Localisation visuelle multimodale à long terme

Fabien Bonardi, Samia Ainouz, Rémi Boutteau, Yohan Dupuis, Xavier Savatier, Pascal Vasseur

Localisation visuelle multimodale à long terme

Fabien BONARDI[1], Samia AINOUZ[1], Rémi BOUTTEAU[2], Yohan DUPUIS[3], Xavier SAVATIER[2], Pascal VASSEUR[1]

[1]Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes
Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France

[2]Institut de Recherche en Systèmes Électroniques Embarqués
Normandie Univ, UNIROUEN, ESIGELEC, IRSEEM, 76000 Rouen, France

[3]Centre d'Études et d'Expertise sur les Risques, l'Environnement, la Mobilité et l'Aménagement
CEREMA, 76000 Rouen, France

{prénom.nom}@litislab.fr, @esigelec.fr, @cerema.fr

Résumé – La localisation visuelle à long terme en milieu extérieur demeure un défi dû aux nombreux changements d'apparence de l'environnement. Ce problème est d'autant plus complexe lorsque l'association d'images doit être réalisée avec des capteurs différents, en particulier de sensibilités spectrales différentes. Dans ce travail, nous nous sommes concentrés sur ces deux types de contraintes et proposons une nouvelle méthode d'extraction de caractéristiques qui s'adapte à la multimodalité. Nous avons conduit une évaluation expérimentale sur des jeux de données acquis sur le long terme et avec des capteurs de natures différentes (optiques, tailles de capteurs et réponse spectrale). Les tests que nous avons menés tendent à démontrer que notre méthode apporte des améliorations par rapport à l'état de l'art.

Abstract – Long-term *place recognition* in outdoor environments remains a challenge due to high appearance changes in the environment. The problem becomes even more difficult when the matching has to be made with different visual sources, particularly with different spectral ranges. In this paper, we emphasized our work on both constraints and proposed a new feature extraction method adapted to multimodal association. We conducted an evaluation on long-term datasets coming from different imaging sources (optics, sensors size and spectral ranges). The tests we performed tend to demonstrate that our method brings a significant improvement.

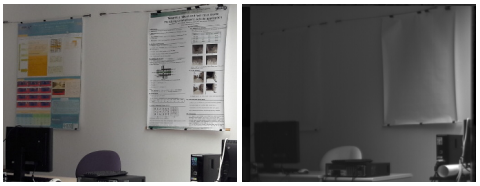
1 Introduction

Les méthodes de localisation et cartographie basées vision ont connu un engouement certains au cours de ces dernières décennies. En guise de solution *ad-hoc* ou en complément d'autres capteurs de différentes natures, les caméras utilisées pour ces tâches sont variées et diffèrent d'un système à l'autre. Il en résulte souvent des méthodes et données qui sont liées au système d'origine et son mode d'apprentissage. Ces méthodes sont ainsi difficilement généralisables à d'autres systèmes et leurs capteurs. La reconnaissance de lieux (*place recognition*) est une étape essentielle d'un processus de localisation : en cherchant à associer des images acquises avec des images géolocalisées issues d'une mémoire composée lors d'une première expérience, elle permet d'estimer une localisation approximative, de réduire la dérive d'un processus de SLAM (*Simultaneous Localization And Mapping*) ou encore d'améliorer la précision de la localisation [1].

Aux changements d'apparence des images dus aux différentes sensibilités des capteurs s'ajoutent les variations à long terme induites par l'environnement : illumination de la scène, changements dus aux saisons et au climat, *etc*. Plusieurs contributions ont démontré les difficultés inhérentes à ces problématiques [2, 3]. Il convient dans ce cas de concevoir une méthode d'extraction de caractéristiques invariantes aux changements d'apparence de l'environnement et de modalités.

Les caractéristiques (*features*) extraites des images pour des tâches de localisation se divisent en deux catégories principales : les caractéristiques globales et les caractéristiques locales. Les auteurs dans [4] par exemple, proposent une méthode de localisation globale transformant dans un premier temps les images brutes afin de réduire l'influence des variations de l'apparence de la scène dues aux changements d'illumination et représentent ensuite les données sous la forme d'une signature compacte. L'usage de «signatures globales» requiert néanmoins un point de vue proche entre les images à associer. C'est pourquoi d'autres approches préfèrent avoir recours aux caractéristiques locales, qui permettent par la suite d'envisager la recherche d'une transformation géométrique entre deux images. Ces méthodes, souvent au cœur des processus de SLAM permettent d'établir une représentation 3D de l'environnement [5]. Les recherches récentes ont remis au goût du jour les réseaux de neurones profonds : ceux-ci peuvent se substituer totalement au processus de reconnaissance d'images comme dans [6], ou bien seulement une partie du processus (*transfert learning*), comme la description et la métrique utilisée pour comparer des points d'intérêt dans [7].

La question de la multimodalité se pose lorsque l'on cherche à associer des données issues de capteurs différents, et particu-



(a) visible (b) infrarouge

FIGURE 1 – Sérigraphie dans les spectres visible et infrarouge : alors que le contenu du poster est visible dans l'image couleur (1a), le poster semble vierge dans l'image infrarouge (1b).

lièrement en imagerie lorsque les réponses spectrales de deux caméras ne sont pas les mêmes. Utiliser une caméra infrarouge peut être intéressant pour envisager une vision nocturne par exemple. Les auteurs de [8] proposent par exemple une modification des extracteurs de points d'intérêt courants de manière à les rendre invariants aux changements d'apparence induits par le changement de modalité.

Dans ce papier, nous proposons une nouvelle méthode d'extraction de caractéristiques invariantes aux changements à long terme de l'environnement ainsi qu'aux changements de modalité. Dans un premier temps, nous conduisons une étude sur la répétabilité des détecteurs courants. Nous proposons par la suite une méthode de description multi-échelle à l'aide d'histogrammes de gradients modifiés de sorte que la composition d'un dictionnaire de mots visuels soit réalisée avec la distance de Hellinger (partie 2). Nous présentons les résultats expérimentaux menés sur différentes base d'images en les comparant avec les méthodes courantes de l'état de l'art dans la partie 3.

2 Méthode proposée

2.1 Détection de points d'intérêt

Comme nous pouvons le constater sur la figure 1, les images acquises dans le spectre infrarouge présentent des différences d'apparence notables en comparaison de la même scène acquise à l'aide d'une caméra visible : les textures des objets ont tendance à disparaître dans l'infrarouge, de même que certaines matières apparaissent très claires dans une modalité mais foncé dans l'autre, et *vice versa*. C'est pourquoi nous avons décidé dans un premier temps d'évaluer la répétabilité des méthodes de détection courantes afin de choisir celle qui est la plus invariante à ces changements de modalité. Nous avons pour cela utilisé le jeu d'images visible/infrarouge proche proposé dans [7]. Ces tests de répétabilité nous ont amenés à la conclusion que les détecteurs de coins étaient plus appropriés dans le cas d'une extraction d'informations issues de deux modalités différentes.

2.2 Motif de description multi-échelle

Afin d'avoir une description du point d'intérêt ainsi détecté à plusieurs échelles, nous établissons une pyramide de Gaussienne de l'image à l'aide d'un noyau gaussien de 5×5 pixels. Cela permet également d'adoucir les contours et de retirer les composantes en hautes fréquences. On sous-échantillonne les images résultantes en ne conservant qu'un pixel sur deux à la fois en hauteur et en largeur. Nous calculons ainsi 5 niveaux de résolutions différentes.

2.3 Histogrammes de gradients réduits

Sur chaque niveau de résolution, nous établissons une description inspirée du motif de SIFT : on considère un voisinage de 4×4 régions de 4×4 pixels. Contrairement au descripteur SIFT, nous n'effectuons pas de pondération sur les zones ainsi délimitées. Pour chaque zone, nous calculons un histogramme de gradients orientés. [8] soulève le problème de l'«inversion des gradients» qui peuvent apparaître lorsque l'on passe d'une image visible à une image infrarouge : en effet, les changements de teintes pour certains matériaux que nous avons évoqués peuvent inverser les zones de noir et de blanc dans des régions à fort contraste. Il en résulte des gradients ayant globalement la même direction et la même amplitude mais des sens opposés d'une modalité à l'autre. De façon à éviter cet écueil, nous sommons les composantes des histogrammes représentant les gradients de sens opposés.

2.4 Application du noyau de Hellinger

Plusieurs études ont démontré que la distance Euclidienne n'est pas la plus appropriée pour comparer des descripteurs qui agrègent l'information sous forme d'histogrammes [9]. Les mesures de Hellinger ou χ^2 sont plus adaptées dans ce cas. Les auteurs de [9] ont ainsi proposé RootSIFT, une évolution de SIFT permettant d'utiliser la distance Euclidienne sur ces nouvelles descriptions de sorte à ce qu'elle soit équivalente à la distance de Hellinger sur des descriptions SIFT traditionnelles. Nous appliquons la même astuce à nos descripteurs : soit $\mathbf{x}_1$ et $\mathbf{x}_2$ deux histogrammes unitaires ($\|\mathbf{x}_i\|_2 = 1$), la distance euclidienne qui les sépare est donnée par l'équation 1 :

$$d_{Eucl}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{2 - 2K_{Eucl}(\mathbf{x}_1, \mathbf{x}_2)} \tag{1}$$

avec $K_{Eucl}(\mathbf{x}_1, \mathbf{x}_2) = {\mathbf{x}_1}^T \mathbf{x}_2$ le noyau euclidien (ou similarité). Nous souhaitons remplacer ce noyau par la similarité de Hellinger donnée à l'équation 2 :

$$K_{Hell}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{n} \sqrt{x_{1j} x_{2j}} \tag{2}$$

avec $\mathbf{x}_1$ et $\mathbf{x}_2$ deux histogrammes normalisés par la norme $L1$ ($\sum_j^n x_{ij} = 1$ et $x_{ij} \geq 0$). On peut donc calculer la similarité de Hellinger entre deux descripteurs en normalisant les histogrammes et en passant chaque élément à la racine carrée. Ainsi, $K_{Eucl}(\sqrt{\mathbf{x}_1}, \sqrt{\mathbf{x}_2}) = \sqrt{\mathbf{x}_1}^T \sqrt{\mathbf{x}_2} = K_{Hell}(\mathbf{x}_1, \mathbf{x}_2)$. Appliquer la distance euclidienne à ces histogrammes modifiées est équivalent à utiliser la similarité de Hellinger sur les histogrammes initiaux.

2.5 «Sac de mots visuels»

Une approche courante pour effectuer des requêtes sur base d'image est de composer un «sac de mots visuels» (*Bag of Visual Words*), présentée par [10]. Nous considérons pour cela la séquence d'images utilisée comme mémoire : les caractéristiques sont extraites de chaque image ; l'ensemble de ces caractéristiques sont agrégées avec un algorithme K-mean afin de constituer un «vocabulaire visuel» de 1000 mots. Chaque image en mémoire se voit alors associé un histogramme de mots visuels qui sera utilisé plus tard comme signature de comparaison avec les images acquises en ligne. Les signatures les plus proches des images en mémoire permettront de trouver potentiellement le même lieu représenté.

3 Résultats expérimentaux

Afin d'évaluer notre méthode, nous avons choisi quatre jeux de données : les deux premiers confrontent des images visibles et infrarouges, respectivement proche infrarouge (*Near Infra-Red*) présenté dans [8] et infrarouge lointain (*Long-Wavelength Infrared*) présenté dans [7] prises au même moment. Le troisième nommé VPrice présente des images visibles de même scènes pendant deux saisons différentes. Nous proposons le dernier jeu d'images qui opposent des images visibles et infrarouges (SWIR) prises à plusieurs mois d'intervalle. Les résultats sont synthétisés dans le tableau 1. Nous avons réalisé une évaluation croisée en alternant la modalité utilisée en mémoire pour chaque ensemble d'images sauf le jeu VPRiCE qui utilise la modalité visible dans les deux cas.

3.1 Association visible/proche infrarouge

Le premier jeu de données est composé de plusieurs sous-ensembles d'images classés par thème. Nous avons choisi d'effectuer nos tests sur les ensembles *urban*, *street* et *country*. Les images de ce jeu de données ont été rectifiées si bien que les points de vues entre mémoire et banque de test sont strictement identiques. Les résultats sont présentés dans le tableau 1 pour chaque sous-groupe. Les résultats avec les descripteurs courants, en particulier l'association FAST-SIFT, présentent un taux d'appariements réussis entre 96% et 100% sur les bases *urban* et *street*. Notre méthode obtient des résultats comparables (entre 94% et 100%) et montre son intérêt sur la base *country* avec un taux de 80% lorsque l'infrarouge est utilisé comme mémoire. Il faut noter que le sous-ensemble *country* présente une végétation fournie, pour laquelle le phénomène d'inversion des gradients est marquant (un exemple de paire est donné en figure 2). C'est pour cette raison que le résultat sur ce sous-ensemble est en baisse par rapport à *urban* et *street*.

3.2 Association visible/infrarouge lointain

Cette deuxième modalité infrarouge est particulièrement appréciée pour son utilisation en vision nocturne. Comme pour le jeu précédent, les images sont rectifiées ; une paire d'images est donnée en exemple en figure 3. Les deux modalités sont ici plus éloignées ; les résultats (colonnes «Dataset Barcelone») montrent que notre méthode apporte ici une réelle amélioration du taux d'appariements réussis avec 61% contre 52% au mieux pour les autres méthodes lorsque les images LWIR sont utilisées comme mémoire et 56% contre 38% au mieux dans le cas où les images visibles composent la mémoire.

FIGURE 2 – Une paire d'images visible/infrarouge proche

FIGURE 3 – Une paire d'images visible/infrarouge lointain

3.3 VPrice : association à long terme

Dans cette partie, nous avons utilisé une partie des images proposées dans le cadre du challenge VPrice [1] ; un exemple est donné en figure 4. Ce challenge a été pensé de façon à évaluer des méthodes de reconnaissance de lieux visuelles à long-terme avec des scènes présentant de fortes modifications d'apparences dues aux saisons. De plus, les images de ce jeu de données n'ont pas été rectifiées si bien que le point de vue n'est pas strictement identique pour deux images de la même paire. Une fois de plus, notre descripteur apporte de réelles améliorations sur les résultats (colonne VPRiCE), avec 73% de réussite contre 68% au mieux pour les autres méthodes.

3.4 Association visible/infrarouge à long-terme

Pour cette dernière partie, nous avons composé notre propre jeu de données afin de cumuler à la fois la contrainte de la multimodalité à celle de l'association à long terme (saisons différentes). Nous avons pour cela utilisé une caméra visible ainsi

1. roboticvision.atlassian.net/wiki/pages/viewpage.action ?pageId= 14188617

FIGURE 4 – Une paire d'images du challenge VPrice

	Dataset EPFL						Dataset Barcelone		Dataset VPRiCE	Notre dataset	
	urban		street		country						
En mémoire	Visible	NIR	Visible	NIR	Visible	NIR	LWIR	Visible		SWIR	Visible
SIFT - SIFT	96%	94%	78%	66%	40%	34%	9%	9%	36%	15%	5%
SIFT - GISIFT	98%	94%	70%	64%	34%	32%	11%	9%	42%	20%	10%
FAST - SIFT	**100%**	**100%**	**96%**	**96%**	**75%**	76%	22%	9%	68%	15%	**20%**
Harris - SIFT	**100%**	98%	88%	92%	73%	61%	18%	20%	52%	15%	10%
Harris - GISIFT	**100%**	**100%**	90%	90%	69%	67%	52%	38%	47%	25%	15%
Notre méthode	**100%**	**100%**	**96%**	94%	73%	**80%**	**61%**	**56%**	**73%**	**35%**	15%

TABLE 1 – Taux des bons appariements sur les différents jeux de données en fonction de la modalité utilisée en mémoire

FIGURE 5 – Une paire d'images du jeu visible/SWIR

qu'une caméra SWIR (*Short WaveLength InfraRed*). Nous n'avons pas rectifié les images, d'autant plus que la résolution des images infrarouges est plus faible que celles des images visibles comme peut en témoigner la paire d'images en figure 5. Le taux de bons appariements atteint ici 35% (contre 25% pour la meilleure) avec notre méthode lorsque la modalité SWIR est utilisée en mémoire et 15% (contre 20% pour la meilleure) lorsque c'est le cas de la modalité visible. On remarque néanmoins que les résultats sont meilleurs lorsque les images issues de la caméra SWIR sont utilisées comme séquence en mémoire : ceci est dû au fait que ces images sont de résolution plus basse et assez bruitées, l'algorithme du K-mean convergeant alors vers une représentation plus générale de chaque mot visuel.

4 Conclusion

La problématique de la localisation visuelle à long terme reste une question ouverte : face à de nombreux changements d'apparence de l'environnement, il est difficile de concevoir une méthode se concentrant sur les informations images pertinentes et suffisamment invariantes à ces changements multiples. Le problème est d'autant plus complexe lorsqu'il s'agit d'apparier des images provenant de modalités différentes.

Dans ce papier, une méthode d'extraction de caractéristiques abordant la problématique de la localisation visuelle sous ces deux contraintes (association à long terme et multimodalité) a été présentée. Nous avons montré que notre méthode, en comparaison des méthodes usuelles de l'état de l'art, présente un réel apport. Les résultats sur les jeux de données les plus ambitieux montrent néanmoins que ces méthodes sont loin d'être parfaites. Il serait envisageable par exemple d'améliorer la méthode de quantification des descripteurs par rapport aux mots du dictionnaire.

Références

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition : A survey," *IEEE Transactions on Robotics*, 2016.

[2] P. Neubert, N. Sunderhauf, and P. Protzel, "Appearance change prediction for long-term navigation across seasons," in *European Conference on Mobile Robots (ECMR)*, 2013.

[3] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *AAAI Conference on Artificial Intelligence*, 2014.

[4] C. McManus, W. Churchill, W. Maddern, A. Stewart, and P. Newman, "Shady dealings : Robust, long- term visual localisation using illumination invariance," in *International Conference on Robotics and Automation (ICRA)*, 2014.

[5] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping : a survey," *Artificial Intelligence Review*, 2015.

[6] T. Weyand, I. Kostrikov, and J. Philbin, "Planet-photo geolocation with convolutional neural networks," *arXiv preprint :1602.05314*, 2016.

[7] C. A. Aguilera, F. J. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Learning cross-spectral similarity measures with deep convolutional neural networks," in *Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[8] D. Firmenichy, M. Brown, and S. Susstrunk, "Multispectral interest points for rgb-nir image registration," in *International Conference on Image Processing (ICIP)*, 2011.

[9] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[10] J. Sivic and A. Zisserman, "Video google : A text retrieval approach to object matching in videos," in *International Conference on Computer Vision (ICCV)*, 2003.