IJCV manuscript No.
(will be inserted by the editor)

A Robust Monocular 3D Object Tracking Method Combining Statistical and Photometric Constraints

Leisheng Zhong · **Li Zhang**

Received: date / Accepted: date

Abstract Both region-based methods and direct methods have become popular in recent years for tracking the 6-dof pose of an object from monocular video sequences. Region-based methods estimate the pose of the object by maximizing the discrimination between statistical foreground and background appearance models, while direct methods aim to minimize the photometric error through direct image alignment. In practice, region-based methods only care about the pixels within a narrow band of the object contour due to the level-set-based probabilistic formulation, leaving the foreground pixels beyond the evaluation band unused. On the other hand, direct methods only utilize the raw pixel information of the object, but ignore the statistical properties of foreground and background regions. In this paper, we find it beneficial to combine these two kinds of methods together. We construct a new probabilistic formulation for 3D object tracking by combining statistical constraints from region-based methods and photometric constraints from direct methods. In this way, we take advantage of both statistical property and raw pixel values of the image in a complementary manner. Moreover, in order to achieve better performance when tracking heterogeneous objects in complex scenes, we propose to increase the distinctiveness of foreground and background statistical models by partitioning the global foreground and background regions into a small number of sub-regions around the object contour. We demonstrate the effectiveness of the proposed novel strategies on a newly constructed real-world dataset containing different types of objects with ground-truth poses. Further experiments on several challenging public datasets also show that our method obtains competitive or even superior tracking results compared to previous works. In comparison with the recent state-of-art region-based method, the proposed hybrid method is proved to be more stable under silhouette pose ambiguities with a slightly lower tracking accuracy.

Keywords 3D object tracking · Region-based method · Direct method · Statistical constraints · Photometric constraints

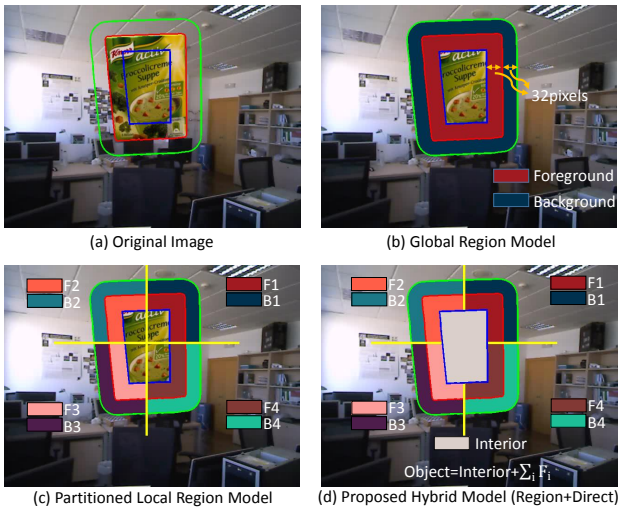
Leisheng Zhong
Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China
E-mail: zls13@mails.tsinghua.edu.cn

Li Zhang
Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China
E-mail: chinazhangli@mail.tsinghua.edu.cn

1 Introduction

3D object pose tracking is an essential problem in computer vision (Lepetit and Fua, 2005). It is the basic problem for augmented reality applications (Lima et al, 2010; Park et al, 2008), and is also widely used in robotic perception tasks (Choi and Christensen, 2010). In recent years, region-based methods (Prisacariu and Reid, 2012; Hexner and Hagege, 2016; Ren et al, 2017) and direct methods (Crivellaro and Lepetit, 2014; Seo and Wuest, 2016; Zhong et al, 2017) have gained increasing popularity because of their ability to track different kinds of objects in complex environment. Both two kinds of methods assume a known 3D object model and try to track the 6-dof pose of the object in monocular video sequences, but they extract and utilize very different information from the images:

Region-based methods focus on the statistical properties of different image regions. They iteratively evolve the projected 2D contour of the 3D object model by optimizing the pose parameters, aiming to maximize the statistical discrepancy between the foreground region and the background region (Prisacariu and Reid, 2012; Tjaden et al, 2016). Region-based methods are especially suitable for tracking homogeneous objects in simple environment, in which case the s-



(a) Original Image (b) Global Region Model (c) Partitioned Local Region Model (d) Proposed Hybrid Model (Region+Direct)

Fig. 1: Illustration of the proposed method. (a) We want to track the 6-dof pose of the box. (b) Traditional region-based methods use global foreground region and background region models. We call the ± 32 pixels band as 'histogram band' since it is used for color histogram calculation. (c) We propose to partition the global foreground and background regions into a small number of sub-regions around the object contour. F1$\sim$F4: Partitioned local foreground; B1$\sim$B4: Partitioned local background. (d) The final model we propose in this paper. We further make use of the interior region and apply photometric constraints to the whole object region (more details in Fig. 4), resulting in a robust hybrid 3D object tracker.

tatistical properties of foreground and background are distinctively different. However, in more complex situations, such as tracking heterogeneous objects in cluttered background, the performance would strongly degrade due to the variation of foreground and background statistics (Hexner and Hagege, 2016). Moreover, region-based methods suffer from pose ambiguities because they determine the 3D object pose only from 2D silhouette information. For example, region-based methods would fail when tracking a symmetrical object, because the projected silhouette will not change when the object rotates around its symmetrical axis (Prisacariu and Reid, 2012). (In contrast, direct methods would succeed in tracking a symmetrically shaped object, provided that the object has varying texture.) These failing cases indicate that using only statistical information is not sufficient for robust pose tracking in complex situations.

On the contrary, direct methods focus on raw pixel values instead of image statistics. They optimize the pose parameters by directly and densely aligning consecutive video frames over a 3D object model to minimize the photometric error of corresponding object foreground pixels. The most important assumption of direct methods is photometric constancy, which assumes that the corresponding pixels between consecutive frames have the same pixel value. However, this assumption is often violated by illumination changes or surface reflectance properties (Zhong et al, 2017). Besides, although direct methods are able to track poor-textured objects, they still favor a certain amount of pixel value variations, because the pose optimization procedure relies on the image gradient. This is also quite different from region-based methods, which favor totally texture-less objects. Based on the above observations, we think using only photometric constraints is also not the optimal choice.

Table 1: Comparison of region-based methods and direct methods. The complementarity of these two kinds of methods makes it reasonable and beneficial to combine them together. In this paper, we take advantage of both methods and propose a robust hybrid 3D object tracker.

	Region-based Method	Direct Method
Constraints	Statistical constraints	Photometric constraints
Utilized information	Statistical information of image regions	Raw pixel values
Utilized pixels	Pixels around the contour	Pixels inside the contour
Considered region	Foreground and background	Foreground only
Good at tracking	Homogeneous / texture-less objects	Heterogeneous / well-textured objects
Failure cases	Heterogeneous objects in complex background, Symmetrical objects	Illumination changes, purely texture-less objects
Recover from pose drifts	Yes	No

To resolve the above problems when using only region-based methods or direct methods alone, we propose a hybrid 3D object tracker which takes advantage of both two kinds of methods by applying statistical constraints and photometric constraints to different image regions, as shown in Fig. 1(d). We partition the image into several foreground-background sub-region pairs and a single interior region. For each foreground-background sub-region pair, we apply statistical constraints based on the partitioned local statistical models. This would make the object contour evolve in the right direction to best fit the local foreground-background statistics. For the object region (including the interior region and the local foreground regions), we apply photometric constraints by minimizing the pixel value difference between consecutive frames. In this way, we properly utilize the information of the interior region, which region-based methods completely ignore during optimization (due to the smoothed Dirac delta function in gradient calculation (Prisacariu and Reid, 2012)).

Very few previous works have mentioned the integration of region-based methods and direct methods. We compare the different characteristics of these two kinds of methods

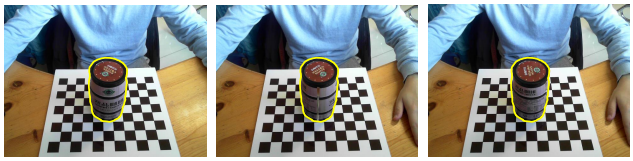


Fig. 2: An example of failure cases for region-based methods. When a symmetrical coffee can rotates around its vertical axis, the projected contour (shown in yellow) in the image does not change, which leads to wrong pose estimations for region-based methods. In this case, the correct pose could be determined by adding in photometric constraints in the proposed method, since the appearance (pixel values) changes while rotating.

in Table 1 and detail the complementary properties of these two kinds of methods as follows:

Direct methods help region-based methods: Firstly, image statistics are not very reliable in complex scenes, where the statistical models of foreground region and background region tend to be very similar. In this situation, it would be beneficial to add in photometric constraints from direct methods because they are based on raw pixel values and could help to alleviate the statistical confusion. Secondly, region-based methods are inherently not suitable for tracking the pose of symmetrical objects because different 3D poses would result in the same 2D contour (Prisacariu and Reid, 2012). As shown in Fig. 2, when a cylindrical object rotates around its vertical axis, the projected contour does not change at all, which means region-based methods are very likely to fail in tracking its pose. However, the pixel values (or texture) of the object change continuously while rotating, so the correct object pose could be determined from photometric consistency. Thirdly, as mentioned above, region-based methods ignore the interior region during pose optimization. As shown in Fig. 3, the smoothed Dirac delta function indicates the influence (or weights) of pixels in the gradient calculation. For pixels outside a narrow evaluation band (± 8 pixels in most previous works), the influence drops to below 5% of the center pixel. Moreover, the foreground and background regions for statistical model calculation and updating are often chosen as local regions around the object contour (e.g. ± 32 pixels) for distinctiveness, as shown in Fig. 1(b). This means a large number of pixels inside the object contour (i.e. beyond +32 pixels) are completely unused in region-based methods, which is a waste of information. By introducing photometric constraints, these pixels are properly utilized and would contribute to a more robust tracking algorithm. Based on the above observations, the proposed hybrid method would be more stable compared to region-based methods when encountering silhouette pose ambiguities or similar foreground-background statistics. For example, our method is shown to be more stable compared to the recent state-of-art region-based method (Tjaden et al, 2017) with a slightly lower tracking accuracy. More details will be discussed in the experiment part.

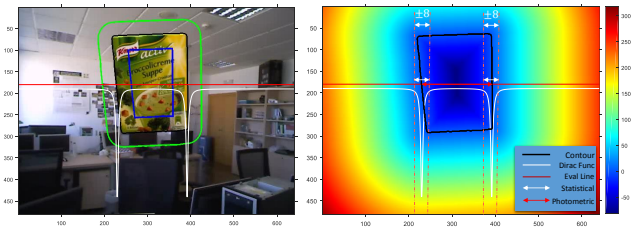


Fig. 3: Analysis of a single horizontal line in the image. Left: We analyze the pixels in a single line (red) of the image frame. Right: Distance transform ($\Phi(\mathbf{x})$) and other annotations: (1) Contour: The image contour, which is the zero level-set ($\Phi(\mathbf{x})=0$). (2) Dirac Func: Plot of the smoothed Dirac delta function ($\delta_\epsilon(\Phi)$, detailed in Sect. 3.4), which determines the weights of the pixels in the gradient calculation of region-based methods. (3) Eval Line: We take this red line as an example of the energy evaluation process. (4) Statistical: Pixels used for statistical energy evaluation and minimization in our implementation, which locate in a ± 8 pixels evaluation band around the contour. (5) Photometric: Pixels used for photometric constraints in our method, which are all the pixels inside the object contour.

Region-based methods help direct methods: On the other hand, region-based methods also help direct methods in the following aspects. Firstly, raw pixel values are not reliable when the photometric constancy assumption is violated. For example, when the illumination changes, corresponding pixels in neighbouring frames may not share the same values, which would disrupt the direct tracking method (Seo and Wuest, 2016). This is partly because direct methods only use raw pixel values of the foreground region. They ignore the background region, and also the image statistics. This issue could be alleviated by introducing statistical constraints from region-based methods, because image statistics are relatively more stable than raw pixel values in the case of illumination changes. Secondly, direct methods rely on image gradients during pose optimization, and are not very suitable for tracking purely homogeneous objects. Just on the contrary, region-based methods are especially good at tracking purely homogeneous objects, which would be of help to direct methods when there are hardly any image gradients in the object region. Thirdly, direct methods are inherently not able to recover from pose drifts. When there is a small pose drift caused by a slightly inaccurate tracking result, direct methods would mistake the involved background pixels as new foreground pixels. As a result, the pose drifts would accumulate with time and finally lead to tracking fail-

ure (Zhong et al, 2017). In this case, the wrongly involved background pixels could be excluded by region-based methods according to the different statistical distributions of foreground and background regions. So region-based methods could help direct methods in recovering from small pose drifts.

To conclude, region-based methods rely on image statistics and ignore raw pixel values, especially for the interior region. Direct methods rely on raw pixel values and ignore the foreground-background statistical information. The complementarity of these two kinds of methods makes it reasonable and beneficial to combine them together.

The main contributions of this paper are:

1. We propose a robust hybrid 3D object pose tracking method by combining statistical constraints from region-based methods and photometric constraints from direct methods. Our new probabilistic formulation leverages both statistical information and photometric information of the image in a complementary way, resulting in a robust hybrid 3D object pose tracker. To the best of our knowledge, we are the first to combine region-based methods and direct methods in a unified 3D object tracking framework.

2. We improve the performance of our method in complex scenes by partitioning the global foreground and background into local sub-region pairs. The partitioned local foreground and background statistical models are more distinguishable for heterogeneous objects and cluttered background, which further increases the robustness of the proposed hybrid tracker.

2 Related Work

A large number of 3D object tracking methods have been proposed in the past several decades (Lepetit and Fua, 2005). These methods can be classified into four major categories: feature-based methods, edge-based methods, region-based methods and direct methods. Feature-based methods require sufficient texture on the object surfaces in order to extract enough number of keypoints to establish feature correspondences, thus they are not able to deal with poor-textured objects. Edge-based methods rely on strong edges, so that they often struggle with the numerous local minima, especially when encountering cluttered background and sensor noise (Hexner and Hagege, 2016). In recent years, region-based methods and direct methods have become popular because of their capability to deal with these drawbacks. In the following, we mainly focus on the recent advances in region-based methods and direct methods, which are closely related to our work.

2.1 Region-based methods

Most of the recent region-based 3D object tracking methods are based on PWP3D (Prisacariu and Reid, 2012), which is the first region-based 3D object tracker that achieves real-time performance and is still among the state-of-art. This famous work utilizes some basic concepts from Bibby and Reid (2008) and Dambreville et al (2008), but the authors use level-set functions as shape embedding function, and define an energy function based on pixel-wise posterior probabilities, which proves to perform better than pixel-wise likelihoods (Prisacariu and Reid, 2012). After that, the authors have extended their system to work on mobile phones (Prisacariu et al, 2013) and RGB-D sensors (Ren et al, 2014).

Several region-based methods which focus on improving the performance of PWP3D have been proposed (Tjaden et al, 2016; Hexner and Hagege, 2016; Zhao et al, 2014). Tjaden et al (2016) address the problem of optimization strategies. The original implementation of PWP3D uses a simple gradient descent method in the pose optimization process, which is not very efficient since the optimal step size depends on model complexity and the size of the projected silhouette, thus is very difficult to decide. So Tjaden et al (2016) present a novel pixel-wise optimization strategy based on a Gauss-Newton-like algorithm. They approximate the Hessian matrix using only first order derivatives, which proves to be efficient in determining the step size automatically. They also introduce a real-time implementation based on CPU parallelization and OpenGL rendering. In this paper, we formulate a novel hybrid energy function based on both statistical and photometric constraints, which also needs efficient optimization strategy. So we adapt the Gauss-Newton-like optimization strategy suggested by Tjaden et al (2016) to our method, which improves the convergence property and tracking robustness.

Hexner and Hagege (2016) address the performance degradation of PWP3D in complex scenes, where the statistical properties of foreground and background vary, such as tracking heterogeneous objects in cluttered background. In the original PWP3D method, the foreground and background regions are described by global statistical models, which leads to tracking failure in complex scenes because a single global appearance model does not sufficiently capture the spatial variation. So Hexner and Hagege (2016) propose a new framework based on multiple local appearance models. The multiple local regions are centered around the 2D contour points. For each local region, they maintain a unique local foreground statistical model and local background statistical model respectively, which capture the local statistical properties better than a single global model. For each point around the contour, a local energy function is evaluated and all the local energies are fused together for final pose estimation. This local region-based method is shown to have

wider basin of attraction and higher probability of convergence to the correct pose when evaluating on a 3D object detection dataset, especially for heterogeneous objects. But the performance of Hexner and Hagege (2016) for 3D object tracking in video sequences is not addressed. This is probably due to the large number of local regions which add to the complexity in computation. Inspired by them, we also make use of local statistics by partitioning the global region into a small number (e.g. 4) of foreground-background sub-region pairs, as shown in Fig. 1(c,d). In our method, the local regions are defined in a more concise manner, which makes it more feasible for practical applications. Moreover, the utilizing of photometric constraints in our method also helps to deal with scene complexity.

Zhao et al (2014) improve the performance of PWP3D by adding a boundary term in the energy function. They also incorporate a particle-filtering module into the pipeline. The performance of their method is evaluated on the Rigid Pose Dataset (Pauwels et al, 2013), which we also adopt for quantitative evaluation and comparison. Some more recent works also achieve real-time performance with both RGB (Tjaden et al, 2017) and RGB-D (Kehl et al, 2017b) settings, but they have not explored the possibility of incorporating photometric constraints in the region-based 3D object tracking pipeline.

2.2 Direct methods

In recent years, direct methods have also been widely used in planar tracking (Alismail et al, 2016; Chen et al, 2017), 3D object tracking (Crivellaro and Lepetit, 2014; Seo and Wuest, 2016) and visual odometry (Engel et al, 2014, 2017). Most of them are built on the famous direct image alignment framework: Lucas-Kanade algorithm (Lucas et al, 1981; Baker and Matthews, 2004). The L-K framework is originally used to estimate 2D image transformations, but is later extended to 3D object tracking by employing a 3D warping function with a known object model (Crivellaro and Lepetit, 2014). Direct methods exploit rich per-pixel information in the image instead of local features, thus are generally more robust than feature-based methods. The most important drawback of direct method is its underlying assumption of photometric constancy, which is often violated by lighting variations (Zhong et al, 2017). Many different approaches have been proposed to tackle this issue and to improve the robustness of direct methods. Crivellaro and Lepetit (2014) propose to use gradient-based Descriptor Fields instead of raw pixel values to improve the tracking performance with poor-textured objects, specular objects and lighting variations. Similarly, Chen et al (2017) use gradient orientation (GO) as the dense image descriptor. Alismail et al (2016) introduce Bit-Planes, an illumination-invariant binary descriptor, which obtains good results in low light and sudden illumination changes. Except for designing robust dense image descriptors, some works propose to use robust similarity measures such as Mutual Information (MI) (Caron et al, 2014), or Normalized Cross Correlation (NCC) (Scandaroli et al, 2012). Other works try to model the illumination changes under Lambertian assumption (Seo and Wuest, 2016).

In this paper, we adopt the idea of using robust local descriptors instead of intensity or color to improve the robustness of photometric constraints when encountering illumination changes and specularity. Specifically, we use the gradient-based Descriptor Fields proposed by Crivellaro and Lepetit (2014) to formulate the photometric constraints, which will be detailed in Sect. 3.

2.3 Other State-of-art Methods

Apart from region-based methods and direct methods, some other recent 3D object tracking approaches also achieve state-of-art results.

Firstly, edge-based methods have been improved by incorporating color information or color statistics to search for the correct edge correspondence. Panin et al (2008) propose an edge-based method integrating color and edge likelihoods. They consider intensity gradients and local color statistics as two complimentary visual modalities for efficient data fusion. Petit et al (2013) develop a robust edge-based 3D object tracker combing geometrical and color edge information. They integrate geometrical and color features along edges by combining the geometrical information provided by the distance between model and image edges with a denser color information through foreground/background color statistics. Seo et al (2014) propose a novel 3D object tracking method based on optimal local searching of 3D-2D correspondences between object model and image edges. In their searching scheme, the region appearance is modeled by HSV histograms on a newly defined local space to ensure the confident searching direction.

Secondly, a new kind of methods which combine model-based approach and SLAM approach also show good results. Loesch et al (2015) propose to estimate the camera pose relative to an object by combing dynamically extracted 3D contour points with a key-frame-based SLAM algorithm. The absolute information of the CAD model is directly integrated in the bundle adjustment (BA) process of the SLAM algorithm. Singhal et al (2016) propose an object-based SLAM algorithm which combines 3D object tracking and visual odometry together. In their pipeline, objects are included in the map as semantic entities to enhance the SLAM algorithm.

The success of these approaches have demonstrated the effectiveness of properly combining information of different

modalities. In this paper, we also propose a hybrid method combining statistical information and photometric information of the image.

2.4 Assumptions of the Proposed Method

In this paper, we find that direct method is ideal for covering the shortage of region-based method, as discussed in Sect. 1. So we apply photometric constraints to the object region, trying to utilize the unused information of region-based method. Experimental results demonstrate that the proposed hybrid tracker obtains better results than using region-based method or direct method alone. In the following discussions, we assume an initial object pose is available for the first video frame, and focus on the frame-to-frame pose tracking problem. 3D object detection methods, such as (Hinterstoisser et al, 2011; Kehl et al, 2017a), can be combined with our method for initialization or reset.

As will be shown in the experiment part, the proposed method is able to handle illumination changes and specularity, as well as partial occlusions. In order to clarify the proper application scenarios of the proposed method, some assumptions need to be made about the camera motion and the viewpoint. Firstly, a relatively small motion between consecutive frames is assumed. Secondly, we assume the object is within a moderate distance to the camera in most of the video frames, so that the object region in the image is neither too small (e.g. less than 30 pixels) nor too big (e.g. bigger than the field of view).

3 Proposed Method

In this section, we present the proposed hybrid 3D object tracking algorithm. In Sect. 3.1, we introduce the basic notations and formulate our new probabilistic model. In Sect. 3.2, we formulate the partitioned local statistical constraints. In Sect. 3.3, we formulate the photometric constraints. In the end, we present the final energy function combining statistical and photometric constraints in Sect. 3.4, together with the pose optimization strategy.

3.1 The Proposed Hybrid 3D Object Tracking Model

We begin by introducing the proposed hybrid 3D object tracking model. Fig. 4 illustrates the generative model of our method. We define the notations as follows:

The RGB image is denoted by $\mathbf{I}$. Every pixel $\mathbf{x} = (x,y)^T$ has a corresponding color vector $\mathbf{y} = \mathbf{I}(\mathbf{x})$. Every pixel belonging to the object region $\mathbf{x} \in \Omega_{obj}$ has a corresponding 3D point $\mathbf{X} = (X,Y,Z)^T$ in the camera coordinate frame and $\mathbf{X}_0 = (X_0,Y_0,Z_0)^T$ in the object coordinate frame. The 3D rigid transformation between this two coordinate frames is defined by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{T}$, which can be encoded in a 6-dof pose vector $\mathbf{p} = (t_x,t_y,t_z, \omega_x,\omega_y,\omega_z)^T$ using Lie algebra representation. We have $\mathbf{X} = \mathbf{R}(\mathbf{p})\mathbf{X_0} + \mathbf{T}(\mathbf{p})$ and $\mathbf{x} = \pi(K\mathbf{X})$, where K is intrinsic matrix of the pre-calibrated camera, and $\pi(\mathbf{X}) = (X/Z, Y/Z)^T$.

Similar to most of region-based methods, we represent the object in the image by its contour C, which is the zero level-set $C = \{\mathbf{x}|\Phi(\mathbf{x}) = 0\}$ of the signed distance function $\Phi(\mathbf{x})$ (Prisacariu and Reid, 2012). In previous works, the foreground and background statistics are usually represented by a global foreground appearance model $P(\mathbf{y}|M_f)$ and a global background appearance model $P(\mathbf{y}|M_b)$, where $\mathbf{y}$ is the RGB values of a certain foreground or background pixel. The conditional distributions $P(\mathbf{y}|M_f)$ and $P(\mathbf{y}|M_b)$ are the likelihoods of a pixel with color $\mathbf{y}$ belonging to foreground or background regions, and are commonly represented with RGB color histograms.

Before introducing the proposed hybrid 3D object tracking model, we first declare that there are two different 'bands' used in this paper. First is the *histogram band*, which is the region around the contour used for color histogram calculation (e.g. ± 32 pixels as shown in Fig. 1 (b)). Second is the *evaluation band*, which is a narrower region around the contour used for statistical energy evaluation and minimization (e.g. ± 8 pixels as shown in Fig. 3). We denote the widths of these two bands by BW_{hist} and BW_{eval} respectively. Theoretically, statistical constraints should be applied to all the pixels in the histogram band. But as with most previous works, we choose to speed up the optimization in real implementation by evaluating the statistical constraints only in a narrower evaluation band. This is a reasonable approximation because for pixels outside the evaluation band, the influence of their gradients drops to below 5% of the center pixel due to the smoothed Dirac delta function (as discussed in Sect. 1 and demonstrated in Fig. 3). Nevertheless, we still assume all the pixels in the histogram band (Ω_{fb}) are used for statistical constraints in the following theoretical derivations.

Now we introduce our hybrid 3D object tracking model. As shown in Fig. 4, the whole foreground and background regions are divided into n local foreground and background sub-region pairs $\{\Omega_{f_i}, \Omega_{b_i}\}_{i=1:n}$ and a single interior region Ω_{in}. The partitioning of the sub-region pairs are defined as follows:

Firstly, the global foreground region Ω_f and background region Ω_b, together with the interior region Ω_{in} are defined according to the distance transform function $\Phi(\mathbf{x})$:

$\Omega_f = \{\mathbf{x}| - BW_{hist} \leq \Phi(\mathbf{x}) < 0\}$,
$\Omega_b = \{\mathbf{x}|0 < \Phi(\mathbf{x}) \leq BW_{hist}\}$,
$\Omega_{in} = \{\mathbf{x}|\Phi(\mathbf{x}) < -BW_{hist}\}$.

Here BW_{hist} is the bandwidth for color histogram calculation, which we set to 32 pixels as shown in Fig. 1 (b).

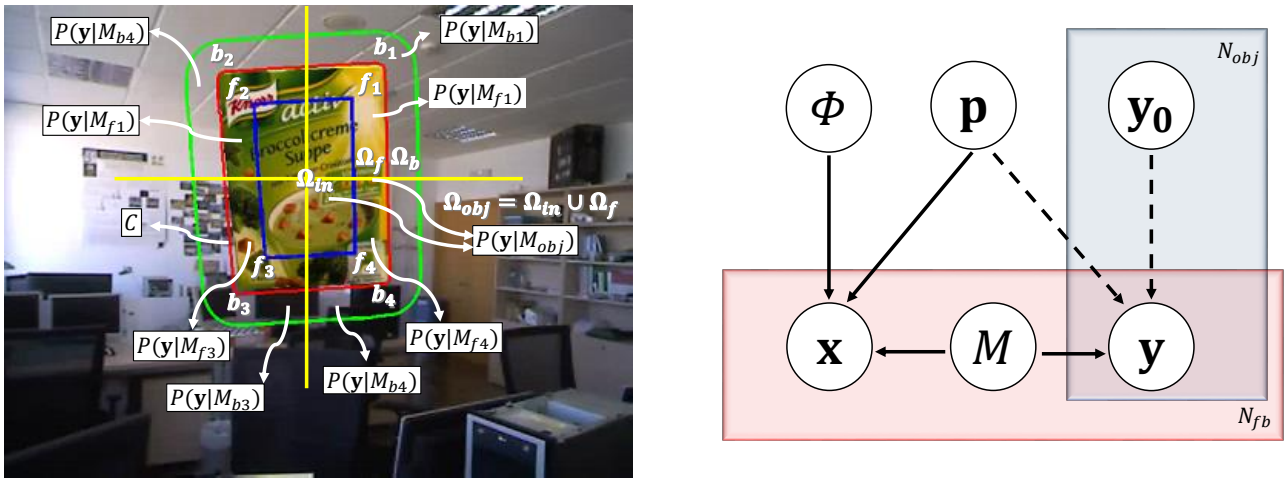


Fig. 4: The notations and graphical model of the proposed method. In the graphical model, the solid edges denote terms that are active for pixels in $\Omega_{fb} = \Omega_f \bigcup \Omega_b = \bigcup_i \{\Omega_{fi}, \Omega_{bi}\}$, and the dotted edges denote terms that are active for pixels in $\Omega_{obj} = \Omega_{in} \bigcup \Omega_f$.

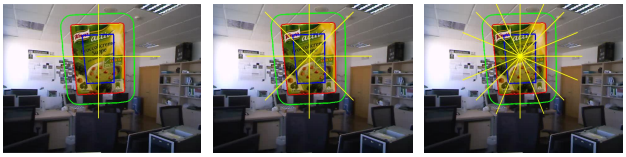


Fig. 5: Some examples of the partitioned sub-regions.

Secondly, the centroid of the object region $\mathbf{x}_c = (x_c, y_c)$ is calculated and is used as the axis center for sub-region division. Then we evenly partition the image region around the axis center $\mathbf{x}_c$ into n parts using polar coordinates. The i-th image region is defined as:

$\Omega_{IMGi} = \{\mathbf{x} \,|\, -\pi + (i-1) \times \frac{2\pi}{n} \leq \theta(\mathbf{x}, \mathbf{x}_c) < -\pi + i \times \frac{2\pi}{n}\}$;

$\theta(\mathbf{x}, \mathbf{x}_c) = \text{atan2}(x - x_c, y - y_c)$; $i = 1, 2, ..., n$. (atan2 is the four-quadrant inverse tangent function.)

Finally, the i-th sub-region pairs are defined as:

$\Omega_{fi} = \Omega_f \cap \Omega_{IMGi}$, $\Omega_{bi} = \Omega_b \cap \Omega_{IMGi}$, $i = 1, 2, ..., n$.

Some examples of the partitioned regions ($n = 4, 8, 16$) are shown in Fig. 5. Compared with (Hexner and Hagege, 2016), in which the local regions are defined as small circles centered around the contour points, here we have offered an alternative way to select the local regions.

The number of local foreground-background sub-region pairs can be chosen according to the spatial variation of the statistical properties. The optimal number of regions will be discussed in the experiment part. For example, in Fig. 4, the global foreground and background regions are partitioned into 4 sub-region pairs $\{\Omega_{f_1}, \Omega_{b_1}\}$, $\{\Omega_{f_2}, \Omega_{b_2}\}$, $\{\Omega_{f_3}, \Omega_{b_3}\}$, $\{\Omega_{f_4}, \Omega_{b_4}\}$ and a single interior region Ω_{in}. For each sub-region pair $\Omega_i = \{\Omega_{f_i}, \Omega_{b_i}\}$, we have a specific local foreground statistical model $P(\mathbf{y}|M_{f_i})$ and a local background statistical model $P(\mathbf{y}|M_{b_i})$. For the object region $\Omega_{obj} = \Omega_{in} \bigcup \Omega_f$, we impose a photometric constraint to each pixel according to the photometric constancy assumption of direct methods, and the likelihood is denoted by $P(\mathbf{y}|M_{obj})$. So in our generative model, the region indicator variable $M \in \{M_{f_i}, M_{b_i}\}_{i=1:n} \cup \{M_{obj}\}$.

As with most region-based methods, we also assume pixel-wise independence and treat the whole image region as a bag-of-pixels $\{\mathbf{x_j}, \mathbf{y_j}\}_{j=1:N_\Omega}$ (Bibby and Reid, 2008). N_Ω is the total number of pixels used in our probabilistic model:

$$N_\Omega = N_{fb} + N_{obj} = \sum_{i=1}^{n} (N_{f_i} + N_{b_i}) + N_{obj}$$

where N_{f_i} and N_{b_i} are the number of pixels belonging to the i-th foreground-background sub-region pair and N_{obj} is the number of pixels belonging to the object region. Note that the pixels in the foreground region Ω_f are used twice for statistical constraints and photometric constraints respectively.

According to the graphical model in Fig. 4, for a single pixel which belongs to one of the foreground-background sub-region pairs, i.e., $\mathbf{x} \in \Omega_{fb} = \bigcup_i \Omega_i$, the joint distribution can be written as:

$$\begin{aligned} &P(\mathbf{x}, \mathbf{y}, \mathbf{p}, \Phi, M) \\ &= P(\mathbf{x}|\mathbf{p}, \Phi, M) P(\mathbf{y}|M) P(M) P(\Phi) P(\mathbf{p}) \end{aligned} \tag{1}$$

For a single pixel which belongs to the object region, i.e., $\mathbf{x} \in \Omega_{obj}$, the joint distribution can be written as:

$$P(\mathbf{y}, \mathbf{y}_0, \mathbf{p}) = P(\mathbf{y}|\mathbf{y}_0, \mathbf{p}) P(\mathbf{y}_0) P(\mathbf{p}) \tag{2}$$

where $\mathbf{y}_0$ is the color of the corresponding pixel $\mathbf{x}_0$ in the previous frame, which will be detailed in Sect. 3.3.

The pose estimation problem can be solved by maximizing the posterior probability over all of the pixels:

$$\mathbf{p}_{MAP} = \underset{\mathbf{p}}{\operatorname{argmax}} \prod_{\mathbf{x}\in\Omega_{fb}} P(\mathbf{p},\Phi|\mathbf{x},\mathbf{y},M) \prod_{\mathbf{x}\in\Omega_{obj}} P(\mathbf{p}|\mathbf{y},\mathbf{y}_0) \tag{3}$$

In the following subsections, we derive the statistical constraints applied to pixels $\mathbf{x}\in\Omega_{fb}$ and the photometric constraints applied to pixels $\mathbf{x}\in\Omega_{obj}$. Then, we combine these two kinds of constraints together and obtain the final energy function.

3.2 Statistical Constraints

We apply a statistical constraint for every pixel $\mathbf{x}\in\Omega_{fb}$. Since we have partitioned the foreground and background regions into several sub-region pairs, we calculate the statistical constraints for each sub-region pair respectively. For a pixel in the i-th sub-region pair $\mathbf{x}\in\Omega_i=\Omega_{f_i}\cup\Omega_{b_i}$, the joint distribution has the same form as Eq. (1), but here $M\in\{M_{f_i},M_{b_i}\}$, meaning only the pixels in the i-th sub-region pair are being considered. We omit $P(\Phi)$ and $P(\mathbf{p})$ because we do not have prior knowledge of the level-set embedding function and the pose parameters. We replace $P(\mathbf{y}|M)P(M)$ with $P(M|\mathbf{y})P(\mathbf{y})$ according to Bayes rule and then marginalize over M (Bibby and Reid, 2008; Prisacariu and Reid, 2012):

$$\begin{aligned}&P(\mathbf{p},\Phi|\mathbf{x},\mathbf{y})\\&=P\left(\mathbf{x}|\Phi,\mathbf{p},M_{f_i}\right)P\left(M_{f_i}|\mathbf{y}\right)+P\left(\mathbf{x}|\Phi,\mathbf{p},M_{b_i}\right)P\left(M_{b_i}|\mathbf{y}\right)\end{aligned} \tag{4}$$

Here, $1/P(\mathbf{x})$ is dropped since it is considered as constant for all pixel locations. $P\left(\mathbf{x}|\Phi,\mathbf{p},M_{f_i}\right)$ and $P\left(\mathbf{x}|\Phi,\mathbf{p},M_{b_i}\right)$ are the spatial priors for a pixel location $\mathbf{x}$:

$$P\left(\mathbf{x}|\Phi,\mathbf{p},M_{f_i}\right)=\frac{H_e\left(\Phi\left(\mathbf{x}\right)\right)}{\eta_{f_i}} \tag{5}$$

$$P\left(\mathbf{x}|\Phi,\mathbf{p},M_{b_i}\right)=\frac{1-H_e\left(\Phi\left(\mathbf{x}\right)\right)}{\eta_{b_i}} \tag{6}$$

where H_e is the smoothed Heaviside step function identical to (Tjaden et al, 2016). $\eta_{f_i}=\sum_{\mathbf{x}\in\Omega_i}H_e\left(\Phi\left(\mathbf{x}\right)\right)$ and $\eta_{b_i}=\sum_{\mathbf{x}\in\Omega_i}\left(1-H_e\left(\Phi\left(\mathbf{x}\right)\right)\right)$ are the number of foreground and background pixels in the i-th sub-region pair (Prisacariu and Reid, 2012).

$P\left(M_{f_i}|\mathbf{y}\right)$ and $P\left(M_{b_i}|\mathbf{y}\right)$ are the color posteriors:

$$P\left(M_{f_i}|\mathbf{y}\right)=\frac{P\left(\mathbf{y}|M_{f_i}\right)P\left(M_{f_i}\right)}{P\left(\mathbf{y}|M_{f_i}\right)P\left(M_{f_i}\right)+P\left(\mathbf{y}|M_{b_i}\right)P\left(M_{b_i}\right)} \tag{7}$$

$$P\left(M_{b_i}|\mathbf{y}\right)=\frac{P\left(\mathbf{y}|M_{b_i}\right)P\left(M_{b_i}\right)}{P\left(\mathbf{y}|M_{f_i}\right)P\left(M_{f_i}\right)+P\left(\mathbf{y}|M_{b_i}\right)P\left(M_{b_i}\right)} \tag{8}$$

where $P\left(\mathbf{y}|M_{f_i}\right)$ and $P\left(\mathbf{y}|M_{b_i}\right)$ are the likelihood functions calculated from the local color histograms of the i-th sub-region pair. The region prior $P\left(M_{f_i}\right)=\eta_{f_i}/\left(\eta_{f_i}+\eta_{b_i}\right)$, $P\left(M_{b_i}\right)=\eta_{b_i}/\left(\eta_{f_i}+\eta_{b_i}\right)$.

Now we have the posterior distribution of a single pixel:

$$P(\mathbf{p},\Phi|\mathbf{x},\mathbf{y})=H_e\left(\Phi\left(\mathbf{x}\right)\right)P_{f_i}+\left(1-H_e\left(\Phi\left(\mathbf{x}\right)\right)\right)P_{b_i} \tag{9}$$

where $P_{f_i}=P\left(M_{f_i}|\mathbf{y}\right)/\eta_{f_i}$, $P_{b_i}=P\left(M_{b_i}|\mathbf{y}\right)/\eta_{b_i}$. We take the negative log posterior probability and sum over the pixels and the sub-region pairs to obtain the energy function for statistical constraints:

$$\begin{aligned}E_{fb}&=-\sum_{i=1:n}\sum_{\mathbf{x}\in\Omega_i}\log P(\mathbf{p},\Phi|\mathbf{x},\mathbf{y})\\&=-\sum_{i=1:n}\sum_{\mathbf{x}\in\Omega_i}\log\left(H_e\left(\Phi\left(\mathbf{x}\right)\right)P_{f_i}+\left(1-H_e\left(\Phi\left(\mathbf{x}\right)\right)\right)P_{b_i}\right)\end{aligned} \tag{10}$$

3.3 Photometric Constraints

We apply a photometric constraint for every pixel $\mathbf{x}\in\Omega_{obj}$ to ensure the photometric constancy between the current frame and the previous frame. Pixel-wise photometric constancy assumes that the corresponding pixels from the previous frame (which serves as the template image in direct method) and the current frame have the same intensity or color. As shown in Fig. 6, we assume $\mathbf{I}\left(\mathbf{x}\right)=\mathbf{I}_0\left(\mathbf{x}_0\right)$ for each pixel in the object region. For a pixel $\mathbf{x}_0$ in the previous frame, we first back-project $\mathbf{x}_0$ to the 3D model in pose $\mathbf{p}_0$, and then re-project the 3D point $\mathbf{X}$ to the current frame in pose $\mathbf{p}$. In this way, we find a corresponding pixel $\mathbf{x}$ for every $\mathbf{x}_0$ through the 3D warping function $\mathbf{x}=W\left(\mathbf{x}_0,\mathbf{p}\right)=Proj\left(Proj^{-1}\left(\mathbf{x}_0,\mathbf{p}_0\right),\mathbf{p}\right)$.

For a given pixel $\mathbf{x}_0$ in the previous frame, its color $\mathbf{y}_0=\mathbf{I}_0\left(\mathbf{x}_0\right)$. Following the joint distribution for a single pixel in the object region (Eq. (2)), we have:

$$P(\mathbf{p}|\mathbf{y},\mathbf{y}_0)=\frac{P(\mathbf{y}|\mathbf{p},\mathbf{y}_0)P(\mathbf{p})}{P(\mathbf{y})} \tag{11}$$

We drop the term $P\left(\mathbf{y}\right)$ because it does not depend on the pose parameter $\mathbf{p}$. We assume the prior distribution of $\mathbf{p}$ is

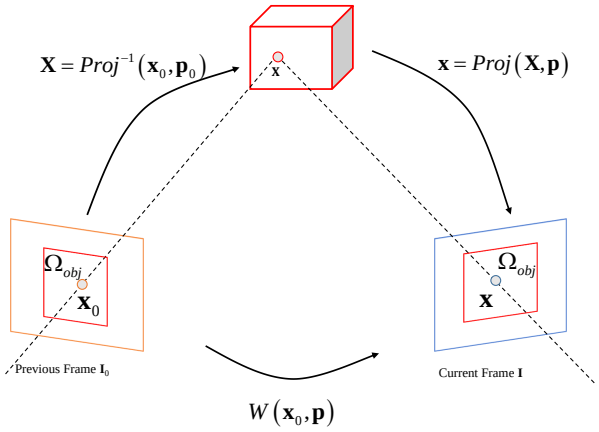


Fig. 6: The 3D warping function. According to the photometric constancy assumption, the corresponding pixels in the object regions of consecutive frames have the same color, i.e., $\mathbf{I}(\mathbf{x}) = \mathbf{I}_0(\mathbf{x}_0)$. The correspondence is found using a 3D warping function $W(\mathbf{x}, \mathbf{p})$.

uniform, i.e., $P(\mathbf{p}) = const$. So that $P(\mathbf{p}|\mathbf{y}, \mathbf{y}_0) \propto P(\mathbf{y}|\mathbf{p}, \mathbf{y}_0)$. Here a common assumption is that $P(\mathbf{y}|\mathbf{p}, \mathbf{y}_0) \sim N(y_0, \sigma)$ (Kerl et al, 2013):

$$P(\mathbf{y}|\mathbf{p}, \mathbf{y}_0) \propto \exp\left\{\frac{-\|\mathbf{y} - \mathbf{y}_0\|^2}{\sigma^2}\right\} \tag{12}$$

Then we obtain the energy function for photometric constraints:

$$\begin{aligned} E_{obj} &= -\sum_{\mathbf{x}\in\Omega_{obj}} \log P(\mathbf{y}|\mathbf{p}, \mathbf{y}_0) \\ &= \lambda \sum_{\mathbf{x}\in\Omega_{obj}} \|\mathbf{y} - \mathbf{y}_0\|^2 \\ &= \lambda \sum_{\mathbf{x}\in\Omega_{obj}} \|\mathbf{I}(\mathbf{x}) - \mathbf{I}(\mathbf{x}_0)\|^2 \end{aligned} \tag{13}$$

where $\mathbf{x} = W(\mathbf{x}_0, \mathbf{p})$, $\lambda = \frac{1}{\sigma^2}$.

As introduced in Sect. 1, the original assumption of photometric constancy is often violated by illumination changes and specularity. Although this could be alleviated by the statistical term in our method, it is still a negative factor when dealing with illumination changes or specular objects. Therefore, we propose to use robust dense local descriptors instead of color to formulate the photometric constraints. Specifically, Crivellaro and Lepetit (2014) have demonstrated that gradient-based Descriptor Fields could significantly improve the robustness of direct alignment in presence of non-Lambertian illumination effects, such as specularities. So we choose to replace the color $\mathbf{I}(\mathbf{x})$ with the gradient-based Descriptor Fields $\mathbf{F}(\mathbf{x})$ and get a slightly altered energy for photometric constraints:

$$E_{obj} = \lambda \sum_{\mathbf{x}\in\Omega_{obj}} \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)\|^2 \tag{14}$$

Here $\mathbf{F}(\mathbf{x})$ denotes the gradient-based Descriptor Fields, which consists of four channels (positive and negative values of image gradients along x-direction and y-direction respectively, details could be found in (Crivellaro and Lepetit, 2014)). With this slightly altered energy function for photometric constraints, our method could better handle illumination changes and specularity.

Another issue is the covisibility of the 3D object points in consecutive frames. A 3D object point $\mathbf{X}$ (back-projected from image point $\mathbf{x}_0$) which is visible in frame $\mathbf{I}_0$ might be invisible in frame $\mathbf{I}$ due to self-occlusion of the object. In this case, $\mathbf{x}$ and $\mathbf{x}_0$ are not projections of the same 3D point, so that they should not be included in the photometric energy. We apply a simple strategy to detect these self-occluded pixels. For every back-projected point $\mathbf{X}$ from frame $\mathbf{I}_0$, we calculate the angle between the sight ray in frame $\mathbf{I}$ (from camera to point $\mathbf{X}$) and the surface normal at $\mathbf{X}$. If the angle is larger than 90°, we decide it is visible in frame $\mathbf{I}$. Otherwise, we treat it as invisible points and discard it from the photometric energy. According to our experiments, the number of invisible points is usually very small, because significant change of viewpoint between consecutive frames does not frequently happen. But this strategy would be necessary when encountering fast rotation or translation of the object.

3.4 Joint Statistical and Photometric Energy Minimization for Pose Optimization

In the previous subsections, we have derived the energy functions for statistical constraints and photometric constraints respectively. Now we combine them together to obtain the final energy function:

$$\begin{aligned} E =& E_{fb} + E_{obj} \\ =& -\sum_{i=1:n} \sum_{\mathbf{x}\in\Omega_i} \log\left(H_e(\Phi(\mathbf{x}))P_{f_i} + (1 - H_e(\Phi(\mathbf{x})))P_{b_i}\right) \\ &+ \lambda \sum_{\mathbf{x}\in\Omega_{obj}} \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)\|^2 \end{aligned} \tag{15}$$

The first term in the energy function is a general non-linear optimization problem w.r.t. pose parameter $\mathbf{p}$, while the second term leads to a standard non-linear least squares problem. λ acts as the weighting factor to balance the statistical term and the photometric term.

We found that using simple gradient descent approach yields bad results for this complex non-linear optimization problem. In order to efficiently optimize pose parameters on this complex energy function, we use an iterative Gauss-Newton-like optimization strategy and split the calculation of Jacobian matrix and Hessian matrix into two parts:

$$\mathbf{J} = \mathbf{J}_{fb} + \mathbf{J}_{obj} \qquad \mathbf{H} = \mathbf{H}_{fb} + \mathbf{H}_{obj} \tag{16}$$

For the first part:

$$\begin{aligned} \mathbf{J}_{fb} =& \frac{\partial E_{fb}}{\partial \mathbf{p}} \\ =& \sum_{i=1:n} \sum_{\mathbf{x} \in \Omega_i} \mathbf{J}_{fb}(\mathbf{x}) \\ =& -\sum_{i=1:n} \sum_{\mathbf{x} \in \Omega_i} \frac{P_{f_i} - P_{b_i}}{H_e(\Phi(\mathbf{x}))P_{f_i} + (1 - H_e(\Phi(\mathbf{x})))P_{b_i}} \\ & \times \frac{\partial H_e(\Phi(\mathbf{x}))}{\partial \mathbf{p}} \end{aligned} \tag{17}$$

and

$$\frac{\partial H_e(\Phi(\mathbf{x}))}{\partial \mathbf{p}} = \frac{\partial H_e}{\partial \Phi} \frac{\partial \Phi}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{p}} \tag{18}$$

Here, $\frac{\partial H_e}{\partial \Phi} = \delta_e(\Phi)$ is the smoothed Dirac delta function, $\frac{\partial \Phi}{\partial \mathbf{x}} = \left[\frac{\partial \Phi}{\partial x}, \frac{\partial \Phi}{\partial y}\right]$ is calculated using centered finite differences. $\frac{\partial \mathbf{x}}{\partial \mathbf{p}}$ can be derived from

$$\mathbf{x} = \pi(K\mathbf{X}) = \pi(K(\mathbf{R}(\mathbf{p})\mathbf{X_0} + \mathbf{T}(\mathbf{p})))$$

and the details can be found in Tjaden et al (2016); Hexner and Hagege (2016). Then, the Hessian is approximated as:

$$\mathbf{H}_{fb} = \sum_{i=1:n} \sum_{\mathbf{x} \in \Omega_i} \mathbf{J}_{fb}(\mathbf{x})^T \mathbf{J}_{fb}(\mathbf{x}) \tag{19}$$

Here the approximation of Hessian from Jacobians for this general non-linear energy term is developed from empirical and experimental studies of some previous works such as (Tjaden et al, 2016; Ren et al, 2017; Kehl et al, 2017b). We also give a mathematical derivation of this approximation in Appendix A.

The Jacobian and Hessian for the second part can be directly calculated from standard Gauss-Newton method:

$$\begin{aligned} \mathbf{J}_{obj} =& \frac{\partial E_{obj}}{\partial \mathbf{p}} \\ =& \lambda \sum_{\mathbf{x} \in \Omega_{obj}} (\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)) \frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{p}} \end{aligned} \tag{20}$$

and

$$\frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{p}} = \frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{p}} \tag{21}$$

where $\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)$ is the residual, $\frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}} = \left[\frac{\partial \mathbf{F}}{\partial x}, \frac{\partial \mathbf{F}}{\partial y}\right]$ is the gradient of each channel in the Descriptor Fields. $\frac{\partial \mathbf{x}}{\partial \mathbf{p}}$ can be calculated from $\mathbf{x} = W(\mathbf{x}_0, \mathbf{p}) = Proj\left(Proj^{-1}(\mathbf{x}_0, \mathbf{p}_0), \mathbf{p}\right)$.

The Hessian:

$$\mathbf{H}_{obj} = \lambda \sum_{\mathbf{x} \in \Omega_{obj}} \left(\frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{p}}\right)^T \left(\frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{p}}\right) \tag{22}$$

In the end, the pose parameter update $\Delta\mathbf{p}$ is calculated in each iteration as follows:

$$\Delta\mathbf{p} = -\mathbf{H}^{-1}\mathbf{J}^T = -\left(\mathbf{H}_{fb} + \mathbf{H}_{obj}\right)^{-1}\left(\mathbf{J}_{fb} + \mathbf{J}_{obj}\right)^T \tag{23}$$

4 Evaluation

4.1 Datasets and Implementation Details

We evaluate the performance of our method through four different set of experiments. For all of the experiments a commodity desktop computer with Intel i7 quad core CPU @4.0 GHz and NVIDIA GeForce GTX970 GPU is used. In our implementation, GPU is only used for rendering purposes, and all the other computations are performed on the CPU. After each frame, the local color histograms are updated using the same rule as in (Tjaden et al, 2017). Our method runs at 20-25 Hz using CPU parallelization.

4.1.1 Datasets

Firstly, we evaluate our method on a newly constructed dataset (HTS Dataset) with ground-truth annotations. The HTS Dataset consists of 6 real-world video sequences containing Heterogeneous objects, Texture-less objects and Symmetrical objects. In this experiment, we demonstrate the improvements of our method compared to the baseline for tracking different types of objects after applying partitioned local statistical model and adding in photometric constraints.

Secondly, we compare the performance of our method with some state-of-art methods on Rigid Pose Dataset (Pauwels et al, 2013). The proposed method achieves competitive or better performance compared to the state-of-art on this challenging dataset.

Thirdly, in order to demonstrate the robustness of our method with respect to illumination changes and specularity,

Table 2: Evaluation results on HTS Dataset. (Tracking Success Rate, in %)

Method	Constraints	hetero-1	hetero-2	t-less-1	t-less-2	sym-1	sym-2
Global Region	Global Statistical	84.8	77.0	98.3	97.6	86.8	77.3
Partitioned Local Region	Local Statistical	91.9	87.6	98.8	**97.9**	92.1	85.4
Hybrid Region+Direct (Proposed)	Local Statistical + Photometric	**97.9**	**99.0**	**98.9**	**97.9**	**98.1**	**94.3**

we evaluate our method on Dense Tracking Dataset (Crivellaro and Lepetit, 2014), which contains strong moving light sources and bright specularities.

Finally, we compare the performance of our method with the recent state-of-art local region-based method (Tjaden et al, 2017) on their semi-synthetic dataset.

4.1.2 Optimal Number of Sub-regions

Choosing the number of local sub-region pairs n is a trade-off problem between discriminative power and invariance. Using a larger n could make the partitioned local statistical models more discriminative, but it would also make them less stable because of the decreasing number of pixels.

After trying different number of sub-regions on a subset of Rigid Pose Dataset (Pauwels et al, 2013), we find that using $n = 4$ is enough for most of the tracking scenarios, while increasing it to $n = 8$ yields virtually the same results. Using $n = 16$ or larger would degrade the tracking performance. So we use $n = 4$ in all of the following experiments. The details about selecting the optimal number of sub-regions will be discussed in Sect. 4.3.3.

4.1.3 Balancing the Tracking Energy with λ

It is very important to choose the proper weighting factor λ in Eq. (15). The weighting factor λ should compensate for the numerical scale and the different number of pixels used in the statistical term and the photometric term (Kehl et al, 2017b). According to our experiment, the numerical scale for the statistical term is $\sim 10^4$, and the numerical scale for the photometric term is $\sim 10^2$. We also consider the different number of pixels used for statistical constraints and photometric constraints. The weighting factor is then set to $\lambda = 10^2 \times \frac{N_{fb}}{N_{obj}}$ in each iteration.

4.1.4 Scale-aware Bandwidth Selection

In the previous sections, the width of the histogram band BW_{hist} is fixed to ± 32 pixels, and the width of the evaluation band BW_{eval} is fixed to ± 8 pixels. Now we discuss the bandwidth selection problem.

As for the histogram band, we find the optimal BW_{hist} is insensitive to the size of the object silhouette. The ± 32 band provides enough number of pixels to construct discriminative color histograms. Using larger BW_{hist} could decrease the discrimination of the calculated color histograms and is not necessary since the energy evaluation is only conducted in a narrower evaluation band. If BW_{hist} is set too small, the small number of pixels will cause over-fitting of the calculated color histograms.

As for the evaluation band, most of the previous region-based methods choose to use a fixed BW_{eval} of ± 8 pixels, and we find it reasonable because it yields good results in most cases. But we also notice that it would be a little more accurate to use a smaller BW_{eval} when the object silhouette is very small in the image. Through our experiments, we develop an empirical rule for choosing a scale-aware BW_{eval} as follows:

$$BW_{eval} = \min\left\{8, \max\left\{2, 2^{\frac{\sqrt{ObjArea}}{50}}\right\}\right\}$$

where $ObjArea$ is the area of the object region.

4.2 Evaluation on HTS Dataset

To evaluate the performance of the proposed hybrid tracker for different types of real-world objects, we have constructed the HTS (Heterogeneous, Texture-less, Symmetrical) Dataset. This dataset consists of 6 video sequences capturing a heterogeneously colored cube, a texture-less bunny, and a symmetrical coffee can. The objects used in HTS Dataset are shown in Fig. 8. For all the videos, we obtain the ground-truth object poses using a 2D marker board provided by the ArUco library (Garrido-Jurado et al, 2014).

In order to validate the effectiveness of the proposed two main contributions, we have tested 3 different versions of our method on HTS Dataset. Firstly, we test a baseline method: Global Region-based method, which applies only global statistical constraints. Secondly, we test a Partitioned Local Region-based method, in which we partition the global region into a small number of local sub-regions and apply local statistical constraints. Finally, we test the proposed hybrid method which combines local statistical constraints and photometric constraints together. The evaluation results are shown in Table 2. We measure the tracking success rate (SR) throughout each sequence. A frame is successfully tracked if the rotation error and translation error are both smaller than the thresholds (e_rot=0.1 radians, e_transl=30 mm). We analyze the tracking performance for different types of objects in detail as follows:

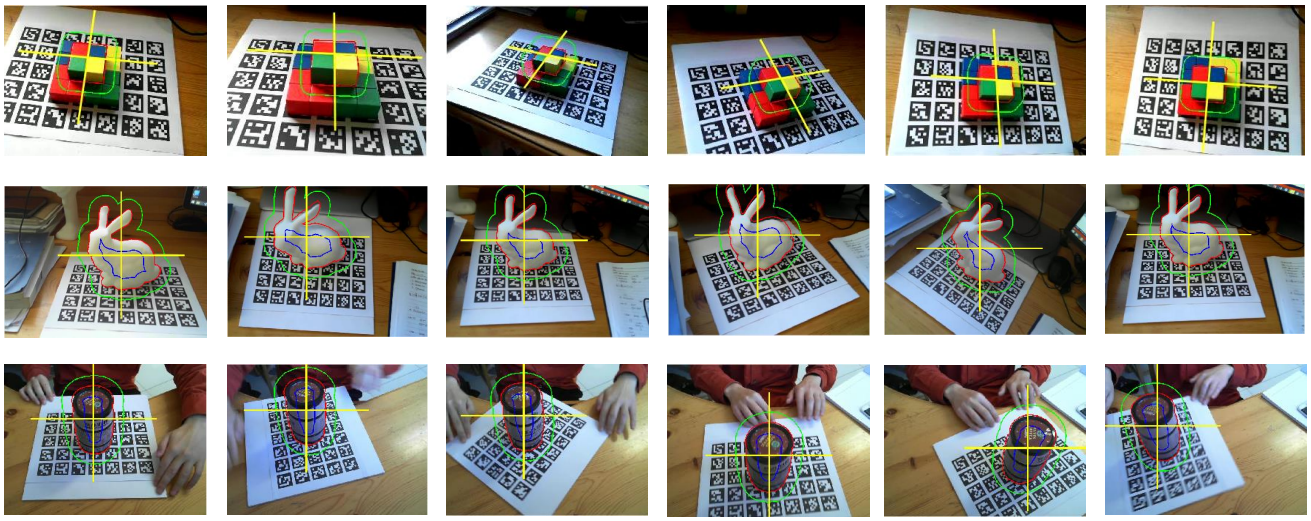


Fig. 7: Sample frames from the HTS Dataset and the corresponding tracking results of the proposed method.

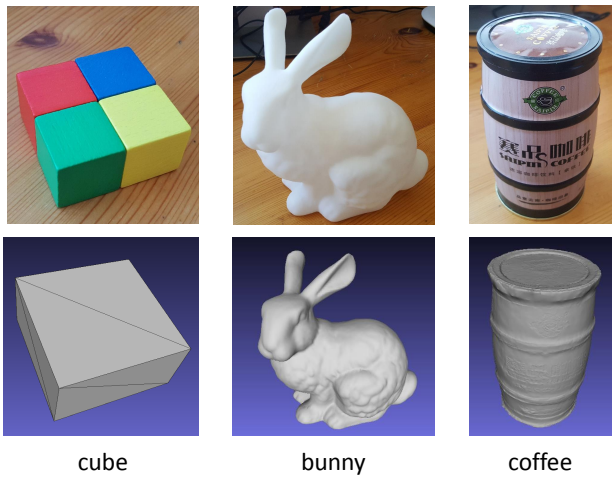


Fig. 8: The newly constructed HTS Dataset. First row: The 3 objects in HTS Dataset. Second row: The corresponding 3D mesh models.

4.2.1 Heterogeneous Object

Compared to the baseline Global Region-based method, our two improved versions perform significantly better for the heterogeneous cube. Specifically, when applying partitioned local statistical constraints, the Local Region-based method obtains 7.1% and 10.6% higher SR scores than the baseline in the two video sequences. After adding in photometric constraints, the proposed Hybrid method again obtains 6.0% and 11.4% improvements in SR score (overall 13.1% and 22.0% improvements compared to the baseline). The reason is that the partitioned sub-regions have more discriminative statistical properties, and the involvement of photometric constraints utilize extra dense pixel-wise information thus making the algorithm more robust. In this experiment, we put the heterogeneously colored cube in a heterogeneous background with very similar color distributions, so that the global foreground and background color histograms also tend to be very similar. As illustrated in Fig. 9, the global foreground and background color histograms tend to be confusing and not discriminative enough for segmenting the foreground and background regions. However, when partitioning the global region into 4 sub-foreground-regions (1∼4) and 4 sub-background-regions (5∼8), the corresponding local color histograms are much more discriminative. This also contributes to a far better foreground (and background) posterior map (Eqs. (7), (8)) as shown in Fig. 10, which makes the pose optimization results more accurate and stable.

Note that this heterogeneously colored cube is specially designed to illustrate the proposed partitioned region-based model. We also use a special way to partition the sub-regions for this cube: we bind the two axes to coincide with the color pattern on the 3D model, and then project the axes onto the image, so that the axes will also coincide with the color pattern in each image frame (as long as the pose estimation is correct). This kind of special partitioning will not work for general objects, and we use it here only for demonstration purpose. For comparison, we also test the performance when using the standard partitioning strategy (which we have introduced in Sect. 3.1) for this cube. The SR scores are 90.1 and 85.7 for the 'Partitioned Local Region' method, and 96.0 and 96.2 for the 'Hybrid Region+Direct' method. The scores indicate that the standard partitioning

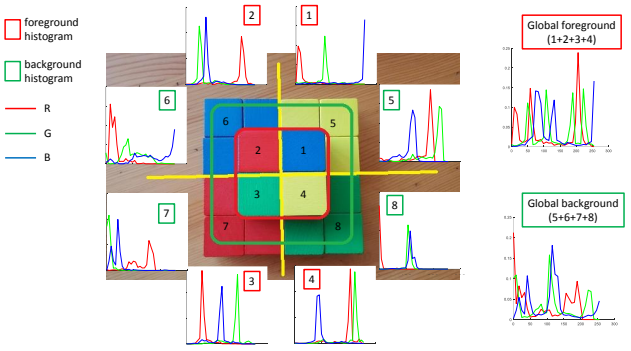


Fig. 9: Comparison of global color histograms and partitioned local color histograms.

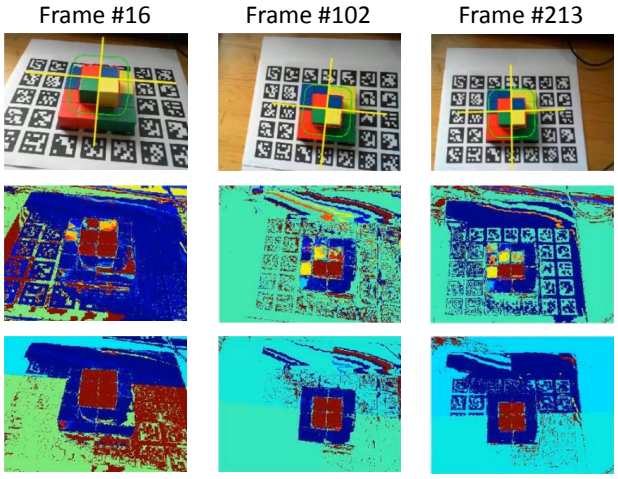


Fig. 10: Comparison of foreground posterior maps of global color model and partitioned local color model. First row: Image frames. Second row: Foreground posterior maps of global color model. Third row: Foreground posterior maps of partitioned local color model. (Red: High probability; Blue: Low probability.)

way is also effective for this cube (although not as effective as the special way here). Except for this colored cube, all the other objects in this dataset and the following datasets are partitioned in the standard way. We clarify that the idea of partitioning the global region into local sub-regions is effective for general heterogeneous objects, as proved in the following experiments.

4.2.2 Texture-less Object

Contrary to the heterogeneous object, the proposed methods show very little improvements for the texture-less (homogeneous) object ('bunny'). This is to be expected, since for homogeneous objects, the global statistical model is already sufficient to describe the foreground region. In this situation, applying partitioned local statistical models only contributes to a slightly better description for the background region. So the Partitioned Local Region-based method only achieves slightly better results. Moreover, as described in Sect. 1, direct methods rely on image gradients, thus they are less suitable for tracking purely homogeneous objects. So adding in photometric constraints does not improve the tracking performance for the texture-less object, either.

4.2.3 Symmetrical Object

For the symmetrical coffee can, the proposed method outperforms the baseline by a large margin. In order to show the advantage of our method for tracking this symmetrical object, we mainly rotate the coffee can around its vertical axis in the 'sym-1' sequence (it also contains small-scale translations), then simultaneously rotate and translate it in the 'sym-2' sequence. In this experiment, the Partitioned Local Region-based method obtains 5.3% and 8.1% higher SR scores than the baseline. After adding in photometric constraints, the proposed Hybrid method again obtains 6.0% and 8.9% improvements in SR score (overall 11.3% and 17.0% improvements compared to the baseline). When a symmetrical object rotates around its axis of symmetry, the projected contour on the image does not change, which would disrupt the traditional global region-based methods. The previous works usually treat the poses around the axis of symmetry identical. But in some applications (e.g. AR applications), it is necessary to distinguish them. In order to tackle this problem, we use more discriminative local statistical models, and distinguish the symmetrical poses by applying photometric constraints. Although the object contour doesn't change while rotating, the texture of the object changes continuously, which could be used to determine the right rotating angle. Note that the Global Region-based method also obtains a relatively high SR score in these two videos. This is because the coffee can isn't a perfectly symmetrical object, some small differences in shape help to distinguish different rotating angles. Moreover, since we rotate it with a relatively low speed, sometimes a frame which is not perfectly tracked will not be detected as tracking failure because the error doesn't exceed the threshold. On the other hand, the SR scores are improved by local region partitioning because it helps to correctly estimate the translation part.

To conclude, the above evaluation results have proven the effectiveness of the two proposed main contributions: the partitioned local statistical model and the combination of statistical + photometric constraints. In particular, the proposed method is shown to be most effective for tracking heterogeneous and symmetrical objects, which solves the inherent problems of global region-based methods.

Some sample frames from the HTS Dataset and the corresponding tracking results of the proposed method are pre-

Table 3: Evaluation results on Original and Noisy Sequences of Rigid Pose Dataset. (Tracking Success Rate, in %)

Method	Constraints	soda		soup		clown		candy		cube		edge		average
		orig	noisy	orig	noisy	orig	noisy	orig	noisy	orig	noisy	orig	noisy	
PWP3D	Statistical	84	84	96	96	96	89	84	84	84	74	85	84	86.7
Bound.Const.	Statistical	96	**97**	95	**98**	95	96	94	**96**	92	93	93	93	94.8
Descriptor Fields	Photometric	92	85	92	93	**98**	93	90	88	96	95	92	94	92.3
Consecutive	Photometric	90	88	95	95	93	92	93	93	95	95	95	95	93.3
Proposed	Stat.+Phot.	**97**	95	**98**	**98**	**98**	**97**	**95**	95	**97**	**96**	**96**	**96**	**96.5**

Table 4: Evaluation results on Occluded Sequences of Rigid Pose Dataset. (Tracking Success Rate, in %)

method	constraints	soda	soup	clown	candy	cube	edge	average
PWP3D	Statistical	44	44	44	39	38	39	41.3
Bound. Cons.	Statistical	73	**82**	81	**84**	**76**	64	76.7
Descriptor Fields	Photometric	50	54	48	66	53	40	51.8
Consecutive	Photometric	66	76	68	81	71	65	71.2
Proposed	Stat. + Phot.	**75**	79	**84**	**84**	73	**67**	**77.0**

Fig. 11: Tracked objects in Rigid Pose Dataset (Pauwels et al, 2013).

sented in Fig. 7. Our method successfully tracks the 3 objects of different types through all the video sequences with high precision.

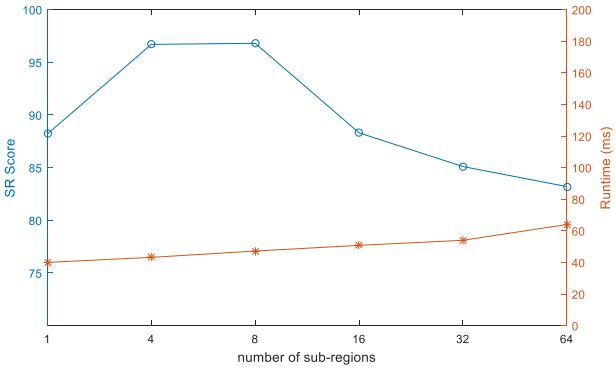


Fig. 12: SR score and runtime (per frame) w.r.t. the number of sub-regions.

4.3 Evaluation on Rigid Pose Dataset

In the second set of experiments, we compare the performance of the proposed method with the state-of-art methods on Rigid Pose Dataset (Pauwels et al, 2013). The Rigid Pose Dataset provides synthetic video sequences of 6 different objects (as shown in Fig. 11) under a variety of realistic conditions. The provided videos compose of original sequences, noisy sequences and occluded sequences, and are featured with fast and wide-range movements, object variability and background cluttering.

4.3.1 Results on Original and Noisy Sequences

Table 3 summarizes the evaluation results on the original and noisy sequences of Rigid Pose Dataset. We compare the results of our method with the other four methods: a state-of-art region-based 3D tracker, PWP3D (Prisacariu and Reid, 2012); a boundary constrained region-based method (Zhao et al, 2014); a gradient-based direct tracker (Crivellaro and Lepetit, 2014); and a method based on direct alignment between consecutive video frames under Lambertian assumption (Seo and Wuest, 2016). We use the same evaluation criteria as in Pauwels et al (2013). We measure the tracking success rate throughout the entire sequence. When tracking is lost, the tracker is automatically reset to the ground truth. The evaluation results in Table 3 show that our method performs averagely better than the other 4 state-of-art methods on original and noisy sequences of this challenging dataset. Specifically speaking, the proposed hybrid method achieves competitive or higher tracking success rate (SR) in tracking both well-textured (soda, soup, candy, cube) and poor-textured (clown, edge) objects. Our method also performs well in the presence of noise. Some indicative video frames and the corresponding tracking results are illustrated in Fig. 13. The proposed method accurately tracks the objects even when they move fast, rotate in a wide range, or move far away from the camera. We also plot the 6-dof tracking results of the (original) 'soup' sequence in Fig. 13 and com-

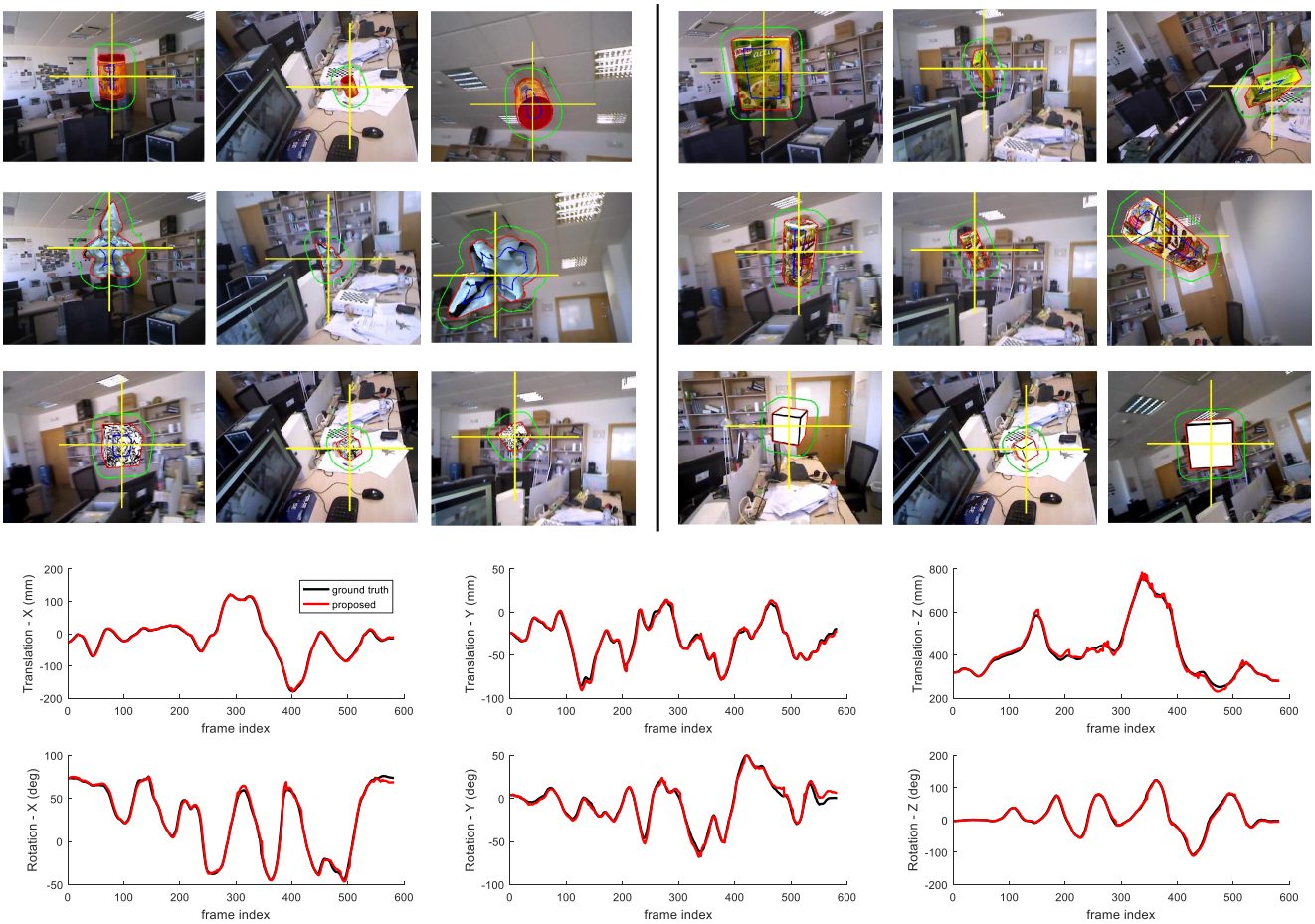


Fig. 13: Top: Sample frames from the Rigid Pose Dataset (Pauwels et al, 2013) and the corresponding tracking results. Bottom: Determined pose parameters of the 'soup' sequence. We compare the 6-dof tracking results (red) with the ground-truth pose parameters (black) to show the tracking accuracy of the proposed method.

pare them with the ground-truth pose parameters, which shows the high tracking accuracy of the proposed method.

As discussed in previous sections, the proposed hybrid tracker combines statistical and photometric constraints together, which increases the robustness of the algorithm. As a result, our hybrid tracker achieves better performance than the region-based PWP3D method (Prisacariu and Reid, 2012) and Boundary Constrained method (Zhao et al, 2014) which utilize only statistical constraints, and also outperforms the two direct methods (Descriptor Fields (Crivellaro and Lepetit, 2014) and Consecutive (Seo and Wuest, 2016)) which utilize only photometric constraints.

4.3.2 Results on Occluded Sequences

We also evaluate our method on the occluded sequences. The results are summarized in Table 4. Our method obtains a slightly higher average SR score on these occluded videos compared to the state-of-art. Although we didn't apply a specific occlusion handling strategy in our method, the localized statistical model itself helps to deal with some certain degrees of partial occlusions (Tjaden et al, 2017). We notice that our method would fail when a large part (e.g. more than half) of the object is occluded, which will be further discussed in Sect. 4.6.

4.3.3 Selecting the Optimal Number of Sub-regions

We demonstrate the impact of using different number of sub-regions n by evaluating on a subset of Rigid Pose Dataset. We test $n = 1, 4, 8, 16, 32, 64$ on the original sequences and record the SR score and runtime w.r.t. each n. The results are plotted in Fig. 12. The curves show that using $n = 4$ or $n = 8$ yields virtually equal best SR scores, while using $n = 16$ or larger degrades the tracking performance. Also, the average runtime per frame slowly increases with n due to the increasing number of local histograms that need to be maintained. So we decide it would be a good choice to use $n = 4$ or $n = 8$, and all the other experimental results in this paper are obtained with $n = 4$.

Table 5: Evaluation results on Dense Tracking Dataset. (Tracking Success Rate, in %)

Method	Constraints	Exp#1	Exp#2	ATLAS#1	ATLAS#2
Global Region	Global Statistical	97.5	88.2	76.1	60.6
Partitioned Local Region	Local Statistical	98.5	95.3	88.5	69.7
Hybrid Region+Direct (Proposed)	Local Statistical + Photometric	**99.7**	**98.4**	**90.4**	**78.8**

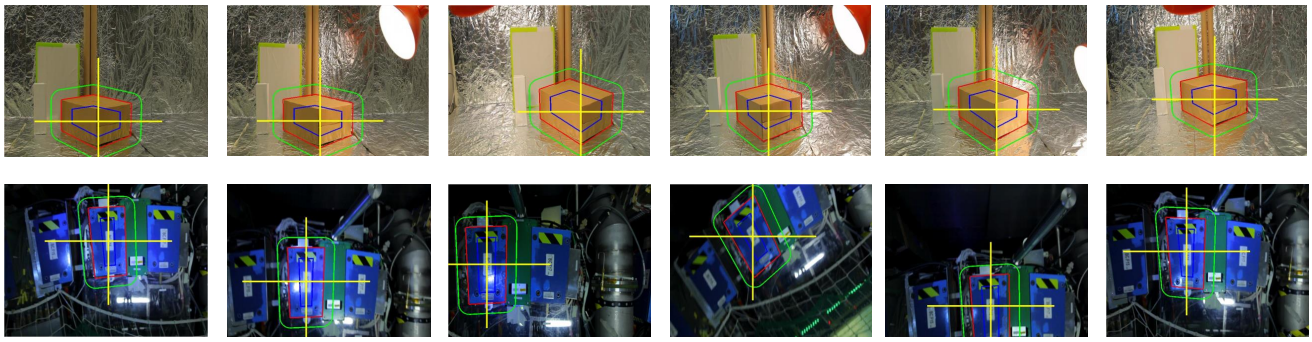


Fig. 14: Sample frames from the Dense Tracking Dataset (Crivellaro and Lepetit, 2014) and the corresponding tracking results.

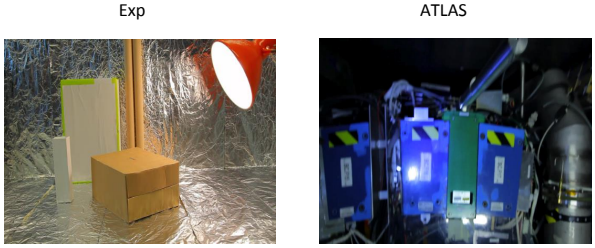


Fig. 15: Two challenging environments in Dense Tracking Dataset (Crivellaro and Lepetit, 2014).

4.4 Evaluation on Dense Tracking Dataset

To better evaluate our method in more complicated and realistic situations, we run a third set of experiments on Dense Tracking Dataset (Crivellaro and Lepetit, 2014). This dataset provides real-world video sequences in two challenging environments, including strong moving light source, bright specularities and motion blur caused by fast movement (as shown in Fig. 15). Moreover, 3D models of the whole scenes are provided in this dataset, so we crop out the main object for tracking evaluation. While both two scenes contain strong moving light sources, the first scene (videos Exp#1 and Exp#2) contains specular background, and the second scene (videos ATLAS#1 and ATLAS#2) contains specular objects.

The evaluation results are shown in Table 5. Some indicative frames together with our tracking results are illustrated in Fig. 14. The proposed Hybrid method obtains better results than Global Region-based method and Partitioned Local Region-based method in the presence of complex illumination changes and specularity (with specular object or specular background), which demonstrates the robustness of our method to non-Lambertian lighting effects. Note that the use of gradient-based Descriptor Fields in our photometric energy helps to overcome the disadvantage of direct methods when dealing with illumination changes and specularity, as discussed in Sect. 3.3.

4.5 Evaluation on the Dataset of (Tjaden et al, 2017)

In order to compare the proposed method with the recent state-of-art method (Tjaden et al, 2017), we conduct the last experiment on the semi-synthetic dataset of (Tjaden et al, 2017). As with in (Tjaden et al, 2017), we calculate the RMSE (Root Mean Squared Error) and STD (Standard Deviation) of translation in (x, y, z) directions and rotation around the (x, y, z) axes. Since (Tjaden et al, 2017) is a purely region-based method, we also evaluate the region-only version of our method (without including the photometric term). The results are shown in Table 6. We compare the 6-dof pose parameters estimated by the proposed (hybrid) method and Tjaden et al (2017) in Fig. 16. The calculated mean errors indicate that both two versions of our method obtain a relatively lower (but comparable) tracking precision compared to (Tjaden et al, 2017). But note that as with in (Tjaden et al, 2017), the errors in Table 6 are calculated from the tracking results of frames 1∼868, because Tjaden et al (2017) suffers from a silhouette pose ambiguity for rotation starting from frame 869, as shown in the second row of Fig. 16. In contrast, this pose ambiguity is successfully avoided by the proposed hybrid method thanks to the added photometric term

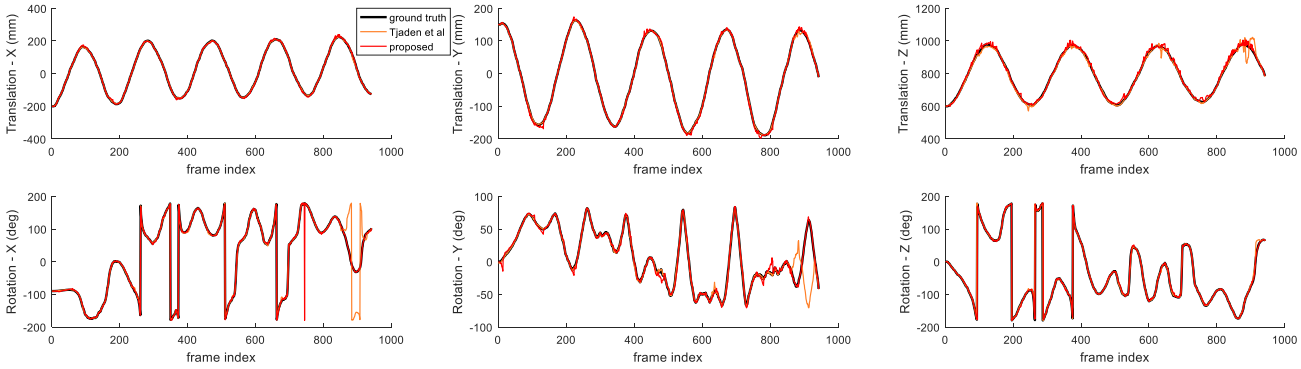


Fig. 16: Comparison of the 6-dof tracking results estimated by our method (red) and Tjaden et al (2017) (orange). The ground-truth pose parameters are in black.

Table 6: Comparison with (Tjaden et al, 2017) (RMSE $\pm$ STD).

	Tjaden et al	Proposed	Proposed (region-only)
x	$1.2 \pm 0.9mm$	$1.8 \pm 1.5mm$	$1.8 \pm 1.6mm$
y	$1.3 \pm 1.1mm$	$2.0 \pm 1.5mm$	$2.2 \pm 1.6mm$
z	$7.5 \pm 5.7mm$	$10.9 \pm 7.8mm$	$10.8 \pm 7.2mm$
r_x	$2.3 \pm 2.3°$	$2.0 \pm 1.3°$	$2.2 \pm 1.3°$
r_y	$1.3 \pm 1.2°$	$1.8 \pm 1.4°$	$1.9 \pm 1.8°$
r_z	$1.1 \pm 2.0°$	$2.0 \pm 1.6°$	$2.0 \pm 1.5°$

in our hybrid energy function, which proves the advantage of our method in dealing with silhouette pose ambiguities.

4.6 Discussions

Regarding the above experiments, several issues need to be discussed in detail.

The limitation of our method. Although we have demonstrated the robustness of our method in several challenging datasets, it still has some limitations. Firstly, an efficient occlusion handling strategy is yet to be developed. As declared in Sect. 4.3.2, our method could handle a certain degree of occlusion, but would fail when encountering heavy occlusions. We expect that the robustness of our method to heavy occlusions could be improved by cooperating with some learning-based strategy (such as training a deep neural network to accurately detect the occluded area). Secondly, we find that the proposed method gets relatively more sensitive to initialization quality after adding in photometric constraints. This is because direct methods are inherently not able to recover from a bad initialization (which could be seen as a pose drift, as mentioned in Table 1). But thanks to the statistical term, our method still has the ability to recover from imperfect initializations. We test on the original sequences of Rigid Pose Dataset by randomly sampling a perturbation angle and a translational offset for each frame. According to our experiments, the proposed method is able to recover from a rotation error of $\pm 10°$, or a translation error of $\pm 10\%$ of the object size. So we think it is acceptable since most of the state-of-art 3D object detection methods could provide an initial pose with much better precision.

The cooperation of the statistical term and the photometric term. As discussed in Sect. 1, the statistical term from region-based methods and the photometric term from direct methods have complementary properties and would help each other in some specific tracking scenarios. The above experiments have also demonstrated that the cooperation of these two terms would contribute to better tracking results. But it is necessary to discuss if there are some cases when one term disturbs the convergence of the second one. For example, since traditional direct methods are sensitive to illumination changes and specularity, photometric term seems to play a negative role in this case. But as declared in Sect. 3.3 and proved by the experimental results in Sect. 4.4, the use of gradient-based Descriptor Fields instead of intensity or color in the photometric term has significantly alleviated this problem. The involvement of the photometric term actually helps to improve the SR scores in videos containing lighting variations. On the other hand, we find that the photometric term do disturb the statistical term in recovering from pose drifts or imperfect initializations (but within an acceptable degree), as discussed right above. Based on the overall performance of the proposed hybrid tracker, we consider these two terms generally cooperate very well, only with some minor interferences.

The influence of the accuracy and complexity of 3D models. The accuracy of 3D models would influence the tracking accuracy of model-based methods. To the best of our knowledge, except for some simple objects (such as cubes) which could be precisely created, most of the 3D models used in the above datasets are reconstructed by some 3D modeling softwares (which are based on off-line Structure-from-Motion and Multi-View Stereo techniques) from photos of different viewpoints. According to our experience, the

Table 7: Model Loading and Rendering Time Analysis

Model	Triangles	T_{load}	T_{render}
ATLAS	2	0.07s	0.9ms
clown	800	0.2s	1.4ms
driller	25306	0.4s	1.6ms
coffee	508645	6.0s	3.6ms

accuracy of the reconstructed model is usually within several millimeters for an everyday object. This kind of 3D models are not as accurate as 3D models acquired from laser scanning, but they are enough for our method to produce good tracking results as shown in the above experiments. On the other hand, the complexity of the 3D models would affect the runtime of our method. We choose 4 representative models from the evaluated datasets, and the number of triangles ranges from 2 to 508645. The model loading and rendering time for each model is listed in Table 7 ('driller' is the 3D model used in the dataset of (Tjaden et al, 2017)). For 3D model with a large number of triangles, the loading time increases significantly. But since the model only needs to be loaded once into the renderer before tracking the first frame, it will not affect the subsequent tracking process. The model rendering time also increases with the number of triangles, but stays within an acceptable range. We find that a general everyday object could be well modeled by 1k∼100k triangles, which we believe is feasible for real-time tracking.

5 Conclusion

We have presented a robust hybrid 3D object tracking method by integrating statistical constraints from region-based methods and photometric constraints from direct methods. Through this novel integration, we are able to make full use of the image information: (1) Both statistical distributions and raw pixel values of the image are utilized; (2) Both pixels around the contour and pixels inside the contour are properly used. Therefore, the proposed hybrid 3D object tracker gains superior robustness over region-based method or direct method alone. Experiments on a newly constructed real-world dataset and several challenging public datasets have demonstrated the competitive or superior performance of our method compared to the state-of-art.

Appendix A The Hessian Approximation

Here we give a mathematical explanation of the Hessian approximation (Eq. (19)) used for the statistical term in our energy function:

$$E_{fb} = -\sum_{i=1:n}\sum_{\mathbf{x}\in\Omega_i} \log\left(H_e\left(\Phi\left(\mathbf{x}\right)\right)P_{f_i} + \left(1-H_e\left(\Phi\left(\mathbf{x}\right)\right)\right)P_{b_i}\right) \tag{24}$$

We first consider the statistical energy for a single pixel $\mathbf{x}$:

$$E_{fb}\left(\mathbf{x}\right) = -\log\left(H_e\left(\Phi\left(\mathbf{x}\right)\right)P_{f_i} + \left(1-H_e\left(\Phi\left(\mathbf{x}\right)\right)\right)P_{b_i}\right) \tag{25}$$

The Jacobian for pixel $\mathbf{x}$:

$$\begin{aligned}\mathbf{J}_{fb}\left(\mathbf{x}\right) &= \frac{\partial E_{fb}\left(\mathbf{x}\right)}{\partial \mathbf{p}} \\ &= -\frac{P_{f_i}-P_{b_i}}{H_e\left(\Phi\left(\mathbf{x}\right)\right)P_{f_i} + \left(1-H_e\left(\Phi\left(\mathbf{x}\right)\right)\right)P_{b_i}}\frac{\partial H_e\left(\Phi\left(\mathbf{x}\right)\right)}{\partial \mathbf{p}}\end{aligned} \tag{26}$$

Here $\mathbf{J}_{fb}\left(\mathbf{x}\right) \in \mathbf{R}^6$, corresponding to the 6-dof pose parameter $\mathbf{p}$. The m-th element can be written as:

$$\begin{aligned}\left[\mathbf{J}_{fb}\left(\mathbf{x}\right)\right]_m &= \frac{\partial E_{fb}\left(\mathbf{x}\right)}{\partial p_m} \\ &= -\frac{P_{f_i}-P_{b_i}}{H_e\left(\Phi\left(\mathbf{x}\right)\right)P_{f_i} + \left(1-H_e\left(\Phi\left(\mathbf{x}\right)\right)\right)P_{b_i}}\frac{\partial H_e\left(\Phi\left(\mathbf{x}\right)\right)}{\partial p_m}\end{aligned} \tag{27}$$

The Hessian for pixel $\mathbf{x}$: $\mathbf{H}_{fb}\left(\mathbf{x}\right) \in \mathbf{R}^{6\times 6}$, and the (m,n)-th element can be written as:

$$\begin{aligned}&\left[\mathbf{H}_{fb}\left(\mathbf{x}\right)\right]_{m,n} = \frac{\partial^2 E_{fb}\left(\mathbf{x}\right)}{\partial p_m \partial p_n} \\ =&\frac{\partial\left[-\frac{P_{f_i}-P_{b_i}}{H_e(\Phi(\mathbf{x}))P_{f_i}+(1-H_e(\Phi(\mathbf{x})))P_{b_i}}\frac{\partial H_e(\Phi(\mathbf{x}))}{\partial p_m}\right]}{\partial p_n} \\ =&\frac{\left(P_{f_i}-P_{b_i}\right)^2\frac{\partial H_e(\Phi(\mathbf{x}))}{\partial p_n}}{\left[H_e\left(\Phi\left(\mathbf{x}\right)\right)P_{f_i} + \left(1-H_e\left(\Phi\left(\mathbf{x}\right)\right)\right)P_{b_i}\right]^2}\frac{\partial H_e\left(\Phi\left(\mathbf{x}\right)\right)}{\partial p_m} - \\ &\frac{P_{f_i}-P_{b_i}}{H_e\left(\Phi\left(\mathbf{x}\right)\right)P_{f_i} + \left(1-H_e\left(\Phi\left(\mathbf{x}\right)\right)\right)P_{b_i}}\frac{\partial^2 H_e\left(\Phi\left(\mathbf{x}\right)\right)}{\partial p_m \partial p_n}\end{aligned} \tag{28}$$

We denote the first term as h_1 (which contains first order derivatives), and the second term as h_2 (which contains second order derivatives):

$$\left[\mathbf{H}_{fb}\left(\mathbf{x}\right)\right]_{m,n} = h_1 - h_2 \tag{29}$$

Comparing to Eq. (27), we have:

$$h_1 = \left[\mathbf{J}_{fb}\left(\mathbf{x}\right)\right]_m\left[\mathbf{J}_{fb}\left(\mathbf{x}\right)\right]_n = \left[\mathbf{J}_{fb}(\mathbf{x})^T\mathbf{J}_{fb}\left(\mathbf{x}\right)\right]_{m,n} \tag{30}$$

As with the standard Gauss-Newton method (which aims to solve non-linear least square problems), we obtain an

approximation of the Hessian matrix by ignoring the second-order derivative term h_2:

$$\left[\mathbf{H}_{fb}(\mathbf{x})\right]_{m,n} = h_1 = \left[\mathbf{J}_{fb}(\mathbf{x})^T \mathbf{J}_{fb}(\mathbf{x})\right]_{m,n} \tag{31}$$

$$\mathbf{H}_{fb}(\mathbf{x}) = \mathbf{J}_{fb}(\mathbf{x})^T \mathbf{J}_{fb}(\mathbf{x}) \tag{32}$$

In the end, we sum over all of the pixels and obtain the Hessian approximation used in Sect. 3.4:

$$\mathbf{H}_{fb} = \sum_{i=1:n} \sum_{\mathbf{x}\in\Omega_i} \mathbf{J}_{fb}(\mathbf{x})^T \mathbf{J}_{fb}(\mathbf{x}) \tag{33}$$

Acknowledgements This work is partly supported by the National Natural Science Foundation of China under Grant U1533132.

References

Alismail H, Browning B, Lucey S (2016) Robust tracking in low light and sudden illumination changes. In: International Conference on 3D Vision (3DV), IEEE, pp 389–398

Baker S, Matthews I (2004) Lucas-kanade 20 years on: A unifying framework. International Journal of Computer Vision 56(3):221–255

Bibby C, Reid I (2008) Robust real-time visual tracking using pixel-wise posteriors. In: European Conference on Computer Vision (ECCV), Springer, pp 831–844

Caron G, Dame A, Marchand E (2014) Direct model based visual tracking and pose estimation using mutual information. Image and Vision Computing 32(1):54–63

Chen L, Zhou F, Shen Y, Tian X, Ling H, Chen Y (2017) Illumination insensitive efficient second-order minimization for planar object tracking. In: IEEE International Conference on Robotics and Automation (ICRA), IEEE

Choi C, Christensen HI (2010) Real-time 3d model-based tracking using edge and keypoint features for robotic manipulation. In: IEEE International Conference on Robotics and Automation (ICRA), pp 4048–4055

Crivellaro A, Lepetit V (2014) Robust 3d tracking with descriptor fields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3414–3421

Dambreville S, Sandhu R, Yezzi A, Tannenbaum A (2008) Robust 3d pose estimation and efficient 2d region-based segmentation from a 3d shape prior. In: European Conference on Computer Vision (ECCV), Springer, pp 169–182

Engel J, Schöps T, Cremers D (2014) Lsd-slam: Large-scale direct monocular slam. In: European Conference on Computer Vision (ECCV), pp 834–849

Engel J, Koltun V, Cremers D (2017) Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence

Garrido-Jurado S, Muñoz-Salinas R, Madrid-Cuevas FJ, Marín-Jiménez MJ (2014) Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognition 47(6):2280–2292

Hexner J, Hagege RR (2016) 2d-3d pose estimation of heterogeneous objects using a region based approach. International Journal of Computer Vision 118(1):95–112

Hinterstoisser S, Holzer S, Cagniart C, Ilic S, Konolige K, Navab N, Lepetit V (2011) Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In: International Conference on Computer Vision (ICCV), pp 858–865

Kehl W, Manhardt F, Tombari F, Ilic S, Navab N (2017a) Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In: International Conference on Computer Vision (ICCV), pp 1521–1529

Kehl W, Tombari F, Ilic S, Navab N (2017b) Real-time 3d model tracking in color and depth on a single cpu core. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 745–753

Kerl C, Sturm J, Cremers D (2013) Robust odometry estimation for rgb-d cameras. In: IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 3748–3754

Lepetit V, Fua P (2005) Monocular model-based 3D tracking of rigid objects. Now Publishers Inc

Lima JP, Simões F, Figueiredo L, Kelner J (2010) Model based markerless 3d tracking applied to augmented reality. Journal on 3D Interactive Systems 1

Loesch A, Bourgeois S, Gay-Bellile V, Dhome M (2015) Generic edgelet-based tracking of 3d objects in real-time. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 6059–6066

Lucas BD, Kanade T, et al (1981) An iterative image registration technique with an application to stereo vision. In: International Joint Conference on Artificial Intelligence (IJCAI), vol 81, pp 674–679

Panin G, Roth E, Knoll A (2008) Robust contour-based object tracking integrating color and edge likelihoods. In: VMV, pp 227–234

Park Y, Lepetit V, Woo W (2008) Multiple 3d object tracking for augmented reality. In: IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR), pp 117–120

Pauwels K, Rubio L, Diaz J, Ros E (2013) Real-time model-based rigid object pose estimation and tracking combining dense and sparse visual cues. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2347–2354

Petit A, Marchand E, Kanani K (2013) A robust model-based tracker combining geometrical and color edge information. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 3719–3724

Prisacariu VA, Reid ID (2012) Pwp3d: Real-time segmentation and tracking of 3d objects. International Journal of Computer Vision 98(3):335–354

Prisacariu VA, Kahler O, Murray DW, Reid ID (2013) Simultaneous 3d tracking and reconstruction on a mobile phone. In: IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, pp 89–98

Ren C, Prisacariu V, Kähler O, Reid I, Murray D (2017) Real-time tracking of single and multiple objects from depth-colour imagery using 3d signed distance functions. International Journal of Computer Vision pp 1–16

Ren CY, Prisacariu V, Kaehler O, Reid I, Murray D (2014) 3d tracking of multiple objects with identical appearance using rgb-d input. In: International Conference on 3D Vision (3DV), IEEE, vol 1, pp 47–54

Scandaroli GG, Meilland M, Richa R (2012) Improving ncc-based direct visual tracking. In: European Conference on Computer Vision (ECCV), Springer, pp 442–455

Seo BK, Wuest H (2016) A direct method for robust model-based 3d object tracking from a monocular rgb image. In: European Conference on Computer Vision Workshop (ECCVW), pp 551–562

Seo BK, Park H, Park JI, Hinterstoisser S, Ilic S (2014) Optimal local searching for fast and robust textureless 3d object tracking in highly cluttered backgrounds. IEEE Transactions on Visualization and Computer Graphics 20(1):99–110

Singhal P, White R, Christensen H (2016) Multi-modal tracking for object based slam. arXiv preprint arXiv:160304117

Tjaden H, Schwanecke U, Schömer E (2016) Real-time monocular segmentation and pose tracking of multiple objects. In: European Conference on Computer Vision (ECCV), Springer, pp 423–438

Tjaden H, Schwanecke U, Schömer E (2017) Real-time monocular pose estimation of 3d objects using temporally consistent local color histograms. In: International Conference on Computer Vision (ICCV), pp 124–132

Zhao S, Wang L, Sui W, Wu Hy, Pan C (2014) 3d object tracking via boundary constrained region-based model. In: IEEE International Conference on Image Processing (ICIP), IEEE, pp 486–490

Zhong L, Lu M, Zhang L (2017) A direct 3d object tracking method based on dynamic textured model rendering and extended dense feature fields. IEEE Transactions on Circuits and Systems for Video Technology