Super Résolution Temporelle de Formes Multi-Vues

Vincent Leroy, Edmond Boyer, Jean-Sébastien Franco

Super Résolution Temporelle de Formes Multi-Vues

V. Leroy J.S. Franco E. Boyer
INRIA Grenoble
655 Avenue de l'Europe, 38330 Montbonnot-Saint-Martin
vincent.leroy@inria.fr

Résumé

Nous considérons le problème de super résolution temporelle de formes, par l'utilisation de multiples observations d'un même modèle déformé. Sans pertes de généralité, nous nous concentrons plus particulièrement au scénario multi-caméras moyenne échelle, c'est à dire des scènes dynamiques, pouvant contenir plusieurs sujets. Ce contexte favorise l'utilisation de caméras couleur, mais nécessite une méthode de reconstruction robuste aux inconsistances photométriques. Dans ce but, nous proposons une nouvelle approche, spécialement dédiée à ce contexte moyenne échelle, utilisant des descripteurs et des schémas de votes adaptés. Cette méthode est étendue à la dimension temporelle de manière à améliorer les reconstructions à chaque instant, en exploitant la redondance des informations dans le temps. Pour cela, les informations photométriques fiables sont accumulées dans le temps à l'aide de champs de déformations combinés à une stratégie de croissance de région. Nous démontrons l'amélioration des reconstructions apportée par notre approche à l'aide de séquences multi-camera synthétiques.

Mots Clef

Stéréo Multi-Vue, Super Résolution Spatio-Temporelle.

Abstract

We consider the problem of temporally superresolving shape models of deformable objects by exploiting multi-view observations over time. With no loss of generality, we especially focus on mid-scale camera settings i.e. dynamic scenes with one to several people, which can particularly benefit from it. This context favors multi color cameras settings but requires reconstruction algorithms robust to inconsistent photoconsistency measures. To this purpose we introduce a novel approach that specifically addresses the mid-scale setting with adapted features and voting schemes. This approach is extended to the time dimension in order to improve 3D reconstruction by taking advantage of temporal information redundancy. To this goal reliable photometric information is accumulated over time using 3D warps combined with a seed growing strategy for robustness. Our approach provably improves reconstruction accuracy by considering multiple frames, hence time superresolving shape models. To conduct fair evaluations we also introduce a multi-camera synthetic dataset that provides ground-truth data for mid-scale dynamic scenes.

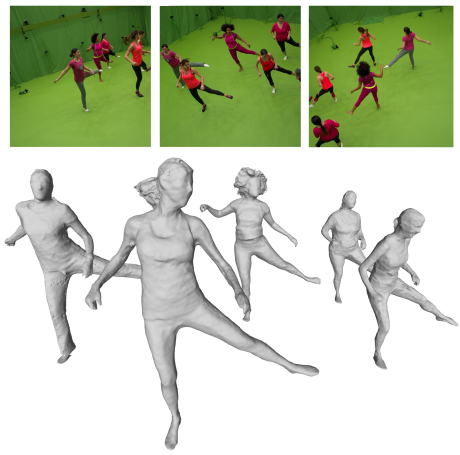

FIGURE 1 – Trois des 68 images d'entrée d'un système multi-caméras moyenne échelle comme exemple de reconstruction d'une scène moyenne échelle dynamique, contenant de multiples sujets en mouvement.

Keywords

Multi-view stereo, Space Time Superresolution.

1 Introduction

Nous proposons une méthode capable de modéliser en 3D une scène dynamique de moyenne échelle observée à l'aide d'un système multi-caméras synchronisé et calibré. Par moyenne échelle, nous faisons référence à une aire de capture d'une douzaine de mètres carrés ou plus, dans laquelle des sujets en mouvement sont observés. Ce scénario diffère des cas standards de Reconstruction Stéréo Multi Vues (MVS), qui traitent, soit de cas petite échelle (petits objets observés dans un volume de capture d'au plus $1m^3$), soit de scènes à grande échelle [43] (bâtiments, villes, etc...). Notre objectif est de généraliser l'applicabilité des méthodes MVS à des scénarios de capture de scènes plus larges, comprenant par exemple de larges mouvements sportifs ou des chorégraphies avec plusieurs danseurs, afin de multiplier les possibilités créatives de beaucoup d'applications associées à la création de contenu 3D réaliste, telles que l'analyse de mouvements sportifs, la création de contenu 3D immersif pour les applications de réalité virtuelle, etc... D'un point de vue technique, ce scénario diffère du problème MVS standard selon plusieurs fronts. Premièrement, le volume de capture

étant plus grand, les caméras possèdent un angle de capture plus large, projetant les surfaces observées sur des régions plus petites de l'image (environ 10%, contrairement au cas de reconstruction standard où l'objet occupe la quasi-totalité de l'image). Deuxièmement, la résolution des caméras permettant de capturer de telles scènes est inférieure d'un ordre de grandeur à celle d'un appareil photo digital utilisé à petite échelle. Troisièmement, la texture des sujets observés est très hétérogène, et souvent quasi-uniforme, avec seulement quelques motifs locaux. Quatrièmement, un plus grand nombre de sujets augmente les inter/auto occultations. La combinaison de tous ces facteurs entraîne des images de plus faible résolution, plus bruitées et une information de texture plus ambigüe, que dans le cas MVS usuel.

Nos contributions résident dans les deux points suivants : puisque le mouvement est une composante centrale dans notre scénario, nous nous tournons de manière naturelle vers l'exploitation de la redondance temporelle de l'information, de manière à améliorer la qualité d'une reconstruction, à l'aide de ses voisines temporelles. D'un autre côté, nous ré-examinons la totalité de l'approche MVS, en utilisant toute l'information à disposition de manière à améliorer la qualité des détails et filtrer le bruit. Pour cela, nous extrayons une carte de profondeur par caméra à l'aide de descripteurs DAISY robustes [35]. Toutes les cartes sont alors fusionnées en une fonction implicite continue par le biais d'une fonction de distance signée tronquée (FDST, représentation robuste d'une surface 3D), de manière similaire aux récents travaux basés sur les capteurs de profondeur [17, 25]. Nous proposons un algorithme capable d'utiliser cette fonction en combinaison avec une estimation de la déformation locale de la surface, afin d'obtenir itérativement une estimation de la surface temporellement super-résolue.

Nous validons notre approche sur plusieurs jeux de données réelles contenant de multiples objets ou personnes, où notre approche permet une réduction du bruit et une amélioration de la complétude dans les régions très occultées. Nous mettons aussi en place un protocole d'évaluation quantitative à l'aide de deux jeux de données synthétiques, moyenne échelle et dynamiques, qui seront mis à la disposition de la communauté. Une amélioration significative de la qualité des reconstructions par rapport à l'état de l'art est mesurée, qui est encore augmentée par la super résolution temporelle.

2 Travaux antérieurs

L'état de l'art en reconstruction stéréo multi-vues étant extrêmement riche, nous ne nous focaliserons ici que sur les travaux les plus pertinents pour situer notre contribution.

La reconstruction stéréo multi-vues statique d'objets inertes a été intensivement étudiée, et des études comparatives et jeux de données de référence [1] sont disponibles et mis à jour régulièrement à petite [30, 18] et grande [33] échelle. Parmi les méthodes les plus performantes, différents moyens de représenter la scène sont utilisés, tels que les surfaces de niveau implicites [28, 10], volumes discrets creusé [5], cartes de profondeur [14, 13] ou échantillons épars obtenus par association de descripteurs qui sont ensuite densifiés en une surface[12, 43, 29]. Nous optons pour une fusion dense de cartes de profondeurs obtenues à l'aide de descripteurs DAISY [35] en utilisant une représentation volumique implicite sous la forme d'une FDST. Nous montrons que cette méthode permet d'obtenir des reconstructions d'une qualité équivalente à l'état de l'art sur un jeu de données standard [18] et supérieure dans le cas moyenne échelle dynamique. Contrairement à toutes ces méthodes, notre approche permet aussi une super-résolution temporelle de nos reconstruction MVS.

La reconstruction de scènes dynamiques, où plusieurs objets en mouvement sont observés par de multiples caméras ou capteurs, a initialement été abordée à chaque instant de temps indépendamment, selon plusieurs approches, telles que [11] (forme à partir de la silhouette), MVS ou point clés, ou alors la combinaison des deux [32]. Plusieurs auteurs ont tenté d'ajouter des contraintes temporelles à la reconstruction, par exemple avec une triangulation de Delaunay en 4 dimensions [3] ou encore à l'aide d'une hypothèse de régularité sur une hypersurface 4D extraite à l'aide de contraintes MVS [15]. D'autres approches mettent en place une propagation locale basée sur le flot optique ou de scène pour guider la régularité d'une inférence [19] où la zone de recherche spatiale pour l'association de descripteurs [23]. Certains travaux mettent en place des contraintes topologiques sur la totalité d'une séquence [27, 23] permettant une extraction consistante d'objets fins tels que des cordes [27] ou encore en assurant une topologie constante sur les silhouettes [23]. Plutôt que de nous concentrer sur des a priori uniquement géométriques, notre méthode permet de propager l'information stéréo observée à l'intérieur d'une fenêtre temporelle.

Parmi les travaux précurseurs, [39] proposent une approche de découpage en 6D innovante, creusant les paires de voxels inconsistantes entre deux instants consécutifs. [37] proposent aussi une amélioration locale, limitée, de la surface en propageant d'un instant à l'autre l'information stéréo le long du flot optique. Notre méthode généralise ce principe dans une fenêtre temporelle par alignement local des formes. [7] démontrent la pertinence d'une technique permettant d'accumuler les silhouettes dans le temps à l'aide de transformations rigides, principe que nous appliquons à notre méthode MVS. [28, 24] estiment simultanément la surface MVS et le flot de scène, mais n'exploitent pas l'alignement local pour propager l'information plusieurs instants autour de la reconstruction comme proposé.

Modalités différentes : Plusieurs approches pertinentes résolvent ce problème avec des capteurs différents, c'est à dire de profondeur [17, 25, 16, 45], par lumière structurée [8], ou encore par stéréo photométrique à l'aide d'un système de lumières actif [41]. [8] construisent un modèle utilisant conjointement information photométrique et lu-

mière structurée et l'animent par suivi de formes sur des petites portions de séquence, mais néanmoins ne l'améliorent pas tel que proposé. [17, 25, 16] sont des inspirations fortes car ils démontrent comment une représentation FDST peut être utilisée pour accumuler l'information, pour des objets rigides ou non. Les différences clef sont que nous généralisons ce principe pour un système multi-caméras couleur, et un scénario contenant d'amples et rapides mouvements. Aussi, pour la plupart des approches précédemment citées, l'aire de capture dans le cas de scènes dynamiques non rigides est limitée à quelques mètres carrés [32, 8, 25, 16, 45, 41], limitation écartée à l'aide de notre méthode. Une exception notable est [6] qui applique la fusion de FDST à grande échelle, néanmoins dans le cas statique.

3 Aperçu de la Méthode

Notre objectif est de reconstruire des modèles temporellement super-résolus de scènes dynamiques moyenne échelle. Nous nous concentrons ici sur un système caméra couleur qui permet la capture de tels scénarios en autorisant une certaine flexibilité dans l'acquisition. D'un autre côté, ce type de système soulève des problématiques spécifiques car l'information photométrique devient facilement inconsistante et ambigüe entre les caméras, possiblement très éloignées. Notre approche exploite la redondance temporelle afin d'améliorer la qualité des modèles dans une fenêtre temporelle dans une séquence. Nous faisons l'hypothèse que dans une fenêtre temporelle sont observées plusieurs instances d'un même modèle déformé, présentant potentiellement des changements topologiques, tel que décrit en figure 2. Notre système de super-résolution temporelle figure 3 considère donc en entrée les images couleur des caméras pour une fenêtre temporelle et produit un seul maillage 3D du modèle super-résolu. Pour cela, l'information de plusieurs instants (typiquement 3 à 7 frames) est fusionnée à la suite d'une alternance entre une estimation de la forme et du mouvement tel que détaillé ci-après.

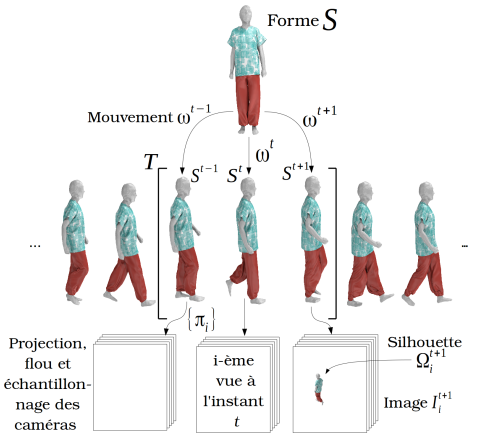

FIGURE 2 – Modèle de formation des images, silhouettes et notations du problème.

4 Reconstructions Initiales

Notre méthode commence tout d'abord par reconstruire chaque instant indépendamment. Ces reconstructions seront ensuite considérées dans la fenêtre temporelle pour la super-résolution du modèle. L'indice temporel est omis pour clarifier les notations. Prenant en entrée une séquence de N images $\{I_i\}_{i=1}^N$ observées par C caméras calibrées, de projections $\{\pi_i\}_{i=1}^N$ et de centres $\{c_i\}_{i=1}^N$.
Nous considérons aussi une séquence de silhouettes $\{\Omega_i\}_{i=1}^N$ obtenues par le biais d'une quelconque méthode de soustraction de fond, donc par conséquent possiblement erronées. De manière à spécifiquement résoudre les problèmes découlant de scénarios à moyenne échelle, c'est à dire une couverture de scène hétérogène, et une grande distance inter-caméra, nous proposons une méthode innovante combinant cartes de profondeur basées stéréo et fusion de FDST robuste. Il est primordial de préciser que la méthode est conçue de manière à faciliter l'intégration temporelle, comme expliqué plus loin. Nous reconstruisons des maillages 3D selon ces trois étapes successives :

1. Construction de cartes de profondeurs lisses par morceaux pour chaque caméra, incluses dans un volume de confiance défini par les silhouettes.
2. Fusion de ces cartes de profondeur à l'aide d'une FDST standard.
3. Extraction de l'isosurface zéro de la fonction implicite, à l'aide d'une nouvelle méthode de maillage, conçue spécialement pour le scénario multi-vue.

4.1 Estimation des Cartes de Profondeur

La première étape des reconstructions initiales consiste à construire une carte de profondeur par image d'entrée. L'objectif ici est d'appliquer une stratégie locale, qui fournit des estimations précises mais bruitées des profondeurs. Le filtrage est assigné à l'étape globale suivante basée sur la FDST. Le principe consiste à échantillonner les profondeurs le long des rayons arrivant dans la caméra et de garder le meilleur candidat potentiel, selon une métrique de photoconsistance basée sur les descripteurs d'image. Afin d'augmenter la précision et de limiter les faux positifs, nous contraignons la recherche de candidats à l'intérieur d'un volume de confiance, basé sur l'information de silhouette.

Volume de confiance Les silhouettes $\{\Omega_i\}_{i=1}^N$ définissent, par extrusion, une enveloppe visuelle 3D qui est supposée contenir l'objet observé. En pratique, les silhouettes sont susceptibles d'être erronées, et de contenir par exemple des trous, ou des parties manquantes, et ne peuvent pas garantir cette propriété. Notre objectif principal étant de réduire la zone de recherche le long des rayons à une région susceptible d'intersecter la surface observée, nous définissons le volume de confiance V tel que :

$$V = \{x \in \mathbb{R}^3 : \exists^{>\alpha} i \, (\pi_i(x) \in I_i) \\ \wedge \exists^{>\beta} i \, (\pi_i(x) \in \Omega_i)\}, \quad (1)$$

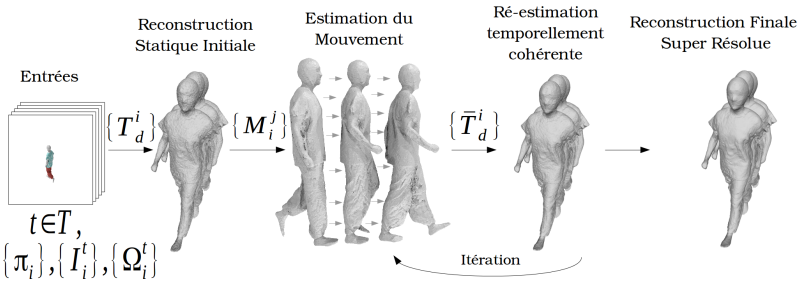

FIGURE 3 – Processus de super résolution temporelle des reconstructions.

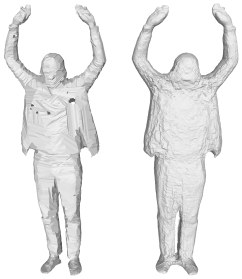

FIGURE 4 – *(gauche)* Enveloppe visuelle (54/68 caméras) et *(droite)* notre volume de confiance $\alpha = 10$, $\beta = 10$.

c'est à dire le lieu géométrique de $\mathbb{R}^3$ qui se reprojette dans au moins α images et au moins β silhouettes. α et β sont des constantes définies par l'utilisateur, avec α restreignant les prédictions peu observées, et autorisant des prédictions éloignées de l'enveloppe visuelle exacte quand β augmente. Intuitivement, V est une version dilatée de l'enveloppe visuelle dans la région de l'espace visible par au moins α images, comme montré en figure 4.

Mesure de la photoconsistance Afin de prédire la profondeur le long d'un rayon, nous utilisons une métrique basée sur le désaccord paire à paire des images. Alors qu'historiquement, la Corrélation Croisée Normalisée s'est imposée comme moyen standard pour comparer la reprojection d'un point sur plusieurs images [12, 27, 10, 42], de récentes avancées dans le domaine des descripteurs d'images ont démontré les bénéfices apportés par les descripteurs basés sur les gradients, tels que SIFT, GLOH, DAISY [21, 22, 35]. Nous choisissons DAISY car il a été expérimentalement montré qu'il s'agit du descripteur le plus robuste dans un contexte dense. Pour un point $x \in \mathbb{R}^3$ et étant donné deux images I_i et I_j, le désaccord photométrique $g_{i,j}(x)$ en x est donné par la distance euclidienne entre les deux descripteurs D_i et D_j associés à la reprojection de x sur les images :

$$g_{i,j}(x) = (D_i(\pi_i(x)) - D_j(\pi_j(x)))^2. \qquad (2)$$

La mesure de photoconsistance $\rho_i(x)$ en x étant donné toutes les images est alors calculée comme un vote robuste normalisé des descripteurs $D_j(\pi_j(x))$ en x similaires à $D_i(\pi_i(x))$:

$$\rho_i(x) = \sum_{j \in C_i} \bar{\omega}_j W(g_{i,j}(x)), \qquad (3)$$

avec : les valeurs normalisées $\bar{\omega}_j$ de $\omega_j = \cos(\theta_{ij})$ pondérant les contributions des caméras autour de la caméra i en utilisant l'angle θ_{ij} entre les axes optiques des caméras i et j ; C_i le sous ensemble des caméras j telles que $\omega_j > 0.7$; et $W()$ est une fonction de vote robuste, une fenêtre de Parzen Gaussienne dans l'espace des descripteurs dans nos expérimentations. On peut noter que 1 représente le meilleur score et 0 est le pire.

La mesure de photoconsistance ci-dessus assume implicitement l'observation de surfaces Lambertiennes, et bien que robuste aux spécularités dans une certaine mesure, peut ne pas fonctionner dans les cas extrêmes. Aussi, en ce qui concerne les occultations, nous nous attendons à ce que ρ présente un maximum local lorsque le rayon intersecte la surface, même dans le cas où la surface est observée par peu de caméras. Aussi, nous limitons la recherche le long d'un rayon afin d'empêcher d'aller trop loin à l'intérieur du volume de confiance(1) comme expliqué ci-après.

Prédiction des profondeurs : pour chaque pixel à l'intérieur de chaque silhouette, la profondeur est prédite le long du rayon sortant de la caméra comme le maximum de la fonction de photoconsistance ρ introduite précédemment. Comme expliqué plus tôt, l'information photométrique est souvent peu fiable dans notre contexte. De manière à prévenir les fausses détections de maxima éloignés de la surface observée, nous adoptons une stratégie conservative, dans laquelle la recherche d'un maximum le long d'une ligne de vue débute à partir du volume de confiance et s'interrompt lorsque la photoconsistance accumulée dépasse une certaine limite, limitant la pénétration des rayons à l'intérieur du volume. L'idée est assez similaire à [26] qui définissent et intègrent une probabilité d'intérieur, néanmoins en utilisant une métrique de photoconsistance similaire à [42].

Plus précisément, le meilleur candidat d_i^p le long du rayon $r_i(p, d)$ sortant de la caméra i à travers le pixel p est déterminé de la manière suivante :

$$d_i^p = \begin{cases} d_V(p) \text{ if } \underset{d\in[d_V(p),d_{max}]}{\arg\max} \ \rho_i(r_i(p,d)) < \tau_{photo}, \\ \underset{d\in[d_V(p),d_{max}]}{\arg\max} \ (\rho_i(r_i(p,d))) \text{ sinon.} \end{cases} \tag{4}$$

Où $d_V(p)$ est la première valeur de profondeur le long de $r_i(p,d)$ à l'intérieur du volume de confiance V, τ_{photo} est une valeur minimale de la photoconsistance en dessous de laquelle on retombe sur l'information de silhouette , et d_{max} la limite de recherche telle que :

$$\int_{x=d_V(p)}^{d_{max}} \rho_i(r_i(p,x))\mathrm{d}x \leq \rho_{max} \tag{5}$$

Afin d'accélérer le calcul des cartes de profondeur, et d'ajouter une consistance spatiale, nous commençons par regrouper les pixels en super-pixels avec [2] , puis sélectionnons aléatoirement des candidats pour chaque super pixel. Nous recherchons la profondeur de ces candidats de manière exhaustive, afin de fournir une approximation des profondeurs à l'intérieur de chaque super pixel. Les autres profondeurs sont alors calculées autour de ces dernières. Comme post-traitement, nous appliquons un filtre bilatéral sur chaque carte. Cette étape permet de filtrer les valeurs aberrantes, tout en ayant un faible impact sur le temps de calcul, motivant notre choix dans un contexte 4D.

4.2 Construction de la Fonction Implicite

étant donné les cartes de profondeur d_i de chaque caméra, à un instant donné, l'étape suivante consiste à récupérer une forme 3D. Similairement à de récents travaux[9, 17, 25], mais dans un contexte différent (petite échelle), nous commençons par fusionner les cartes de profondeur en une fonction implicite, en exploitant les avantages apportés par une stratégie basée sur la FDST. Une autre raison motivant ce choix est la capacité naturelle de la FDST à pouvoir accumuler un nombre arbitraire de cartes de profondeurs, plus précisément, les voisines temporelles d'un instant donné, facilitant l'intégration dans notre cas.
Pour un point $x \in \mathbb{R}^3$, la distance signée tronquée $TD(x) \in \mathbb{R}$ à la surface est définie comme la moyenne pondérée de toutes les prédictions des caméras $F_i(x), i \in C$:

$$\begin{aligned} F_i(x) &= \begin{cases} \min(\mu, \eta(x)) & iif\ \eta(x) \geq -\mu, \\ \emptyset & \text{sinon,} \end{cases} \\ \eta(x) &= d_i(\pi_i(x)) - \|c_i - x\|, \end{aligned} \tag{6}$$

$$TD(x) = \frac{\sum\limits_{i\in C_x} \rho_i(d_i(\pi_i(x)))F_i(x)}{\sum\limits_{i\in C_x} \rho_i(d_i(\pi_i(x)))}, \tag{7}$$

uù $C_x = \{i \in C : F_i(x) \neq \emptyset\}$. Si d_i n'est pas défini en x, c'est à dire x est à l'extérieur du domaine de visibilité de la caméra, alors la caméra i ne contribue pas à la FDST. Lorsqu'aucune caméra ne contribue en x, mais x est à l'intérieur du volume de confiance V, alors le point x est considéré comme appartenant à l'intérieur de la surface, c'est à dire $TD(x) < 0$.

4.3 Génération du Maillage

Comme précédemment défini, on peut extraire une surface 3D comme l'isosurface 0 de la fonction implicite caractérisant la forme. Une grande majorité des méthodes existantes considèrent une approche basée sur les Marching Cubes [20] (MC) pour réaliser cela [12, 17, 26]. Bien qu'une telle approche fonctionnerait dans notre cas, nous considérons plutôt une approche palliant à certaines limitations de MC. En effet, MC est basée sur une discrétisation régulière de l'espace, et limite par conséquent la précision maximale, à moins de mettre en place une stratégie de subdivision spécifique.
Nous construisons une méthode simple basée sur de récents travaux utilisant la tesselation de Voronoï [44] démontrant qu'une meilleure précision peut être atteinte par une discrétisation des formes et non pas de l'espace. La méthode consiste à appliquer les opérations suivantes :

1. échantillonnage de points à l'intérieur de la surface définie par la FDST. Cette étape est effectuée en sélectionnant aléatoirement des pixels appartenant à la silhouette dans toutes les images. Pour chaque pixel sélectionné, le rayon sortant est parcouru et le premier point à l'intérieur de la surface mais proche de l'interface est conservé. Nous réitérons jusqu'à obtenir un nombre de points 3D défini par l'utilisateur.
2. Construction du diagramme de Voronoï à partir de l'ensemble des points à l'intérieur de la forme.
3. Découpage du diagramme de Voronoï avec le niveau 0 de la FDST. Cette opération extrait l'intersection entre les cellules de Voronoï à l'interface et la surface.

échantillonner des points proches de la surface à partir des différents points de vue assure une discrétisation dépendante du nombre d'observations. Ceci permet d'obtenir une meilleure précision de reconstruction dans les zones de l'espace observées par plus de caméras.

5 Super Résolution Temporelle

Jusque là, nous avons effectué la reconstruction de chaque instant de temps, de manière indépendante, d'une séquence S^t. L'objectif de la super résolution temporelle est de raffiner chaque reconstruction en prenant en compte ses voisines temporelles. Idéalement, cela nous permettrait de récupérer et d'intégrer un niveau de détail inaccessible à partir de l'information d'un seul instant. Comme démontré dans la section d'évaluation, notre approche atteint cet objectif, par propagation des profondeurs correctement détectées entre les instants. Comme illustré en figure 3, notre approche calcule chaque instant à l'aide de ses voisins dans une fenêtre temporelle de n trames $\{S^t\}_{t=1}^n$. Nous faisons l'hypothèse que chaque instant t à l'intérieur de la fenêtre temporelle, correspond à l'observation d'une instance de la même forme S^{ref}, avec $S^{ref} = S^1$ ou bien $S^{ref} = S^{n/2}$,

déformée par un champ de mouvement 3D W^t_{ref}. L'approche consiste alors à alterner entre les deux opérations suivantes :

1. étant donné $\{S^t\}$, estimer le champ de mouvement $\{W^t_1\}^n_{t=2}$ avec $S^{ref} = S^1$, sans perte de généralité.
2. étant donné $\{W^t_1\}^n_{t=2}$ ré-estimer S_1 par fusion de toutes les FDST de chaque instant, déformées par $\{W^t_1\}^n_{t=2}$.

En pratique, S^{ref} correspond au centre de la fenêtre temporelle de taille impaire. à chaque itération, la fenêtre temporelle est glissée le long de toute la séquence, puis les mouvements d'un instant à l'autre sont entièrement ré-estimés. Le processus complet est réitéré plusieurs fois, typiquement 3 dans nos expériences.

5.1 Estimation du Mouvement

Considérons deux maillages S^k et S^l obtenus aux instants k et l, notre objectif est de récupérer un champ de mouvement dense $W^l_k : \mathbb{R}^3 \to \mathbb{R}^3$ déformant S^k en S^l.
Notre but consiste uniquement à améliorer les formes, par conséquent, nous ne désirons par récupérer le mouvement complet entre les deux formes, tel que dans le suivi de formes ou l'estimation du flot de scène. Nous cherchons en réalité à extraire les déplacements épars les plus fiables, dans des régions de la surface qui pourront alors bénéficier de l'intégration temporelle. Par conséquent, le champ de mouvement 3D ne doit pas forcément reproduire parfaitement le mouvement, néanmoins, il doit être équipé d'une mesure de la confiance, identifiant les mouvements valides et autorisant à ignorer le reste lors de l'intégration d'une frame à l'autre.
Plusieurs méthodes existent pour récupérer cette information pour les formes en mouvement. Dépendant de l'a-priori sur le modèle de déformation, elle vont de modèles faiblement contraints [38] basés sur le flot de scène, à des modèles localement rigides type ARAP [4], tel que dans les méthodes Kinect et Dynamic Fusion [17, 25] ou encore [8]. Des modèles plus fortement contraints existent, mettant en place des squelettes articulés [40].
Notre contexte ne nécessitant pas un modèle de mouvement complet , nous préférons opter pour une stratégie localement contrainte. De plus, les scènes dynamiques de moyenne échelle peuvent contenir des mouvements de grande ampleur entre les instants, prônant par conséquent pour une méthode d'association éparse mais robuste. Pour ces raisons, nous optons pour l'utilisation de descripteurs de surfaces afin de récupérer des associations 3D robustes, qui seront ensuite densifiées par l'alternance des étapes décrites précédemment. Nous choisissons MeshHog [46], afin de détecter et extraire des descripteurs de la surface, basés sur sa géométrie et son apparence, car ils offrent un bon compromis entre robustesse, complétude et précision , parmi d'autres méthodes comme la diffusion de chaleur [34] ou bien Harris 3D [31].
En notant $\{M^k\}$ l'ensemble des paires de descripteurs correspondantes entre S^k et S^{k+1}, obtenues via MeshHog, et $m \in \{M^k\}$ une telle paire. On associe à m un facteur de confiance λ_m privilégiant les régions contenant des associations denses et cohérentes entre elles. Dans ce but, nous sélectionnons les k-plus proches voisins m_j de m dans $\{M^k\}$. Soit δ^j_m le désaccord entre les vecteurs de déplacement associés à m et m_j. δ^j_m est alors la médiane des j valeurs $\mathcal{G}(\delta^j_m)$, où $\mathcal{G}$ est un noyau gaussien. Cette stratégie conservative favorise localement les régions de S^k où m et ses voisins présentent le même vecteur de déplacement. Puisqu'au fil des itérations, de plus en plus d'associations seront détectées, on peut voir cette approche comme une croissance qui étend progressivement le champ de déplacement autour des régions où les associations sont détectées de manière consistante.
Avec les paires de descripteurs MeshHog correspondantes $m \in \{M^k\}$, leurs vecteurs de déplacement $\{T_m\}$ de S^k à S^{k+1} et leurs facteurs de confiance λ_m, on définit le champ de mouvement vers l'avant $W^+_k : \mathbb{R}^3 \to \mathbb{R}^3$ tel que :

$$W^+_k(x) = \sum_{m \in \{M^k\}} \lambda_m \, \mathcal{G}_m(x) \, T_m, \tag{8}$$

avec $\mathcal{G}_m()$ un noyau gaussien pondérant les contributions de m relativement à la distance spatiale entre x et le descripteur de m sur S^k.
Le champ de mouvement vers l'arrière $W^-_k(x)$ qui déforme S^k en S^{k-1} est défini de manière similaire en utilisant les descripteurs MeshHog entre S^k et S^{k-1}. Le champ de mouvement W^l_k est alors défini tel que :

$$W^l_k(x) = \begin{cases} \sum\limits_{t \in [k,l-1]} W^+_t(x) & if\ k < l, \\ \sum\limits_{t \in [k,l+1]} W^-_t(x) & if\ k > l, \\ 0 & if\ k = l, \end{cases} \tag{9}$$

5.2 Intégration Temporelle

étant donné les champs de mouvement denses $\{W^l_k\}$ tels que définis dans la section précédente, l'approche de super résolution temporelle consiste à raffiner la reconstruction de l'instant de référence $S^{ref} = S^k$. Pour cela, on intègre la dimension temporelle dans la FDST $\overline{TD} : \mathbb{R}^3 \to \mathbb{R}$:

$$\overline{TD}(x) = \frac{\sum\limits_{t \in T} \sum\limits_{i \in C^t_x} \rho^t_i d^t_i(\pi_i(\mathcal{W}^t_k(x))) F^t_i(\mathcal{W}^t_k(x))}{\sum\limits_{t \in T} \sum\limits_{i \in C^t_x} \rho^t_i d^t_i(\pi_i(\mathcal{W}^t_k(x)))} \tag{10}$$

où $T = [k - n/2, k + n/2]$, $C^t_x = \{i \in C : F^t_i(x) \neq \emptyset\}$ et ρ^t_i, d^t_i et F^t_i sont respectivement la mesure de photoconsistance, la prédiction de profondeur et la fonction de distance signée tronquée introduite en section 4.1 et 4.2, à l'instant t. $\mathcal{W}^t_k : \mathbb{R}^3 \to \mathbb{R}$ est simplement le champ de déformation : $\mathcal{W}^t_k(x) = x + W^t_k(x)$. Notons que dans l'intégration ci présentée, la contribution des instants voisins peut aussi être pondérée proportionnellement à la distance temporelle à l'instant de référence.
Finalement, la forme implicite ci-dessus est utilisée pour générer un maillage, comme expliqué en section 4.3.

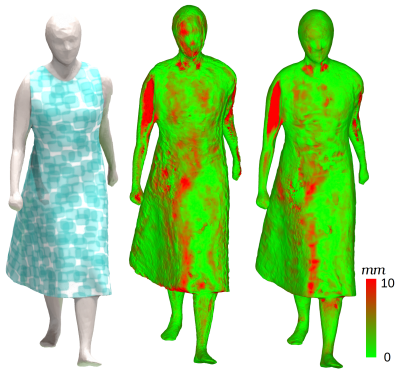

FIGURE 5 – *De gauche à droite :* Précision sur la séquence synthétique *dress*. Reconstruction statique ($T = 1$). Super résolution temporelle, 3 itérations et $T = 5$.

6 Résultats

Afin de démontrer les bénéfices apportés par notre méthode, nous avons conduit plusieurs expérimentations. Premièrement, des expériences quantitatives ont été menées, pour évaluer l'amélioration apportée par la super résolution temporelle. Dans ce but, et étant donné l'absence de jeux de données de référence pour les scènes dynamiques moyenne échelle, nous avons créé un tel jeu de données, fournissant une vérité terrain tant dans la géométrie que dans l'apparence. Dans un second temps, des résultats qualitatifs sur des données réelles ont été obtenus, pour illustrer l'ajout de détails absents dans les reconstructions statiques par l'intégration temporelle. Finalement, bien que notre méthode ne soit pas conçue dans ce but, nous la comparons à l'état de l'art sur un jeu de données statique petite échelle standard, afin de démontrer que cette approche obtient des résultats de qualité similaire. Le code source et les données seront mises à disposition à la communauté.

Les temps de calculs de notre implémentation C++ multi processeur, sur un Xeon 16 coeurs 3.00GHz, 32Gb RAM, sur un jeu de données 86 caméras 4M pixels sont les suivants : 5-20min/frame pour construire la FDST implicite, dépendant du nombre total de pixels à l'intérieur des silhouettes, 5min/frame pour l'estimation du mouvement, et 5min pour l'extraction de surface, pour un maillage final de 3M faces. Une implémentation GPU serait facilement mise en place, permettant une diminution significative des temps de calcul.

6.1 Données Synthétiques

Jeu de données De nombreux jeux de données sont disponibles en ligne pour le cas de la reconstruction stéréo multi-vues, tels que [30], ou [18]. Néanmoins, aucun n'existe à notre connaissance en ce qui concerne le problème dynamique à échelle moyenne, contenant des surfaces se déformant dans le temps. Pour cette raison, nous construisons un jeu de données d'évaluation, avec comme objectif d'être le plus proche possible de la réalité. On peut aussi noter que l'utilité d'un tel jeu de données ne se limite pas qu'au problème de reconstruction, mais est aussi extrêmement intéressant pour l'évaluation des problèmes de suivi de formes en mouvement ou de modélisation de l'apparence. Les données sont des surfaces simulées, typiquement des vêtements, animées avec des données réelles capturées, typiquement des formes de corps en mouvement, sur lesquelles un suivi de forme dans le temps a été appliqué. Leurs principales caractéristiques sont :

1. La simulation de capture des images reproduit la disposition d'une plate-forme de capture multi-caméras réelle.
2. Les formes et leurs mouvements ont été capturés, et par conséquent reproduisent des situations dynamiques réelles.
3. La déformation locale des formes est générée et permet de simuler aussi bien des vêtements, que d'autre types de déformations.
4. L'apparence est elle aussi générée, et permet d'étudier la robustesse des méthodes dans des cas variés, tels que des textures très / peu contrastée, des spécularités, la diffusion de couleur, du flou de mouvement etc...

évaluation étant donné la vérité terrain mentionnée plus haut, nous effectuons nos évaluations en utilisant des mesures standard du domaine [30, 18], c'est à dire, précision et complétude. Les reconstructions statiques et temporellement super résolues ont été faites sur 20 trames d'une séquence synthétique, avec des déformations locales de la robe (voir fig. 5), observée par 60 caméras, avec un volume de capture d'approximativement 8mx4mx6m, pour des images d'une résolution de $2048x2048$. La précision est montrée figure 5 à l'instant le moins bien reconstruit, illustrant les bénéfices de l'intégration temporelle sur une grande partie de la surface.

La Figure 7 montre comment la complétude moyenne sur 5 trames augmente avec une taille de fenêtre temporelle de 1, 5 et 7. Sur cette figure apparaît aussi la complétude obtenue avec [13], une des méthodes les plus performantes sur le DTU [18]. Alors que la comparaison de précision serait injuste envers cette méthode, puisqu'elle ne prend pas en compte les silhouettes, et donc reconstruit des points éloignés de la véritable surface, la complétude reste quand même un indicateur pertinent. Cette figure montre aussi les valeurs minimales et maximales de la complétude sur 20 trames de la séquence *dress*. Il est intéressant de remarquer que la complétude minimale est améliorée d'environ 15% à une distance à la surface de 3mm, alors que la meilleure résolution des pixels d'une caméra sur la surface varie de 2x2 à 5x5mm.

6.2 Données Réelles

Nous avons aussi effectué des reconstructions à partir de multiples jeux de données dynamiques moyenne échelle réels, contenant plusieurs sujets, pas uniquement humains. Chaque séquence a été capturée à l'aide de 68 caméras couleur calibrées (2048x2048) de longueurs focales variant de 15 à 22mm. Fig. 6 montre la résolution et la qualité d'une image d'entrée, notre reconstruction statique et l'amélioration par super résolution temporelle. Elle illustre

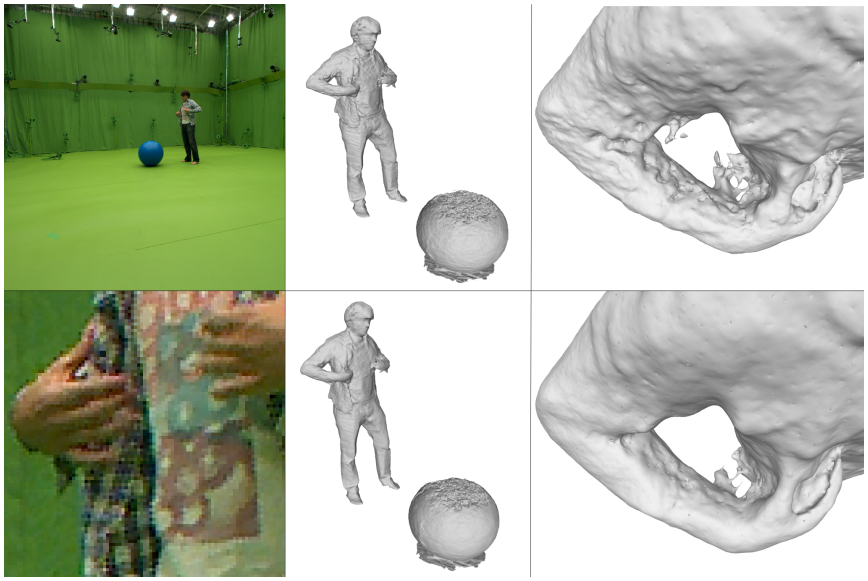

FIGURE 6 – Amélioration par intégration temporelle des détails sur des données réelles. De haut en bas, Gauche : une image d'entrée et gros plan au niveau du pixel. Milieu : reconstruction statique et super résolution temporelle, avec $T = 5$ et 2 itérations. Droite : plan rapproché sur la région du bras du modèle.

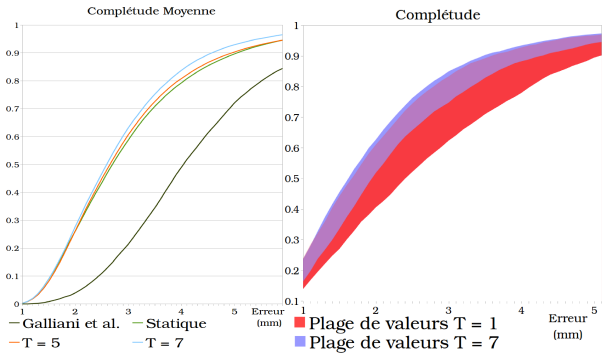

FIGURE 7 – (*gauche*) Complétude moyenne entre [13] et notre approche, 5 trames de *dress*, (*droite*) Min et max de la complétude sur 20 trames de *dress*, T = 7, itérations = 3.

aussi les difficultés rencontrées dans le scénario dynamique moyenne échelle et montre l'intérêt de l'intégration temporelle, qui améliore la reconstruction de manière claire dans la région du bras, peu observée par le système.

6.3 Données Standard

Dans un souci d'hexaustivité, nous évaluons aussi nos reconstructions statiques sur le DTU [18]. Alors que notre méthode, n'a pas été conçue pour résoudre ce cas de reconstruction stéréo multi-vues, cette comparaison reste informative et montre que notre approche obtient des résultats compétitifs avec l'état de l'art.

7 Conclusion

Nous avons présenté une méthode de super résolution temporelle de reconstructions, spécialement efficace dans les scénarios complexes dynamiques de moyenne échelle, capturés à l'aide d'un système multi-caméras. Notre approche

	Précision		Complétude	
	Moy.	Med.	Moy.	Med.
Points				
Méthode proposée	0.588	0.262	1.228	1.039
Tola et al. [36]	0.245	0.186	0.518	0.381
Furukawa et al. [12]	0.367	0.246	0.462	0.350
Campbell et al. [5]	0.576	0.391	0.220	0.154
Surfaces				
Méthode proposée	0.575	0.245	0.679	0.319
Tola et al. [36]	0.306	0.195	0.460	0.327
Furukawa et al. [12]	0.548	0.251	0.438	0.350
Campbell et al. [5]	0.750	0.354	0.413	0.351

TABLE 1 – Comparaisons sur 2 objets de [18] (mm)

est tout d'abord capable d'obtenir des reconstructions indépendantes précises, puis de les raffiner par exploitation de la redondance temporelle de l'information. Cette étape permet d'obtenir des reconstructions moins bruitées et plus précises, plus particulièrement dans les régions très occultées à un instant donné. Ceci est réalisé en propageant les information photométriques fiables dans le temps, en accumulant les formes FDST implicites, et en extrayant les surfaces à l'aide d'une solution indépendante du volume. Nous introduisons une stratégie de croissance de région, pour graduellement estimer le mouvement des sujets, alternant entre une accumulation temporelle prudente des observations et une ré-estimation du mouvement de la scène. Les comparaisons à l'état de l'art démontrent aussi bien l'efficacité de notre approche à reconstruire des surfaces dans le cas statique standard, que dans le cas dynamique moyenne échelle, comme validé à l'aide du jeu de données synthétique proposé.

Références

[1] Middleburry multi-view stereo evaluation dataset. vision.middlebury.edu/mview/.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC Superpixels. Technical report, EPFL, 2010.

[3] E. Aganj, J. Pons, F. Ségonne, and R. Keriven. Spatio-temporal shape from silhouette using four-dimensional delaunay meshing. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8, 2007.

[4] M. Alexa, D. Cohen-Or, and D. Levin. As-rigid-as-possible shape interpolation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, New Orleans, LA, USA, July 23-28, 2000*, pages 157–164, 2000.

[5] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I*, pages 766–779, 2008.

[6] J. Chen, D. Bautembach, and S. Izadi. Scalable real-time volumetric surface reconstruction. *ACM Trans. Graph.*, 32(4) :113 :1–113 :16, 2013.

[7] G. K. M. Cheung, S. Baker, and T. Kanade. Visual hull alignment and refinement across time : A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA*, pages 375–382, 2003.

[8] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. G. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4) :69, 2015.

[9] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 303–312, 1996.

[10] C. H. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3) :367–392, 2004.

[11] J. Franco and E. Boyer. Efficient polyhedral modeling from silhouettes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(3) :414–427, 2009.

[12] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007.

[13] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 873–881, 2015.

[14] M. Goesele, B. Curless, and S. M. Seitz. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 2402–2409, 2006.

[15] B. Goldlücke and M. A. Magnor. Space-time isosurface evolution for temporally coherent 3d reconstruction. In *CVPR (1)*, pages 350–355, 2004.

[16] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. Volumedeform : Real-time volumetric non-rigid reconstruction. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 362–379, 2016.

[17] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. A. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. J. Davison, and A. W. Fitzgibbon. Kinectfusion : real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011*, pages 559–568, 2011.

[18] R. R. Jensen, A. L. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 406–413, 2014.

[19] E. S. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8, 2007.

[20] W. E. Lorensen and H. E. Cline. Marching cubes : A high resolution 3d surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1987, Anaheim, California, USA, July 27-31, 1987*, pages 163–169, 1987.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[22] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA*, pages 257–263, 2003.

[23] A. Mustafa, H. Kim, J. Guillemaut, and A. Hilton. Temporally coherent 4d reconstruction of complex dynamic scenes. *CoRR*, abs/1603.03381, 2016.

[24] J. Neumann and Y. Aloimonos. Spatio-temporal stereo using multi-resolution subdivision surfaces. *International Journal of Computer Vision*, 47(1-3) :181–193, 2002.

[25] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion : Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR) 2015, Boston, MA, USA, June 7-12, 2015*, pages 343–352, 2015.

[26] M. R. Oswald and D. Cremers. A convex relaxation approach to space time multi-view 3d reconstruction. In *ICCV Workshop on Dynamic Shape Capture and Analysis (4DMOD)*, 2013.

[27] M. R. Oswald, J. Stühmer, and D. Cremers. Generalized connectivity constraints for spatio-temporal 3d reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 32–46, 2014.

[28] J. Pons, R. Keriven, and O. D. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global

image-based matching score. *International Journal of Computer Vision*, 72(2) :179–193, 2007.

[29] A. Romanoni, A. Delaunoy, M. Pollefeys, and M. Matteucci. Automatic 3d reconstruction of manifold meshes via delaunay triangulation and mesh sweeping. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, pages 1–8, 2016.

[30] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 519–528, 2006.

[31] I. Sipiran and B. Bustos. A robust 3d interest points detector based on harris operator. In *Eurographics Workshop on 3D Object Retrieval, Norrköping, Sweden, May 2, 2010, Proceedings*, pages 7–14, 2010.

[32] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3) :21–31, 2007.

[33] C. Strecha, W. von Hansen, L. J. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*. IEEE Computer Society, 2008.

[34] J. Sun, M. Ovsjanikov, and L. J. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. *Comput. Graph. Forum*, 28(5) :1383–1392, 2009.

[35] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*, 2008.

[36] E. Tola, V. Lepetit, and P. Fua. DAISY : an efficient dense descriptor applied to wide-baseline stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(5) :815–830, 2010.

[37] T. Tung, S. Nobuhara, and T. Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 1709–1716, 2009.

[38] S. Vedula, S. Baker, P. Rander, R. T. Collins, and T. Kanade. Three-dimensional scene flow. In *ICCV*, pages 722–729, 1999.

[39] S. Vedula, S. Baker, S. M. Seitz, and T. Kanade. Shape and motion carving in 6d. In *2000 Conference on Computer Vision and Pattern Recognition (CVPR 2000), 13-15 June 2000, Hilton Head, SC, USA*, pages 2592–2598, 2000.

[40] D. Vlasic, I. Baran, W. Matusik, and J. Popovic. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.*, 27(3), 2008.

[41] D. Vlasic, P. Peers, I. Baran, P. E. Debevec, J. Popovic, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Trans. Graph.*, 28(5) :174 :1–174 :11, 2009.

[42] G. Vogiatzis, C. H. Esteban, P. H. S. Torr, and R. Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12) :2241–2246, 2007.

[43] H. Vu, R. Keriven, P. Labatut, and J. Pons. Towards high-resolution large-scale multi-view stereo. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 1430–1437, 2009.

[44] L. Wang, F. Hétroy-Wheeler, and E. Boyer. On volumetric shape reconstruction from implicit forms. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, pages 173–188, 2016.

[45] R. Wang, L. Wei, E. Vouga, Q. Huang, D. Ceylan, G. G. Medioni, and H. Li. Capturing dynamic textured surfaces of moving targets. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, pages 271–288, 2016.

[46] A. Zaharescu, E. Boyer, and R. Horaud. Keypoints and local descriptors of scalar functions on 2d manifolds. *International Journal of Computer Vision*, 100(1) :78–98, 2012.