# Improving NCC-Based Direct Visual Tracking

Glauco Garcia Scandaroli[1,⋆], Maxime Meilland[2], and Rogério Richa[3]

[1] AROLAG, INRIA Sophia Antipolis-Méditerranée
Glauco.Scandaroli@inria.fr
[2] CNRS-I3S, Université de Nice Sophia-Antipolis
[3] LCSR, Johns Hopkins University

**Abstract.** Direct visual tracking can be impaired by changes in illumination if the right choice of similarity function and photometric model is not made. Tracking using the sum of squared differences, for instance, often needs to be coupled with a photometric model to mitigate illumination changes. More sophisticated similarities, *e.g.* mutual information and cross cumulative residual entropy, however, can cope with complex illumination variations at the cost of a reduction of the convergence radius, and an increase of the computational effort. In this context, the normalized cross correlation (NCC) represents an interesting alternative. The NCC is intrinsically invariant to affine illumination changes, and also presents low computational cost. This article proposes a new direct visual tracking method based on the NCC. Two techniques have been developed to improve the robustness to complex illumination variations and partial occlusions. These techniques are based on subregion clusterization, and weighting by a residue invariant to affine illumination changes. The last contribution is an efficient Newton-style optimization procedure that does not require the explicit computation of the Hessian. The proposed method is compared against the state of the art using a benchmark database with ground-truth, as well as real-world sequences.

## 1 Introduction

Direct visual tracking (DVT) can be defined as the problem of finding the transformation parameters that best align a reference image to the following frames in a video stream. This is a basic task that must be solved in different computer vision problems, *e.g.* mosaicking [1], visual odometry [2], SLAM [3]. In contrast to feature-based methods, which are built on the extraction and matching of a sparse set of characteristics from the image [4], direct visual tracking methods exploit each individual pixel's intensity to solve the visual tracking. In this work, we focus on the latter class of techniques.

In DVT methods, the quality of the match between two images is measured by a similarity function. Typically, DVT methods are built upon the sum of squared differences (SSD), which usually assumes brightness constancy.

⋆ Glauco Garcia Scandaroli is funded by grants from the *Conseil Régional Provence-Alpes-Côte d'Azur*, and the *DGCIS* Rapace project leaded by Geocean.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VI, LNCS 7577, pp. 442–455, 2012.

The SSD has proven to be very efficient, mainly because the optimization can be much simplified due to numerous solutions to nonlinear least squares [5,6]. SSD tracking, however, is severely impaired when brightness constancy is violated, since motion and photometric variations are dealt in the same way by the similarity function. Different works improve the SSD-based tracking by estimating online [7] or offline [8] photometric parameters, and the problem of partial occlusion is usually treated by a robustly weighted SSD [8]. Nevertheless, as we show in this paper, the SSD does not perform well under concurrent illumination changes and partial occlusions.

More general examples of similarity measures include the mutual information [9], and cross cumulative residual entropy [10]. These similarities have also been applied to DVT, and they relax the brightness constancy to more complex photometric variations. The counterpart is an overwhelming increase of the complexity of the solution, and the reduction on the radius of convergence. A review on different similarity measures and their respective MATLAB implementation is available in [11]. The normalized cross correlation (NCC) is a similarity measure invariant to affine illumination changes, which is simpler than those solutions and whose radius of convergence is comparable to the SSD. Typically, gradient-based solutions for the NCC have to resort to computationally expensive Newton's method [12], or approximations [13] that do not imply a well defined optimization (*c.f.* §3).

We propose a novel DVT solution using the NCC as similarity measure. This similarity is chosen because of its simplicity. It is intrinsically invariant to affine illumination changes, which is a powerful characteristic that allied with two techniques here presented can improve the robustness to nonlinear illumination and partial occlusion. These techniques are based on subregion partitioning, and weighting using a residue invariant to affine illumination variations. We also propose a method to improve the gradient solution while having a well defined optimization problem.

The proposed method is extensively tested and compared against other state of the art methods using a benchmark dataset [14] and challenging real-world video sequences. The obtained results show that our method is suited for visual tracking under complex illumination variation, and tracking can still be performed for partially occluded target under extreme illumination settings.

## 2 Notations and Background

The matrix $\boldsymbol{I} \in \mathbb{R}^{m \times n}$ denotes an image with $m$ rows and $n$ columns, and each element of $\boldsymbol{I}$ represents a pixel intensity. Let the point $p = (u, v) \in \mathbb{R}^2$, then the function $\boldsymbol{I}(p) \colon \mathbb{R}^2 \mapsto \mathbb{R}$ maps a point $p$ to the respective intensity in $\boldsymbol{I}$, which is obtained using interpolation when $u$ and $v$ are not integers. A *warp* function $w(\mathbf{x}, p) \colon \mathcal{X} \times \mathbb{R}^2 \mapsto \mathbb{R}^2$ defines a differentiable transformation with $\mathbf{x} \in \mathcal{X}$, with $\mathcal{X}$ being a Lie Group of dimension $N$, and, in this work, we consider warps that satisfy $w\big(\mathbf{x}_1, w(\mathbf{x}_2, p)\big) = w(\mathbf{x}_1 \circ \mathbf{x}_2, p)$, and $w\big(\mathbf{x}_1, w({\mathbf{x}_1}^{-1}, p)\big) = w\big(\mathbf{x}_1 \circ {\mathbf{x}_1}^{-1}, p\big)\big) = p$. An example is given by planar homographic warps, which

can be written using the $\mathbb{SL}(3)$ [6]. The warp $w$ maps pixel positions from the *reference image* $\boldsymbol{I}_{\mathrm{R}}(p)$ of a target object to a *current image* $\boldsymbol{I}_{\mathrm{C}}(p)$ later obtained for this object. We denote $\boldsymbol{I}(\mathbf{x})$ as the resulting image after warping *all* pixels of $\boldsymbol{I}(w(\mathbf{x}, p))$.

### 2.1 Illumination Changes

According to the Blinn–Phong model, the illumination over a surface can be modeled by components of specular $\gamma_s$, diffuse $\gamma_d$ and ambient $\gamma_a$ reflections:

$$\boldsymbol{I}_{\mathrm{C}} = \gamma_s(\boldsymbol{I}_{\mathrm{R}}) + \gamma_d(\boldsymbol{I}_{\mathrm{R}}) + \gamma_a(\boldsymbol{I}_{\mathrm{R}})\,.$$

These functions are in general nonlinear, and also depend on the object's material and camera viewpoint. The illumination model, however, is simplified if we assume a planar and Lambertian surface and neglecting specular reflections. Considering such simplifications, the photometric model is reduced to an affine transformation

$$\boldsymbol{I}_{\mathrm{C}} = \alpha\boldsymbol{I}_{\mathrm{R}} + \beta\,, \tag{1}$$

where $\alpha, \beta \in \mathbb{R}$. This approximation holds, at least locally, for most applications.

### 2.2 Normalized Cross–Correlation

The Normalized Cross–Correlation (NCC) is defined as:

$$\mathcal{N}_{\times}\big(\boldsymbol{I}_{\mathrm{R}}, \boldsymbol{I}_{\mathrm{C}}\big) = \frac{\sum_i \big(\boldsymbol{I}_{\mathrm{R}}(p_i) - I_{\mathrm{R}}\big)\big(\boldsymbol{I}_{\mathrm{C}}(p_i) - I_{\mathrm{C}}\big)}{\sqrt{\sum_i \big(\boldsymbol{I}_{\mathrm{R}}(p_i) - I_{\mathrm{R}}\big)^2}\sqrt{\sum_i \big(\boldsymbol{I}_{\mathrm{C}}(p_i) - I_{\mathrm{C}}\big)^2}}\,,$$

where the correlation coefficient $\mathcal{N}_{\times} \in [-1, 1]$, and $I_{\mathrm{R}} = \frac{1}{mn}\sum_p \boldsymbol{I}_{\mathrm{R}}(p)$, $I_{\mathrm{C}} = \frac{1}{mn}\sum_p \boldsymbol{I}_{\mathrm{C}}(p)$ represent the mean of $\boldsymbol{I}_{\mathrm{R}}$, $\boldsymbol{I}_{\mathrm{C}}$ pixel intensities, respectively. We can also write the NCC using vector notation. It follows directly that:

$$\mathcal{N}_{\times}\big(\boldsymbol{I}_{\mathrm{R}}, \boldsymbol{I}_{\mathrm{C}}\big) = \frac{\mathbf{i}_{\mathrm{R}}{}^{T}\mathbf{i}_{\mathrm{C}}}{|\mathbf{i}_{\mathrm{R}}||\mathbf{i}_{\mathrm{C}}|}\,. \tag{2}$$

where $\mathbf{i}_{\mathrm{R}}$, and $\mathbf{i}_{\mathrm{C}}$ are the vectors obtained by stacking the intensities of $\boldsymbol{I}_{\mathrm{R}}$ and $\boldsymbol{I}_{\mathrm{C}}$, respectively, such that the $i$-th element of each vector writes $\mathbf{i}_{\mathrm{R}(i)} = \boldsymbol{I}_{\mathrm{R}}(p_i) - I_{\mathrm{R}}$, and $\mathbf{i}_{\mathrm{C}(i)} = \boldsymbol{I}_{\mathrm{C}}(p_i) - I_{\mathrm{C}}$. Notice that this representation yields an easier interpretation of the NCC.

The correlation coefficient $\mathcal{N}_{\times}\big(\boldsymbol{I}_{\mathrm{R}}, \boldsymbol{I}_{\mathrm{C}}\big) = 0$ implies that the vectors $\mathbf{i}_{\mathrm{R}}$ and $\mathbf{i}_{\mathrm{C}}$ are *orthogonal*, thus the images share no information. Furthermore, a coefficient $\mathcal{N}_{\times}\big(\boldsymbol{I}_{\mathrm{R}}, \boldsymbol{I}_{\mathrm{C}}\big) = 1$ implies that the vectors are *parallel*, therefore the images are perfectly aligned. Recalling inner product properties, we have that the correlation remains unaffected after any shift and/or (positive) scale. Note that the absolute value of NCC remains the same after a negative scaling, however, the sign of the resulting correlation coefficient is inverted. Scales and shifts on $\mathbf{i}_{\mathrm{C}}$ are directly related to illumination variations, *i.e.* $\alpha$ and $\beta$ of photometric model (1). Invariance to such effects is indeed a good property for a similarity measure.

### 2.3 Direct Visual Tracking

The direct visual tracking (DVT) problem corresponds to finding the parameters $\mathbf{x}$ that yield a current image $\boldsymbol{I}_{\mathrm{C}}\big(w(\mathbf{x},p)\big)=\boldsymbol{I}_{\mathrm{C}}(\mathbf{x})$ to best match a given reference $\boldsymbol{I}_{\mathrm{R}}$. Similarity functions $\mathcal{S}\big(\boldsymbol{I}_{\mathrm{R}},\boldsymbol{I}_{\mathrm{C}}\big(\mathbf{x},p)\big)$ are scores of how good is the matching between two images. The DVT writes the following optimization for some $\mathcal{S}$ that is maximized when the images are best matched:

$$\widehat{\mathbf{x}} = \arg\max_{\mathbf{x}} \mathcal{S}\big(\boldsymbol{I}_{\mathrm{R}},\boldsymbol{I}_{\mathrm{C}}(\mathbf{x})\big)\,. \tag{3}$$

We identify from now on the group element $\mathbf{x}\in\mathcal{X}$ with $\varphi(\mathbf{x})\in\mathbb{R}^N$, where $\varphi$ is a minimal parametrization sending the identity element in $\mathcal{X}$ to the null vector $\mathbf{0}\in\mathbb{R}^N$. Such a parametrization is given, *e.g.*, by $\varphi=\log$. The optimization (3) is not globally concave, therefore finding the optimal solution is not an easy task. Gradient-based optimization has been widely employed to solve the problem, since this method presents a good trade between region of convergence and computational cost. The solution of (3) can be given by Newton's method [15], *i.e.*, the function $\mathcal{S}\big(\boldsymbol{I}_{\mathrm{R}},\boldsymbol{I}_{\mathrm{C}}(\mathbf{x})\big)$ is approximated by a parabola around $\widehat{\mathbf{x}}_0$:

$$\begin{aligned}\mathcal{S}\big(\boldsymbol{I}_{\mathrm{R}},\boldsymbol{I}_{\mathrm{C}}(\mathbf{x})\big) \approx{}& \mathcal{S}\big(\boldsymbol{I}_{\mathrm{R}},\boldsymbol{I}_{\mathrm{C}}(\widehat{\mathbf{x}}_0)\big)\\ &+\frac{\partial\mathcal{S}(\boldsymbol{I}_{\mathrm{R}},\boldsymbol{I}_{\mathrm{C}}(\mathbf{x}))}{\partial\mathbf{x}}\Big|_{\mathbf{x}=\widehat{\mathbf{x}}_0}\widetilde{\mathbf{x}}+\frac{1}{2}\widetilde{\mathbf{x}}^T\frac{\partial^2\mathcal{S}(\boldsymbol{I}_{\mathrm{R}},\boldsymbol{I}_{\mathrm{C}}(\mathbf{x}))}{\partial\mathbf{x}^T\partial\mathbf{x}}\Big|_{\mathbf{x}=\widehat{\mathbf{x}}_0}\widetilde{\mathbf{x}}\,,\end{aligned}$$

where $\widetilde{\mathbf{x}}=\widehat{\mathbf{x}}_0{}^{-1}\circ\mathbf{x}$ and the maximum is obtained at $\partial_{\widetilde{\mathbf{x}}}\mathcal{S}=0$:

$$\widetilde{\mathbf{x}}^* = -\left(\frac{\partial^2\mathcal{S}(\boldsymbol{I}_{\mathrm{R}},\boldsymbol{I}_{\mathrm{C}}(\mathbf{x}))}{\partial\mathbf{x}^T\partial\mathbf{x}}\Big|_{\mathbf{x}=\widehat{\mathbf{x}}_0}\right)^{-1}\frac{\partial\mathcal{S}(\boldsymbol{I}_{\mathrm{R}},\boldsymbol{I}_{\mathrm{C}}(\mathbf{x}))}{\partial\mathbf{x}}^T\Big|_{\mathbf{x}=\widehat{\mathbf{x}}_0}. \tag{4}$$

Note that an increment $\widetilde{\mathbf{x}}$ is obtained, and the solution is given by $\widehat{\mathbf{x}}=\widehat{\mathbf{x}}_0\circ\widetilde{\mathbf{x}}^*$. Newton's method converges in one iteration for quadratic functions. For non-quadratic functions, however, the solution is obtained computing (4) iteratively, until the increment $\widetilde{\mathbf{x}}$ is conveniently small, *i.e.*, $|\widetilde{\mathbf{x}}|<\varepsilon$. According to the classification of Baker and Matthews [5], Eq. (4) yields a *forward compositional* method. Some similarity measures can be computed efficiently inverting the roles of reference and current images, *i.e.* computing the solution for $\mathcal{S}\big(\boldsymbol{I}_{\mathrm{R}}(\mathbf{x}\circ\widetilde{\mathbf{x}}^{-1}),\boldsymbol{I}_{\mathrm{C}}(\mathbf{x})\big)$. This latter procedure yields an *inverse compositional*, and it can be solved as (4).

## 3 Visual Tracking Using NCC as Similarity Measure

Recall the NCC vector formulation (2) for two images $\boldsymbol{I}_{\mathrm{R}}$, $\boldsymbol{I}_{\mathrm{C}}(\mathbf{x})$:

$$\mathcal{N}_{\times}\big(\boldsymbol{I}_{\mathrm{R}},\boldsymbol{I}_{\mathrm{C}}(\mathbf{x})\big)=\frac{\mathbf{i}_{\mathrm{R}}{}^T\mathbf{i}_{\mathrm{C}}(\mathbf{x})}{|\mathbf{i}_{\mathrm{R}}||\mathbf{i}_{\mathrm{C}}(\mathbf{x})|}\,.$$

For the sake of simplicity, we denote $\mathcal{N}_{\times}\big(\boldsymbol{I}_{\mathrm{R}},\boldsymbol{I}_{\mathrm{C}}(\mathbf{x})\big)$ by its shorter $\mathcal{N}_{\times}$, and recall that $|\mathbf{v}|=\sqrt{\mathbf{v}^T\mathbf{v}}$, and $\partial_{\mathbf{u}}\frac{1}{|\mathbf{v}(\mathbf{u})|}=\frac{-\mathbf{v}(\mathbf{u})^T}{|\mathbf{v}(\mathbf{u})|^3}\partial_{\mathbf{u}}\mathbf{v}(\mathbf{u})$. We obtain the gradient of $\mathcal{N}_{\times}$ as:

$$\frac{\partial\mathcal{N}_{\times}}{\partial\mathbf{x}}=\mathbf{g}_{\mathrm{FC}}(\mathbf{x})=\left(\frac{\mathbf{i}_{\mathrm{R}}}{|\mathbf{i}_{\mathrm{R}}|}-\mathcal{N}_{\times}\frac{\mathbf{i}_{\mathrm{C}}(\mathbf{x})}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|}\right)^T\frac{\mathbf{J}_{\mathrm{C}}^{\times}(\mathbf{x})}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|}\,, \tag{5}$$

where $\mathbf{J}_{\mathrm{C}}^{\times}(\mathbf{x})$ is obtained after stacking the Jacobians of $\boldsymbol{I}_{\mathrm{C}}$, and the $i$-th element writes $\mathbf{J}_{\mathrm{C}(i)}^{\times}(\mathbf{x}) = \mathbf{J}_{\mathrm{C}(i)}(\mathbf{x}) - \frac{1}{mn}\sum_i \mathbf{J}_{\mathrm{C}(i)}(\mathbf{x})$, with $\mathbf{J}_{\mathrm{C}(i)}(\mathbf{x}) = \frac{\partial \boldsymbol{I}_{\mathrm{C}}(w(\mathbf{x},p_i))}{\partial \mathbf{x}}$ being the image Jacobian as found in [6]. Evaluating the partial derivative w.r.t. $\mathbf{x}^T$ of Eq. (5), we thus obtain the full expression for the Hessian as:

$$\frac{\partial^2 \mathcal{N}_{\times}}{\partial \mathbf{x}^T \partial \mathbf{x}} = -\mathcal{N}_{\times}\frac{\mathbf{J}_{\mathrm{C}}^{\times}(\mathbf{x})^T\mathbf{J}_{\mathrm{C}}^{\times}(\mathbf{x})}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|^2} - \frac{\mathbf{J}_{\mathrm{C}}^{\times}(\mathbf{x})^T}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|}\left(\frac{\mathbf{i}_{\mathrm{R}}\mathbf{i}_{\mathrm{C}}(\mathbf{x})^T}{|\mathbf{i}_{\mathrm{R}}||\mathbf{i}_{\mathrm{C}}(\mathbf{x})|} + \frac{\mathbf{i}_{\mathrm{C}}(\mathbf{x})\mathbf{i}_{\mathrm{R}}{}^T}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})||\mathbf{i}_{\mathrm{R}}|}\right)\frac{\mathbf{J}_{\mathrm{C}}^{\times}(\mathbf{x})}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|}$$
$$+ 3\frac{\mathbf{J}_{\mathrm{C}}^{\times}(\mathbf{x})^T}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|}\frac{\mathbf{i}_{\mathrm{C}}(\mathbf{x})\mathbf{i}_{\mathrm{C}}(\mathbf{x})^T}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|^2}\frac{\mathbf{J}_{\mathrm{C}}^{\times}(\mathbf{x})}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|} + \sum_{i=0}^{mn}\frac{\mathbf{H}_{\mathrm{C}(i)}^{\times}(\mathbf{x})}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|}\left(\frac{\mathbf{i}_{\mathrm{R}(i)}}{|\mathbf{i}_{\mathrm{R}}|} - \mathcal{N}_{\times}\frac{\mathbf{i}_{\mathrm{C}(i)}(\mathbf{x})}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|}\right), \quad (6)$$

where $\mathbf{H}_{\mathrm{C}(i)}^{\times}(\mathbf{x}) = \partial_{\mathbf{x}^T}\mathbf{J}_{\mathrm{C}(i)}^{\times}(\mathbf{x})$. Computing (6) at each iteration is not an easy task mainly because $\mathbf{H}_{\mathrm{C}(i)}^{\times}(\mathbf{x})$ involves more complex computations, and it is not as stable as the Jacobians. Therefore, any simplification provided for the Hessian computation would be very welcome.

Dame and Marchand [9] remark that the Hessian of the mutual information should not be approximated by neglecting only the term that involves the image Laplacian, *e.g.* the NCC's term $\sum\frac{\mathbf{H}_{\mathrm{C}(i)}^{\times}(\mathbf{x})}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|}\left(\frac{\mathbf{i}_{\mathrm{R}(i)}}{|\mathbf{i}_{\mathrm{R}}|} - \mathcal{N}_{\times}\frac{\mathbf{i}_{\mathrm{C}(i)}(\mathbf{x})}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|}\right)$. It can be shown, however, that such an approximation is not always negative definite, which is a problem for a maximization problem. A similar approximation is suggested by Brooks and Arbel [13], that might lead to an unexpected behavior caused by an unstable Hessian. There are practical Newton methods [15] that can improve the conditioning of the Hessian, but these modifications add computational burden without a guarantee of increased speed nor basin of convergence. Dame and Marchand [9] suggest an interesting approximation of the Hessian at the solution, *i.e.* approximating the Hessian around $\mathcal{N}_{\times}(\boldsymbol{I}_{\mathrm{C}}(\mathbf{x}), \boldsymbol{I}_{\mathrm{C}}(\mathbf{x}))$ such that

$$\frac{\partial^2 \mathcal{N}_{\times}}{\partial \mathbf{x}^T \partial \mathbf{x}} \approx \mathbf{M}_{\mathrm{FC}} = -\frac{\mathbf{J}_{\mathrm{C}}^{\times}(\widehat{\mathbf{x}}_{\mathrm{k}})^T\mathbf{J}_{\mathrm{C}}^{\times}(\widehat{\mathbf{x}}_{\mathrm{k}})}{|\mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})|^2} + \frac{\mathbf{J}_{\mathrm{C}}^{\times}(\widehat{\mathbf{x}}_{\mathrm{k}})^T}{|\mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})|}\frac{\mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})\mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})^T}{|\mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})|^2}\frac{\mathbf{J}_{\mathrm{C}}^{\times}(\widehat{\mathbf{x}}_{\mathrm{k}})}{|\mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})|} \quad (7)$$

yielding a definite negative matrix.[1] The solution for the DVT is obtained after computing (4) iteratively using Eqs. (5) and (7):

$$\widetilde{\mathbf{x}}_{\mathrm{FC}}^{*} = -{\mathbf{M}_{\mathrm{FC}}}^{-1}\mathbf{g}_{\mathrm{FC}}(\widehat{\mathbf{x}}_{\mathrm{k}})^T, \quad (8)$$

where $\widetilde{\mathbf{x}}_{\mathrm{FC}}^{*}$ is the increment for $\widehat{\mathbf{x}}_{\mathrm{k}+1} = \widehat{\mathbf{x}}_{\mathrm{k}} \circ \widetilde{\mathbf{x}}_{\mathrm{FC}}^{*}$ until a convenient $\widetilde{\mathbf{x}}_{\mathrm{FC}}^{*}\colon |\widetilde{\mathbf{x}}_{\mathrm{FC}}^{*}| < \varepsilon$ is obtained. The same procedure[2] can be used to compute an *inverse compositional* solution for the NCC:

$$\mathbf{g}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}}) = -\left(\mathcal{N}_{\times}\frac{\mathbf{i}_{\mathrm{R}}}{|\mathbf{i}_{\mathrm{R}}|} - \frac{\mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})}{|\mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})|}\right)^T\frac{\mathbf{J}_{\mathrm{R}}^{\times}}{|\mathbf{i}_{\mathrm{R}}|}, \quad (9a)$$

$$\mathbf{M}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}}) = -\frac{\mathbf{J}_{\mathrm{R}}^{\times T}\mathbf{J}_{\mathrm{R}}^{\times}}{|\mathbf{i}_{\mathrm{R}}|^2} + \frac{\mathbf{J}_{\mathrm{R}}^{\times T}}{|\mathbf{i}_{\mathrm{R}}|}\frac{\mathbf{i}_{\mathrm{R}}\mathbf{i}_{\mathrm{R}}{}^T}{|\mathbf{i}_{\mathrm{R}}|^2}\frac{\mathbf{J}_{\mathrm{R}}^{\times}}{|\mathbf{i}_{\mathrm{R}}|}, \quad (9b)$$

$$\widetilde{\mathbf{x}}_{\mathrm{IC}}^{*} = \mathbf{M}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}})^{-1}\mathbf{g}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}})^T, \quad (9c)$$

[1] Technically, this matrix can be semi-definite negative. It can be verified from the eigenvectors of $\mathbf{I}_{mn} - \mathbf{i}_{\mathrm{C}}\mathbf{i}_{\mathrm{C}}{}^T$ that semi-definiteness happens iff $\boldsymbol{I}_{\mathrm{C}}$'s gradients are zero, *i.e.* the reference image is not textured, which is not practical for visual tracking.

[2] Remarking that locally $\widetilde{\mathbf{x}}^{-1} \approx -\widetilde{\mathbf{x}}$.

where $\widetilde{\mathbf{x}}^*_{\mathrm{IC}}$ is the increment for $\widehat{\mathbf{x}}_{\mathrm{k}+1} = \widehat{\mathbf{x}}_{\mathrm{k}} \circ \widetilde{\mathbf{x}}^*_{\mathrm{IC}}$ until a convenient $\widetilde{\mathbf{x}}^*_{\mathrm{IC}} : |\widetilde{\mathbf{x}}^*_{\mathrm{IC}}| < \varepsilon$ is obtained. This approach can be considered as an *improved* version for the *steepest descent*, and this solution is very interesting as $\mathbf{M}_{\mathrm{IC}}$ and its inverse can be computed only once, thus reducing the computational cost of each iteration. Nevertheless, the basin of convergence can be smaller compared to the *forward compositional*. Despite these techniques for the NCC DVT can be quite adequate for some applications, it is still possible to improve the solution of the problem.

## 4 Revisiting the NCC

The NCC is intrinsically robust against affine illumination changes, but there is not a simple and transparent approach to reject occlusion and unmodeled illumination, *e.g.* specular reflections. We propose to redefine the NCC:

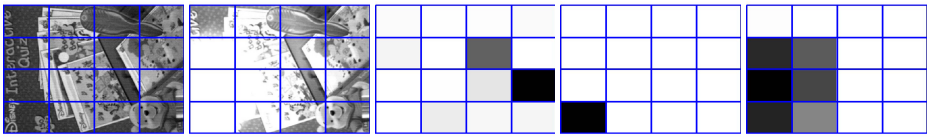
$$\mathcal{N}^{\mathbf{W}}_{\times}\big(\boldsymbol{I}_{\mathrm{R}}, \boldsymbol{I}_{\mathrm{C}}(\mathbf{x})\big) = \frac{\mathbf{i}_{\mathrm{R}}{}^{T}\mathbf{W}\mathbf{i}_{\mathrm{C}}(\mathbf{x})}{|\mathbf{i}_{\mathrm{R}}|_{\mathbf{W}}|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|_{\mathbf{W}}}\,, \tag{10}$$

where $\mathbf{W}$ is symmetric positive definite weighting matrix and $|\mathbf{v}|_{\mathbf{P}} = \sqrt{\mathbf{v}^T\mathbf{P}\mathbf{v}}$. A simpler option suggests $\mathbf{W}$ be written as a diagonal matrix with elements $\mu_i{>}0$.

### 4.1 Local Illumination Changes

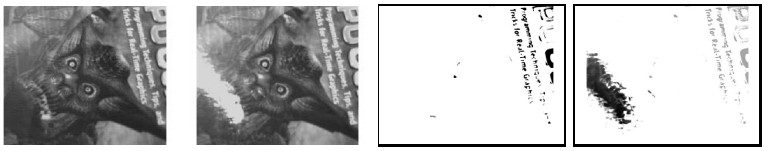
Maximizing the NCC of the whole reference image makes the implicit assumption that the same affine illumination parameters in (1) are shared by every pixel. This hypothesis is but seldom satisfied due to reflective properties of the target, and local illumination sources. Instead of assuming that the reference image represents a target with constant reflective properties, we split $\boldsymbol{I}_{\mathrm{R}}$ in several subregions $\mathcal{G}_i$. Irani and Anandan [12] proposed this grid approach to improve the robustness adding two simple steps: only the concave subregions are taken into account (*i.e.*, all the Hessian eigenvalues are negative), and each subregion is weighted by its determinant. This technique can be helpful whilst using Newton's method, but weighting by the determinant may not be very robust with the approximations presented in §3. Fig. 1 (a) shows a reference image and (b) the same image corrupted by non-uniform illumination. We illustrate the effects of the determinant weighting in Fig. 1 (c) and (d). Remark that the corrupted region in Fig. 1 (b) is not well identified using any of the constant Hessians.

We propose a technique to improve the optimization based on Hessian approximations. The *k-means* algorithm [16] is employed to partition the subregions *two* clusters. We classify the cluster with absolute NCC closer to the unit as *good* $\mathcal{G}^+$, and the other as *bad* $\mathcal{G}^-$. Afterwards, we assign a weight to every subregion. For values lower than $\mathcal{G}^+$'s centroid, weights $\mu^g_i$ are assigned from the current distance to the centroid using Huber's influence function [17]. The other subregions have a weight assigned to $\mu^g_i = 1$. Fig. 1 (e) displays the grid weighting using the proposed technique. We can verify that the proposed method is able to identify the degraded portion of the image, and reduce their corresponding influence in the optimization.



(a) $\boldsymbol{I}_{\mathrm{R}}$ (b) $\boldsymbol{I}_{\mathrm{C}}$ (c) $\mu_i^g : \mathbf{M}_{\mathrm{IC}}$+[12] (d) $\mu_i^g : \mathbf{M}_{\mathrm{FC}}$+[12] (e) $\mu_i^g$ : proposed

**Fig. 1.** Grid weighting procedure; (a) reference image; (b) current image; (c) weights from [12] using $\mathbf{M}_{\mathrm{FC}}$; (d) weights from [12] using $\mathbf{M}_{\mathrm{IC}}$; (e) proposed weights. All weights vary from black: $\mu_i = 0$ to white: $\mu_i = 1$.



(a) $\boldsymbol{I}_{\mathrm{R}}$ (b) $\boldsymbol{I}_{\mathrm{C}}$ (c) $\mu_i^p$ : [18] (d) $\mu_i^p$ : proposed

**Fig. 2.** Pixel weighting procedure; (a) reference image; (b) current image; (c) weights from [18] vary from black: $\mu_i^p = 0.8$; to white: $\mu_i^p = 1$ (d) proposed weights vary from black: $\mu_i^p = 0.25$ to white: $\mu_i^p = 1$

### 4.2 Specular Reflections and Occlusion

Other types of unmodelled changes in the current image can impair direct visual tracking applications, *e.g.* specular reflections and partial occlusions. These effects can indeed be treated by the technique proposed in §4.1. Moreover, if the reference image is already small, the local approach it not very recommended. Arya *et al.* [18] treats such local variations by weighting each pixel from $\boldsymbol{I}_{\mathrm{R}}$ and $\boldsymbol{I}_{\mathrm{C}}$ using their histograms and Huber's influence function. This technique tries to approximate the images by mono-modal distributions. Nevertheless, this weighting might not present the desired effect depending on the degradation level. As an illustration, Fig. 2 (a) represents a reference image and (b) the same image corrupted by specular reflection. Fig. 2 (c) represents the weights using the method [18]. Remark that the specular reflection was not detected, and only the pixels with larger gradients are affected.

Our approach is directly connected to the NCC gradient (5). Observe that the residue $\mathbf{r} \triangleq \frac{\mathbf{i}_{\mathrm{R}}}{|\mathbf{i}_{\mathrm{R}}|} - \mathcal{N}_{\times} \frac{\mathbf{i}_{\mathrm{C}}(\mathbf{x})}{|\mathbf{i}_{\mathrm{C}}(\mathbf{x})|}$ can be defined. This residue $\mathbf{r}$ defines a new distribution. We compute the weights $\mu_i^p$ using Huber's influence function together with the median and the median absolute deviation of $\mathbf{r}$. This weighting approach is similar to the one employed by robust least-squares, however, the NCC defines a distribution invariant to affine illumination changes. Fig. 2 (d) displays the weights computed using the proposed approach. Despite weighting the strong gradients from the right side, the specular reflection is well identified.

### 4.3 Improving the Gradient Solution

It is well known in the literature that robust estimators reduce convergence speed of the optimization in favor of the robustness against outliers. Furthermore, using solely the *inverse* or *forward* solution neglects all the *gradient information* that could be provided either by the *current* or *reference* images. We propose a method to improve the solution. First, we heuristically approximate the parabolas for the *forward* and *inverse* compositional methods using constant Hessians:

$$\begin{cases} \mathcal{N}_{\times}^{\mathrm{FC}} \approx \mathcal{N}_{\times}^{\mathrm{FC}}(\widehat{\mathbf{x}}_{\mathrm{k}}) + \mathbf{g}_{\mathrm{FC}}(\widehat{\mathbf{x}}_{\mathrm{k}})\widetilde{\mathbf{x}} + \frac{1}{2}\widetilde{\mathbf{x}}^T\mathbf{M}_{\mathrm{FC}}\widetilde{\mathbf{x}}\,, \\ \mathcal{N}_{\times}^{\mathrm{IC}} \approx \mathcal{N}_{\times}^{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}}) - \mathbf{g}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}})\widetilde{\mathbf{x}} + \frac{1}{2}\widetilde{\mathbf{x}}^T\mathbf{M}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}})\widetilde{\mathbf{x}}\,. \end{cases} \tag{11}$$

where $\mathcal{N}_{\times}^{\mathrm{FC}} = \mathcal{N}_{\times}(\boldsymbol{I}_{\mathrm{R}}, \boldsymbol{I}_{\mathrm{C}}(\mathbf{x}))$, and $\mathcal{N}_{\times}^{\mathrm{IC}} = \mathcal{N}_{\times}\big(\boldsymbol{I}_{\mathrm{R}}(\mathbf{x} \circ \widetilde{\mathbf{x}}^{-1}), \boldsymbol{I}_{\mathrm{C}}(\mathbf{x})\big)$. Thus, to obtain the maximum we compute the partial derivative w.r.t. $\widetilde{\mathbf{x}}$ for (11) that equals to zero and obtain:

$$\begin{cases} 0 = \mathbf{g}_{\mathrm{FC}}(\widehat{\mathbf{x}}_{\mathrm{k}})^T + \mathbf{M}_{\mathrm{FC}}\widetilde{\mathbf{x}}^*\,, \\ 0 = -\mathbf{g}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}})^T + \mathbf{M}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}})\widetilde{\mathbf{x}}^*\,. \end{cases} \tag{12}$$

Ideally, under the assumption that the similarity function is quadratic, the *inverse* and the *forward* solutions are the same. Nevertheless, in practice, these solutions give complementary information that we propose to exploit. Summing both right hand sides of (12), we obtain the optimal increment:

$$\widetilde{\mathbf{x}}^* = -\big(\mathbf{M}_{\mathrm{FC}} + \mathbf{M}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}})\big)^{-1}\big(\mathbf{g}_{\mathrm{FC}}(\widehat{\mathbf{x}}_{\mathrm{k}}) - \mathbf{g}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}})\big)^T\,. \tag{13}$$

The computational effort of the proposed solution is increased comparing to the *inverse compositional*, however, we double the information employed to solve the optimization. This is at the expense of recomputing $\mathbf{M}_{\mathrm{FC}}(\widehat{\mathbf{x}}_{\mathrm{k}})$ and $\mathbf{g}_{\mathrm{FC}}(\widehat{\mathbf{x}}_k)$ at each iteration. Solution (13) is inspired by the ESM [6], which achieves a second order convergence rate for the SSD without computing the Hessian explicitly. The ESM uses the information of both reference and current image Jacobians, nevertheless, the estimation of the photometric parameters is a necessary task to accomplish a similar result with the NCC. Brooks and Arbel [13] propose an ESM extension for the other similarities than the SSD by directly adding reference and current Jacobians, as proposed by the ESM.[3] Such a solution is contradicted by considering the, theoretically possible, case where the photometric gradients have inverse signals. Note that solution (13) is still valid for that case.

### 4.4 Summary of the Proposed Method

This Section summarizes the contributions of the proposed method in Algorithm 1. The weighting matrix $\mathbf{W}$ is defined as a diagonal matrix, and the elements $\mathbf{W}_{ii}$ are obtained by the multiplication of the weights $\mu_i^g$ from §4.1

[3] The primal ESM solution [6] assumes brightness constancy, therefore it is sound to directly sum the Jacobians, but one must be careful under illumination changes.

**Algorithm 1.** Visual tracking using proposed method

**Input:** threshold $\varepsilon$, maximum number of iterations $\overline{\mathrm{k}}$, grid $\mathcal{G}$.

1: **for each** subregion $\mathcal{G}_i \in \mathcal{G}$ **do**
2: $\quad \mathbf{i}_{\mathrm{R}} : \mathbf{i}_{\mathrm{R}_i} = \boldsymbol{I}_{\mathrm{R}_i} - \frac{1}{mn}\sum_p \boldsymbol{I}_{\mathrm{R}}(p)$,
$\quad \mathbf{J}^{\times}_{\mathrm{R}} : \mathbf{J}^{\times}_{\mathrm{R}_i} = \mathbf{J}_{\mathrm{R}_i} - \frac{1}{mn}\sum_i \mathbf{J}_{\mathrm{R}_i}$, $\mathbf{J}_{\mathrm{R}_i} = \left.\dfrac{\partial \boldsymbol{I}_{\mathrm{R}}(w(\mathbf{x}, p_i)}{\partial \mathbf{x}}\right|_{\mathbf{x}=0}$.
3: **end for**
4: **for each** new image $\boldsymbol{I}_{\mathrm{C}}$ **do**
5: $\quad$ **repeat**
6: $\quad\quad$ **for each** subregion $\mathcal{G}_i$ **do**
7: $\quad\quad\quad \mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}}) : \mathbf{i}_{\mathrm{C}_i}(\widehat{\mathbf{x}}_{\mathrm{k}}) = \boldsymbol{I}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}}) - \frac{1}{mn}\sum_p \boldsymbol{I}_{\mathrm{C}_i}\big(w(\widehat{\mathbf{x}}_{\mathrm{k}}, p)\big)$,
$\quad\quad\quad \mathbf{J}^{\times}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}}) : \mathbf{J}^{\times}_{\mathrm{C}_i}(\widehat{\mathbf{x}}_{\mathrm{k}}) = \mathbf{J}_{\mathrm{C}_i}(\widehat{\mathbf{x}}_{\mathrm{k}}) - \frac{1}{mn}\sum_i \mathbf{J}_{\mathrm{C}_i}(\widehat{\mathbf{x}}_{\mathrm{k}}), \mathbf{J}_{\mathrm{C}_i} = \left.\dfrac{\partial \boldsymbol{I}_{\mathrm{C}}(w(\mathbf{x}, p_i))}{\partial \mathbf{x}}\right|_{\mathbf{x}=\widehat{\mathbf{x}}_{\mathrm{k}}}$.
8: $\quad\quad$ **end for**
9: $\quad\quad$ Cluster good $\mathcal{G}^+$ and bad $\mathcal{G}^-$ subregions by $\mathcal{N}_{\times} = \frac{\mathbf{i}_{\mathrm{R}}{}^T \mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})}{|\mathbf{i}_{\mathrm{R}}||\mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})|}$,
$\quad\quad$ Compute weights $\mu_i^{\mathcal{G}}$ of every subregion $\mathcal{G}_i$, *c.f.* §4.1.
10: $\quad\quad$ Compute weights $\mu_i^p$ of every pixel; dist. $\mathbf{r} = \frac{\mathbf{i}_{\mathrm{R}}}{|\mathbf{i}_{\mathrm{R}}|} - \mathcal{N}_{\times}\frac{\mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})}{|\mathbf{i}_{\mathrm{C}}(\widehat{\mathbf{x}}_{\mathrm{k}})|}$ of $\mathcal{G}^+$, *c.f.* §4.2.
11: $\quad\quad$ Compute $\mathbf{W}_{ii} = \mu_i^g \mu_i^p$; compute $\widetilde{\mathbf{x}}^*$ via (14), then update $\widehat{\mathbf{x}}_{\mathrm{k}+1} = \widehat{\mathbf{x}}_{\mathrm{k}} \circ \widetilde{\mathbf{x}}^*$.
12: $\quad$ **until** $|\widetilde{\mathbf{x}}_{\mathrm{k}}| < \varepsilon$ or $\mathrm{k} > \overline{\mathrm{k}}$
13: **end for**

and $\mu_i^p$ from §4.2. Remark that the use of the weighting is optional, and grid or pixel weighting are neglected by setting $\mu_i^g = 1$ or $\mu_i^p = 1$, respectively. Using the same procedure as §3, it is direct to obtain the explicit forms of $\mathbf{g}^{\mathbf{W}}_{\mathrm{FC}}(\widehat{\mathbf{x}}_{\mathrm{k}})$, $\mathbf{M}^{\mathbf{W}}_{\mathrm{FC}}$, $\mathbf{g}^{\mathbf{W}}_{\mathrm{FC}}(\widehat{\mathbf{x}}_{\mathrm{k}})$, and $\mathbf{g}^{\mathbf{W}}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}})$ needed by the optimization of the revisited NCC (10). The solution is given by:

$$\widetilde{\mathbf{x}}^* = -\big(\mathbf{M}^{\mathbf{W}}_{\mathrm{FC}} + \mathbf{M}^{\mathbf{W}}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}})\big)^{-1}\big(\mathbf{g}^{\mathbf{W}}_{\mathrm{FC}}(\widehat{\mathbf{x}}_{\mathrm{k}}) - \mathbf{g}^{\mathbf{W}}_{\mathrm{IC}}(\widehat{\mathbf{x}}_{\mathrm{k}})\big)\,, \tag{14}$$

where $\widetilde{\mathbf{x}}^*$ is the increment that composes $\widehat{\mathbf{x}}_{\mathrm{k}+1} = \widehat{\mathbf{x}}_{\mathrm{k}} \circ \widetilde{\mathbf{x}}^*$.

## 5 Experiments

We evaluate the tracking accuracy and the robustness w.r.t. illumination variations and partial occlusions using a benchmark dataset and challenging video sequences. Our objective is to show that the proposed techniques can improve the NCC tracking, in terms of computational effort and speed, to the same level as other state of the art methods whilst improving the robustness to concurrent illumination changes and partial occlusions. For all of the experiments, we compute the homographic warp using the $\mathbb{SL}(3)$ parametrization as [6].

### 5.1 Evaluation on Benchmark Dataset

First, we evaluate the proposed method using a planar tracking benchmark [14]. This benchmark consists of 8 different reference images classified among low,

repetitive, normal and high texture. There are sequences of 5 different motion types for each target: high angles, distance range, fast far motion, fast close motion and, at last, illumination changes. The estimated positions obtained from the visual tracking are compared to a ground truth database, and the tracking is considered successful if the template position error is lower than 10 pixels. The results are given as rate of successfully tracked frames.

We tested four different methods using this dataset: two methods obtained from the literature, and two implementations of the method we propose. The first from the literature is an implementation of the SSD with local illumination changes (*SSD+i*) from Silveira and Malis [7], and the second is an NCC with *inverse compositional*-like *steepest descent* optimization (*NCC+ICS*), *c.f.* §3 and Eq. (9). Two implementations of the proposed method are evaluated, first, without the pixel-wise robust weighting (Proposed-I), *i.e.* Algorithm 1 without computing line 10; and secondly, with pixel-wise robust weighting (Proposed-II), *i.e.* Algorithm 1 fully implemented.

We use the same stopping criteria for every method: the optimization is stopped with $\varepsilon = 10^{-3}$, and each method is allowed to run at most 200 iterations to provide the same opportunity to the four tested methods. The warps are computed using bilinear interpolation, and we use a downsized template to 320×240 pixels instead of the original $640 \times 480$ pixels so to avoid oversampling in most of the sequences. The only difference between the parameters refers to the local grid. The *SSD+i* uses a grid of 5×5 local patches, where the proposed methods use a grid of 3×3 local patches. The scores obtained by the four methods are presented in Table 1, the numbers in bold refer to the best result obtained for the dataset, and the underlined refer to a unique best result. For the sake of comparison, we consider any score difference below 5% to be irrelevant.

Among the methods implemented from the literature, we can verify that the *NCC+ICS* presents the worst results since it did not outperform any of the other three methods a single time. Additionally, it could only achieve scores similar to the other methods in 25% of the sequences. The results obtained from the other three methods were similar in 52.5% of the sequences, however, for this dataset, we could classify the *SSD+i* performance as *worse* than the proposed methods. The *SSD+i* was the best tracker throughout 10% of the sequences, that yielded an average improvement of 22% of tracked frames, *i.e.*, 264 frames.

The two proposed methods obtained similar results for 80% of sequences, and one or both methods outperformed the *SSD+i* in 37.5% sequences, that yielded an average improvement of 21% tracked frames, *i.e.* 252 frames. We remark the outstanding performance of Proposed-I for sequences with illumination changes, where the method was able to track more than 99.8% of all images. This method performed better in 25% of illumination sequences probably because the robust weighting reduces the influence of pixels with strong gradients, and these are specially responsible for the method's accuracy. Other authors have evaluated their methods using this benchmark, *e.g.* the mutual information from Dame and Marchand [9], the interested reader can verify that the results for the proposed method are of the same order.

**Table 1.** Scores obtained for benchmark [14]. Methods: (1) *SSD+i*, (2) *NCC+ICS*, (3) PROPOSED-I, (4) PROPOSED-II. Rows display textures: L. stands for Low, R. for Repetitive, N. for Normal and H. for High. Columns display sequences: A. stands for Angle, R. for Range, F.F. for Fast Far, F.C. for Fast Close, I. for Illumination.

| (1) | A. | R. | F.F. | F.C. | I. |
|---|---|---|---|---|---|
| L. | 99.8% | 77.3% | **62.1%** | **39.6%** | 97.8% |
| | 100% | 98.9% | **54.9%** | 51.0% | 91.1% |
| R. | 100% | **87.6%** | 26.4% | 72.6% | 100% |
| | 91.2% | 64.9% | 12.7% | 72.6% | 61.4% |
| N. | 99.8% | 98.9% | 38.8% | 50.1% | 99.6% |
| | 100% | 99.9% | 14.7% | 84.5% | 100% |
| H. | 66.3% | 28.6% | 7.0% | 11.2% | 34.8% |
| | 100% | 51.8% | 13.2% | 36.2% | 90.2% |

| (2) | A. | R. | F.F. | F.C. | I. |
|---|---|---|---|---|---|
| L. | 99.9% | 76.8% | 16.8% | 26.7% | 97.8% |
| | 99.8% | 92.8% | 16.8% | 28.7% | 75.5% |
| R. | 24.8% | 22.6% | 11.6% | 31.4% | 85.4% |
| | 35.6% | 18.3% | 7.7% | 37.6% | 34.6% |
| N. | 83.2% | 71.3% | 15.6% | 64.2% | 100% |
| | 63.7% | 53.7% | 11.5% | 24.2% | 96.6% |
| H. | 13.0% | 9.7% | 5.7% | 3.8% | 15.5% |
| | 51.7% | 25.8% | 11.8% | 11.8% | 67.9% |

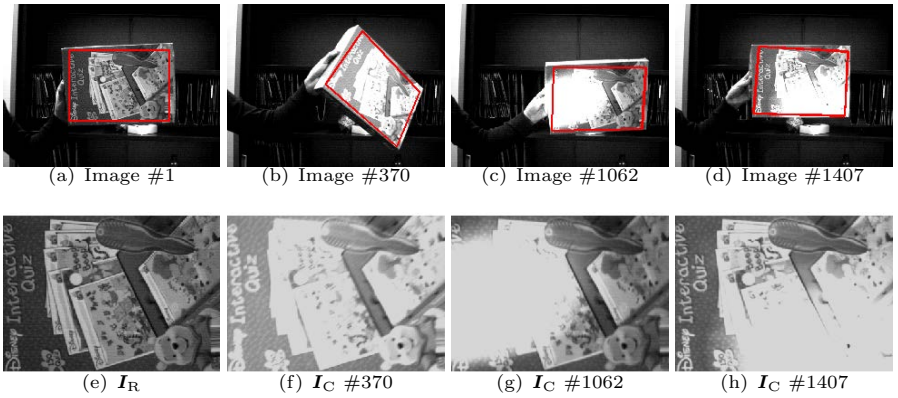
| (3) | A. | R. | F.F. | F.C. | I. |
|---|---|---|---|---|---|
| L. | 99.7% | 76.8% | 52.7% | 27.6% | 100% |
| | 100% | 99.9% | 21.6% | **66.0%** | **100%** |
| R. | 100% | 57.7% | 22.2% | 68.2% | 100% |
| | **100%** | **81.3%** | 12.2% | 53.6% | **100%** |
| N. | 100% | 96.8% | **58.2%** | **90.5%** | 100% |
| | 99.9% | 99.9% | **20.1%** | 80.5% | 100% |
| H. | **93.6%** | **52.3%** | 9.2% | 14% | **98.9%** |
| | 100% | 51.5% | **22.0%** | **75.0%** | **100%** |

| (4) | A. | R. | F.F. | F.C. | I. |
|---|---|---|---|---|---|
| L. | 99.8% | **92.2%** | 51.8% | 31.6% | 100% |
| | 100% | 95.8% | 13.5% | 42.1% | 85.2% |
| R. | 100% | 59.1% | 22.3% | 68.1% | 100% |
| | **100%** | **81.1%** | 10.5% | 69.1% | **100%** |
| N. | 100% | 96.1% | **58.5%** | **86.1%** | 100% |
| | 99.8% | 99.9% | **20.5%** | 85.3% | 100% |
| H. | 76.4% | 16.9% | 7.2% | 9.7% | 59.8% |
| | 100% | **69.7%** | **19.7%** | 42.8% | **100%** |

### 5.2 Evaluation under Challenging Illumination

Secondly, we evaluate the proposed method on two sequences from Silveira and Malis [7]. These sequences represent extreme real-world situations with challenging illumination and targets from different materials and sizes. The obtained results are compared against the *SSD+i* for BEAR and BOOK sequences. Besides 8 geometric parameters from the $\mathbb{SL}(3)$, the *SSD+i* must estimate other photometric parameters that increase with the number of grids. We evaluate the methods using the same reference images, minimum step size $\varepsilon = 10^{-3}$, and 50 maximum iterations. These are reasonable parameters for most real-time applications. Furthermore, we found the BOOK to be more complex because none of the methods was able to complete the tracking using the standard parameters. We thus reevaluate this sequence using the proposed method with $\varepsilon = 10^{-4}$ and 500 iterations. Table 2 presents the comparative result in terms of total tracked images, median of iterations per image, number of estimated parameters, and the median NCC of the final $\boldsymbol{I}_{\mathrm{C}}$ and $\boldsymbol{I}_{\mathrm{R}}$. Figs. 3, and 4 present key samples results obtained for the proposed method in sequences BEAR, and BOOK respectively. The full videos of these sequences can be found in `http://goo.gl/qYZu0`.

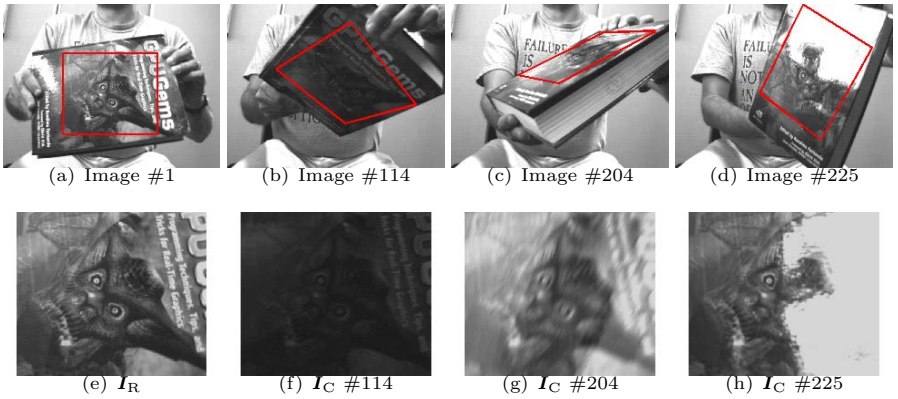
We can verify that the proposed method performed at least similarly to the *SSD+i*. Note that the proposed method presents a slight increase in the number of iterations, however, we obtained a median of 23 iterations for the *NCC+ICS* (using the robust techniques). This result highlights the importance of the improvement proposed in §4.3, since, using the information from *inverse* and *forward* solutions, the NCC compares to a second order method in terms of iterations. Remark that the proposed method performed better than the *SSD+i* for the BOOK sequence. The *SSD+i* gets stuck in a local minimum at frame 163, however, the proposed method is able to continue until 203 using the same parameters. The decrease of the median NCC and increase on the iteration



(a) Image #1 (b) Image #370 (c) Image #1062 (d) Image #1407

(e) $\boldsymbol{I}_{\mathrm{R}}$ (f) $\boldsymbol{I}_{\mathrm{C}}$ #370 (g) $\boldsymbol{I}_{\mathrm{C}}$ #1062 (h) $\boldsymbol{I}_{\mathrm{C}}$ #1407

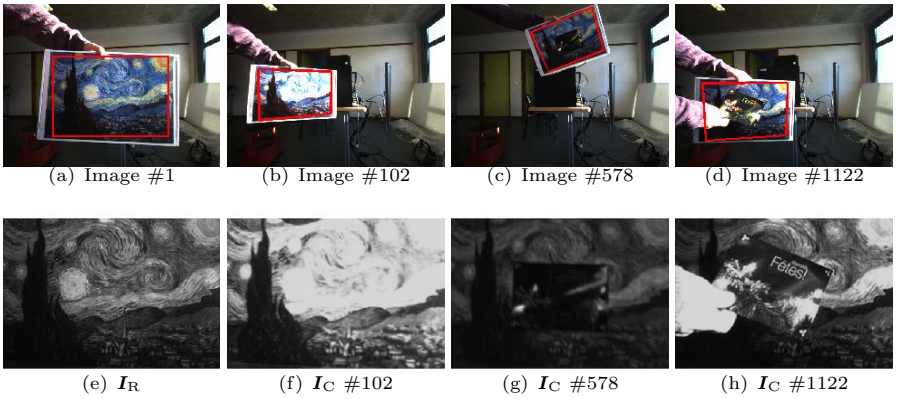
**Fig. 3.** Samples from BEAR sequence – Total of 1573 images



(a) Image #1 (b) Image #114 (c) Image #204 (d) Image #225

(e) $\boldsymbol{I}_{\mathrm{R}}$ (f) $\boldsymbol{I}_{\mathrm{C}}$ #114 (g) $\boldsymbol{I}_{\mathrm{C}}$ #204 (h) $\boldsymbol{I}_{\mathrm{C}}$ #225

**Fig. 4.** Samples from BOOK sequence – Total of 283 images



(a) Image #1 (b) Image #102 (c) Image #578 (d) Image #1122

(e) $\boldsymbol{I}_{\mathrm{R}}$ (f) $\boldsymbol{I}_{\mathrm{C}}$ #102 (g) $\boldsymbol{I}_{\mathrm{C}}$ #578 (h) $\boldsymbol{I}_{\mathrm{C}}$ #1122

**Fig. 5.** Samples from STARRY NIGHT sequence – Total of 1600 images

**Table 2.** Comparative results under challenging illumination and occlusion

BEAR

| Method | #Img. | #Its. | #Parms. | RMS | $\mathcal{N}_\times$ |
|---|---|---|---|---|---|
| *SSD+i* | 1573 | 10 | 8G+36P | 20 | 0.807 |
| P.M. | 1573 | 12 | 8G | – | 0.807 |

BOOK

| Method | #Img. | #Its. | #Parms. | RMS | $\mathcal{N}_\times$ |
|---|---|---|---|---|---|
| *SSD+i* | 163 | 11 | 8G+49P | 10.1 | 0.962 |
| P.M. | 203 | 14 | 8G | – | 0.957 |
| P.M. $\varepsilon$=$10^{-4}$ | 283 | 19 | 8G | – | 0.959 |

STARRY-NIGHT

| Method | #Img. | #Its. | #Parms. | RMS | $\mathcal{N}_\times$ |
|---|---|---|---|---|---|
| *SSD+i* | 400 | 12 | 8G+43P | 13.3 | 0.939 |
| *SSD+i*(M) | 680 | 21 | 8G+43P | 21.4 | 0.876 |
| P.M. | 1600 | 22 | 8G | – | 0.849 |

numbers is directly related to the sequences that the *SSD+i* was unable to track. We consider frame 204 to be the most difficult from this sequence. Our method was only capable of completing the sequence without the real-time constraint imposed by the iterations. The *SSD+i* still failed in the same local minimum (frame 163). We invite the reader to revisit Fig. 4 and examine the samples with extreme tilt and illumination changes.

### 5.3 Evaluation under Partial Occlusion and Illumination Changes

The last experiment evaluates the *SSD+i* and the proposed method in a real-world situation where part of the reference image is partially occluded and illumination is varying throughout the experiment. Again, we use the same parameters: reference image, minimum step and maximum iterations for both methods. Table 2 presents the comparative results from *SSD+i* and its implementation with a robust M-estimator, and the proposed method. Fig. 5 presents some key frames results obtained for the proposed method in the sequence STARRY NIGHT, The full video can be found in `http://goo.gl/qYZu0`. The proposed method was the only capable of completing the full sequence. The *SSD+i* was unable to cope with the partial occlusion, and this is the main reason why it presents less median iterations per frame and a larger NCC than the other two methods. The *SSD+i* with M-estimator was capable of tracking the occluded patch as long as there were no illumination changes. We can infer that the *SSD+i* is unable to handle partial occlusion and illumination changes at the same time. The capability of facing illumination changes and occlusion supports the weighting techniques presented in §4. To the authors knowledge, this is the only approach capable of dealing with such extreme situations.

## 6 Discussion

We have proposed a novel solution to the problem of direct visual tracking under illumination changes and partial occlusions. The direct tracking is solved using the NCC, mainly because this similarity is intrinsically invariant to affine illumination changes. Two techniques are presented to increase the robustness

against non-modeled effects and, compared to the state of the art, these techniques highly improve the rejection of degraded areas. We also address how to improve the gradient solution using both *inverse* and *forward compositional* approaches. Our method is compared against state of the art methods in a benchmark dataset and challenging real-world video sequences. We verify that the proposed method is able to cope with tracking partially occluded objects even under severe illumination changes.

## References


1. Szeliski, R.: Image alignment and stitching: a tutorial. Foundations and Trends in Computer Graphics and Vision 2 (2006)
2. Comport, A.I., Malis, E., Rives, P.: Real-time quadrifocal visual odometry. International Journal of Robotics Research 29 (2010)
3. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: International Symposium on Mixed and Augmented Reality (2007)
4. Roth, P.M., Winter, M.: Survey of appearance-based methods for object recognition. Technical Report ICG-TR-01/08 (2008)
5. Baker, S., Matthews, I.: Equivalence and efficiency of image alignment algorithms. In: IEEE Conference on Computer Vision and Pattern Recognition (2001)
6. Benhimane, S., Malis, E.: Homography-based 2D visual tracking and servoing. International Journal of Robotics Research 26 (2007)
7. Silveira, G., Malis, E.: Unified direct visual tracking of rigid and deformable surfaces under generic illumination changes in grayscale and color images. International Journal of Computer Vision 89 (2010)
8. Hager, G., Belhumeur, P.: Efficient region tracking with parametric models of geometry and illumination. IEEE Transactions Pattern Analysis and Machine Intelligence 20 (1998)
9. Dame, A., Marchand, E.: Accurate real-time tracking using mutual information. In: International Symposium on Mixed and Augmented Reality (2010)
10. Wang, F., Vemuri, B.C.: Non-rigid multi-modal image registration using cross-cumulative residual entropy. International Journal of Computer Vision 74 (2007)
11. Richa, R., Hager, G.: Robust similarity measures for direct gradient-based visual tracking. Technical report (2012), `http://www.cs.jhu.edu/~richa/robust.html`
12. Irani, M., Anandan, P.: Robust multi-sensor image alignment. In: International Conference on Computer Vision (1998)
13. Brooks, R., Arbel, T.: Generalizing inverse compositional and ESM image alignment. International Journal of Computer Vision 87 (2010)
14. Lieberknecht, S., Benhimane, S., Meier, P., Navab, N.: A dataset and evaluation methodology for template-based tracking algorithms. In: International Symposium on Mixed and Augmented Reality (2009)
15. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer (2000)
16. Lloyd, S.: Least squares quantization in PCM. IEEE Transactions on Information Theory 28 (1982)
17. Huber, P.J.: Robust Statistics. John Wiley and Sons, New York (1981)
18. Arya, K., Gupta, P., Kalra, P., Mitra, P.: Image registration using robust M-estimators. Pattern Recognition Letters 28 (2007)