SLAM visuel actif pour la cartographie en grande gamme dynamique

Christian BARAT, Andrew COMPORT

Laboratoire I3S ,UNS-CNRS,
2000, route des lucioles, 06903 Sophia-Antipolis, FRANCE
barat@i3s.unice.fr,Andrew.comport@cnrs.fr

Résumé – L'imagerie à grande gamme dynamique (HDR) à partir de plusieurs images avec des temps d'exposition différents existe depuis plusieurs années dans le commerce, mais seulement pour une camera fixe. Dans le cadre de la robotique mobile cela se complique car le capteur se déplace. L'approche visuelle dense de localisation et cartographie (SLAM) permet d'estimer la pose entre chaque acquisition et de recaler les images et ainsi de construire un environnement 3D. Mais, peu de travaux ont été effectués pour faire du SLAM visuel en grande gamme dynamique. Ce travail propose une solution pour construire des cartes 3D de l'environnement en HDR. L'approche est basée sur théorie de l'information avec pour objectif de contrôler le temps d'exposition du capteur en essayant de maximiser l'information de la carte de l'environnement 3D.

Abstract – Acquiring High Dynamic Range (HDR) photos from several images, with an active shutter providing different exposures (sensor integration periods), has been widely commercialised in photography for static camera positions. In the case of a mobile video sensor (as is the case in robotics), this problem is more difficult due to real-time motion of the sensor. Recent dense visual SLAM approaches provide solution to estimate the motion, however, few works have attempted to perform HDR visual SLAM. In this paper a new approach is proposed that enables 3D HDR environment maps to be acquired actively from a dynamic set of images in real-time. The 6 dof pose, the dense scene structure and the HDR texture map will be estimated simultaneously with the objective of maximizing the dynamic range. In particular, a method is proposed to actively control the shutter based on information theory to optimise the information content of the 3D HDR environment map for RGB-D sensors.

1 Introduction

Dans le domaine de la robotique mobile il est indispensable de localiser le robot avec précision et robustesse. La plupart des approches font l'hypothèse de conservation de l'intensité des pixels durant les différentes acquisitions. La solution employée généralement est de fixer le temps d'exposition du capteur. Cela, devient problématique pour l'acquisition d'un environnement comportant une grande dynamique, avec des zones sombre et des zones très lumineuses. Le capteur peut saturer où donner des images sombres. La solution proposée est de contrôler le temps d'exposition du capteur de façon à acquérir toute la dynamique de la scène. Un premier travail pour améliorer la dynamique proposé par [1] est de combiner différentes images acquises avec des temps d'expositions différents. Les temps d'expositions n'étant pas contrôlés et inconnus, ils sont estimés en même temps que la cartographie 3D est effectuée. Une amélioration de ce travail est proposée par [2] qui a effectué le suivi dans le domaine des radiances normalisées, mais toujours avec un temps d'exposition géré par défaut par le capteur. Pour améliorer la carte 3D HDR, ce travail propose de contrôler de manière active le temps d'exposition du capteur. Pour cela, il est nécessaire d'obtenir deux modèles. Le premier permet d'estimer l'éclairement (irradiance) à partir de l'intensité des pixels (modèle inverse) et le second estime la transformation des intensités des pixels en fonction de la valeur du temps d'exposition (modèle direct). Dans la littérature, le contrôle du temps d'exposition se base généralement seulement sur l'intensité des pixels [3] et non pas sur la radiance. D'autres proposent une méthode de combinaison d'images en basse dynamique pour obtenir des images à haute gamme dynamique en utilisant une sélection prédéfinies de temps d'exposition [4], ce qui limite le choix. Dans le travail présenté, il est proposé d'estimer les temps d'expositions successifs optimaux de façon à maximiser la quantité d'information. Plus précisément l'objectif est de maximiser l'entropie combinée des images d'irradiance. Les contributions principales de ce travail sont les suivantes :

- Contrôle actif du temps d'exposition en temps réel.
- Construction d'une cartographie 3D en grande gamme dynamique optimale.
- Robustesse du SLAM aux variations de luminosité.

Le paragraphe 2 présente la méthode classique de suivi dense visuel avec un capteur RGB-D (couleur et profondeur) pour des images en basse gamme dynamique (LDR). Le paragraphe 3 montre l'adaptation de celle-ci aux images HDR. Le paragraphe 4 décrit l'approche pour déterminer le temps d'exposition optimal. Et, finalement les résultats sont présentés pour illustrer et valider la méthode dans le paragraphe 5.

2 Suivi Visuel Dense

Considérons un capteur RGB-D calibré (voir paragraphe 5) avec une fonction de luminance $\mathbf{I} : \Omega \times \mathbb{R}^+ \rightarrow \mathbb{L}; (\mathbf{p}, t) \mapsto \mathbf{I}(\mathbf{p}, t)$ et une fonction de profondeur $\mathbf{D} : \Omega \times \mathbb{R}^+ \rightarrow \mathbb{R}^+; (\mathbf{p}, t) \mapsto \mathbf{D}(\mathbf{p}, t)$ où $\Omega = [1, n] \times [1, m] \subset \mathbb{R}^2$,$\mathbf{P} = (p_1, p_2, .., p_{nm})^T \in \mathbb{R}^{mn \times 2} \subset \Omega$ sont les positions des pixels dans l'image acquise au temps t, et $n \times m$ est la dimension des images et $\mathbb{L} = [0, 1]$ est la gamme de luminance normalisée. Considérons l'ensemble des mesures sous la forme d'un vecteur tel que $\mathbf{I}(\mathbf{P}, t) \in \mathbb{L}^{nm \times 1}$ et $\mathbf{D}(\mathbf{P}, t) \in \mathbb{R}^{nm \times 1}$. Dans la suite t et $\mathbf{P}$ sont omis par soucis de clarté. $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_{nm})^T \in \mathbb{R}^{nm \times 3}$ est définie comme la matrice de sommets correspondant la rétro-projection des points de l'image de profondeur :

$$\mathbf{v}_i = \mathbf{K}^{-1} \overline{\mathbf{p}}_i \mathbf{D}(\mathbf{p}_i) \tag{1}$$

où $\mathbf{K} \in \mathbb{R}^{3 \times 3}$ est la matrice intrinsèque de la caméra. $\overline{\mathbf{p}}_i$ sont les coordonnées homogènes.

$\mathbf{V}$ est aussi définie comme une fonction de sommets 3D tel que $\mathbf{V} : \Omega \times \mathbb{R}^+ \rightarrow \mathbb{R}^3; ((\mathbf{p}, t) \mapsto \mathbf{V}(\mathbf{p}, t)$. L'ensemble $\boldsymbol{\mathcal{I}} = \{\mathbf{I}, \mathbf{V}\}$ est défini comme une image augmentée contenant les intensités et les sommets pour chaque pixel.

2.1 Modèle 3D basé Image

La représentation 3D est basée sur un graphe de N images augmentées $\mathcal{J} = \{\boldsymbol{\mathcal{I}}_1, .., \boldsymbol{\mathcal{I}}_N\}$ où chaque arête du graphe est un vecteur à 6 degrés de liberté (*twist*) $\mathbf{x} = (\boldsymbol{\nu}, \boldsymbol{\omega}) \in \mathbb{R}^6$ qui connecte deux images dans le graphe. Ce *twist* est lié à une pose 3D, $T = (\mathbf{R}, \mathbf{t}) \in \mathbb{SE}(3)$ à l'aide de l'application exponentielle $\mathbf{T} = e^{[x]\wedge}$ avec l'opérateur $[.]_\wedge$ définit tel que :

$$[x]_\wedge = \begin{bmatrix} [\boldsymbol{\omega}]_\times & \boldsymbol{\nu} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{2}$$

où $[.]_\times$ est l'opérateur matriciel symétrique *skew* , $\mathbf{R} \in \mathbb{SO}(3)$ est une matrice de rotation et $\mathbf{t} \in \mathbb{R}(3)$ un vecteur de translation.

Le modèle est construit de manière incrémentale en utilisant une approche SLAM [5]. Et, il est utilisé pour prédire une image virtuelle augmentée à partir de k images de référence extraites du modèle 3D. Cette image $\boldsymbol{\mathcal{I}}^* = \{\mathbf{I}^*, \mathbf{V}^*\}$ peut être utilisée pour réaliser l'estimation de la pose avec l'image courante $\boldsymbol{\mathcal{I}}$. Le caractère $*$ est utilisé par la suite pour désigner les variables relatives à l'image de référence prédite.

2.2 Cartographie en basse gamme dynamique

Dans le cas classique de suivi basé image, le temps d'exposition du capteur reste fixe. L'objectif est ici d'estimer la transformation 3D, $\widetilde{\mathbf{T}}$ entre l'image courante $\mathbf{I}$ et une image $\boldsymbol{\mathcal{I}}^*$ augmentée de référence prédite à partir du modèle 3D. En fait, on estime le vecteur (*twist*) de transformation $\mathbf{x}$ tel que :

$$\widetilde{\mathbf{T}} = \widehat{\mathbf{T}} \mathbf{T}(x) \tag{3}$$

où $\widehat{\mathbf{T}}$ est une estimée de la pose initiale $\widetilde{\mathbf{T}}$. Avec pour hypothèse que la luminosité est conservée (Lambertien) et que le temps d'exposition est constant pour $\mathbf{I}$ et $\mathbf{I}^*$, la variable (6 ddl) $\mathbf{x}$ peut être estimée en minimisant l'erreur d'intensité :

$$e(x)_{LDR} = \left[\boldsymbol{I} \left(w \left(\widehat{\boldsymbol{T}} \boldsymbol{T}(x), \boldsymbol{V}^* \right) \right) - \boldsymbol{I}^*(\boldsymbol{P})^* \right] \tag{4}$$

où la fonction de projection $w\left(\widehat{\boldsymbol{T}}\boldsymbol{T}(x), \boldsymbol{V}^*\right)$ projette les sommets $\boldsymbol{V}^*$ associés aux pixels rétro-projetés $\boldsymbol{P}^*$ de l'équation 1, avec la transformation $\widehat{\boldsymbol{T}}\boldsymbol{T}(x)$ sur le plan de l'image normalisée :

$$\overline{\boldsymbol{p}}^w = \boldsymbol{K} \boldsymbol{\Pi} \widetilde{\boldsymbol{T}} \overline{\mathbf{v}}^* \tag{5}$$

La matrice $\boldsymbol{\Pi} = [\boldsymbol{I}_{3\times3}, \mathbf{0}] \in \mathbb{R}^{3\times4}$ est une matrice de projection de la dimension 4 vers la dimension 3. Un trait suscrit ($\overline{\mathbf{v}}$) sera utilisé pour indiquer des coordonnées homogènes normalisées sur le dernier composant.

3 Cartographie en grande gamme dynamique

Ce paragraphe présente maintenant le suivi en HDR en remplaçant les images de référence du modèle 3D en LDR par des images en haute gamme dynamique (HDR), c'est à dire dans le domaine de l'irradiance. Ces images ne seront plus saturées $\boldsymbol{I}^*_{HDR} \in [0, \inf]$. L'exposition $\mathbf{X}$ est définie comme le produit de l'irradiance $\mathbf{E}$ par le temps d'exposition Δt, voir Eq.6. L'intensité du pixel $\mathbf{I}$ dépend de la fonction de réponse de la caméra f. L'irradiance $\mathbf{L}$ est proportionnelle à la radiance $\mathbf{E}$ pour chaque pixel, mais le facteur de proportionnalité peut être différent en fonction de la position sur le capteur . Par la suite la radiance $\mathbf{E}$ sera utilisée plutôt que l'irradiance $\mathbf{L}$ sachant que l'on peut toujours obtenir $\mathbf{L}$ après une étape de calibration.

$$\boldsymbol{I}_{ij} = f\left(\boldsymbol{E}_i \Delta t_j\right) \tag{6}$$

Si f est monotone, elle est donc inversible, (6) devient :

$$f^{-1}\left(\boldsymbol{I}_{ij}\right) = \boldsymbol{E}_i \Delta t_j \tag{7}$$

En utilisant le logarithme :

$$ln f^{-1}\left(\boldsymbol{I}_{ij}\right) = ln \boldsymbol{E}_i + ln \Delta t_j \tag{8}$$

Finalement avec $g = ln f^{-1}$:

$$ln \mathbf{E}_i = g\left(\boldsymbol{I}_{ij}\right) - ln \Delta t_j \tag{9}$$

où i est l'indice des pixels et j l'indice des temps d'exposition.

On peut obtenir le modèle de g (modèle inverse) en utilisant la méthode présentée dans [6]. Pour cela, il faut acquérir plusieurs images à la même position pour différent temps d'exposition. Par la suite nous utiliserons le modèle inverse moyenné suivant $g = \left(g_{rouge} + g_{vert} + g_{bleu}\right)/3$ plutôt que 3 modèles en fonction des 3 canaux.

4 Optimisation de la dynamique de la radiance

Maintenant pendant le SLAM, le modèle 3D (graphe d'images de référence) est mis à jour avec les nouvelles images dans le

domaine des log radiance : $ln(E)$. Pour mettre à jour le modèle 3D, on mélange la nouvelle image avec les images de référence. Chaque Image de référence $\boldsymbol{I}^*_{HDR}$ et ses poids cumulatifs $\boldsymbol{C}^*_{HDR}$ (Eq.10) sont mis à jour de manière incrémentale entre le temps $t-1$ et le temps t par :

$$\boldsymbol{C}^*_{HDR}(\boldsymbol{p},t) \leftarrow \boldsymbol{C}^*_{HDR}(\boldsymbol{p},t-1) + h(\boldsymbol{I}^w(\boldsymbol{p},t)) \quad (10)$$

$$\boldsymbol{I}^*_{HDR}(\boldsymbol{p},t) \leftarrow \frac{h(\boldsymbol{I}^w(\boldsymbol{p},t))\boldsymbol{I}^w_{HDR}(\boldsymbol{p},t) + \boldsymbol{C}^*_{HDR}(\boldsymbol{p},t-1)\boldsymbol{I}^*_{HDR}(\boldsymbol{p},t-1)}{\boldsymbol{C}^*_{HDR}(\boldsymbol{p},t)} \quad (11)$$

où $\boldsymbol{I}^w$ et $\boldsymbol{I}^w_{HDR}$ sont respectivement l'image courante LDR et HDR projetées sur l'image de référence. La fonction de pondération utilisée est proportionnelle à la sensibilité du capteur :

$$h(\boldsymbol{I}_{ij}) = \frac{d\boldsymbol{I}_{ij}}{dg(\boldsymbol{I}_{ij})} \quad (12)$$

où g est définit dans l' Eq.(9). L'objectif principal est d'optimiser la dynamique du log-Radiance pendant l'acquisition en temps réel. Cela est équivalent à maximiser l'entropie. L'idée est de sélectionner le prochain meilleur temps d'exposition qui maximisera l'entropie de l'image de log-Radiance mise à jour (mélangée). S'il est possible de prédire la réponse du capteur pour différentes valeurs du temps d'exposition à partir d'une image, il est par conséquent possible de prédire quel temps d'exposition optimisera l'entropie. Pour cela, il faut identifier un modèle direct qui permettra d'estimer des images pour différents temps d'exposition à partir de l'image courante.

4.1 Modèle direct

Dans ce paragraphe il est présenté une méthode pour estimer la transformation d'une image d'un temps d'exposition à un autre. La transformation correspond à une amplification de l'intensité du pixel $\boldsymbol{I}_{ij}$ due au changement de temps d'exposition. Cette amplification est estimée à partir d'une série d'images acquises avec des temps d'exposition variant de 1ms à 30ms avec un pas de 1ms et cela pour chaque niveau d'intensité de pixel (0 à 255). Par conséquent la relation linéaire entre une image obtenue avec un temps d'exposition s_1 par rapport à une image obtenue avec un temps d'exposition s_2 est une matrice A de dimension 256×30 (niveaux d'intensité$\times$nombre de temps d'expositions) :

$$\boldsymbol{I}^{s_2}_{ij} = A(\boldsymbol{I}^{s_1}_{ij}, s_1)\boldsymbol{I}^{s_1}_{ij} \quad (13)$$

avec $s_1 = n$, $s_2 = n+1$ et $n \in [1, 30]$. Pour estimer une image avec une variation de temps d'exposition égale à k ms il faut effectuer la transformation de proche en proche k fois. Pour diminuer le temps de calcul, plutôt que de calculer la nouvelle image, les calculs vont être effectués directement sur les histogrammes, ce qui sera suffisant pour l'estimation d'entropie. Un estimateur de densité à base de noyaux gaussien est utilisé pour avoir une estimation lissée :

$$\widehat{f_h}(\boldsymbol{I}) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{\boldsymbol{I}-\boldsymbol{I}_i}{h}\right) \quad (14)$$

où N est le nombre de noyaux (256 niveaux d'intensité), h est un paramètre de lissage , K est un noyau gaussien :

$$K(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad (15)$$

4.2 Prédiction d'histogramme

Le modèle direct est appliqué sur l'histogramme de l'image courante pour estimer l'histogramme de l'image prédite pour un temps d'exposition différent. Puis, on peut calculer l'entropie directement sur la somme pondéré de l'histogramme courant et de l'histogramme prédit. Le nouvel histogramme pour un temps d'exposition $s \in [1..30]$ est calculé en utilisant l'histogramme du précédent temps d'exposition $s+1$ (16) $s-1$ (17) en utilisant la matrice d'amplification $A(\mathbf{I}, s)$ avec $\mathbf{I}$ l'intensité du pixel.

$$\widehat{f_s}(\mathbf{I}) = \sum_{\mathbf{I}=1}^{255} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(A(\mathbf{I},s)\widehat{f}_{s+1}(\mathbf{I})-\mathbf{I}\right)^2} \quad (16)$$

$$\widehat{f_s}(\mathbf{I}) = \sum_{\mathbf{I}=1}^{255} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\widehat{f}_{s-1}(\mathbf{I})}{A(\mathbf{I},s)}-\mathbf{I}\right)^2} \quad (17)$$

Avec $\widehat{f}$ calculé par Eq.(14) pour un temps d'exposition $s \in [2..29]$ms. Une fois que nous avons l'histogramme d'intensité des pixels on peut obtenir facilement l'histogramme des log-Radiance. Il est plus intéressant de travailler dans le domaine des log-Radiance que dans le domaine des radiance car il est moins épart.

5 Résultats

Un capteur Asus Xtion (RGB-D) est utilisé dans les expérimentations. Il a une caméra 640×480 RGB et un capteur 640×480 de profondeur. On peut contrôler le temps d'exposition de 1ms à 3000ms. On se limitera à la zone de 1ms à 30ms pour rester temps réel à 30hz. Pour valider notre approche un première expérimentation est présentée avec un caméra statique puis ensuite dans une application mobile de SLAM.

5.1 Capteur Statique avec 3 images

Des résultats sont présentés pour l'acquisition de 3 images successives avec des Δt optimisés. La première image est acquise avec un Δt de 15ms. Un exemple est présenté Fig.1. Pour cet exemple on une augmentation de l'entropie de 6.41 à 6.95. On peut voir l'amélioration de l'image au niveau du néon qui est saturé dans la première image et en ayant choisi un temps d'exposition court (1ms) on ne sature presque plus et on obtient une image mixée de meilleure qualité. L'apport de la troisième image est plus faible, car avec les 2 premières on a déjà capturé une grande partie de la dynamique.

5.2 Capteur en mouvement :HDR SLAM

Cette fois ci la camera se déplace pendant l'acquisition. On commence de la même manière avec un Δt=15ms, puis on optimise les Δt suivant. Sur la figure 2 on peut remarquer l'amélioration de la partie sombre sur la gauche grâce à la combinaison des images avec des temps d'exposition Δt successifs de

FIGURE 1 – A gauche les 3 images LDR sources acquises et à droite les images résultantes HDR. L'entropie pour la première image dans le domaine des log-Radiance est de 6.41 puis après mélange avec la deuxième elle est de 6.93 pour finir à 6.95 avec l'ajout de la troisième image. Les temps d'expositions successifs sont 15ms, 1ms, et 10ms.

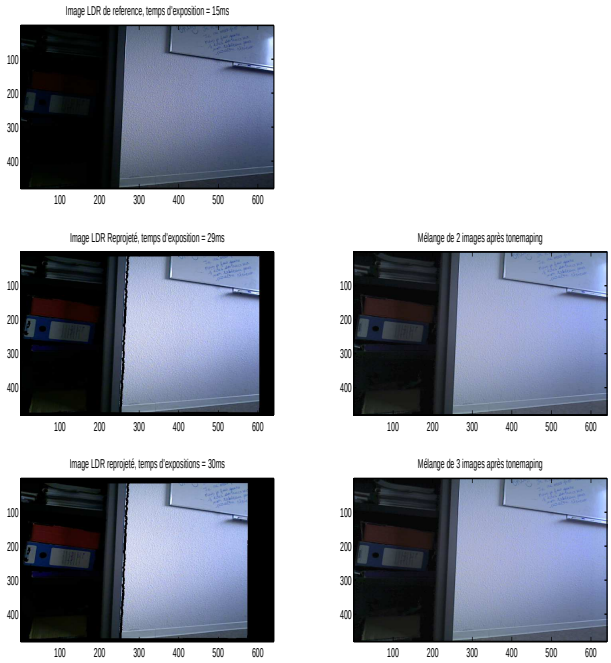

FIGURE 2 – A gauche les 3 images LDR sources acquises et à droite les images résultantes HDR.

15, 29 et 30 ms. L'entropie évolue de 5.69 à 6.47. Après test, la méthode classique de SLAM LDR (paragraphe 2.2) n'arrive pas à converger dans tous les cas en utilisant les images obtenues avec des temps d'expositions différents, car on n'a pas conservation de l'intensité des pixels et dans les cas où elle converge la vitesse se dégrade quand on s'éloigne du temps d'exposition de l'image de référence.

6 Conclusion

Dans ce travail Il a été présenté une méthode de contrôle du temps d'exposition d'une caméra en temps réel basée sur la théorie de l'information. Grâce à cette approche, il est possible de cartographier l'environnement en grande gamme dynamique et cela de manière optimisée. Des résultats illustrent l'amélioration apportée par cette nouvelle approche.

Références

[1] M. Meilland, C. Barat, and A. Comport, "3d high dynamic range dense visual slam and its application to real-time object re-lighting," in *Proceedings of International Symposium on Mixed and Augmented Reality, ISMAR*, Adelaide, Australia, 2013.

[2] S. Li, A. Handa, Y. Zhang, and A. Calway, "Hdrfusion : HDR SLAM using a low-cost auto-exposure RGB-D sensor," *CoRR*, vol. abs/1604.00895, 2016.

[3] Q. K. Vuong, S. hwan Yun, and S. Kim, "A new auto exposure and auto white-balance algorithm to detect high dynamic range conditions using cmos technology," in *Proceedings of the World Congress on Engineering and Computer Science*, San Francisco,USA, 2008.

[4] M. A. Ali and S. Mann, "Comparametric image compositing : Computationally efficient high dynamic range imaging," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 913–916.

[5] M. Meilland and A. I. Comport, "On unifying image and model-based real-time dense mapping at large scales," in *Int. Conf. on Intelli- gent Robots and Systems*. IEEE, 2013.

[6] M. Grossberg and S. Nayar, "Determining the Camera Response from Images : What is Knowable ?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1455–1467, Nov 2003.