malis. Improving vision-based control using efficient second-order minimization techniques. In *ICRA*, 2004. 1

[30] R. Mur-Artal and J. D. Tardós. Fast relocalisation and loop closing in keyframe-based slam. In *ICRA*, 2014. 1

[31] D. Oberkampf, D. F. DeMenthon, and L. S. Davis. Iterative pose estimation using coplanar points. In *CVPR*, 1993. 2

[32] O. Pele and M. Werman. Accelerating pattern matching or how much can you slide? In *ACCV*. 2007. 2

[33] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: an efficient alternative to sift or surf. In *ICCV*, 2011. 1

[34] G. Schweighofer and A. Pinz. Robust Pose Estimation from a Planar Target. *IEEE TPAMI*, 28(12), 2006. 1, 2, 5

[35] H.-Y. Shum and R. Szeliski. Construction of panoramic image mosaics with global and local alignment. In *Panoramic Vision*, pages 227–268. 2001. 1

[36] P.-C. Wu, Y.-H. Tsai, and S.-Y. Chien. Stable pose tracking from a planar target with an analytical motion model in real-time applications. In *MMSP*, 2014. 2

[37] G. Yu and J.-M. Morel. Asift: A new framework for fully affine invariant image comparison. *Image Processing On Line*, 2011. 1, 5

[38] Y. Zheng, Y. Kuang, S. Sugimoto, K. Astrom, and M. Okutomi. Revisiting the PnP Problem: A Fast, General and Optimal Solution. In *ICCV*, 2013. 1, 2, 4, 5, 6

[39] C. Zhu, X. Lin, and L.-P. Chau. Hexagon-based search pattern for fast block motion estimation. *IEEE TCSVT*, 12(5):349–355, 2002. 2

[40] S. Zhu and K.-K. Ma. A new diamond search algorithm for fast block-matching motion estimation. *IEEE TIP*, 9(2):287–290, 2000. 2