Accurate Real-time Tracking Using Mutual Information

Amaury Dame*
CNRS, IRISA,
INRIA Rennes Bretagne-Atlantique

Eric Marchand†
Université de Rennes 1, IRISA,
INRIA Rennes Bretagne-Atlantique

Figure 1: Augmenting an aerial image through MI-based tracking. An image template (a) extracted from a geographic map is registered in real-time with an aerial image (b). The information extracted from a geographic information system is used to augment the aerial image (c).

ABSTRACT

In this paper we present a direct tracking approach that uses Mutual Information (MI) as a metric for alignment. The proposed approach is robust, real-time and gives an accurate estimation of the displacement that makes it adapted to augmented reality applications. MI is a measure of the quantity of information shared by signals that has been widely used in medical applications. Since then, and although MI has the ability to perform robust alignment with illumination changes, multi-modality and partial occlusions, few works propose MI-based applications related to object tracking in image sequences due to some optimization problems.

In this work, we propose an optimization method that is adapted to the MI cost function and gives a practical solution for augmented reality application. We show that by refining the computation of the Hessian matrix and using a specific optimization approach, the tracking results are far more robust and accurate than the existing solutions. A new approach is also proposed to speed up the computation of the derivatives and keep the same optimization efficiency.

To validate the advantages of the proposed approach, several experiments are performed. The ESM and the proposed MI tracking approaches are compared on a standard dataset. We also show the robustness of the proposed approach on registration applications with different sensor modalities: map versus satellite images and satellite images versus airborne infrared images within different AR applications.

1 INTRODUCTION

Typical augmented reality applications require real-time tracking performances. Indeed, to allow a precise positioning of the virtual scene within the real images, knowing the actual position of the camera is necessary. As for many problems in computer vision, the motion estimation remains therefore one of the key issues.

*e-mail: amaury.dame@irisa.fr
†e-mail: eric.marchand@irisa.f

IEEE International Symposium on Mixed and Augmented Reality 2010
Science and Technolgy Proceedings
13 -16 October, Seoul, Korea
978-1-4244-9346-3/10/$26.00 ©2010 IEEE

Major difficulties in such a tracking process are image noise, illumination changes and occlusions. Along with robustness to such perturbations, our motivation is to focus on tracking and registration considering different sensor modalities. For example, registrations will be performed between a map and an airborne image sequence (see figure 1) or between infra-red and visible images (see figure 10).

Most of the available tracking techniques can be divided into two main classes: feature-based and model-based tracking. 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 tracked objects. This model can be a 3D model leading, mainly, to a pose estimation process corresponding to a registration process between measures in the image and the forward projection of the 3D model [7][4]. One can also consider 2D models. Within this category, the object to be tracked can be represented by a descriptor. These descriptors can be object histograms leading to mean shift like approaches [3] or point neighborhood leading to keypoint tracking by matching approaches [12][10]. Such approaches are usually very robust to illumination variation, occlusions, etc. It is also possible to consider that this 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 tracking methods such as the KLT [13] or [8][1][2]. 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 estimate its 3D motion.

The approach described in this paper is related to the later category of trackers. In this context, 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 (SSD) function [13][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 or add new parameters to estimate the illumination variations [8][20]. Nevertheless those approaches leads to

complex models.

In this paper, our goal is first to have a visual tracking approach that is robust to occlusions and illumination variations, but also to track an object with its appearance model acquired in another modality than the one used in the current image sequence. The proposed solution is then to replace the SSD function by a more robust alignment function.

One can consider local normalized cross correlation (NCC) [9] to replace SSD, but our results show that it is not applicable to different image modalities. The proposed solution is then to maximize the information shared between the reference image and the sequence by maximizing the Mutual Information (MI) function [19, 22, 17]. MI has also proved to be robust to occlusions and illumination variations and can therefore be considered as a good alignment measure for tracking [6, 15]. However the existing approaches are not taking full advantage of the accuracy of MI and thus are not appropriate for augmented reality applications.

In this paper we present a MI-based tracker where an important contribution is to propose an optimization process adapted to the MI cost function. The optimization process that we propose is an inverse compositional approach where an important part of the derivatives needed in the optimization can be precomputed, resulting in small computation times. A precise, complete and efficient computation of the Hessian matrix is described. The inverse compositional approach allows the estimation of the Hessian matrix after convergence. We show that this Hessian matrix can be used in a Newton's like approach to give an accurate and fast estimation of the displacement parameters that will prove its reliability in augmented reality applications. Finally a new approach is proposed to speed up the computation of the derivatives through a selection of the used reference pixels that makes the mutual information tracking process possible at video-rate meeting AR requirements.

In the remainder of this paper, Section 2 presents an overview of the differential approaches. In section 3, a brief introduction on information theory is given with the definition of mutual information, then a formulation adapted to the differential tracking method is presented. Section 4 deals with the optimization of the resulting mutual information function with respect to the motion parameters to estimate. Finally section 5 presents tracking results including the Metaio benchmark and presents augmented reality experiments that demonstrate the new multimodal capability of the approach.

2 Differential template-based tracking

Differential 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 $I_0..I_t$. In the case of a similarity function f, the problem can be written as :

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

where we search the displacement $\widehat{\mathbf{p}}_t$ 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 $\mathbf{p}$. Indeed, its correct formulation $w(\mathbf{x},\mathbf{p})$ gives the function that moves a point $\mathbf{x}$ from the reference image to its coordinates in the current image.

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 $\widehat{\mathbf{p}}$ are thus too expensive if not impossible.

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

Multiple solutions exists to compute the update of the current displacement parameters and perform the optimization. Indeed Baker and Matthews showed that two formulations were equivalent [1]. 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 $\Delta\mathbf{p}$. 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 will classically be 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:

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

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

where the summation is computed on each point $\mathbf{x}$ of the reference template that is the region of interest (ROI) 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 robustifies the least squared problem toward occlusions [8] and a model of illumination changes can be coupled with the motion model to create a tracker robust to lighting changes [20].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 2(a)). Considering these two modalities is obviously an extreme case, but it will emphasize the robustness of the proposed approach. The value of SSD is computed with respect to the translations between the map and the satellite image. It is clear that the two images are showing the same place (at least for a human

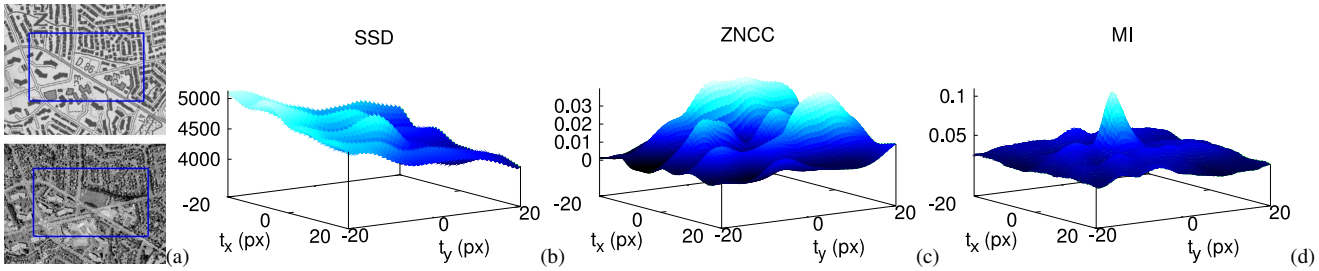

Figure 2: Alignment functions wrt. translations between two images from the same area: (a) aerial image and the map reference. MI shows a maximum near zero translation at the alignment position whereas SSD and ZNCC gives no clear information on the alignment quality.

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 2(b) gives no information on the alignment between the two images. The NCC has shown some very good results in multimodal alignment problems [9]. The efficiency of the zero-mean normalized cross correlation (ZNCC) has been evaluated on the mutlimodal example in figure 2(c). We can see that the case is too extreme and that there is also no significant optimum. We can conclude that even ZNCC is not sufficient to give a good measure of alignment in this case.

To deal with occlusions, illumination variations and multimodality, we propose to use the mutual information [19, 22] as the alignment function, that is, as we will see, robust to all this variations of appearance.

3 Mutual Information

3.1 Information theory

Mutual information is an alignment function that was first introduced in the context of information theory [19]. Some essential notions such as entropy and joint entropy are required for a good understanding of this alignment measure.

3.1.1 Entropy

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:

$$h(I) = -\sum_r p_I(r) \log\left(p_I(r)\right). \tag{8}$$

The log basis only changes the entropy value with a scale factor, therefore it has no interest in our tracking problem and will be omitted since we only seek the maximum of the cost function but not a particular value.

Since our goal is to focus on images, let us consider I as an image and $r = I(\mathbf{x})$ as the possible gray-level intensities of the image pixels $\mathbf{x}$. The probability distribution function of the gray-level values is then simply given by a the normalized histogram of the image I. The entropy can therefore be considered as a measure of dispersion of the image histogram.

3.1.2 Joint entropy

Following the same principle, 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:

$$h(I,I^*) = -\sum_{r,t} p_{II^*}(r,t) \log\left(p_{II^*}(r,t)\right) \tag{9}$$

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 bidimensional histogram of the two images. As for entropy, joint entropy corresponds to a measure of dispersion of the joint histogram of (I,I^*).

At first sight the joint entropy could be considered as a good alignment measure: if the dispersion of the joint histogram is small then the correlation between the two images is strong and we can suppose that the two images are aligned. Nevertheless the dependencies on the entropies of I and I^* makes it not adapted. Indeed if one of the images has a constant gray-level value then the joint histogram would be very focused and the entropy value very small despite the fact that the two images are not aligned.

3.1.3 Original Mutual information

The definition of mutual information (MI) solves the above mentioned problem [19, 22]. Subtracting the random variable's entropies from their joint entropy yields to an alignment measure that is not depending on the variable marginal entropies. The MI of two random variables I and I^* is then given by the following equation:

$$MI(I,I^*) = h(I) + h(I^*) - h(I,I^*). \tag{10}$$

MI is then the quantity of information shared between two random variables. If the two variables/images are aligned then their mutual information is maximal.

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 notation as in section 2, the mutual information can thus be written with respect to $\mathbf{p}$:

$$MI(\mathbf{p}) = MI(w(I,\mathbf{p}),I^*) = h(w(I,\mathbf{p})) + h(I^*) - h(w(I,\mathbf{p}),I^*). \tag{11}$$

The final expression of MI is obtained by developing the previous equation using the entropy equations (8) and (9):

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

Let us consider a simple example, in figure 3 mutual information has been computed with respect to a translational displacement $\mathbf{p} = (t_x,t_y)$ using its classical definition. A white noise has been added to the reference image I^*. The ground truth displacement between the two images is known and is $\mathbf{p} = \mathbf{0}$. The blue rectangle drawn in the images represents the region of the reference image that is used to compute the reference histograms. On the left is represented this histogram that contains 256 gray level values. As we can see, the original definition of MI proposed by Shannon shows a large maximum at the ground truth position but also shows many local maxima known as interpolation artifacts.

3.2 Smoothing Mutual Information

The differential approach consists of using the function and its derivatives to bring the estimated parameters to the optimum of the similarity function. The smoother the function the more efficient the optimization. Thus, preliminary modifications have to be applied to the original formulation to modify the shape of mutual information function and smooth it.

3.2.1 Histograms binning

The computation of MI on histograms of 256 entries presents problems due to the large number of empty bins that have strong repercussions on the entropies measures [16]. Moreover the computation of those histograms are expensive in memory and time.

Starting from this observations, one obvious solution is to decrease the number of histogram bins. The analytical formulation of a normalized histogram of an image I^* is classically written as follows:

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

where $\mathbf{x}$ are the points of the region of interest in the image, $N_{\mathbf{x}}$ is the number of points and t are the possible values of $I^*(\mathbf{x})$, i.e. $t \in [0,255]$. In the classical formulation ϕ is a Kronecker's function: $\phi(x) = 1$ for $x = 0$ and $\phi(x) = 0$ otherwise. So that each time $I^*(\mathbf{x}) = i$ the i^{th} histogram bin value is incremented.

The number of bins corresponds to the maximal gray level intensity of the image $N_{c_{I^*}} = 256$. To reduce it, the image intensities are simply scaled as follows:

$$\overline{I^*}(\mathbf{x}) = I^*(\mathbf{x}) \frac{N_c - 1}{N_{c_{I^*}} - 1} \quad (14)$$

where N_c is the new number of histogram bins. The resulting intensities are no longer integer values. Thus the ϕ function has to be modified to keep the information on these real values. Several solutions have been proposed to simultaneously smooth the mutual information function and keep its accuracy [22][14]. Our approach is based on the use of B-spline functions [14] that are approximations of Gaussian functions and have the advantages of their fast computation and differentiability.

The final analytical formulation of the normalized histogram becomes:

$$p_{I^*}(t) = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \phi\left(t - \overline{I^*}(\mathbf{x})\right) \quad (15)$$

where the possible gray-level values are now $t \in [0, N_c]$.

The probability distribution function of I^* and the joint probability of (I, I^*) are modified using the same approach that yields to:

$$p_I(r, \mathbf{p}) = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \phi\left(r - \bar{I}(w(\mathbf{x}, \mathbf{p}))\right) \quad (16)$$

$$p_{II^*}(r, t, \mathbf{p}) = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \phi\left(r - \bar{I}(w(\mathbf{x}, \mathbf{p}))\right) \phi\left(t - \overline{I^*}(\mathbf{x})\right) \quad (17)$$

Several solutions have been proposed to estimate an optimal number of histogram bins such as Sturges' rule or Scott's rule [18]. Nevertheless, a constant number of bin set with $N_c = 8$, that keeps a small value and avoids loosing information, has always given satisfying results in our experiments. Note that the final number of bins is higher than N_c due to the side-effect of the B-spline functions.

If we compare the mutual information values between the original formulation and the new one, the benefits of the histogram binning operation are obvious. As figure 3 shows, mutual information function is convexified. Nevertheless MI is still subject to small interpolation artifacts.

3.2.2 Image interpolation

The image interpolation problem is similar to the binning interpolation one. In binning interpolation we put a real value on an integer array, in image interpolation a real value is extracted from an integer array. Indeed the position resulting from the warp of a point $w(\mathbf{x}, \mathbf{p})$ is usually not an integer value.

One classical solution that has been used in the previous computation of the MI in figure 3 is to choose a bilinear interpolation. This is typically similar to the use of first-order B-splines in the binning interpolation problem. The typical solution to solve this problem would be to use a cubic or quadratic image interpolation. Nevertheless such methods are highly time consuming since for each warped point, the computation of its intensity would require to use the intensities of the 9 or 16 neighboring pixels.

The proposed solution that is less time consuming and as far as we know equally efficient is to convolute the current image with a Gaussian filter to smooth the pixel intensities and then use bilinear interpolation. The corresponding MI results have been represented on the 2D translational example that was previously showing interpolation artifacts in figure 3. Using both histogram binning and image filtering, the mutual information function's shape becomes perfectly smooth and thus adapted to work with its derivatives in an optimization method.

The proposed MI formulation is also appropriate if we consider the alignment between the map image and aerial image shown in figure 2. Indeed MI is maximal at $\mathbf{p} = 0$ and the shape of the function remains smooth.

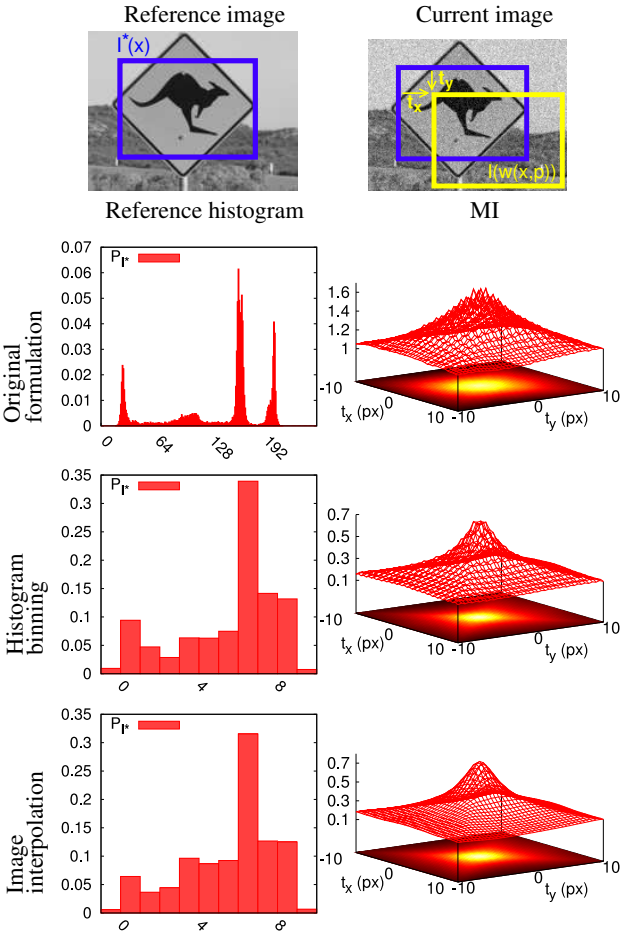

Figure 3: Smoothing mutual information.

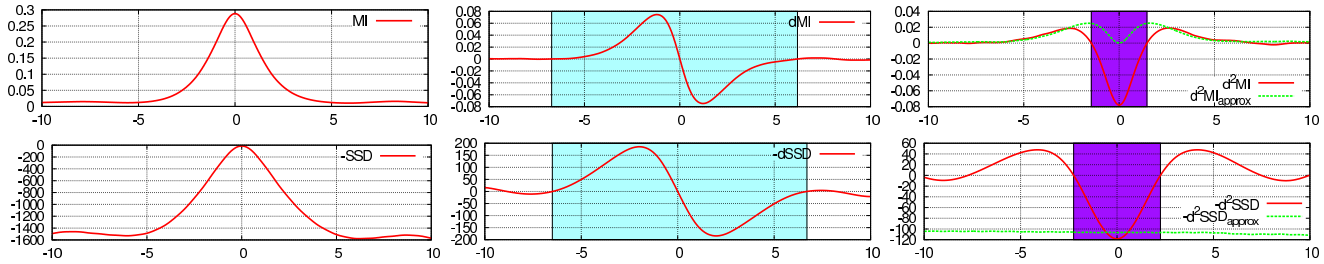

Figure 4: SSD, MI and their derivatives with respect to one translation (px). The purple area is the convergence domain using a classical Newton's method, the blue one is the convergence domain of a Gradient descent method. The proposed method keeps the wider convergence domain of the gradient's method in blue.

4 Mutual Information-based Tracker

In this section we will see how to use the MI cost function with the differential trackers presented in section 2. 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.

4.1 Derivative function analysis

In this work we ought to track planar objects through 3D displacements. This problem implies a strong correlation between the elements of the vector $\mathbf{p}$. Therefore, the use of 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:

$$\Delta\mathbf{p} = -\mathbf{H}^{-1}\mathbf{G}^{\top} \tag{18}$$

where $\mathbf{G}$ and $\mathbf{H}$ are respectively the Hessian and gradient 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:

$$\mathbf{G} = \frac{\partial MI(w(I^*,\Delta\mathbf{p}),w(I,\mathbf{p}))}{\partial\Delta\mathbf{p}} \tag{19}$$

$$\mathbf{H} = \frac{\partial^2 MI(w(I^*,\Delta\mathbf{p}),w(I,\mathbf{p}))}{\partial\Delta\mathbf{p}^2} \tag{20}$$

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

$$\mathbf{G} = \sum_{r,t}\frac{\partial p_{II^*}}{\partial\Delta\mathbf{p}}\left(1+\log\left(\frac{p_{II^*}}{p_{I^*}}\right)\right) \tag{21}$$

$$\mathbf{H} = \sum_{r,t}\frac{\partial p_{II^*}}{\partial\Delta\mathbf{p}}^{\top}\frac{\partial p_{II^*}}{\partial\Delta\mathbf{p}}\left(\frac{1}{p_{II^*}}-\frac{1}{p_{I^*}}\right)+\frac{\partial^2 p_{II^*}}{\partial\Delta\mathbf{p}^2}\left(1+\log\frac{p_{II^*}}{p_{I^*}}\right) \tag{22}$$

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 (19) to equation (22) can be found in [5].

By analogy with classical Hessian computation in SSD minimization, second order derivatives are usually neglected in the Hessian matrix computation [21, 5, 6]. In our approach we compute the Hessian matrix using the second order derivatives that are, in our point of view, required to obtain a precise estimation of the motion. More details are given in appendix A to highlight the problems induced by this classical approximation.

As we can see in equation (21) and equation (22), the derivatives of the mutual information depend on the derivatives of the joint probability. Using the previous definition in (17) and passing the derivative operator through the summation yields the following expressions:

$$\frac{\partial p_{II^*}}{\partial\Delta\mathbf{p}} = \frac{1}{N_{\mathbf{x}}}\sum_{\mathbf{x}}\phi\left(t-\bar{I}(w(\mathbf{x},\mathbf{p}))\right)\frac{\partial\phi\left(r-\overline{I^*}(w(\mathbf{x},\Delta\mathbf{p}))\right)}{\partial\Delta\mathbf{p}} \tag{23}$$

$$\frac{\partial^2 p_{II^*}}{\partial\Delta\mathbf{p}^2} = \frac{1}{N_{\mathbf{x}}}\sum_{\mathbf{x}}\phi\left(t-\bar{I}(w(\mathbf{x},\mathbf{p}))\right)\frac{\partial^2\phi\left(r-\overline{I^*}(w(\mathbf{x},\Delta\mathbf{p}))\right)}{\partial\Delta\mathbf{p}^2} \tag{24}$$

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:

$$\frac{\partial\phi(r-\overline{I^*}(w(\mathbf{x},\Delta\mathbf{p})))}{\partial\Delta\mathbf{p}} = -\frac{\partial\phi}{\partial r}\frac{\partial\overline{I^*}}{\partial\Delta\mathbf{p}} \tag{25}$$

$$\frac{\partial^2\phi\left(r-\overline{I^*}(w(\mathbf{x},\Delta\mathbf{p}))\right)}{\partial\Delta\mathbf{p}^2} = \frac{\partial^2\phi}{\partial t^2}\frac{\partial\overline{I^*}}{\partial\Delta\mathbf{p}}^{\top}\frac{\partial\overline{I^*}}{\partial\Delta\mathbf{p}}-\frac{\partial\phi}{\partial r}\frac{\partial^2\overline{I^*}}{\partial\Delta\mathbf{p}^2}. \tag{26}$$

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

$$\frac{\partial\overline{I^*}}{\partial\Delta\mathbf{p}} = \nabla\overline{I^*}\frac{\partial w(\mathbf{x},\mathbf{p})}{\partial\Delta\mathbf{p}} \tag{27}$$

$$\frac{\partial^2\overline{I^*}}{\partial\Delta\mathbf{p}^2} = \frac{\partial w}{\partial\Delta\mathbf{p}}^{\top}\nabla^2\overline{I^*}\frac{\partial w}{\partial\Delta\mathbf{p}}+\nabla\overline{I^*}_x\frac{\partial^2 w_x}{\partial\Delta\mathbf{p}^2}+\nabla\overline{I^*}_y\frac{\partial^2 w_y}{\partial\Delta\mathbf{p}^2} \tag{28}$$

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 is constant, all the expressions from equation (25) to equation (28) are constants and have to be precomputed only one time.

In our work we focus on planar object tracking. 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 but can also be applied on pose estimation on $\mathbb{SE}(3)$ and other motion models, thus details will not be given on the warp derivatives. All details regarding the derivatives of the chosen warp function can be found in [2].

4.2 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 (see figure 3). The violation could cause the Newton's method to fail, thus a better approach has to be chosen.

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 5(a)) from many initial position parameters (see figure 5(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 $w(\mathbf{x}_i, \mathbf{p})$ [11]. The reference points are simply chosen as the 4 corners of the template so that the error becomes:

$$err(\mathbf{p}) = \sqrt{\sum_{i=1}^{4} \|\mathbf{x}_i^* - w(\mathbf{x}_i, \mathbf{p})\|} \tag{29}$$

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. We represent the convergence rate and the average number of iterations required to reach convergence. 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 4.1). 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 5.

4.2.1 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 4 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). The convergence domain using a classical Newton's method would be very small.

As figure 5(c) shows, the convergence domain of the Newton's method is indeed 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 becomes very small.

However considering the 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 4).

4.2.2 Conditioning the optimization

In tracking problem formulated with a SSD function, the Gauss-Newton approximation condition the problem by estimating a Hessian matrix that is always definite positive (see the green curve in figure 4) 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).

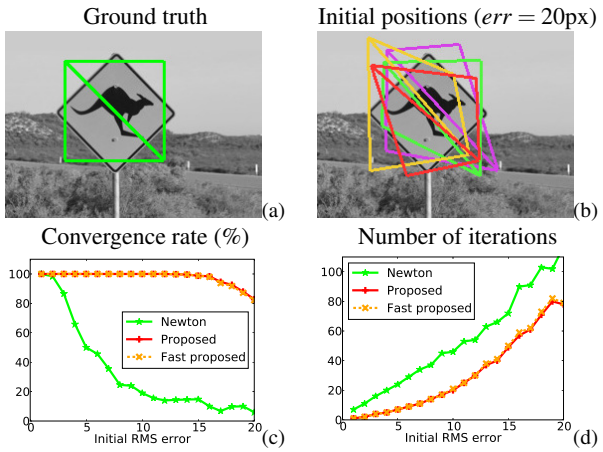

Figure 5: Empirical convergence analysis of the optimization methods. The proposed methods (blue and green curves) have a very high convergence rate compared to the classical Newton's methods (red curve).

In the mutual information maximization, the problem is different. Indeed approximating the Hessian matrix as it is proposed in [21, 5, 6] do not gives an estimation of the Hessian matrix after convergence (see the green curve in 4 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 $\overline{I}(w(\mathbf{x}, \mathbf{p})) = \overline{I^*}(\mathbf{x})$.

This solution has several advantages:

- It gives a definite negative Hessian matrix that yields to have a wide convergence domain (blue area in figure 4). 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 5.1.2, 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 5. 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 $err > 16$.

Figure 5(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.

4.2.3 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 (27) and (28). 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:

$$\|\nabla I^*(\mathbf{x})\| > \alpha \tag{30}$$

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 5. 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 5(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.

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

5 Visual tracking experimental results

The visual tracking 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 convergence of the optimization. We limit our experiments to the estimation of the displacement of planar objects. The estimated homography can be decomposed to find the rotations and translations of the plane and its normal up to a distance factor, which is sufficient for augmented reality applications.

5.1 Monomodal tracking

The robustness and accuracy of the proposed mutual information tracker have been evaluated on various image sequences.

5.1.1 Tracking through natural variations

This experiment concerns an indoor sequence acquired at video rate (25Hz). The initialization of the tracker 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 7. 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 7 frame 200). The object is moved from its initial position using wide angle and wide range motions (figure 7 frame 400). And finally the object is subject to fast motion causing a significant blur in many images (figure 7 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 tracker. 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 4.2.2), the images are processed at video rate (25Hz). Using the fast computation (section 4.2.3) it is about 40Hz. All the corresponding sequences are presented in the attached video.

5.1.2 Evaluation on benchmark datasets

To have a quantitative measure of its accuracy and robustness, the tracker has been evaluated on some very demanding reference datasets proposed by Metaio GmbH [11]. 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 6 shows the results that have been obtained using the proposed approach. The tracker is considered in convergence 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 (29), a detailed definition is available in [11]. The mutual information based tracker proves its robustness and accuracy on most of the sequences.

The results obtained using the ESM approach [2] reported from [11] are also represented in the lower table of figure 6 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 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 tracker well suited for the augmented reality problem.

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 %

Figure 6: Ratio of successfully tracked images for our approach compared to the ESM [11].

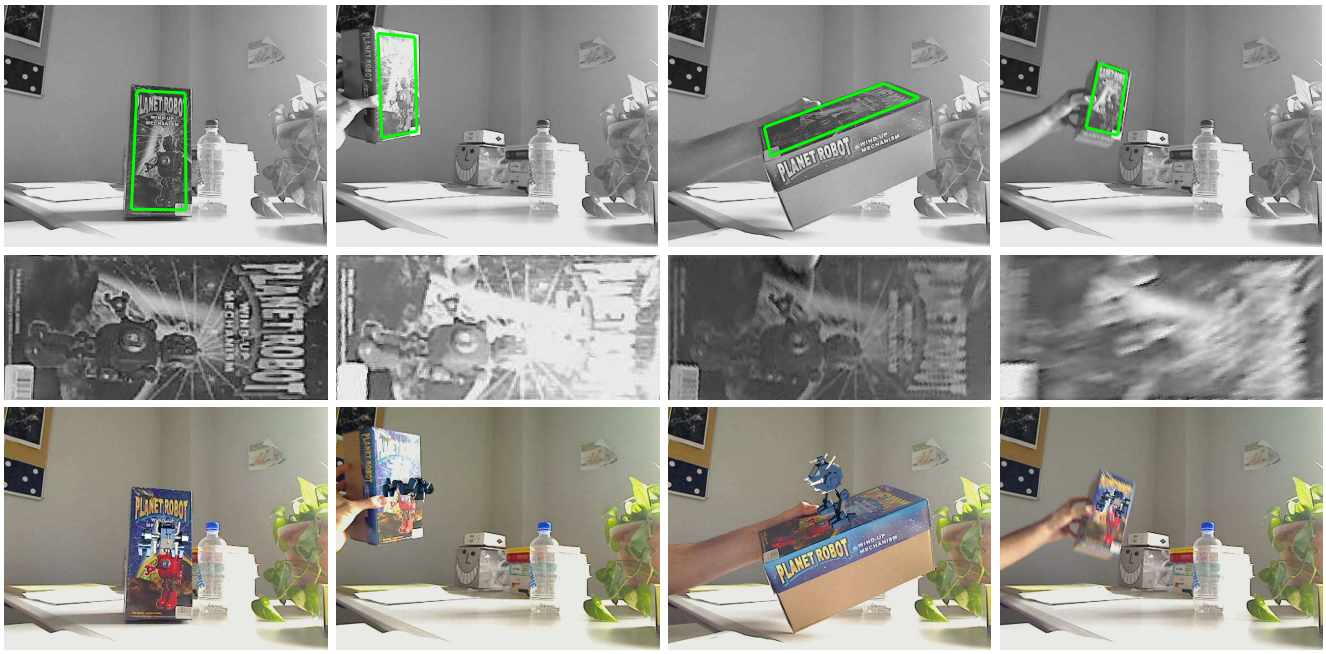

Figure 7: Tracking of a planar object through illumination variations. First row: frame 0, 200, 400 and 600. The green rectangle represents the rectangle from the template image transformed using the estimated homography. Second row: projection of the templates for the same iterations in the reference image. Third row: augmenting with a virtual robot placed on the top of the box.

5.2 Multimodal tracking

5.2.1 Satellite images versus map

This experiment illustrates the capabilities of the presented mutual information-based tracker 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 camera focusing on a poster representing the satellite image corresponding to the map.

As it has been previously noticed in figure 2, a non-linear relationship exists between the intensities of the map and aerial image and this link can be evaluated by the MI functions. Mutual information can therefore allow for tracking the satellite image using the map image. Figure 9 shows the reference image and some image 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 tracker converges despite some strong blurring effects. To validate the accuracy, we also used the estimated homography in an augmented reality application. Since the IGN map are linked with a GIS, some virtual information such as road, hydrographic network, or house footprint can be overlaid on the original satellite image in a consistent way.

5.2.2 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. The initial homography is manually defined.

As we can expect, although very different, the two images shown in figure 10 are sharing a lot of information and thus MI can handle the tracking of the infrared sequence. The warp function is still a homography. The satellite scene is then supposed to be planar leading to an approximation. Nevertheless the proposed method remains robust. No ground truth is available, but the overlaid images as well as the augmented reality application qualitatively validates the accuracy of the tracker. As figure 10 shows, the satellite image of the airport is well tracked on the sequence.

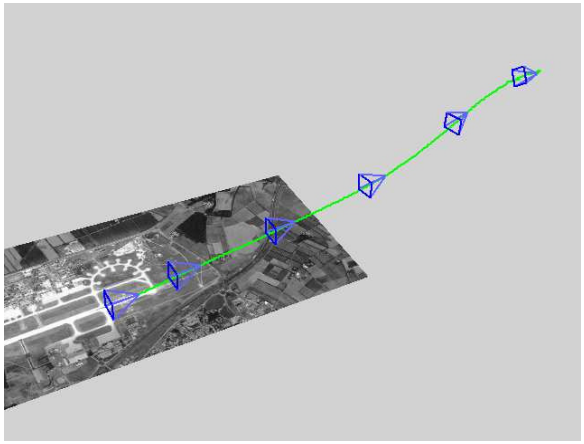

Figure 8: From the homography to the estimation of the camera position. Green curve: estimated camera trajectory in the 3D space, blue: the 6 estimated camera positions corresponding to the frames represented in figure 10.

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 8, 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 10 and 11 also shows some tracked images and some augmented images that validate the accuracy of the motion estimation. The complete sequences are visible in the attached video.

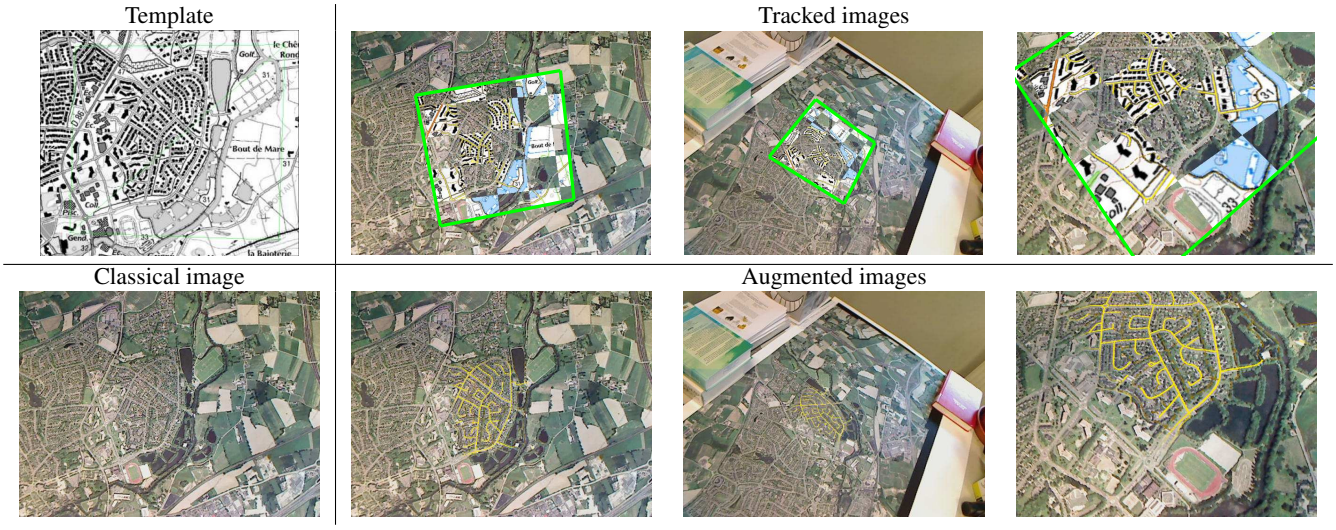

Figure 9: Tracking of an aerial sequence using a map template image by MI: frames 1, 250 and 500 are represented with the overimposed satellite reference (inside the green rectangle) projected using the estimated homography (image and map source: IGN) and augmented with the roads positions.

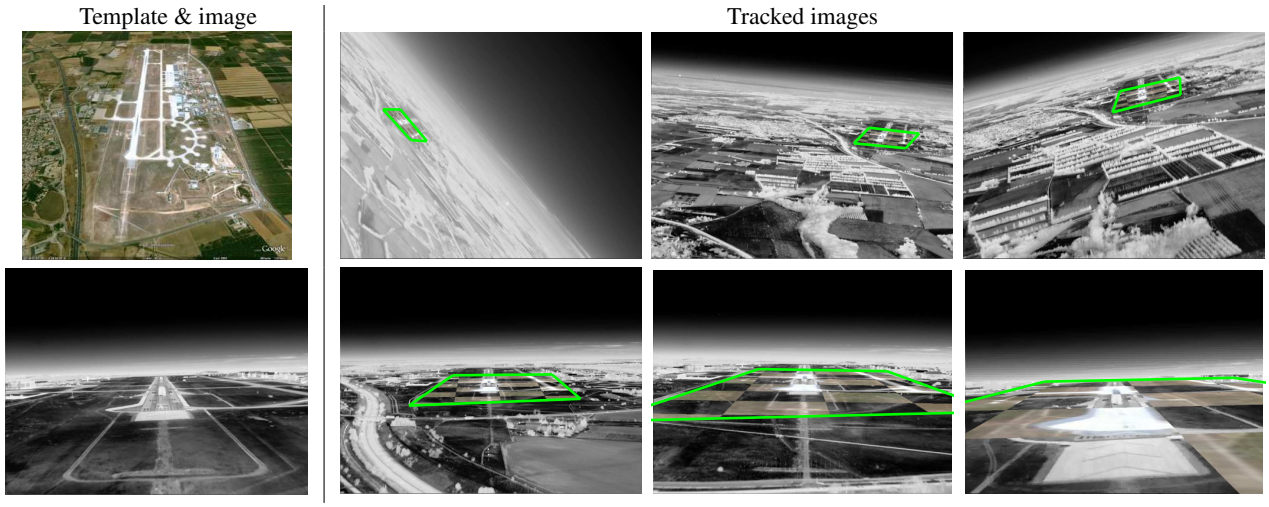

Figure 10: Tracking of a satellite template image using MI on an airborne infrared sequence. 6 frames are represented with the overimposed aerial reference (inside the green rectangle) projected using the estimated homography (Infrared images courtesy of Thales Optronic, optical image is obtained from google earth).

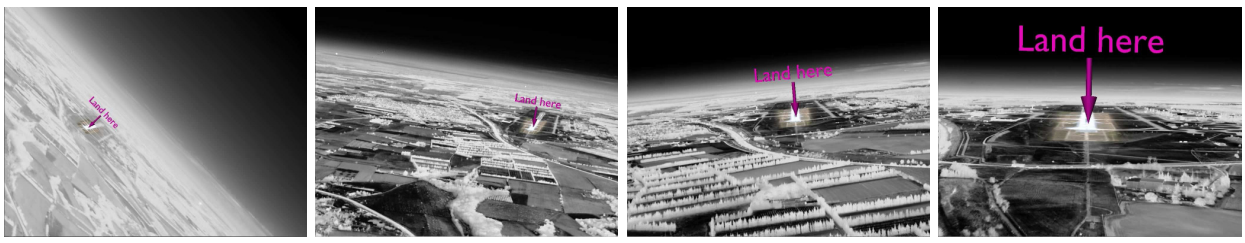

Figure 11: Augmenting the infrared images with the satellite appearance of the runway and an additional “land here” sign.

5.2.3 Potential AR applications of multimodal registration

Registering a map and an aerial image sequence is an extreme case, but registration between aerial and satellite (or any combination of such modalities), acquired at different time (and thus different) can be considered. Potential applications include visual odometry, aircraft or drone localization, pilot assistance, etc.

Infrared cameras (although still expensive) are widely used by civilians and, obviously, military aircraft. Such a registration process with a simple satellite image may prove to be very helpful for the pilots especially when landing (night or day) on a small and ILS free airport. Considering that aircraft position is fully known, additional information about runway, other aircraft positions or military targets may thus be easily displayed in the pilot helmet.

Although, we mentioned here applications in the aeronautic area, it is clear that other domains may be targeted such as energy monitoring, robotics, urbanism, architecture, defense, ...

6 Conclusion

This paper presented a robust and accurate template based-tracker that was defined using a new approach based on the mutual information alignment function. The definition of MI has been adapted to the differential tracking 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 tracker suitable for augmented reality applications.

Finally the proposed tracker 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 tracking. Some new applications are also proposed to use a model image acquired from another modality than the tracked sequence that are significant in flying, for example, in vehicle localization applications.

The algorithm presented here has been limited to planar object tracking. Nevertheless the proposed approach could similarly be applied to more complex model-based tracking applications where we could directly estimate the position of the object on $\mathbb{SE}(3)$. The method could also be extended to non-rigid object tracking.

Appendix

A Why the Hessian matrix must not be approximated

It is common to find the Hessian matrix of MI given in equation (22) approximated by the following expression [21][5]:

$$\mathbf{H} \simeq \sum_{r,t} \frac{\partial p_{II^*}}{\partial \Delta\mathbf{p}}^{\top} \frac{\partial p_{II^*}}{\partial \Delta\mathbf{p}} \left(\frac{1}{p_{II^*}} - \frac{1}{p_{I^*}} \right). \tag{31}$$

where the second order derivative of the joint probability has been neglected. The approximation is inspired from the one that is made in the Gauss-Newton's method for a least squared problem that is assuming that the neglected term is null after convergence.

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 (31) is positive. The goal is to maximize MI. The Hessian matrix after convergence would then be supposed to be negative by definition. The common approximation of (31) is thus not suited for the optimization of MI.

Acknowledgment

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

References

[1] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1090 – 1097, December 2001.

[2] S. Benhimane and E. Malis. Homography-based 2d visual tracking and servoing. *Int. Journal of Computer Vision*, 26(7):661–676, July 2007. Special IJCV/IJRR issue on vision for robots.

[3] 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*, volume 2, pages 142–149, 2000.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *IEEE Trans. on PAMI*, 28(9):1465–1479, Sept. 2006.

[11] S. Lieberknecht, S. Benhimane, P. G. Meier, and N. Navab. A dataset and evaluation methodology for template-based tracking algorithms. In G. Klinker, H. Saito, and T. Höllerer, editors, *ISMAR*, pages 145–151. IEEE Computer Society, 2009.

[12] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.

[13] 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.

[14] 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.

[15] G. Panin and A. Knoll. Mutual information-based 3d object tracking. *Int. Journal of Computer Vision*, 78(1):107–118, 2008.

[16] J. Pluim, J. Maintz, and M. Viergever. Mutual information matching and interpolation artefacts. In K. Hanson, editor, *SPIE Medical Imaging*, volume 3661, pages 56–65. SPIE Press, 1999.

[17] 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.

[18] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, December 1979.

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

[20] 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.

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

[22] P. Viola and W. Wells. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.