2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Camera Localization using Mutual Information-based Multiplane Tracking

Bertrand Delabarre, Eric Marchand

Abstract— **This paper deals with dense visual tracking robust towards scene perturbations using 3D information to provide a space-time coherency. The proposed method is based on an piecewise-planar scenes visual tracking algorithm which aims to minimize an error between an observed image and a reference template by estimating the parameters of a rigid 3D transformation taking into acount the relative positions of the planes in the scene. The major drawback of this approch stems from the registration function used to perform the minimization (the sum of squared differences) as it is very poorly robust towards scene variations. In this paper, the tracking process is adapted to take into account two more complex registration functions. First, the sum of conditional variance. Since it is invariant to global illumination variations, the proposed algorithm is robust with relation to those conditions whilst keeping a low computation complexity. Then, the mutual information is considered. In that case the complexity is greater but so is the robustness towards non global illumination variations, specularities or occlusions. The proposed approaches, after being described, are tested on different scenes under varying illumination conditions to assess their respective efficiency.**

I. Introduction

Visual tracking is a fundamental step of robotics vision. Its field of application is vast, including for example visual servoing [5], pose estimation [4] or augmented reality [6]. Visual tracking approaches can be divided in several branches. It is possible for instance to differenciate approaches based on visual features extracted from the images such as keypoints or lines and dense methods also called template-based registration methods relying on a template extracted from a reference image. This paper deals with the latter category. When performing such visual tracking, the goal is to optimize a registration function representing the difference or similarity between a reference template and the current image. Several works have focused on different registration functions from the most simple, the sum of squared differences (SSD) [2], which compares the luminance of each pixel and is therefore very poorly robust to variations of the scene to sophisticated ones such as the mutual information (MI) [8], [7], very robust towards scene perturbations but quite complex to implement. Other functions have also been considered which can be put in between the two previously named such as the sum of conditional variance (SCV) [14] or the normalised cross correlation (NCC) [15]. They both are easier to use than the MI and more robust to global illumination variations than the SSD. Those approaches lead to visual tracking algorithms optimizing, for most of them, the parameters of a 2D displacement (translation, affine motion, homography) in the image frame as in [2], [8], [14] but can also be based on a rigid 3D displacement as in [3] or even on the parameters of a camera pose as in [4].

The aim of this paper is to introduce a visual tracking process robust towards scene variations such as illumination variations (global or local) and occlusions while integrating the Euclidean constraints of the observed scene so as to keep spatial coherency between the followed template planes. To that end, the optimization is performed on $\mathrm{SE}(3)$, a space where those constraints are known. The proposed algorithm is based on the approach introduced by the authors of [3] which is adapted to both the SCV and MI. Our main contribution is to improve the approach proposed in [3] to more robust and complex similarity functions and the use of SCV and MI in a 3D optimization scheme to track complex objects with appearing or disappearing faces contrary to what was done in [14] and [8] who only considered 2D displacement in the image. The use of the SCV in the algorithm should allow robustness towards global illumination variations which are frequent in real life scenes, in particular exterior scenes, while keeping computation time low as the only difference with the SSD is the actualisation of the reference template at each new frame. A version using the mutual information is also considered, as the invariance of MI towards all kinds of scene perturbations such as specularities or occlusions should insure a greater robustness of the tracking. The main drawback should be the complexity of integrating the registration function into the algorithm. Both approaches optimize the parameters of the 3D displacement in $\mathrm{SE}(3)$ which allows, in addition of keeping a spatial coherency of the scene, to directly gather the displacement of the camera without any further computation.

The paper is organized as follows. First, the main principles of differential template tracking are recalled and the algorithm used in [3] is introduced. The two other considered registration functions are then expressed and their integration into the algorithm are detailled. Finally, experimental results are shown that validate the approach in different environments.

II. Differential template tracking

Differential template tracking [2] is a class of approaches based on the optimization of an image registration function. They aim to estimate the displacement $\mathbf{p}$ of a template I^* (that is a set of pixels) in an image sequence. To define the template I^*, the usual method is to extract it from the

B. Delabarre is with Université de Rennes 1, IRISA, INRIA, Lagadic team, Rennes, France. bertrand.delabarre@irisa.fr

E. Marchand is with Université de Rennes 1, IRISA, INRIA Lagadic team, Rennes, France. eric.marchand@irisa.fr

first image of the sequence. Then, considering a difference function f, the problem can be written as:

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} f(I^*, w(I_t, \mathbf{p})). \tag{1}$$

In that case, the goal is to find the displacement $\hat{\mathbf{p}}$ that minimizes the difference between the template I^* and the current image in the sequence I_t warped with relation to the last known displacement $\mathbf{p}$. Please note that the global warp of the image $w(I_t, \mathbf{p})$ is used as an abuse of the proper notation $w(\mathbf{x}, \mathbf{p})$ representing the position of a single warped point $\mathbf{x}$.

L'espace des paramètres $\mathbf{x}$ à estimer est variable puisque $\mathbf{T}(x)$ peut The dimension and nature of the displacement $\mathbf{p}$ is variable as it can be used to represent several types of transformations $\mathbf{T}(\mathbf{p})$. For example, the approach brought on in [11], [2] began considering only simple translations leading to $\mathbf{p} \in \mathbb{R}^2$. Later on, other models were considered such as affine transformation in [9] giving $\mathbf{p} \in \mathbb{R}^6$, homographies as in [2] creating $\mathbf{p} \in \mathrm{SL}(3)$ or even [3] leading to $\mathbf{p} \in \mathrm{SE}(3)$.

III. Multi-planar scene tracking

To consider a multiplane tracking approach, the choice has been made to optimize displacament parameters of a camera pose in SE(3). This allows to keep a space-time coherency, as the Euclidean constraints are constant and easy to determine in a 3D space. The following section will recall the methodology introduced in [3] to perform such a visual tracking.

A. Notations

The rigid 3D transformation between two frames $\mathbf{T}(\mathbf{r})$ can be expressed as a homogeneous matrix:

$$\mathbf{T}(\mathbf{r}) = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \tag{2}$$

where $\mathbf{r}$ are the six parameters of a 3D displacement, $\mathbf{R}$ is a rotation matrix ($\mathbf{R} \in \mathrm{SO}(3)$) and $\mathbf{t}$ is a translation vector ($\mathbf{t} \in \mathbb{R}^3$). Let us then define $\mathbf{H}(\mathbf{T}(\mathbf{r}))$ the homography linking the projection of a plane in each frame:

$$\mathbf{H}(\mathbf{T}(\mathbf{r})) = \mathbf{R} + \mathbf{t}\mathbf{n}_d^{*\top} \tag{3}$$

where ${\mathbf{n}^*}_{d^*} = \frac{\mathbf{n}}{d^*}$ is the ratio between the unitary normal to the plane in the origin frame and d^* the distance between the plane and the origin of the frame. Let us note that one can divide $\mathbf{H}(\mathbf{T}(\mathbf{r}))$ by $\sqrt[3]{1 + \mathbf{t}^\top \mathbf{R}\mathbf{n}_d^{*\top}}$ in order to have a normalised homography. This homography can also be expressed in the camera frame:

$$\mathbf{G}(\mathbf{r}) = \mathbf{K}\mathbf{H}(\mathbf{T}(\mathbf{r}))\mathbf{K}^{-1} \tag{4}$$

where $\mathbf{K}$ is the matrix containing the intrinsic parameters of the camera:

$$\mathbf{K} = \begin{bmatrix} px & 0 & u_0 \\ 0 & py & v_0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

The transformation can therefore be written in the image frame as:

$$\mathbf{x}_2 = \mathbf{G}(\mathbf{r})\mathbf{x}_1 \tag{6}$$

with $\mathbf{x}_1$=$(u1, v1, 1)^\top$ and $\mathbf{x}_2$=$(u2, v2, 1)^\top$ respectively the origin point and the resulting point of the transformation. This can also be expressed using a warp function:

$$\mathbf{x}_2 = w(\mathbf{x}_1, \mathbf{G}(\mathbf{r})) = w(\mathbf{x}_1, \mathbf{r}). \tag{7}$$

Given equation (7) it is also possible to find the original point from the warped one with:

$$\begin{aligned} \mathbf{x}_1 &= w^{-1}(\mathbf{x}_2, \mathbf{r}) = w(\mathbf{x}_2, \mathbf{r}^{-1}) \\ &= \mathbf{G}(\mathbf{r})^{-1}\mathbf{x}_1. \end{aligned} \tag{8}$$

B. Tracking algorithm

The authors of [3] have proposed a method that includes euclidean constraints between several planes into the tracking task. In order to do that, the optimisation is done on 6 parameters representing a 3D transformation in SE(3) using the sum of squared differences (SSD) as the registration function. Let us consider a template I^* of size $N_{\mathbf{x}}$ pixels representing the projection of a plane in the 3D frame. The tracking process then consists in finding the parameters of the transformation $\mathbf{T}(\mathbf{r})^k \in \mathrm{SE}(3)$ traducing the displacement of the considered scene at the iteration k. Considering an inverse compositional approach [1], the goal of the optimization is to find the optimal increment of parameters $\Delta\mathbf{r}$ which verifies $\forall \mathbf{x}_i \in I$:

$$\begin{aligned} I(w(\mathbf{x}_i, \mathbf{r}^{k-1})) &= I^*(w^{-1}(\mathbf{x}_i, \Delta\mathbf{r})) \\ &= I^*(\mathbf{G}(\Delta\mathbf{r})^{-1}\mathbf{x}). \end{aligned} \tag{9}$$

The 3D transformation $\mathbf{T}(\mathbf{r})$ being the same for evey plane in the scene the process can therefore track several planes in a unique optimization loop, the difference between planes coming from the different matrices $\mathbf{G}$ induced by $\Delta\mathbf{r}$. For the reminder of this paper the warp associated to a plane l will be noted w_l. The optimal increment of parameters $\widehat{\Delta\mathbf{r}}$ is obtained by minimizing the SSD between the current image warped with the displacement parameters computed at the last frame $\mathbf{r}$ and the template warped with current parameters $\widehat{\Delta\mathbf{r}}$:

$$\begin{aligned} \widehat{\Delta\mathbf{r}} = \arg\min_{\Delta\mathbf{r}} \sum_l \sum_{i=1}^{N_{\mathbf{x}_l}} \Big[& I^*(w_l(\mathbf{x}_i, \Delta\mathbf{r})) - \\ & I(w_l(\mathbf{x}_i, \mathbf{r}^{k-1})) \Big]^2. \end{aligned} \tag{10}$$

The displacement is subsequently updated as follows:

$$\mathbf{T}(\mathbf{r})^k \leftarrow \mathbf{T}(\mathbf{r})^{k-1}\mathbf{T}(\widehat{\Delta\mathbf{r}})^{-1}. \tag{11}$$

To perform this minimization as was proposed in [1] in SL(3) and [3] in SE(3), let us start by expressing the first order Taylor expansion associated to the chosen registration function:

$$SSD(\Delta\mathbf{r}) = \sum_l \sum_{i=1}^{N_{\mathbf{x}_l}} \left[I^*(w_l(\mathbf{x}_i, \Delta\mathbf{r})) - I(w_l(\mathbf{x}_i, \mathbf{r}^{k-1})) \right]^2 \tag{12}$$

which is defined as:

$$SSD(\Delta\mathbf{r}) \simeq \sum_{l}\sum_{i=1}^{N_{\mathbf{x}_l}} \left[I^*(\mathbf{x}_i) - I(w_l(\mathbf{x}_i, \mathbf{r}^{k-1}))\right]^2 + \mathbf{J}(\Delta\mathbf{r})\Delta\mathbf{r} \quad (13)$$

where $\mathbf{J}(\Delta\mathbf{r})$ is the Jacobian matrix of $SSD(\Delta\mathbf{r})$.

Decomposing the Jacobian matrix thanks to the different transformations applied to each pixel gives:

$$\begin{aligned}\mathbf{J}(\Delta\mathbf{r}) &= \frac{\partial I^*}{\partial w_l}\frac{\partial w_l}{\partial \mathbf{K}}\frac{\partial \mathbf{K}}{\partial \mathbf{T}}\frac{\partial \mathbf{T}}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial \Delta\mathbf{r}} \\ &= \mathbf{J}_{I^*}\mathbf{J}_{w_l}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\mathbf{T}}\mathbf{J}_{\mathbf{x}}(\Delta\mathbf{r})\end{aligned} \quad (14)$$

leading to [3]:

$$\widehat{\Delta\mathbf{r}} = -\left(\mathbf{J}_{I^*}\mathbf{J}_{w_l}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\mathbf{T}}\mathbf{J}_{\mathbf{x}}(\mathbf{0})\right)^{+} SSD(\mathbf{0}). \quad (15)$$

Let us note that an ESM approach can also be chosen to perform the optimization as it was chosen in [3]. In this eventuality, the update of the displacement is given by:

$$\widehat{\Delta\mathbf{r}} = -\left(\left(\frac{\mathbf{J}_I + \mathbf{J}_{I^*}}{2}\right)\mathbf{J}_{w_l}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\mathbf{T}}\mathbf{J}_{\mathbf{x}}(\mathbf{0})\right)^{+} SSD(\mathbf{0}). \quad (16)$$

The problem when using the SSD as the registration function of a tracking process is the fact that it is not adapted to the perturbations that are usually undergone by real life scenes such as illumination variations or occlusions. This is why in this paper the sum of conditional variance is used to enhance the robustness of the approach.

C. Sum of conditionnal variance

In [14], it has been proposed a tracking algorithm based on the sum of conditionnal variance (SCV) for 2D homography estimation. The SCV is a template-based difference function but rather than using the raw template I^* as the SSD, it is adapted at each step of the tracking process to the illumination conditions of the current image I, creating an adapted template $\hat{I}$ thanks to an expectation operator $\mathcal{E}$:

$$\hat{I}(\mathbf{x}) = \mathcal{E}(I(\mathbf{x}) \mid I^*(\mathbf{x})). \quad (17)$$

This operator computes, for each grey level in I^*, an adapted one which reflects the changes the template would undergo given the current illumination conditions of I:

$$\hat{\mathbf{I}}(j) = \sum_i i\frac{p_{II^*}(i,j)}{p_{I^*}(j)} \quad (18)$$

where p_{I^*} and p_{II^*} are respectively the probability density function and joint probability density function of I^* and I:

$$\begin{aligned}p_{II^*}(i,j) &= P(I(\mathbf{x}) = i, I^*(\mathbf{x}) = j) \\ &= \frac{1}{N_{\mathbf{x}}}\sum_{k=1}^{N_{\mathbf{x}}}\alpha(I(\mathbf{x}_k) - i)\alpha(I^*(\mathbf{x}_k) - j)\end{aligned} \quad (19)$$

where $\alpha(u) = 1$ if and only if $u = 0$. From this, the probability density function of I^* is given by:

$$p_{I^*}(j) = \sum_i p_{II^*}(i,j). \quad (20)$$

Finally, the difference function is given by:

$$SCV = \sum_{i=1}^{N_{\mathbf{x}}}\left[\hat{I}(\mathbf{x}_i) - I(\mathbf{x}_i)\right]^2. \quad (21)$$

The algorithm described earlier on is not impacted by the choice of this registration function since the only difference between the SSD and the SCV versions is the need to compute the adapted template $\hat{I}$ at each new frame in the sequence to replace I^* in the equations.

IV. Mutual information

The SCV is a good compromise when trying to perform visual tracking on scenes where the variations of illumination are global. Nevertheless, as it is not invariant to local changes, the common scene perturbations that are for example occlusions and specularities cause the tracking process to fail. To handle that, the tracking algorithm has been adapted to use the MI as its registration function to insure greater robustness. To do that, let us first introduce the mutual information and then redefine the optimization process with relation to the new parameters and adapted to the use of several templates in the same algorithm.

1) Registration function: The mutual information, as defined by Shannon [16], represents the quantity of information shared by two signals. It is not a difference based on intensities like the SSD and SCV but a similarity criterion based on the entropies of the considered sources:

$$\mathrm{MI}(I, I^*) = \mathrm{H}(I) + \mathrm{H}(I^*) - \mathrm{H}(I, I^*). \quad (22)$$

The entropy $\mathrm{H}(I)$ is a measure of the randomness of a random variable. Given a discrete variable I with a dynamic d, its entropy is given by the following equation:

$$\mathrm{H}(I) = -\sum_{r=0}^{d} p_I(r)\log\left(p_I(r)\right) \quad (23)$$

where $p_I(r)$ represents the probability distribution function of I (the probability for a given pixel of I to have an intensity r). Following the same principle, the joint entropy $\mathrm{H}(I, I^*)$ of two sources I and I^* is defined by:

$$\mathrm{H}(I, I^*) = -\sum_{r,t=0}^{d} p_{II^*}(r,t)\log\left(p_{II^*}(r,t)\right) \quad (24)$$

where $p_{II^*}(r,t)$ is the joint probability distibution function of I and I^*.

2) Integration into the multiplane tracking algorithm: In the task at hand, the two considered random variables I and I^* are the chosen template and current view as defined in section III-B. The equation of the mutual information (22) therefore becomes:

$$\begin{aligned}\mathrm{MI}(\Delta\mathbf{r}) &= \mathrm{MI}(\bigcup_l w_l(I, \mathbf{r}), \bigcup_l w_l(I^*, \Delta\mathbf{r})) \\ &= \mathrm{H}(\bigcup_l w_l(I, \mathbf{r})) + \mathrm{H}(\bigcup_l w_l(I^*, \Delta\mathbf{r})) \\ &\quad -\mathrm{H}(\bigcup_l w_l(I, \mathbf{r}), \bigcup_l w_l(I^*, \Delta\mathbf{r}))\end{aligned} \quad (25)$$

where $\text{MI}(\bigcup_l w_l(I,\mathbf{r}), \bigcup_l w_l(I^*,\Delta\mathbf{r}))$ is the the mutual information computed on the union of all followed planes projected in the image. Once the different entropies are developped, the equation can be simplified to (see [8] for more details):

$$\text{MI}(\Delta\mathbf{r}) = \sum_{r,t=0}^{d} p_{II^*}(r,t,\Delta\mathbf{r}) \log\left(\frac{p_{II^*}(r,t,\Delta\mathbf{r})}{p_I(r)p_{I^*}(t,\Delta\mathbf{r})}\right). \quad (26)$$

To compute the needed probabilities, histogram binning is necessary to insure the derivability of the equation. This also permits a smoother cost function, enhancing the optimization process, and a faster computation time. First, the image is scaled from its original dynamic d (usually 256 for grey level images) to a chosen number of bins Nc:

$$\overline{I}(\mathbf{x}) = I(\mathbf{x})\frac{(Nc-1)}{d-1} \quad (27)$$

then the probabilities are computed using a kernel function. Several kernel function were discussed in [8], and third order B-splines were chosen [12], [17]:

$$p_{I^*}(t,\Delta\mathbf{r}) = \frac{1}{N_\mathbf{x}}\sum_{i=0}^{N_\mathbf{x}} \phi\left(t-\overline{I^*}(w_l(\mathbf{x}_i,\Delta\mathbf{r}))\right) \quad (28)$$

$$p_I(r) = \frac{1}{N_\mathbf{x}}\sum_{i=0}^{N_\mathbf{x}} \phi\left(r-\overline{I}(w_l(\mathbf{x}_i,\mathbf{r}))\right) \quad (29)$$

$$p_{II^*}(r,t,\Delta\mathbf{r}) = \frac{1}{N_\mathbf{x}}\sum_{i=0}^{N_\mathbf{x}} \phi\left(r-\overline{I}(\mathbf{x}_i,\mathbf{r})\right)\phi\left(t-\overline{I^*}(w_l(\mathbf{x}_i,\Delta\mathbf{r}))\right)$$

where ϕ is the third order B-spline function. Let us note that from this point, once the probabilities are computed from the templates, the development of the method is similar to what was done in [8] up to the expression of the image Jacobian which represents the derivation of the image with repesct to the pose parameters $\mathbf{r}$. The main difference between the two first registration functions and this one is that, being a similarity function, the MI must be maximized to an unknwon value instead of minimized to zero:

$$\widehat{\Delta\mathbf{r}} = \arg\max_{\Delta\mathbf{r}} \text{MI}\left(I(\mathbf{x}),\overline{I^*}(w_l(\mathbf{x},\Delta\mathbf{r}))\right) \quad (30)$$

which means that the optimization process will be different. In this paper, the tracking task will be performed as in [8], by minimizing the Jacobian of the MI:

$$\widehat{\Delta\mathbf{r}} = -\mathbf{H}_{MI}^{-1}\mathbf{G}_{MI}^\top \quad (31)$$

the Jacobian and the Hessian of the MI being defined as in the following equation:

$$\mathbf{G}_{MI} = \frac{\partial \text{MI}(w_l(I^*,\Delta\mathbf{r}),w_l(I,\mathbf{r}))}{\partial\Delta\mathbf{r}} \quad (32)$$

$$\mathbf{H}_{MI} = \frac{\partial^2 \text{MI}(w_l(I^*,\Delta\mathbf{r}),w_l(I,\mathbf{r}))}{\partial\Delta\mathbf{r}^2}. \quad (33)$$

The probability density functions being derivable thanks to the B-spline binning, those matrices can be expressed as:

$$\mathbf{G}_{MI} = \sum_{r,t}\frac{\partial p_{II^*}}{\partial\Delta\mathbf{r}}\left(1+\log\left(\frac{p_{II^*}}{p_{I^*}}\right)\right) \quad (34)$$

$$\mathbf{H}_{MI} = \sum_{r,t}\frac{\partial p_{II^*}}{\partial\Delta\mathbf{r}}^\top\frac{\partial p_{II^*}}{\partial\Delta\mathbf{r}}\left(\frac{1}{p_{II^*}}-\frac{1}{p_{I^*}}\right) + \frac{\partial^2 p_{II^*}}{\partial\Delta\mathbf{r}^2}\left(1+\log\frac{p_{II^*}}{p_{I^*}}\right) \quad (35)$$

with:

$$\frac{\partial p_{II^*}}{\partial\Delta\mathbf{r}} = \frac{1}{N_\mathbf{x}}\sum_{i=0}^{N_\mathbf{x}}\phi\left(r-\overline{I}(w_l(\mathbf{x}_i,\mathbf{r}))\right)\frac{\partial\phi\left(t-\overline{I^*}(w_l(\mathbf{x}_i,\Delta\mathbf{r}))\right)}{\partial\Delta\mathbf{r}}$$

$$\frac{\partial^2 p_{II^*}}{\partial\Delta\mathbf{r}^2} = \frac{1}{N_\mathbf{x}}\sum_{i=0}^{N_\mathbf{x}}\phi\left(r-\overline{I}(w_l(\mathbf{x}_i,\mathbf{r}))\right)\frac{\partial^2\phi\left(t-\overline{I^*}(w_l(\mathbf{x}_i,\Delta\mathbf{r}))\right)}{\partial\Delta\mathbf{r}^2}.$$

The derivatives of the B-spline are given by:

$$\frac{\partial\phi(t-\overline{I^*}(w_l(\mathbf{x}_i,\Delta\mathbf{r})))}{\partial\Delta\mathbf{r}} = -\frac{\partial\phi}{\partial t}\frac{\partial\overline{I^*}}{\partial\Delta\mathbf{r}}$$

$$\frac{\partial^2\phi\left(t-\overline{I^*}(w_l(\mathbf{x}_i,\Delta\mathbf{r}))\right)}{\partial\Delta\mathbf{r}^2} \simeq \frac{\partial^2\phi}{\partial r^2}\frac{\partial\overline{I^*}}{\partial\Delta\mathbf{r}}^\top\frac{\partial\overline{I^*}}{\partial\Delta\mathbf{r}}$$

where:

$$\frac{\partial\overline{I^*}}{\partial\Delta\mathbf{r}} = \mathbf{J}_{I^*}\mathbf{J}_{w_l}\mathbf{J}_\mathbf{K}\mathbf{J}_\mathbf{T}\mathbf{J}_\mathbf{x}(\mathbf{0}).$$

V. Experimental results

Several experiments have been realized to validate the proposed approaches. The methodology is the same for every registration function. First, an original pose is computed from the first image of the considered sequence by matching four points in the image with their 3D correspondances. That pose can afterwards be easily updated with the results of the tracking process. Each tracked plane is then projected thanks to the initial pose an the algorithm can be initialized. From that point the tracking can be launched on the image sequence. Some optimizations have been implemented into the tracker. First, the three algorithms follow a pyramidal scheme to increase their efficiency and robustness towards important displacements. The trackers based on the SSD and SCV also use M-estimators [10], [13] to prevent outliers from perturbing the optimization process.

A. Empirical convergence analysis

A first experiment was realized to analyse the convergence domain of each approach in different conditions. Once the tracker has been initialised with the first frame of the sequence, it is started from an image in the sequence and the pose parameters are set to the corresponding ground truth. The parameters are then perturbed with white Gaussian noise on the pose of chosen σ and the tracking is launched. After the tracking is over, the resulting pose parameters and the ground truth are compared and if the error is small enough, the tracking is considered successfull. The process is repeted 500 times for each method in each situation. The results are shown on figure 2. When adding noise to the

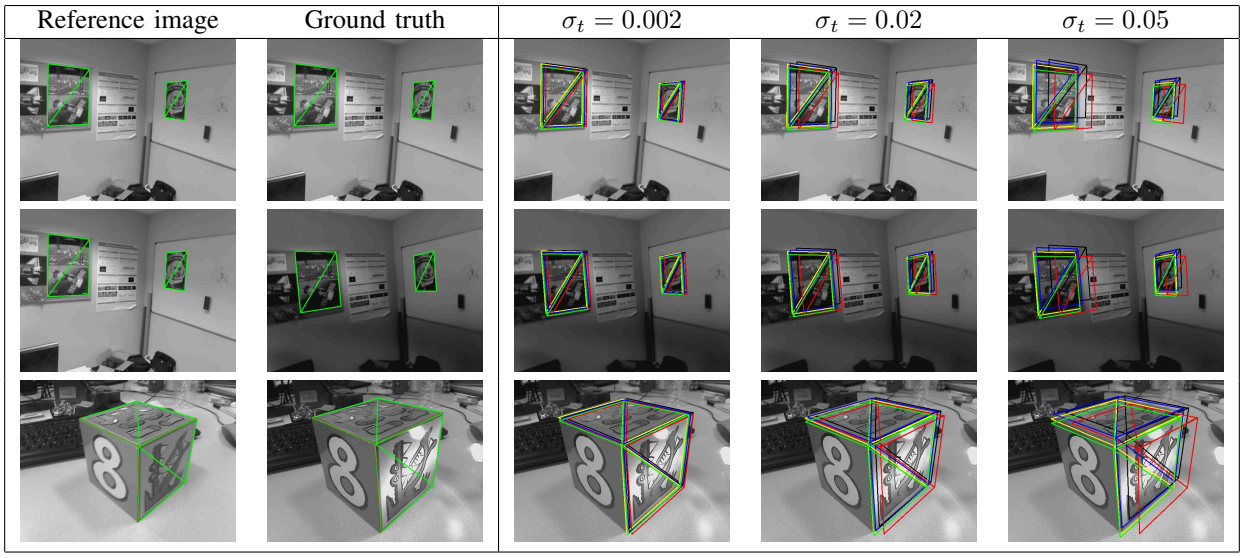


Fig. 1. Examples of starting positions with different σ_t. For all experiments $\sigma_R = 0.01$.

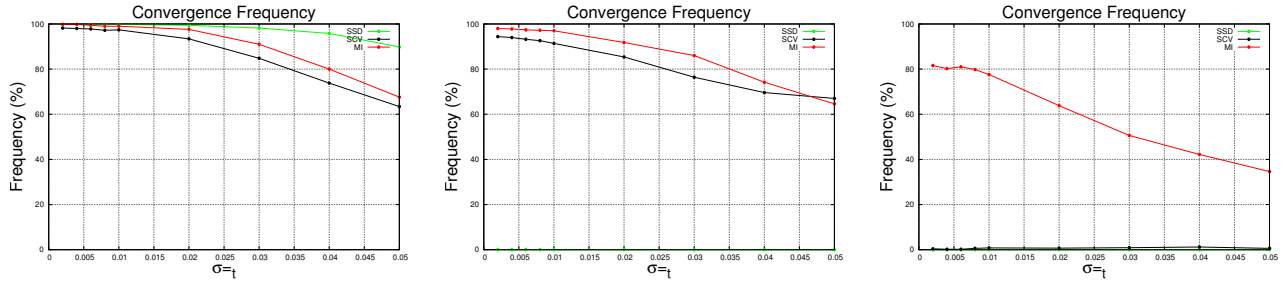


Fig. 2. From left to right : convergence frequency in nominal conditions (row 1 of fig. 1), in light varrying conditions (row 2 of fig. 1) and when confornted to sepcularities (row 3 of fig. 1).

pose parameters, a σ_R of 0.01 rad is chosen for the rotations since their impact is very important on the pose and σ_t is chosen in a range from 0.002 to 0.05 m to see the impact of the starting position (see figure 1 for examples of starting positions). The curves show that mutual information is the only possible solution when confronted to important scene variations such as large occlusions or specularities. They also show that in nominal conditions the convergence domain of both SCV and MI, whilst not being as important as the SSD, is very wide which show that both registration functions are suited for tracking purposes in that case.

B. Robustness towards illumination variations

An experiment was realized to compare the three algorithms in conditions where the illumination of the scene is varying non-linearly. This was done by turning the light of a room on and off. The sequence also contains importantly blured images due to fast camera motion. The results are shown on figure 4. The SSD method fails at the first illumination variation, which was to be expected since the template is no longer a good reference for the current frame. The SCV, adapting the template before each tracking iteration succeeded on the sequence, up to a point where the combination of blurry images and illumination variations is too important and the tracking fails. As for the MI, it reaches the end of the sequence without problems, even where the SCV failed. Let us notice that both the SCV and MI methods are impacted on images where the displacement is too brutal

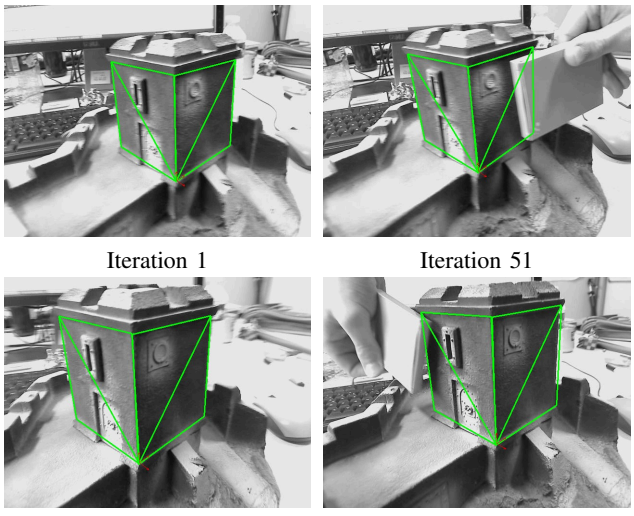
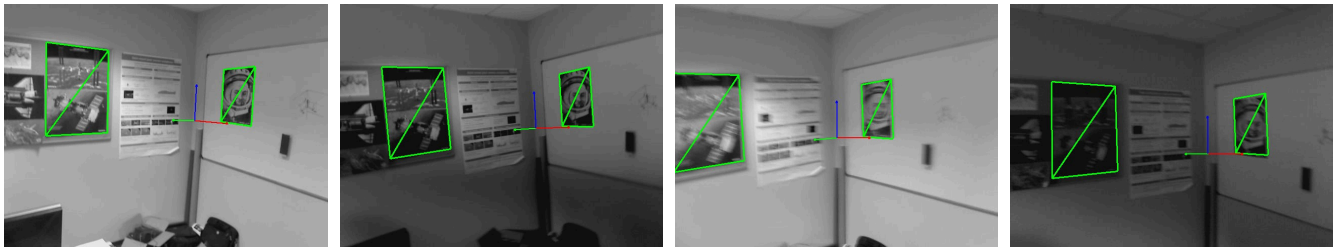


Fig. 3. Tracking results on a scene with big occlusions. The MI is a little bit impacted on iteration 210 when the occluding object is taken out from the template area but the camera pose stays correct (see attached video).



Iteration 1 Iteration 13 Iteration 84 Iteration 119

Fig. 4. Tracking results on a scene with illumination variations using the MI based algorithm. The SCV tracks the template up to iteration 119 whereas the MI is impacted but recovers at the next frame and goes without problems to the end of the sequence (see attached video).

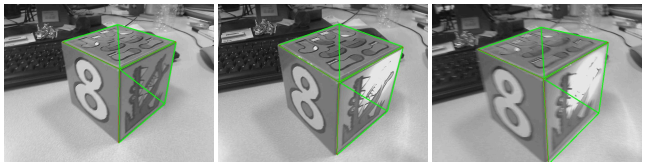
and the image too blurry, but recover the exact position of the template on the next correct images.

C. Robustness towards scene occlusions

The next experimentation was done on a scene which contained large occlusions. The results for the MI version of the algoritm can be seen on figure 3. As expected both the SSD and SCV methods, although M-estimators were used in the estimation process, failed to track the templates since it was occluded and the tracking failed. But as far as the MI is concerned, the tracking process coped well with the situation and, even if it was a little bit impacted during the occlusions, it recovered immediatly after and went to the end of the sequence without any problem.

D. Robustness towards specularities

Finally, a test was realized on a scene where specularities impact the followed planes. The results are shown on figure 5. The specularities were created by pointing a light source on a reflective object, creating big white areas and reflections. Again, the SSD based method was not able to track the patch correctly and was lost. The SCV based method also failed, since the change in illumination was not a global one and could not be taken into account properly. But again, the MI based algorithm realized a good tracking process without any problems as it is robust in those conditions.



Iteration 1 Iteration 64 Iteration 162

Fig. 5. Tracking results on a scene with specularities. The tracking is not impacted even with a reflection and a big white area. It is impacted on iteration 162 due to the combination of both specularity and important blur but recovers immediatly afterwards and continues untill the end of the sequence without problems (see attached video).

VI. Conclusion

This paper introduces a new way to use two robust registration functions for visual tracking. It is a template based differential tracking process that can follow piecewise planar scenes and keep a space-time coherency of the followed templates while directly estimating the camera pose. It is shown to be robust in importantly perturbed conditions. Moreover, it gives a good way of obtaining a pose estimation at each frame without any additionnal computation, hence insuring a better precision. The method could be extended to non-rigid registration processes or future works could allow a detection of the scene geometry to automatically adapt it to any environment without any "a priori".

References

[1] S. Baker, I. Matthews. Equivalence and efficiency of image alignment algorithms. In *CVPR'01*, pages 1090 – 1097, December 2001.

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

[3] S. Benhimane and E. Malis. Integration of Euclidean constraints in template-based visual tracking of piecewise-planar scenes. In *IEEE/RSJ IROS*, 2006.

[4] G. Caron, A. Dame, E. Marchand. L'information mutuelle pour l'estimation visuelle directe de pose. In *18e congrès francophone AFRIF-AFIA RFIA 2012*, Lyon, France, January 2012.

[5] F. Chaumette and S. Hutchinson. Visual servo control, Part I: Basic approaches. *IEEE Robotics and Automation Magazine*, 13(4):82–90, December 2006.

[6] A.I. Comport, E. Marchand, M. Pressigout, and F. Chaumette. Real-time markerless tracking for augmented reality: the virtual visual servoing framework. *IEEE TVCG*, 12(4):615–628, July 2006.

[7] A. Dame, E. Marchand. Accurate real-time tracking using mutual information. In *IEEE ISMAR'10*, pages 47–56, Seoul, Korea, October 2010.

[8] A. Dame, E. Marchand. Second order optimization of mutual information for real-time image registration. *IEEE Trans. on Image Processing*, 21(9):4190–4203, September 2012.

[9] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE PAMI*, 20(10):1025–1039, October 1998.

[10] P.-J. Huber. *Robust Statistics*. Wiler, New York, 1981.

[11] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *JCAI'81*, pages 674–679, 1981.

[12] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE trans. on Medical Imaging*, 16(2):187–198, 1997.

[13] E. Malis and E. Marchand. Experiments with robust estimation techniques in real-time robot vision. In *IEEE/RSJ IROS'06*, pages 223–228, Beijing, China, October 2006.

[14] R. Richa, R. Sznitman, R. Taylor, and G. Hager. Visual tracking using the sum of conditional variance. In *IEEE IROS'11*, pages 2953–2958, San Francisco, September 2011.

[15] G. Scandaroli, M. Meilland, and R. Richa. Improving ncc-based direct visual tracking. In *ECCV'12*, pages 442–455, 2012.

[16] C.E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001.

[17] P. Thévenaz and M. Unser. Optimization of Mutual Information for Multiresolution Image Registration. *IEEE Trans. on Image Processing*, 9(12):2083–2099, 2000.