Illumination Insensitive Efficient Second-order Minimization for Planar Object Tracking

Lin Chen[1], Fan Zhou[1,2], Yu Shen[3], Xiang Tian[1,4], Haibin Ling[3,5], Yaowu Chen[1,4]

***Abstract*— Tracking for planar objects is an important issue to vision-based robotic applications. In direct visual tracking (DVT) methods, the similarity between two images is often measured through the sum of squared differences (SSD) especially with the efficient second-order minimization (ESM) due to its simplicity and efficiency. However, SSD-based ESM is not robust to illumination changes since it is usually built upon the brightness constancy assumption. Contrast to image brightness, gradient orientations (GO) are invariant to both linear and non-linear illumination changes as verified in practice. Based on GO, we propose an illumination insensitive ESM method for planar object tracking in this paper. In order to introduce GO into the ESM, we generalized the original ESM formulas for multi-dimensional features. In addition, a denoising method based on the Perona-Malik function and a mask image were suggested to improve GO's robustness against image noise and low texture. Our experimental results on dataset for planar objects with illumination changes and a benchmark dataset confirm the proposed method is robust to illumination variations and capable to deal with the general tracking challenges.**

I. Introduction

Visual tracking for planar objects is a fundamental technique of many vision-based robotic applications, such as visual odometry [1], [2], visual servoing [3] and visual SLAM [4]. Feature-based method and direct method are the two main categories of visual tracking techniques. The former solves tracking problem by detecting and matching distinguishable features (e.g. corners or edges), while for the latter only image intensity information is used. In this paper, the direct visual tracking (DVT) methods will be mainly addressed.

Typically, the objective of DVT methods is to obtain the parameters of a transformation model as to minimize the sum of squared differences (SSD) between the reference image and the current image. For this nonlinear least squares problem, the efficient second-order minimization (ESM) [3], [5] method is the most popularly used optimization technology because of its high convergence rate and low computation cost. However, such SSD-based ESM is not robust under illumination variations, since it is usually built upon the brightness constancy assumption (BCA) [3], [6].

Currently, there are two main strategies to improve the robustness of SSD-based ESM to illumination variations. The first one is modelling the illumination changes. According to the references of [7]–[10], the illumination changes were modeled as an affine transformation which only takes global changes of illumination into consideration, while other researches employed a more complete model based on the thin-plate spline to compensate local illumination variations [11], [12]. The second strategy is applying robust similarity measures to replace the SSD. To handle complex illumination variations, Scandaroli et al. [13] calculated the local Normalized Cross Correlation (NCC), while Richa et al. [14] used the sum of conditional variance (SCV) to cope with non-linear illumination changes. Later, Richa et al. [15] proposed LSCV to decrease the SCV's sensitivity to local illumination changes. In addition, mutual information (MI) [16]–[18] and cross cumulative residual entropy (CCRE) [19], [20] applied in the medical image registration domain were also successfully introduced to improve the robustness to illumination changes.

Unlike the strategies mentioned above, we introduced robust dense features - the gradient orientations (GO) into DVT methods for the first time, to our knowledge. This choice was motivated by GO's insensitivities to global illumination changes [21], non-linear illumination changes [22] and changes in illumination directions [23], which has been verified in practice. Due to its robustness, GO is widely used in computer vision applications, e.g. stereo matching [24] and motion estimation [21], [25] and also used as the component in common descriptors, e.g SIFT [22] and HOG [26]. By using GO as features, here we propose a novel robust DVT method GO-ESM based on ESM for planar object tracking which is insensitive to illumination. Since GO can be considered as a two-dimensional vector, we generalized the original ESM [3] formulas for multi-dimensional features, which allowed us to combine the robustness of GO with the advantages of ESM method. To introduce GO, we also suggested an anisotropic diffusion denoising method based on the Perona-Malik function [27] to handle image noise and applied a mask image to deal with lox texture challenges. Fig. 1 shows an overview of GO-ESM.

The GO-ESM was compared to the state-of-the-art tracking methods on dataset for planar objects with illumination

*This work is supported by the Fundamental Research Funds for the Central Universities, National Key Research and Development Plan (Grant No. 2016YFB1001200), National Natural Science Foundation of China Grant No. 61528204 and National Science Foundation Grant 1350521

[1]Authors are with Institute of Advanced Digital Technology and Instrument, Zhejiang University, China. linlindedream@zju.edu.cn, {fanzhou,tianx,cyw}@mail.bme.zju.edu.cn

[2]Fan Zhou is with Key Laboratory for Biomedical Engineering of Ministry of Education of China, Zhejiang University, China.

[3]Authors are with Meitu HiScene Lab, HiScene Information Technologies, Shanghai, China. sheny@hiscene.com, hbling@temple.edu

[4]Authors are with Zhejiang Provincial Key Laboratory for Network Multimedia Technologies, Hangzhou, China.

[5]Haibin Ling is with Computer & Information Sciences Department, Temple University, USA.

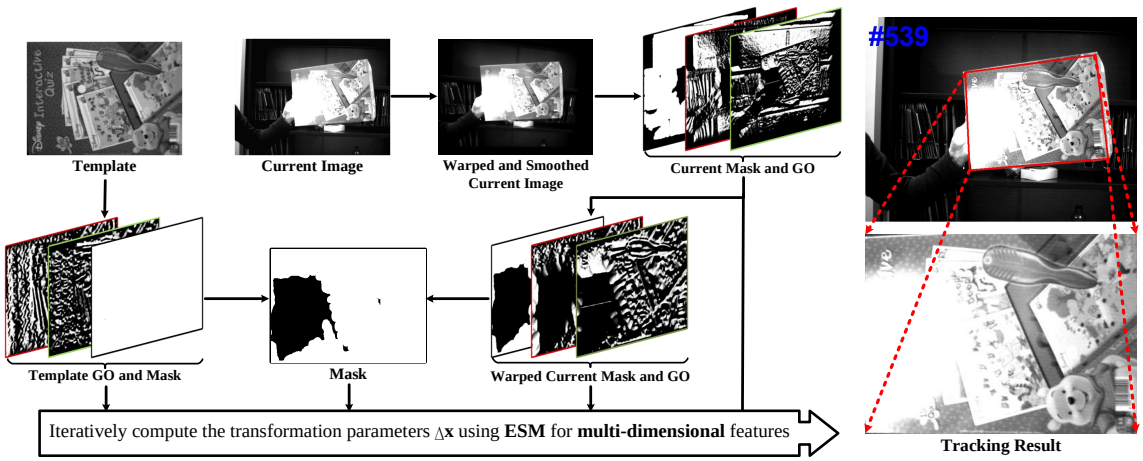


Fig. 1. Overview of the proposed GO-ESM tracking method.

changes (POIC) and the TMT benchmark dataset [28]. The obtained results demonstrate that GO-ESM is robust to illumination variations as well as general tracking challenges.

This paper is organized as follows. Followed by notations and background in Sec. II, Sec. III introduces the proposed GO-ESM tracking method. Then Sec. IV evaluates and analyses experimental results and Sec. V concludes.

II. Notations and Background

Let $I \in \mathbb{R}^{n\times m}$ be an image matrix and $\mathbf{p} = (u, v)^\top \in \{1, 2, \ldots, n\} \times \{1, 2, \ldots, m\}$ be pixel coordinates, $I(\mathbf{p})$ is therefore the intensity of the pixel $\mathbf{p}$. Let warp function $w(\mathbf{x}, \mathbf{p})$ be an image transformation function (e.g. a translation or a projective transformation) where $\mathbf{x} \in \mathbb{R}^p$ is transformation parameters.

DVT methods formulate tracking as an image registration problem. For a planar object (also called the template) which is selected in the reference image I_R in some region of $q = n \cdot m$ pixels, there are parameters $\mathbf{x}$ of a planar homographic warp that map the pixel of the template $I_\mathrm{R}(\mathbf{p})$ into its corresponding pixel $I_\mathrm{C}(w(\mathbf{x}, \mathbf{p}))$ in the current image I_C. Then, the goal in DVT is to obtain the estimation of parameters $\widehat{\mathbf{x}}$ that optimizes the similarity measures $S(I_\mathrm{R}(\mathbf{p}), I_\mathrm{C}(w(\mathbf{x}, \mathbf{p})))$, which can be simplified as $S(\mathbf{x})$ since I_R is constant, between the template and the warped current image. As mentioned above, the SSD is the most used similarity measure because of its simplicity and efficiency, which is usually built upon the brightness constancy assumption (BCA) [3], [6]:

$$I_\mathrm{C}(w(\mathbf{x}, \mathbf{p}_i)) = I_\mathrm{R}(\mathbf{p}_i),\ i \in \{1, 2, \ldots, q\} \tag{1}$$

Then, the tracking system can be expressed as an nonlinear least squares problem:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} S_\mathrm{SSD}(\mathbf{x}) \tag{2}$$

where

$$S_\mathrm{SSD}(\mathbf{x}) = \sum_{i=1}^{q} (I_\mathrm{C}(w(\mathbf{x}, \mathbf{p}_i)) - I_\mathrm{R}(\mathbf{p}_i))^2 \tag{3}$$

The ESM [3], [5] method is popular to solve (2) since its high convergence rate and low computation cost. According to (1), we have:

$$d_i(\mathbf{x}) = I_\mathrm{C}(w(\mathbf{x}, \mathbf{p}_i)) - I_\mathrm{R}(\mathbf{p}_i) = 0 \tag{4}$$

Let $\mathbf{d}(\mathbf{x}) \in \mathbb{R}^q$ contain the image differences:

$$\mathbf{d}(\mathbf{x}) = \begin{bmatrix} d_1(\mathbf{x}) & d_2(\mathbf{x}) & \ldots & d_q(\mathbf{x}) \end{bmatrix}^\top \tag{5}$$

Consider the second-order Taylor series of $\mathbf{d}(\mathbf{x})$ about $\widehat{\mathbf{x}}_c$ and then evaluated at $\mathbf{x} = \widehat{\mathbf{x}}_c \circ \Delta\mathbf{x}$ (where $\circ$ is the binary operation following the definition in [5]):

$$\mathbf{d}(\widehat{\mathbf{x}}_c \circ \Delta\mathbf{x}) \approx \mathbf{d}(\widehat{\mathbf{x}}_c) + \mathbf{J}_\mathbf{d}(\widehat{\mathbf{x}}_c)\Delta\mathbf{x} + \frac{1}{2}\mathbf{M}(\widehat{\mathbf{x}}_c, \Delta\mathbf{x})\Delta\mathbf{x} \tag{6}$$

where $\widehat{\mathbf{x}}_c$ and $\Delta\mathbf{x}$ are the current estimation and increment of parameters $\mathbf{x}$, respectively; $\mathbf{J}_\mathbf{d}(\mathbf{x}) \in \mathbb{R}^{q\times p}$ is the Jacobian matrix of $\mathbf{d}(\mathbf{x})$; while $\mathbf{M}(\widehat{\mathbf{x}}_c, \Delta\mathbf{x})$ is a matrix depending on $\Delta\mathbf{x}$ and the Hessian matrices of $\mathbf{d}(\mathbf{x})$ evaluated at $\widehat{\mathbf{x}}_c$. ESM is considered to be able to achieve a convergence rate as high as the Newton's method [29] and to avoid the computation of the Hessian at the same time by using a first-order Taylor series of $\mathbf{J}_\mathbf{d}(\mathbf{x})$ about $\widehat{\mathbf{x}}_c$ evaluated at $(\widehat{\mathbf{x}}_c \circ \Delta\mathbf{x})$:

$$\mathbf{J}_\mathbf{d}(\widehat{\mathbf{x}}_c \circ \Delta\mathbf{x}) \approx \mathbf{J}_\mathbf{d}(\widehat{\mathbf{x}}_c) + \mathbf{M}(\widehat{\mathbf{x}}_c, \Delta\mathbf{x}) \tag{7}$$

Besides, using the BCA (1), we also have:

$$\mathbf{d}(\widehat{\mathbf{x}}_c \circ \Delta\mathbf{x}) \approx \mathbf{0} \tag{8}$$

and

$$\mathbf{J}_\mathbf{d}(\widehat{\mathbf{x}}_c \circ \Delta\mathbf{x}) \approx \mathbf{J}_\mathrm{R} \tag{9}$$

where $\mathbf{J}_\mathrm{R} \in \mathbb{R}^{q\times p}$ denotes the Jacobian matrix of $I_\mathrm{R}(w(\mathbf{x}, \mathbf{p}))$ at $\mathbf{x} = \mathbf{e}$ (where $\mathbf{e}$ is the parameters of an identity warp). Inserting (7), (8) and (9) into (6), we can obtain $\Delta\mathbf{x}$ without computing $\mathbf{M}(\widehat{\mathbf{x}}_c, \Delta\mathbf{x})$:

$$\Delta\mathbf{x} = -2(\mathbf{J}_\mathrm{R} + \mathbf{J}_\mathbf{d}(\widehat{\mathbf{x}}_c))^+ \mathbf{d}(\widehat{\mathbf{x}}_c) \tag{10}$$

where $(\cdot)^+$ is the pseudo-inverse of a matrix. The Jacobian $\mathbf{J}_\mathrm{R}$ is constant which can be precomputed, while the Jacobian $\mathbf{J}_\mathbf{d}(\widehat{\mathbf{x}}_c)$ needs to be updated depending on $\widehat{\mathbf{x}}_c$ (refer to [3] for

details of the Jacobian computations). The increment of the parameters $\Delta\mathbf{x}$ is estimated iteratively according to the ESM update law (10), until it is below a threshold, i.e., $\|\Delta\mathbf{x}\| < \varepsilon$. At each iteration, the parameters $\mathbf{x}$ is updated as:

$$\widehat{\mathbf{x}} = \widehat{\mathbf{x}}_c \circ \Delta\mathbf{x} \tag{11}$$

III. The Proposed Visual Tracking Method

As mentioned in Sec. I, SSD-based ESM [3] is not robust to illumination changes because it usually assumes brightness constancy [3], [6]. Here, we propose a novel robust DVT method on the basis of the illumination insensitive ESM which uses gradient orientations (GO) as image features (shown in Fig. 1).

The GO of image I can be defined as two-dimensional feature images obtained by dividing the gradient vectors by their magnitudes at each pixel:

$$\begin{bmatrix} \mathcal{O}_x \\ \mathcal{O}_y \end{bmatrix} = \begin{bmatrix} \frac{\nabla_x I}{\|\nabla I\|} & \frac{\nabla_y I}{\|\nabla I\|} \end{bmatrix}^\top \tag{12}$$

where ∇ denotes the gradient with the subscript x and y denoting the x and y orientations, respectively. It should be noted that GO is considered as a two-dimensional vector $(\mathcal{O}_x, \mathcal{O}_y)$ instead of angular values θ (rad), because based on this definition the extra precaution needs not to be taken to compute GO's derivatives (difference between two angles cannot exceed π [21]). However, the original ESM [3] formulas do not apply to multi-dimensional features. To address this problem, we generalize the ESM to the condition of multi-dimension, which allows the combination of ESM with GO. On the other hand, image noise and texture need additional treatments for using GO since GO is sensitive to them according to (12). To this end, we also suggest an anisotropic diffusion denoising method which can preserve the image structures while reducing noise, and employ a mask image to handle lox texture.

To introduce GO-ESM method, the rest of this section is divided into the following four subsections. First, the robustness of using GO to image noise and low texture is detailed in Sec. III-A and Sec. III-B, respectively. Then, a generalization of ESM for multi-dimensional features is derived in Sec. III-C. Finally, the GO-ESM is summarized in Sec. III-D.

A. Robustness to image noise

It is obvious that GO is sensitive to image noise due to the computation of (12). To handle this problem, the input images need to be smoothed before extracting GO. However, simply using a Gaussian filter with a fixed kernel could not obtain a general good performance for different sequences (will be shown in Fig. 4), since the image texture information which the extraction of GO depends on is blurred by the linear filter method.

The Perona-Malik function [27] is a kind of anisotropic diffusion filter method in which the rate of diffusion is controlled by the image gradient:

$$\begin{cases} \dfrac{\partial I}{\partial t} = \operatorname{div}(c(\|\nabla I\|)\nabla I) \\ I(u, v, 0) = I_0(u, v) \end{cases} \tag{13}$$

where I_0 and I are the original image and the image over time, respectively, $\operatorname{div}(\cdot)$ denotes the divergence operator and $c(\|\nabla I\|)$ denotes the diffusion coefficient which can be defined as follows (according to the proposal in [27]):

$$c(\|\nabla I\|) = \frac{1}{1 + \left(\frac{\|\nabla I\|}{\lambda}\right)^2} \tag{14}$$

where λ is a constant usually set experimentally or by a function of image noise. Using (14), large diffusion rate performs strong smoothing for small gradient magnitude, while small diffusion rate protects image structures (e.g. edges or lines) in regions with big gradient magnitude.

In this work, we choose the Perona-Malik method [27] for image denoising to preserve image texture information while reducing noise.

B. Robustness to low texture

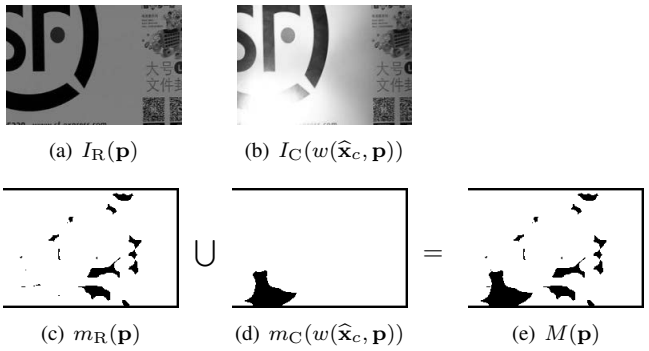
According to (12), GO is meaningless in image areas with low texture where the gradient magnitudes are close to zero. Usually, zeros are assigned to those invalid pixels. This method can be formulated as masking the feature images of GO by a mask image:

$$m(\mathbf{p}) = \begin{cases} 1, & \|\nabla I(\mathbf{p})\| > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

where τ is the threshold of gradient magnitude usually fixed at a small value.

There are two main cases of presenting low texture: the texture of target object inherently is flat (Fig. 2(a)), and the specular reflections blur or damage the texture information of the target object (Fig. 2(b)). In order to handle both of the two cases, the intersection of the mask images is employed (Fig. 2(e)):

$$M(\mathbf{p}) = m_{\mathrm{R}}(\mathbf{p}) \cup m_{\mathrm{C}}(w(\widehat{\mathbf{x}}_c, \mathbf{p})) \tag{16}$$



(a) $I_{\mathrm{R}}(\mathbf{p})$ (b) $I_{\mathrm{C}}(w(\widehat{\mathbf{x}}_c, \mathbf{p}))$ (c) $m_{\mathrm{R}}(\mathbf{p})$ (d) $m_{\mathrm{C}}(w(\widehat{\mathbf{x}}_c, \mathbf{p}))$ (e) $M(\mathbf{p})$

Fig. 2. The mask images. (a) the template; (b) the warped current image; (c) the mask image of the template; (d) the warped mask image of the current image; (e) the proposed mask image.

where the mask image of the template $m_{\mathrm{R}}(\mathbf{p})$ is constant, which can be precomputed, while the warped mask image of the current image $m_{\mathrm{C}}(w(\widehat{\mathbf{x}}_c, \mathbf{p}))$ needs to be updated depending on $\widehat{\mathbf{x}}_c$. Fig. 2(b) also shows that illumination variations may produce confusing texture information in the low texture areas. From Fig. 2(e), it can be seen that M masks the low texture areas while discarding this confusing texture information.

C. ESM for multi-dimensional features

Let $\mathcal{F} \in \mathbb{R}^{n\times m\times k}$ be the multi-dimensional feature images of image $I \in \mathbb{R}^{n\times m}$, where k is the total number of feature dimensions. Let superscript i be the i^{th} dimension of image features, then $\mathcal{F}^i(\mathbf{P})$ denotes the pixel $\mathbf{p}$'s feature value of the i^{th} dimension such that $\forall i \in \{1, 2, \ldots, k\}$. In order to handle multi-dimensional features, we redefine the constancy assumption (1), given by:

$$\mathcal{F}^i_{\mathrm{C}}(w(\mathbf{x}, \mathbf{p}_j)) = \mathcal{F}^i_{\mathrm{R}}(\mathbf{p}_j),\ j \in \{1, 2, \ldots, q\} \tag{17}$$

Then the SSD between the multi-dimensional feature images of the template $\mathcal{F}_{\mathrm{R}}(\mathbf{p})$ and the warped multi-dimensional feature images of the current image $\mathcal{F}_{\mathrm{C}}(w(\mathbf{x}, \mathbf{p}))$ can be computed as:

$$\begin{aligned} S_{\mathrm{mSSD}}(\mathbf{x}) &= \sum_{j=1}^{q}\sum_{i=1}^{k}(\mathcal{F}^i_{\mathrm{C}}(w(\mathbf{x}, \mathbf{p}_j)) - \mathcal{F}^i_{\mathrm{R}}(\mathbf{p}_j))^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{q}(\mathcal{F}^i_{\mathrm{C}}(w(\mathbf{x}, \mathbf{p}_j)) - \mathcal{F}^i_{\mathrm{R}}(\mathbf{p}_j))^2 \end{aligned} \tag{18}$$

This generalization of SSD on dimensionality (18) allows the original ESM [3] method to be used on each feature dimension for the Jacobian computations. According to (17), we have:

$$d^i_j(\mathbf{x}) = \mathcal{F}^i_{\mathrm{C}}(w(\mathbf{x}, \mathbf{p}_j)) - \mathcal{F}^i_{\mathrm{R}}(\mathbf{p}_j) = 0 \tag{19}$$

then (5) can be rewritten as:

$$\mathbf{d}(\mathbf{x}) = \begin{bmatrix} \mathbf{d}^1(\mathbf{x})^\top & \mathbf{d}^2(\mathbf{x})^\top & \ldots & \mathbf{d}^k(\mathbf{x})^\top \end{bmatrix}^\top \tag{20}$$

where $\mathbf{d}^i(\mathbf{x})$ is the image feature differences of the i^{th} dimension:

$$\mathbf{d}^i(\mathbf{x}) = \begin{bmatrix} d^i_1(\mathbf{x}) & d^i_2(\mathbf{x}) & \ldots & d^i_q(\mathbf{x}) \end{bmatrix}^\top \tag{21}$$

Let $\mathbf{J}_{\mathbf{d}^i}(\mathbf{x}) \in \mathbb{R}^{q\times p}$ be the Jacobian matrix of $\mathbf{d}^i(\mathbf{x})$ and $\mathbf{J}_{\mathrm{R}^i} \in \mathbb{R}^{q\times p}$ be the Jacobian matrix of $\mathcal{F}^i_{\mathrm{R}}(w(\mathbf{x}, \mathbf{p}))$ at $\mathbf{x} = \mathbf{e}$. Accordingly, $\mathbf{J}_{\mathbf{d}}(\mathbf{x})$ and $\mathbf{J}_{\mathrm{R}}$ can be computed, respectively, given by:

$$\mathbf{J}_{\mathbf{d}}(\mathbf{x}) = \begin{bmatrix} \mathbf{J}_{\mathbf{d}^1}(\mathbf{x})^\top & \mathbf{J}_{\mathbf{d}^2}(\mathbf{x})^\top & \ldots & \mathbf{J}_{\mathbf{d}^k}(\mathbf{x})^\top \end{bmatrix}^\top \tag{22}$$

and

$$\mathbf{J}_{\mathrm{R}} = \begin{bmatrix} \mathbf{J}_{\mathrm{R}^1}{}^\top & \mathbf{J}_{\mathrm{R}^2}{}^\top & \ldots & \mathbf{J}_{\mathrm{R}^k}{}^\top \end{bmatrix}^\top \tag{23}$$

Then, using the constancy assumption (17), we also have:

$$\mathbf{d}^i(\widehat{\mathbf{x}}_c \circ \Delta\mathbf{x}) \approx \mathbf{0} \tag{24}$$

and

$$\mathbf{J}_{\mathbf{d}^i}(\widehat{\mathbf{x}}_c \circ \Delta\mathbf{x}) \approx \mathbf{J}_{\mathrm{R}^i} \tag{25}$$

Algorithm 1. The proposed GO-ESM method

Input: the reference image I_{R}, threshold ε, maximum iteration $\bar{l}$
1: denoise I_{R}, c.f. Sec. III-A
2: compute $\mathcal{F}_{\mathrm{R}}(\mathbf{p})$ and $m_{\mathrm{R}}(\mathbf{p})$ via (12) and (15), respectively
3: compute $\mathbf{J}_{\mathrm{R}}$ via (23)
4: **for each** new image I_{C} **do**
5: set number of iterations $l = 0$
6: warp *all* pixels of I_{C} using $\widehat{\mathbf{x}}_0$ and denoise it, c.f. Sec. III-A
7: compute $\mathcal{F}_{\mathrm{C}}$ and m_{C} via (12) and (15), respectively
8: **repeat**
9: warp $\mathcal{F}_{\mathrm{C}}$ and m_{C} using $\widehat{\mathbf{x}}_l$
10: compute $M(w(\widehat{\mathbf{x}}_l, \mathbf{p}))$ via (16)
11: compute $\mathbf{J}(\widehat{\mathbf{x}}_l)$ and $\Delta\mathbf{x}$ via (22) and (10), respectively
12: update $\widehat{\mathbf{x}}_{l+1} = \widehat{\mathbf{x}}_l \circ \Delta\mathbf{x}$ and $l = l + 1$
13: **until** $\|\Delta\mathbf{x}\| < \varepsilon$ or $l > \bar{l}$
14: **end for**

Equations (24) and (25) assure that the ESM update law (10) can be applied to multi-dimensional image features such as GO.

D. Summary of the GO-ESM

Algorithm 1 summarizes the proposed DVT method with Fig. 1 showing its overview. Some implementation details are highlighted as follows:

- To reduce effect of changes in orientations of gradient vectors resulting from image rotation, we warp all pixels of the current image I_{C} using the initial estimation of the transformation parameters $\widehat{\mathbf{x}}_0$ before extracting GO.
- For each new image I_{C}, both the extraction of GO and the computation of m_{C} are executed only once outside the iteration, and then the warp function is directly operated on the GO images $\mathcal{F}_{\mathrm{C}}$ or the mask image m_{C} using the current estimation $\widehat{\mathbf{x}}_l$ at the $(l+1)^{th}$ iteration.
- In this work, we set $\lambda = 5$ for the diffusion coefficient function (14). Besides, the threshold of gradient magnitude τ is set to 0.005 from experience.

IV. Experimental Results

To evaluate the robustness of the proposed method to illumination changes as well as to general tracking challenges, the GO-ESM is tested on two datasets: our POIC dataset and the TMT benchmark dataset [28] with ground truth. According to the Modular Tracking Framework (MTF) [30] (which decomposes a registration based tracker into three sub modules - appearance model (AM), state space model (SSM) and search method (SM)), the GO-ESM is compared against trackers with ESM [3] AM and $\mathbb{SL}(3)$ Homography SSM using the state-of-the-art SMs - NCC [13], SCV [14], LSCV [15], MI [16], [17] and CCRE [19], [20]. In addition to these, the original SSD-based ESM [3] and the ESM with a gain and bias (GB) illumination compensation model [9] are also included into the comparison.

All results are generated using a fixed sampling resolution of 100×100. For all trackers, we set $\varepsilon = 0.01$, and use maximum number of iterations of 200 and 30 for our POIC dataset and the TMT benchmark dataset, respectively. All tests are run on an Intel i5 3.3GHz PC with 8G RAM.

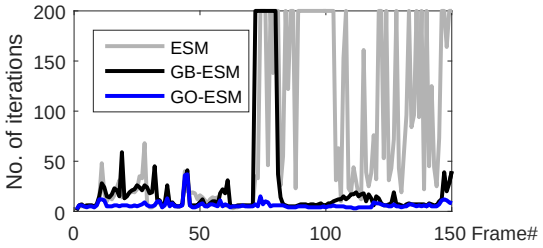


Fig. 3. The numbers of iterations for ESM [3], GB-ESM [9] and GO-ESM to converge on sequence BOOK [12] (only show the first 150 frames). For the first 60 frames, the ESM, GB-ESM and GO-ESM take in average number of iterations of 14.22, 14.85, 7.25, respectively, to converge. Note that ESM and GB-ESM lose track at frame #72 and #147, respectively.

A non-optimized implementation in Matlab of the GO-ESM runs at about 15.8ms/iteration for an image region of 100×100, which is about 1.7 times slower than the original ESM [3] running under the same condition. Although the proposed method is somewhat more complex, it is more robust and need fewer iterations than the ESM and GB-ESM when severe illumination changes occur (see Fig. 3).

A. Effectiveness of the proposed denoising method

Fig. 4 shows some tracking results using the GO-ESM with and without denoise processing on BOOK and BEAR sequences [12]. It can be seen that the performance with denoise processing is superior than without denoise processing, and the Perona-Malik method [27] is superior to the Gauss.

B. Evaluation on the POIC dataset

To evaluate the performance of GO-ESM with difficult illumination environments, experiments were made on our POIC dataset which contains ten video sequences with a total of 6,663 frames. Among the ten sequences, BEAR and BOOK sequences are from reference [12], while the other eight sequences were recorded by the author. The objects with varying texture and lambertian/specular materials were chosen to have a full spectrum of challenges. In addition, our POIC dataset presents extreme real-world situations with various types of illumination changes. All sequences are resized into 600×800 pixels for test.

Fig. 5 shows some key frames from our method together with the other seven methods compared. Clearly, GO-ESM outperforms all the other methods, indicating its robustness to kinds of illumination variations, including linear and non-linear changes.

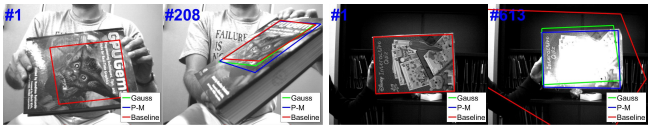


Fig. 4. Comparison of the denoising methods. Gauss and P-M denote the Gauss filter and Perona-Malik method [27], respectively. The baseline is obtained without any denoise processing. From left to right, samples are from BOOK and BEAR sequences [12].

1) Diffuse and specular reflections and inter reflection: Though the LSCV, MI and GO-ESM completed the notebook sequence with slight diffuse and specular reflections, the LSCV and MI had intermittent failures during tracking resulting from specular reflections (e.g., #109, and #159), while GO-ESM succeeded for all frames. For the BEAR sequence where the target underwent severe specular reflections, the GB-ESM, NCC, LSCV and GO-ESM completed the entire sequence. However, all the methods except GO-ESM suffered a lot of intermittent failures during tracking (e.g., #536, #754 and #1179). In addition, the GO-ESM was the only one to succeed on the boxI sequence, whereas others had different problems caused by severe diffuse and specular reflections (e.g., #308 and #331).

The disk and tea sequences with plastics and metal at surface, respectively, presented inter-reflection (e.g., #192 and #349 for disk, #206 and #303 for tea) and irregular specular reflection (e.g., #84, and #463 for disk, #133 for tea). Again only the GO-ESM fully completed both of them.

The above superior performance of GO-ESM is believed to benefit from GO's insensitivity to kinds of illumination changes. Besides, the robustness of GO-ESM to specular reflections confirms the effectiveness of the proposed mask image M.

2) Shadow: From Fig. 5, the boxII sequence presented large surface obliquity with irregular specular reflection (e.g., #616) and shadows (e.g., #675). Except ESM, GB-ESM and SCV, all the other methods could complete the full sequence. The BOOK sequence suffered from variable illumination and shadows (e.g., #118 and #248). The MI and GO-ESM performed better than other trackers, while the MI failed around frame #248 due to shadows. These comparison results demonstrate that GO-ESM is robust to shadows which can be attributed to the robustness of GO.

3) Global and local illumination changes: All methods accurately tracked magazine sequence with slight global and local illumination changes, except the original ESM [3] (e.g., #364). For the boxIII sequence presenting severe and instantaneous changes of global intensity (e.g., #247 and #572), the GB-ESM, NCC and GO-ESM were able to succeed on it. Contrast to boxIII, the envelope sequence underwent severe and gradual changes both global (e.g., #391 and #415) and local (e.g., #245, #340 and #382) intensity, the NCC and GO-ESM still outperform other methods, while the GB-ESM somehow got stuck from frame #245. The above observations illustrate that while the GB illumination compensation model [9] is not effective for non-linear lighting changes due to the limitation of the model itself, our method does not have this limitation.

C. Evaluation on the TMT Benchmark Dataset

To illustrate the GO-ESM's robustness to general tracking challenges, we choose the TMT benchmark dataset [28] for the test. This choice is motivated by that the TMT benchmark covers a wider range of challenges [28] (see Fig. 8) and is more suitable for the 2D planar object tracking than most of the other publicly reported datasets (e.g. [31] dose

Fig. 5. Comparison of tracking results for ESM [3], GB-ESM [9], NCC [13], SCV [14], LSCV [15], MI [16], [17], CCRE [19], [20] and GO-ESM on the POIC dataset. For all the ten full video of tracking results, one can contact the authors.

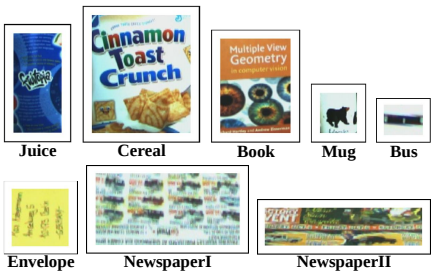


Fig. 6. Planar projections of the objects used in TMT [28].

not consider perspective challenge and only covers 2 DOF transformation). Besides, the TMT uses eight different target objects (shown in Fig. 6) and has 109 sequences with a total of 70,592 frames.

The object positions estimated by the tracker were compared to the ground truth in **Alignment Error** (E_{AL}) [32], and the tracking was considered successful if E_{AL} was less than a threshold t_p which was set to 10 through all tests. Accuracy of a tracker was measured through its **success rate** (SR):

$$\mathrm{SR} = \frac{|S|}{|F|}, \; S = \{f^{(i)} \in F : E_{\mathrm{AL}}^{(i)} < t_p\} \tag{26}$$

where F is the set of all frames and $E_{\mathrm{AL}}^{(i)}$ is the error in the i^{th} frame $f^{(i)}$ [30].

Fig. 7 plots for the overall success rate (SR) of the eight comparing methods. It can be noted that the GO-ESM ranks first among all trackers, while NCC ranks second better than SCV which is slightly superior to LSCV. These observations are in agreement with the reports from reference [30].

Table I shows the SR scores of all methods for each object as well as for the overall TMT dataset, with the best results marked in red and the second best in blue. It can be seen that GO-ESM achieves the best for the entire dataset and ranks among top two for five objects. Although the NCC ranks among top two for seven objects, which seems to be superior to GO-ESM, its SR scores are 13% and 9% less than GO-ESM for newspaperI and book objects, respectively. While GO-ESM's SR scores are only less than NCC in a small percent, e.g. 2% for cereal object. In other words, our method is more consistent and robust than NCC for different objects.

Fig. 8 shows the SR plots of each method on eight tracking challenges. It can be seen that for all tracking challenges GO-ESM ranks among top tree except for perspective deformation (PR). In the case of occlusion (OC) only GO-ESM achieves good performance (Fig. 8(e)), while for low texture (TX) the GB-ESM and GO-ESM are significantly better than the others (see Fig. 8(g)), which benefits from the advantages of the proposed mask image.

It is worth noting that although the NCC is only slightly worse performance than GO-ESM on TMT benchmark, it is much inferior than the proposed method on our POIC dataset (see Fig. 5). These results of the evaluation verify that GO-ESM is not only particularly robust to illumination changes but also robust against general challenges in visual tracking with competitive tracking performance.

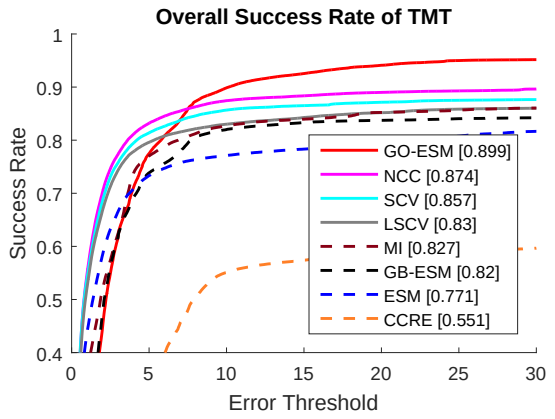


Fig. 7. Comparison of the overall success rate of the TMT [28] benchmark dataset (the legends are ranked by SR scores on all 109 sequences).

V. Conclusion

We have proposed an illumination-insensitive ESM for planar object tracking by introducing GO into DVT problem. To address image noise and low texture challenges, the Perona-Malik method [27] and mask images were suggested. In addition, we generalized the original ESM method to combine GO features with ESM method. The results of the experimental test on both our POIC dataset and the benchmark dataset achieved excellent performance, which clearly indicate robustness of the proposed method to illumination changes and to general tracking challenges.

Acknowledgment

We'd like to thank Rogerio Richa for his help with the code. Also, we'd like to thank Ankush Roy et al. from University of Alberta for their open source tracking framework and tracking benchmark.

References

[1] A. I. Comport, E. Malis, and P. Rives, "Real-time quadrifocal visual odometry," *The International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 245–266, 2010.

[2] C. Kerl, J. Sturm, and D. Cremers, "Robust odometry estimation for rgb-d cameras," in *ICRA*, 2013, pp. 3748–3754.

[3] S. Benhimane and E. Malis, "Homography-based 2d visual tracking and servoing," *The International Journal of Robotics Research*, vol. 26, no. 7, pp. 661–676, 2007.

[4] H. Lim, J. Lim, and H. J. Kim, "Real-time 6-dof monocular visual slam in a large-scale environment," in *ICRA*, 2014, pp. 1532–1539.

[5] S. Benhimane and E. Malis, "Real-time image-based tracking of planes using efficient second-order minimization," in *IROS*, vol. 1. IEEE, 2004, pp. 943–948.

[6] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision." in *IJCAI*, vol. 81, 1981, pp. 674–679.

[7] H. Jin, P. Favaro, and S. Soatto, "A semi-direct approach to structure from motion," *The Visual Computer*, vol. 19, no. 6, pp. 377–394, 2003.

[8] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International journal of computer vision*, vol. 56, no. 3, pp. 221–255, 2004.

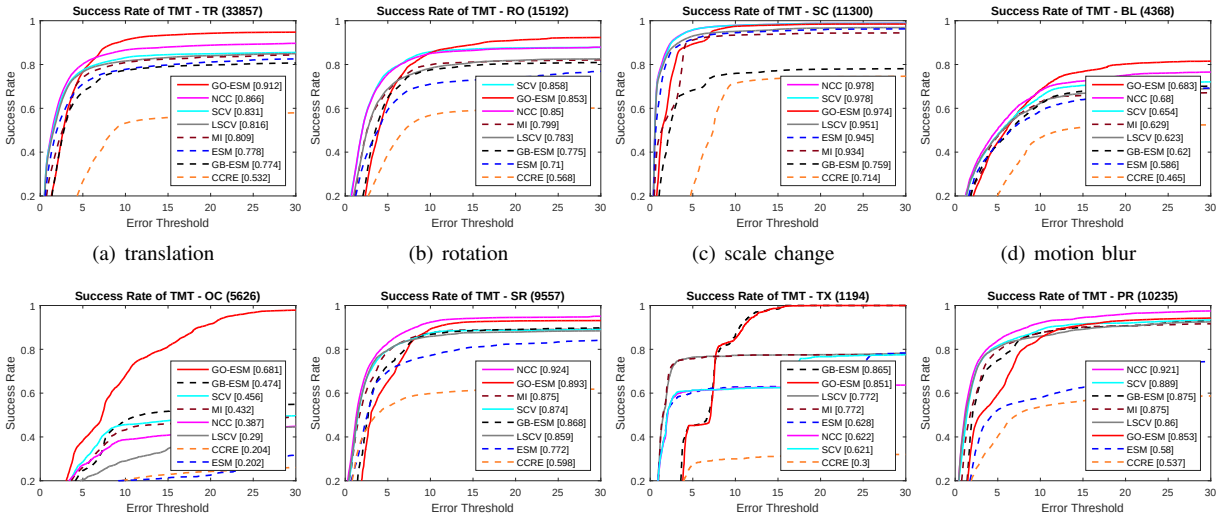
[9] A. Bartoli, "Direct image registration with gain and bias," *Contributions au recalage dimages et la reconstruction 3D de scnes rigides et dformables*, 2006: 4.

TABLE I

OVERALL SUCCESS RATE OF TMT [28]

Object	ESM	GB-ESM	NCC	SCV	LSCV	MI	CCRE	GO-ESM
Juice (13)	0.90	0.85	0.89	0.85	0.85	0.85	0.81	0.87
Cereal(13)	0.90	0.87	0.91	0.87	0.87	0.88	0.86	0.89
Book (36)	0.79	0.77	0.86	0.82	0.82	0.85	0.74	0.95
Mug (39)	0.67	0.82	0.85	0.87	0.79	0.75	0.16	0.82
Bus (2)	0.70	0.75	1	1	0.74	0.58	N/A	1
envelope (2)	0.34	1	1	1	1	0.86	0.03	0.94
NewspaperI (2)	0.88	0.99	0.87	0.82	0.94	0.94	0.50	1
NewspaperII (2)	0.99	0.99	1	1	1	1	0.55	0.99
overall (109)	0.77	0.82	0.87	0.86	0.83	0.83	0.55	0.90

N/A = There were no frames that had E_{AL} less than t_p



(a) translation (b) rotation (c) scale change (d) motion blur

(e) occlusion (f) specular reflection (g) low texture (h) perspective deformation

Fig. 8. Plots of success rate over eight tracking challenges. Total number of frames for each challenges is shown on top of each plot while the legends are arranged by SR scores.

[10] ——, "Groupwise geometric and photometric direct image registration," *TPAMI*, vol. 30, no. 12, pp. 2098–2108, 2008.

[11] G. Silveira and E. Malis, "Real-time visual tracking under arbitrary illumination changes," in *CVPR*. IEEE, 2007, pp. 1–6.

[12] ——, "Unified direct visual tracking of rigid and deformable surfaces under generic illumination changes in grayscale and color images," *International journal of computer vision*, vol. 89, no. 1, pp. 84–105, 2010.

[13] G. G. Scandaroli, M. Meilland, and R. Richa, "Improving ncc-based direct visual tracking," in *ECCV*. Springer, 2012, pp. 442–455.

[14] R. Richa, R. Sznitman, R. Taylor, and G. Hager, "Visual tracking using the sum of conditional variance," in *IROS*. IEEE, 2011, pp. 2953–2958.

[15] R. Richa, M. Souza, G. Scandaroli, E. Comunello, and A. Von Wangenheim, "Direct visual tracking under extreme illumination variations using the sum of conditional variance," in *ICIP*. IEEE, 2014, pp. 373–377.

[16] N. Dowson and R. Bowden, "Mutual information for lucas-kanade tracking (milk): An inverse compositional formulation," *TPAMI*, vol. 30, no. 1, pp. 180–185, 2008.

[17] A. Dame and E. Marchand, "Accurate real-time tracking using mutual information," in *ISMAR*. IEEE, 2010, pp. 47–56.

[18] ——, "Second-order optimization of mutual information for real-time image registration," *Image Processing, IEEE Transactions on*, vol. 21, no. 9, pp. 4190–4203, 2012.

[19] F. Wang and B. C. Vemuri, "Non-rigid multi-modal image registration using cross-cumulative residual entropy," *International journal of computer vision*, vol. 74, no. 2, pp. 201–215, 2007.

[20] R. Richa, R. Sznitman, and G. Hager, "Robust similarity measures for gradient-based direct visual tracking," *CIRL Technical report*, 2012.

[21] P.-Y. Burgi, "Motion estimation based on the direction of intensity gradient," *Image and Vision Computing*, vol. 22, no. 8, pp. 637–653, 2004.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[23] H. F. Chen, P. N. Belhumeur, and D. W. Jacobs, "In search of illumination invariants," in *CVPR*, vol. 1. IEEE, 2000, pp. 254–261.

[24] N. Baha and S. Larabi, "Accurate real-time neural disparity map estimation with fpga," *Pattern Recognition*, vol. 45, no. 3, pp. 1195–1204, 2012.

[25] T. Kondo, "Motion estimation using gradient orientation structure tensors," in *ICICIC*, 2007, pp. 450–450.

[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1. IEEE, 2005, pp. 886–893.

[27] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *TPAMI*, vol. 12, no. 7, pp. 629–639, 1990.

[28] A. Roy, X. Zhang, N. Wolleb, C. Perez Quintero, and M. Jagersand, "Tracking benchmark and evaluation for manipulation tasks," in *ICRA*. IEEE, 2015, pp. 2448–2453.

[29] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science and Business Media, 2006.

[30] A. Singh, A. Roy, X. Zhang, and M. Jagersand, "Modular decomposition and analysis of registration based trackers," *arXiv preprint arXiv:1603.01292*, 2016.

[31] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *CVPR*, 2013, pp. 2411–2418.

[32] S. Lieberknecht, S. Benhimane, P. Meier, and N. Navab, "A dataset and evaluation methodology for template-based tracking algorithms," in *ISMAR*. IEEE, 2009, pp. 145–151.