A Direct Method for Robust Model-Based 3D Object Tracking from a Monocular RGB Image

Byung-Kuk Seo(✉) and Harald Wuest

Fraunhofer IGD, Darmstadt, Germany
byung-kuk.seo@igd-extern.fraunhofer.de, harald.wuest@igd.fraunhofer.de

Abstract. This paper proposes a novel method for robust 3D object tracking from a monocular RGB image when an object model is available. The proposed method is based on direct image alignment between consecutive frames over a 3D target object. Unlike conventional direct methods that only rely on image intensity, we newly model intensity variations using the surface normal of the object under the Lambertian assumption. From the prediction about image intensity in this model, we also employ a constrained objective function, which significantly alleviates degradation of the tracking performance. In experiments, we evaluate our method using datasets that consist of test sequences under challenging conditions, and demonstrate its benefits compared to other methods.

Keywords: Pose estimation · Object tracking · Model-based · Direct image alignment · Motion model

1 Introduction

3D tracking (or 6D pose estimation) of target objects is a crucial issue in computer vision, robotics, and augmented reality. Over the last decade, numerous methods have been proposed and successfully demonstrated for 3D object tracking. Despite that, achieving accurate, robust, and fast tracking is still challenging in everyday environments where there exists a large range of 3D objects under various backgrounds, illuminations, occlusions, and motions.

In early methods, feature points have prominently been used to handle pose estimation problems of 2D/3D objects [27,30], but such feature-based methods require that the objects have sufficient texture on their surfaces. For poorly textured 3D objects, strong edges have been popular and are still promising in many industrial applications [7,11]. However, they are often troublesome against heavy background clutter due to the nature of edge property. As recent RGBD cameras enable to obtain more dense information about 3D scenes including objects,

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-49409-8_48) contains supplementary material, which is available to authorized users.

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part III, LNCS 9915, pp. 551–562, 2016.
DOI: 10.1007/978-3-319-49409-8_48

Fig. 1. Tracking results using the proposed method under challenging conditions (green lines visualize object models projected on images with estimated poses). (Color figure online)

RGBD-based methods have been boosted to tackle challenging pose estimation problems [15,18,29]. Nevertheless, RGBD cameras have several issues need to be considered: depth information is quite noisy and only available within limited ranges with material difficulties (such as specular and transparent materials). Moreover, they are not commonly supported yet in real application domains, compared to RGB cameras.

On the other hand, direct methods have been attractive because they allow that rich information in an image can be contributed to pose estimation, instead of being limited by local features [1,4,5,8,12]. In direct methods, the brightness (intensity) constancy is commonly assumed, but it is often violated by intensity variations, which are induced by illumination changes, surface reflectance properties, or even changes in camera gain. In this paper, we propose a direct method for robust 3D object tracking from a monocular RGB image when an object model is available. In our method, we model intensity variations by deriving differential entities from image formation under the Lambertian assumption, and define a compensation parameter using the surface normal of a 3D target object. From the prediction about image intensity in this model, we also employ a constrained objective function, resulting in suppressing the error accumulation and converging with less iteration. In addition, we provide new datasets that comprise challenging conditions such as partial occlusions, background clutters, and illumination changes (see Fig. 3), in order to evaluate our method in an intensive manner and explicitly demonstrate its advantages.

Our main contribution is a novel direct method based on an elaborate motion model between consecutive frames over a 3D target object, leading to robust 3D object tracking as shown in Fig. 1. Here, we clarify that this paper focuses on robust frame-to-frame pose estimation when an initial pose is only given without preparing or training a set of reference images; thus, the initialization (or reinitialization) issue is out of scope in this paper, and if available, its relevant methods can be combined with the proposed method.

2 Related Work

In the literature, lots of methods have been proposed for dealing with 6D pose estimation problems. As more relevant works in our interest, we briefly highlight model-based approaches for 3D object tracking in a monocular RGB view where 3D knowledge of a target object, such as a 3D model or a set of registered patches, is known as *a prior*.[1]

In a typical way of model-based approaches, the pose estimation is performed by establishing 3D-2D correspondences between 3D knowledge and 2D observation in an image (such as feature points [27,30] or edges [7,11]); thus, most of methods in this manner highly depend on how to extract and match distinctive local features, which are not trivial tasks under challenging conditions.

Region-based methods have been of interest in terms of 6D pose estimation. In particular, level set based segmentation methods have successfully been demonstrated for 3D object tracking [9,24]. These methods follow a general statistical representation of a level set function and evolve a contour of a 3D model over a camera pose, without considering the correspondence problem. In principle, however, such region segmentation requires intensive tasks because the contour is evolved in an infinite-dimensional space, and it is also difficult to guarantee good segmentation results according to scene complexity, even though these issues have been improved [22,26].

To date, direct methods have actively been adopted for 2D object tracking. Similar to region-based methods, direct methods exploit rich information in an image instead of local features, but they directly align image intensity with one of registered templates. As a pioneer work, the Lucas and Kanade framework [19] has played a prominent role and has brought about many variants with improvements; for example, efficient optimization algorithms [1,4] or robust similarity measures [10,16,23,25]. With the availability of 3D knowledge of a target object, direct methods have also successfully been applied for 3D object tracking [5,8].

Even though direct methods have been promising for pose estimation, its underlying assumption (i.e., brightness constancy) is often violated by intensity variations. To tackle such errors, intensity variations have been modeled not only in a typical form, represented by a multiplicative term [32], an additive term [6], or both terms [2,17,20], but also in a more generalized framework, described by complex physical processes [14]. In an alternative way, direct methods have also employed learning-based approaches to handle 3D objects with complex shapes as well as challenging conditions such as illumination changes and occlusions [21,28], but they require computationally demanding training stages with a significant number of data, which are quite cumbersome in practice.

[1] In model-based approaches, simultaneous localization and mapping-based methods can be considered for pose estimation in unknown 3D environments, but they are not suitable for 3D object tracking that aims at estimating poses relative to target objects; thus, we do not detail methods in this category.

3 Proposed Method

We start by briefly defining fundamental relations between consecutive frames over a 3D target object under a camera motion in Sect. 3.1. The proposed method is then detailed in Sect. 3.2 (modeling intensity variations) through Sect. 3.3 (objective function and optimization).

3.1 Motion Model Between Consecutive Frames

Consider a camera is moving relative to a 3D target object as shown in Fig. 2. Under the perspective projection Π, the 3D point on the object surface in the camera coordinate system $\mathbf{s}^{\mathcal{C}} = (X^{\mathcal{C}}, Y^{\mathcal{C}}, Z^{\mathcal{C}})^\top$ is mapped to the 2D point on the image plane $\mathbf{u} = (u, v)^\top$:

$$\mathbf{u} = \Pi(\mathbf{s}^{\mathcal{C}}) = \left(\frac{X^{\mathcal{C}} f_u}{Z^{\mathcal{C}}} + u_0, \frac{Y^{\mathcal{C}} f_v}{Z^{\mathcal{C}}} + v_0 \right)^\top, \tag{1}$$

where (f_u, f_v) are the focal lengths of the camera, and (u_0, v_0) are the principal points of the camera. Under the rigid body transformation $\mathcal{G} \in \mathrm{SE}(3)$, the $\mathbf{s}^{\mathcal{C}}$ is transformed from the 3D point on the object surface in the world coordinate system $\mathbf{s}^{\mathcal{O}}$:

$$\mathbf{u} = \Pi(\mathcal{G}(\mathbf{s}^{\mathcal{O}}; \boldsymbol{\xi})) \quad \text{with} \quad \boldsymbol{\xi} = (\boldsymbol{\omega}^\top, \boldsymbol{\tau}^\top)^\top, \tag{2}$$

where $\boldsymbol{\xi}$ is the parameter associated with the Lie algebra $\mathrm{se}(3)$[2], described by the translational velocity $\boldsymbol{\omega}$ and the rotational velocity $\boldsymbol{\tau}$. In a monocular RGB view, the $\mathbf{s}^{\mathcal{O}}$ is in general unknown, but it can be determined when the object model $\mathbf{m}^{\mathcal{O}}$ is given; thus, Eq. (2) can be rewritten as

$$\mathbf{u} = \Pi(\mathcal{G}(\mathbf{s}^{\mathcal{O}} \rightarrow \mathbf{m}^{\mathcal{O}}(\mathbf{u}); \boldsymbol{\xi})) \quad \text{with} \quad \mathbf{m}^{\mathcal{O}}(\mathbf{u}) = \mathcal{G}^{-1}(\Pi^{-1}(\mathbf{u}, d^{\mathcal{C}}); \boldsymbol{\xi}), \tag{3}$$

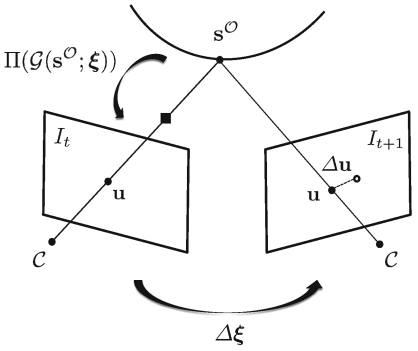

Fig. 2. Illustration and notations of image information between consecutive frames over a 3D target object under a camera motion.

[2] Since the relation of the $\mathcal{G}$ to the camera pose is straightforward, this paper retains the same notation $\boldsymbol{\xi}$ for the camera pose or motion.

where $d^{\mathcal{C}}$ is the depth buffer to be rendered using the object model with respect to the $\boldsymbol{\xi}$.

Under the camera motion between consecutive frames $\Delta\boldsymbol{\xi}$, on the other hand, the 2D point on the image plane at time t is mapped to the corresponding 2D point on the image plane at time $t+1$, and it can be defined as the following consistency constraint (under the brightness constancy assumption):

$$I_{t+1}(\mathbf{u}+\Delta\mathbf{u}) = I_t(\mathbf{u}), \tag{4}$$

where $I_t(\mathbf{u})$ is the image intensity of the 2D point $\mathbf{u}$ at time t, and $\Delta\mathbf{u}$ is the displacement of the point on the image plane. Since the $\Delta\mathbf{u}$ can be represented by the $\Delta\boldsymbol{\xi}$ from Eq. (2), it can also be redefined as a consistency constraint with respect to the $\Delta\boldsymbol{\xi}$ by coupling both of relations (Eqs. (3) and (4)):

$$I_{t+1}(\Pi(\mathcal{G}(\mathbf{m}^{\mathcal{O}}(\mathbf{u}); \boldsymbol{\xi}+\Delta\boldsymbol{\xi}))) = I_t(\mathbf{u}). \tag{5}$$

3.2 Modeling Intensity Variations

Assuming that a 3D target object is rigid and has a Lambertian surface; the object is illuminated by a distant point light; and a camera undergoes a rigid motion relative to the target object, the observed image intensity at a 2D point $\mathbf{u}$ on the image plane is given by

$$I(\mathbf{u}) = \sigma(\mathbf{s})\mathbf{n}(\mathbf{s})^{\top}\mathbf{l}, \tag{6}$$

where σ is the surface albedo, $\mathbf{n}$ is the unit surface normal, $\mathbf{l}$ is the unknown scaled light vector representing the light direction and intensity, and $\mathbf{s}$ is the surface point corresponding to the $\mathbf{u}$. For the differential change of the intensity, we take the total derivative of the intensity with respect to t:

$$\frac{dI}{dt} = (\mathbf{n}^{\top}\mathbf{l})\frac{d\sigma}{dt} + \sigma\frac{d}{dt}(\mathbf{n}^{\top}\mathbf{l}). \tag{7}$$

Since the $\frac{d\sigma}{dt}$ is an entity on the surface, it is constant over time, and the differential change of the intensity is then simplified to $dI = \sigma d(\mathbf{n}^{\top}\mathbf{l})$. Therefore, the intensity variations between consecutive frames ΔI_t^{t+1} can be described by

$$\begin{aligned}\Delta I_t^{t+1} &= \sigma\Delta(\mathbf{n}^{\top}\mathbf{l})_t^{t+1} = \sigma(\mathbf{n}_{t+1}^{\top}\mathbf{l}_{t+1} - \mathbf{n}_t^{\top}\mathbf{l}_t) \\ &= \sigma(\mathbf{n}_t^{\top}\mathbf{l}_t)(\kappa-1) = I_t(\kappa-1),\end{aligned} \tag{8}$$

where κ is the compensation parameter, given by $\kappa = \mathbf{n}_t^{\top}\mathbf{l}_{t+1}/\mathbf{n}_t^{\top}\mathbf{l}_t$ under the given object rigidity assumption ($\mathbf{n}_{t+1} = \mathbf{n}_t$). Here the conventional brightness constancy assumption is satisfied when $\kappa = 1$.

In general, the illumination is unknown, and its complete modeling is nearly impossible. In this model, however, the κ can be estimated using the surface normal of the object:

$$\kappa = \frac{\mathbf{n}_t^{\top}\mathbf{l}_{t+1}}{\mathbf{n}_t^{\top}\mathbf{l}_t} \approx \frac{\mathcal{E}[I_{t+1}|\mathbf{n}_t]}{\mathcal{E}[I_t|\mathbf{n}_t]}, \tag{9}$$

where $\mathcal{E}[I|\mathbf{n}]$ is the conditional expectation of I given $\mathbf{n}$, modeled by the first-order approximation of the radiance model from any Lambertian object under the general distant light distribution [3]: $I \approx \mathcal{E}[I|\mathbf{n}] = \sigma(l_0 + n_x l_x + n_y l_y + n_z l_z)$, where $\mathbf{n} = (n_x, n_y, n_z)^\top$ and $\mathbf{l} = (l_x, l_y, l_z)^\top$ are the surface normal and illumination vectors, and l_0 is the additional offset. Here, the conditional expectation is computed using the multivariate linear regression when the surface normal of the object is given.

3.3 Objective Function and Optimization

In the proposed method, the 6D pose estimation is formulated by the minimization of an objective function, including an error term and a stability term. By combining Eqs. (5) and (8), the error term $e_{\mathcal{I}}(\mathbf{u};\boldsymbol{\xi})$ is defined as

$$e_{\mathcal{I}}(\mathbf{u};\boldsymbol{\xi}) = I_{t+1}(\Pi(\mathcal{G}(\mathbf{m}^{\mathcal{O}}(\mathbf{u});\boldsymbol{\xi}))) - \kappa I_t(\mathbf{u}). \tag{10}$$

From the prediction about image intensity in Eq. (9), on the other hand, it follows that, for any function Θ of $\mathbf{n}$,

$$\begin{aligned}\mathcal{E}[(I-\Theta(\mathbf{n}))^2] &= \mathcal{E}[(I-\mathcal{E}[I|\mathbf{n}]+\mathcal{E}[I|\mathbf{n}]-\Theta(\mathbf{n}))^2] \\ &= \mathcal{E}[(I-\mathcal{E}[I|\mathbf{n}])^2] + \mathcal{E}[(\mathcal{E}[I|\mathbf{n}]-\Theta(\mathbf{n}))^2] \\ &\geq \mathcal{E}[(I-\mathcal{E}[I|\mathbf{n}])^2],\end{aligned} \tag{11}$$

where the cross term is zero, so that we define the stability term $e_{\mathcal{S}}(\mathbf{u};\boldsymbol{\xi})$ as follows:

$$e_{\mathcal{S}}(\mathbf{u};\boldsymbol{\xi}) = I_{t+1}(\Pi(\mathcal{G}(\mathbf{m}^{\mathcal{O}}(\mathbf{u});\boldsymbol{\xi}))) - \mathcal{E}[I_t|\mathbf{n}](\mathbf{u}). \tag{12}$$

Therefore, the minimization becomes

$$\min_{\boldsymbol{\xi}} \sum_{\mathbf{u}\in\mathcal{R}} \psi\big(e_{\mathcal{I}}(\mathbf{u};\boldsymbol{\xi})^2 + \gamma(\kappa) e_{\mathcal{S}}(\mathbf{u};\boldsymbol{\xi})^2\big), \tag{13}$$

where $\Psi(\cdot)$ is the robust estimator to penalize outliers, $\gamma(\kappa)$ is the weight function to balance both of terms, and $\mathcal{R}$ is the object region where the depth information is available. For robust estimation, we adopt the Charbonnier penalty function: $\psi(e) = (e^2 + \epsilon^2)^{0.5}$. For balancing the weight, we define an exponential decay function: $\gamma(\kappa) = \exp(-\lambda|\kappa - 1|)$, where λ is the constant.

For optimization, the minimization can be linearized using the first-order Taylor series expansion under a small camera motion $\delta\boldsymbol{\xi}$. In the form of Eqs. (10) and (12), moreover, efficient optimization algorithms [1] can be applied with less modification because the $\Pi(\mathcal{G}(\cdot))$ can be considered as a family of warps (i.e., a warp between image planes over a 3D object). In the proposed method, we adopt the forward compositional (FC) algorithm, which is efficiently compatible with our objective function, and then the minimization is written as (under the assumption that $\Pi(\mathcal{G}(\mathbf{m}^{\mathcal{O}}(\mathbf{u});0)) = \mathbf{u}$)

$$\min_{\delta\boldsymbol{\xi}} \sum_{\mathbf{u}\subseteq\mathcal{R}} \psi\Big(||e_{\mathcal{I}}(\mathbf{u};\boldsymbol{\xi}) + \mathbf{J}_{\mathcal{I}}\delta\boldsymbol{\xi}||^2 + \gamma(\kappa)||e_{\mathcal{S}}(\mathbf{u};\boldsymbol{\xi}) + \mathbf{J}_{\mathcal{I}}\delta\boldsymbol{\xi}||^2\Big), \tag{14}$$

where $\mathbf{J}_{\mathcal{T}}$ is the chain of the Jacobian matrices, detailed by

$$\nabla I_{t+1}(\Pi(\mathcal{G}(\mathbf{m}^{\mathcal{O}}(\mathbf{u});\boldsymbol{\xi})))\frac{\partial \Pi(\mathcal{G}(\mathbf{m}^{\mathcal{O}}(\mathbf{u});\boldsymbol{\xi}))}{\partial \boldsymbol{\xi}}\Big|_{\boldsymbol{\xi}=0}. \tag{15}$$

Here, pixels with small gradient are filtered out because they do not much contribute to the optimization, and then remaining pixels are regularly sampled with a grid ($\mathbf{u} \subseteq \mathcal{R}$). The final pose is therefore determined by iteratively updating the latest pose with the estimated motion until the norm of the estimated parameters is small or the maximum number of iterations is reached.

4 Experiments

4.1 Details on Implementation

To handle large motions, we adopted a multiscale strategy. In our implementation, we used four levels (the image resolution at the finest level was 640×480), and each level was downsampled by a factor of two with a Gaussian smoothing (5×5 kernel) and a bilinear interpolation. The depth and surface normal information were not scaled down to avoid interpolation across boundaries, but they were directly interpolated from ones at the finest level. The optimization was started at the coarsest level, and the estimated pose was used as the initial pose for the next fine level. The pixel filtering and sampling were only performed at the finest level. For all evaluations, several parameters were set: $\lambda = 1.0, \epsilon = 0.001$, the minimum magnitude of the gradient = 0.1, the grid interval = 4 pixels, the minimum norm of the estimated parameters = 10e-6, and the maximum number of iterations = 200. Here, the grid interval was adaptively set (with an increment or decrement of one pixel) relative to the object area, which is changed according to the distance between the camera and object. In addition, several steps to acquire information, such as depth, surface normal, object silhouette, and image gradient, were implemented using the rendering pipeline of a GPU.

4.2 Datasets

For intensive evaluations, we created new datasets that provide test sequences (RGB images), ground truth poses, camera intrinsic parameters, and 3D object models. As target objects, we chose two 3D objects (**Gear** and **Wheel**) that have complex shapes and no texture, rather than other objects that have common shapes (like a box or cup) and/or sufficient texture. For their 3D models, wireframe models were prepared without texture maps. Test sequences were captured with various camera motions under three different conditions such as controlled scene (**Seq1**), partial occlusions and background clutters (**Seq2**), and illumination changes (**Seq3**). For obtaining ground truth poses, each object was manually registered on multiple ARUCO markers [13]. Prior to capturing test sequences, the camera (standard USB RGB camera) was calibrated once. Automatic camera settings related to exposure time, gain, and white balance were not controlled during the capturing, except for an automatic focus. Figure 3 shows examples of test sequences in our datasets.

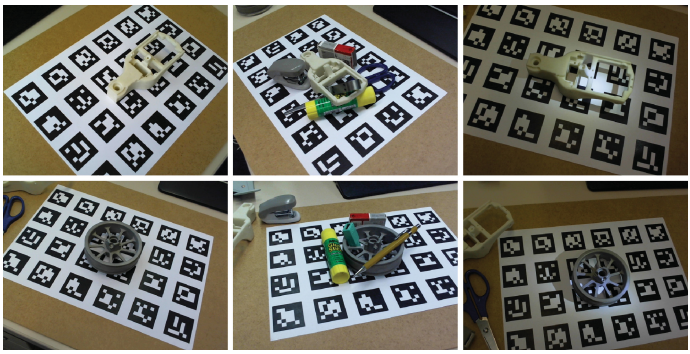

Fig. 3. Examples of test sequences in our datasets: (Top-Row) Gear-Seq1, Seq2, Seq3, (Bottom-Row) Wheel-Seq1, Seq2, Seq3.

4.3 Evaluations and Comparisons

We evaluated our method with several different methods, which can be derived from both terms in our objective function, in order to validate its benefits:

- BCC ($\kappa = 1$ and $\gamma(\kappa) = 0$): An objective function has a single error term based on the brightness consistency constraint.
- BCC+S ($\kappa = 1$ and $\gamma(\kappa) = 1$): An objective function has an error term based on the brightness consistency constraint and a stability term, where both terms are equally weighted.
- Ours (κ and $\gamma(\kappa)$ are estimated): An objective function has an error term, compensated by the κ and a stability term, weighted by the $\gamma(\kappa)$.

Here, these methods were tested in combination with the FC algorithm and the efficient second order minimization (ESM) algorithm [4] as well, wheter a better convergence rate can be expected. We also compared our method with edge-based tracking using a Gaussian mixture model (EBT-GM) [31], which is one of promising approaches in model-based 3D object tracking. Note that, in all tests, the initial poses were set with the ground truth poses.

For evaluations, we computed the average distance (AD) of all model points transformed with the estimated pose and the ground truth pose, which was defined in [15]. In this metric, we decided that the estimated pose was correct when the average distance was below 10 % of the object model diameter and calculated the success rate. To detail error profiles, we also computed the distances of rotation and translation parameters between the estimated pose and the ground truth pose. In addition, we computed the average processing times and the average iteration numbers in each case to examine the runtime performance and computational efficiency.

Table 1 summarizes results of our evaluations. Overall, the proposed method consistently performed well in every case. In particular, it outperformed other methods on challenging scenes (Seq2 and Seq3). The BCCs were often drifted and unstable due to the error accumulation (some details are shown by error

Table 1. Evaluation results: (First Rows) success rates based on the AD criterion [15] and (Second Rows) average processing times (ms) (the highest scores in the success rates are bold; asterisks denote that the tracking totally failed from certain sequences; and numbers in parentheses indicate the total numbers of test sequences).

Method	Gear			Wheel			Average
	Seq1 (850)	Seq2 (934)	Seq3 (1046)	Seq1 (811)	Seq2 (981)	Seq3 (941)	
BCC$_{(FC)}$	0.701	0.320	0.165	0.459	0.235	0.180*	0.344
	32.74	32.49	34.77	32.49	32.51	33.51	33.08
BCC$_{(ESM)}$	0.611	0.338	0.079	0.859	0.411	0.054	0.392
	33.43	30.41	32.86	32.73	32.42	37.07	33.15
BCC+S$_{(FC)}$	0.961	0.916	0.155*	**1.000**	0.919	0.410*	0.727
	30.86	29.06	34.10	30.55	30.27	31.98	31.14
Ours$_{(FC)}$	0.962	0.919	0.992	**1.000**	0.920	**1.000**	**0.966**
	46.28	41.00	49.11	44.20	44.44	46.43	45.24
Ours$_{(ESM)}$	0.974	**0.941**	**0.999**	**1.000**	**0.955**	0.868	0.956
	41.90	38.07	45.09	39.88	37.30	42.71	40.83
EBT-GM [31]	**0.981**	0.027*	0.770*	0.352*	0.004*	0.191*	0.388
	23.23	44.02	22.05	44.00	61.13	40.15	39.10

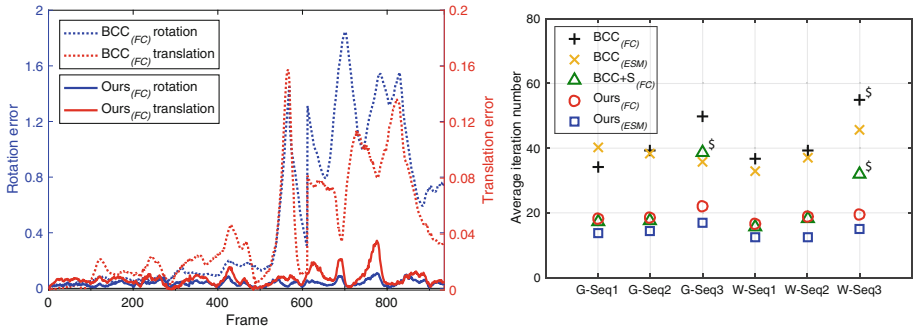

Fig. 4. (Left) Error profiles of BCC$_{(FC)}$ and Ours$_{(FC)}$ in Gear-Seq2, (Right) Average iteration numbers (asterisks denote that the tracking totally failed from certain sequences).

profiles in Fig. 4-(Left)). The EBT-GM was very sensitive to partial occlusions, background clutters, and even object clutters (e.g., which are caused by near edges in thin parts of the objects), so that in most of cases, it totally failed from certain sequences. From results of the BCC+S, it was verified that both of terms in our objective function contributed not only to significantly alleviate degradation of the tracking performance, but also to provide a better convergence rate (see Fig. 4-(Right)). Figures 5 and 6 show several comparison results in Seq2 and Seq3, and more results are shown in a supplementary video. On the

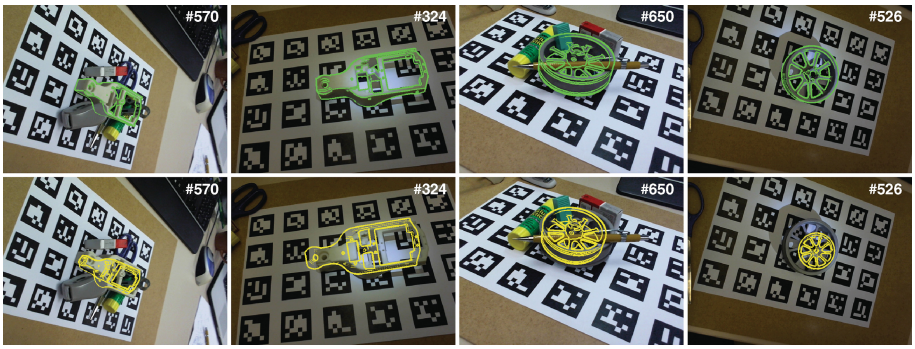

Fig. 5. Comparisons of (green lines) Ours$_{(FC)}$ and (yellow lines) BCC$_{(FC)}$ in (Left to Right) Gear-Seq2, Seq3 and Wheel-Seq2, Seq3 (numbers with hashtags indicate test sequence numbers, and color lines visualize object models projected on images with estimated poses). (Color figure online)

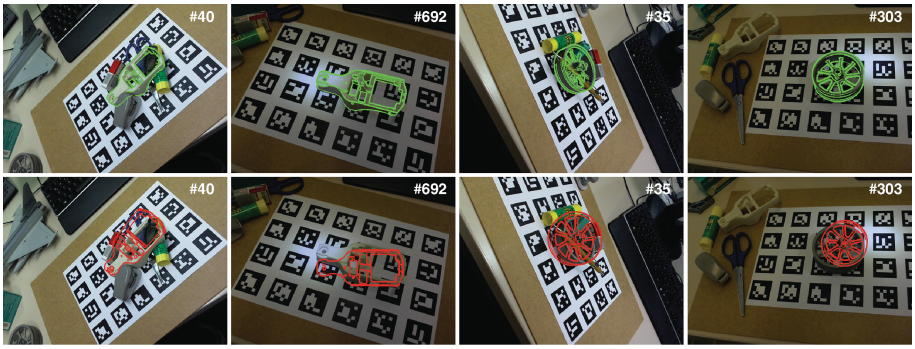

Fig. 6. Comparisons of (green lines) Ours$_{(FC)}$ and (red lines) EBT-GM in (Left to Right) Gear-Seq2, Seq3 and Wheel-Seq2, Seq3 (numbers with hashtags indicate test sequence numbers, and color lines visualize object models projected on images with estimated poses). (Color figure online)

other hand, the proposed method relatively needed more computations, but the average processing times were fairly acceptable for real-time applications (about 20–24 fps on a desktop with a 2.93 GHz CPU). Moreover, our implementation was not fully optimized and can be improved for more speed-up; for example, pixel-wise computations can obviously be parallelized using a GPU.

5 Conclusion

This paper proposed a new direct method for robust 3D object tracking. In our method, the image alignment was newly formulated by modeling intensity variations using surface normal information of an object and defining a stability term based on the prediction model about image intensity. Experimental results showed that our method successfully performed even on challenging scenes.

In this paper, we focused on 6D pose estimation of a single 3D object instance in a monocular RGB view, but it would be very interesting to explore further improvements and extensions of our method.

Acknowledgements. This work was carried out during the tenure of an ERCIM 'Alain Bensoussan' Fellowship Programme.

References

1. Baker, S., Matthews, I.: Lucas-Kanade 20 years on: a unifying framework. Int. J. Comput. Vis. **56**(3), 221–255 (2004)
2. Bartoli, A.: Groupwise geometric and photometric direct image registration. IEEE Trans. Pattern Anal. Mach. Intell. **30**(12), 2098–2108 (2008)
3. Basri, R., Jacobs, D.W.: Lambertian reflectance and linear subspaces. IEEE Trans. Pattern Anal. Mach. Intell. **25**(2), 218–233 (2003)
4. Benhimane, S., Malis, E.: Real-time image-based tracking of planes using efficient second-order minimization. In: International Conference on Robotics and Automation, pp. 943–948 (2004)
5. Caron, G., Dame, A., Marchand, E.: Direct model based visual tracking and pose estimation using mutual information. Image Vis. Comput. **32**(1), 54–63 (2014)
6. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2010)
7. Comport, A.I., Marchand, E., Pressigout, M., Chaumette, F.: Real-time markerless tracking for augmented reality: the virtual visual servoing framework. IEEE Trans. Visual Comput. Graph. **12**(4), 615–628 (2006)
8. Crivellaro, A., Lepetit, V.: Robust 3D tracking with descriptor fields. In: International Conference on Computer Vision and Pattern Recognition, pp. 3414–3421 (2014)
9. Dambreville, S., Sandhu, R., Yezzi, A., Tannenbaum, A.: Robust 3D pose estimation and efficient 2D region-based segmentation from a 3D shape prior. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 169–182. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88688-4_13
10. Dame, A., Marchand, E.: Second-order optimization of mutual information for real-time image registration. IEEE Trans. Image Process. **21**(9), 4190–4203 (2012)
11. Drummond, T., Cipolla, R.: Real-time visual tracking of complex structures. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 932–946 (2002)
12. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: large-scale direct monocular SLAM. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 834–849. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10605-2_54
13. Garrido-Jurado, S., Muñoz Salinas, R., Madrid-Cuevas, F.J., Marín-Jiménez, M.J.: Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recogn. **47**(6), 2280–2292 (2014)
14. Haussecker, H.W., Fleet, D.J.: Computing optical flow with physical models of brightness variation. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 661–673 (2001)
15. 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: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 548–562. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37331-2_42

16. Irani, M., Anandan, P.: Robust multi-sensor image alignment. In: International Conference on Computer Vision, pp. 959–966 (1998)
17. Jin, H., Favaro, P., Soatto, S.: Real-time feature tracking and outlier rejection with changes in illumination. In: International Conference on Computer Vision, pp. 684–689 (2001)
18. Krull, A., Brachmann, E., Michel, F., Ying Yang, M., Gumhold, S., Rother, C.: Learning analysis-by-synthesis for 6D pose estimation in RGB-D images. In: International Conference on Computer Vision, pp. 954–962 (2015)
19. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: International Joint Conference on Artificial Intelligence, pp. 674–679 (1981)
20. Negahdaripour, S.: Revised definition of optical flow: integration of radiometric and geometric cues for dynamic scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(9), 961–979 (1998)
21. Nguyen, M.H., de la Torre, F.: Metric learning for image alignment. Int. J. Comput. Vis. **88**(1), 69–84 (2010)
22. Prisacariu, V.A., Reid, I.D.: PWP3D: real-time segmentation and tracking of 3D objects. Int. J. Comput. Vis. **98**(3), 335–354 (2012)
23. Richa, R., Sznitman, R., Taylor, R., Hager, G.: Visual tracking using the sum of conditional variance. In: International Conference on Intelligent Robots and Systems, pp. 2953–2958 (2011)
24. Rosenhahn, B., Brox, T., Weickert, J.: Three-dimensional shape knowledge for joint image segmentation and pose tracking. Int. J. Comput. Vis. **73**(3), 243–262 (2007)
25. Scandaroli, G.G., Meilland, M., Richa, R.: Improving NCC-based direct visual tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 442–455. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33783-3_32
26. Schmaltz, C., Rosenhahn, B., Brox, T., Weickert, J.: Region-based pose tracking with occlusions using 3D models. Mach. Vis. Appl. **23**(3), 557–577 (2012)
27. Skrypnyk, I., Lowe, D.G.: Scene modelling, recognition and tracking with invariant image features. In: International Symposium on Mixed and Augmented Reality, pp. 110–119 (2004)
28. Tan, D.J., Ilic, S.: Multi-forest tracker: a chameleon in tracking. In: International Conference on Computer Vision and Pattern Recognition, pp. 1202–1209 (2014)
29. Tejani, A., Tang, D., Kouskouridas, R., Kim, T.-K.: Latent-class hough forests for 3D object detection and pose estimation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 462–477. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10599-4_30
30. Vacchetti, L., Lepetit, V., Fua, P.: Stable real-time 3D tracking using online and offline information. IEEE Trans. Pattern Anal. Mach. Intell. **26**(10), 1385–1391 (2004)
31. Wuest, H., Wientapper, F., Stricker, D.: Adaptable model-based tracking using analysis-by-synthesis techniques. In: International Conference on Computer Analysis of Images and Patterns, pp. 20–27 (2007)
32. Zhang, L., Curless, B., Hertzmann, A., Seitz, S.M.: Shape and motion under varying illumination: unifying structure from motion, photometric stereo, and multi-view stereo. In: International Conference on Computer Vision, pp. 618–625 (2003)