



Second order optimization of mutual information for real-time image registration

Amaury Dame, Eric Marchand

► To cite this version:

Amaury Dame, Eric Marchand. Second order optimization of mutual information for real-time image registration. IEEE Transactions on Image Processing, 2012, 21 (9), pp.4190-4203. 10.1109/TIP.2012.2199124 . hal-00750528

HAL Id: hal-00750528

<https://inria.hal.science/hal-00750528>

Submitted on 12 Nov 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Second order optimization of mutual information for real-time image registration

Amaury Dame, Eric Marchand

Abstract—In this paper we present a direct image registration approach that uses Mutual Information (MI) as a metric for alignment. The proposed approach is robust, real-time and gives an accurate estimation of a set of 2D motion parameters. MI is a measure of the quantity of information shared by signals. Although it has the ability to perform robust alignment with illumination changes, multi-modality and partial occlusions, few works propose MI-based applications related to spatio-temporal image registration or object tracking in image sequences due to some optimization problems that we will explain.

In this work, we propose a new optimization method that is adapted to the MI cost function and gives a practical solution for real time tracking. We show that by refining the computation of the Hessian matrix and using a specific optimization approach, the registration results are far more robust and accurate than the existing solutions while the computation is cheaper. A new approach is also proposed to speed up the computation of the derivatives and keep an equivalent optimization efficiency. To validate the advantages of the proposed approach, several experiments are performed.

Index Terms—Mutual information, registration, tracking, optimization.

I. INTRODUCTION

Image registration goal is to geometrically align two images acquired at different time and from different camera viewpoints [5], [40]. Considering a displacement model, this alignment process requires the optimization of a similarity measure. Various registration problems can be considered. First, one can consider the case where a wide baseline between two viewpoints is available. In this case, most of the approaches consist of the following steps: features or landmarks detection, features matching, displacement/transformation estimation. Possible applications for such registration methods include stereo-mapping to recover depth from disparities, remote sensing, mosaicing of a large area, medical image registration, etc. A good survey of such techniques is proposed in [40]. The second group of registration problem, also known as tracking, considers image sequence analysis where images differ only slightly and assumptions about smooth changes are justified. Although wide baseline registration techniques still apply, since a continuous motion is assumed from frame-to-frame, other methods can be proposed and only a small

increment of the transformation parameters has then to be estimated. Possible applications are motion estimation, video mosaicing, augmented reality, etc.

In this paper, we shall consider only registration in image sequences. Such approach which can be seen as a 2D motion estimation issue is also often referred as direct tracking or region tracking methods. Major difficulties in such a registration process are image noise, illumination changes and occlusions. Along with robustness to such perturbations, our motivation is to focus on registration and tracking considering different sensor modalities (eg, infra-red and visible images). The choice of a robust similarity measure is then fundamental. In this paper a process based on mutual information [32], [39] is proposed.

Most of the available direct tracking techniques can be divided into two main classes: feature-based and model-based registration method. The former approach focuses on tracking 2D features such as geometrical primitives (point, segments, circles, etc.) or object contours (such as active contours). The latter explicitly uses a model of the scene. This model can be a 3D model leading to a pose estimation process defined as a registration between the measures in the image and the forward projection of the 3D model [15][11]. One can also consider 2D models. Within this category, the features to be tracked can be represented by a descriptor. These descriptors can be image histograms leading to mean shift like approaches [10] or point neighborhood leading to keypoint tracking by matching approaches [23][21]. Following a statistical approach, [4] proposes an approach that merges both a level set approach and histogram-based approach to solve the registration problem. While very robust, these approaches are nevertheless not prone to the estimation of complex movements. In this attempt, it is possible to consider that the 2D model is a reference image (or a template). In that case, the goal is to estimate the motion (or warp) between the current image and a reference template. An example of such approaches are differential image registration methods such as the KLT [24] or its sequels [28][17][1][2][20][3]. Those approaches are not limited to 2D motion estimation, considering for example the motion of a planar object in the image, it is indeed possible to retrieve its 3D motion. The approach described in this paper is related to this last category of registration methods.

In the context of “KLT-like approaches”, a measure of the alignment between the reference image and the current image and its derivatives with respect to the motion (warp) parameters is used within a non-linear estimation process to estimate the current object motion. What seems to be a well adapted measure is the standard Sum of Squared Differences

Copyright (c) 2012 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Amaury Dame was with CNRS, IRISA, INRIA Rennes-Bretagne Atlantique. He is now with the Active Vision Group, Department of Engineering Science, University of Oxford.

Eric Marchand is with Université de Rennes 1, IRISA, INRIA Rennes-Bretagne Atlantique E-mail: Eric.Marchand@irisa.fr

(SSD) function [24][1]. But such approaches are not effective in the case of illumination changes and occlusions. Several solutions have been proposed to add robustness toward those variations. Some include the use of M-estimators to deal with occlusions [28][17] or add new parameters to estimate the illumination variations [17][33]. One can also consider local normalized cross correlation (NCC) [18] or ZNCC to replace SSD.

In this paper, our goal is: first to have an image registration approach that is robust to environmental variations and second, that can handle multi-modality. The proposed solution is then no longer to minimize the SSD but to consider a more robust alignment function, the Mutual Information (MI) between the reference image and the current image, that is defined by the information shared by them and maximize it. MI has been introduced in the context of information theory by Shannon [32]. It has been later considered as an image similarity measure back in the mid ninety's independently by Collignon [9] for tomographic image registration, Studholme [34] for MR and CT image, and by Viola [38] for projection image. Since then MI has become a classical similarity measure especially for multi-modal registration techniques [30] (eg, for medical or remote sensing applications). MI has proved to be robust to occlusions and illumination variations and, therefore, it can be also considered as a good alignment measure for image tracking. Nevertheless, to date only few works feature mutual information within a frame-to-frame tracking method [8], [13], [14], [29]. To be used efficiently within a template tracking algorithm (such as [2]), an optimization technique has to be considered. Various approaches have been proposed: first-order gradient descent [39], multi-resolution hill climbing algorithm [34] or simulated annealing techniques [31]. Powell's [9], [25] or Simplex [27], [8] methods (which do not require function derivatives to be analytically expressed) have been very popular in MI optimization but the former is sensitive to local optima in the registration criterion whereas the latter is known to be computationally inefficient. Considering that MI computation is evaluated from the joint image intensity histogram, an analytic derivative of the mutual information is difficult to obtain. In order to compute MI derivatives, [26] introduces partial volume interpolation for the construction of the joint histogram leading to an analytic computation of MI gradients. In [36], the authors formulate the mutual-information criterion as a continuous and differentiable function of the registration parameters using B-Spline Parzen windows. These derivatives are then used in a Levenberg-Marquardt like minimization method. Such a formulation has been considered within a motion estimation process (or camera tracking) in [13], [14], [29]. However the existing approaches are not taking full advantage of the accuracy of MI nor are they appropriate for real-time applications especially if a complex motion model is considered.

In this paper we present a MI-based image alignment. An important contribution is to propose an optimization process adapted to the MI cost function. We propose an inverse compositional optimization approach [2] where an important part of the required derivatives can be precomputed, resulting in small computation times. A precise, complete and efficient

computation of the Hessian matrix is described. We show that the inverse compositional approach allows the estimation of the Hessian matrix after convergence that can be used in a Newton's like approach to give an accurate and fast estimation of the displacement parameters. Finally a new approach is proposed to speed up the computation of the derivatives through a selection of the reference pixels making the image alignment process possible at video-rate.

In the remainder of this paper, Section II presents an overview of the differential image registration approaches. In section III, a brief introduction on information theory is given with the definition of mutual information, then a formulation adapted to the differential alignment method is presented. Section IV deals with the optimization of the resulting mutual information function with respect to the motion parameters to estimate. Finally section V presents several image registration or template tracking experiments including the Metaio benchmark and demonstrates the multi-modal capability of the approach.

II. DIFFERENTIAL TEMPLATE-BASED IMAGE REGISTRATION

Differential image alignment [2] (or template tracking) is a class of approaches based on the optimization of an image registration function.

The goal is to estimate the displacement \mathbf{p} of an image template I^* in a sequence of images. Considering a frame-to-frame tracking process, the template I^* is usually a region of interest extracted from the very first image of the sequence. In the case of a similarity function f , the problem can be written as :

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

where we search the displacement $\hat{\mathbf{p}}_t$ that maximizes the similarity between the template I^* and the warped current image I_t . In the case of a dissimilarity function the problem would be simply inverted in the sense that we would search the minimum of the function f . For the purpose of clarity, the warping function w is here used in an abuse of notation to define the overall transformation of the image I by the parameters \mathbf{p} . Afterwards, its proper formulation will be preferred using $w(\mathbf{x}, \mathbf{p})$ to denote the position of the point \mathbf{x} transformed using the parameter \mathbf{p} .

The displacement parameters \mathbf{p} can be of high dimension. For instance, the experiments that will be presented at the end of the paper consider a homography transformation that corresponds to $\mathbf{p} \in \mathfrak{sl}(3)$ that is 8 parameters. Approaches such as an exhaustive search of $\hat{\mathbf{p}}$ are thus too expensive if not impossible.

To solve the maximization problem, the assumption made in the differential image registration approaches is that the displacement of the object between two consecutive frames is quite small. The previous estimated displacement $\hat{\mathbf{p}}_{t-1}$ can therefore be used as first estimation of the current displacement to perform the optimization of f and incrementally reach the best estimation $\hat{\mathbf{p}}_t$.

Multiple solutions exists to compute the update of the current displacement parameters and perform the optimization. Baker and Matthews showed that two formulations were equivalent [1] depending on whether the update is acting on the current image or the reference. The former is the direct compositional formulation which considers that the update is applied to the current image, thus we search the update $\Delta \mathbf{p}$ that maximize f as:

$$\Delta \mathbf{p}^k = \arg \max_{\Delta \mathbf{p}} f(I^*, w(w(I_t, \Delta \mathbf{p}), \mathbf{p}^k)). \quad (2)$$

This equation is typically solved using a Taylor expansion where the update is computed with the function derivatives with respect to $\Delta \mathbf{p}$. For a pixel \mathbf{x} , the update of the current parameters \mathbf{p}^k is then applied as follows:

$$w(w(\mathbf{x}, \Delta \mathbf{p}), \mathbf{p}^k) \rightarrow w(\mathbf{x}, \mathbf{p}^{k+1}). \quad (3)$$

A second equivalent formulation is the inverse compositional formulation which considers that the update modifies the reference image, so that $\Delta \mathbf{p}$ is chosen to maximize:

$$\Delta \mathbf{p}^k = \arg \max_{\Delta \mathbf{p}} f(w(I^*, \Delta \mathbf{p}), w(I_t, \mathbf{p}^k)). \quad (4)$$

In this case the current parameters will be updated using:

$$w(w^{-1}(\mathbf{x}, \Delta \mathbf{p}^k), \mathbf{p}^k) \rightarrow w(\mathbf{x}, \mathbf{p}^{k+1}). \quad (5)$$

In the inverse compositional formulation, since the update parameters are applied to the reference image, the derivatives with respect to the displacement parameters are computed using the gradient of the reference image. Thus, these derivatives can be partially precomputed and the algorithm is far less time consuming. Since we are interested in a fast estimation of the displacement parameters, the remainder of the paper will focus on the later inverse compositional approach.

One essential choice remains the one of the alignment function f . One natural solution is to choose the function f as the sum of squared differences (SSD) of the pixel intensities between the reference image and the transformed current image:

$$\hat{\mathbf{p}}_t = \arg \min_{\mathbf{p}} (SSD(I^*, w(I_t, \mathbf{p}))) \quad (6)$$

$$= \arg \min_{\mathbf{p}} \sum_{\mathbf{x} \in \mathcal{W}} (I^*(\mathbf{x}) - I_t(w(\mathbf{x}, \mathbf{p})))^2 \quad (7)$$

where the summation is computed on each point \mathbf{x} of the reference template that is the region of interest (\mathcal{W}) of the reference image. As suggested by its definition, this dissimilarity function is very sensitive to occlusions and illumination variations. Many solutions have been proposed to robustify the SSD. M-estimators can robustify the least squared problem toward occlusions [17], while a model of illumination changes can be coupled with the motion model to create a tracker robust to lighting changes [33]. Nevertheless those solutions are complex since additional parameters have to be estimated and aligning two images acquired using different modalities of acquisition remains impossible.

Let us for example consider an aerial image and a map template (see figure 1(a)). Considering these two modalities is obviously an extreme case, but it will emphasize the robustness

of the MI with respect to other similarity function. The value of SSD is computed with respect to the translations between the map and the satellite image. The two images are showing the same place (at least for a human eye they contain the same “information”), however, since the link between the intensities of the pixels is not linear, the SSD function represented in figure 1(b) gives no information on the alignment between the two images. Since NCC and ZNCC have shown some very good results in multi-modal alignment problems [18], we also evaluated the ZNCC efficiency in that matter. We can see in figure 1(c) that the case is too extreme and that ZNCC has also no significant optimum and therefore can not be used in this registration task.

To deal with occlusions, illumination variations and multi-modality, we propose to define our alignment function as the mutual information [32], [39]. Originating from the information theory, MI is a measure of statistical dependency between two signals (or two images in our case) that is, as we will see, robust to all this variations of appearance.

III. MUTUAL INFORMATION

Rather than comparing intensities, mutual information is the quantity of information shared between two random variables. Mutual information of two random variable I and I^* is then given by the following equation[32]:

$$MI(I, I^*) = h(I) + h(I^*) - h(I, I^*). \quad (8)$$

where the entropy $h(I)$ is a measure of variability of a random variable I (signal, image...). If r are the possible values of I and $p_I(r) = P(I = r)$ is the probability distribution function of r , then the Shannon entropy $h(I)$ of a discrete variable I is given by the following expression:

$$h(I) = - \sum_r p_I(r) \log(p_I(r)). \quad (9)$$

The probability distribution function of the gray-level values is then simply given by the normalized histogram of the image I . The entropy can therefore be considered as a dispersion measure of the image histogram.

Following the same principle, the joint entropy $h(I, I^*)$ of two random variables I and I^* can be defined as the variability of the couple of variables (I, I^*) . The Shannon joint entropy expression is given by:

$$h(I, I^*) = - \sum_{r,t} p_{II^*}(r,t) \log(p_{II^*}(r,t)) \quad (10)$$

where r and t are respectively the possible values of the variables I and I^* , and $p_{II^*}(r,t) = P(I = r \cap I^* = t)$ is the joint probability distribution function. In our problem I and I^* are images. Then r and t are the gray-level values of the two images and the joint probability distribution function is a normalized bi-dimensional histogram of the two images. As for entropy, joint entropy corresponds to a dispersion measure of the joint histogram of (I, I^*) . If this expression is combined with the previously defined differential motion estimation problem, we can consider that the image I is depending on the displacement parameters \mathbf{p} . If we use the same warp function

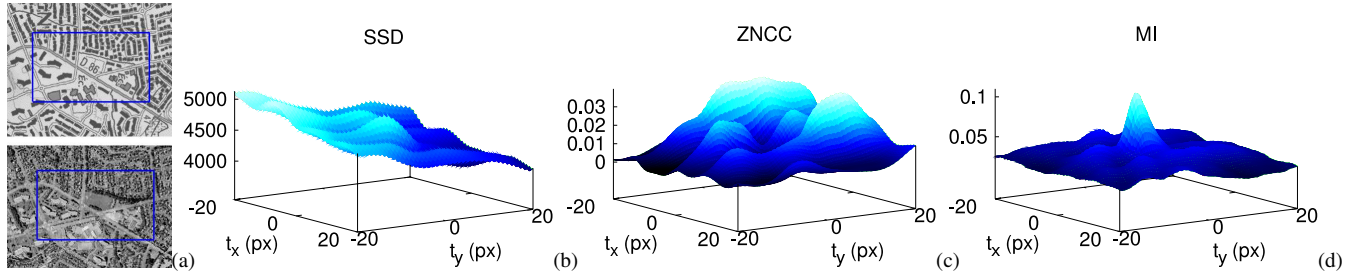


Fig. 1. Alignment functions wrt. translations between two images from the same area: (a) aerial image and the map reference. MI shows a maximum near zero translation at the alignment position whereas SSD and ZNCC gives no clear information on the alignment quality.

notation as in section II, the mutual information can thus be written with respect to \mathbf{p} :

$$\begin{aligned} MI(\mathbf{p}) &= MI(w(I, \mathbf{p}), I^*) \\ &= h(w(I, \mathbf{p})) + h(I^*) - h(w(I, \mathbf{p}), I^*). \end{aligned} \quad (11)$$

The final expression of MI is obtained by developing the previous equation using the entropy equations (9) and (10):

$$MI(\mathbf{p}) = \sum_{r,t} p_{II^*}(r, t, \mathbf{p}) \log \left(\frac{p_{II^*}(r, t, \mathbf{p})}{p_I(r, \mathbf{p}) p_{I^*}(t)} \right) \quad (12)$$

The analytical formulation of a normalized histogram of an image I^* is classically written as follows:

$$p_{I^*}(t) = \frac{1}{N_x} \sum_{\mathbf{x}} \phi(t - \bar{I}^*(\mathbf{x})) \quad (13)$$

$$p_I(r, \mathbf{p}) = \frac{1}{N_x} \sum_{\mathbf{x}} \phi(r - \bar{I}(w(\mathbf{x}, \mathbf{p}))) \quad (14)$$

$$p_{II^*}(r, t, \mathbf{p}) = \frac{1}{N_x} \sum_{\mathbf{x}} \phi(r - \bar{I}(w(\mathbf{x}, \mathbf{p}))) \phi(t - \bar{I}^*(\mathbf{x}))$$

where \mathbf{x} are the points of the region of interest in the image and N_x is the number of points. r and t are the possible values of $I(\mathbf{x})$ and $I^*(\mathbf{x})$ that are the scaled version of the original images, so that $\{r, t\} \in [0, N_c]^2$. Let us note that to have a smooth mutual information it is important to maintain the number of histogram bins N_c low ($N_c = 8$ in our implementation). In the classical formulation ϕ is a Kronecker's function: $\phi(x) = 1$ for $x = 0$ and $\phi(x) = 0$ otherwise. So that, for instance, each time $\bar{I}^*(\mathbf{x}) = i$ the i th histogram bin value of p_{I^*} is incremented. However this formulation does not take advantage of the decimal part of the scaled intensities, therefore several solutions have been proposed to simultaneously smooth the mutual information function, make its formulation differentiable and keep its accuracy [39][25]. Several approaches propose to use Gaussian function, however, in our approach, we focus on the use of B-splines functions for ϕ [25], [36]. It has indeed been shown that these functions provides a good approximation of Gaussian functions while their computation and the one of their derivatives is cheaper. As it will be discussed later, this also permits to have a smooth, accurate but computationally cheap gradient of MI to perform its optimization.

IV. MUTUAL INFORMATION-BASED MOTION ESTIMATION

In this section we will see how to use the MI cost function with the differential image registration formulation presented

in section II. Once our approach is fully defined, a pseudo-code of the algorithm is given to summarize the proposed method.

A. Overview

The goal of our tracking problem is to align an image template I^* with an input image I . If we assume that the reference template appears in I , the goal is to search for the transformation that aligns the pixels \mathbf{x} of the reference image I^* to the corresponding pixels \mathbf{x}' of I in the sense of our chosen similarity measure. Assuming that the transformation from the reference points to the input image can be modeled by a warp function $\mathbf{x}' = w(\mathbf{x}, \mathbf{p})$, the problem can be formulated as:

$$\hat{\mathbf{p}} = \arg \max_{\mathbf{p}} MI(I^*(\mathbf{x}), I(w(\mathbf{x}, \mathbf{p}))) \quad (15)$$

Since this problem is impossible to solve linearly, a non-linear optimization is performed. To initialize the optimization, a first guess of the displacement parameters is required. Since we suppose that the displacement of the object between two consecutive frames is small, a good approximation is to approximate the parameter \mathbf{p}_t of the input image I_t at a time t using the parameters estimated for the previous frame: $\mathbf{p}_t = \widehat{\mathbf{p}_{t-1}}$.

To initialize the whole tracking approach, the position of the template in the first image I_0 has to be known coarsely. Since the first image of the sequence is usually the one that defines the template I^* , the first displacement parameters \mathbf{p}_0 between the template and the first image simply correspond to an identity transformation (considering the warp functions defined in the first chapter it yields: $\mathbf{p}_0 = \mathbf{0}$). Otherwise, the first estimation can be performed using some matching process, such as a keypoints matching approach [23], [21] or other wide baseline registration method. The first approximation of the displacement $\mathbf{p}_t^0 = \widehat{\mathbf{p}_{t-1}}$ is then refined using the numerical resolution of the equation (15). To solve the maximization, an iterative optimization method is used that successively goes closer and closer to the optimum of the cost function $\hat{\mathbf{p}}_t$ (see Figure 2). For a clarity purpose, let us now consider the maximization peculiar to one image I (we drop the subscript t) and focus on the iteration number noted using the superscript k .

Let us recall that for efficiency issue, we chose to consider an inverse compositional approach. The difference with the forward compositional approach comes from the optimization process where the updating steps from the current guess to

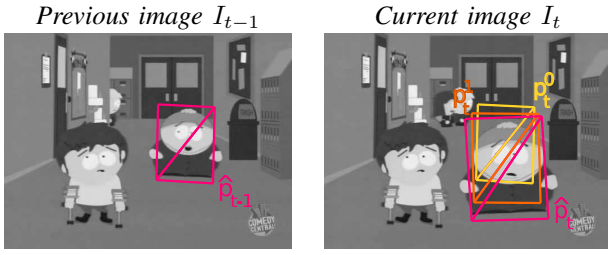


Fig. 2. The first approximation of the position \mathbf{p}_t^0 , given by the previous position $\widehat{\mathbf{p}}_{t-1}$, is iteratively refined to find the optimal parameters $\widehat{\mathbf{p}}_t$.

the optimal displacement parameters are modified. Instead of searching the update parameters that will bring the warped points of the current image into the points of the template image, the formulation of the problem is inverted so that we search the “inverse” update that brings the points of the template image into the warped points of the current image. In the inverse compositional approach [2], the goal is then formulated as finding the update $\Delta \mathbf{p}$ that leads to the optimum, so that, at each iteration k , we seek (see Section II for details):

$$\Delta \mathbf{p}^k = \arg \max_{\Delta \mathbf{p}} MI(I^*(w(\mathbf{x}, \Delta \mathbf{p})), I(w(\mathbf{x}, \mathbf{p}^k))) \quad (16)$$

The optimization using this formulation is similar to the optimization using the forward compositional approach. Nevertheless, since the update is considered to affect the reference image, we will see that more elements of the mutual information derivatives with respect to the update can be precomputed.

B. Derivative function analysis

Let us remind that the goal is to estimate the displacement parameters \mathbf{p}_t that maximizes the MI using a first estimation of the parameters \mathbf{p}_{t-1} and an iterative update of the parameters. In this work we ought to register planar regions through 3D displacements. This problem implies a strong correlation between the elements of the vector \mathbf{p} . Therefore, estimating the update using first-order optimization method such as a steepest gradient descent is not adapted. Such non-linear optimization are usually performed using a Newton’s method that assume the shape of the function to be parabolic.

Newton’s method uses a second order Taylor expansion at the current position \mathbf{p}^{k-1} to estimate the update $\Delta \mathbf{p}$ required to reach the optimum of the function (where the gradient of the function is null). The same estimation and update are performed until the parameter \mathbf{p}^k effectively reaches the optimum. The update is estimated following the equation:

$$\Delta \mathbf{p} = -\mathbf{H}^{-1} \mathbf{G}^\top \quad (17)$$

where \mathbf{G} and \mathbf{H} are respectively the gradient and Hessian matrices of the mutual information with respect to the update $\Delta \mathbf{p}$. Following the inverse compositional formulation defined in equation (4) those matrices are equal to:

$$\mathbf{G} = \frac{\partial MI(w(I^*, \Delta \mathbf{p}), w(I, \mathbf{p}))}{\partial \Delta \mathbf{p}} \quad (18)$$

$$\mathbf{H} = \frac{\partial^2 MI(w(I^*, \Delta \mathbf{p}), w(I, \mathbf{p}))}{\partial \Delta \mathbf{p}^2} \quad (19)$$

Applying the derivative chain rules to equation (12) yields the following gradient and Hessian matrices:

$$\mathbf{G} = \sum_{r,t} \frac{\partial p_{II^*}}{\partial \Delta \mathbf{p}} \left(1 + \log \left(\frac{p_{II^*}}{p_{I^*}} \right) \right) \quad (20)$$

$$\mathbf{H} = \sum_{r,t} \frac{\partial p_{II^*}}{\partial \Delta \mathbf{p}}^\top \frac{\partial p_{II^*}}{\partial \Delta \mathbf{p}} \left(\frac{1}{p_{II^*}} - \frac{1}{p_{I^*}} \right) + \frac{\partial^2 p_{II^*}}{\partial \Delta \mathbf{p}^2} \left(1 + \log \frac{p_{II^*}}{p_{I^*}} \right) \quad (21)$$

For the purpose of clarity, the marginal probabilities and joint probability that are actually depending on r, t, \mathbf{p}^* and $\Delta \mathbf{p}$ are simply denoted as p_I, p_{I^*} and p_{II^*} . The details of the calculation from equation (18) to equation (21) can be found in [13].

By analogy with the Hessian computation in a Gauss-Newton’s method for a least squared problem that is assuming that the neglected term is null after convergence, second order derivatives are usually neglected in the Hessian matrix computation [36], [37], [13], [14] leading to:

$$\mathbf{H} \simeq \sum_{r,t} \frac{\partial p_{II^*}}{\partial \Delta \mathbf{p}}^\top \frac{\partial p_{II^*}}{\partial \Delta \mathbf{p}} \left(\frac{1}{p_{II^*}} - \frac{1}{p_{I^*}} \right). \quad (22)$$

In our approach we compute the Hessian matrix using the second order derivatives. In our point of view, they are required to obtain a precise estimation of the motion. Indeed, let us consider the approximation made in (22). Considering the expression of the marginal probability $p_{I^*}(t) = \sum_r p_{II^*}(r, t)$, it is clear that $p_{I^*}(t) > p_{II^*}(r, t)$ so $1/p_{II^*}(r, t) - 1/p_{I^*}(t) > 0$. Since $\frac{\partial p_{II^*}}{\partial \Delta \mathbf{p}}^\top \frac{\partial p_{II^*}}{\partial \Delta \mathbf{p}}$ is a positive matrix then the final Hessian matrix given by (22) is positive. Since the goal is to maximize MI, the Hessian matrix after convergence is supposed to be negative by definition. The common approximation of (22) is thus not suited for the optimization of MI.

As we can see in equation (20) and equation (21), the derivatives of the mutual information depend on the derivatives of the joint probability. Using the previous definition in (14) and passing the derivative operator through the summation yields the following expressions:

$$\begin{aligned} \frac{\partial p_{II^*}}{\partial \Delta \mathbf{p}} &= \frac{1}{N_x} \sum_x \phi(t - \bar{I}(w(\mathbf{x}, \mathbf{p}))) \frac{\partial \phi(r - \bar{I}^*(w(\mathbf{x}, \Delta \mathbf{p})))}{\partial \Delta \mathbf{p}} \\ \frac{\partial^2 p_{II^*}}{\partial \Delta \mathbf{p}^2} &= \frac{1}{N_x} \sum_x \phi(t - \bar{I}(w(\mathbf{x}, \mathbf{p}))) \frac{\partial^2 \phi(r - \bar{I}^*(w(\mathbf{x}, \Delta \mathbf{p})))}{\partial \Delta \mathbf{p}^2}. \end{aligned} \quad (23)$$

The remaining expressions to evaluate are the variations of the B-spline function ϕ with respect to the update. Their derivatives are obtained using the chain rule leading to:

$$\frac{\partial \phi(r - \bar{I}^*(w(\mathbf{x}, \Delta \mathbf{p})))}{\partial \Delta \mathbf{p}} = -\frac{\partial \phi}{\partial r} \frac{\partial \bar{I}^*}{\partial \Delta \mathbf{p}} \quad (24)$$

$$\frac{\partial^2 \phi(r - \bar{I}^*(w(\mathbf{x}, \Delta \mathbf{p})))}{\partial \Delta \mathbf{p}^2} = \frac{\partial^2 \phi}{\partial r^2} \frac{\partial \bar{I}^*}{\partial \Delta \mathbf{p}}^\top \frac{\partial \bar{I}^*}{\partial \Delta \mathbf{p}} - \frac{\partial \phi}{\partial r} \frac{\partial^2 \bar{I}^*}{\partial \Delta \mathbf{p}^2}. \quad (25)$$

Finally the derivatives of the reference image intensity with respect to the update parameters $\Delta \mathbf{p}$ is given by the following expressions:

$$\frac{\partial \bar{I}^*}{\partial \Delta \mathbf{p}} = \nabla \bar{I}^* \frac{\partial w(\mathbf{x}, \mathbf{p})}{\partial \Delta \mathbf{p}} \quad (26)$$

$$\frac{\partial^2 \bar{I}^*}{\partial \Delta \mathbf{p}^2} = \frac{\partial w}{\partial \Delta \mathbf{p}}^\top \nabla^2 \bar{I}^* \frac{\partial w}{\partial \Delta \mathbf{p}} + \nabla \bar{I}^*_x \frac{\partial^2 w_x}{\partial \Delta \mathbf{p}^2} + \nabla \bar{I}^*_y \frac{\partial^2 w_y}{\partial \Delta \mathbf{p}^2} \quad (27)$$

where $\nabla \bar{I}^*$ are the image gradients of the reference image, obtained using the convolution of a Gaussian filter and a derivative filter, the Gaussian filter allowing for a smoother version of the gradients. The motivation for using the inverse compositional formulation is then obvious. The derivatives of the warp function are all computed at $\Delta \mathbf{p} = 0$, their values are then constant for each pixels of the template. Moreover, since the reference image \bar{I}^* is constant, its gradients and all the expressions from equation (24) to equation (27) are constants and have to be precomputed only one time.

In our work we focus on planar region registration. The warp function is thus defined by the group action $w : \mathbb{SL}(3) \times \mathbb{P}^2$ with $\mathbf{x} \in \mathbb{P}^2$ and \mathbf{p} defines the 8 parameters of the $\mathfrak{sl}(3)$ lie algebra associated to the $\mathbb{SL}(3)$ group. However, this research is not limited to such a warp function thus details will not be given on the warp derivatives. All details regarding the derivatives of the chosen warp function can be found in [3].

Let us emphasize that any kind of warp model can be considered. Although homography have been considered in this paper, it can also be applied on affine motion model [12], pose parameters $\mathbb{SE}(3)$ [29][7] and other motion models. The method could also be extended to non-rigid registration process. In that case, specific local distortions have to be considered. Radial basis functions (such as Wendland's function or thin-plate splines) are able to handle locally varying geometric distortions and can be considered within the proposed framework. In any cases, although the main algorithm will remain unchanged, the only modification will be to redefine the warp derivatives.

C. Optimization approach

The Newton's method that can be used to perform the estimation of the update parameters $\Delta \mathbf{p}$ is based on the assumption of a similarity function with a parabolic shape. One can immediately notice that this assumption can be easily violated by looking at the function's shape where we see that the assumption is correct only near the maximum. Since the violation could cause the Newton's method to fail, a better approach has to be found.

To evaluate the efficiency of the following optimization methods, a set of alignment experiments has been realized. The goal is to estimate the known position \mathbf{p}^* of a template in an image (see figure 4(a)) from many initial position parameters (see figure 4(b)). The initial parameters are automatically generated applying a random noise to the ground truth position.

The convergence rate of the optimization method are then evaluated with respect to the initial positioning error. The positioning error err is defined as the RMS distance between the correct position of some reference points $\mathbf{x}_i^* = w(\mathbf{x}_i, \mathbf{p}^*)$

and the current position of the points $w(\mathbf{x}_i, \mathbf{p})$ [22]. The reference points are simply chosen as the 4 corners of the template so that the error becomes:

$$err(\mathbf{p}) = \sqrt{\sum_{i=1}^4 \|\mathbf{x}_i^* - w(\mathbf{x}_i, \mathbf{p})\|^2} \quad (28)$$

We consider that the optimization converges as soon as the error err is below 0.5 px. 500 alignment experiments are performed for each initial positioning error err from 1 to 20 that is a total of 10000 experiments. As output we retrieve the convergence rate, the average number of iterations required to reach convergence, the final residues after convergence and the computation time of each iteration. Indeed, those values gives a good overview of the efficiency of the optimization methods.

The Gradient descent method cannot estimate an accurate estimation of the homography (see section IV-B). Indeed its use gives a final estimation with an error always above 0.5 px for the all set of experiments (that is a 0% convergence rate). Thus the results have not been included in figure 4.

1) *Newton's method*: Mutual information function is a quasi-concave function, thus the parabolic hypothesis of the Newton's method is only valid near the convergence. As soon as the displacement in the sequence is important, the initial parameters \mathbf{p}_{t-1} would be on the convex part of the cost function that will cause the optimization to diverge.

The problem is in fact equivalent using a SSD function. One example of the values obtained on the estimation of a translational displacement is presented in figure 3 for both the MI function and the minus of the SSD function. For the purpose of clarity, we choose to analyze the minus of the SSD function to deal with a maximization for both functions. The quasi-concave shape of both functions is obvious. The parabolic assumption is only correct for the concave part of the function, that is where their second order derivatives are negative (the area highlighted in purple). Therefore the convergence domain using a classical Newton's method would be very small.

Figure 4(c) shows the convergence results obtained using our set of convergence tests. The convergence domain of the Newton's method is indeed practically very small in the case of the homography estimation. As soon as the initial error exceeds 2 px, the initial parameters are, most of the time, out of the convergence domain of the Newton's method and the convergence rate decreases drastically. Considering the converging experiments, then once the convergence is achieved, the parabolic shape assumption is verified and the method gives good quality estimation with a mean final residue of 0.06 px. However, it is rarely the case and the computation of the Hessian at each iteration makes the process really time consuming (see figure 5 for the time per iteration).

Considering the simple one dimensional example, one could expect an optimization that has a convergence domain as wide as the one of the gradient descent method (the blue area in figure 3).

2) *Conditioning the optimization*: In this section, we show how to combine the convergence domain of the gradient

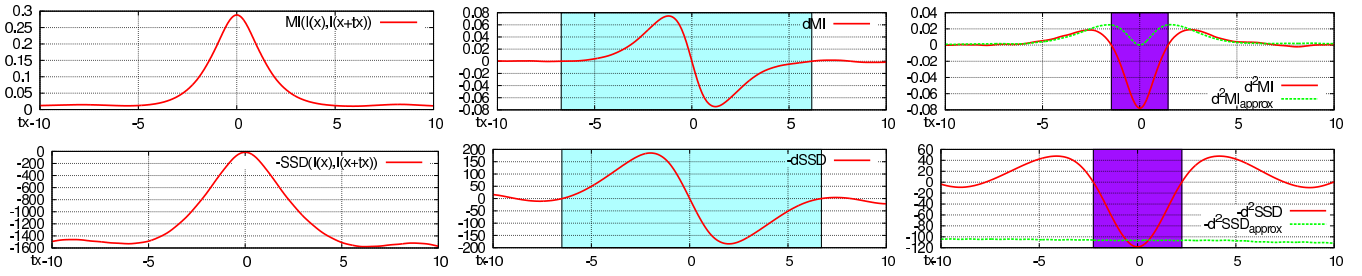


Fig. 3. SSD, MI and their derivatives with respect to one translation (px). The purple area is the convergence domain using a classical Newton's method, the blue one is the convergence domain of a Gradient descent method. The proposed method keeps the wider convergence domain of the gradient's method in blue, while having the convergence properties of the Newton's method near the optimum, allowing an accurate estimation of complex transformations.

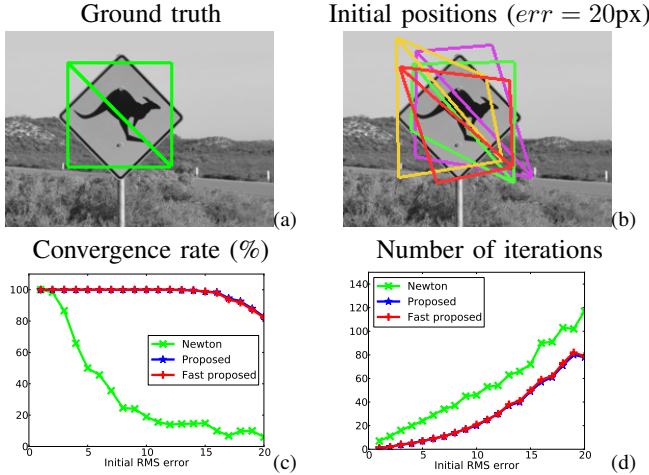


Fig. 4. Empirical convergence analysis of the optimization methods. The proposed methods (blue and green curves) have a very high convergence rate compared to the classical Newton's methods (red curve).

| | Newton | Inverse | Inverse fast |
|---------------------|--------|---------|--------------|
| Time/iteration (ms) | 13.4 | 5.1 | 3.5 |

Fig. 5. Average time in ms required to compute one update $\Delta \mathbf{p}$ of the optimization methods.

descent with the accuracy and efficiency of the Newton's method for the optimization of MI. In registration problems formulated with a SSD function, the Gauss-Newton approximation condition the problem by estimating a Hessian matrix that is always definite positive (see section IV-B and the green curve in figure 3) and that is a good approximation of the exact Hessian matrix after convergence. Therefore its use permits to have a convergence domain as wide as the one with a gradient method (blue area) and a good convergence behavior next to the optimum yielding to accurate estimations.

In the mutual information maximization, the problem is different. Indeed approximating the Hessian matrix as it is proposed in [36], [13], [14] do not gives an estimation of the Hessian matrix after convergence (see the green curve in 3 for the MI function). No approximation on the Hessian of MI simplifies the problem as the Gauss-Newton approach does for the SSD.

The solution that we propose is inspired from the Gauss-Newton approach. The idea remains to use an estimation of the Hessian matrix after convergence. To compute this estimation

we consider that after convergence the alignment between the template and the warped current image is perfect. Therefore we simply assume that at the optimum we have $\bar{I}(w(\mathbf{x}, \mathbf{p})) = \bar{I}^*(\mathbf{x})$.

This solution has several advantages:

- It gives a definite negative Hessian matrix that yields to have a wide convergence domain (blue area in figure 3). We can notice that the resulting convergence domain is as wide as the one of the SSD function in the considered 1D example. In section V-A2, further experiments will show that it is also the case for a homography estimation.
- Since the Hessian matrix used in the Newton's method is the Hessian matrix after convergence, the behavior of the optimization near convergence is optimal and the final estimated displacement parameters are very accurate.
- This approach has the advantage of its computation time. In the classical Newton's method the Hessian and Jacobian are computed for each iterations. In the proposed approach the Hessian matrix is computed one time in the whole experiment.

The proposed optimization has been evaluated on the set of experiment presented in figure 4. As expected, the convergence domain is larger than the one using the classical Newton's method. The optimization converges for all the experiments with an initial error below $16px$ and the convergence rate slightly decreases for $err > 16$. Since we use the Hessian estimated after convergence, then the behavior near the optimum is suited to reach an accurate solution yielding to final residues that have a mean of 0.06 px.

Figure 4(d) shows the number of iterations to reach convergence. The number of iterations with the proposed method is fewer than the one with the classical Newton's method while its computation is much cheaper (see figure 5).

3) *Improving the computation time:* Compared to a simple least squared problem, mutual information can still be considered as a very complex function to compute. The proposed approach offers already a practical solution. Nevertheless, faster performance is sometimes desired.

To compute the MI between the two images, all the information is required, so all the reference pixels must be used to compute the marginal and joint probabilities. As for the variation of the mutual information computation, only the motion of the pixels that are not in a uniform region will have a strong effect. This fact is obvious from equation (26)

and (27). One very simple modification is then to perform the computation of the gradient and Hessian using only a selection of pixels in the template.

A simple measure to determine if a point is in a uniform region of the template is given by the norm of the reference image gradients. Therefore the selection condition can be written as:

$$\|\nabla I^*(\mathbf{x})\| > \alpha \quad (29)$$

where α is a given threshold. The summation in equation (23) is therefore computed on the reference pixels that respect this condition.

The efficiency of the proposed approach has been compared to the previous one using the set of experiments represented in figure 4. Using a threshold $\alpha = 6$, the selected number of points corresponds to 18% of the total number of reference points. We can see on figure 4(c & d) that the convergence rate and the number of required iterations is equal to the ones of the previous method up to few percent and iterations. Curves “proposed” and “fast proposed” are superimposed meaning precision efficiency remains the same (the final residues still has a mean value of 0.06 px) whereas computational efficiency greatly improves: 3.5ms per iteration vs 5.1ms (see figure 5) for the same number of iterations (see figure 4(d)).

In summary, for a similar efficiency, the computation time of the proposed method is 30% smaller (see figure 5). Such a selection method is therefore highly recommended in MI derivatives computation.

To summarize the whole process, a pseudocode of the algorithm is presented in Figure 6 in the case of a classical image registration task.

V. EXPERIMENTAL RESULTS

The differential image registration method that is presented in this paper has been implemented on a laptop with a 2.4GHz processor. The evaluation of the displacement parameters has been performed using the presented inverse compositional scheme combined with a pyramidal approach that increases the convergence domain and speeds up the convergence of the optimization. We limit our experiments to the estimation of the displacement of planar image regions.

A. Mono-modal image alignment

The robustness and accuracy of the proposed method have been evaluated on various image sequences.

1) *Image alignment through natural variations*: This experiment concerns an indoor sequence acquired at video rate (25Hz). The initialization of the registration process has been performed by learning the reference image from the first image of the sequence and setting the initial homography to an identity. The template includes 16000 reference pixels.

The sequence has been chosen to illustrate the robustness of the motion estimation through many perturbation. Some images of the sequence are shown in figure 9. Firstly, the object is subject to several illumination variations: the artificial light produced an oscillation on the global illumination of the captured sequence. Moreover the object is not Lambertian,

thus the sequence is subject to saturation and specularities (see figure 9 frame 200). The object is moved from its initial position using wide angle and wide range motions (figure 9 frame 400). And finally the object is subject to fast motion causing a significant blur in many images (figure 9 frame 600).

The frames of the sequence are presented with the corresponding estimated positions of the reference image. No ground truth of the object position is known, however, the projection of the tracked image on the reference image has been performed and qualitatively attests the accuracy of the registration process. Indeed the reconstructed templates show strong variations in terms of appearance but not in terms of position. We can conclude that the estimation of the motion is robust and accurate despite the strong illumination variations and blurring effects.

Concerning the processing time, using the proposed approach with no selection of the reference points (section IV-C2), the images are processed at video rate (25Hz). Using the fast computation (section IV-C3) it is about 40Hz. All the corresponding sequences are presented in the attached video.

In this experiment, we see that even if nothing guaranties that the optimization reaches the global maximum, the proposed computation of MI has such a wide maximum that it yields to a really robust approach. If nonetheless the convergence had to be verified, a solution could be to use a parallel tracking by matching approach [23] and check if we can find a better match (an higher MI score) than the one estimated in the non-linear optimization.

2) *Evaluation on benchmark datasets*: To have a quantitative measure of its accuracy and robustness, the registration process has been evaluated on some very demanding reference datasets proposed by Metaio GmbH [22]. Those datasets include a large set of sequences with the typical motions that we are suppose to face in augmented reality applications. Indeed sequences using eight reference images from low repetitive texture to highly repetitive texture are included. And for each reference image is a set of four sequences depicting wide angle, high range, fast far and fast close motion and one sequence with illumination variations.

The estimated motion has been compared with the ground truth for each sequences. The percentages given in the tables have been computed by Metaio relative to their ground truth. The upper table on figure 8 shows the results that have been obtained using the proposed approach. The tracker is considered as converging if the error between the estimation and the ground truth is below a given threshold. The error measure is similar to the one defined in equation (28). A detailed definition is available in [22]. Some images of the sequences are shown in Figure 7 with the estimated position of the reference template. The mutual information based tracker proves its robustness and accuracy on most of the sequences.

The results obtained using the ESM approach [3] reported from [22] are also represented in the lower table of figure 8 where better convergence results are in bold characters. If we compare the results of the two methods we can see that both have similar convergence rates in most cases. But MI has an undeniable advantage in the cases of illumination variations

```

%Definition of the reference:
 $I^* = I_0$ ;
 $\mathbf{p}_0 = \text{Id}$ ;

%Precomputation of the derivatives:
for  $\mathbf{x} \setminus \{\|\nabla I^*(\mathbf{x})\| > \alpha\}$  do
    Compute  $\frac{\partial \phi(r - \bar{I}^*(w(\mathbf{x}, \Delta \mathbf{p})))}{\partial \Delta \mathbf{p}}$ ;
end for
Compute  $\mathbf{H}^{-1}$  using  $\bar{I}(w(\mathbf{x}, \mathbf{p})) = \bar{I}^*(\mathbf{x})$ ;

for each image  $I_t$  do
     $\mathbf{p}_t^0 = \mathbf{p}_{t-1}$ ;
     $k = 0$ ;
    while  $\Delta \mathbf{p}$  is significant do
        Compute  $\mathbf{G}$ ;
         $\Delta \mathbf{p} = -\mathbf{H}^{-1} \mathbf{G}^\top$ ;
         $\mathbf{p}_t^{k+1} = \mathbf{p}_t^k \oplus \Delta \mathbf{p}^{-1}$ ; %See equation (5)
    end while
end for

```

Fig. 6. Pseudo-code of the proposed method to solve a classical image registration task.

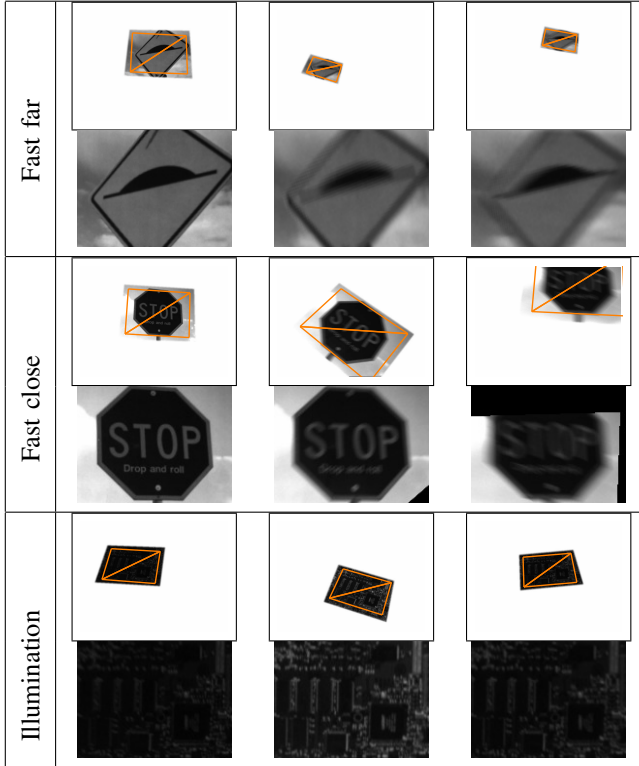


Fig. 7. Various image sequences from the Metaio dataset [22]: fast motions (blurred images) and illumination variations. The first line represents the image with the estimated position of the reference (in green). The second line represents inverse projection from the image to the reference image.

experiments.

From this demanding set of experiments, we can conclude that the proposed MI computation has a large convergence domain (at least as large as the one in the least squared problem) and that the proposed optimization is adapted to use the potential of the MI function leading to a very efficient image registration method.

B. Multi-modal image alignment

1) *Satellite images versus map*: This experiment illustrates the capabilities of the presented mutual information-based image registration process in alignment applications between map and aerial images. The reference image is a map template provided by IGN (Institut Géographique National) that can easily be linked to Geographic Information System (GIS) and the sequence has been acquired using a moving USB

| MI | Angle | Range | Fast Far | Fast Close | Illumination |
|------------|--------------|--------------|-------------|-------------|--------------|
| Low | 100.0 | 94.1 | 75.2 | 56.5 | 99.5 |
| | 100.0 | 98.1 | 69.9 | 43.7 | 93.0 |
| Repetitive | 76.9 | 67.9 | 22.8 | 63.6 | 100.0 |
| | 91.3 | 67.1 | 10.4 | 70.5 | 96.2 |
| Normal | 99.2 | 99.3 | 43.9 | 86.7 | 99.6 |
| | 100.0 | 100.0 | 14.8 | 84.5 | 100.0 |
| High | 47.1 | 23.2 | 7.2 | 10.0 | 50.6 |
| | 100.0 | 69.8 | 20.8 | 83.8 | 100.0 |
| ESM | Angle | Range | Fast Far | Fast Close | Illumination |
| Low | 100.0 | 92.3 | 35.0 | 21.6 | 71.1 |
| | 100.0 | 64.2 | 10.6 | 26.8 | 56.3 |
| Repetitive | 61.9 | 50.4 | 22.5 | 50.2 | 34.5 |
| | 2.9 | 11.3 | 6.8 | 35.8 | 11.3 |
| Normal | 95.4 | 77.8 | 7.5 | 67.1 | 76.8 |
| | 99.6 | 99.0 | 15.7 | 86.8 | 90.7 |
| High | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | 100.0 | 61.4 | 22.8 | 45.5 | 79.7 |

Fig. 8. Ratio of successfully registered images for our approach compared to the ESM [22].

camera focusing on a poster representing the satellite image corresponding to the map.

As it has been previously noticed in figure 1, a non-linear relationship exists between the intensities of the map and aerial image and this link can be evaluated by the MI function. Mutual information can therefore allow for aligning the satellite image using the map image. As Figure 10 shows, the selected initial position can be rather far from the correct position. Figure 11 shows the reference image and some images of the sequence with the corresponding overlaid results. There is no available ground truth for this experiment, nevertheless the overlaid results give a good overview of the alignment accuracy. We can also see in the attached video that the alignment converges despite some strong blurring effects.

2) *Airborne infrared image versus satellite images*: The same method has been evaluated with another current modality. This time the reference is a satellite image and the sequence is an airborne infrared sequence provided by Thales Optronic.

As expected, although very different, the two images shown in figure 13 are sharing a lot of information and thus MI can handle the registration process between the infrared sequence and the satellite image template. The warp function is still a homography. The airport scene is then supposed to be planar leading to an approximation. Nevertheless the proposed



Fig. 9. registration of a planar template through illumination variations. First row: frame 0, 200, 400 and 600. The green rectangle represents the rectangle from the template image transformed using the estimated homography. Second row: projection of the templates for the same iterations in the reference image.

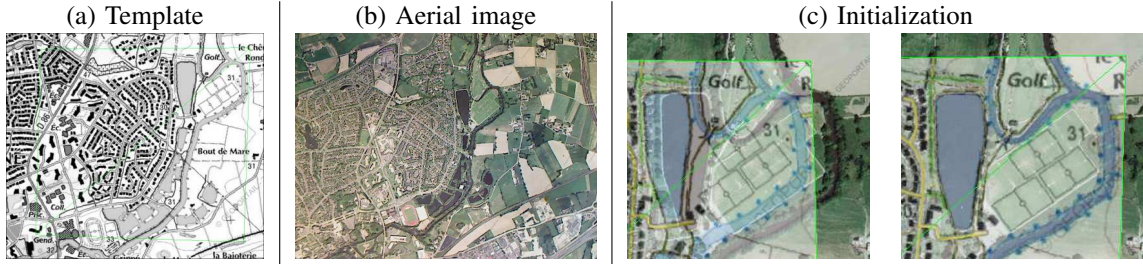


Fig. 10. Registration of an aerial sequence with a map template image by MI: (a) the considered template to be registered with an image of an aerial view of the same area (b) (image and map source: IGN). (c) Shows the initialization process (corresponding to one step of the tracking algorithm). It allows to show that the convergence domain is quite large despite the fact that the images are very different.



Fig. 11. Registration of an aerial sequence with a map template image by MI: frames 1, 250, 500 and 750 are represented with the over-imposed satellite reference (inside the green rectangle) projected using the estimated homography (image and map source: IGN).

method remains robust. No ground truth is available, but the overlaid images qualitatively validates the accuracy of the registration process.

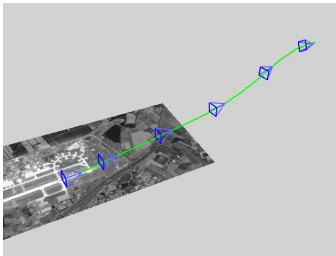


Fig. 12. From the homography to the estimation of the camera position. Green curve: estimated camera trajectory in the 3D space, blue: the 6 estimated camera positions corresponding to the frames represented in figure 13.

The homographies have been decomposed to estimate the position of the plane with respect to the airport. The resulting 3D trajectory of the camera is represented in figure 12, as we can see the trajectory is smooth and has the expected

behavior that shows the approach of a plane with respect to the runway. The trajectory of the camera with respect to the time is presented in the attached video. Figures 13 also shows some registered images that validate the accuracy of the motion estimation. The complete sequences are visible in the attached video.

To illustrate the improvements led by our approach Figure 14 shows the difference between a classical first order maximization approach using MI [13] (first row) and the proposed one (second row) that considers the full computation of the Hessian. Small registration errors can be observed when considering the classical approach while using a complete Hessian allows a better estimation of the transformation. Plots on the right of Figure 14 show the estimated altitude of the camera/plane (up to a scale factor) during the landing step. One can see, that in the first case, the estimation of the trajectory is noisy while with our approach one can clearly identify the classical three different steps of a landing process: airplane reduces downward slope, classically from 5 to 3 degree, this

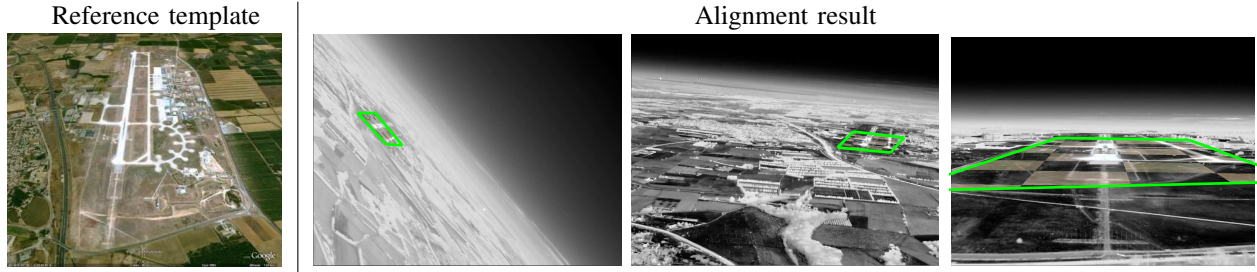


Fig. 13. Registration of a satellite template image using MI on an airborne infrared sequence. 6 frames are represented with the over-imposed aerial reference (inside the green rectangle) projected using the estimated homography (Infrared images courtesy of Thales, optical image is obtained from google earth).

can be seen at 460 on the abscissa axis and finally executes a flare (350). Figure 13 shows the reference template (left), and six images of the sequence (overlaid with the projected template).

When estimating an homography (that is 8 independent parameters), template size may be important especially when considering multi-modal images. To sufficiently populate the joint probability histogram necessary for the computation of the mutual information and to maintain a good accuracy, we have always considered at least 100×100 pixels template. Nevertheless, although the computational cost will obviously increase with the number of pixels, it has to be remembered that most of the computation (Gradient, Hessian) are precomputed and the approach remains cost effective.

C. Mosaicing application

Image mosaics are a collection of overlapping images. The goal of the mosaicing problem is to find the transformations that relate the different image coordinates. Once the transformation between all the images is known, an image of the whole scene can be constructed. This problem requires to find a warping function that maps the coordinates of one image into the coordinate system of another image. When considering a video, one has to warp each new image into the coordinate system of the very first image of the video [35][6][19][16]. This is basically a motion estimation process performed on the whole image. One can consider to estimate this motion using matched keypoints as in [6] or using SSD based motion estimation as in [19][16]. The latter approach is very efficient when image sequences are considered, that is, when displacements between one frame to another are small but shows its limits in case of image noise or occlusions. This section shows the benefits of using the presented approach to solve the registration problem.

These experiments show the application of the MI-based motion estimation algorithm to the mosaicing problem. In these sequences, since some parts of the scene completely disappear, it is necessary to define multiple reference images along the sequence. The approach is build as follows:

- Initialization: the first image is chosen as reference image, i.e. $I_0^* = I_0$.
- Registration: for every frame, we compute the displacement \mathbf{p}_k between I_t and I_k^* .
- Reference Update: every 30 images, the reference image I_k^* is changed and defined as the current image, i.e. $I_k^* = I_t$ for $t = 30k$ (involving a small drift).

Using the homography from the current image to the current reference image and the homographies between the references, we retrieve the homography between the current image and the first image. Using this homography, we can project all the images of the sequence into the mosaic image and construct the global image of the whole scene.

In the first experiment (Figure 15 and 16), the overlapping images are simply a compressed sequence of 1000 images. The aerial scene is acquired from a camera embedded on a flying UAV and shows the ground that is approximately 500 meters away from the camera. Since this distance is very large, the scene can be approximated as a plane and registered using homographies. During the acquisition of the sequence, the camera is moving forward and is rotating around the vertical axis.

In Figure 15 we show some images from the sequence. This sequence has been downloaded on Youtube and is affected by the H264 coding artifacts. We can also note the poor quality of the images. Despite this poor quality, the resulting mosaic presented in Figure 16 shows the accuracy of the MI based method. Since the camera is making an entire revolution, the first and last images are overlapping. A very small drift occurs between the first and last estimated positions. Let us note that nothing has been performed to reduce the drift (such as the bundle adjustment approach proposed by [6]). Considering the template update problem and the planar assumption, the estimated homographies are accurate. A second similar experiment presents a mosaic build using more than 10000 images. The images are extracted from a highly compressed video. The camera was attached to a free flying balloon flying over Paris. Figure 17 shows three steps of the mosaic construction.

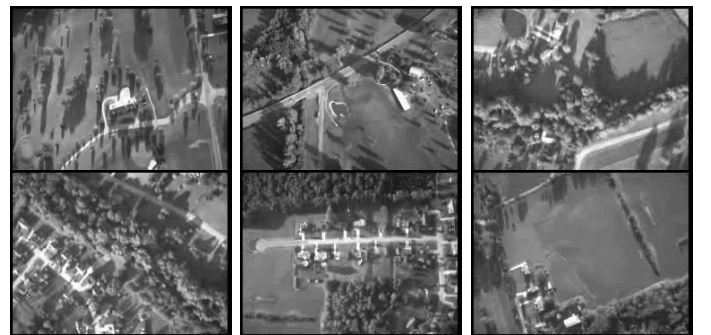


Fig. 15. Some overlapping key images used for the mosaicing application.

In the last experiment (see mosaic in Figure 19), we consider

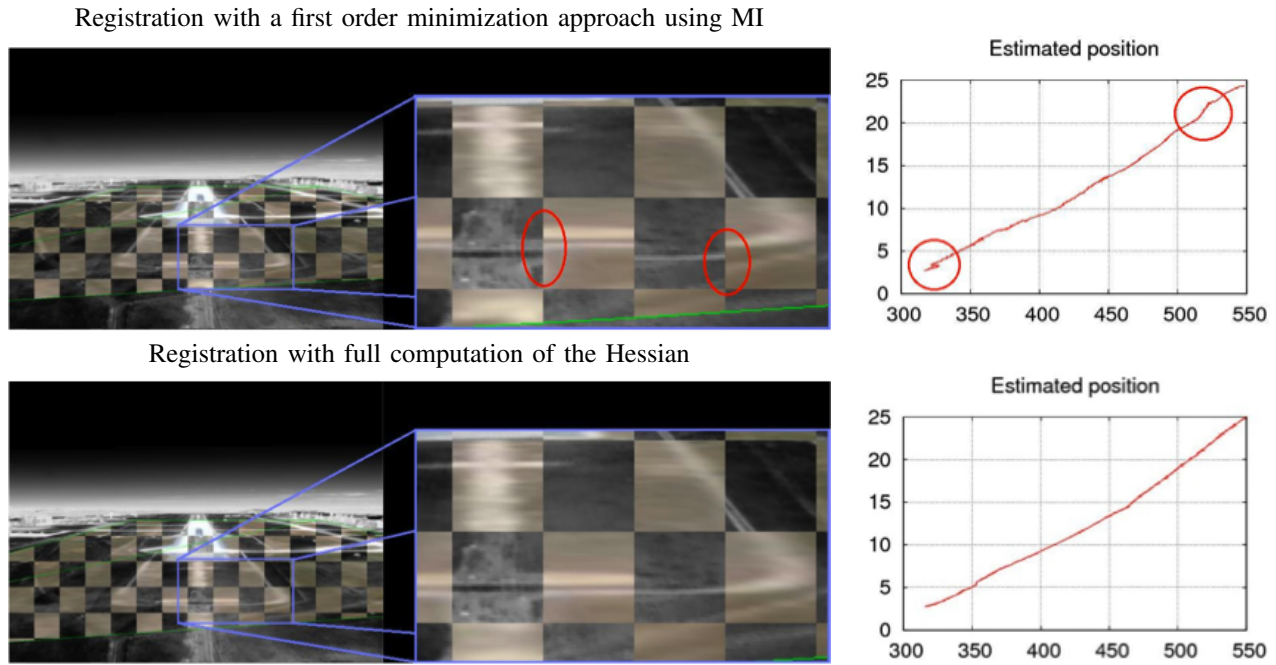


Fig. 14. Registration of a satellite template image using MI on an airborne infrared sequence. First row corresponds to a registration process with a first order minimization approach while second row depicts our approach with a full computation of the Hessian. On the left is the estimated altitude versus distance to the runway (up to a scale factor).



Fig. 19. Mosaic created from the John Ford movie “*she wore a yellow ribbon*”. An affine motion model was considered. Note that some cavalymen are moving all along the sequence. Despite these disturbances, motion is correctly estimated.

a sequence extracted from the John Ford movie “*she wore a yellow ribbon*”. To build this mosaic, an affine motion model was considered. The interest of this sequence is that some cavalymen are moving all along the sequence and, therefore, act as important occlusions as can be seen on Figure 18. A comparison with a mosaic built using the SSD criterion is proposed in Figure 20 and demonstrates the robustness of our approach.

VI. CONCLUSION

This paper presents a robust and accurate template based-motion estimation process that was defined using a new approach based on the mutual information alignment function. The definition of MI has been adapted to the differential image alignment problem so that the function is smooth and as concave as possible. The proposed definition preserves the advantages of MI with respect to its robustness toward occlusions, illumination variations and images from different

modalities. A new optimization approach has been defined to deal with the quasi-concave shape of MI. The proposed approach is taking advantage of both the wide convergence domain of MI and its accurate maximum and besides is not computationally expensive. Moreover the time consumption is greatly reduced using a new approach based on the reference pixels selection that yields to an accurate, fast and robust registration process.

Finally the proposed method has been evaluated using several experiments. Its robustness and accuracy is verified using reference datasets and shows its advantages compared with classical approaches on monomodal image registration method. Some new applications are also proposed to use a model image acquired from another modality than the original sequence.

The algorithm presented here has been limited to planar scene. Nevertheless the proposed approach could similarly be applied to more complex model-based tracking applications

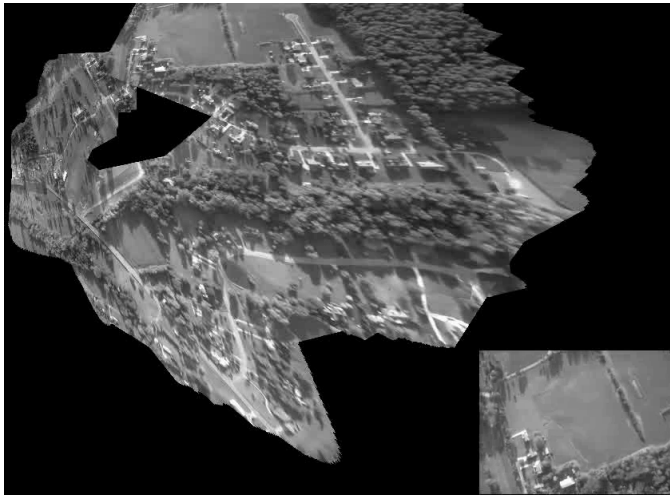


Fig. 16. Resulting mosaic image: despite the poor quality of the sequence and the approximation that the scene is planar, the final displacement between the first and last image is accurate.

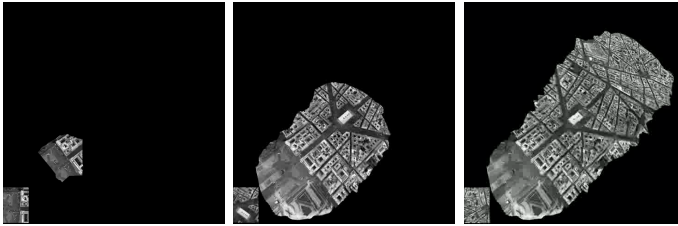


Fig. 17. Three steps of the "Paris" mosaic construction. The sequence feature more than 10000 images acquired from a camera attached to a free-flying balloon.



Fig. 18. Three images used for the "yellow ribbon" mosaic construction.



Fig. 20. Mosaic created from the John Ford movie but the similarity function is the SSD.

where we could directly estimate the position of the object on $\mathbb{SE}(3)$ [7][29]. The method could also be extended to non-rigid registration process.

ACKNOWLEDGMENT

This work was supported by DGA under contribution to student grant.

REFERENCES

- [1] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition, 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. Homography-based 2d visual tracking and servoing. *Int. Journal of Robotics Research*, 26(7):661–676, July 2007.
- [4] C. Bibby and I. Reid. Robust real-time visual tracking using pixel-wise posteriors. In *European Conference on Computer Vision, ECCV'08*, pages 831–844, 2008.
- [5] L. Brown. A survey of image registration techniques. *ACM Comput. Surv.*, 24:325–376, Dec. 1992.
- [6] M. Brown and D. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007.
- [7] G. Caron, A. Dame, and E. Marchand. L'information mutuelle pour l'estimation visuelle directe de pose. In *18e congrès francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle, RFIA 2012*, Lyon, France, January 2012.
- [8] H. Chen, P. Varshney, and M.-A. Slamani. On registration of regions of interest (roi) in video sequences. In *IEEE Conf. on Advanced Video and Signal Based Surveillance.*, pages 313 – 318, July 2003.
- [9] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal. Automated multi-modality image registration based on information theory. In *Int. Conf. on Information Processing in Medical Imaging, IPMI'95*, Ile de Berder, France, June 1995.
- [10] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 142–149, 2000.
- [11] A. Comport, E. Marchand, M. Pressigout, and F. Chaumette. Real-time markerless tracking for augmented reality: the virtual visual servoing framework. *IEEE Trans. on Visualization and Computer Graphics*, 12(4):615–628, July 2006.
- [12] A. Dame. A unified direct approach for visual servoing and visual tracking using mutual information. PhD thesis, Université de Rennes 1, Dec. 2010.
- [13] N. Dowson and R. Bowden. A unifying framework for mutual information methods for use in non-linear optimisation. In *European Conference on Computer Vision, ECCV'06*, volume 1, pages 365–378, June 2006.
- [14] N. Dowson and R. Bowden. Mutual information for lucas-kanade tracking (milk): An inverse compositional formulation. *IEEE Trans. on PAMI*, 30(1):180–185, Jan. 2008.
- [15] T. Drummond and R. Cipolla. Real-time visual tracking of complex structures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7):932–946, July 2002.
- [16] N. Gracias and J. Santos-Victor. Underwater video mosaics as visual navigation maps. *Computer Vision and Image Understanding*, 79(1):66 – 91, 2000.
- [17] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, Oct. 1998.
- [18] M. Irani and P. Anandan. Robust multi-sensor image alignment. In *IEEE Int. Conf. on Computer Vision, ICCV'98*, pages 959–966, Bombay, India, 1998.
- [19] M. Irani, P. Anandan, and S. Hsu. Mosaic based representations of video sequences and their applications. In *IEEE Int. Conf. on Computer Vision*, page 605, 1995.
- [20] F. Jurie and M. Dhome. Hyperplane approximation for template matching. *IEEE Trans. on PAMI*, 24(7):996–1000, July 2002.
- [21] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *IEEE Trans. on PAMI*, 28(9):1465–1479, Sept. 2006.
- [22] S. Lieberknecht, S. Benhimane, P. Meier, and N. Navab. A dataset and evaluation methodology for template-based tracking algorithms. In *IEEE Int. Symp on Mixed and Augmented Reality, ISMAR'09*, pages 145–151, 2009.
- [23] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.
- [24] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Int. Joint Conf. on Artificial Intelligence, IJCAI'81*, pages 674–679, 1981.
- [25] 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.
- [26] F. Maes, D. Vandermeulen, and P. Suetens. Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information. *Medical Image Analysis*, 3(4):373–386, 1999.
- [27] C. Meyer, J. Boes, B. Kim, P. Bland, K. Zasadny, P. Kison, K. Koral, K. Frey, and R. Wahl. Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using

- affine and thin-plate spline warped geometric deformations. *Medical Image Analysis*, 1(3):195 – 206, 1997.
- [28] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, Dec. 1995.
 - [29] G. Panin and A. Knoll. Mutual information-based 3d object tracking. *Int. Journal of Computer Vision*, 78(1):107–118, 2008.
 - [30] J. Pluim, J. Maintz, and M. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Trans. on Medical Imaging*, 22(8):986–1004, Aug. 2003.
 - [31] N. Ritter, R. Owens, J. Cooper, R. Eikelboom, and P. Van Saarloos. Registration of stereo and temporal images of the retina. *IEEE Trans. on Medical Imaging*, 18(5):404–418, May 1999.
 - [32] C. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, Jan. 2001.
 - [33] G. Silveira and E. Malis. Real-time visual tracking under arbitrary illumination changes. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition, CVPR'07*, Minneapolis, USA, June 2007.
 - [34] C. Studholme, D. Hill, and D. J. Hawkes. Automated 3D registration of truncated mr and ct images of the head. In *British Machine Vision Conference, BMVC'95*, pages 27–36, Birmingham, Surrey, UK, Sept. 1995.
 - [35] R. Szeliski. Image alignment and stitching: a tutorial. *Found. Trends. Comput. Graph. Vis.*, 2:1–104, January 2006.
 - [36] P. Thévenaz and M. Unser. Optimization of Mutual Information for Multiresolution Image Registration. *IEEE trans. on Image Processing*, 9(12):2083–2099, 2000.
 - [37] P. Thévenaz and M. Unser. Optimization of mutual information for multiresolution image registration. *IEEE Transactions on Image Processing*, 9(12):2083 – 2099, Dec. 2000.
 - [38] P. Viola and W. Wells. Alignment by maximization of mutual information. In *Int. Conf. on Computer Vision, ICCV'95*, Washington, DC, 1995.
 - [39] P. Viola and W. Wells. Alignment by maximization of mutual information. *Int. Journal of Computer Vision*, 24(2):137–154, 1997.
 - [40] B. Zitová and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977 – 1000, 2003.



Amaury Dame graduated from the Institut National des Sciences Appliquées de Rennes in 2007 and received a PhD in computer science from the Université de Rennes 1 in 2010 as a member of the INRIA Lagadic team. He is now research assistant in the Active Vision Group, Department of Engineering Science, University of Oxford. His research interests include computer vision, slam and visual servoing. He received a Best paper runner-up awards at IEEE ISMAR 2010.



Eric Marchand is professor of computer science at Université de Rennes 1 in France and a member of the INRIA/IRISA Lagadic team. He received a Ph.D degree and a "Habilitation Diriger des Recherches" in Computer Science from the Université de Rennes 1 in 1996 and 2004 respectively. He spent one year as a Postdoctoral Associates in the AI lab of the Dpt of Computer Science at Yale University. He has been an INRIA research scientist ("Chargé de recherche") at INRIA Rennes-Bretagne Atlantique from 1997 to 2009. His research interests include robotics, visual servoing, real-time object tracking and augmented reality. He is an associate editor for IEEE Trans. on Robotics.