

L'information mutuelle pour l'estimation visuelle directe de pose

Guillaume Caron¹

Amaury Dame²

Eric Marchand³

¹ INRIA Rennes/IRISA, Lagadic, Rennes

² IRISA / CNRS, Lagadic, Rennes

³ Université de Rennes 1, IRISA, Lagadic, Rennes

IRISA/INRIA Rennes, campus universitaire de Beaulieu, 35042 Rennes Cedex
{guillaume.caron, amaury.dame}@inria.fr, eric.marchand@irisa.fr

Résumé

Cet article aborde l'estimation de pose basée modèle (ou localisation de caméra). Nous proposons une approche directe qui considère l'image comme un tout. Pour cela, nous exploitons une mesure de similarité, l'information mutuelle qui mesure la quantité d'information partagée par deux signaux, ou par deux images, dans notre cas.

L'avantage de cette mesure de similarité est de permettre de traiter des images de différentes modalités, en particulier, des images réelles et des images de synthèse. De plus, cette mesure gère les occultations et les changements d'illumination.

Les résultats avec des séquences d'images de synthèse (benchmark) et réelles, en caméra statique ou mobile, démontrent la robustesse de la méthode par rapport aux perturbations et la capacité de notre méthode à faire des estimations de pose stables et précises.

Mots Clef

Estimation de pose, suivi visuel, information mutuelle

Abstract

This paper deals with model-based pose estimation (or camera localization). We propose a direct approach that considers the image as a whole. For this, we consider a similarity measure, the mutual information. Mutual information is a measure of the quantity of information shared by two signals (or two images in our case).

The advantage of this similarity measure is to be able to deal with different image modalities. In particular one can consider a similarity between real and synthetic images. Furthermore, it can handle occlusions and illumination changes.

Results with synthetic (benchmark) and real image sequences, with static or mobile camera, demonstrate the robustness of the method with respect to perturbations and the ability of our method to produce stable and precise pose estimations.

Keywords

Pose estimation, visual tracking, mutual information

1 Introduction

Estimer la pose d'un objet, et le suivre, est critique pour de nombreuses applications en robotique telles que la localisation, la navigation, le positionnement, etc. La vision artificielle, dans ces contextes, présente un grand potentiel puisque les images apportent une information très riche sur l'environnement. Le problème d'estimation de pose d'objet est alors équivalent à la localisation de caméra. L'objectif de ce travail est de concevoir une nouvelle méthode d'estimation de pose de caméra.

La localisation de caméra a été très étudiée dans les dernières décennies. L'estimation de structure à partir du mouvement (SFM) avec optimisation par ajustement de faisceaux [29, 20] ou, dans la communauté robotique, la localisation et la cartographie visuelles simultanées (vS-LAM) [16, 18, 28], sont des approches usuelles pour estimer la pose, relative ou non, de la caméra. Ces approches reconstruisent l'environnement et estiment simultanément la position de la caméra mais a besoin de faire une boucle pour corriger la dérive. Enfin, l'odométrie visuelle est une autre technique permettant de retrouver la pose relative de la caméra [7] mais les estimations dérivent irrémédiablement.

Cependant, si un modèle 3D de l'environnement est déjà connu par le robot, les problèmes d'exploration et de fermeture de boucle sont évités. Dans [26], il a été montré que l'utilisation d'informations 3D sur l'environnement assure une meilleure précision dans l'estimation de pose. Cette information 3D rend l'estimation de pose de caméra, embarquée sur une plateforme mobile, précise et sans dérive, si le robot se déplace près de marqueurs référencés [11], ou même geo-référencés [13].

Depuis quelques années, de plus en plus de modèles 3D de villes ou d'environnements urbains sont disponibles. Le présent travail est réalisé dans le contexte du projet CityVIP de l'Agence Nationale de la Recherche (ANR) dans le cadre duquel l'Institut Géographique National (IGN) a nu-

mérisé les rues et les bâtiments du XII^{ème} arrondissement de Paris (Fig. 1(a)). Ainsi, nous avons pour objectif d'exploiter ce modèle 3D texturé pour localiser un véhicule par la vision, c'est-à-dire estimer la pose de la caméra dans la scène virtuelle en combinant les informations apportées par l'image réelle (Fig. 1(b)) et le monde virtuel (Fig. 1(c)), dans une approche multi-modale.

L'estimation de pose basée modèle est un problème abordé depuis plusieurs années en travaillant avec divers types de primitives : les points [21, 19], les droites [15], leur combinaison [25] ou encore des modèles 3D filaires [8]. Ces travaux ont exploité des primitives géométriques mais seuls certains autres travaux prennent en compte l'information photométrique explicitement dans l'estimation de pose et le suivi. Certains d'entre eux associent des primitives géométriques et photométriques [14, 24]. La primitive photométrique (les intensités de l'image) peut être considérée directement pour estimer l'homographie et ensuite la position relative entre l'image courante et celle de référence [4]. Une approche plus récente propose d'estimer une telle transformation en utilisant une approche fondée sur la théorie de l'information. Dans [9], l'information mutuelle partagée par un modèle plan texturé et des images acquises par la caméra est utilisé pour estimer une homographie (et ensuite une transformation 3D). Nous proposons de généraliser ce dernier travail aux modèles 3D en général, définis par un maillage, puisque c'est une manière courante en vision par ordinateur ou en synthèse d'images, de représenter une scène virtuelle. Par conséquent, ce travail formule

le problème d'estimation de pose comme la maximisation de l'information mutuelle partagée par une image réelle et une vue virtuelle rendue à partir d'une pose donnée.

La suite de l'article est organisée en trois parties principales. Premièrement, la formulation générale de l'estimation visuelle de pose basée modèle en un problème d'optimisation non linéaire est introduit dans la partie 2. Ensuite, dans la section 3, la maximisation de l'information mutuelle pour optimiser la pose est détaillée. Enfin, les résultats illustrent, dans les parties 4 et 5, le comportement de la méthode d'optimisation, sa précision et sa robustesse, avant de conclure.

2 Définition du problème d'estimation de pose

L'estimation de pose est considérée dans ce travail comme un problème d'optimisation non linéaire. Par conséquent, pour une nouvelle image acquise par la caméra, la pose est calculée en minimisant l'erreur entre les mesures réalisées dans l'image et la projection d'un modèle 3D de la scène pour une pose donnée. Puisque le mouvement de la caméra entre deux images est supposé faible, la pose obtenue pour l'image précédente est une bonne initialisation de la pose pour la nouvelle image. Le problème d'initialisation se retrouve seulement pour la première image acquise par la caméra. Ce problème est plus du ressort de la détection, de l'appariement et de la reconnaissance, et est en dehors de l'objet de cet article.

2.1 Estimation de pose basée primitive géométrique

L'estimation visuelle de pose est particulièrement connue à travers les approches basée primitives géométriques. Le but est, à partir d'une pose initiale arbitraire, que la caméra atteigne la pose désirée \mathbf{r}^* , telle que \mathbf{r} est une représentation vectorielle de la pose à six degrés de liberté ($\mathbf{r} = [t_X, t_Y, t_Z, \theta_X, \theta_Y, \theta_Z]$). \mathbf{r}^* doit satisfaire certaines propriétés mesurées dans les images. En considérant $\mathbf{s}(\mathbf{r})$, la projection de la scène 3D pour la pose \mathbf{r} , la méthode d'optimisation est conçue de manière à ce que l'erreur entre $\mathbf{s}(\mathbf{r})$ et \mathbf{s}^* (l'observation dans l'image) soit minimale. Le problème d'optimisation peut alors être écrit :

$$\hat{\mathbf{r}} = \arg \min_{\mathbf{r}} \|\mathbf{s}(\mathbf{r}) - \mathbf{s}^*\|. \quad (1)$$

Le modèle 3D est classiquement fait de primitives géométriques comme les points, les droites, etc. Dans ce cas, le problème principal est de déterminer dans chaque image les correspondances entre la projection du modèle et les primitives extraites de l'image \mathbf{s}^* , mais aussi de suivre ces primitives dans chaque image.

Les erreurs ou les imprécisions du suivi bas niveau engendrent des erreurs importantes dans le processus de suivi.

2.2 Estimation de pose directe

Pour éviter ces problèmes de suivi et d'appariement de primitives géométriques, et aussi la perte de précision que ces

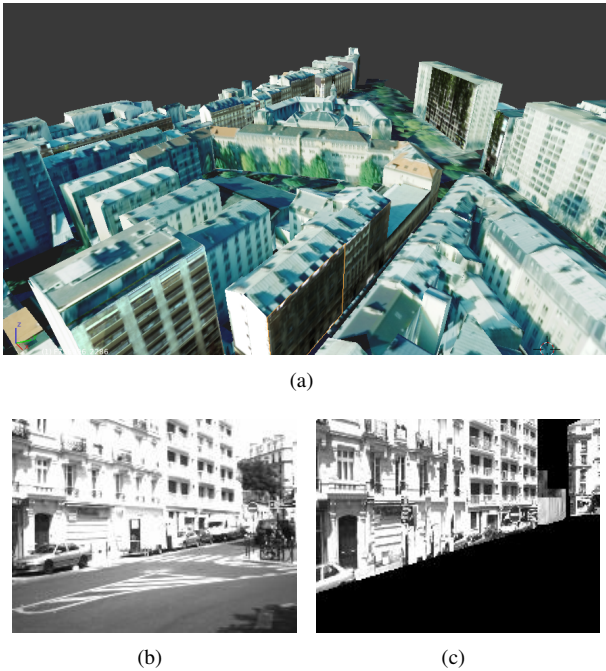


FIGURE 1 – (a) Le modèle 3D texturé du XII^{ème} arrondissement de Paris, (b) une image réelle acquise dans une rue et (c) la vue de synthèse correspondante.

approchent introduisent, d'autres formulations qui utilisent les images comme un tout doivent être proposées. Notons que des approches directes ont été largement étudiées pour le suivi 2D ou l'estimation de mouvement [3, 4]. Dans ces approches, l'idée est de minimiser directement l'erreur, la somme des carrés des différences (ou SSD), entre la région de référence d'une image \mathbf{I}^* et l'image courante \mathbf{I} transférée dans l'espace de la région de référence en utilisant un modèle de mouvement donné (généralement une homographie).

Théoriquement, en admettant qu'un modèle 3D de la scène soit disponible, ce procédé peut se transposer à celui de l'estimation de pose. En effet, dans ce cas, la pose peut être déterminée en minimisant l'erreur entre l'image acquise par la caméra \mathbf{I}^* et la projection de la scène pour une pose donnée $\mathbf{I}(\mathbf{r})$. La fonction de coût peut s'écrire sous la forme :

$$\hat{\mathbf{r}} = \arg \min_{\mathbf{r}} \sum_{\mathbf{x}} (\mathbf{I}(\mathbf{r}, \mathbf{x}) - \mathbf{I}^*(\mathbf{x}))^2. \quad (2)$$

Dans l'équation (2), $\mathbf{I}(\mathbf{r})$ peut être obtenu en utilisant un moteur de rendu (comme OpenGL). Ce modèle virtuel, même tapissé de textures photoréalistes est rendu par un quelconque moteur 3D et l'image obtenue n'est autre qu'une image de synthèse. Par conséquent, même si la fonction de coût de l'équation (2) ne comporte pas de suivi ou d'appariement de primitives, les variations d'illumination ou les occultations ont un grand impact sur la fonction de coût engendrant l'échec du suivi visuel.

Nous proposons de formuler un autre critère d'optimisation qui compare directement et entièrement les images courante et désirée. Plutôt que d'utiliser une fonction de coût basée SSD, nous définissons une fonction d'alignement entre les deux images : l'Information Mutuelle (MI) entre $\mathbf{I}(\mathbf{r})$ et \mathbf{I}^* [27, 30]. La fonction MI est une mesure de la quantité d'information partagée par deux images. Quand la fonction MI est maximale, les deux images sont recalées. La mesure de similarité MI a été utilisée dans des travaux de recalage [30] et plus récemment pour suivre des plans dans des séquences d'images [9]. Cette primitive s'est montrée robuste au bruit, aux réflexions spéculaires et même à différentes modalités entre l'image de référence et la courante. Cette dernière caractéristique est particulièrement intéressante dans notre travail puisque nous voulons aligner une vue de synthèse sur une image réelle.

Notons, de plus, que l'estimation de pose et l'asservissement visuel sont des problèmes duaux [8] et qu'en asservissement visuel, les approches directes utilisant un critère de type SSD [6] ou MI [10] se sont montrées très efficaces et ont inspiré ce travail.

Nous proposons alors une extension significative de [9, 10] au cas de l'estimation de pose basée modèle non plan et au suivi.

3 L'information mutuelle sur SE(3)

Comme abordé en section 2, les fonctions de coût plus ou moins classiques (eq. (1) et (2)) doivent être reformulées.

L'objectif est de réaliser le recalage du modèle par rapport à l'image et cela peut être formulé comme l'optimisation de l'information mutuelle partagée par l'image de référence \mathbf{I}^* et la projection du modèle \mathcal{M} . Si γ est l'ensemble des paramètres intrinsèques de la caméra (focale et point principal) et \mathbf{r} sa pose, le problème d'estimation peut s'écrire de la façon suivante [22] :

$$\hat{\mathbf{r}} = \arg \max_{\mathbf{r}} \text{MI}(\mathbf{I}^*, \mathbf{I}_{\gamma}(\mathcal{M}, \mathbf{r})). \quad (3)$$

L'image $\mathbf{I}_{\gamma}(\mathcal{M}, \mathbf{r})$ est le résultat de la projection du modèle \mathcal{M} à la pose donnée \mathbf{r} . Par une approximation de Taylor au premier ordre de la fonction MI à la pose courante \mathbf{r} , le lien entre la variation de l'information mutuelle et la variation de la pose est exprimé. L'incrément à appliquer à la pose est ensuite obtenu par une méthode d'optimisation de type Newton.

Pour résoudre cette fonction, un modèle 3D texturé de l'objet à suivre est nécessaire et doit être projeté pour chaque pose de caméra \mathbf{r} . Pour générer les images du modèle 3D, nous avons utilisé OpenGL comme moteur de rendu 3D et plus particulièrement la librairie Ogre3D [1]. OpenGL permet non seulement de générer des images de synthèse mais aussi des images de profondeur. Plus précisément, on obtient une image où chaque pixel contient la coordonnée en Z du point 3D projeté en ce pixel. C'est particulièrement intéressant puisque le Z de chaque point visible apparaît dans la matrice jacobienne liant les variations de l'information mutuelle et de la pose (cf. 3.3).

3.1 L'information mutuelle

La fonction MI est définie à partir de l'entropie \mathbf{H} de deux images \mathbf{I} et \mathbf{I}^* et leur entropie jointe :

$$\text{MI}(\mathbf{I}, \mathbf{I}^*) = \mathbf{H}(\mathbf{I}) + \mathbf{H}(\mathbf{I}^*) - \mathbf{H}(\mathbf{I}, \mathbf{I}^*) \quad (4)$$

Les entropies $\mathbf{H}(\mathbf{I})$ et $\mathbf{H}(\mathbf{I}^*)$ et l'entropie jointe $\mathbf{H}(\mathbf{I}, \mathbf{I}^*)$ sont une mesure de variabilité d'une, respectivement de deux, variable aléatoire \mathbf{I} , respectivement \mathbf{I} et \mathbf{I}^* . Pour $\mathbf{H}(\mathbf{I})$, si i sont les valeurs possibles de $\mathbf{I}(\mathbf{x})$ ($i \in [0, N_c]$) avec $N_c = 255$ et si $p_{\mathbf{I}}(i) = Pr(\mathbf{I}(\mathbf{x}) = i)$ est la densité de probabilité de i , alors l'entropie au sens de Shannon, $\mathbf{H}(\mathbf{I})$, d'une variable discrète \mathbf{I} est donnée par :

$$\mathbf{H}(\mathbf{I}) = - \sum_{i=0}^{N_c} p_{\mathbf{I}}(i) \log(p_{\mathbf{I}}(i)). \quad (5)$$

D'une manière similaire, on a l'expression de l'entropie jointe :

$$\mathbf{H}(\mathbf{I}, \mathbf{I}^*) = - \sum_{i=0}^{N_c} \sum_{j=0}^{N_{c^*}} p_{\mathbf{II}^*}(i, j) \log(p_{\mathbf{II}^*}(i, j)). \quad (6)$$

3.2 Optimisation de pose basée information mutuelle

Il est bien connu que les rotations et les translations de la caméra sont corrélées, avec, par exemple, et de manière

évidente, pour une translation selon l'axe des X et une rotation autour des Y . Par conséquent, une simple optimisation par descente de gradient utilisant la direction donnée par la jacobienne de l'image liée à l'information mutuelle ne fournirait pas une estimation précise de l'optimum de la fonction MI. Alors, une technique d'optimisation du second ordre de type Newton est nécessaire.

En utilisant une approximation de Taylor au premier ordre de la fonction de similarité MI, à la pose courante \mathbf{r}_k dans un schéma d'estimation non linéaire de pose donne :

$$\text{MI}(\mathbf{r}_{k+1}) \approx \text{MI}(\mathbf{r}_k) + \mathbf{L}_{\text{MI}}^T \dot{\mathbf{r}} \Delta_t. \quad (7)$$

Δ_t est le laps de temps nécessaire pour transformer \mathbf{r}_k en \mathbf{r}_{k+1} en utilisant la variation de la pose $\dot{\mathbf{r}}$ (qui peut être vue comme la vitesse de la caméra virtuelle $\mathbf{v} = \dot{\mathbf{r}}$). La pose est mise à jour grâce à $e^{[\mathbf{v}]}$, l'application exponentielle sur SE(3) :

$$\mathbf{r}_{k+1} = e^{[\mathbf{v}]} \mathbf{r}_k. \quad (8)$$

\mathbf{L}_{MI} (éq. (7)) est la jacobienne de l'image (ou la matrice d'interaction dans un contexte d'asservissement visuel [5]) liée à la fonction MI, c'est-à-dire la matrice liant la variation de la fonction MI et la variation de la pose. Cela mène à :

$$\mathbf{L}_{\text{MI}}^T(\mathbf{r}_{k+1}) \approx \mathbf{L}_{\text{MI}}^T(\mathbf{r}_k) + \mathbf{H}_{\text{MI}}(\mathbf{r}_k) \mathbf{v} \Delta_t, \quad (9)$$

où $\mathbf{H}_{\text{MI}}(\mathbf{r}_k)$ est la matrice d'interaction de la matrice d'interaction MI : la matrice hessienne de la fonction MI, par abus de notation. Le but est de maximiser l'information mutuelle alors nous voulons que le système atteigne la pose \mathbf{r}_{k+1} pour laquelle la variation de la fonction MI par rapport à la variation de la pose soit nulle : $\mathbf{L}_{\text{MI}}(\mathbf{r}_{k+1}) = 0$. En fixant $\Delta_t = 1$ dans l'équation (9), sans perte de généralité, l'incrément approché qui mène à une variation nulle de MI est :

$$\mathbf{v} = -\mathbf{H}_{\text{MI}}^{-1}(\mathbf{r}_k) \mathbf{L}_{\text{MI}}^T(\mathbf{r}_k). \quad (10)$$

À la manière de [9], dans le but d'avoir une bonne estimation de la hessienne après convergence, plutôt qu'utiliser la hessienne $\mathbf{H}_{\text{MI}}^{-1}(\mathbf{r}_k)$, nous utilisons $\mathbf{H}_{\text{MI}}^{*-1}$ estimée à la pose désirée \mathbf{r}^* ($\mathbf{H}_{\text{MI}}^{*-1} = \mathbf{H}_{\text{MI}}^{-1}(\mathbf{r}^*)$) :

$$\mathbf{v} = -\mathbf{H}_{\text{MI}}^{*-1} \mathbf{L}_{\text{MI}}^T. \quad (11)$$

\mathbf{L}_{MI} est la matrice d'interaction liée à la fonction MI calculée à la position courante \mathbf{r}_k . Bien entendu, la pose optimale \mathbf{r}^* est inconnue mais la matrice hessienne à l'optimum $\mathbf{H}_{\text{MI}}^{*-1}$ peut être estimée sans connaître \mathbf{r}^* , en posant $Z = Z^*$ (cf. fin de la partie 3.3) puisque les poses consécutives sont proches.

3.3 La matrice d'interaction de l'information mutuelle

Sachant les expressions de l'entropie et de l'entropie jointe (éq. (5) et (6)), la fonction MI (éq. (4)) se développe en :

$$\text{MI}(\mathbf{I}, \mathbf{I}^*) = \sum_{i,j} p_{\text{II}^*}(i,j) \log \left(\frac{p_{\text{II}^*}(i,j)}{p_{\text{I}}(i)p_{\text{I}^*}(j)} \right). \quad (12)$$

À partir de l'équation (12) et de simplifications permises par la règle de dérivation en chaîne [12], l'expression de la matrice d'interaction \mathbf{L}_{MI} et de la hessienne \mathbf{H}_{MI} sont :

$$\mathbf{L}_{\text{MI}} = \sum_{i,j} \mathbf{L}_{p_{\text{II}^*}} \left(1 + \log \left(\frac{p_{\text{II}^*}}{p_{\text{I}^*}} \right) \right) \quad (13)$$

$$\text{et } \mathbf{H}_{\text{MI}} = \sum_{i,j} \mathbf{L}_{p_{\text{II}^*}}^T \mathbf{L}_{p_{\text{II}^*}} \left(\frac{1}{p_{\text{II}^*}} - \frac{1}{p_{\text{I}^*}} \right) + \mathbf{H}_{p_{\text{II}^*}} \left(1 + \log \left(\frac{p_{\text{II}^*}}{p_{\text{I}^*}} \right) \right), \quad (14)$$

où l'ensemble des valeurs possibles a été omis par souci de clarté. La mesure MI impose le calcul complet de la matrice hessienne (pas d'approximation comme c'est généralement le cas avec les primitives plus conventionnelles) [9]. Pour respecter les conditions de dérivation de la fonction MI, les probabilités $p_{\text{I}}(i)$ sont interpolées par des fonctions B-splines, notées ϕ dans un souci de clarté. Ainsi, la formulation analytique finale de $p_{\text{I}}(i)$, qui peut être considérée comme un histogramme normalisé, devient :

$$p_{\text{I}}(i) = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \phi(i - \bar{\mathbf{I}}(\mathbf{x})), \quad (15)$$

où les valeurs possibles de niveau de gris sont maintenant $\bar{\mathbf{I}}(\mathbf{x}) \in [0, N_c - 1]$. Par conséquent, cette formulation permet aussi la réduction du nombre d'entrées de l'histogramme [23] afin de réduire la dimensionnalité du problème mais aussi de lisser le profil de la fonction de coût MI [10].

Ainsi, à partir de l'expression de la probabilité jointe :

$$p_{\text{II}^*}(i,j,\mathbf{r}) = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \phi(i - \bar{\mathbf{I}}(\mathbf{x},\mathbf{r})) \phi(j - \bar{\mathbf{I}}^*(\mathbf{x})), \quad (16)$$

on déduit ses variations par rapport à la pose de la caméra, c'est-à-dire la matrice d'interaction (éq. (13)) et la hessienne (eq. (14)) :

$$\mathbf{L}_{p_{\text{II}^*}(i,j,\mathbf{r})} = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \mathbf{L}_{\phi(i - \bar{\mathbf{I}}(\mathbf{x},\mathbf{r}))} \phi(j - \bar{\mathbf{I}}^*(\mathbf{x})) \quad (17)$$

$$\mathbf{H}_{p_{\text{II}^*}(i,j,\mathbf{r})} = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \mathbf{H}_{\phi(i - \bar{\mathbf{I}}(\mathbf{x},\mathbf{r}))} \phi(j - \bar{\mathbf{I}}^*(\mathbf{x})) \quad (18)$$

La variation de ϕ s'obtient par la règle de dérivation en chaîne :

$$\mathbf{L}_{\phi(i - \bar{\mathbf{I}}(\mathbf{x},\mathbf{r}))} = -\frac{\partial \phi}{\partial i} \mathbf{L}_{\bar{\mathbf{I}}} \quad (19)$$

$$\mathbf{H}_{\phi(i - \bar{\mathbf{I}}(\mathbf{x},\mathbf{r}))} = \frac{\partial^2 \phi}{\partial i^2} \mathbf{L}_{\bar{\mathbf{I}}}^T \mathbf{L}_{\bar{\mathbf{I}}} - \frac{\partial \phi}{\partial i} \mathbf{H}_{\bar{\mathbf{I}}}. \quad (20)$$

En admettant une scène lambertienne, au moins pour de faibles déplacements, la matrice d'interaction de l'intensité d'un point $\mathbf{L}_{\bar{\mathbf{I}}}$ et sa matrice hessienne $\mathbf{H}_{\bar{\mathbf{I}}}$ sont obtenus par [6] :

$$\mathbf{L}_{\bar{\mathbf{I}}} = \nabla \bar{\mathbf{I}} \mathbf{L}_{\mathbf{x}} \quad (21)$$

$$\mathbf{H}_{\bar{\mathbf{I}}} = \mathbf{L}_{\mathbf{x}}^T \nabla^2 \bar{\mathbf{I}} \mathbf{L}_{\mathbf{x}} + \nabla_x \bar{\mathbf{I}} \mathbf{H}_x + \nabla_y \bar{\mathbf{I}} \mathbf{H}_y, \quad (22)$$

où $\nabla \bar{\mathbf{I}} = (\nabla_x \bar{\mathbf{I}}, \nabla_y \bar{\mathbf{I}})$ sont les gradients de l'image, $\nabla^2 \bar{\mathbf{I}}$ sont les gradients des gradients de l'image et \mathbf{L}_x est la matrice d'interaction d'un point qui lie son déplacement dans le plan image normalisé à la vitesse de la caméra. \mathbf{H}_x et \mathbf{H}_y sont les matrices hessiennes des deux coordonnées d'un point par rapport à la vitesse de la caméra. La matrice d'interaction \mathbf{L}_x est donnée par [5] :

$$\mathbf{L}_x = \begin{bmatrix} -1/Z & 0 & x/Z & xy & -(1+x^2) & y \\ 0 & -1/Z & y/Z & 1+y^2 & -xy & -x \end{bmatrix}. \quad (23)$$

La matrice d'interaction dépend à la fois de la position (x, y) du point dans le plan image normalisé et de sa profondeur Z dans le repère caméra. Z est obtenu à partir du rendu fait par le moteur 3D de notre modèle 3D texturé, en utilisant le Z-buffer traditionnel, pour chaque point.

Les matrices hessiennes \mathbf{H}_x et \mathbf{H}_y du point sont données dans [17] mais ne sont pas rappelées ici à cause de la taille des matrices.

Pendant le processus itératif de l'optimisation, la caméra virtuelle se déplace, conduisant au changement de profondeur de chaque point. Les matrices d'interaction et hessienne changent alors à chaque itération. La connaissance de Z^* est nécessaire (éq. (11)), alors on admet que la profondeur des points entre les poses courante et désirée sont assez similaires et on pose $Z^* = Z$, puisque les poses consécutives sont proches. Ainsi, à convergence, l'estimation de \mathbf{H}^* sera précise.

L'algorithme présenté en figure 2 résume tous les processus de l'estimation de pose et le suivi basés information mutuelle.

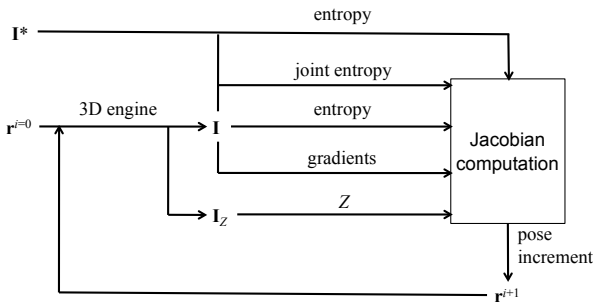


FIGURE 2 – Synoptique de l'algorithme d'estimation de pose basé information mutuelle. Le processus boucle jusqu'à ce que l'information mutuelle partagée par \mathbf{I} et \mathbf{I}^* soit stable.

4 Résultats de simulation

Notre méthode d'estimation de pose basée information mutuelle a été évaluée sur une séquence d'images de synthèse. Un jeu de données d'un benchmark de TrakMark [2] est utilisé pour cela (Fig. 3). Le jeu de données "Conference Venue Package 01" est utilisé et le mouvement de caméra dans la séquence de référence est composé de translations 3D, et de rotations selon l'axe X et l'axe Y de la caméra.



(a) image 0

(b) image 725

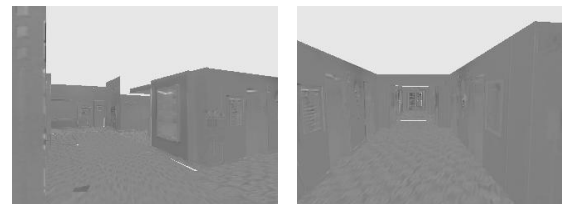
FIGURE 3 – Images pour l'évaluation de l'estimation de pose basée information mutuelle en simulation.

Notre algorithme a réussi à retrouver les poses de la caméra tout au long de la séquence de test de 1210 images. La précision d'estimation est évaluée aussi bien dans l'image qu'en 3D par la suite.

Les différences d'images, entre les images de références d'un jeu de données et les images obtenues aux poses optimales calculées par notre méthode, sont un bon moyen d'évaluer qualitativement les poses estimées. Les différences d'images doivent être grises quand les deux images sont identiques. La figure 4 présente quelques différences d'images en plusieurs endroits le long de la trajectoire. Puisque nos images virtuelles ne sont pas rendues à l'aide du même moteur 3D que celui utilisé par TrakMark pour générer les jeux de données, certaines propriétés de rendu (filtrage de texture, couleurs, etc) sont différentes. C'est pourquoi les différences d'images ne sont pas parfaitement grises à convergence. Notons cependant que malgré ce problème, le recalage réussit. Ce succès est dû à la mesure d'information mutuelle qui est robuste à de tels problèmes, tandis que les fonctions de coût plus classiques, comme la SSD, ne le sont pas.

Après avoir évalué les résultats qualitativement dans les images, l'évaluation des estimations est faite quantitativement en 3D. La trajectoire estimée est extrêmement proche de la vérité terrain qui est incluse dans un volume de $5m \times 8m \times 0.7m$ (cf. les erreurs en translation et en rotation de la figure 5). L'erreur en translation est la norme de la différence entre la translation référence et l'estimée. L'erreur en rotation est calculée comme suit, en considérant \mathbf{R}^* la matrice de rotation de la vérité terrain et \mathbf{R} l'estimée :

1. calcul de la matrice de "différence" de rotation $\mathbf{R}_d = \mathbf{R}^* \mathbf{R}^T$



(a) différence 0

(b) différence 725

FIGURE 4 – Différence entre les images de références et les images optimales.

2. décomposition de R_d en un axe et un angle de rotation à l'aide de la formule de Rodrigues
3. L'erreur en rotation entre la vérité terrain et l'estimation est la valeur absolue de cet angle

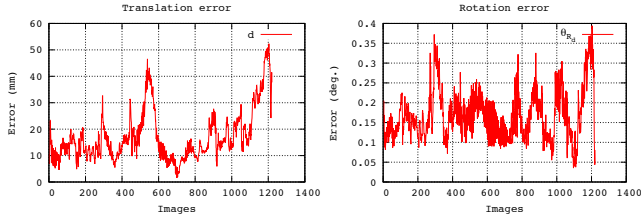


FIGURE 5 – Erreurs d'estimation en (a) position et en (b) orientation sur toute la séquence, par rapport à la vérité terrain.

5 Résultats expérimentaux

5.1 Validation sur une “boîte de thé”

La première évaluation sur images réelles est menée avec une caméra statique, dans le champ de vue de laquelle une boîte est déplacée avec un mélange de translations et de rotations. Les faces de la boîte ont été scannées pour tapisser de textures de modèle 3D. La boîte réelle, dans la séquence, présente une illumination différente du modèle avec aussi des réflexions spéculaires et des occultations partielles avec les doigts. Malgré ces perturbations, le suivi est un succès tout au long des 500 images de la séquence. La figure 6 présente trois extraits de la séquence d'images avec, pour chaque, l'image réelle, l'image de synthèse à la pose optimale et le z-buffer nécessaire au calcul de la partie géométrique de la matrice d'interaction (éq. (23)).

Les différences d'images de la figure 6 permettent d'évaluer la qualité du suivi puisque nous n'avons pas de vérité terrain pour cette expérimentation. Notons cependant que le modèle virtuel est parfaitement aligné avec la boîte réelle dans l'image, quel qu'en soit l'orientation.

Comme il a été mentionné au début de l'article, le but du travail est de réaliser un suivi et une estimation de pose directs, et non pas la détection et la mise en correspondance. Ainsi, pour la première image de la séquence, la pose initiale de l'objet est manuellement déterminée ou de façon semi automatique en cliquant sur quatre sommets de la boîte, puis en faisant une estimation de pose classique. Cette pose initiale déterminée manuellement est généralement plus éloignée de l'optimale que dans le processus de suivi où la pose initiale pour une nouvelle image est l'optimale pour l'image précédente. Dans cette expérimentation, la pose initiale manuelle est distante de 2.5cm et 3.3° de l'optimale. Il est intéressant d'observer l'évolution de la fonction MI en fonction des itérations de l'optimisation de la pose (Fig. 7) puisqu'elle est lisse et atteint de façon logarithmique sa valeur maximale.

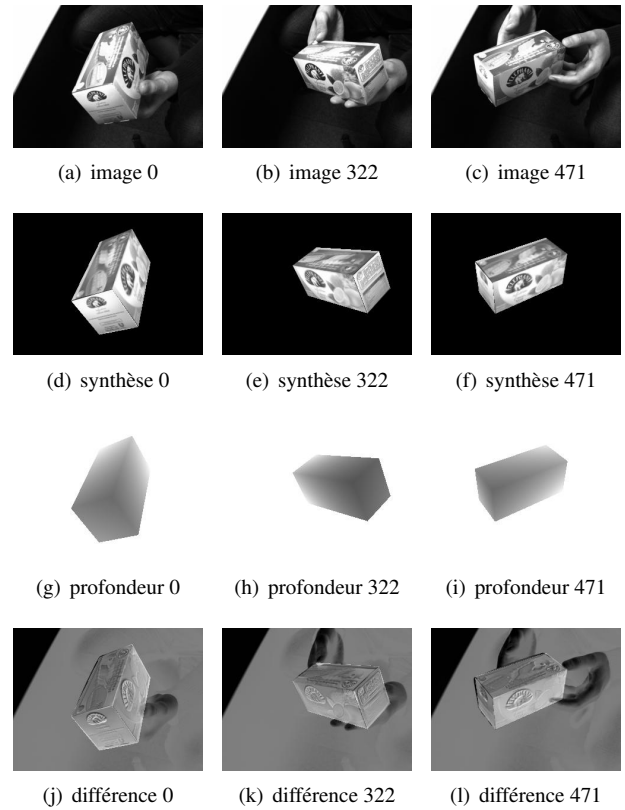


FIGURE 6 – Suivi d'une boîte de thé dans 500 images. (a-c) Trois images sur lesquelles (d-f) la vue de synthèse est recalée, avec l'information du Z de chaque pixel de l'objet obtenu à partir du (g-i) buffer (tampon) de profondeur. Pour évaluer la précision du suivi, la différence entre les images réelle et de synthèse est calculée (j-l).

5.2 Application à la localisation de véhicule

Un autre objectif de notre méthode d'estimation de pose basée information mutuelle est d'estimer la pose d'une caméra en mouvement, embarquée sur un véhicule, en utilisant ses images (Fig. 8) et un modèle 3D texturé de la ville dans laquelle la voiture est conduite. Contrairement à l'expérimentation de la boîte de thé où aucune vérité terrain n'est disponible, on peut superposer la trajectoire estimée

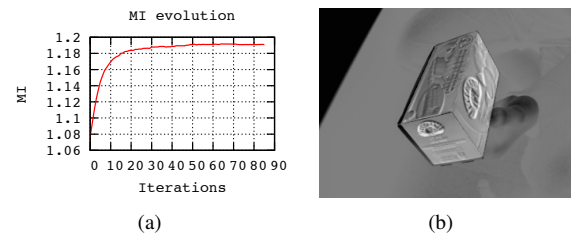


FIGURE 7 – (a) Evolution de l'information mutuelle en fonction des itérations pour la première image de la séquence. (b) La pose initiale est distante de 2.5cm et 3.3° de l'optimale.



FIGURE 8 – Exemples d’images de la séquence du XII^{ème} arrondissement de Paris avec occultations des bâtiments par des gens et par des voitures en mouvement ou non (voir aussi Fig. 1). (c) montre un cas où l’algorithme diverge à cause d’une occultation trop importante (75% de l’image) par rapport au modèle, par un store et une camionnette.

sur la vue satellite de la ville pour évaluer la précision d’estimation (Fig. 9). Le fait que la trajectoire estimée soit bien alignée avec la rue et bien au centre (rue à sens unique) montre la stabilité et la précision des poses estimées, en dépit des occultations des bâtiments par les voitures (Fig. 1(b) et 1(c)), les changements d’illumination, les vibrations de la camera ou le virage au début de la séquence (en bas à gauche de la figure 9).

6 Conclusion et travaux futurs

Nous avons abordé une nouvelle méthode de suivi visuel direct et d’estimation de pose impliquant la mesure d’information mutuelle partagée par deux images : une image réelle de référence et une vue virtuelle évoluant à mesure que la pose est optimisée, en maximisant l’information mutuelle. La difficulté a été de réussir à lier la variation de la mesure d’information mutuelle à la variation de la pose de la caméra ou de l’objet. Les résultats montrent, en simulation comme dans les conditions réelles, en particulier avec une caméra embarquée sur une voiture, que la méthode est robuste et très précise.

L’implémentation actuelle a cependant l’inconvénient de ne pas être temps-réel avec environ quatre secondes de traitement pour chaque image. Cependant, une implémentation en multi-résolution avec un schéma incrémental de niveau de transformation, tout cela implémenté sur GPU peut réduire considérablement le temps de traitement. Nous allons aussi aborder le problème de modèle 3D texturé d’une faible qualité dans nos prochains travaux.

Remerciements

Les auteurs remercient la contribution de l’Agence Nationale de la Recherche (ANR) pour les projets CityVIP et ReV-TV et la contribution de la DGA (Direction Générale de l’Armement).

Références

- [1] *Ogre3d, open source 3d graphics engine*, <http://www.ogre3d.org>.
- [2] *Trakmark benchmarking*, <http://trakmark.net>.

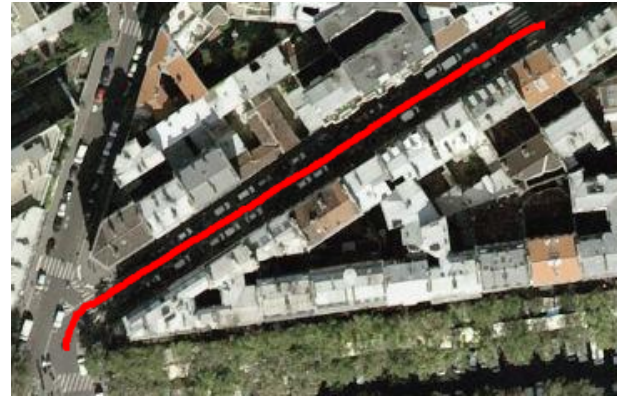


FIGURE 9 – Chemin (en rouge) obtenu par l’estimation de pose basée information mutuelle, sans aucun filtrage. La stabilité d’estimation de la trajectoire est évidente et sa précision est montrée par le fait que la trajectoire soit bien alignée sur la rue et qu’il n’y ait pas de chevauchement de la trajectoire sur les voitures garées ou les bâtiments.

- [3] S. Baker and I. Matthews, *Lucas-kanade 20 years on : A unifying framework*, *IJCV* **56** (2004), no. 3, 221–255.
- [4] S. Benhimane and E. Malis, *Real-time image-based tracking of planes using efficient second-order minimization*, *IEEE/RSJ Int. Conf. on Intelligent Robots Systems (Sendai, Japan)*, vol. 943-948, October 2004, p. 1.
- [5] F. Chaumette and S. Hutchinson, *Visual servo control, part i : Basic approaches*, *IEEE Robotics and Automation Magazine* **13** (2006), no. 4, 82–90.
- [6] C. Collewet and E. Marchand, *Photometric visual servoing*, *IEEE Trans. on Robotics* **27** (2011), no. 4, 828–834.
- [7] A.I. Comport, E. Malis, and P. Rives, *Real-time quadrifocal visual odometry*, *Int. J. of Robotics Research, Special issue on Robot Vision* **29** (2010), no. 2-3, 245–266.
- [8] A.I. 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** (2006), no. 4, 615–628.
- [9] A. Dame and E. Marchand, *Accurate real-time tracking using mutual information*, *IEEE Int. Symp. on Mixed and Augmented Reality, ISMAR (Seoul, Korea)*, October 2010, pp. 47–56.
- [10] _____, *Mutual information-based visual servoing*, *IEEE Trans. on Robotics* **27** (2011), no. 5.
- [11] P. David, *Vision-based localization in urban environments*, *Army Science Conference*, november 2010, pp. 428–433.

- [12] N. Dowson and R. Bowden, *A unifying framework for mutual information methods for use in non-linear optimisation*, European Conf. Computer Vision (Graz, Austria), may 2006, pp. 365–378.
- [13] E. Frontoni, A. Ascani, A. Mancini, and P. Zingaretti, *Robot localization in urban environments using omnidirectional vision sensors and partial heterogeneous a priori knowledge*, Int. Conf. on Mechatronics and Embedded Systems and Applications, MESA (Qingdao, China), july 2010, pp. 428–433.
- [14] P. Georgel, S. Benhimane, and N. Navab, *A unified approach combining photometric and geometric information for pose estimation*, British Machine Vision Conf., BMVC, september 2008.
- [15] B. Jiang, *Calibration-free line-based tracking for video augmentation*, Int. Conf. on Computer Graphics & Virtual Reality, CGVR (Las Vegas, USA), june 2006, pp. 104–110.
- [16] N. Karlsson, E. Di Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian, and M.E. Munich, *The vslam algorithm for robust localization and mapping*, IEEE Int. Conf. on Robotics and Automation (Barcelona, Spain), April 2005.
- [17] J.T. Lapresté and Y. Mezouar, *A Hessian approach to visual servoing*, IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS (Sendai, Japan), September 2004, pp. 998–1003.
- [18] T. Lemaire and S. Lacroix, *Monocular-vision based SLAM using line segments*, Int. Conf. on Robotics and Automation (Roma, Italy), 2007.
- [19] V. Lepetit and P. Fua, *Monocular model-based 3d tracking of rigid objects : A survey*, Foundations and Trends in Computer Graphics and Vision **1** (2005), no. 1, 1–89.
- [20] Maxime Lhuillier, *Automatic scene structure and camera motion using a catadioptric system*, Computer Vision and Image Understanding **109** (2008), no. 2, 186–203.
- [21] E. Marchand and F. Chaumette, *Virtual visual servoing : a framework for real-time augmented reality*, EUROGRAPHICS’02 Conf. Proceeding (Saarebrücken, Germany) (G. Drettakis and H.-P. Seidel, eds.), Computer Graphics Forum, vol. 21(3), September 2002, pp. 289–298.
- [22] G. Panin and A. Knoll, *Mutual information-based 3d object tracking*, Int. J. Comput. Vision **78** (2008), 107–118.
- [23] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever, *Mutual information matching and interpolation artefacts*, SPIE Medical Imaging (K.M. Hanson, ed.), vol. 3661, SPIE Press, 1999, pp. 56–65.
- [24] M. Pressigout and E. Marchand, *Real-time hybrid tracking using edge and texture information*, Int. Journal of Robotics Research, IJRR **26** (2007), no. 7, 689–713.
- [25] E. Rosten and T. Drummond, *Fusing points and lines for high performance tracking.*, IEEE Int. Conf. on Computer Vision, vol. 2, October 2005, pp. 1508–1511.
- [26] E. Royer, M. Lhuillier, Dhome M., and T. Chateau, *Localization in urban environments : Monocular vision compared to a differential gps sensor*, IEEE Conf. on Computer Vision and Pattern Recognition, CVPR (Washington, DC, USA), vol. 2, 2005, pp. 114–121.
- [27] CE. Shannon, *A mathematical theory of communication*, Bell system technical journal **27** (1948).
- [28] G. Silveira, E. Malis, and P. Rives, *Monocular-vision based SLAM using line segments*, IEEE Trans. on Robotics **24** (2008), no. 5, 969–979.
- [29] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, *Bundle adjustment : A modern synthesis*, Springer, 2000.
- [30] P. Viola and W. Wells, *Alignment by maximization of mutual information.*, Int. Journal of Computer Vision **24** (1997), no. 2, 137–154.