DUNE-M : un estimateur local d'incertitude sur la position des points d'intérêt évolutif dans le temps

Katia Sousa Lillo[1,3] Andrea De maio[2] Simon Lacroix[2] Amaury Negre[1]

[1] GIPSA-Lab, Université Grenoble Alpes, CNRS, Grenoble INP, 38000 Grenoble, France

[2]LAAS-CNRS, Université de Toulouse, CNRS 7, Avenue du Colonel Roche, 31031 Toulouse, France

[3]Thales Avionics, THALES group, 25 rue Jules Védrines, 26000 Valence, France

Résumé – L'estimation de l'incertitude des points d'intérêt est essentielle pour les systèmes basés sur la vision, tels que la navigation visuelle. Nous montrons que les erreurs inhérentes au suivi visuel, en particulier en utilisant le tracker KLT, sont temporellement corrélées et peuvent être capturées par un réseau de neurones récurrent. Le système proposé intègre tout l'historique de suivi d'un point d'intérêt. DUNE-M est entraîné et évalué sur le jeu de données KITTI [4], mettant en évidence de bons résultats par rapport à l'état de l'art. Une application d'odométrie visuelle est présentée illustrant les bénéfices de l'utilisation d'une mémoire.

Abstract – Uncertainty estimation of visual feature is essential for vision-based systems, such as visual navigation. We show that the errors inherent in visual tracking, especially using the KLT tracker, are temporally correlated and can be captured by a recurrent neural network. The proposed system integrates the entire tracking history of a feature position. DUNE-M is trained and evaluated on KITTI [4] datasets, showing good results compared to the state of the art. A visual odometry application is presented to illustrate the benefits of using a memory.

1 Introduction

La détection et le suivi de points dans une séquence d'images est un processus essentiel pour de nombreuses applications robotiques, en particulier, l'odométrie visuelle qui consiste en l'extraction de caractéristiques, le suivi de points d'intérêt et l'estimation du mouvement. La possibilité d'estimer précisément l'erreur issue du suivi des points d'intérêt est un avantage dans la sélection des meilleures points pour l'estimation du mouvement. De plus, dans un système intégrant des données inertielles et visuelles, tel que [6], l'utilisation d'un modèle d'erreur de suivi des points pourrait permettre de propager correctement les erreurs tout au long du processus de fusion.

La plupart des approches d'estimation du mouvement issue de points d'intérêt supposent que l'incertitude sur la position de ces points est constante. L'estimateur d'incertitude [7] sur la position des points d'intérêt permet de qualifier la précision des observations entre deux instants k et $k+1$. Toutefois, l'incertitude locale des points d'intérêt est estimée de façon indépendante de l'évolution de la position des points au cours du temps au delà de $k+1$. [7] suppose donc que l'incertitude sur la position des points de $k+1$ à $k+2$ est indépendante et décorrélée de l'incertitude sur la position des points de k à $k+1$.

L'évolution de la position d'un point étant dépendante des positions précédentes du point, l'évolution de l'incertitude sur les positions d'un point dépend également de l'incertitude sur les positions précédentes du point.

Quelques travaux ont été consacrés à l'intégration de la mémoire, en particulier des couches LSTM (Long-Short Term Memory), dans l'estimation de la position d'un système mobile [8]. Pose-LSTM [8] estime la pose d'un système mobile à partir des poses précédentes prises par le système mobile. En parallèle, des études sur la position des points étudient les LSTM, notamment SURF-LSTM [2] qui propose un descripteur robuste SURF intégrant des cellules mémoires LSTM. Dans les solutions existantes, aucune ne s'intéresse au problème d'estimation de l'incertitude.

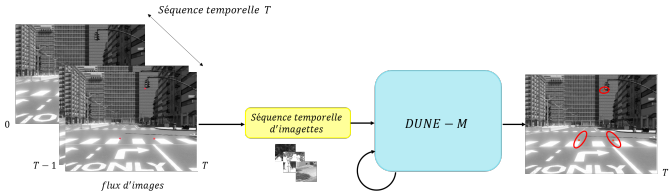


Figure 1 : Deep UNcertainty Estimation avec Mémoire (DUNE-M). Le système prend en entrée une séquence temporelle d'imagettes envoyées itérativement à DUNE-M pour prédire la covariance sur la position des points d'intérêt relative à l'image finale de la séquence temporelle T.

Nous introduisons DUNE-M (Deep UNcertainty Estimation-Memory) un estimateur d'incertitude local basé sur un modèle de réseau de neurones (RdN) évolutif dans le temps intégrant une mémoire comme illustré figure 1. Ainsi, l'ensemble de l'historique de suivi de la position d'un point est exploité pour fournir des incertitudes plus précises. Nous montrons que l'intégration d'une mémoire dans notre estimateur permet d'affiner la précision sur l'estimation de l'incertitude de la position des points de suivi au cours du temps.

2 DUNE-M

La section suivante est consacrée à définir le fonctionnement de DUNE-M, en particulier définir le format des entrées et des sorties ainsi que l'architecture du RdN.

2.1 Format des données

L'estimateur de covariance proposé est un modèle basé sur un RdN qui estime les incertitudes locales sur la position des points d'intérêt en apprenant l'erreur entre la position des points observés et leurs positions réelles. L'erreur ${}^{uv}\boldsymbol{e}_k$ est définie par la distance euclidienne entre la position réelle du point ${}^{uv}\boldsymbol{m}_k \in \mathbb{R}^2$ avec des coordonnées de pixels (u, v) à l'instant k et le point mesuré ${}^{uv}\tilde{\boldsymbol{m}}_k \in \mathbb{R}^2$:

$$ {}^{uv}\boldsymbol{e}_k = {}^{uv}\,\tilde{\boldsymbol{m}}_k - {}^{uv}\,\boldsymbol{m}_k \tag{1}$$

Un point d'intérêt peut être observé à plusieurs instants. Dans [7], les données d'entrée sont des stacks de deux imagettes $\boldsymbol{W}_k$ et $\boldsymbol{W}_{k+1}$, c'est-à-dire une portion de l'image de dimension (21×21), centrées en ${}^{uv}\tilde{\boldsymbol{m}}$ le point tracké entre deux instants k et $k+1$. Considérons une séquence d'imagettes successives, les données d'entrées de DUNE-M sont représentées par la concaténation de l'ensemble des stacks de deux imagettes entre deux instants $\forall k \in |[0; T]|$ de la séquence temporelle T, tel que décrit Fig. 2.

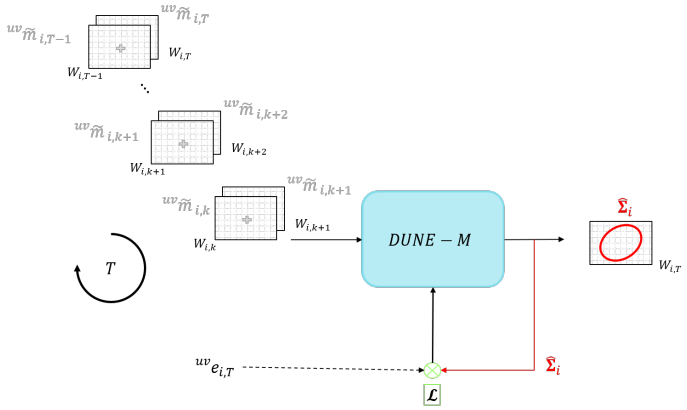


FIGURE 2 : DUNE-M prend en entrée un échantillon i décrit par une séquence imagettes concaténées $[{}^{uv}\boldsymbol{W}_{i,k}\,;\,{}^{uv}\boldsymbol{W}_{i,k+1}]$ centrées en ${}^{uv}\tilde{\boldsymbol{m}}_{i,k}$ et ${}^{uv}\tilde{\boldsymbol{m}}_{i,k+1}$ $\forall k \in |[0; T]|$, et prédit la matrice de covariance $\tilde{\boldsymbol{\Sigma}}_i$ associée à ${}^{uv}\boldsymbol{W}_{i,T}$. La fonction de coût appliquée $\mathcal{L}$ compare l'erreur mesurée à T ${}^{uv}\boldsymbol{e}_{i,T}$ avec la matrice de covariance prédite $\tilde{\boldsymbol{\Sigma}}_i$

DUNE-M fournit alors une covariance locale $\tilde{\boldsymbol{\Sigma}}_i \in \mathbb{R}^{2\times2}$ dans le repère image $\boldsymbol{\mathcal{F}}_{pix}$ [7] sur la position de la dernière observation d'un point d'intérêt suivi ${}^{uv}\tilde{\boldsymbol{m}}_{i,k=T}$ d'un échantillon i évaluée à partir de tout l'historique de suivi du point d'intérêt le long de cette séquence.

La fonction de coût $\mathcal{L}$ appliquée, issue de [7], maximise la vraisemblance de l'erreur ${}^{uv}\boldsymbol{e}$ suivant une distribution gaussienne. Avec cette approche, la matrice de covariance est apprise par rapport à l'entrée, ce qui permet de capturer l'incertitude hétéroscédastique de chaque échantillon.

2.2 Structure du réseau de neurones

Par choix, DUNE-M utilise des paires d'images successives. La structure est inspirée de [7] et permet d'évaluer directement l'impact de l'utilisation d'une mémoire LSTM. Les paires d'imagettes sont itérées dans le réseau de neurones récurrent DUNE-M.

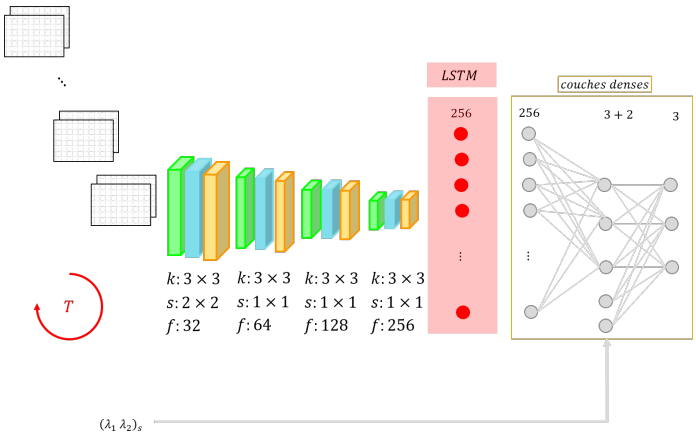


FIGURE 3 : Structure de l'estimateur DUNE-M d'incertitude de la position du suivi des points d'intérêt d'une séquence temporelle intégrant les paramètres de Harris $(\lambda_1, \lambda_2)_i$ - T est la taille de la séquence temporelle, i un échantillon, k indique la dimension des Kernels ; s la dimension des strides ; et f le nombre de filtres. Les couches vertes représentent les convolutions, les bleues les couches de normalisation et les jaunes les dropouts. Chaque couche du RdN, à l'exception de la dernière, est suivie d'une fonction d'activation *leakyrelu*

La construction de DUNE-M a été réalisée en intégrant une couche LSTM à la sortie du bloc convolutif et en entrée des couches denses (Fig. 3). L'objectif est de capturer les caractéristiques sortantes des couches convolutionnelles et de corréler ces caractéristiques temporellement. Intuitivement, nous faisons l'hypothèse que l'évolution de l'incertitude sur la position d'un point de suivi dépend de ses positions antérieures. L'utilisation d'une mémoire LSTM est donc le moyen d'observer l'évolution des caractéristiques relatives à un point suivi le long d'une séquence temporelle et, en particulier, de pouvoir apprendre l'incertitude sur la position d'un point suivi à partir de l'historique complet de l'évolution du point depuis sa détection. La prise en compte de l'historique du suivi de chaque point est un levier d'amélioration à l'estimation de la précision d'un point où l'incertitude n'est plus cumulée au cours du temps mais adaptée à la situation (*i.e.*, dégradation ou amélioration du champ de vue).

3 Résultats

Tous nos résultats ont été réalisé sur le jeu de données KITTI [3]. Le modèle a été appris à partir du scénario `2011_09_26_0002`. `2011_09_26_0005` a été utilisé pour tester le modèle au cours de l'apprentissage. Enfin, le scénario `2011_09_26_0113` est adopté pour la phase d'évaluation de DUNE-M.

Au cours de chaque scénario, les points sont détectés et trackés. Par choix [7], notre étude est réalisée à partir du tracker Kanade Lucas Tomatsi (KLT) [1] après une initialisation avec le détecteur de Harris [5]. La structure 3D exacte du monde (*i.e.*, carte de profondeur, points géolocalisés, etc ..) n'est pas disponible directement. Aussi, l'application du tracker en rétropropagation (*i.e.*, sur un aller-retour) permet de mesurer le biais induit par la méthode de tracking et en particulier, d'évaluer l'incertitude de position sur ce point. En particulier, on se place à $T = 5$, ainsi on opère un aller retour sur un total de 10 images (5 allers et 5 retours).

L'erreur sur la séquence est définie par la distance entre la position du point initialement détecté et la position du point à l'issue du suivi sur l'aller-retour, idéalement identique à celle de la première détection. Dès lors, on suppose une répartition de l'évolution uniforme.

3.1 Métriques

La performance du modèle est évaluée en comparant l'erreur réelle ${}^{uv}\boldsymbol{e}_k$ et l'incertitude prédite $\boldsymbol{\Sigma}_k$ sur la base de deux métriques : l'erreur normalisée (NNE) et la distance de Mahalanobis (MD), telles que définies dans [7]. Le modèle d'incertitude optimal est obtenu pour des valeurs égales à 1. Les valeurs inférieures correspondent à une estimation pessimiste de l'incertitude, et les valeurs supérieures à un modèle d'estimation optimiste.

3.2 Préparation des données

70% des données sont utilisées pour entraîner le réseau neuronal, 15% sont conservées pour la phase de test et 15% pour l'évaluation. Nous effectuons un apprentissage sur 150 epochs avec une taille de bacth de 64. Le pas d'apprentissage initial est identique ainsi que la méthode d'optimisation Nadam. Le modèle final est sélectionné comme étant celui qui donne les meilleurs résultats MD et NNE sur les données de test.

3.3 Analyse des résultats

DUNE-M est évalué à partir des données d'évaluations issues du scénario `2011_09_26_0113`. Les incertitudes prédites sont analysées à 1σ, 2σ et 3σ et référencées dans le tableau Tab. 1. Pour une distribution normale, les quantiles devraient inclure respectivement 68%, 95% et 99, 7% des erreurs observées ${}^{uv}\boldsymbol{e}$.

Nous comparons nos résultats (Tab. 1) à la méthode [9] et à un modèle de covariance fixe (Fix-C) égale à $0.5 \cdot 5 \cdot 2$ pixel.

La méthode [9] donne des résultats trop optimistes avec une distance de mahalanobis $MD \sim 1.5$, validée également par $NNE \sim 1.4$. Ce résultat est visible dans la répartition des quantiles entre 84% et 90% des erreurs observées à 1σ. À l'inverse, la méthode Fix-C est pessmiste comme précédemment avec une distance de mahalanobis $MD \sim 0.3$ et $NNE \sim 0.3$. DUNE-M montre les meilleurs résultats notamment avec $MD \sim 1.2$ et $NNE \sim 1$. De plus, 90% des erreurs observées sont comprises dans 3σ. Ces résultats sont très satisfaisants et constituent les meilleurs résultats disponibles dans la littérature sur l'estimation d'incertitude de la position des points d'intérêt.

Métriques	DUNE-M	[9]	Fix-C
$3\sigma_u$	90.03	90.63	98.3
$3\sigma_v$	90.76	96.0	99.9
$2\sigma_u$	88.33	95.21	96.46
$2\sigma_v$	88.26	88.7	99.5
$1\sigma_u$	83.56	84.56	94.43
$1\sigma_v$	82.80	91.23	98.5
MD	**1.17**	1.46	0.26
NNE	**1.07**	1.37	0.26

TABLE 1 : Comparaison des résultats de DUNE-M par rapport à l'état de l'art

3.4 Évaluation des prédictions de DUNE sur l'estimation du mouvement

Nous appliquons une validation géométrique dont la vocation est d'évaluer l'impact des variances sur la reconstruction du mouvement.

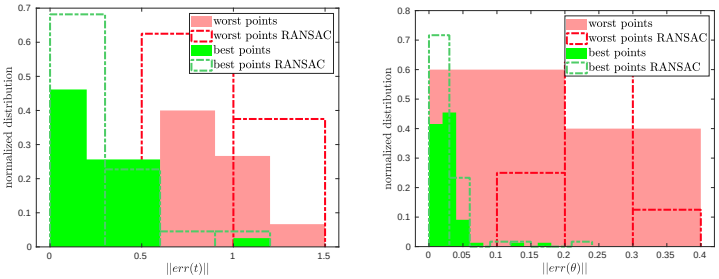
La pose de la caméra n'est pas connue. Toutefois, l'approche de la rétropropagation permet de revenir à l'état initial permettant de caractériser la vérité terrain du mouvement de la caméra par la transformation ${}_{k+1}T_{\text{ground truth}} = \begin{bmatrix} \mathcal{I}_{3\times3} & \not\vdash_{3\times1} \end{bmatrix}$.

Par ailleurs, le mouvement de la caméra peut également être estimé sur l'aller retour par stéréovision. La transformation prédite ${}_{k}T_{\text{estimated}}$ est évaluée. L'erreur sur la transformation rigide peut ainsi être calculée avec :

$$ {}_{k+1}T_{\text{err}} = {}_{k+1}T_{\text{ground truth}} \cdot {}_{k+1}T^{-1}_{\text{estimated}} \qquad (2) $$

où ${}_{k+1}T_{\text{err}}$ définit l'erreur de transformation rigide à l'instant $k+1$.

Les meilleurs points (*i.e.*, *best points*) sont sélectionnés à partir du critère $\sqrt{tr(\boldsymbol{\Sigma}_i)} \leq 0.5 * T$, $i \in |[0; s]|$ et les pires points (*i.e.*, *worst points*) vérifient $\sqrt{tr(\boldsymbol{\Sigma}_i)} > 0.5 * T$. Les résultats sont présentés Fig. 4.



(a) (b)

FIGURE 4 : Impact sur l'estimation du mouvement - (a) Erreur sur la translation (en m) ; (b) Erreur sur la rotation (en deg).

Les résultats (Fig. 4) montrent une erreur moyenne de 0.3017 m en translation pour les meilleurs points et 0.7604 pour les pires points. Ces métriques sont comparables à celles issues de RANSAC avec une erreur moyenne

en translation de 0.3017 m pour les meilleurs points et 0.8521 pour les pires (Fig. 4a). La sélection des meilleurs ayant une meilleure incertitude sur sa position met en évidence une erreur moindre lors de l'estimation de la pose d'un système mobile et montre de meilleurs résultats. Ces résultats sont également visibles sur les erreurs de rotation (Fig. 4b) avec une erreur moyenne de 0.0298 deg pour les meilleurs points et 0.1627 deg pour les pires ; avec RANSAC on estime une erreur moyenne de 0.0272 deg pour les meilleurs points et 0.2321 deg pour les pires.

4 Discussion

4.1 Influence de la mémoire

Nous comparons notre méthode DUNE-M à [7]. Pour chaque séquence temporelle, [7] retourne T prédictions d'incertitude. La prédiction finale de la séquence est la somme des T prédictions d'incertitude fournies par [7]. Les résultats sont mis en relation Tab. 2 avec les prédictions issues de DUNE-M, estimant directement la prédiction issue de la séquence.

Métriques	[7]	DUNE-M
$3\sigma_u$	93.46	90.03
$3\sigma_v$	96.16	90.76
$2\sigma_u$	91.63	88.33
$2\sigma_v$	94.73	88.26
$1\sigma_u$	88.4	73.56
$1\sigma_v$	90.93	72.80
MD	0.70	1.17
NNE	0.68	1.07

Table 2 : Comparaison des résultats DUNE-M et [7]

Le modèle d'estimation issu de [7], comme étant la somme des incertitudes des entrées de la séquence temporelle est pessimiste. En effet, $NNE \sim 0.7$, de même que la distance de mahalanobis $MD = 0.7$. $DUNE - M$ montre de meilleurs résultats, notamment avec $NNE \sim 1$ et $MD = 1.17$. Ainsi, l'influence et l'impact de l'ajout d'une couche LSTM permet d'affiner l'estimation des covariances à partir de l'historique d'observation des points d'intérêt.

Par ailleurs, la mise en place d'un réseau récurrent offre une certaine flexibilité du format des données d'entrée. En effet, elle rend possible l'utilisation de séquence à taille variable en entrée comme en sortie du réseau. Dans notre cas, DUNE-M permet d'accueillir des jeux de données à partir de points observés ayant un nombre d'observation variable.

5 Conclusion

Nous avons présenté une nouvelle architecture de réseau de neurones récurrent qui produit des estimations de l'incertitude sur la position des points d'intérêt suivis par le tracker KLT intégrant des corrélations temporelles avec les positions antérieures des points. Nous avons montré qu'il est possible de capturer l'évolution temporelle dynamique d'un point suivi. Le modèle prédictif DUNE-M a été comparé et validé par deux métriques dédiées.

Par la suite, DUNE-M peut être intégré à diverses applications robotiques. En particulier, le format de données avec des séquences à taille variable offre plus de souplesse dans le choix de la sélection de points d'intérêt, ce qui constitue la plus grande force de l'architecture mise en place avec DUNE-M.

Références

[1] D. Lucas Bruce et Takeo Kanade : An iterative image registration technique with an application to stereo VIsion. 1981.

[2] Ahmed Elmoogy, Xiaodai Dong, Tao Lu, Robert Westendorp et Kishore Reddy : Surf-lstm : A descriptor enhanced recurrent neural network for indoor localization. *In 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, pages 1–5. IEEE, 2020.

[3] Andreas Geiger, Philip Lenz, Christoph Stiller et Raquel Urtasun : Vision meets robotics : The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[4] Andreas Geiger, Philip Lenz et Raquel Urtasun : Are we ready for autonomous driving ? the kitti vision benchmark suite. *In Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[5] C. Harris et M. Stephens : A combined corner and edge detector. *In Procedings of the Alvey Vision Conference 1988*, 1988.

[6] Anastasios I. Mourikis et Stergios I. Roumeliotis : A multi-state constraint kalman filter for vision-aided inertial navigation. *In Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3565–3572, 2007.

[7] Katia Sousa Lillo, Andrea de Maio, Simon Lacroix, Amaury Nègre, Michèle Rombaut, Nicolas Marchand et Vercier Nicolas Vercier : Dune : Deep uncertainty estimation for tracked visual features. *In IPAS 2022-5th IEEE International Conference on Image Processing, Applications and Systems (IPAS 2022)*. IEEE, 2022.

[8] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck et Daniel Cremers : Image-based localization using lstms for structured feature correlation. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 627–637, 2017.

[9] Xue Iuan Wong et Manoranjan Majji : Uncertainty quantification of lucas kanade feature track and application to visual odometry. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 950–958, 2017.