



DOCTORAT DE L'UNIVERSITÉ DE LORRAINE
RAPPORT DE SOUTENANCE

Nom et prénom du doctorant : ZINS MATTHIEU

Titre de la thèse : Contributions à la précision et à la robustesse de la localisation visuelle dans un monde d'objets

Doctorat : Informatique

Date de la soutenance : 9 décembre 2022

Mr Matthieu Zins a fait une présentation claire, synthétique et élégante de ses travaux de thèse. Il a su replacer ses résultats dans la dynamique actuelle du sujet, en y montrant l'importance de ses contributions personnelles. Le jury tient à souligner le caractère extrêmement pédagogique tant du manuscrit que de l'exposé oral.

La contribution principale de Mr. Zins porte sur l'amélioration de la précision et robustesse de la localisation visuelle de caméras. Il a montré comment améliorer la localisation visuelle grâce à une modélisation légère d'objets 3D par des ellipsoïdes, combinés avec les progrès récents de l'apprentissage profond pour la détection d'objets dans des images. Il a présenté une nouvelle méthode de localisation qui fonctionne en temps réel, ce qui est vraiment remarquable. Il a également su évaluer et comparer ses contributions à l'état de l'art avec beaucoup de rigueur.

Le candidat a fait preuve d'une grande qualité d'ouverture ainsi que de sérieux dans la poursuite de ses travaux. Le jury a beaucoup apprécié l'originalité de l'approche et des solutions proposées. Ces travaux ouvrent également des pistes de recherche particulièrement prometteuses pour le futur.

D'autre part, le jury a apprécié la qualité de ses réponses aux nombreuses questions, démontrant sa grande maîtrise du sujet.

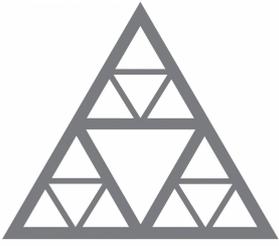
Le jury félicite donc Mr Matthieu Zins pour la très grande qualité de ses travaux et décide de lui décerner le titre de docteur de l'Université de Lorraine, spécialité informatique.

Président du jury : *Sylvain Lazard* (A COMPLÉTER)

Membres du jury :

Nom	Signature	Nom	Signature
Marie-Odile BERGER		Gilles SIMON	
Vincent LEPETIT		Eric MARCHAND	
Gabriela CSURKA	 P.O.	Sylvain LAZARD	

Si le rapport comporte plusieurs pages ou s'il est rédigé sur un document distinct, il devra être paraphé sur chaque page et signé par le Président du jury.



École des Ponts
ParisTech

Champs s/ Marne, le 29 octobre 2022

Rapport sur le mémoire de thèse de M. Mathieu ZINS

Résumé du mémoire

Vincent Lepetit

Chercheur
ENPC ParisTech
6-8, Avenue Blaise Pascal
Champs-sur-Marne
77455 Marne-la-Vallée
☎ 07.69.33.75.62

✉ vincent.lepetit@enpc.fr
🔗 [vincentlepetit.github.io/](https://github.com/vincentlepetit)

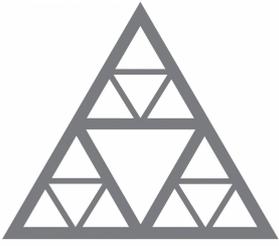
La thèse de M. Mathieu ZINS porte sur la localisation spatiale de caméra à partir d'images, un problème important de la vision par ordinateur. Le fil conducteur est la représentation des objets d'une scène par des ellipsoïdes et la détection d'objets dans des images sous forme d'ellipses. La thèse est motivée par les travaux précédents de l'équipe qui ont montré qu'il était possible de calculer la pose d'une caméra à partir d'une correspondance ellipsoïde-ellipse. La thèse de M. Mathieu ZINS bâtit sur ces travaux pour développer des méthodes originales et performantes pour l'estimation de poses de caméra.

Après une introduction qui introduit clairement la motivation pour la thèse et ses applications potentielles, le premier chapitre décrit l'état de l'art en positionnement visuel. Ce chapitre donne une très bonne vision d'ensemble de cet état de l'art qui est pourtant très vaste, car beaucoup de directions différentes ont déjà été considérées : points d'intérêt, régression directe avec un réseau profond, considération d'objets. Le chapitre se termine par la description de travaux sur les correspondances ellipsoïde-ellipse pour calcul une pose de caméra et en particulier les équations et les méthodes développées précédemment par l'équipe de recherche du candidat.

Le **chapitre 2** propose des architectures profondes pour la prédiction d'ellipses correspondant à des objets. Le chapitre montre d'abord pourquoi simplement utiliser l'ellipse inscrite dans une boîte de détection classique peut entraîner une grande erreur. Plusieurs paramétrisations d'une ellipse sont considérées. Ces paramétrisations peuvent avoir des discontinuités, ce qui complique l'entraînement de l'architecture, et une première fonction de coût qui combine régression et classification pour résoudre ce problème est introduite. Une première architecture utilisant cette fonction de coût appliquée aux détections d'objets de FasterRCNN est introduite pour prédire les ellipses correspondant à des objets à partir d'une image.

Une seconde option pour la fonction de coût est ensuite proposée : Elle est basée sur la distance entre les fonctions de plongement (au final, une distance entre "level sets") de l'ellipse prédite et de l'ellipse à prédire est proposée pour obtenir une fonction de coût qui ne dépend pas de plusieurs termes à combiner. Pour s'adapter à cette nouvelle fonction de coût, l'architecture a été légèrement modifiée.

Étant donnée une scène, les ellipsoïdes correspondant aux objets sont reconstruits à partir des boîtes 2D englobantes dans 3 vues ou plus. Ces ellipsoïdes sont ensuite reprojétés dans des images de la scène pour obtenir des images d'entraînement et entraîner l'architecture profonde à détecter les objets de la scène cible sous forme d'ellipses. Une première méthode simple pour le calcul de pose basée sur le résultat classique à partir



École des Ponts
ParisTech

Vincent Lepetit

Chercheur
ENPC ParisTech
6-8, Avenue Blaise Pascal
Champs-sur-Marne
77455 Marne-la-Vallée
☎ 07.69.33.75.62

✉ vincent.lepetit@enpc.fr
🔗 [vincentlepetit.github.io/](https://github.com/vincentlepetit)

de 3 correspondances et l'approximation que le centre d'un ellipsoïde se reprojete sur le centre de sa reprojction est introduite. Une étude de l'erreur pour évaluer l'impact de cette approximation est fournie.

L'approche proposée est ensuite évaluée et comparée avec plusieurs méthodes de l'état de l'art sur plusieurs jeux de données classiques ainsi que sur un jeu de données d'une scène synthétique créé par le candidat. Les évaluations sont détaillées et montrent l'avantage de l'approche proposée. Les résultats sur les séquences de WatchPose, où les autres approches ont sans doute beaucoup de difficulté, sont impressionnants. Finalement, une dernière architecture basée sur le détecteur d'objets YOLO est proposée pour permettre l'entraînement *end-to-end* de la prédiction d'ellipses.

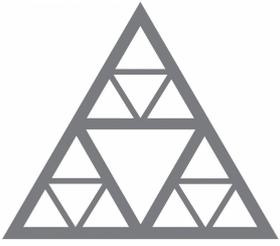
Le **chapitre 3** porte sur l'utilisation des ellipses détectées pour calculer la pose de la caméra, la méthode utilisée au chapitre 2 étant basée sur une approximation. L'idée est de minimiser une fonction objective qui est une somme robuste de distances entre la reprojction d'un ellipsoïde et son ellipse associée. L'initialisation est faite à l'aide de la méthode du chapitre 2. Plusieurs fonctions de distance entre ellipses : intersection-over-union, distance entre boîtes, distance algébrique, distance entre *level sets* comme pour l'entraînement de la détection des ellipses, distances Wasserstein et de Bhattacharyya entre les distributions normales définies par les ellipses sont proposées. L'impact du choix de cette distance sur la précision du calcul de pose est ensuite évalué. L'impact des occultations partielles entre objets est également évalué.

La seconde partie du chapitre porte sur l'estimation de l'incertitude dans la détection des ellipses et de la propagation de cette incertitude vers l'incertitude de la pose calculée. Les méthodes classiques de prédiction d'incertitude par un réseau profond sont présentées. Le choix fait dans la thèse est la prédiction de la variance de la distance entre l'ellipse à prédire et l'ellipse prédite. Ce choix me paraît être une bonne idée, parce que compact tout en fournissant suffisamment d'information : Prédire une matrice d'incertitude sur l'ensemble des paramètres croît de façon exponentielle, alors que prédire une seule valeur est suffisant pour pondérer les observations dans la fonction objective. Les incertitudes prédites sont évaluées d'abord visuellement pour vérifier qu'elles sont cohérentes avec la précision des ellipses prédites. Elles sont ensuite introduites dans la fonction objective. Il est alors vérifié qu'elles permettent effectivement une estimation plus précise de la pose, notamment en présence d'occultation.

Le **chapitre 4** porte sur le SLAM (*Simultaneous Localisation and Mapping*) en se basant sur des points d'intérêt, comme il est classiquement fait, mais aussi sur des objets. Une description des différentes approches de SLAM est d'abord présentée, depuis les premières méthodes de *bundle adjustment* jusqu'aux dernières méthodes basées sur l'apprentissage profond et les quelques méthodes utilisant des objets pour le suivi et la localisation de la caméra.

La méthode proposée intègre la détection d'objets et leur représentation sous forme d'ellipsoïdes et d'ellipses à ORB-SLAM2, une implémentation populaire de SLAM utilisant des points d'intérêt. Tout comme pour les points d'intérêt, la méthode proposée intègre les objets au fur et à mesure qu'ils apparaissent, affine leur estimation, et les utilise pour estimer la pose de la caméra et relocaliser la caméra quand le suivi est perdu.

Plus précisément, les objets sont d'abord détectés à l'aide de YOLO. Une nouvelle boîte



École des Ponts
ParisTech

Vincent Lepetit

Chercheur
ENPC ParisTech
6-8, Avenue Blaise Pascal
Champs-sur-Marne
77455 Marne-la-Vallée
☎ 07.69.33.75.62

✉ vincent.lepetit@enpc.fr
🔗 [vincentlepetit.github.io/](https://github.com/vincentlepetit)

est suivie dans la séquence jusqu'à ce qu'une différence de point de vue suffisante soit atteinte. Un ellipsoïde est alors reconstruit à partir de les ellipses inscrites dans les boites successives. Cet ellipsoïde est utilisé ensuite pour retrouver la boite détectée par YOLO qui correspond à l'objet. Les ellipsoïdes sont ensuite affinés à chaque image en minimisant la distance de Wasserstein entre la reprojction de l'ellipsoïde et l'ellipse inscrite dans la boite détectée par YOLO. Plusieurs façons d'utiliser les objets pour l'estimation de la pose de caméra sont également proposées.

La méthode proposée a été implémentée en C++ à partir de ORBSLAM2, et évaluée et comparée à l'état de l'art à la fois sur un jeu de données classique et des séquences capturées et calibrées par le candidat. Comparée à l'utilisation de points d'intérêt seuls, la méthode proposée est beaucoup plus robuste pour des scènes peu texturées ou des grands changements d'échelle. Comparée à une autre méthode (EAO-SLAM) utilisant les objets, les cartes produites sont plus précises (alors que EAO-SLAM a besoin de plus d'informations). Enfin, il est montré que la méthode peut servir à d'abord reconstruire une scène puis à se relocaliser dans cette scène pour une application de réalité augmentée.

Enfin, le manuscrit se termine par un résumé des contributions, et sur une suggestion d'extension avec le remplacement des ellipsoïdes par des superquadriques, pour pouvoir mieux approximer la forme des objets.

Discussion

J'ai été très impressionné par le travail présenté par le manuscrit, à plusieurs niveaux :

- l'approche présentée est très originale. Alors que beaucoup de travaux ont déjà été réalisés en SLAM, la thèse parvient à proposer une nouvelle direction qui permet d'avancer l'état de l'art en termes de robustesse et de précision.
- il y a un énorme travail d'implémentation. En particulier, la dernière méthode présentée fonctionne en temps réel, ce qui est vraiment remarquable. Il est également à noter que le candidat a mis ses implémentations en accès public.
- enfin, il y a également un énorme travail en termes d'évaluation et de comparaison avec des méthodes existantes.

L'importance des contributions et la qualité du travail ont été reconnues au travers de plusieurs publications, à 3DV, IROS, ISMAR, et IJCV, qui sont des conférences et revues très importantes du domaine. IJCV est une revue générale en vision par ordinateur, 3DV est une conférence spécialisée en vision 3D, ISMAR en réalité augmentée et IROS en robotique, ce qui montre aussi l'étendue de l'impact des contributions de la thèse.

Je suis donc extrêmement favorable à la soutenance du mémoire sous sa forme actuelle en vue de l'obtention d'un doctorat de l'Université de Lorraine.

Eric Marchand

Professeur des Universités

Université de Rennes 1, IRISA UMR 6074

Eric.Marchand@irisa.fr

Tel. +33 (0)2 99 84 74 27

Rapport sur le mémoire de thèse intitulé :

**Contribution à la précision et à la robustesse de la localisation visuelle
dans un monde d'objets**

Présenté par **Matthieu Zins**

en vue de l'obtention du titre de docteur de l'Université de Lorraine

Discipline : « Informatique »

Le manuscrit de M. Matthieu Zins porte sur des problèmes de vision par ordinateur en abordant plus précisément la problématique de la localisation d'une caméra dans un environnement structuré par des « objets ». Le champ applicatif concerné est principalement celui de la réalité augmentée mais on voit bien que d'autres applications, comme la robotique, sont évidemment envisageables.

Au cœur de cette thèse, on trouve la notion de représentation de la scène sous la forme d'ellipsoïdes. En détectant des ellipses dans les images (projection de ces ellipsoïdes) il est possible de remonter à la position (translation et rotation) de la caméra. Ce principe est utilisé pour un calcul de pose à la fois sous la forme d'un P3P (*perspective 3 point*), P2E (*Perspective 2 Ellipse*) et de la minimisation d'une erreur de reprojection permettant de prendre en compte un nombre important d'objets et donc d'ellipses. L'auteur utilise aussi cette représentation au sein d'un algorithme de SLAM (*Simultaneous Localization and Mapping*) monoculaire intégrant cette représentation sous forme d'objets/ellipsoïdes.

Synthèse du manuscrit

Le manuscrit soumis comporte 155 pages (plus les références) rédigées en Français, réparties en 4 chapitres, une introduction et une conclusion. Le manuscrit se termine par une conclusion rappelant les principales contributions et dégageant quelques pistes pour des travaux ultérieurs.

Le chapitre introductif pose bien le problème abordé, à savoir le souhait d'avoir des systèmes de localisation en vision par ordinateur et en particulier la volonté d'introduire la notion d'objet dans le système et leur modélisation approximative sous forme d'ellipsoïdes.

Le chapitre 1 porte sur un état de l'art des techniques de positionnement visuel et en particulier sur la problématique du calcul de pose qui vise à calculer la position d'un objet par rapport à la caméra (translation et rotation). L'auteur passe, à mon sens, un peu rapidement sur les approches géométriques (PnP et globalement sur la minimisation de l'erreur de reprojection) ce qui est un peu dommage car ces techniques ont une place importante dans le reste de la thèse. Il insiste plus sur les approches reposant sur le *deep learning* et sur la notion d'objet. Naturellement, les algorithmes reposant sur une représentation des objets sous formes d'ellipsoïdes sont largement mis en avant. Après des rappels bienvenus sur la représentation des ellipsoïdes et des ellipses, des approches P1E (Perspective 1 Ellipse) et P2E sont décrites. Une conclusion à ce chapitre, très dense en références bibliographiques, permettrait sans doute de mieux guider le lecteur vers les contributions à venir de l'auteur.

Le second chapitre porte donc sur la représentation des objets sous forme d'ellipsoïdes et sur la manière d'extraire leur projection (ellipses) des images pour le calcul de pose. L'auteur propose un réseau de neurones permettant d'extraire les ellipses orientées de manière cohérente (en termes de projection perspective) avec la représentation 3D de la scène sous forme d'ellipsoïdes. Une attention particulière est portée à la définition des fonctions de *loss* et donc à la représentation mathématique des ellipses orientées. En particulier une représentation sous forme de Level Sets semble donner de meilleurs résultats que des approches basées sur des histogramme. La première étape de détection des ellipses s'effectue en deux étapes : une étape de détection des objets (Faster R-CNN) puis une étape d'estimation des paramètres de l'ellipse correspondant à cet objet (le tout reposant donc sur deux CNN distincts). Le modèle de la scènes (sous formes d'ellipsoïdes) est lui obtenu à partir de 3 ellipses dans 3 images. Une fois ce modèle connu, il peut être utilisé pour annoter automatiquement les images à fournir au réseau. A partir des ellipses et donc du modèle de la scène, il est bien sur possible de calculer la pose de la caméra (par un P3P sur le centre des ellipses/ellipsoïdes). Un second réseau portant sur une approche jointe de la détection/estimation des ellipses est ensuite proposée modifiant YOLO pour estimer directement la position des ellipses. Les expériences montrent que les réseaux proposés sont capables d'inférer la position d'ellipses cohérentes avec la projection des ellipsoïdes reconstruites. Elles montrent aussi que la représentation des ellipses sous formes de level set est particulièrement adaptée au problème. Les contributions de ce chapitre ont essentiellement été publiés dans 3DV en 2020 et IJCV en 2022.

Le chapitre 3 porte plus précisément sur l'amélioration du calcul de pose à partir de ces ellipses. L'idée semble naturelle de parvenir à un processus de calcul de pose reposant sur la minimisation de l'erreur de reprojection ellipses-ellipsoïdes par une optimisation non-linéaire. Ce processus nécessite le calcul d'une distance entre ellipses projetées et ellipses extraites de

l'image. Différentes métriques donc sont considérées : distance entre boîtes, distance algébrique, distance de Wasserstein entre distribution de probabilité (earth moving distance), et finalement une distance reposant sur des level set et une représentation implicite des ellipses. C'est cette dernière représentation qui s'avère donner les meilleurs résultats en particulier quand les objets sont partiellement visibles ou occultés. Un avantage de cette approche de calcul de pose par rapport à celle du chapitre 2 est que l'on peut considérer un nombre important d'ellipses dans le calcul de pose. Une incertitude associée aux paramètres de chaque ellipse est aussi estimée et considérée dans la fonction de coût à optimiser pour donner plus ou moins d'importance à chaque objet. Les résultats obtenus sont tout à fait convaincants. Il y a cependant peu de détails sur la manière dont l'optimisation est réalisée concrètement, sur l'algorithme utilisé (GN, LM), sur le calcul ou l'estimation des Jacobiens, etc. Plus d'information sur ces points auraient été intéressants. Les contributions de ce chapitre ont essentiellement été publiés dans IEEE/RSJ IROS 2022.

La dernière étape consiste à étendre le calcul de pose au SLAM. Il s'agit donc non seulement de calculer la position de la caméra pour chaque image mais aussi de remonter à un modèle précis de la structure 3D de la scène. L'idée consiste donc à coupler ORB-SLAM2 (un SLAM basé points d'intérêt et utilisant un ajustement de faisceaux) à un SLAM basé objet introduisant ainsi un certain degré de sémantique dans le SLAM. Ces SLAM étant finalement très proches de la problématique du calcul de pose, l'intégration des travaux précédents dans un SLAM visuel se fait de manière naturelle via la définition d'une erreur de reprojection intégrant points et ellipses. J'ai trouvé étonnant de voir que c'était la distance de Wasserstein entre ellipse plutôt que celle, reposant sur une représentation sous forme de *level set* et mise en avant dans les chapitres précédents qui était utilisées. L'un des avantages du SLAM proposé mis en avant et bien illustré dans les résultats est l'apport de la notion d'objets (ellipsoïde) par rapport aux seuls points d'intérêt pour la relocalisation et la reprise de la localisation en cas de perte temporaire d'information. Une fois les résultats présentés sont tout à fait convaincants. Les contributions de ce chapitre ont essentiellement été publiés dans IEEE ISMAR 2022.

Avis sur le travail présenté

Qualité du manuscrit. Le manuscrit est écrit dans un français clair, et concis. Il est, de manière générale, bien rédigé et bien organisé. Il permet globalement de comprendre les contributions de l'auteur. La lecture en parallèle des articles en anglais rédigé par l'auteur apporte cependant parfois des éclaircissements bienvenus sur les méthodes.

Positionnement vis à vis de l'état de l'art. Outre le chapitre 2 sur le positionnement visuel, les états de l'art réalisés dans le document sont propres à chaque chapitre et sont relativement complets. Peut-être une section d'état de l'art assez générale sur les techniques de *deep learning* aurait été le bienvenu car elles sont souvent utilisées dans le cadre de cette thèse.

Validation expérimentale. Chaque contribution est validée à travers des évaluations réalisées sur des benchmark mais aussi sur des séquences d'images ad-hoc. Les choix algorithmiques sont validés et justifiés expérimentalement. L'auteur présente systématiquement des tests d'ablation très complets et pertinents.

Valorisation des travaux. Finalement précisons que les travaux de Matthieu Zins ont été publiés dans plusieurs conférences internationales (*ISMAR 2022*, *IROS 2022*, *3DV 2020*) et un journal *The Int. Journal of Computer Vision*. Ce sont trois excellentes conférences de robotique

et vision par ordinateur et un journal de référence dans le domaine. Il s'agit là sans aucun doute d'une excellente valorisation des travaux présentés. La plupart des codes informatiques permettant la reproduction de ces travaux sont aussi disponible en ligne sur *github* ce qui représente aussi une valorisation notable.

Conclusion

En conclusion de ce rapport, malgré les quelques remarques que j'ai pu faire ici ou là, les travaux de thèse de Matthieu Zins traduisent une très bonne maîtrise des formalismes de la localisation en vision artificielle. La représentation des objets sous la forme d'ellipsoïdes et de leur analyse sous forme d'ellipses dans les images pour remonter à la localisation de la caméra est une approche je considère comme élégante et efficace.

En conséquence, je donne un avis très favorable pour que les travaux de Matthieu Zins soient présentés publiquement en vue de l'obtention du grade de docteur de l'université de Lorraine.

Rennes, le 17 novembre 2022



Eric Marchand
Professeur des universités
Université de Rennes 1,
IRISA UMR 6074, Inria Rennes