Adéquation algorithme architecture pour la localisation basée image sur système embarqué

David Vandergucht[2] Mehdi Darouich[1] Valérie Gouet-Brunet[2] Thibaut Vercueil[1]
Laurent Schneider[2] Thomas Peyret[1] Benoit Tain[1] Maria Lepecq[1]

[1] CEA, LIST, 91191 Gif-Sur-Yvette, France
[2] Univ. Paris-Est, LASTIG MATIS, IGN, ENSG, F-94160 Saint-Mande, France

1 Résumé

Dans cet article, nous présentons l'adaptation sur système embarqué d'une chaîne de traitements dédiée à la localisation basée image pour l'aide à la navigation pédestre par réalité augmentée. L'étape d'estimation de la pose du capteur d'acquisition de l'image est une étape clé et sensible, en termes de qualité de la pose attendue comme en complexité et temps d'estimation de cette pose. Nous décrivons l'approche d'estimation de pose et la solution hardware choisit, ainsi que leurs adaptations en vue d'une adéquation optimale entre algorithme et architecture. Des expériences en conditions réalistes de navigation pédestre, où les mouvements sont complexes, montrent la robustesse et l'efficacité de la solution proposée, qui est prometteuse pour une future application sur système léger impliquant l'estimation précise de la pose.

2 Localisation visuelle temps réel sur système embarqué

2.1 La localisation, différentes possibilités

Où suis-je ? Comment caractériser mon emplacement, ma posture ? Ces questions peuvent avoir plusieurs réponses en fonction du contexte applicatif. Aujourd'hui, pour un piéton, l'utilisation du GPS du téléphone semble évidente pour répondre globalement à cette question. Cependant, en intérieur ou dans le canyon urbain, le GPS se dégrade vite. Il est suffisamment précis pour une localisation à l'échelle de la rue, mais pas assez pour permettre des applications de réalité augmentée à l'échelle du piéton, qui requièrent une haute précision en termes de position comme d'orientation (pose). Face à ces limitations, d'autres solutions existent, parmi lesquelles l'analyse d'image, démocratisée par l'omniprésence des capteurs photo et par les capacités de traitement informatique actuelles. De nombreuses techniques d'analyse exploitent l'image comme un GPS visuel, permettant la localisation à des niveaux de précision et d'échelle d'exploration divers [4]. Les techniques offrant le plus de précision sont le dernier maillon de la chaîne de localisation des systèmes d'aide à la navigation (véhicule comme piéton), mais elles supposent généralement la connaissance d'une pose initiale, qui peut être apportée par les approches moins précises opérant à des échelles d'exploration plus larges.

2.2 De la localisation du véhicule à la localisation du piéton sur système embarqué

Le problème de la localisation des véhicules est très étudié en raison des enjeux liés à la conduite de véhicule autonome. Qu et al [5] ont notamment développé une méthode permettant d'estimer la pose d'un véhicule en utilisant l'ajustement de faisceaux local, la propagation de l'incertitude liée à l'estimation de poses successives et son exploitation dans l'optimisation de la solution et la détection contrainte d'amers visuels géolocalisés (des panneaux routiers), afin de réduire la dérive. Cette technique offre de bonnes performances, mais elle a été développée autour de jeux de données routiers (Kitti [3], Stéréopolis [2]), qui présentent des mouvements linéaires vers l'avant ou en virage, sans variations brusques. Cet aspect n'impose pas une collecte très fréquente d'images (1-5Hz), ce qui laisse du temps pour les traiter, d'autant plus que la technique a été conçue pour du traitement hors ligne, privilégiant ainsi la robustesse, au détriment de sa compatibilité avec le temps réel. Le travail présenté ici a consisté à adapter cette technique en vue de l'aide à la navigation piétonne par réalité augmentée, et en se focalisant sur la localisation fine du piéton dans l'environnement, plus précisément la capture des mouvements - peu contraints, rapides et complexes - associés aux déplacements et rotations de sa tête, dans un contexte temps réel.

La réalité augmentée implique un couplage très serré entre la récupération des données caractérisant la scène (pose de l'utilisateur, amers visuels) et le retour vers l'utilisateur (incrustation des données dans le retour visuel affiché). La latence maximale entre le mouvement réel de la scène et son impact sur ce qui est affiché doit être en dessous de la latence du système vestibulo-oculaire, de l'ordre de 7 a 15 ms [8], afin de prévenir les problèmes de cybermalaise (maux de tête, nausée, etc). De plus la cadence minimale du flux vidéo "augmenté" doit être au-delà du seuil perceptible pour l'être humain, soit 20 images par seconde (ips). Cette faible latence implique une co-localité des données et des traitements très forte, afin d'éviter tout délai dû aux communications, disqualifiant l'utilisation d'un PC standard. L'utilisation d'un matériel porté, proche du capteur et de l'afficheur, est la solution retenue dans la plupart des systèmes. Pour un système de lunettes intelligentes par exemple, les unités de calcul impliquées doivent être embarquées et répondre à des contraintes fortes d'encombrement et de poids réduits pour le confort d'utilisation et de consommation électrique faible pour garantir l'autonomie du système (typiquement 3cm x 5cm, 200g batterie incluse, 10W).

3 Notre approche

3.1 Architecture préexistante

Le système de localisation étudié repose sur la solution proposée dans [5]. Il est découpé en 3 blocs fonctionnels (voir Figure 1) : 1) Détection et appariement de points d'intérêt de type SIFT ou SURF, 2) estimation de la pose, création d'une carte 3D et optimisation de cette carte à partir des points d'intérêt, et détection des panneaux routiers comme amers visuels géolocalisés permettant de raffiner la trajectoire.

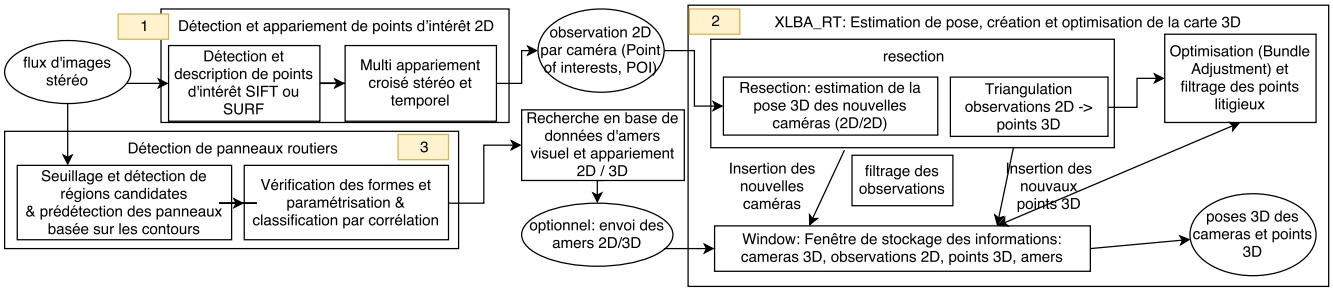

FIGURE 1 – Vue générale de l'architecture de traitement préexistante [5].

3.2 Opportunités d'accélération

Afin de rendre le système plus efficace et en mesure de traiter un flux d'images à plus haute fréquence, nous avons proposé plusieurs aménagements :

1. Partitionner et distribuer l'application entre une partie à très basse latence, exécutée sur un Système sur Puce (System On Chip - SoC) proche des capteurs d'image, et une partie plus conventionnelle apte aux traitements plus complexes. L'objectif est de réduire au plus tôt la bande passante nécessaire en localisant certaines opérations de traitement de l'information au plus près des pixels ;
2. Développer un système de détection et d'appariement de points 2D accéléré matériellement, très basse latence et déterministe, offrant une alternative rapide à l'appariement croisé robuste utilisé dans [5] ;
3. Restructurer l'application pour qu'elle fournisse une pose 3D à basse latence, pendant qu'un autre processus en arrière-plan s'occupe de la construction et de l'optimisation de la carte 3D, dans un contexte "en ligne".

Les améliorations 1 et 2 tirent parti d'un SoC intégrant une brique d'accélération matérielle (IP) permettant de détecter, décrire et suivre des points d'intérêt 2D dans un flux d'images stéréoscopiques ainsi que d'un processeur multi-coeur embarqué. Un redesign en profondeur de l'application ainsi que l'utilisation de la plateforme ROS [6] ont permis d'adresser le point 3.

3.3 Description de la brique d'accélération matérielle de suivi de points d'intérêt

L'IP matérielle de suivi de points a été conçue pour détecter, décrire et suivre des points d'intérêt dans un flux d'images stéréoscopiques de manière performante et efficace. La détection des points s'effectue avec l'algorithme de Harris ; ils sont triés par sous-zone d'image selon une grille homogène. Le descripteur des points est similaire à SURF (analyse pondérée des gradients autour du point), et leur appariement minimise la distance L1 entre descripteurs. Cet accélérateur matériel, par conception, est capable de détecter, décrire et suivre un point 2D avec une latence intra-image (une centaine de lignes d'image) tout en gardant un comportement déterministe et une répartition homogène des points, quelle que soit la scène. Des travaux similaires existent [7], mais la garantie de latence passe par le bypass des points en cas de flux trop important, ce qui ne garantit pas la cohérence des points dans les zones très fortement peuplées.
Plusieurs adaptations par rapport à l'approche initiale ont dû être effectuées pour rester compatible avec l'IP matérielle. Le multi appariement croisé a du être simplifié pour limiter le nombre de plans mémoire impliqués (mémoire = surface = consommation électrique). Aussi, la résolution des images et le nombre de points d'intérêt ont été adaptés : passage d'images HD avec 2048 points à des images QHD (960x540) avec 1024 points, permettant de limiter la complexité de la brique de calcul matérielle. L'IP de suivi de points a été décrite en langage de description matériel (VHDL) et implémentée sur un FPGA (circuits intégrés en silicium reprogrammables), permettant la mise en place d'un démonstrateur sur carte proFPGA munie d'un Virtex 7. Sur ce modèle, l'IP de suivi occupe 10% des ressources de calcul et fonctionne à 100 MHz.
D'autre part, l'intérêt pour une solution basse consommation et basse latence nous a poussés à effectuer une caractérisation en technologie circuit intégré pour déterminer l'intérêt d'une telle solution sur un composant dédié offrant plus de performance et d'efficacité énergétique. Nous avons eu accès à la technologie FDSOI 22nm,offrant un très bon compromis performance/consommation. Les synthèses montrent que cette IP peut fonctionner à une fréquence de 1,1 GHz, occupant une surface de $2.8\,\mathrm{mm}^2$ et consommant 377mW.

4 Résultats

Cette nouvelle chaîne de traitement hybride a été testée avec succès sur plusieurs jeux de données comme Kitti [3] ou surtout EuRocMav [1] qui dispose de mouvements plus complexes que Kitti. Un jeu de donnée stéréo piéton a également été réalisé et testé. Les évaluations en termes de précision sont toujours en cours, mais le nuage de point 3D reconstruit est cohérent (Figure 2), les trajectoires sont conformes aux vérités terrain (Figure 3), avec une moyenne d'écart de distance de 0.042m et un écart-type de 0.026m, sans fermeture de boucle ni relocalisation, tout en assurant désormais une estimation de pose à 20Hz. L'implémentation de l'IP de suivi de points sur FPGA traite une image QHD en 21 ms (soit 47 images par seconde), ce qui est compatible avec la performance de la chaîne globale. En implémentation ASIC, intégré dans une puce dédiée, cette IP pourrait calculer des images QHD en moins de 2ms. La première phase de la détection de panneaux routiers (Figure 1), compatible avec une exécution sur processeur embarqué du fait de sa régularité, a aussi été portée sur le processeur embarqué. Les tests effectués montrent que le traitement d'une image QHD nécessite moins de 50 Mcycles, soit 16 ips.

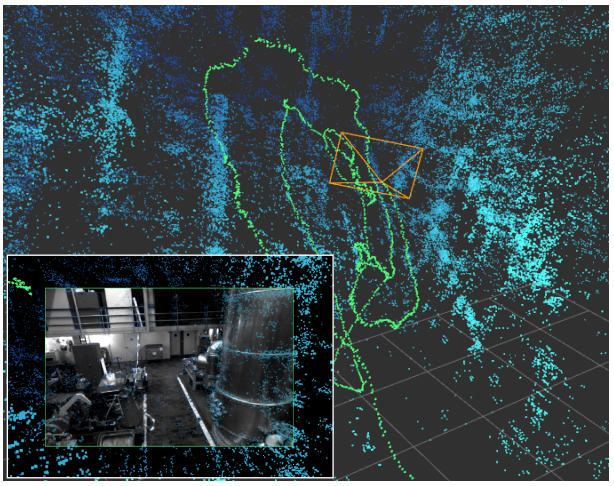

FIGURE 2 – Reconstruction 3D de la scène MH_01_easy de EuRocMav[1] et augmentation 3D de la prise de vue.

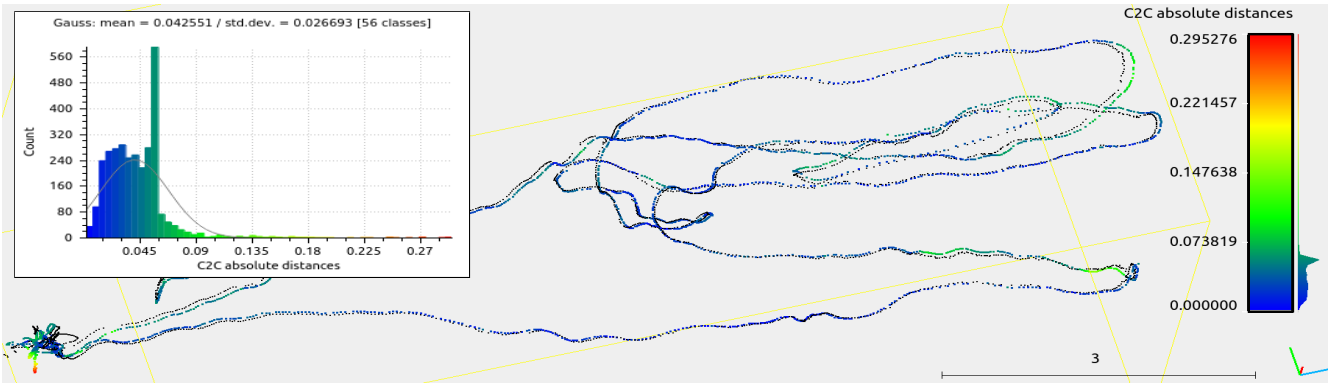

FIGURE 3 – Comparaison de la trajectoire estimée (en noir) avec la vérité terrain (en couleur) sur la séquence MH_01_easy de EuRocMav[1]. Le parcours total fait 80m. L'échelle de couleurs indique la distance entre la vérité terrain et l'estimation, et l'histogramme à gauche la distribution de ces distances.

5 Perspectives et remerciements

Les perspectives de ce travail sont nombreuses, elles visent en premier lieu à poursuivre l'accélération des briques logicielles sur système embarqué, mais également à le robustifier : exploitation d'autres sources (suivi multi-capteur couplé image et centrale inertielle), détection d'amers embarquée par réseaux de neurones profonds, relocalisation dans l'environnement, etc. Ces travaux ont été réalisés dans le cadre du projet européen KET ENIAC Things2do qui vise à faire émerger un écosystème technologique européen autour de la technologie FDSOI.

Références

[1] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 2016.

[2] Nicolas Paparoditis et Jean-Pierre Papelard et Bertrand Cannelle et Alexandre Devaux et Bahman Soheilian et Nicolas David et Erwan Houzay. Stereopolis ii : A multi-purpose and multi-sensor 3d mobile mapping system for street visualisation and 3d metrology. *Revue Française de Photogrammétrie et de Télédétection*, (200) :69–79, 2014.

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets Robotics : The KITTI Dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[4] Nathan Piasco, Désiré Sidibé, Cédric Demonceaux, and Valérie Gouet-Brunet. A survey on visual-based localization : On the benefit of heterogeneous data. *Pattern Recognition*, 74 :90–109, 2018.

[5] Xiaozhi Qu, Bahman Soheilian, and Nicolas Paparoditis. Landmark based localization in urban environment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017.

[6] Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. ROS : an open-source Robot Operating System. In *ICRA Workshop on Open Source Software*, 2009.

[7] J. Wang, S. Zhong, L. Yan, and Z. Cao. An embedded system-on-chip architecture for real-time visual detection and matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(3) :525–538, March 2014.

[8] Feng Zheng, Turner Whitted, Anselmo Lastra, Peter Lincoln, Andrei State, Andrew Maimone, and Henry Fuchs. Minimizing latency for augmented reality displays : Frames considered harmful. In *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*, pages 195–200. IEEE, 2014.