
Cadre générique pour le recalage dense combinant un coût dense et un coût basé sur des correspondances de primitives

Jim Braux-Zin¹, Romain Dupont¹, Adrien Bartoli²

1. CEA SACLAY

91191 Gif sur Yvette, France

j.brauxzin@gmail.com, dupont.romain@gmail.com

2. ISIT - Université d'Auvergne/CNRS

63000 Clermont-Ferrand, France

adrien.bartoli@gmail.com

RÉSUMÉ. L'estimation dense de correspondances entre deux images est un sujet essentiel de la vision par ordinateur et s'exprime sous plusieurs formes : déformations rigides ou flexibles avec de faibles ou grandes amplitudes de déplacements. De nombreuses solutions spécifiques existent mais aucune méthodologie unifiée n'a été formulée. Cet article propose une nouvelle approche générale qui combine de manière robuste un coût dense par pixel et un coût basé sur des correspondances de primitives. Ce dernier utilise une distance robuste permettant d'exploiter des correspondances de points ou de segments. Les correspondances permettent d'empêcher l'optimisation dense de tomber dans un minimum local. En utilisant un coût dense robuste, associé à une régularisation au second ordre et une détection explicite des auto-occultations, nous obtenons des résultats égalant ou surpassant l'état de l'art pour les applications de flot optique 2D, stéréo à fortes disparité et recalage de surfaces déformables.

ABSTRACT. Dense motion field estimation is a key computer vision problem. Many solutions have been proposed to compute small or large displacements, narrow or wide baseline stereo disparity, or non-rigid surface registration, but a unified methodology is still lacking. We here introduce a general framework that robustly combines direct and feature-based matching. The feature-based cost is built around a novel robust distance function that handles keypoints and weak features such as segments. It allows us to use putative feature matches to guide dense motion estimation out of local minima. Our framework uses a robust direct data term with a powerful second order regularization. Our framework achieves state of the art performance in several cases (standard optical flow benchmarks, wide-baseline stereo and non-rigid surface registration).

MOTS-CLÉS : flot optique, stéréo, correspondances, primitives.

KEYWORDS: optical flow, stereo, matching, features.

DOI:10.3166/TS.32.195-213 © 2015 Lavoisier

1. Introduction

Un champ dense de correspondances, appelé flot optique dans le cas général 2D, est une information très utile. Les méthodes actuelles de l'état de l'art utilisent des algorithmes d'optimisation variationnelle étendant le modèle de Horn et Schunk (Horn, Schunck, 1981). Elles consistent en la minimisation couplée d'un terme de donnée, par exemple basé sur l'hypothèse de constance de l'intensité, et d'un terme de régularisation. L'utilisation de normes non quadratiques (Werlberger *et al.*, 2009 ; Ranftl *et al.*, 2012), de termes de données robustes aux variations de luminosité (Zabih, Woodfill, 1994 ; Mei *et al.*, 2011 ; Ranftl *et al.*, 2012) et de régularisations favorisant des solutions affines par morceaux (Bredies *et al.*, 2010 ; Ranftl *et al.*, 2012) a fait progresser la précision et la robustesse des résultats. Le problème des auto-occultations peut être mitigé par l'emploi d'une régularisation anisotropique (Werlberger *et al.*, 2009) mais celles-ci sont très sensibles. La meilleure approche consiste à détecter explicitement les auto-occultations (Xu *et al.*, 2010 ; Gay-Bellile *et al.*, 2010). Effectuer la mise en correspondance de manière pyramidale, de basse à haute résolution améliore la convergence. Cependant cela ne permet pas d'éviter les minima locaux dans les cas suivants : un petit élément se déplaçant rapidement dans l'image ou de fortes distorsions, perspectives ou non rigides, qui rendent les images trop différentes même à basse résolution.

Contrairement aux méthodes denses, la mise en correspondance de primitives (points, segments...) par détection et description est mature et beaucoup moins sujette à des minima locaux : sans *a priori* fort sur la scène observée, les fausses correspondances ne peuvent toutefois pas être totalement supprimées. Les descripteurs de points sont pour la plupart basés sur le concept d'histogramme de gradients (Lowe, 2004). Nous voulons insister ici sur un autre type de primitives spécialement utiles pour les scènes d'origine humaine : les segments. Wang *et al.* (2009) ont mis au point un descripteur de groupe de segments qui encode des propriétés géométriques semi-locales, bien adapté au cas d'images peu texturées ou en présence de distorsions perspectives importantes.

Le domaine du recalage de surfaces non rigides constitue un bon exemple de la dualité entre les primitives et l'information dense. Le suivi de surface consiste à la mise à jour image par image dans une vidéo des déformations d'une surface non rigide. Entre deux images les déformations sont faibles ce qui permet aux approches denses de converger sans problème et d'offrir la meilleure précision (Gay-Bellile *et al.*, 2010). Quand la déformation entre deux images est importante, on parle de détection de surface non rigide, et là seules des approches basées sur des correspondances de points sont en concurrence (Pilet *et al.*, 2008 ; Pizarro, Bartoli, 2012). Cette même dualité est présente dans la reconstruction 3D où dans le cas de faibles disparités les cartes de profondeurs sont générées par des méthodes denses. Mais dans le cas de changement important de point de vue, ce sont les approches *Structure-from-Motion* (Hartley, Zisserman, 2000) éparses qui sont utilisées.

Il n'y a eu que quelques tentatives pour tirer parti en même temps de l'image et des correspondances de primitives. Xu *et al.* (2010) effectuent une optimisation

discrète parmi des déplacements candidats induits par, entre autres, des correspondances SIFT (Lowe, 2004). Le résultat est ensuite régularisé. Les résultats obtenus sont très précis mais le coût de traitement est important. Brox et Malik (2011) ont inspiré notre travail par leur approche couplant des correspondances de points à un terme basé image au sein d'une optimisation variationnelle. Leur méthode est encore à ce jour une référence, mais plusieurs limitations restreignent son potentiel.

Notre contribution consiste en l'élargissement considérable du domaine d'applicabilité des méthodes variationnelles d'estimation dense de flot optique. Pour cela les coûts basés image et correspondances éparées sont combinés de manière robuste et flexible, compatible avec tous les détecteurs et descripteurs de primitives. La mise en correspondance est effectuée une seule fois sur les images à résolution maximale mais permet d'éviter les minima locaux pendant tout le processus d'estimation. Après une courte introduction sur la méthode LDOF (Brox, Malik, 2011) en section 2, notre cadre de travail est introduit. Il est construit autour d'une régularisation variation totale généralisée (Bredies *et al.*, 2010) (abrégée TGV pour *Total Generalized Variation*) en section 3.1, d'un coût dense robuste prenant en compte les occultations en section 3.2, et d'un nouveau coût basé sur les primitives en section 3.3, robuste aux fausses correspondances et capable d'intégrer aussi bien des points d'intérêt que des segments. Chacun de ces blocs peut être indépendamment remplacé pour tirer parti des progrès effectués dans les différents domaines impliqués sans remettre en cause l'architecture générale ni aucun des autres blocs.

Notations

Nous considérons une paire d'images I_0 et I_1 . Le champ de déplacements est estimé sur le domaine $\Omega \subset \mathbb{R}^2$ de I_0 . Les images sont considérées comme des fonctions continues $\Omega \rightarrow \mathbb{R}$ en interpolant les intensités. La norme euclidienne L^2 est notée $\|\cdot\|$ et la norme L^1 $|\cdot|$. Les vecteurs et fonctions vectorielles s'écrivent en gras avec des lettres minuscules (\mathbf{x}), et les matrices en gras et majuscules (\mathbf{J}). Le champ de déplacements estimé est appelé $\mathbf{u} : \Omega \rightarrow \mathbb{R}^2$ tel que pour tout $\mathbf{x} \in \Omega$, $I_0(\mathbf{x}) \approx I_1(\mathbf{x} + \mathbf{u}(\mathbf{x}))$.

2. Large displacement optical flow

Cette section consiste en un résumé de l'approche de Brox et al. (Brox, Malik, 2011) qui introduisent des descripteurs de points d'intérêt dans une méthode variationnelle d'estimation de flot optique en ajoutant un nouveau terme dans la fonction de coût :

$$C_{\text{LDOF}}(\mathbf{u}) = C_{\text{couleur}}(\mathbf{u}) + \gamma C_{\text{grad}}(\mathbf{u}) + \alpha C_{\text{lissage}}(\mathbf{u}) \quad (1)$$

$$+ \beta \sum_{i=1}^N C_{\text{corr}}(\mathbf{u}, \mathbf{u}_{\text{corr}}^{(i)}) + \underbrace{C_{\text{desc}}(\mathbf{u}_{\text{corr}}^{(i)})}_{\text{plus proches voisins}}$$

où $C_{\text{couleur}}(\mathbf{u}) + \gamma C_{\text{grad}}(\mathbf{u})$ est le coût dense, $C_{\text{lissage}}(\mathbf{u})$ le terme de régularisation, et la deuxième ligne le coût basé sur les correspondances. $C_{\text{corr}}(\mathbf{u}, \mathbf{u}_{\text{corr}}^{(i)})$ est défini par

$$\int_{\Omega} \delta_i(\mathbf{x}) \rho_i(\mathbf{u}_{\text{corr}}^{(i)}) \Psi(\|\mathbf{u}(x) - \mathbf{u}_{\text{corr}}^{(i)}\|) d\mathbf{x} \quad (2)$$

où pour chaque correspondance $i \in 1 \dots N$, $\mathbf{u}_{\text{corr}}^{(i)}$ est le déplacement induit par la correspondance, δ_i est une fonction indicateur pour les pixels affectés et $\rho_i(\mathbf{u}_{\text{corr}}^{(i)})$ le score de la correspondance. Ψ est une fonction de coût robuste, approximation convexe de la norme L^1 . Brox et al. ont proposé deux types de primitives pour fournir un déplacement *a priori* aux pixels affectés : des régions associées à une description riche (SIFT et couleur), ou des points échantillonnés sur une grille régulière fine et associés à de simples descripteurs basés gradient. Les correspondances sont effectuées par recherche des plus proches voisins dans l'espace des descripteurs ; c'est cette étape qui est représentée par le terme $C_{\text{desc}}(\mathbf{u}_{\text{corr}})$ qui n'intervient pas dans le reste de l'optimisation. Le terme de coût dense est la somme des différences d'intensité et de gradient. Le terme de régularisation est la variation totale du champ de déplacements.

Notre formulation ne partage que peu de détails avec LDOF mais a été inspirée par deux principes mis à jour par Brox et al. D'abord, ils ont démontré que si les correspondances n'améliorent pas la précision de l'estimation dans les cas "faciles", elles permettent d'éviter les minima locaux quand les hypothèses de l'approche pyramidale ne sont pas respectées. Plus important, ils ont observé le comportement notable de l'approche pyramidale qui équilibre automatiquement les termes de la fonction de coût. En effet, les correspondances ont une grande influence à basse résolution où elles sont quasi-denses et de plus en plus faible au fur et à mesure que la résolution augmente et que le terme dense gagne en importance.

En dépit des bons résultats obtenus par LDOF, nous considérons leur modèle trop restrictif. L'équation (2) nécessite que les correspondances soient bien localisées (pour être traduites en déplacement *a priori*) et qu'elles soient associées à un score pour minimiser l'impact des fausses correspondances. Cela limite grandement les types de primitives et de descripteurs utilisables. De plus, cela introduit un couplage indésirable entre le descripteur et le reste de l'algorithme. L'idée de départ consistait à laisser l'optimisation variationnelle "choisir" parmi les voisins grâce au terme direct et à la fonction de coût robuste. Cependant leur conclusion (Brox, Malik, 2011) est qu'incorporer seulement le plus proche voisin donne de meilleurs résultats que d'incorporer plusieurs coûts incohérents. Nous proposons une approche généralisée qui relâche ces contraintes pour une meilleure compatibilité avec les techniques actuelles et futures de mise en correspondance de primitives, voir la section 3.3.

3. Formulation proposée

Nous utilisons un modèle variationnel construit autour de la régularisation variation totale généralisée (Bredies *et al.*, 2010) à l'ordre 2 (TGV²), récemment introduite, qui

favorise des champ affines par morceaux et est détaillée en section 3.1. Plusieurs coûts denses peuvent être utilisés, dont trois sont listés en section 3.2. Nous ajoutons ensuite un coût basé correspondances de primitives, expliqué en section 3.3 et obtenons la forme suivante :

$$C(\mathbf{u}) = \lambda C_{\text{dense}}(\mathbf{u}) + \text{TGV}^2(\mathbf{u}, \alpha_0, \alpha_1) \quad (3)$$

$$+ \beta \sum_{i=1}^N C_{\text{corr}}(\mathbf{u}, \mathcal{F}_i).$$

3.1. Variation totale généralisée

3.1.1. Définition

Les composantes u_x et u_y du flot optique sont régularisées indépendamment. Dans cette section, u désigne au choix l'un de ces champs scalaires. La variation totale $\text{TV}(u) = \int_{\Omega} |\nabla u(\mathbf{x})| \, d\mathbf{x}$ est l'une des régularisations les plus utilisées grâce à sa formulation simple et au développement d'algorithmes efficaces pour la norme L^1 qui permet la préservation des discontinuités. Cependant, en privilégiant des champs constants par morceaux, cette régularisation a tendance à introduire des effets d'escalier lorsque les variations sont progressives. Plusieurs alternatives ont été proposées telles que la régularisation de Huber (Werlberger *et al.*, 2009) mais nous nous concentrons ici sur la variation totale généralisée (Bredies *et al.*, 2010) qui étend la variation totale aux dérivées plus élevées. Elle est définie dans l'espace dual de Legendre-Fenchel mais la formulation primale permet de faire le lien avec la variation totale. En particulier, aux ordres un et deux :

$$\text{TGV}_{\alpha}^1(u) = \alpha \text{TV}(u) = \alpha \int_{\Omega} |\nabla u(\mathbf{x})| \, d\mathbf{x}$$

$$\text{TGV}_{\alpha}^2(u) = \min_{\mathbf{w} \in \mathcal{W}} \left\{ \alpha_1 \int_{\Omega} |\nabla u - \mathbf{w}| \, d\mathbf{x} + \alpha_0 \int_{\Omega} |\nabla \mathbf{w}| \, d\mathbf{x} \right\}.$$

TGV^1 tend à produire des solutions constantes par morceaux alors que TGV^2 favorise les solutions affines par morceaux, plus adaptées pour l'estimation de flot optique ou de disparités stéréoscopiques.

3.1.2. Optimisation TGV^2 par primal-dual

La méthode primal-dual de Chambolle & Pock (Chambolle, Pock, 2011), très efficace, peut être utilisée pour optimiser le modèle TGV^2 discrétisé. Le lecteur intéressé pourra se référer à (Bredies, 2012 ; Ranftl *et al.*, 2012) pour une introduction détaillée à l'algorithme avec les informations pratiques telles que les versions discrètes des opérateurs de différentiation. Une approche pyramidale (facteur de décimation $s \in [0.5, 1]$) permet d'accélérer la convergence, de limiter les minima locaux et, comme nous l'avons vu, d'intégrer de manière élégante les correspondances éparses.

3.2. Coût dense basé image

La minimisation du coût dense est basée sur une linéarisation itérative. N'importe quelle fonction de coût définie sur tout le domaine image Ω et aux variations suffisamment progressives peut être utilisée. Il est par contre important de prendre en compte la complexité du calcul du coût étant donné qu'il sera évalué des centaines de fois par pixel.

$$C_{AD}(\mathbf{x}, \mathbf{u}) = |I_0(\mathbf{x}) - I_1(\mathbf{x} + \mathbf{u}(\mathbf{x}))| \quad (4)$$

La différence absolue des intensités (4) est la plus simple des fonctions de coût et la plus utilisée. Elle est robuste aux déformations de l'image. Cependant l'hypothèse implicite de constance de l'intensité est très rarement respectée en pratique.

$$C_{Census}(\mathbf{x}, \mathbf{u}) = \Delta(C(I_0, \mathbf{x}), C(I_1, \mathbf{x} + \mathbf{u}(\mathbf{x}))) \quad (5)$$

La transformée Census (Zabih, Woodfill, 1994), dont nous utilisons la variante ternaire (Ranftl *et al.*, 2012), encode la structure locale d'une fenêtre de taille fixe. Dans l'équation (5), Δ est la distance de Hamming et $C(I, \mathbf{x})$ la transformée Census de l'image I au pixel \mathbf{x} . Elle est robuste à tout changement de luminosité monotone. En revanche la discrétisation des différences de pixel introduit une perte en précision. Une grande fenêtre rend la transformée plus discriminante mais moins robuste aux distorsions et plus lente à calculer.

$$C_{ADC}(\mathbf{x}, \mathbf{u}) = 2 - \exp\left(-\frac{C_{AD}}{\mu_0}\right) - \exp\left(-\frac{C_{Census}}{\mu_1}\right) \quad (6)$$

L'approche AD-Census (Mei *et al.*, 2011) combine les deux précédentes (6) pour être à la fois robuste et présenter des minima bien localisés.

3.2.1. Gestion des occultations

Les occultations doivent être prises en compte pour l'estimation de champs de déplacements non triviaux. Deux types d'occultations peuvent être dégagées : les occultations externes et les auto-occultations. Les occultations externes sont causées par la présence d'un objet occultant dans une seule des deux images. Elles sont gérées en tronquant le terme de donnée au dessus d'un seuil choisi, ce qui empêche les incohérences d'affecter la solution finale. Empiriquement, ce seuil est peu sensible et une valeur de 50 % du maximum du terme de donnée améliore notablement la robustesse sans dégrader la précision. Les auto-occultations apparaissent quand une scène rigide est observée depuis des points de vue éloignés ou quand une surface déformable se replie sur elle même. Suivant Gay-Bellile *et al.* (Gay-Bellile *et al.*, 2010), il est possible de les détecter en utilisant le fait que la dérivée du champ de correspondances s'annule dans une direction. Cette observation faite pour les surfaces déformables est également valide pour les scènes rigides et nous permet d'obtenir une probabilité d'occultation $\mathcal{P}_{occ}(\mathbf{x})$. Le coût dense n'a pas de sens dans les zones

occultées, il est donc multiplié par $1 - \mathcal{P}_{\text{occ}}$ avant de l'inclure dans la fonction de coût globale (3).

3.3. Coût basé correspondances éparses

Comme expliqué en section 2, notre fonction de coût est inspirée par celle utilisée par LDOF (2) mais plus souple et générale :

$$C_{\text{corr}}(\mathbf{u}, \mathcal{F}_i) = \int_{\Omega} \rho_i(\mathbf{x}) \Gamma_{\sigma} [D_f(\mathbf{x} + \mathbf{u}(\mathbf{x}), \mathcal{F}_i)] \, d\mathbf{x} \quad (7)$$

où ρ_i est l'influence de la correspondance, Γ_{σ} est un estimateur robuste, D_f est la distance principale associée à la correspondance. La suite de la section apporte une explication détaillée pour chaque fonction.

3.3.1. Influence ρ_i .

Notre approche n'est pas restreinte à une grille régulière de descripteurs, et la plupart des détecteurs de primitives produisent des coordonnées non entières. Nous traduisons cette propriété en considérant qu'un point affecte les 4 pixels les plus proches de lui. Le poids associé à chaque pixel voisin dépend linéairement de la distance entre son centre et le point considéré, à la manière d'une interpolation bilinéaire. Soit un point d'intérêt i localisé en $\mathbf{x}_f = \mathbf{x}_{f_0} + \mathbf{dx}$, $\mathbf{x}_{f_0} = \text{partie entière}(\mathbf{x}_f)$, $\overline{\mathbf{dx}} = (1, 1)^T - \mathbf{dx}$, son influence ρ_i est définie pour les quatre pixels affectés par :

$$\begin{aligned} \rho_i(\mathbf{x}_{f_0}) &= \overline{\mathbf{dx}}_x \overline{\mathbf{dx}}_y \\ \rho_i(\mathbf{x}_{f_0} + (0, 1)^T) &= \overline{\mathbf{dx}}_x \mathbf{dx}_y \\ \rho_i(\mathbf{x}_{f_0} + (1, 0)^T) &= \mathbf{dx}_x \overline{\mathbf{dx}}_y \\ \rho_i(\mathbf{x}_{f_0} + (1, 1)^T) &= \mathbf{dx}_x \mathbf{dx}_y. \end{aligned}$$

Les segments sont considérés comme un ensemble de points pour le calcul de leur influence. Celle-ci correspond donc à une représentation anticrénelée du segment.

3.3.2. Fonction de coût robuste Γ_{σ}

Contrairement à LDOF (Brox, Malik, 2011), notre approche ne fait pas appel à une mesure externe de la qualité des correspondances, souvent peu fiable ou même indisponible. L'hypothèse utilisée à la place est qu'une correspondance est correcte si elle est cohérente avec le mouvement global. Pour ceci nous utilisons l'estimateur non convexe de Geman McClure $\Gamma_{\sigma}(x) = \frac{x^2}{\sigma + x^2}$ dont l'influence $\frac{d\Gamma_{\sigma}(x)}{dx} \propto \frac{x}{(\sigma + x^2)^2}$ tend vers zéro pour des valeurs de x importantes. Nous choisissons une faible valeur $\sigma = 0.2$ pour un filtrage fort. À basse résolution, plusieurs correspondances sont affectées à chaque pixel, le processus peut alors être vu comme un vote dont les "perdants" voient leur influence durablement diminuer. À plus haute résolution, le coût dense et la régularisation prennent le relais pour converger vers l'optimum visé.

3.3.3. Distances associées aux correspondances D_f

Contrairement à LDOF (2), C_{corr} dépend de la distance entre $\mathbf{x} + \mathbf{u}(\mathbf{x})$ et la primitive correspondante, sans convertir la correspondance en déplacement *a priori*. Cette distinction permet d'utiliser différentes distances pour incorporer des primitives. Nous définissons ici les distances associées aux correspondances de points ou de segments. La méthode est aisément transposable à d'autres types de correspondances.

La mise en correspondance de points est un domaine de recherche mature et toujours actif. Un large choix de descripteurs permet de trouver le meilleur compromis selon l'application visée entre rapidité et robustesse. L'utilisation de détecteurs de points d'intérêt permet d'obtenir des localisations plus précises que la grille régulière proposée par (Brox, Malik, 2011).

On présente ici deux distances, représentées sur la figure 1, pour les points et pour les segments. La distance adaptée aux points est la norme euclidienne. Pour un point d'intérêt $\mathcal{F}_i = \mathbf{x}_i$:

$$D_f^{(\text{point})}(\mathbf{x}, \mathcal{F}_i) = \|\mathbf{x} - \mathbf{x}_i\|. \quad (8)$$

Comme affirmé dans (Brox, Malik, 2011), l'influence des correspondances décroît naturellement à chaque niveau de la pyramide. En effet, l'influence des correspondances est liée au nombre de pixels affectés. A chaque étape de décimation avec un facteur $s < 1$, la proportion de l'image couverte par un pixel est multipliée par $s^2 < 1$.

L'algorithme de mise en correspondance de segments (Wang *et al.*, 2009) est intéressant car il ne s'appuie pas sur des similarités photométriques mais représente la structure géométrique semi-globale et est robuste à d'importantes distorsions perspectives. Un exemple de mise en correspondance est montré sur la figure 2. Les segments mis en correspondance appartiennent à la même ligne 3D mais leurs extrémités ne correspondent pas forcément. La distance appropriée à utiliser est donc la distance point-ligne orthogonale, qui contraint une seule dimension. Soit un segment défini par ses extrémités $\mathcal{F}_i = (\mathbf{x}_{i_b}, \mathbf{x}_{i_e})$:

$$D_f^{(\text{segment})}(\mathbf{x}, \mathcal{F}_i) = \frac{\|(\mathbf{x}_{i_e} - \mathbf{x}_{i_b}) \times (\mathbf{x} - \mathbf{x}_{i_b})\|}{\|\mathbf{x}_{i_e} - \mathbf{x}_{i_b}\|}. \quad (9)$$

Après décimation, la surface de l'image affectée par les segments ne diminue que dans une seule dimension. Pour que leur influence ait la même évolution que celle des points, la fonction d'influence ρ_i est multipliée par s à chaque niveau de pyramide.

3.3.4. Correspondances ponctuelles *a priori*

Dans le cas d'amplitudes importantes de mouvement entre les deux images, le degré de liberté restant ne peut pas être laissé complètement libre. Nous introduisons le concept de correspondance ponctuelle *a priori* basé sur l'hypothèse d'une correspondance linéaire entre les segments. Soit un segment $\mathcal{F}_i^{(0)} = (\mathbf{x}_{i_b}^{(0)}, \mathbf{x}_{i_e}^{(0)})$ dans I_0 et

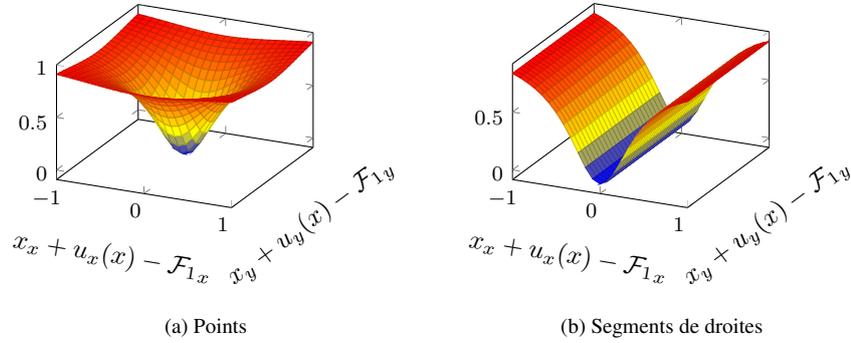


Figure 1. Distances point-primitive avec l'estimateur de Geman McClure : $\Psi_\sigma \circ D$. Les graphes représentent les distances pour un pixel $x = (x_x, x_y) \in I_1$, associé à un déplacement $u(x) = (u_x(x), u_y(x))$ et deux primitives en correspondance $(\mathcal{F}_0, \mathcal{F}_1)$



Figure 2. Exemple de correspondances de segments, I_0 à gauche, I_1 à droite (à voir en couleur)

sa correspondance $\mathcal{F}_i^{(1)} = (\mathbf{x}_{i_b}^{(1)}, \mathbf{x}_{i_e}^{(1)})$ dans I_1 , la correspondance ponctuelle *a priori* $\mathbf{x}_{\text{ap}}^{(1)}$ du point $\mathbf{x}^{(0)} \in \mathcal{F}_i^{(0)}$ est définie par :

$$t = \frac{\langle (\mathbf{x}^{(0)} - \mathbf{x}_{i_b}^{(0)}), (\mathbf{x}_{i_e}^{(0)} - \mathbf{x}_{i_b}^{(0)}) \rangle}{\|\mathbf{x}_{i_e}^{(0)} - \mathbf{x}_{i_b}^{(0)}\|^2}$$

$$\mathbf{x}_{\text{ap}}^{(1)} = \mathbf{x}_{i_b}^{(1)} + t \cdot (\mathbf{x}_{i_e}^{(1)} - \mathbf{x}_{i_b}^{(1)})$$

et la distance associée est :

$$D_{\text{ap}}^{(\text{segment})}(\mathbf{x}, \mathbf{u}(\mathbf{x}), \mathcal{F}_i) = \|\mathbf{x} + \mathbf{u}(\mathbf{x}) - \mathbf{x}_{\text{ap}}^{(1)}\|. \quad (10)$$

L'hypothèse est vérifiée seulement dans le cas de segments fronto-parallèles dont les extrémités correspondent parfaitement. Cependant, c'est la plupart du temps peu éloigné de la vérité et cela peut servir comme contrainte faible pour guider le flot optique à basse résolution. Comme on peut le voir dans le tableau 1, les correspondances ponctuelles ne

*Tableau 1. Influence des correspondances ponctuelles a priori sur la proportion de profondeurs correctes estimées (erreur inférieure à 5 % de la fourchette totale) avec les correspondances de segments. Les résultats pour distorsions faibles sont obtenus avec les images 1 – 2 du jeu de données *herzjesu*. Les résultats pour fortes distorsions sont obtenus avec les images 6 et 1. Détails en section 3.3.4.*

c		Faibles distorsions	Fortes distorsions
0	avec	92.6%	0%
0.5		93.9%	41.7%
1	sans	93.3%	41.7%

sont pas fiables et dégradent les résultats que ce soit pour des disparités importantes ou non. Cependant, avec un coefficient de pondération $c = 0.5$, les résultats sont meilleurs dans le cas des disparités faibles, sans dégrader le cas des disparités élevées. Notons enfin que cette approche peut facilement être étendue à d'autres primitives locales telles que les régions et les contours.

4. Résultats expérimentaux

Cette section est consacrée à la démonstration de la validité et de la flexibilité de notre approche sur différents jeux de données : du flot optique entre deux images proches jusqu'au recalage de surfaces déformables en passant par de la stéréovision à fortes disparités. Sauf avis contraire, les paramètres utilisés sont $\lambda = 6$, $\beta = 0.5$, $\alpha_0 = 4$, $\alpha_1 = 1$, 20 itérations externes, 40 itérations internes et un facteur de décimation $s = 0.8$. Le coût dense utilisé est AD-Census avec une fenêtre 3×3 , $\mu_0 = 1$, $\mu_1 = 0.25$. Les correspondances de points sont effectuées par le détecteur (Rosten *et al.*, 2010) et le descripteur SIFT (Lowe, 2004) implémentés dans la bibliothèque OpenCV (Bradski, 2000), en gardant les paramètres par défaut pour faciliter la reproductibilité. Un filtrage simple est effectué (*cross-check filter*) : les correspondances sont calculées de la première image vers la deuxième et de la deuxième vers la première ; seules les correspondances cohérentes sont conservées pour supprimer les ambiguïtés les plus évidentes.

4.1. Flot optique 2D

4.1.1. Jeu de données KITTI

Nous évaluons notre méthode sur le jeu de données KITTI (Geiger *et al.*, 2012), composé d'images réelles capturées depuis un véhicule. Ces images présentent des difficultés telles que la présence de spéularités, le fait que la majorité des surfaces soient parallèles à l'axe optique et une grande variabilité dans l'amplitude des déplacements et des conditions de luminosité. À ce jour, notre méthode obtient la première

Tableau 2. Résultats sur KITTI. Les méthodes 1 à 4 utilisent la stéréo soit en utilisant les paires stéréos fournies, soit en utilisant les poses des caméras pour contraindre le mouvement à une seule direction. Notre méthode est la meilleure parmi les méthodes de flot optique pur

R	Méthode	Out-Noc	Out-All	Avg-Noc	Avg-All	Time
1	PR-Sf+E	4.08 %	7.79 %	0.9 px	1.7 px	200 s
2	PCBP-Flow[ms]	4.08 %	8.70 %	0.9 px	2.2 px	3 min
3	MotionSLIC[ms]	4.36 %	10.91 %	1.0 px	2.7 px	11 s
4	PR-Sceneflow	4.48 %	8.98 %	1.3 px	3.3 px	150 s
5	TGV2ADCSIFT	6.55 %	15.35 %	1.6 px	4.5 px	12 s (GPU)
6	Data-Flow	8.22 %	15.78 %	2.3 px	5.7 px	3 min
7	fSGM	11.03 %	22.90 %	3.2 px	12.2 px	60 s
8	TGV2CENSUS	11.14 %	18.42 %	2.9 px	6.6 px	4 s (GPU)

place du classement public parmi les méthodes de flot optique pur¹. Notre méthode améliore donc l'état de l'art même pour de faibles déplacements où les minima locaux sont moins problématiques. Pour analyser l'impact de nos contributions, nous nous basons sur la sélection "large displacements"² (*Special Session on Robust Optical Flow*, 2013). La figure 3 compare notre méthode à LDOF (Brox, Malik, 2011), notre inspiration pour l'intégration des correspondances, et TGV2CENSUS (Ranftl *et al.*, 2012) qui utilise une régularisation TGV^2 et un terme de données Census. Il apparaît que la régularisation TGV^2 est l'amélioration la plus importante. Ensuite nos paramètres : en diminuant α_0 , la régularisation est moins contrainte. AD-Census apporte un gain en précision sauf pour la trame 181. Même dans ce contexte de faibles déplacements, la gestion des auto-occultations et des correspondances de point apportent une amélioration substantielle. La mise en correspondance OpenCV FAST-SIFT calcule environ 4000 correspondances par paire d'images en 4 secondes sur un processeur Intel Xeon 6 × 2.40GHz et l'estimation variationnelle prend en moyenne 8s avec une carte graphique NVidia GeForce GTX 460.

4.1.2. Jeu de données Middlebury

Le jeu de données Middlebury (Baker *et al.*, 2011) a été pendant longtemps la référence. Cependant, les paires d'images proposées sont issues d'un environnement très contrôlé et ne reflètent pas les cas observés en situations réelles. Nous l'utilisons quand même pour montrer que la présence de fausses correspondances dans notre méthode ne dégrade pas le résultat. Nous utilisons des correspondances SURF (Bay *et al.*, 2006) non filtrées, de très mauvaise qualité comme on peut le voir sur la figure 4. La régularisation TGV^2 n'est pas adaptée aux paires d'images Middlebury où le mouvement est constant par morceaux (la figure 5 en offre une illustration) donc

1. Les méthodes mieux classées utilisent la connaissance de la pose des caméras (stéréo monoculaire) ou les paires stéréo associées aux image (*scene flow*).

2. Les déformations de cette sélection "large displacement" restent de faible amplitude par rapport aux expériences suivantes.

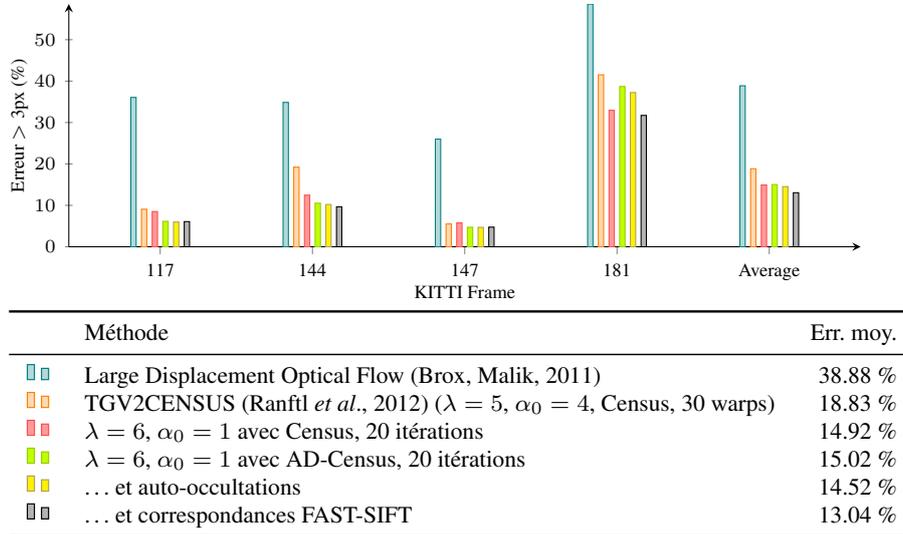


Figure 3. Évaluation quantitative sur la sélection "large déplacements" des images du jeu de données KITTI

Méthode	Dimetrodon	Grove2	Grove3	Hydrangea	RubberWhale	Urban2	Urban3	Venus	Average
LDOF (Brox, Malik, 2011)	0.12	0.18	0.70	0.18	0.13	0.38	0.82	0.38	0.36
Notre méthode	0.12	0.18	0.71	0.18	0.13	0.46	0.60	0.26	0.33
... avec correspondances	0.13	0.18	0.72	0.18	0.13	0.45	0.59	0.26	0.33

Figure 4. Évaluation de l'erreur moyenne en pixels sur le jeu de donnée Middlebury (avec un exemple des correspondances de mauvaise qualité utilisées à droite)

nous utilisons la régularisation TV dans cette expérience en choisissant $\alpha_0 = \infty$. Les résultats sont listés dans le tableau de la figure 4.

4.2. Stéréovision à fortes disparités

La robustesse accrue de notre approche permet d'explorer de nouvelles applications telles que la stéréo en présence de fortes disparités, jusqu'ici impossible à traiter avec des approches de type flot optique. Tola *et al.* (2010) ont publié un jeu de données intéressant pour démontrer les qualités de leur descripteur dense DAISY. Nous comparons notre méthode à leur approche utilisant une optimisation discrète sur le jeu de données *herzjesu* en utilisant des correspondances de segments (Wang *et al.*, 2009) et un poids $\gamma = 5$ tout en projetant les déplacements sur les lignes épipolaires. Nous adoptons la même disposition en matrice que dans leur publication pour faciliter la comparaison en figure 6 et tableau 3. Pour des distorsions extrêmes, le terme de donnée AD-Census montre ses limites mais il est clair que les segments permettent d'élargir le bassin de convergence et d'obtenir des résultats comparables sur la plupart

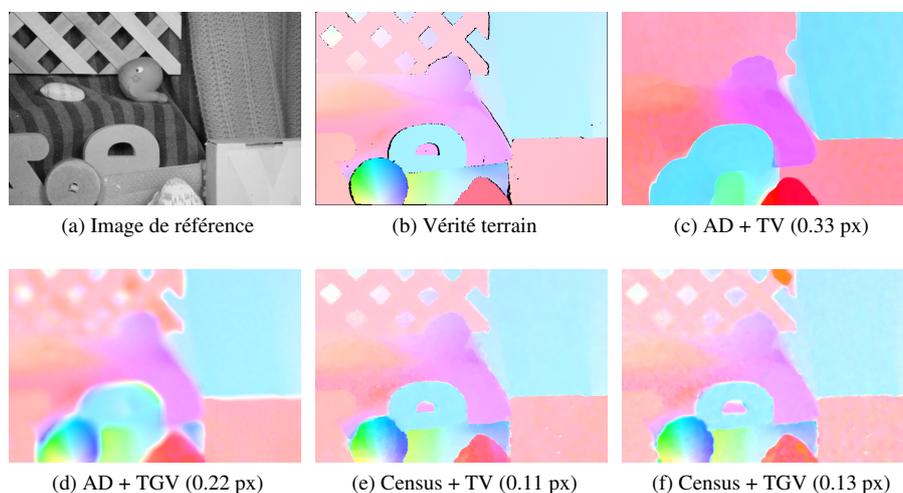


Figure 5. Résultats qualitatifs et erreur moyenne en pixels sur la paire d'image RubberWhale du jeu de données Middlebury avec différents termes de données et régularisations. La régularisation TV est plus adaptée à ces images, à condition que le terme de données utilisé soit assez discriminant pour bien estimer les contours d'occultations : les meilleurs résultats sur ce jeu de données sont ainsi obtenus avec le couple TV + Census et les moins bons résultats avec le couple TV + AD

Tableau 3. Évaluation quantitative sur le jeu de données DAISY *herzjesu*. Le tableau montre le pourcentage d'erreur supérieure à 5 % de l'amplitude totale des profondeurs pour, dans l'ordre : notre algorithme sans correspondances / notre algorithme avec correspondances de segments / Graph-Cut et DAISY. La disposition est la même que dans la figure 6

	1	2	3	4	5
1	-	96.4 / 90.9 / 90.8	85.2 / 86.3 / 88.7	86.6 / 84.8 / 86.9	85.1 / 85.0 / 87.5
2	93.3 / 93.3 / 90.8	-	93.9 / 92.6 / 94.3	78.2 / 91.4 / 92.4	88.1 / 86.7 / 91.4
3	83.6 / 83.2 / 86.9	90.8 / 88.9 / 90.3	-	91.4 / 91.7 / 93.2	89.3 / 88.0 / 93.9
4	80.9 / 73.2 / 85.5	88.4 / 84.8 / 87.1	92.3 / 92.2 / 93.0	-	90.2 / 90.8 / 95.4
5	00.0 / 73.6 / 83.6	76.1 / 74.3 / 86.3	86.9 / 85.4 / 91.7	89.9 / 89.9 / 93.3	-

des images. De plus, notre approche continue variationnelle est plus rapide que les approches discrètes. On remarque également que la gestion des auto-occultations, issue du recalage de surface déformable (Gay-Bellile *et al.*, 2010), s'applique très bien dans le cas de déformations rigides.

La figure 7 contient une comparaison quantitative entre la méthode dite « de base », c'est-à-dire sans terme basé correspondances éparses, la méthode proposée, et les résultats annoncés par (Tola *et al.*, 2010) avec la méthode DAISY. La première conclusion à tirer est que l'ajout de notre terme basé correspondances de segments élargit grande-

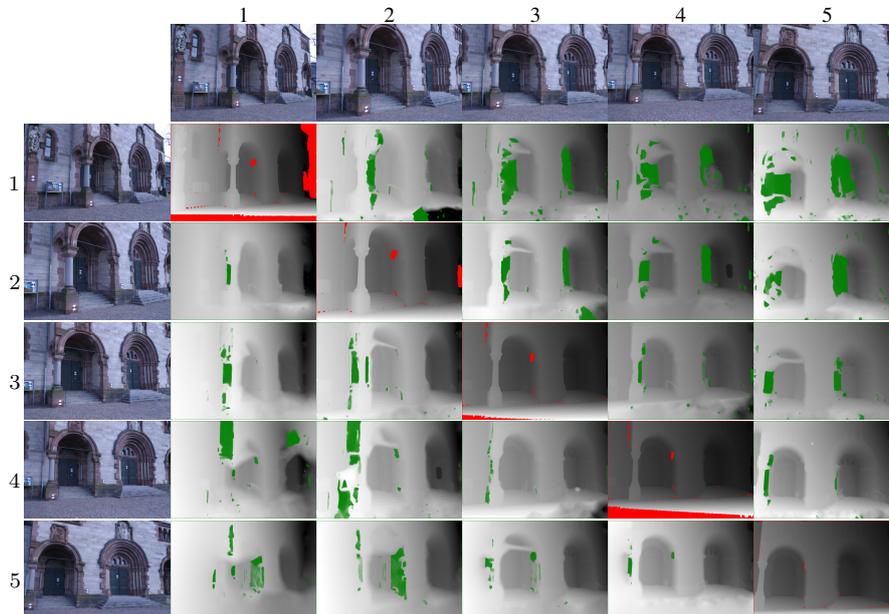


Figure 6. Évaluation sur le jeu de données DAISY herzjesu. Les cartes de profondeurs sur la diagonale sont les vérités terrains. Les autres cartes de profondeur sont calculées en utilisant l'image de la ligne et celle de la colonne comme référence. Les occultations estimées sont coloriées en vert. Voir le tableau 3 pour les résultats quantitatifs

ment le bassin de convergence et divise l'erreur moyenne par presque 5. La deuxième conclusion est que nos résultats étant à moins de 2 % de ceux de DAISY, notre approche permet de réutiliser des méthodes génériques pour des problèmes spécifiques.

Deux exemples qualitatifs, sans vérité terrain, viennent illustrer le bon fonctionnement de la méthode avec des correspondances de segments en figure 8.

4.2.1. Alignement de surfaces déformables

Si des méthodes paramétriques sont plus adaptées au problème d'alignement de surfaces déformables, nous vérifions la généralité de l'approche non paramétrique par quelques résultats qualitatifs en figure 9. Nous comparons nos résultats à la méthode FBDS (Pizarro, Bartoli, 2012), constituant l'état de l'art des méthodes basées correspondances de primitives. Nous utilisons l'implémentation publique C++ de (Alcantarilla, Bartoli, 2012) et utilisons les mêmes correspondances SURF (Bay *et al.*, 2006) en entrée de notre algorithme.

La déformation est estimée en utilisant le modèle plan comme image de référence I_1 car le champ de déplacements est ainsi défini sur toute l'image ce qui facilite l'estimation. Le champ de déplacements est ensuite inversé et appliqué à une grille colorée

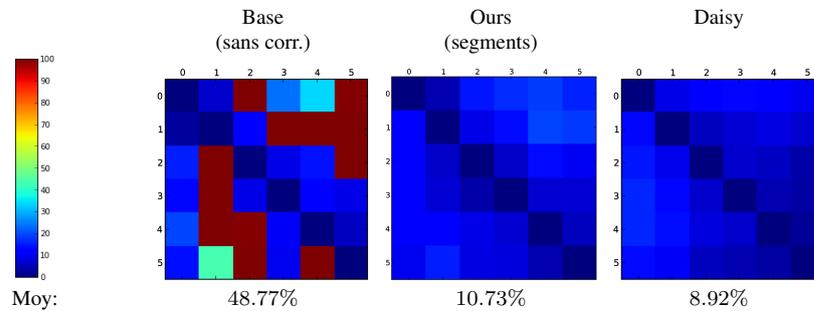


Figure 7. Proportion de profondeurs erronées sur le jeu de données herzjesu. Comme (Tola 2010), nous considérons qu'une profondeur est erronée si l'erreur par rapport à la vérité terrain est supérieure à 5 % de la variation de profondeur sur l'image (différence entre profondeurs maximale et minimale). La première ligne est une représentation en couleur de l'erreur pour chaque couple d'images, la deuxième ligne affiche l'erreur moyenne sur tous les couples

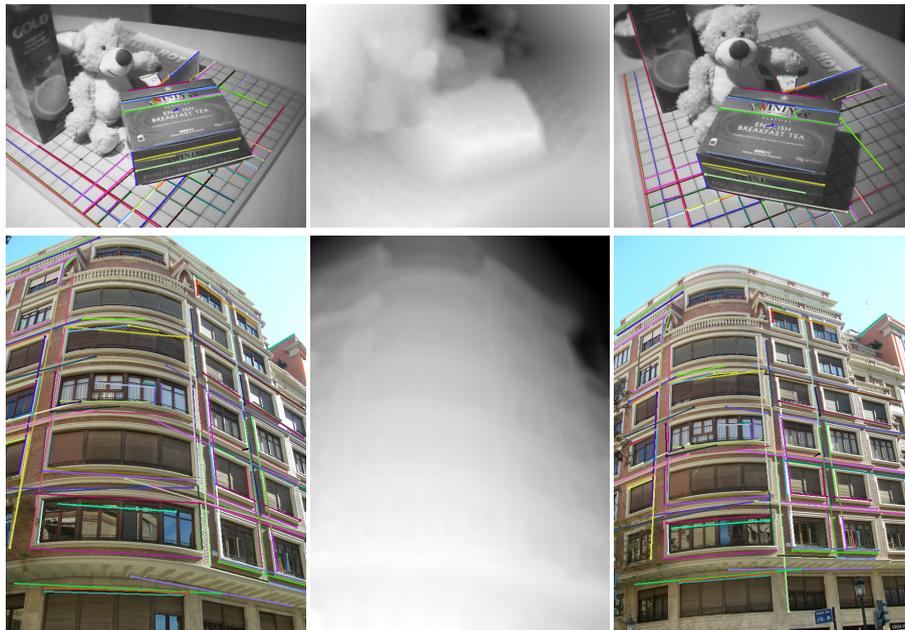


Figure 8. Exemples d'estimation de profondeur par stéréo-vision. De gauche à droite : image de référence, carte de profondeur, deuxième image. Respectivement 81 et 111 correspondances de segments ont été utilisées

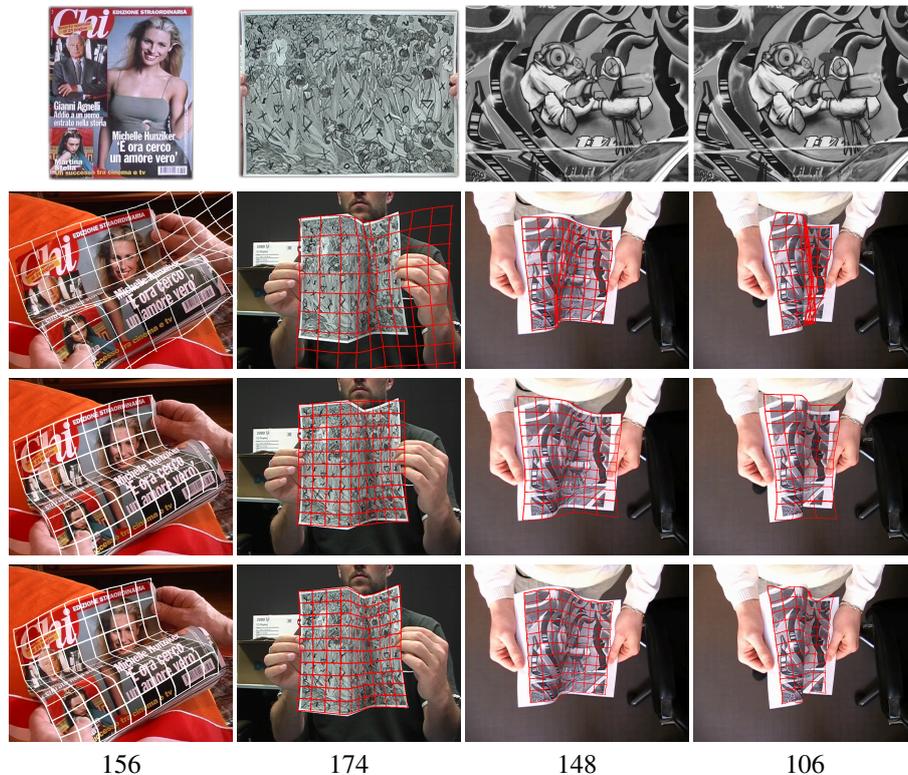


Figure 9. Exemples d'alignement de surfaces déformables avec la méthode non paramétrique. De haut en bas, modèles plans, résultats de FBDS (Pizarro 2012), résultats de notre méthode sans terme dense ($\lambda = 0$), résultats de notre méthode avec $\lambda = 1$, nombre de correspondances SURF. De gauche à droite, les paires d'images proviennent des publications suivantes : Ferrari 2004, Salzmann 2007 et GB 2010

qui est superposée à l'image de la surface déformée. Les importantes déformations sont susceptibles d'introduire de nombreux minimums locaux pour le terme dense, nous réduisons donc son influence avec $\lambda = 1$. Pour évaluer l'apport du terme dense, nous effectuons également une estimation en supprimant totalement son influence avec $\lambda = 0$.

Les résultats montrent que notre terme basé correspondances de primitives permet de rendre les approches variationnelles classiques viables pour de tels problèmes. Les déformations sont bien estimées et les auto-occultations correctement gérées. Même sans le terme dense, nos résultats sont meilleurs que FBDS pour les deux premières paires, ce qui prouve la supériorité du filtrage implicite. Le terme dense apporte toutefois des gains significatifs. Quelques défauts sont toujours présents mais sont surtout dûs à la régularisation.

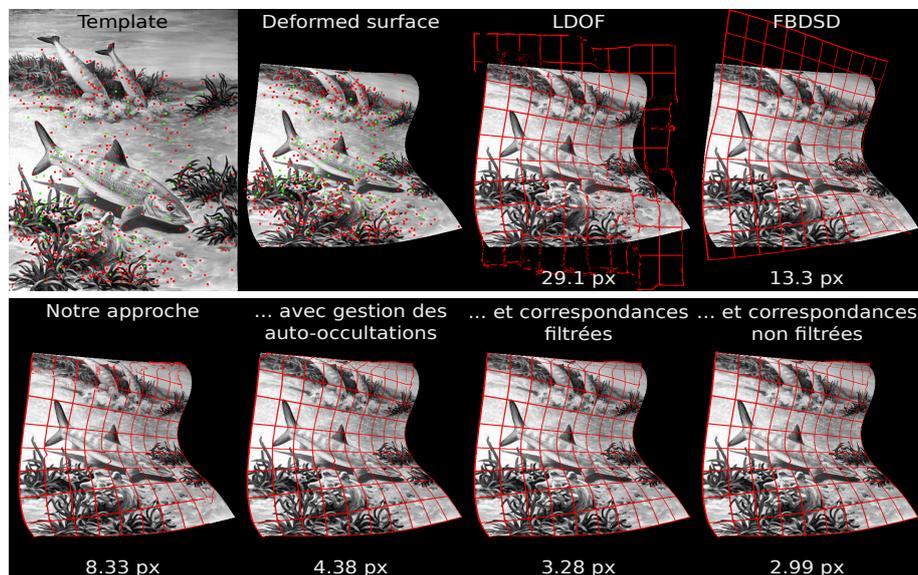


Figure 10. Résultats de la détection de surface déformable sur des données synthétiques. Les 613 correspondances non filtrées et les 103 correspondances filtrées (en vert) sont affichées sur la paire d'image utilisée. Les grilles représentent la transformation inverse estimée par chaque méthode. L'erreur moyenne en pixels est également indiquée

4.3. Détection de surfaces déformables

Un autre domaine intéressant qui met en échec les méthodes variationnelles denses est la détection de surfaces déformables. Étant donné un modèle plan d'une surface texturée et une image de cette même surface déformée, le problème est d'estimer la nouvelle position de chaque pixel de la surface. Les méthodes denses existantes doivent être initialisées proche de la solution pour converger (Pizarro, Bartoli, 2012), ce qui est habituellement réalisé par une étape préliminaire basée correspondances de points d'intérêt. Les points sont d'abord filtrés pour supprimer les fausses correspondances, puis un modèle de déformation est ajusté à ces correspondances.

Notre approche permet d'utiliser conjointement dans une même étape d'optimisation l'information dense et les correspondances de primitives. Pour démontrer quantitativement les avantages de cette méthode, nous générons une déformation synthétique grâce à la boîte à outils Matlab de (Perriollat, Bartoli, 2012). Les correspondances de point sont obtenues avec le détecteur et le descripteur SIFT. Notre méthode est comparée à l'approche basée correspondances puis raffinement FBDS (Pizarro, Bartoli, 2012) ainsi qu'à l'approche flot optique robuste LDOF (Brox, Malik, 2011). On observe dans la figure 10 que LDOF ne peut estimer les déplacements trop importants, et que les approches basées correspondances souffrent d'une faible précision près des

bords de l'image où il y a peu de correspondances. Nos résultats sont meilleurs en utilisant les correspondances non filtrées, ce qui révèle que le filtrage a supprimé de l'information utile. Notre filtrage implicite pendant l'optimisation par l'emploi d'un estimateur robuste est donc préférable à une étape séparée de filtrage.

5. Conclusion

Nous avons introduit un cadre général permettant de largement étendre le domaine d'applicabilité des méthodes variationnelles d'estimation du flot optique. Nous avons combiné une régularisation moderne permettant de préserver les discontinuités avec un terme direct robuste et l'intégration de correspondances de primitives. Notre modèle est compatible avec des primitives faiblement localisées telles que les segments et est robuste aux fausses correspondances. La gestion des auto-occultations et des occultations externes améliore encore la robustesse de notre approche. Ces contributions permettent d'élargir considérablement le bassin de convergence des méthodes variationnelles d'estimation de flot optique : nous obtenons des résultats compétitifs avec l'état de l'art pour le recalage 1D ou 2D, rigide ou flexible, avec des déformations d'amplitudes faibles ou fortes. Pour la suite, nous considérerons l'amélioration de chacun des blocs élémentaires : régularisation d'ordre plus élevé, terme de donnée plus riche et nouvelles primitives : régions ou contours.

Remerciements

Ce travail a été financé en partie par la bourse de recherche ERC 307483 du programme FP7 de l'union européenne.

Bibliographie

- Alcantarilla P. F., Bartoli A. (2012). Deformable 3d reconstruction with an object database. In *British machine vision conference (bmvc)*, p. 1–12.
- Baker S., Scharstein D., Lewis J., Roth S., Black M., Szeliski R. (2011). A database and evaluation methodology for optical flow. *International Journal of Computer Vision (IJCV)*.
- Bay H., Tuytelaars T., Van Gool L. (2006). SURF: Speeded up robust features. In *European conference on computer vision (eccv)*.
- Bradski G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*.
- Bredies K. (2012). Recovering piecewise smooth multichannel images by minimization of convex functionals with total generalized variation penalty. *SFB Report*, vol. 6.
- Bredies K., Kunisch K., Pock T. (2010). Total generalized variation. *SIAM Journal on Imaging Sciences*.
- Brox T., Malik J. (2011). Large displacement optical flow: descriptor matching in variational motion estimation. *International Journal of Pattern Analysis and Machine Intelligence (PAMI)*.
- Chambolle A., Pock T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*.

- Gay-Bellile V., Bartoli A., Sayd P. (2010). Direct estimation of nonrigid registrations with image-based self-occlusion reasoning. *International Journal of Pattern Analysis and Machine Intelligence (PAMI)*.
- Geiger A., Lenz P., Urtasun R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Computer vision and pattern recognition (cvpr)*.
- Hartley R., Zisserman A. (2000). *Multiple view geometry in computer vision* (vol. 2). Cambridge Univ Press.
- Horn B. K. P., Schunck B. G. (1981). Determining optical flow. *Artificial Intelligence*.
- Lowe D. G. (2004, novembre). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, vol. 60, n° 2, p. 91–110.
- Mei X., Sun X., Zhou M., Jiao S., Wang H., Zhang X. (2011). On building an accurate stereo matching system on graphics hardware. In *Third iccv workshop on gpus for computer vision*.
- Perriollat M., Bartoli A. (2012). A computational model of bounded developable surfaces with application to image-based three-dimensional reconstruction. *Computer Animation and Virtual Worlds*.
- Pilet J., Lepetit V., Fua P. (2008). Fast non-rigid surface detection, registration and realistic augmentation. *International Journal of Computer Vision (IJCV)*.
- Pizarro D., Bartoli A. (2012). Feature-based deformable surface detection with self-occlusion reasoning. *International Journal of Computer Vision (IJCV)*.
- Ranftl R., Gehrig S., Pock T., Bischof H. (2012). Pushing the limits of stereo using variational stereo estimation. In *Intelligent vehicles symposium (iv)*.
- Rosten E., Porter R., Drummond T. (2010). Faster and better: A machine learning approach to corner detection. *International Journal of Pattern Analysis and Machine Intelligence (PAMI)*.
- Special session on robust optical flow*. (2013). (German Conference on Pattern Recognition)
- Tola E., Lepetit V., Fua P. (2010). Daisy: An efficient dense descriptor applied to wide-baseline stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, n° 5.
- Wang L., Neumann U., You S. (2009). Wide-baseline image matching using line signatures. In *International conference on computer vision (iccv)*.
- Werlberger M., Trobin W., Pock T., Wedel A., Cremers D., Bischof H. (2009). Anisotropic Huber-L1 optical flow. In *British machine vision conference (bmvc)*.
- Xu L., Jia J., Matsushita Y. (2010). Motion detail preserving optical flow estimation. In *Computer vision and pattern recognition (cvpr)*.
- Zabih R., Woodfill J. (1994). Non-parametric local transforms for computing visual correspondence. In *European conference on computer vision (eccv)*.

Reçu le 18/11/2014
 Accepté le 2/06/2015

