Estimation dense de profondeur combinant approches variationnelles et observateurs asymptotiques

Nadège Zarrouati[1,2] Emanuel Aldea[3] Pierre Rouchon[1]

[1] Mines-ParisTech, Centre Automatique et Systèmes, Unité Mathématiques et Systèmes
60, boulevard Saint-Michel, 75272 Paris Cedex, France
[2] DGA, 7-9 rue des Mathurins, 92220 Bagneux, France
[3] SYSNAV, 57 rue de Montigny, 27200 Vernon, France

nadege.zarrouati | pierre.rouchon@mines-paristech.fr, emanuel.aldea@sysnav.fr

Résumé

Cet article propose une nouvelle approche pour estimer en temps réel la carte de profondeur instantanée à partir de données inertielles et images fournies par une caméra en mouvement libre dans une scène statique. Une fonction coût invariante par rotation est introduite. Sa minimisation conduit à une estimation de la carte de profondeur, solution d'une équation de diffusion sur la sphère Riemannienne de l'espace à trois dimensions. Transcrite en coordonnées pinhole, cette équation est résolue numériquement et donne une première estimation de la carte de profondeur sur la totalité du champ couvert par la caméra. Un observateur asymptotique reposant sur un modèle géométrique de l'évolution du champ de profondeur à partir des données inertielles, permet ensuite d'affiner continûment cette estimation. Cette approche diffère notablement de la plupart des méthodes actuelles qui estiment la carte de profondeur à partir de plusieurs vues stéréo combinées avec des expansions de régions, ou encore des stratégies probabilistes d'affinement incrémental. Des analyses quantitatives des estimations obtenues sur des données de synthèse illustrent l'intérêt de la méthode proposée. De premiers résultats sur données réelles confirment les simulations sur données synthétiques.

Mots Clef

Carte de profondeur, flot optique, observateur asymptotique, invariance par le groupe des rotations.

Abstract

In this paper, we describe a novel approach for estimating and refining the instantaneous depth map for a camera moving freely in a static scene. We propose a $SO(3)$-invariant cost function minimized by the depth map which provides, at each time step, a diffusion equation on the unit Riemannian sphere. Written in pinhole coordinates, this scalar diffusion equation is numerically solved to provide a depth map estimation of the entire field of view. We then propose a rigurous method based on asymptotic observers for continuously refining the depth map. This approach is fundamentally different from most methods currently available that provide a depth map estimate by multiview stereo with region growing, or by probabilistic strategies of optical flow incremental refinement. Quantitative estimations on synthetic data of the asymptotic observers merging optical flow and camera motion illustrate the performance of the proposed method. We also provide an example of qualitative results on a real sequence of images.

Keywords

Depth map, optical flow, asymptotic observer, $SO(3)$-invariance

1 Introduction

Depuis quelques années, un intérêt grandissant a été porté à l'élaboration de cartes de profondeur. En effet, les efforts fournis (e.g. en vision par ordinateur, robotique, photogrammétrie) ont permis des progrès importants vers la reconstruction 3D à partir de nuages de points. Dans le cadre d'applications à la robotique par exemple, la reconstruction d'environnement est indissociable de la méthode de SLAM, que l'on traite généralement par filtrage nonlinéaire de positions de points d'intérêt détectés dans les images (e.g. [1, 2]), ou par ajustement de faisceaux ([3]). L'estimation d'un nuage de points dispersés permet alors la localisation du robot. Elle reste souvent limitée à des zones localisées et ne couvre pas en général tout le champ couvert par la caméra. Passer d'une distribution discrète de positions 3D à une estimation d'une carte de profondeur continue de l'environnement reste une question ouverte comme en témoignent des études récentes : par exemple [4] pour une approche par expansion de régions ; [5] pour une méthode exploitant le flot optique entre image-clés ; ou encore [6, 7] pour des approches probabilistes.

Pour traiter le problème d'estimation de la carte de profondeur en se passant du traitement MVS des images, une

approche moins usitée consiste à estimer le flot optique associé à deux images dont les poses sont très proches, comme dans le cas de séquences video. Une telle analyse dans cet esprit permet de contourner les problèmes typiquement liés au MVS, comme les occultations, les incertitudes liées à l'appariement de points ou les variations de perspective. En contrepartie, cette approche nécessite un affinement temporel incrémental des mesures brutes de flot optique. Parmi les approches probabilistes, nous soulignons [8] qui propose un schéma d'estimation de profondeur à l'échelle du pixel qui détecte des maxima locaux de la corrélation le long des lignes épipolaires.

Une autre manière d'utiliser le flot optique dans des systèmes dynamiques est présentée dans l'article-référence [9] où les auteurs appliquent un filtre de Kalman à l'échelle du pixel pour affiner incrémentalement la carte de profondeur dans la totalité du champ de vue de la caméra. Tout comme les approches probabilistes, les systèmes dynamiques sont parfaitement adaptés à l'estimation incrémentale de la profondeur à partir de la connaissance du mouvement de la caméra et d'informations visuelles : en effet, l'état courant du système est entièrement déterminé par la condition initiale et sa dynamique ; on évite ainsi tout recours au traitement en bloc des images. Cependant, il a été montré [10] que pour des systèmes projectifs, toute linéarisation détériore l'estimation de la géométrie sous-jacente. Construire un observateur non-linéaire pour estimer la profondeur de points d'intérêt *uniques ou isolés* a été longuement étudié, spécifiquement dans le cas où le mouvement de la caméra est connu ([11, 12, 13]).

Nous proposons une nouvelle méthode, reposant sur un système d'équations aux dérivées partielles qui décrivent les dynamiques invariantes par le groupe des rotations $SO(3)$ de l'intensité lumineuse perçue par la caméra couplée à la carte de profondeur de l'environnement. A partir de ce modèle cinématique et de la connaissance du mouvement de la caméra, nous proposons une méthode variationnelle inspirée de l'algorithme de Horn-Schunck. Cette méthode préserve l'invariance par rotation et fournit à chaque instant une estimation de la carte de profondeur. Si l'on compare avec [8], cela nous permet de prendre en compte tout mouvement de la caméra tout en imposant des contraintes de régularité, mais pas d'épipolarité explicites. Cette approche diffère fondamentalement des méthodes précédemment citées qui construisent des observateurs asymptotiques pour l'estimation de la profondeur de points isolés. Notre méthode permet d'obtenir des estimations de profondeur dans la totalité du champ de vue de la caméra, sans contrainte particulière sur le mouvement de celle-ci.

Cet article s'articule de la manière suivante. Les équations différentielles régissant la dynamique de l'intensité lumineuse et de la profondeur sont explicitées dans la Section 2, ainsi que leur formulation en coordonnées pinhole. En Section 3, nous exposons la méthode variationnelle invariante pour estimer la profondeur. En Section 4, nous proposons deux observateurs asymptotiques : le premier est utilisé pour filtrer des données brutes de flot optique, et le second des cartes de profondeur issues de la méthode variationnelle précédemment mentionnée. Dans le cadre d'hypothèses physiquement plausibles concernant le mouvement de la caméra et l'environnement, nous prouvons la convergence de l'estimation fournie par ces observateurs. Dans la Section 5, nous appliquons ces méthodes sur des images de synthèse et nous analysons en simulation leur précision, leur robustesse au bruit et leur vitesse de convergence. Enfin, une application à des données réelles est traitée succinctement.

2 Le modèle invariant par $SO(3)$

2.1 Le système d'EDP sur $\mathbb{S}^2$

Le modèle s'appuie sur les hypothèses géométriques introduites dans [14]. On considère une caméra sphérique, dont le déplacement est connu. Les vitesses linéaires et angulaires $v(t)$ et $\omega(t)$ sont exprimées dans le repère caméra. La position du centre optique dans le référentiel externe $\mathcal{R}$ est indiquée par $C(t)$. L'orientation par rapport à $\mathcal{R}$ est donnée par le quaternion $q(t)$: un vecteur ς dans le repère caméra correspond à un vecteur $q\varsigma q^*$ dans le référentiel externe $\mathcal{R}$ en utilisant la représentation des vecteurs comme des quaternions purement imaginaires. On obtient : $\frac{d}{dt}q = \frac{1}{2}q\omega$. Chaque pixel est associé à un vecteur η dans le repère caméra : η appartient à la sphère $\mathbb{S}^2$ et reçoit la luminosité $y(t,\eta)$. Par conséquent, à chaque instant l'image produite par la caméra est décrite par le champ scalaire $\mathbb{S}^2 \ni \eta \mapsto y(t,\eta) \in \mathbb{R}$.
La scène est représentée comme une surface fermée, C^1 et convexe Σ dans $\mathbb{R}^3$, difféomorphe à $\mathbb{S}^2$. La caméra se trouve à l'intérieur du domaine $\Omega \subset \mathbb{R}^3$ délimité par $\Sigma = \partial\Omega$. A un point $M \in \Sigma$ on associe un et un seul pixel de la caméra : si les points de Σ sont représentés par $s \in \mathbb{S}^2$, à tout instant t une transformation continue et inversible ϕ nous permet d'exprimer le couple de variables (t,η) comme une fonction de (t,s) : $(t,\eta) = (t,\phi(t,s))$.
L'intensité lumineuse émise par un point $M(s) \in \Sigma$ ne dépend pas de la direction d'émission (Σ est une surface Lambertienne) et est indépendante de t (la scène est statique). Ceci implique que $y(t,\eta)$ ne dépend que de s : ainsi l'intensité y est soit une fonction de (t,η), soit, par la transformation ϕ, une fonction de s. La profondeur $C(t)M(s)$ entre le centre optique et l'objet visualisé dans la direction $\eta = \phi(t,s)$ est notée par $D(t,\eta)$, et son inverse par $\Gamma = 1/D$. Fig.1 rappelle le modèle et les notations. On suppose que $s \mapsto y(s)$ est une fonction C^1. Pour tout t, $s \mapsto D(t,s)$ est C^1 car Σ est une surface C^1 de $\mathbb{R}^3$.
Sous ces hypothèses, on a d'une part :

$$\left.\frac{\partial y}{\partial t}\right|_s = 0, \quad \left.\frac{\partial \Gamma}{\partial t}\right|_s = \Gamma^2 v \cdot \eta \tag{1}$$

et d'autre part, par composition de la différentiation :

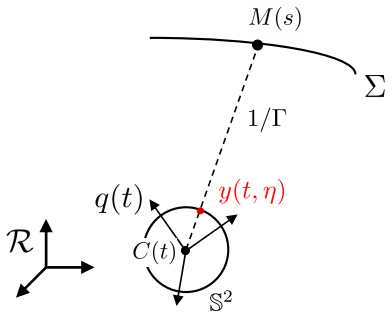

FIGURE 1 – Modèle et notations pour une caméra sphérique dans un environnement statique.

$$\left.\frac{\partial h}{\partial t}\right|_s = \left.\frac{\partial h}{\partial t}\right|_\eta + \left.\frac{\partial h}{\partial \eta}\right|_t \left.\frac{\partial \eta}{\partial t}\right|_s = \left.\frac{\partial h}{\partial t}\right|_\eta + \nabla h \cdot \left.\frac{\partial \eta}{\partial t}\right|_s \quad (2)$$

où h est un champ scalaire défini sur $\mathbb{S}^2$ et ∇h son gradient par rapport à la métrique Riemannienne sur $\mathbb{S}^2$. La valeur de ∇h dans $\eta \in \mathbb{S}^2$ est associée à un vecteur de $\mathbb{R}^3$ tangent à la sphère au point η, lui-même identifié à un vecteur unitaire de $\mathbb{R}^3$ dans le repère caméra. On note $a \cdot b$ le produit scalaire des deux vecteurs a et b de $\mathbb{R}^3$, et $a \times b$ leur produit vectoriel. Par différentiation, l'identité $q\eta q^* = \frac{\overrightarrow{C(t)M(s)}}{||C(t)M(s)||}$, où * représente le conjugué et η est assimilé à un quaternion imaginaire, donne

$$\left.\frac{\partial \eta}{\partial t}\right|_s = \eta \times (\omega + \Gamma \eta \times v) \quad (3)$$

car le vecteur $\eta \times \omega$ correspond au quaternion imaginaire $(\omega\eta - \eta\omega)/2$. Ainsi, en appliquant (2) aux champs scalaires $y(t,\eta)$ et $\Gamma(t,\eta)$, et en identifiant grâce à (1), l'intensité et la profondeur inverse satisfont les équations suivantes :

$$\frac{\partial y}{\partial t} = -\nabla y \cdot (\eta \times (\omega + \Gamma \eta \times v)) \quad (4)$$

$$\frac{\partial \Gamma}{\partial t} = -\nabla \Gamma \cdot (\eta \times (\omega + \Gamma \eta \times v)) + \Gamma^2 v \cdot \eta \quad (5)$$

Les équations (4) et (5) sont invariantes par $SO(3)$: elles restent inchangées par toute rotation déterminée par le quaternion σ qui transforme (η, ω, v) en $(\sigma\eta\sigma^*, \sigma\omega\sigma^*, \sigma v \sigma^*)$. L'équation (4) est l'équation de flot optique qui peut être retrouvée sous de nombreuses formes dans la littérature (voir [15] ou [13] par exemple), alors que (5) est moins standard (voir [14]).

2.2 Le système en coordonnées pinhole

Afin d'appliquer le modèle 2.1 aux données visuelles, on doit exprimer les équations invariantes (4) et (5) dans les coordonnées locales de $\mathbb{S}^2$ qui correspondent à une grille de pixels rectangulaire. Une solution couramment utilisée est de faire appel au modèle pinhole (de type projection centrale), où le pixel de coordonnées (z_1, z_2) correspond au vecteur unitaire $\eta \in \mathbb{S}^2$ exprimé dans $\mathbb{R}^3$ comme : $\left(1 + z_1^2 + z_2^2\right)^{-1/2} (z_1, z_2, 1)^T$. A noter le fait que l'axe optique de la caméra (le pixel $(z_1, z_2) = (0,0)$) correspond ici à la direction z_3. Les directions 1 et 2 correspondent respectivement à l'axe horizontal (de gauche à droite) et à l'axe vertical (du haut vers le bas) de l'image.
Les gradients ∇y et $\nabla \Gamma$ doivent etre exprimés par rapport à z_1 et z_2. Développons ce calcul pour y. Tout d'abord, ∇y est tangent à $\mathbb{S}^2$, donc $\nabla y \cdot \eta = 0$. Deuxièmement, la différentielle dy correspond à $\nabla y \cdot d\eta$ d'une part et à $\frac{\partial y}{\partial z_1} dz_1 + \frac{\partial y}{\partial z_2} dz_2$ d'autre part. Par identification, on obtient les coordonnées de ∇y en $\mathbb{R}^3$. De manière similaire, on retrouve les trois coordonnées de $\nabla \Gamma$. Par l'injection de ces expressions en (4) et (5), on obtient l'EDP suivante qui correspond à (4) et (5) en coordonnées pinhole :

$$\begin{aligned}\frac{\partial y}{\partial t} = &- \frac{\partial y}{\partial z_1} \begin{bmatrix} z_1 z_2 \omega_1 - (1 + {z_1}^2)\omega_2 + z_2 \omega_3 \\ +\Gamma\sqrt{1 + z_1^2 + z_2^2}(-v_1 + z_1 v_3) \end{bmatrix} \\ &- \frac{\partial y}{\partial z_2} \begin{bmatrix} (1 + {z_2}^2)\omega_1 - z_1 z_2 \omega_2 - z_1 \omega_3 \\ +\Gamma\sqrt{1 + z_1^2 + z_2^2}(-v_2 + z_2 v_3) \end{bmatrix}\end{aligned} \quad (6)$$

$$\begin{aligned}\frac{\partial \Gamma}{\partial t} = &- \frac{\partial \Gamma}{\partial z_1} \begin{bmatrix} z_1 z_2 \omega_1 - (1 + {z_1}^2)\omega_2 + z_2 \omega_3 \\ +\Gamma\sqrt{1 + z_1^2 + z_2^2}(-v_1 + z_1 v_3) \end{bmatrix} \\ &- \frac{\partial \Gamma}{\partial z_2} \begin{bmatrix} (1 + {z_2}^2)\omega_1 - z_1 z_2 \omega_2 - z_1 \omega_3 \\ +\Gamma\sqrt{1 + z_1^2 + z_2^2}(-v_2 + z_2 v_3) \end{bmatrix} \\ &+ \Gamma^2(z_1 v_1 + z_2 v_2 + v_3)\end{aligned} \quad (7)$$

où v_1, v_2, v_3, ω_1, ω_2, ω_3 sont les composants des vitesses linéaires et angulaires dans le repère caméra.

3 Une méthode variationnelle pour l'estimation directe de profondeur

3.1 La méthode Horn-Schunk

Horn et Schunck ont décrit en [15] un algorithme pour le calcul du flot optique, défini comme "la distribution des vitesses de déplacement apparentes des patterns lumineux dans une image". Leur proposition s'appuie sur la contrainte de flot optique écrite sous forme compacte :

$$\frac{\partial y}{\partial t} + V_1 \frac{\partial y}{\partial z_1} + V_2 \frac{\partial y}{\partial z_2} = 0 \quad (8)$$

L'identification avec (6) donne

$$V_i(t,z) = f_i(z, \omega(t)) + \Gamma(t,z) g_i(z, v(t)),\ i \in \{1,2\} \quad (9)$$

avec

$$\begin{aligned} f_1(z,\omega) &= z_1 z_2 \omega_1 - (1 + {z_1}^2)\omega_2 + z_2 \omega_3 \\ f_2(z,\omega) &= (1 + {z_2}^2)\omega_1 - z_1 z_2 \omega_2 - z_1 \omega_3 \\ g_i(z,v) &= \sqrt{1 + z_1^2 + z_2^2}(-v_i + z_i v_3),\ i \in \{1,2\}. \end{aligned} \quad (10)$$

A tout instant t, le champ de vitesses apparent $V = (V_1\ \ V_2)^T$ est alors estimé en minimisant par rapport à

$W = (W_1\ W_2)^T$ la fonction coût suivante (l'image $\mathcal{I}$ est ici un rectangle de $\mathbb{R}^2$) :

$$I(W) = \iint_{\mathcal{I}} \left(\left(\frac{\partial y}{\partial t} + W^T \nabla y\right)^2 + \alpha^2 |\nabla W|^2 \right) dz_1 dz_2 \tag{11}$$

où $\alpha > 0$ est un paramètre de régularisation. Notons $V_{\text{HS}}(t,z) = (V_{\text{HS}1}(t,z)\ V_{\text{HS}2}(t,z))^T$ l'estimation de V à l'instant t par la résolution de (11).

3.2 Adaptation à l'estimation directe de profondeur

Au lieu de minimiser le coût I donné par (11) par rapport à W_1 et W_2, définissons un nouveaut coût invariant J,

$$J(\Upsilon) = \iint_{\mathcal{J}} \left(\left(\frac{\partial y}{\partial t} + \nabla y \cdot (\eta \times (\omega + \Upsilon \eta \times v))\right)^2 + \alpha^2 \nabla \Upsilon^2 \right) d\sigma_\eta \tag{12}$$

qu'on minimise par rapport au profil de profondeur $\mathcal{J} \ni \eta \mapsto \Upsilon(t,\eta) \in \mathbb{R}$. Le temps t est ici fixé et $d\sigma_\eta$ est l'élément surfacique infinitésimal sur $\mathbb{S}^2$. $\mathcal{J} \subset \mathbb{S}^2$ est le domaine où y est mesurée et $\alpha > 0$ est le paramètre de régularisation.

La condition de stationnarité de premier ordre pour J et pour toute variation de Υ nous fournit l'EDP suivante caractérisant l'estimation Γ_{HS} du vrai profil Γ :

$$\alpha^2 \Delta \Gamma_{\text{HS}} = \left(\frac{\partial y}{\partial t} + \nabla y \cdot (\eta \times (\omega + \Gamma_{\text{HS}} \eta \times v))\right) \ldots \ldots (\nabla y \cdot (\eta \times (\eta \times v))) \quad \text{sur } \mathcal{J} \tag{13}$$

avec $\frac{\partial \Gamma_{\text{HS}}}{\partial n} = 0$ sur $\partial\mathcal{J}$, où $\Delta\Gamma_{\text{HS}}$ est le Laplacien Γ_{HS} sur la sphère Riemannienne $\mathbb{S}^2$ et $\partial\mathcal{J}$ est la frontière de $\mathcal{J}$, supposée continue par morceaux et de normale n.

En coordonnées pinhole (z_1, z_2), on a

$$d\sigma_\eta = \left(1 + z_1^2 + z_2^2\right)^{-3/2} dz_1 dz_2$$

$$\nabla \Upsilon^2 = \left(1 + z_1^2 + z_2^2\right) \left(\frac{\partial \Upsilon}{\partial z_1}^2 + \frac{\partial \Upsilon}{\partial z_2}^2 + (z_1 \frac{\partial \Upsilon}{\partial z_1} + z_2 \frac{\partial \Upsilon}{\partial z_2})^2 \right)$$

$$\left(\frac{\partial y}{\partial t} + \nabla y \cdot (\eta \times (\omega + \Upsilon \eta \times v))\right)^2 = (F + \Upsilon G)^2$$

où

$$\begin{aligned} F &= \frac{\partial y}{\partial t} + f_1(z,\omega) \frac{\partial y}{\partial z_1} + f_2(z,\omega) \frac{\partial y}{\partial z_2} \\ G &= g_1(z,v) \frac{\partial y}{\partial z_1} + g_2(z,v) \frac{\partial y}{\partial z_2}. \end{aligned} \tag{14}$$

Par conséquent, la condition de stationnarité au premier ordre (13) devient en fonction de (z_1, z_2) :

$$\begin{aligned} \Gamma_{\text{HS}} G^2 + FG = \alpha^2 \Bigg[& \frac{\partial}{z_1} \left(\frac{1 + z_1^2}{\sqrt{1 + z_1^2 + z_2^2}} \frac{\partial \Gamma_{\text{HS}}}{\partial z_1} \right) \\ & + \frac{\partial}{z_2} \left(\frac{1 + z_2^2}{\sqrt{1 + z_1^2 + z_2^2}} \frac{\partial \Gamma_{\text{HS}}}{\partial z_2} \right) + \frac{\partial}{z_2} \left(\frac{z_1 z_2}{\sqrt{1 + z_1^2 + z_2^2}} \frac{\partial \Gamma_{\text{HS}}}{\partial z_1} \right) \\ & + \frac{\partial}{z_1} \left(\frac{z_1 z_2}{\sqrt{1 + z_1^2 + z_2^2}} \frac{\partial \Gamma_{\text{HS}}}{\partial z_2} \right) \Bigg] \end{aligned} \tag{15}$$

Le terme droit de (15) correspond au Laplacien sur la sphère Riemannienne $\mathbb{S}^2$, écrit en coordonnées pinhole. La résolution numérique de cette diffusion scalaire qui fournit l'estimation Γ_{HS} de Γ est similaire à celle qu'on utilise pour l'estimation Horn-Schunck V_{HS} de V.

La fonctionnelle $I(W)$ definie en (11) est minimisée par rapport à W, alors que pour une scène statique il n'y a qu'une seule variable implicite : la carte de profondeur inverse Γ. Pour un η donné, les contraintes géométriques lient W à une ligne épipolaire spécifique. Les approches existantes [16, 17] prennent en compte la contrainte épipolaire par l'inclusion de termes additionnels d'attache aux données qui sont liés à la matrice fondamentale. En revanche, la fonctionnelle $J(\Upsilon)$ s'appuie directement sur les informations de dynamique de la caméra et sur son invariance par $SO(3)$, en ramenant la paramétrisation à Υ seulement.

Alternativement, il est possible d'adapter de manière immédiate cette approche pour employer un terme de régularisation en norme L^1 à la place de la régularisation Tikhonov explicitée dans cette section.

4 Affinement de la profondeur par observateurs asymptotiques

4.1 Observateur asymptotique à base de flot optique

A partir d'une estimation de flot optique, comme V_{HS} par exemple, il est raisonnable de supposer qu'on a accès à tout instant t aux composantes en coordonnées pinhole du champ de vecteurs

$$\varpi_t : \mathbb{S}^2 \ni \eta \mapsto \varpi_t(\eta) = \eta \times (\omega + \Gamma \eta \times v) \in T_\eta \mathbb{S}^2 \tag{16}$$

figurant en (5). Ce champ de vecteurs peut être considéré comme une entrée pour (5), exprimé comme $\varpi_t(\eta) = f_t(\eta) + \Gamma(t,\eta) g_t(\eta)$, où f_t et g_t sont les champs de vecteurs

$$f_t : \mathbb{S}^2 \ni \eta \mapsto f_t(\eta) = \eta \times \omega \in T_\eta \mathbb{S}^2 \tag{17}$$

$$g_t : \mathbb{S}^2 \ni \eta \mapsto g_t(\eta) = \eta \times (\eta \times v) \in T_\eta \mathbb{S}^2. \tag{18}$$

Cela nous permet de proposer l'observateur asymptotique suivant pour $D = 1/\Gamma$: à partir des estimations comme V_{HS}, assimilées à une mesure d'entrée ϖ, et à partir de la dynamique connue de la caméra retenue en f_t et g_t, (19)

fournit une estimation $\widehat{D}$ de la carte de profondeur :

$$\frac{\partial \widehat{D}}{\partial t} = -\nabla \widehat{D} \cdot \varpi_t - v \cdot \eta + k g_t \cdot (\widehat{D} f_t + g_t - \widehat{D}\varpi_t) \quad (19)$$

où ϖ_t, f_t et g_t sont des champs de vecteurs connus définis sur $\mathbb{S}^2$, et $k > 0$ est un gain à régler. En coordonnées pinhole, cet observateur invariant par $SO(3)$ devient

$$\begin{aligned}\frac{\partial \widehat{D}}{\partial t} &= -\frac{\partial \widehat{D}}{\partial z_1} V_1 - \frac{\partial \widehat{D}}{\partial z_2} V_2 - (z_1 v_1 + z_2 v_2 + v_3) \quad (20)\\ &+ k\big(g_1(\widehat{D} f_1 + g_1 - \widehat{D} V_1) + g_2(\widehat{D} f_2 + g_2 - \widehat{D} V_2)\big)\end{aligned}$$

où V est fourni par une estimation plus ou moins précise du flot optique et (f_1, f_2, g_1, g_2) sont définis par (10).
En accord avec l'hypothèse de 2.1, à tout instant t il y a une correspondance bijective et lisse entre un $\eta \in \mathbb{S}^2$ associé à un pixel de la caméra et le point de la scène $M(s)$ correspondant à ce pixel. Par conséquent, pour tout $t \geq 0$, le flot $\phi(t, s)$ défini par

$$\left.\frac{\partial \phi}{\partial t}\right|_{(t,s)} = \varpi_t(\phi(t, s)), \quad \phi(0, s) = s \in \mathbb{S}^2 \quad (21)$$

détermine un difféomorphisme sur $\mathbb{S}^2$ dépendant du temps. Notons par ϕ^{-1} le difféomorphisme inverse : $\phi(t, \phi^{-1}(t, \eta)) \equiv \eta$. Supposons que $\Gamma(t, \eta) > 0$, $v(t)$ et $\omega(t)$ sont bornés uniformément $t \geq 0$ et $\eta \in \mathbb{S}^2$. Ainsi, la trajectoire du centre de la caméra $C(t)$ reste strictement à l'intérieur de la surface convexe Σ, à une distance $\min_{(t,\eta)} D(t, \eta)$. Ces observations justifient les suppositions utilisées dans le théorème suivant [1].

Théorème 1 *On considère $\Gamma(t, \eta)$ associé au déplacement de la caméra à l'intérieur du domaine Ω délimité par la scène Σ (une surface fermée, C^1 et convexe comme présentée en 2.1). Supposons qu'il existe $\bar{v} > 0$, $\bar{\omega} > 0$, $\bar{\gamma} > 0$ et $\bar{\Gamma} > 0$ tels que*

$$\forall t \geq 0,\ \forall \eta \in \mathbb{S}^2,\ |v(t)| \leq \bar{v},\ |\omega(t)| \leq \bar{\omega},\ \bar{\gamma} \leq \Gamma(t, \eta) \leq \bar{\Gamma}.$$

Alors, pour $t \geq 0$, $\Gamma(t, \eta)$ est une solution C^1 de (5)*. Considérons l'observateur* (19) *avec une condition initiale C^1 par rapport à η, $\widehat{D}(0, \eta)$. Nous avons les implications suivantes :*

- *$\forall t \geq 0$, la solution $\widehat{D}(t, \eta)$ de* (19) *existe, est unique et reste C^1 par rapport à η. De plus*

$$t \mapsto \|\widehat{D}(t, .) - D(t, .)\|_{L^\infty} = \max_{\eta \in \mathbb{S}^2} \left|\widehat{D}(t, \eta) - D(t, \eta)\right|$$

  est décroissante (stabilité L^∞).
- *si en plus pour tout $s \in \mathbb{S}^2$, $\int_0^{+\infty} \|g_\tau(\phi(\tau, s))\|^2 d\tau = +\infty$, on a pour tout $p > 0$,*

$$\lim_{t \mapsto +\infty} \int_{\mathbb{S}^2} \left|\widehat{D}(t, \eta) - D(t, \eta)\right|^p d\sigma_\eta = 0$$

  (convergence dans toute topologie L^p)
- *si en plus il existe $\lambda > 0$ et $T > 0$ tels que, pour tout $t \geq T$ et $s \in \mathbb{S}^2$, $\int_0^t \|g_\tau(\phi(\tau, s))\|^2 d\tau \geq \lambda t$, on a, pour tout $t \geq T$,*

$$\|\widehat{D}(t, .) - D(t, .)\|_{L^\infty} \leq e^{-k\bar{\gamma}\lambda t} \|\widehat{D}(0, .) - D(0, .)\|_{L^\infty}$$

  (convergence exponentielle dans L^∞).

Les hypothèses concernant $\int \|g_t(\phi(t, s))\|^2 dt$ peuvent être interprétées comme une condition d'excitation persistante. Elles devraient être satisfaites pour un mouvement générique de la caméra.

4.2 Observateur asymptotique pour les cartes brutes de profondeur

Au lieu de filtrer V_{HS}, on peut construire un observateur filtrant une estimation imprécise de Γ_{HS}. Dans ce cas, (19) devient ($k > 0$)

$$\frac{\partial \widehat{D}}{\partial t} = -\nabla \widehat{D} \cdot (f_t + \Gamma_{\text{HS}} g_t) - v \cdot \eta + k(1 - \widehat{D}\Gamma_{\text{HS}}) \quad (22)$$

qui s'exprime en coordonnées pinhole :

$$\begin{aligned}\frac{\partial \widehat{D}}{\partial t} = &-\frac{\partial \widehat{D}}{\partial z_1}(f_1 + \Gamma_{\text{HS}} g_1) - \frac{\partial \widehat{D}}{\partial z_2}(f_2 + \Gamma_{\text{HS}} g_2)\\ &- (z_1 v_1 + z_2 v_2 + v_3) + k(1 - \widehat{D}\Gamma_{\text{HS}}). \quad (23)\end{aligned}$$

Pour cet observateur, nous avons le résultat de convergence suivant.

Théorème 2 *Considérons les hypothèses du théorème 1 concernant la surface de la scène Σ, $\Gamma = 1/D$, v and ω. On considère l'observateur* (22) *où Γ_{HS} correspond à Γ et où la condition initiale est C^1 par rapport à η. Alors $\forall t \geq 0$, la solution $\widehat{D}(t, \eta)$ de* (22) *existe, est unique, reste C^1 par rapport à η et*

$$\|\widehat{D}(t, .) - D(t, .)\|_{L^\infty} \leq e^{-k\bar{\gamma}t} \|\widehat{D}(0, .) - D(0, .)\|_{L^\infty}$$

(convergence exponentielle dans L^∞)

5 Résultats expérimentaux

Afin d'estimer les performances des méthodes proposées, on définit un taux d'erreur global associé à une estimation de D comme

$$E = \int_{\mathcal{I}} \frac{|\widehat{D}(t, \eta) - D(t, \eta)|}{D(t, \eta)} d\sigma_\eta \quad (24)$$

où D est la vraie valeur de la carte de profondeur, $\widehat{D}$ est l'estimation par une des méthodes proposées et $\mathcal{I}$ représente l'espace image.

5.1 Descriptif des données de synthèse

Les observateurs non-linéaires asymptotiques introduits dans la section 4 sont testés sur une séquence d'images de synthèse caractérisée par :

1. Les contraintes de place nous empêchent de présenter ici la preuve formelle.

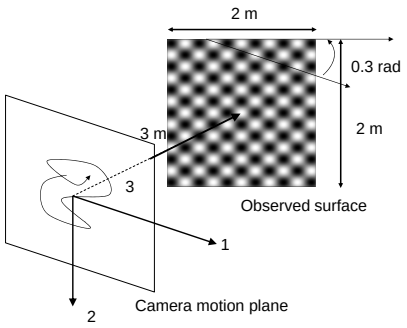

FIGURE 2 – Le cadre virtuel utilisé pour générer les images de synthèse traitées en 5

– *caméra virtuelle* : la résolution des images est 640 par 480 pixels, la fréquence est 60 Hz et le champ de vue est de 50 deg. par 40 deg. ;
– *dynamique de la caméra* : il s'agit de deux translations superposées dans le plan vertical ($v_3 = \omega_1 = \omega_2 = \omega_3 = 0$), dont le profil de vitesse est sinusoïdal, d'amplitude 1 $m.s^{-1}$, et pulsations différentes (π pour v_1 et 3π pour v_2) ;
– *scène virtuelle* : un plan texturé par un motif sinusoïdal placé à 3 m et incliné de 0.3 rad par rapport au plan de mouvement de la caméra
– *synthèse des images* : chaque pixel code une valeur entière comprise entre 0 et 255, qui dépend de l'intensité de la surface observée dans la direction indexée par le pixel, à laquelle on rajoute un bruit Gaussien $\mathcal{N}(0, \sigma^2)$

Le cadre virtuel utilisé pour générer la séquence d'image est représenté en Fig.2.

5.2 Estimation de profondeur à partir du flot optique brut

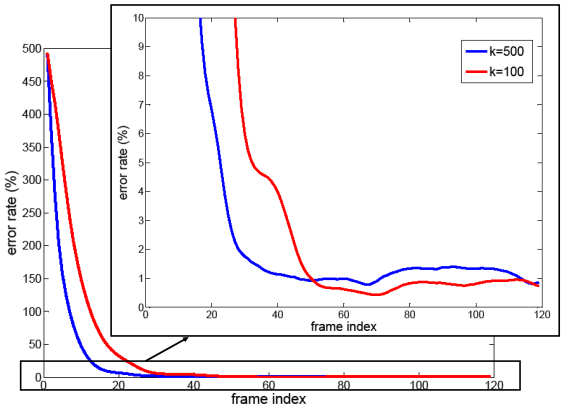

FIGURE 3 – Erreurs constatées pour les cartes de profondeur D estimées par l'observateur asymptotique (20) appliqué au flot optique brut V_{HS} pour differents gains k. Le bruit est fixé à $\mathcal{N}(0, 1)$.

On teste sur la séquence 5.1 l'estimation de profondeur décrite par l'observateur (20) introduit en 4.1. Le flot brut V_{HS} est estimé par la méthode standard Horn-Schunck. Des conditions de type Neumann sont choisies au bord où le flot est dirigé vers l'intérieur de l'image : $\frac{\partial \widehat{D}}{\partial n} = 0$ si $n \cdot V_{\text{HS}} < 0$. Le gain k de l'observateur est choisi par des considérations de mise à l'échelle : $k = 500\ s.m^{-2}$ assure une convergence rapide (3 ou 4 iterations sont suffisantes) mais $k = 20$ ou $30\ s.m^{-2}$ sont des valeurs plus raisonnables si on traite des données bruitées (environ 50 itérations nécessaires).
L'écart-type σ du bruit rajouté aux images de synthèse est fixé à $\sigma = 1$. Les gains $k = 500$ et $k = 100$ sont testés, et les erreurs associées pour $\widehat{D}$ sont présentées en Fig. 3. Les deux erreurs diminuent progressivement : comme prévu, la convergence est plus rapide lorsque le gain est plus grand, et plus stable pour un gain plus petit. Dans les deux cas, le taux d'erreur reste en dessous de 1.5% au bout de 40 images. Pour valider la robustesse de l'algorithme par rapport

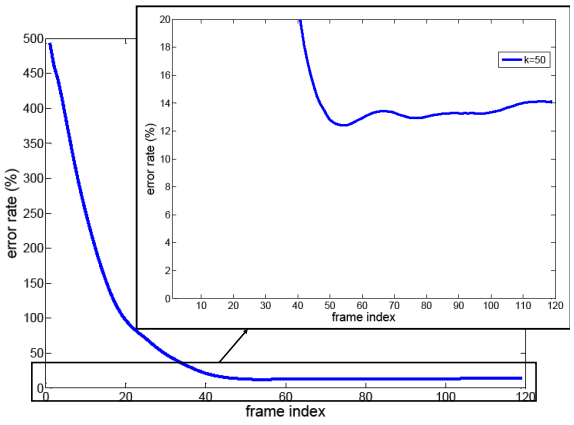

FIGURE 4 – Erreur constatée pour la carte de profondeur D estimée par l'observateur asymptotique (20) filtrant le flot optique brut V_{HS}. Le bruit de l'image est fixé à $\mathcal{N}(0, 20^2)$.

au bruit, l'écart-type σ est multiplié par 20, et le gain est fixé à $k = 50\ s.m^{-2}$. Le taux d'erreur augmente et se stabilise (au bout d'environ 40 images traitées) entre 12 et 14 %. A noter le fait que l'erreur ne peut pas diminuer, car le niveau élevé du bruit détériore la mesure V_{HS} de l'observateur (par rapport à sa vraie valeur V, le taux d'erreur en norme de V_{HS} est d'environ 15%). Les résultats soulignent le fait que cette approche est sensible au bruit qui agit sur les données visuelles par le biais des erreurs présentes dans l'estimation du flot optique.

5.3 Estimation directe de profondeur

Par la suite, l'observateur décrit en (23) a été appliqué à la même séquence. L'entrée Γ_{HS} de l'observateur est obtenue en sortie de (15). Pour adapter le schéma numérique à ce modèle, on fait une approximation des petits angles en négligeant les termes de second ordre en z (on néglige la courbure de $\mathbb{S}^2$ et on considère que l'image correspondant à une petite partie de $\mathbb{S}^2$ peut être approchée par un petit rectangle

Euclidien ; l'erreur de cette approximation est inférieure à 3% pour les paramètres de cette séquence) : (15) devient

$$G^2\Gamma + FG = \alpha^2 \left(\frac{\partial^2 \Gamma}{\partial {z_1}^2} + \frac{\partial^2 \Gamma}{\partial {z_2}^2} \right) \qquad (25)$$

La carte de profondeur inverse Γ est estimée par un schéma

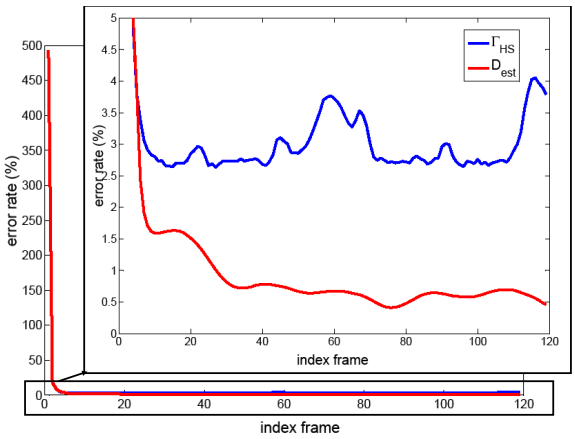

FIGURE 5 – Erreur constatée pour l'estimation directe de $1/\Gamma_{HS}$ (bleu), décrite en 3.2, et de D (rouge) par l'observateur asymptotique (23) filtrant Γ_{HS}. Le bruit est $\mathcal{N}(0,1)$.

itératif, avec une initialisation fournie par l'estimation antérieure, grâce à notre approche invariante. Le paramètre de régularisation α est choisi en prenant en compte l'amplitude du bruit attendu dans les dérivées spatio-temporelles : $\alpha = 300\ m.s^{-1}$. En ce qui concerne l'observateur (23), un gain $k = 50\ s.m^{-1}$ entraîne une convergence en 15-20 images. On teste l'observateur (23) pour différents niveaux

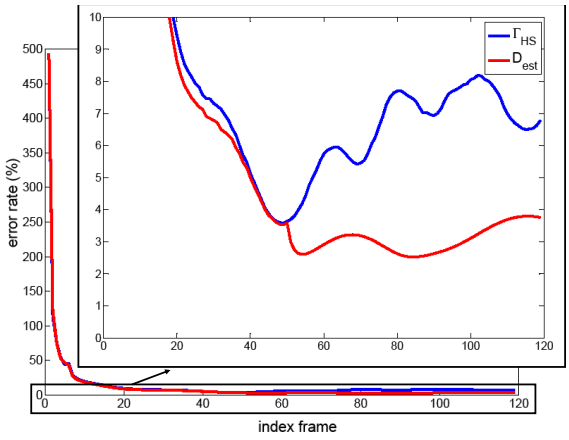

FIGURE 6 – Erreur constatée pour l'estimation directe de $1/\Gamma_{HS}$ (bleu), décrite en 3.2, et de D (rouge) par l'observateur asymptotique (23) filtrant Γ_{HS}. Le bruit est $\mathcal{N}(0, 20^2)$.

de bruit. Pour $\sigma = 1$, les taux d'erreur pour Γ_{HS} et la carte de profondeur estimée $\widehat{D}$ sont présentés en Fig. 5. L'erreur pour Γ_{HS} se stabilise en dessous de 4% mais au dessus de 2.5%. Au contraire, l'erreur associée à la carte de profondeur $\widehat{D}$ filtrée par l'observateur non-linéaire continue à diminuer, et atteint la valeur minimale de 0.5 %.

Pour $\sigma = 20$, les taux d'erreurs associés à Γ_{HS} et à $\widehat{D}$ sont présentés en Fig. 6. Pour ce niveau de bruit, on fixe $\alpha = 1000\ m.s^{-1}$ pour le calcul de Γ_{HS}.

L'observateur filtre l'erreur associée à Γ_{HS} (entre 4 et 8 %) et parvient à un taux de 3 %. Ces performances soulignent l'excellente robustesse au bruit de cet observateur.

5.4 Données réelles

FIGURE 7 – La scène statique analysée

Afin de disposer d'une information de déplacement fiable, une caméra est attachée à un chariot motorisé qui effectue un aller-retour sur environ 2m en 6s, avec une précision de localisation micro-métrique. La vitesse est uniforme, d'environ 0.3 $m.s^{-1}$. On souligne le fait que ce déplacement linéaire a été choisi pour des raisons pratiques, et que la méthode proposée est capable de gérer la dynamique de la caméra sans contraintes spécifiques. La caméra est un modèle Flea2 PGR VGA qui tourne à une fréquence de 20.83 fps, ayant un objectif avec une ouverture de 50 par 40 deg., et orientée perpendiculairement au rail.

La carte de profondeur a été estimée par l'observateur asymptotique qui filtre le flot optique brut (20). Les composants du flot ont été calculés par un algorithme employant une régularisation TV-L^1 (voir également [18]) qui est mieux adapté pour ce type de scène. Le gain à été fixé à $k = 100$. En Fig.7, on présente une image prise alors que la caméra est retournée au point de départ ; la profondeur associée au champ de vue est présentée en Fig.8. Certaines estimations sont mises en évidence en noir ; ces valeurs sont comparées aux mesures réelles prises à l'aide d'un télémètre laser (en rouge). En prenant en compte le fait que les hypothèses théoriques concernant le champ de vue et la propriété Lambertienne des surfaces ne peuvent pas être entièrement satisfaites dans l'environnement présenté, les estimations montrent une forte corrélation avec les valeurs de référence, et l'aspect global de la carte de profondeur est assez réaliste.

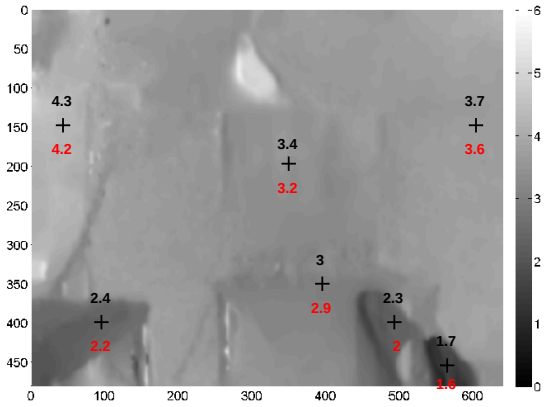

FIGURE 8 – Carte de profondeur (en m) associée à l'image 7. Certaines estimations sont explicitées (en noir) et comparées aux valeurs réelles (en rouge).

6 Conclusions et travaux futurs

Nous avons présenté dans cet article une nouvelle méthode d'estimation et affinement de carte de profondeur connaissant le mouvement de la caméra. La démarche repose sur un système d'équations aux dérivées partielles décrivant la dynamique de l'intensité lumineuse perçue par la caméra et de la profondeur. Ces équations aux dérivées partielles conduisent à une méthode variationnelle pour l'estimation de la carte de profondeur, méthode qui respecte l'invariance par rotation et qui utilise la connaissance du mouvement de la caméra. Ces équations permettent aussi de construire deux observateurs asymptotiques pour estimer le champ de profondeur à partir d'une première estimation de médiocre qualité soit du flot optique, soit du champ de profondeur. Sur des images de synthèse, nous avons montré que la performance du premier observateur dépend de la précision des estimations de flot optique, tandis que le deuxième semble plus robuste. Implémentée sur des données réelles, cette méthode donne des estimations qualitativement proches de la vérité terrain. Cette méthode d'estimation et d'affinement semble une alternative intéressante aux techniques existantes de densification à partir de séquences vidéo. Dans sa forme la plus simple, cette méthode ne requiert pas de temps de calcul important, et pourrait être implémentée sur des systèmes embarqués pour du traitement en temps réel. Cette approche doit être prochainement adaptée pour mieux prendre en compte les discontinuités de profondeur. Ses résultats doivent également être validés quantitativement pour le traitement de données réelles à des mouvements plus agressifs de la caméra.

Références

[1] J. Civera, A. Davison, and J. Montiel, "Inverse depth parametrization for monocular slam," *Robotics, IEEE Transactions on*, vol. 24, no. 5, pp. 932 –945, 2008.

[2] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM 2.0 : An improved particle filtering algorithm for simultaneous localization and mapping that provably converges," in *IJCAI*, 2003.

[3] H. Strasdat, J. M. M. Montiel, and A. Davison, "Scale drift-aware large scale monocular slam," in *Robotics : Science and Systems*, 2010.

[4] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, 2010.

[5] R. A. Newcombe and A. J. Davison, "Live dense reconstruction with a single moving camera," in *CVPR*, 2010, pp. 1498–1505.

[6] C. Strecha, R. Fransens, and L. J. V. Gool, "Wide-baseline stereo from multiple views : A probabilistic account," in *CVPR (1)*, 2004, pp. 552–559.

[7] P. Gargallo and P. F. Sturm, "Bayesian 3d modeling from images using multiple depth maps," in *CVPR (2)*, 2005, pp. 885–891.

[8] G. Vogiatzis and C. Hernández, "Video-based, real-time multi-view stereo," *Image Vision Comput.*, vol. 29, no. 7, pp. 434–441, 2011.

[9] L. Matthies, T. Kanade, and R. Szeliski, "Kalman filter-based algorithms for estimating depth from image sequences," *IJCV*, vol. 3, no. 3, pp. 209–238, 1989.

[10] B. Ghosh, M. Jankovic, and Y. Wu, "Some problems in perspective system theory and its application to machine vision," in *Intell. Robots and Systems*, vol. 1, 1992, pp. 139 –146.

[11] X. Chen and H. Kano, "A new state observer for perspective systems," *Aut. Control, IEEE Trans. on*, vol. 47, no. 4, pp. 658 –663, 2002.

[12] D. Karagiannis and A. Astolfi, "A new solution to the problem of range identification in perspective vision systems," *Aut. Control, IEEE Trans. on*, vol. 50, no. 12, pp. 2074 – 2077, 2005.

[13] M. Sassano, D. Carnevale, and A. Astolfi, "Observer design for range and orientation identification," *Automatica*, vol. 46, no. 8, pp. 1369–1375, 2010.

[14] S. Bonnabel and P. Rouchon, "Fusion of inertial and visual : a geometrical observer-based approach," *CISA*, vol. 1107, pp. 54–58, 2009.

[15] B. Horn and B. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

[16] A. Wedel, T. Pock, J. Braun, U. Franke, and D. Cremers, "Duality tv-l1 flow with fundamental matrix prior," in *IVCNZ 2008*, 2008, pp. 1–6.

[17] D. Bitton, G. Rosman, T. Nir, A. M. Bruckstein, A. Feuer, and R. Kimmel, "Over-parameterized optical flow using a stereoscopic constraint," in *SSVM*, 2011.

[18] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *JMIV*, vol. 40, pp. 120–145, 2010.