Une solution numérique générique pour la reconstruction 3D d'objets non texturés

Jean MÉLOU [1,3] Yvain QUÉAU [2] Fabien CASTAN [3] Jean-Denis DUROU [1]

[1] IRIT, UMR CNRS 5505, Université de Toulouse
[2] GREYC, UMR CNRS 6072, Caen
[3] Mikros Image, Paris

jean.melou@mikrosimage.com

Résumé

Nous proposons une stratégie, simple mais efficace, permettant d'estimer la profondeur à partir de données multi-vues. Ce problème classique, qui consiste à minimiser la somme pondérée de la cohérence photométrique et d'un terme de régularisation, est reformulé sous la forme d'une séquence de problèmes faciles à résoudre. La méthode présentée dans cet article permet de traiter une grande variété de mesures de cohérence photométrique et de termes de régularisation. Elle peut être utilisée, par exemple, pour estimer une solution dite de surface minimale, ou encore une solution fondée sur l'ombrage. Ceci rend l'approche proposée très efficace pour résoudre le problème classique de la reconstruction 3D d'objets non texturés.

Mots Clef

Stéréoscopie multi-vues, reconstruction 3D, *shape-from-shading*.

Abstract

We put forward a simple, yet effective splitting strategy for multi-view stereopsis. It recasts the minimization of the classic photo-consistency + regularization functional as a sequence of simple problems which can be solved efficiently. This framework is able to handle various photo-consistency measures and regularization terms, and can be used for instance to estimate either a minimal-surface or a shading-aware solution. This makes the proposed approach very effective for dealing with the well-known problem of textureless objects 3D-reconstruction.

Keywords

Multi-view stereo, 3D-reconstruction, shape-from-shading.

1 Introduction

La stéréoscopie multi-vues vise à reconstruire la surface d'un objet à partir d'images acquises sous différents angles. La méthode de résolution classique de ce problème consiste à estimer la carte de profondeur associée à une image de référence, en maximisant la cohérence photométrique entre celle-ci et les autres images, dites images témoins. En effet, en utilisant les paramètres des caméras, supposés connus, et la profondeur estimée, il est possible de déterminer, dans chacune des images, les pixels associés à un même point 3D. On peut alors comparer le niveau de gris en un pixel de l'image de référence à ceux des pixels « homologues » dans les images témoins. Cependant, la mesure de la cohérence photométrique n'est pas pertinente pour les surfaces non texturées (cf. figure 1) : le problème d'optimisation doit être régularisé en contraignant les variations de profondeur.

Si l'on note $z : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^+\backslash\{0\}$ la fonction de profondeur à estimer, ces variations peuvent être, en projection perspective, définies comme une fonction de $\frac{\nabla z}{z} = \nabla \log z : \Omega \rightarrow \mathbb{R}^2$. Le problème de la stéréoscopie multi-vues peut alors être formulé, de façon générique, comme la minimisation de la somme d'un terme d'attache aux données f, inversement proportionnel à la cohérence photométrique et d'un terme de régularisation g. Nous nous intéressons donc dorénavant au problème variationnel suivant :

$$\min_z f(z) + g(\nabla \log z) \tag{1}$$

Le choix d'appliquer la régularisation au logarithme sera justifié dans la partie 2.

Nous discutons du choix de f et de g pour le problème (1) dans la partie 2. La partie 3 présente notre principale contribution : un algorithme générique de stéréoscopie multi-vues, qui consiste à reformuler (1) en une série de sous-problèmes plus simples. Dans la partie 4, nous discutons des choix possibles pour le terme de régularisation qui soient adaptés à la reconstruction 3D d'objets non texturés (surface minimale, ombrage), avant d'évaluer empiriquement le potentiel de notre algorithme dans la partie 5. Pour finir, nous résumons nos travaux et esquissons quelques pistes pour de futures recherches dans la partie 6.

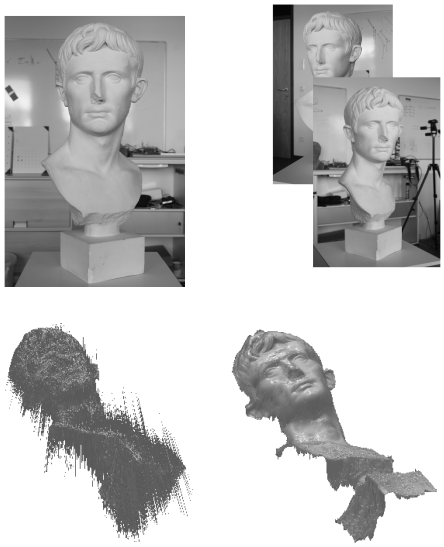

Figure 1 – À partir d'une image de référence d'un objet non texturé (en haut à gauche) et d'un ensemble de $t \geq 1$ images témoins (en haut à droite), l'optimisation de la cohérence photométrique seule ne parvient pas à évaluer une carte de profondeur satisfaisante (en bas à gauche). De bien meilleurs résultats sont obtenus en ajoutant des termes de régularisation fondés sur la surface minimale et sur l'ombrage (en bas à droite).

2 Préliminaires

Dans les travaux présentés ici, nous nous penchons sur la résolution du problème discret associé à (1). Dorénavant, z désigne donc un vecteur de $\mathbb{R}^p$ contenant les p valeurs de la profondeur à estimer, $\log z$ le vecteur des logarithmes de ces p valeurs, et $\nabla \in \mathbb{R}^{2p \times p}$ la matrice des différences finies du premier ordre, de telle sorte que $\nabla z \in \mathbb{R}^{2p}$ constitue une approximation du gradient de profondeur.

Attache aux données. La connaissance a priori des paramètres extrinsèques et intrinsèques des caméras nous permet de définir, pour chacune d'elles, une fonction de projection des points de l'espace 3D vers les pixels de l'image associée. Nous notons cette fonction π^j, $j \in \{1, \ldots, t\}$, pour les $t \geq 1$ caméras témoins. D'autre part, pour une carte de profondeur z donnée, nous notons π_z^{-1} la fonction associant à un pixel i, $i \in \{1, \ldots, p\}$, de l'image de référence le point 3D qui lui est conjugué.

La cohérence photométrique peut viser à comparer divers éléments : de manière générale, nous notons $v_i \in \mathbb{R}^m$ un vecteur caractérisant le pixel i dans l'image de référence. Un tel vecteur peut être simplement défini comme le niveau de gris de ce pixel ($m = 1$), ses valeurs RVB ($m = 3$), la concaténation des niveaux de gris sur un voisinage de taille 3×3 centré en i ($m = 9$), etc.

Pour une caméra témoin $j \in \{1, \ldots, t\}$ donnée, la cohérence photométrique mesure donc l'adéquation, au moyen d'une fonction de coût ρ, entre v_i et le vecteur caractéristique $v^j_{\pi^j \circ \pi_z^{-1}(i)} \in \mathbb{R}^m$ au pixel $\pi^j \circ \pi_z^{-1}(i)$ dans cette image témoin. Le terme f est alors construit en effectuant la somme de ces mesures sur l'ensemble des pixels de l'image de référence, puis en moyennant sur l'ensemble des images témoins :

$$f(z) = \frac{1}{t} \sum_{j=1}^{t} \sum_{i=1}^{p} \rho \left(v_i, v^j_{\pi^j \circ \pi_z^{-1}(i)} \right) \qquad (2)$$

La fonction de coût ρ peut prendre différentes formes [5, chapitre 2]. On peut, par exemple, choisir la moyenne des écarts au carré (*Sum of Squared Deviations*) $\rho_{\text{SSD}}(x, y) = \frac{1}{m} \|x - y\|^2$ ou une variante robuste, puis linéariser 2 par un développement de Taylor au premier ordre, comme cela est proposé dans [7, 13]. Cependant, une telle linéarisation fait l'hypothèse de petits incréments de profondeur. La robustesse peut également s'obtenir en remplaçant SSD par la moyenne des différences absolues (*Sum of Absolute Deviations*) $\rho_{\text{SAD}}(x, y) = \frac{1}{m} \|x - y\|_1$ ou une fonction de coût fondée sur la corrélation croisée (*Zero-mean Normalized Cross Correlation*) $\rho_{\text{ZNCC}}(x, y) = \frac{1}{2} \left[1 - \frac{(x - \overline{x})^\top (y - \overline{y})}{\|x - \overline{x}\| \|y - \overline{y}\|} \right]$. Les valeurs de cohérence photométrique sont ensuite normalisées par un opérateur non linéaire, par exemple la transformée exponentielle $\rho(x, y) := 1 - \exp \left\{ -\frac{\rho(x,y)^2}{\sigma^2} \right\}$, où σ est défini par l'utilisateur. Considérant toutes ces possibilités, le terme d'attache aux données peut se révéler non linéaire, non lisse ou non convexe, et donc rendre l'optimisation difficile. C'est la raison pour laquelle, de manière générale, la minimisation de f est effectuée par une recherche exhaustive (« force brute ») sur un ensemble échantillonné de valeurs de profondeur. La minimisation de l'erreur de reprojection permet alors de sélectionner la profondeur recherchée. Cette stratégie, présentée initialement dans [8], peut paraître simple mais se révèle très efficace pour la reconstruction de cartes de profondeur de scènes fortement texturées [6].

Régularisation. Néanmoins, le terme de cohérence photométrique n'est pas pertinent pour les parties non texturées : en chaque pixel, il existe généralement plusieurs minima globaux. Il est évident qu'en augmentant le nombre d'images témoins, on ne résout en rien ce problème, et que l'ajout de termes de régularisation s'impose. On peut, par exemple, introduire un terme de variation totale [17]. Cependant, en projection perspective, la variation totale n'est pas une mesure pertinente. Elle doit être remplacée par la mesure de l'aire (hypothèse de surface minimale [7]). Le terme de régularisation devient une forme bilinéaire impliquant à la fois la profondeur z et son gradient ∇z, ce qui rend l'optimisation complexe. Cette difficulté peut être évitée en introduisant le changement de variable $\tilde{z} = \sqrt{z}$, la mesure de l'aire pouvant alors être réécrite comme une fonction $g(\nabla \tilde{z})$ [7]. Le changement de variable logarithmique $\tilde{z} = \log z$ offre les mêmes avantages. Il permet en outre d'introduire un

terme de régularisation physiquement réaliste, fondé sur le *shape-from-shading* sous éclairage naturel [14]. Cela peut se révéler particulièrement profitable, puisque les algorithmes de *shape-from-shading* supposent généralement que la réflectance de la scène est uniforme, c'est-à-dire que la scène est non texturée. Comme il nous semble intéressant de comparer la régularisation qui découle de l'hypothèse de surface minimale (lissage arbitraire) à celle qui découle du *shape-from-shading* (lissage guidé par un modèle photométrique), nous optons pour le deuxième changement de variable dans le terme de régularisation. Ceci explique la forme du modèle variationnel (1).

La combinaison entre reconstruction 3D multi-vues et photométrie a été depuis longtemps identifiée comme une piste prometteuse [3], et des preuves d'unicité existent [4]. Les solutions numériques sont cependant peu nombreuses : Jin *et al*. ont proposé une solution variationnelle dans [9], qui suppose l'éclairage constitué d'une seule source lumineuse située à l'infini, ce qui est rarement le cas d'un éclairage naturel. D'autre méthodes combinant les deux approches ont été proposées récemment [10, 12, 18], mais elles considèrent plutôt la photométrie comme une méthode permettant d'affiner la reconstruction 3D obtenue par un critère de cohérence multi-vues. Nous proposons ici une approche jointe, comme cela a été fait dans un article très récent [13]. Par rapport à [13], une particularité de notre travail consiste à ne pas linéariser le terme d'attache aux données (cf. partie 3). Ceci rend l'approche proposée plus générique, puisque n'importe quelle mesure robuste de cohérence photométrique peut être prise en compte, y compris celles qui recourent à des fonctions de coût non lisses ou non convexes.

3 Une approche générique découplée de la stéréoscopie multi-vues

Dans cette partie, nous montrons comment transformer la version discrète du problème variationnel complexe (1) en un problème (6) plus simple, qui peut être résolu par l'algorithme 1.

Modèle variationnel. Comme nous l'avons déjà dit, la fonction f qui caractérise le terme d'attache aux données du problème (1) est généralement non lisse. Le couplage entre pixels voisins induit par l'opérateur gradient dans le terme de régularisation g constitue une difficulté supplémentaire. Nous proposons de séparer ces deux difficultés en découplant les optimisations de f et de g. Pour cette raison, nous introduisons la variable auxiliaire $u = z \in \mathbb{R}^p$, et réécrivons le problème (1) sous la forme équivalente :

$$\begin{array}{ll} \min\limits_{u,z} & f(u) + g(\nabla \log z) \\ \text{s.c.} & u = z \end{array} \tag{3}$$

Dans l'équation (3), le sous-problème en u est encore non dérivable et potentiellement non convexe, mais il peut être résolu localement (et est donc parallélisable). En effet, f est séparable : chaque terme de la somme dans (2) implique la profondeur en un unique pixel. Il est donc possible de rechercher le minimum global de f de façon exhaustive (« force brute ») sur des valeurs échantillonnées de la profondeur. De plus, en supposant g lisse, une simple méthode à directions de descente devrait suffire à minimiser le terme de régularisation. Cependant, la contrainte $u = z$ ne permettrait pas à z de décrire les variations fines du relief. Pour cette raison, nous préférons suivre une approche mixte, où le sous-problème en u est combinatoire, tandis que le sous-problème en z est continu, de sorte que le résultat final z puisse refléter au mieux les détails de la surface. Ainsi, les deux variables z et u diffèrent légèrement par essence, ce qui explique pourquoi nous transformons la contrainte dure $u = z$ du problème (3) en un terme de pénalisation quadratique :

$$\min_{u,z} f(u) + g(\nabla \log z) + \beta \left\| \log u - \log z \right\|^2 \tag{4}$$

où $\beta > 0$ est un hyper-paramètre. Notons que cette pénalisation est appliquée aux logarithmes, ce qui permet de ne faire apparaître l'inconnue z dans (4) que par le biais de son logarithme. Nous définissons donc une nouvelle variable $\tilde{z} = \log z$ et reformulons le problème (4) ainsi :

$$\min_{u,\tilde{z}} f(u) + g(\nabla \tilde{z}) + \beta \left\| \log u - \tilde{z} \right\|^2 \tag{5}$$

Il suffira ensuite de calculer $z = \exp \tilde{z}$ à la fin du processus.

Comme nous l'avons déjà signalé dans la partie précédente, des études récentes plaident pour l'utilisation de termes de régularisation non linéaires. Le sous-problème de (5) en $\tilde{z}$ demeurant donc complexe, nous le simplifions en effectuant un nouveau découplage. En introduisant une deuxième variable auxiliaire $\theta = \nabla \tilde{z} \in \mathbb{R}^{2p}$, le problème (5) s'écrit de manière équivalente :

$$\begin{array}{ll} \min\limits_{u,\theta,\tilde{z}} & f(u) + g(\theta) + \beta \left\| \log u - \tilde{z} \right\|^2 \\ \text{s.c.} & \theta = \nabla \tilde{z} \end{array} \tag{6}$$

Résolution numérique du problème (6)**.** La contrainte linéaire de l'équation (6) pourrait être prise en compte en ayant recours, par exemple, à une approche de type Lagrangien augmenté. Nous préférons, dans ce travail exploratoire, suivre une stratégie plus simple consistant à approcher la solution de (6) par résolution, à chaque itération, d'un problème qui comporte une pénalisation quadratique de la forme :

$$\min_{u,\theta,\tilde{z}} f(u) + g(\theta) + \alpha^{(k)} \left\| \theta - \nabla \tilde{z} \right\|^2 + \beta \left\| \log u - \tilde{z} \right\|^2 \tag{7}$$

où les valeurs successives $\alpha^{(k)} > 0$ augmentent au fil des itérations k. Nous souhaitons, en effet, que la contrainte dure de (6) soit satisfaite à la convergence, c'est-à-dire lorsque $k \to +\infty$, par opposition à la contrainte de (3), qui a été sciemment relaxée avec un paramètre β fixe.

Pour chaque valeur $\alpha^{(k)}$, la solution du problème (7) est approchée par une étape d'optimisation alternée des différents sous-problèmes. Comme nous l'avons déjà signalé, le sous-problème en u est résolu par une recherche exhaustive du minimum (force brute). Nous considérons dans ce travail des termes de régularisation g séparables et lisses (cf. partie 4), de telle sorte que le sous-problème en θ puisse être résolu par une méthode de gradient parallélisable. Enfin, le sous-problème en $\tilde{z}$ est un problème de moindres carrés linéaires, qui peut être résolu par la méthode du gradient conjugué. Le processus est répété jusqu'à ce que l'écart relatif entre deux estimations successives de $z = \exp \tilde{z}$ soit inférieur à un seuil égal à 10^{-4}. L'algorithme 1 estime, dans un premier temps, une carte de profondeur grossière par optimisation de la cohérence photométrique (équation (8)), puis régularise la surface (équation (9)), avant d'intégrer le gradient ainsi obtenu en une *log*-carte de profondeur (équation (10)). Les valeurs $\alpha^{(0)} = 1$ et $\beta = 0,1$, qui ont été déterminées empiriquement, permettent d'obtenir des résultats satisfaisants. Elles sont donc utilisées dans l'ensemble des tests. Enfin, la carte de profondeur est initialisée à une valeur uniforme $z^{(0)}$ égale à la profondeur moyenne de la vérité terrain, qui correspond à un plan fronto-parallèle.

Algorithme 1 : Algorithme générique de reconstruction 3D multi-vues.

Entrées : $z^{(0)}$, $\alpha^{(0)} > 0$, $\beta > 0$
Sorties : Carte de profondeur affinée z

$\tilde{z}^{(0)} = \log z^{(0)}$, $k = 0$, $r^{(0)} = +\infty$;
tant que $r^{(k)} > 10^{-4}$ **faire**
  `// Optimisation de la cohérence photométrique`
  $u^{(k+1)} = \underset{u}{\operatorname{argmin}}\ f(u) + \beta \left\| \log u - \tilde{z}^{(k)} \right\|^2 ;$ (8)
  `// Régularisation des variations de profondeur`
  $\theta^{(k+1)} = \underset{\theta}{\operatorname{argmin}}\ g(\theta) + \alpha^{(k)} \left\| \theta - \nabla \tilde{z}^{(k)} \right\|^2 ;$ (9)
  `// Intégration`
  $\tilde{z}^{(k+1)} = \underset{\tilde{z}}{\operatorname{argmin}}\ \alpha^{(k)} \left\| \nabla \tilde{z} - \theta^{(k+1)} \right\|^2 + \beta \left\| \tilde{z} - \log u^{(k+1)} \right\|^2 ;$ (10)
  `// Mise à jour des variables auxiliaires`
  $\alpha^{(k+1)} = 1,5\, \alpha^{(k)}$;
  $z^{(k+1)} = \exp \tilde{z}^{(k+1)}$;
  $r^{(k)} = \frac{\left\| z^{(k+1)} - z^{(k)} \right\|}{\left\| z^{(k)} \right\|}$;
  $k = k + 1$;
fin

4 Termes de régularisation pour la reconstruction 3D d'objets non texturés

Comme f et g peuvent être non convexes, il n'est pas évident d'analyser la convergence de l'algorithme 1. Cette analyse théorique est donc laissée momentanément de côté, au profit d'une évaluation empirique de notre algorithme sur des données multi-vues réelles. Dorénavant, nous nous intéressons tout particulièrement au problème délicat de la reconstruction 3D d'objets peu texturés. Afin de justifier le choix de tel ou tel terme de régularisation, il nous faut clarifier la notion d'objet non texturé. Commençons donc par rappeler quelques notions de photométrie.

Modèle lambertien de formation de l'image. Dans (2), le terme d'attache aux données traduit l'hypothèse classique que la luminance d'un point de la surface est invariante par changement de point de vue, ce qui est le propre des surfaces lambertiennes, pour lesquelles la réflectance est entièrement caractérisée par l'albédo. Dans le cas où l'éclairage est constitué d'une seule source lumineuse située à l'infini, le niveau de gris I_i au pixel i dans la vue de référence est le produit de l'albédo par l'ombrage :

$$I_i = a_i \max\left\{0, n_i^\top l\right\} \tag{11}$$

où $a_i > 0$ désigne l'albédo au point 3D $\pi_z^{-1}(i)$ conjugué de i, $n_i \in \mathbb{S}^2 \subset \mathbb{R}^3$ est le vecteur unitaire normal à la surface en ce point 3D, et $l \in \mathbb{R}^3$ caractérise le vecteur d'éclairage, en intensité et en orientation. La normale à la surface dépend du gradient de la *log*-profondeur, c'est-à-dire de θ, selon la relation (voir, par exemple [14]) :

$$n_i := n(\theta_i) = \frac{1}{d(\theta_i)} \begin{bmatrix} \mathrm{f}\,\theta_i \\ -1 - [x,y]^\top \cdot \theta_i \end{bmatrix} \tag{12}$$

où nous notons $\theta_i = \begin{bmatrix} \theta_i^1 \\ \theta_i^2 \end{bmatrix} \in \mathbb{R}^2$ le gradient de profondeur au pixel i (le vecteur $\theta \in \mathbb{R}^{2p}$ étant la concaténation des θ_i, $i \in \{1, \ldots, p\}$), où f est la distance focale de la caméra et $(x, y) \in \mathbb{R}^2$ sont les coordonnées centrées du pixel i, et où la contrainte de longueur unitaire du vecteur n_i est garantie par le coefficient de normalisation :

$$d(\theta_i) = \sqrt{\mathrm{f}^2 \|\theta_i\|^2 + \left(1 + [x,y]^\top \cdot \theta_i\right)^2} \tag{13}$$

Le modèle (11) est valide sous l'hypothèse peu réaliste d'une seule source lumineuse située à l'infini. Un éclairage naturel peut cependant être vu comme un ensemble de sources lumineuses situées à l'infini, et la luminance au pixel i est alors obtenue en intégrant le membre droit de l'équation (11) sur l'hémisphère extérieur à la surface. En approchant cette intégrale par des harmoniques sphériques du second ordre, nous obtenons (voir [2] pour plus de détails) :

$$I_i = a_i\, \tilde{n}_i^\top \tilde{l} \tag{14}$$

où $\tilde{l} \in \mathbb{R}^9$ est une représentation de l'éclairage qui peut être étalonnée à l'aide d'un objet de géométrie et de réflectance connues, et où $\tilde{n}_i \in \mathbb{R}^9$ est la « pseudo-normale », qui dépend uniquement des trois composantes de $n_i = \left[n_i^1, n_i^2, n_i^3\right]^\top$, donc de θ_i :

$$\tilde{n}_i := \tilde{n}(\theta_i) = \begin{bmatrix} n_i \\ 1 \\ n_i^1\, n_i^2 \\ n_i^1\, n_i^3 \\ n_i^2\, n_i^3 \\ \left(n_i^1\right)^2 - \left(n_i^2\right)^2 \\ 3\left(n_i^3\right)^2 - 1 \end{bmatrix} \overset{(12)}{=} \begin{bmatrix} \frac{\mathrm{f}\,\theta_i}{d(\theta_i)} \\ \frac{-1-[x,y]^\top\cdot\theta_i}{d(\theta_i)} \\ 1 \\ \frac{\mathrm{f}^2\theta_i^1\theta_i^2}{d(\theta_i)^2} \\ \frac{\mathrm{f}\theta_i^1\left(-1-[x,y]^\top\cdot\theta_i\right)}{d(\theta_i)^2} \\ \frac{\mathrm{f}\theta_i^2\left(-1-[x,y]^\top\cdot\theta_i\right)}{d(\theta_i)^2} \\ \frac{\mathrm{f}^2\left(\left(\theta_i^1\right)^2-\left(\theta_i^2\right)^2\right)}{d(\theta_i)^2} \\ \frac{3\left(-1-[x,y]^\top\cdot\theta_i\right)^2}{d(\theta_i)^2} - 1 \end{bmatrix} \tag{15}$$

Si la scène observée est très texturée, les valeurs a_i de l'albédo dans (14) varient fortement d'un pixel i à l'autre. Il en va donc de même pour les valeurs du niveau de gris I_i et pour les vecteurs caractéristiques v_i, ce qui rend l'optimisation du terme d'attache aux données $f(z)$ dans (2) pertinente, même en l'absence de régularisation. Cependant, si la scène n'est pas texturée, l'albédo est uniforme, par exemple égal à 1 :

$$a_i = 1 \quad \forall i \in \{1, \ldots, p\} \tag{16}$$

ce qui, d'après (14) et (15), rend les variations de luminance purement géométriques, c'est-à-dire dépendant uniquement des variations de n_i. Ces variations peuvent être très subtiles, et donc inadaptées à un terme d'attache aux données tel que celui de l'équation (2), ce qui justifie le recours à la régularisation. Nous présentons dans ce qui suit deux termes de régularisation plausibles.

Régularisation fondée sur l'ombrage. L'équation (14) peut servir de guide pour la stéréoscopie multi-vues, afin que le modèle lambertien puisse lever l'ambiguïté du problème de mise en correspondance des parties non texturées. Par exemple, si on suppose que le modèle lambertien est satisfait à un bruit blanc gaussien près, on peut minimiser les résidus de l'équation (14) au sens des moindres carrés, dans l'esprit de l'approche variationnelle du *shape-from-shading* sous éclairage naturel présentée dans [14]. Ceci nous amène à la régularisation suivante (rappelons que $a_i = 1$) :

$$g_{\text{SFS}}(\theta) = \lambda \sum_{i=1}^{p} \left(\tilde{n}(\theta_i)^\top \tilde{l} - I_i\right)^2 \tag{17}$$

où $\lambda > 0$ est un hyper-paramètre. L'utilisation du terme de régularisation présenté dans (17) dans l'algorithme 1 fournit une solution fondée sur l'ombrage de la stéréoscopie multi-vues. On notera que $g_{\text{SFS}}(\theta)$ est lisse et séparable (chaque terme de la somme n'implique que $\theta_i = [\theta_i^1, \theta_i^2]^\top \in \mathbb{R}^2$), et donc (9) peut être reformulé comme p problèmes non linéaires en deux dimensions, qui peuvent être résolus en parallèle, en utilisant par exemple la méthode BFGS [11, 15].

Régularisation de surface minimale. La régularisation précédente requiert la connaissance du vecteur d'éclairage $\tilde{l}$. Dans certains cas, l'étalonnage de l'éclairage peut se révéler compliqué, voire impossible, et il peut être préférable de ne pas utiliser un modèle explicite de formation de l'image. Dans ce cas, il est possible de simplement limiter les variations de la surface en pénalisant, par exemple, son aire. Suivant ce qui a été proposé dans [14], nous pénalisons la norme ℓ^1 des valeurs du champ d qui a pour expression (13). Nous définissons donc le terme de régularisation de surface minimale suivant :

$$g_{\text{MS}}(\theta) = \mu \sum_{i=1}^{p} d(\theta_i) \tag{18}$$

À nouveau, la fonction $g_{\text{MS}}(\theta)$ est lisse et séparable, de telle sorte que l'équation (9) peut être résolue en parallèle par BFGS.

Régularisation combinée. Bien entendu, le terme de régularisation de surface minimale (18) favorisera les surfaces lisses, et risquera donc de ne pas restituer les détails de la surface. À l'inverse, le terme de *shape-from-shading* (17) cherchera à expliquer les moindres variations du niveau de gris de l'image par des petites variations de profondeur, ce qui risque de provoquer une mauvaise interprétation du bruit. C'est pourquoi il semble intéressant de combiner ces deux termes de régularisation. Nous utiliserons donc, dans les tests, le terme de régularisation suivant :

$$g(\theta) = \underbrace{\lambda \sum_{i=1}^{p} \left(\tilde{n}(\theta_i)^\top \tilde{l} - I_i\right)^2}_{g_{\text{SFS}}(\theta)} + \underbrace{\mu \sum_{i=1}^{p} d(\theta_i)}_{g_{\text{MS}}(\theta)} \tag{19}$$

qui reste lisse et séparable, et qui permet d'obtenir la solution guidée par l'ombrage si $\lambda > 0$ et $\mu = 0$, et la solution de surface minimale si $\lambda = 0$ et $\mu > 0$.

5 Tests

Dans l'ensemble des tests, le vecteur caractéristique v_i est égal à la concaténation des niveaux de gris dans un voisinage du pixel i de taille 3×3. Sans indication contraire, la fonction de coût ρ utilisée dans (2) est la transformée exponentielle de SAD (avec $\sigma = 0,2$). Dans un premier temps, nous testons le modèle proposé sur des données de synthèse. Les images de « Stanford's Bunny », obtenues à l'aide d'un logiciel de rendu, sont de taille 540×540. L'albédo est supposé uniforme, et l'éclairage $\tilde{l}$, comme les paramètres des caméras, sont supposés connus. Un bruit gaussien d'écart-type égal à 1% du niveau de gris maximal a été ajouté, afin de se rapprocher des images réelles.

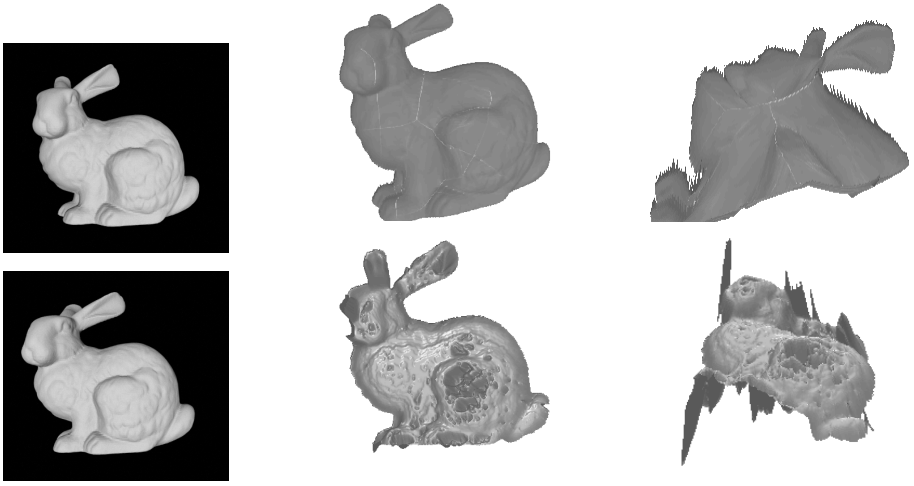

Figure 2 – Première ligne : le *shape-from-shading* permet de retrouver les détails fins (au centre, la surface estimée est vue de face) de l'image de synthèse de gauche, mais la forme globale de l'objet est biaisée à cause de l'ambigüité concave/convexe (à droite, la surface est vue sous un autre angle). Deuxième ligne : stéréoscopie multi-vues non régularisée ($t = 1$), où l'image témoin (à gauche) est obtenue par translation de la caméra perspective. Si la forme globale de l'objet est satisfaisante, les détails de la surface ont été gommés et des artéfacts apparaissent dans les zones non texturées.

Nous commençons par montrer sur la figure 2 le principal inconvénient du *shape-from-shading* ($f(z) = 0$, $\lambda = 1$ et $\mu = 0$) : bien que l'image de référence soit bien expliquée, la profondeur estimée est sujette à l'ambiguïté bien connue concave/convexe. La stéréoscopie multi-vues non régularisée ($f(z) =$ (2), et $\lambda = \mu = 0$) n'est pas satisfaisante, elle non plus : l'ajout d'une deuxième vue (ici, la caméra perspective a effectué une simple translation) et l'optimisation de la cohérence photométrique produisent un résultat bruité, à cause des ambiguïtés de mise en correspondance dans les parties non texturées.

Comme le montre la figure 3, les résultats sont améliorés par l'ajout de termes de régularisation. Quand ils ne sont pas nuls, les hyper-paramètres sont fixés aux valeurs suivantes : $\lambda = 5.10^{-4}$ et $\mu = 5.10^{-5}$ (ces valeurs ont été déterminées de manière empirique). Comme on pouvait s'y attendre, le terme de surface minimale permet d'estimer une carte de profondeur sans bruit et globalement satisfaisante, bien que les détails fins ne soient pas retrouvés. L'utilisation du terme de régularisation fondé sur l'ombrage permet de retrouver les détails de la surface, mais une seule image témoin ($t = 1$) ne permet pas de lever toutes les ambiguïtés de type concave/convexe. Enfin, l'approche conjointe offre les meilleurs résultats, étant entendu que les avantages des deux termes se combinent.

Notons que nous n'avons pas pris en compte les problèmes de visibilité de façon explicite : la profondeur estimée ne peut donc pas être valide pour les parties non visibles dans l'image témoin : ce problème survient en particulier au niveau du côté droit du lapin (cf. figure 3). Afin d'éviter ce problème, on peut simplement augmenter le nombre t d'images témoins ($t = 6$ dans l'exemple de la figure 4),

RMSE = 27 RMSE = 28,8 RMSE = **24,4**

Figure 3 – Stéréoscopie multi-vues avec une seule image témoin (les deux images utilisées sont celles de la figure 2). De gauche à droite : terme de surface minimale seul ($\lambda = 0$), terme fondé sur l'ombrage seul ($\mu = 0$) et combinaison des deux termes ($\lambda > 0$ et $\mu > 0$).

de telle sorte que chaque partie de l'objet visible dans l'image de référence soit également visible dans au moins une image témoin. Les inévitables occultations sont traitées comme des données aberrantes par le terme d'attache aux données robuste $f(z)$. Cela nous permet d'améliorer sensiblement les résultats, comme le confirme la mesure de la RMSE (exprimée en millimètres, les valeurs de la vérité terrain variant dans un intervalle de 800 millimètres) entre la reconstruction 3D et la vérité terrain. Notons que le terme de régularisation par l'ombrage seul semble suffisant. Le terme de surface minimale lisse exagérément la surface et efface ainsi les détails fins. Ceci est confirmé par les courbes de la figure 5 : augmenter le nombre d'images témoins supprime toutes les ambiguïtés concave/convexe du *shape-from-shading*, de sorte que le poids du terme de surface minimale doit être réduit, dans la mesure où il n'est pas fondé sur un modèle physique et où il tend à aplatir systématiquement la surface.

RMSE = 25,3 RMSE = **19,5** RMSE = 23,1

Figure 4 – Stéréoscopie multi-vues avec $t = 6$ images témoins. De gauche à droite : terme de surface minimale seul ($\lambda = 0$), terme d'ombrage seul ($\mu = 0$) et combinaison des deux termes ($\lambda > 0$ et $\mu > 0$). Les erreurs dues aux occultations, visibles sur la figure 3, ont été sensiblement réduites.

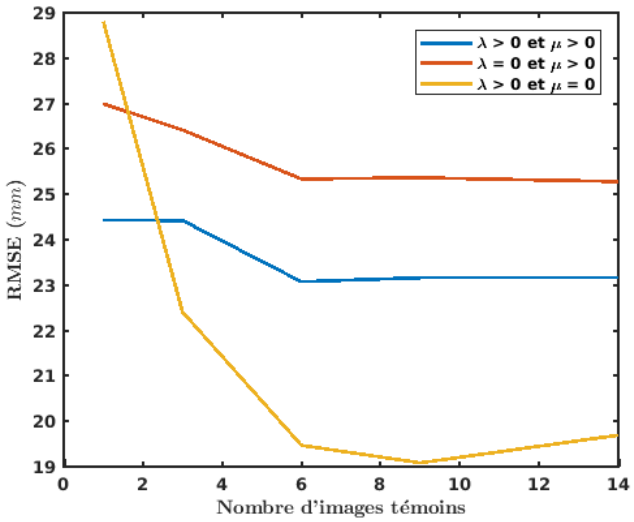

Figure 5 – RMSE pour différents termes de régularisation, et pour différentes valeurs du nombre t d'images témoins. Si l'approche combinée offre de meilleurs résultats pour un petit nombre d'images témoins, l'ombrage seul fonctionne mieux dès lors que $t > 3$.

Pour finir, nous replaçons ces travaux dans le contexte de prises de vue réelles, en utilisant des images d'un buste de l'empereur Auguste [18]. Afin d'estimer les paramètres des caméras ainsi qu'une carte de profondeur grossière, à partir de laquelle nous évaluons l'éclairage et la position du plan fronto-parallèle initial, nous avons utilisé un pipeline de photogrammétrie [1]. Il est notable que cette carte de profondeur grossière n'est pas utilisée comme relief initial. Pour illustrer la possibilité offerte par notre approche d'utiliser différentes mesures de cohérence photométrique, nous présentons les résultats obtenus à partir des transformées exponentielles de SAD et de ZNCC. Parmi les résultats de la figure 6, la dernière reconstruction (en bas à droite), qui utilise la transformée exponentielle ZNCC et une combinaison des termes de régularisation, est la plus satisfaisante, du moins d'un point de vue qualitatif.

6 Conclusion et perspectives

Dans cet article, nous avons présenté un algorithme générique de stéréoscopie multi-vues fondée sur un double découplage. Cet algorithme est suffisamment générique pour pouvoir s'adapter à différentes mesures de cohérence photométrique et à différents termes de régularisation. Notre approche permet donc bien d'atteindre le but visé, à savoir la reconstruction 3D d'objets non texturés guidée par l'ombrage ou par l'hypothèse de surface minimale. Elle présente, en outre, l'avantage de comporter très peu de paramètres à régler.

Le schéma numérique proposé pourrait être étendu à des régularisations d'ordres supérieurs [16], ainsi qu'à l'estimation jointe de la profondeur, de la réflectance et de l'éclairage, comme cela a été fait récemment dans [13]. Enfin, notre approche pourrait être transformée en une approche volumétrique, afin de reconstruire un modèle 3D complet, comme dans [12], et non plus seulement une carte de profondeur.

References

[1] AliceVision. `https://github.com/alicevision/AliceVision`

[2] Basri, R., Jacobs, D.P.: Lambertian reflectances and linear subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(2), 218–233 (2003)

[3] Blake, A., Zisserman, A., Knowles, G.: Surface descriptions from stereo and shading. Image and Vision Computing 3(4), 183–191 (1985)

[4] Chambolle, A.: A uniqueness result in the theory of stereo vision: coupling shape from shading and binocular information allows unambiguous depth reconstruction. Annales de l'IHP - Analyse non linéaire 11(1), 1–16 (1994)

[5] Furukawa, Y., Hernández, C.: Multi-view stereo: A tutorial. Foundations and Trends in Computer Graphics and Vision 9(1-2), 1–148 (2015)

[6] Goesele, M., Curless, B., Seitz, S.M.: Multi-view stereo revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 2402–2409 (2006)

[7] Graber, G., Balzer, J., Soatto, S., Pock, T.: Efficient minimal-surface regularization of perspective depth maps in variational stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 511–520 (2015)

[8] Hernández, C., Schmitt, F.: Silhouette and stereo fusion for 3D object modeling. Computer Vision and Image Understanding 96(3), 367–392 (2004)

[9] Jin, H., Cremers, D., Wang, D., Yezzi, A., Prados, E., Soatto, S.: 3-D Reconstruction of Shaded Objects from Multiple Images Under Unknown Illumination.

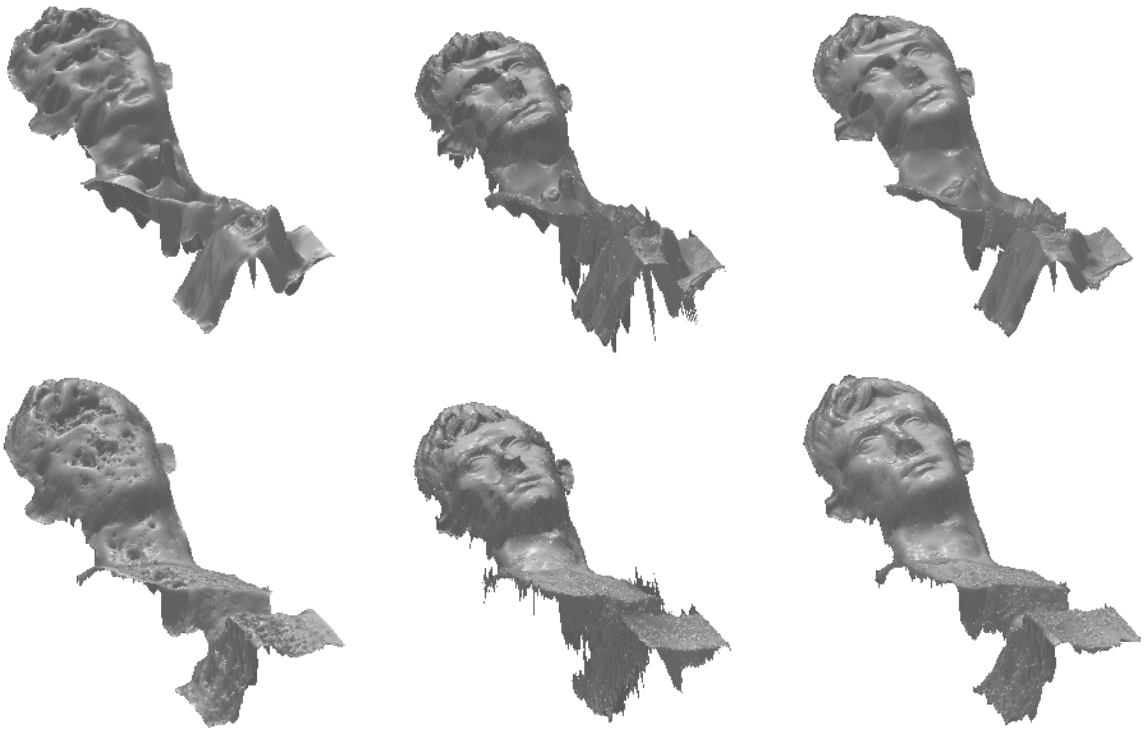

Figure 6 – Résultats de notre algorithme sur des données réelles, avec $t = 6$ images témoins (deux d'entre elles sont représentées sur la figure 1), en utilisant les transformations exponentielles des fonctions de coût SAD (première ligne) et ZNCC (deuxième ligne). De gauche à droite : terme de surface minimale seul ($\lambda = 0$, $\mu = 1.10^{-5}$), terme d'ombrage seul ($\lambda = 5.10^{-3}$, $\mu = 0$) et combinaison des deux termes de régularisation ($\lambda = 5.10^{-3}$, $\mu = 1.10^{-5}$).

International Journal of Computer Vision 76(3), 245–256 (2008)

[10] Langguth, F., Sunkavalli, K., Hadap, S., Goesele, M.: Shading-aware Multi-view Stereo. In: Proceedings of the European Conference on Computer Vision. pp. 469–485 (2016)

[11] Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical Programming 45(1-3), 503–528 (1989)

[12] Maier, R., Kim, K., Cremers, D., Kautz, J., Nießner, M.: Intrinsic3D: High-Quality 3D Reconstruction by Joint Appearance and Geometry Optimization with Spatially-Varying Lighting. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3133–3141 (2017)

[13] Maurer, D., Ju, Y.C., Breuß, M., Bruhn, A.: Combining shape from shading and stereo: A joint variational method for estimating depth, illumination and albedo. International Journal of Computer Vision 126(12), 1342–1366 (2018)

[14] Quéau, Y., Mélou, J., Castan, F., Cremers, D., Durou, J.D.: A variational approach to shape-from-shading under natural illumination. In: International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. pp. 342–357 (2017)

[15] Schmidt, M.: minFunc: unconstrained differentiable multivariate optimization in Matlab (2005), `http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html`

[16] Schroers, C., Hafner, D., Weickert, J.: Multiview Depth Parameterisation with Second Order Regularisation. In: Proceedings of the International Conference on Scale Space and Variational Methods in Computer Vision (2015)

[17] Wendel, A., Maurer, M., Graber, G., Pock, T., Bischof, H.: Dense reconstruction on-the-fly. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1450–1457 (2012)

[18] Zollhöfer, M., Dai, A., Innman, M., Wu, C., Stamminger, M., Theobalt, C., Nießner, M.: Shading-based refinement on volumetric signed distance functions. ACM Transactions on Graphics 34(4), 96:1–96:14 (2015)