Journées Françaises d'Informatique Graphique,
Grenoble, 30 nov - 2 dec 2016

Réalité diminuée : "couper/coller" interactif pour l'aménagement d'intérieur

Julien Fayer[1,2], Géraldine Morin[1], Simone Gasparini[1], Benjamin Coudrin[2]

[1] Université de Toulouse, IRIT-INPT
[2] SAS InnerSense, Toulouse

Résumé

L'arrivée des capteurs RGB-D comme la Kinect a permis de développer de nouvelles approches en matière de vision par ordinateur et en traitement d'image pour comprendre et interpréter une scène d'intérieur. Nous proposons de mettre en place un scénario de Réalité Diminuée qui est capable non seulement de supprimer des éléments de la scène mais aussi de les déplacer. Nous présentons d'abord un état de l'art sur la compréhension puis l'édition de la scène. Nous abordons ensuite les limites de ces approches. Enfin, nous mettons en évidence les problématiques d'« inpainting » propres à notre application et les apports qui tiennent compte des résultats issus de la compréhension tridimensionnelle pour améliorer le processus d'édition (suppression ou déplacement).

Mots clé : RGB-D, réalité diminuée, inpainting, sémantique 3D, édition 3D

1. Introduction et contexte

La Réalité Diminuée est une branche de la Réalité Augmentée qui consiste à modifier (généralement supprimer) des éléments réels d'un environnement numérisé. L'objectif final est de pouvoir faire de l'aménagement virtuel d'un environnement d'intérieur (suppression puis déplacement de meubles, modification de textures du sol ou des murs). Au début, les scénarios de réalité diminuée sont essentiellement de l'ordre du traitement de l'image avec notamment l'utilisation de l'« inpainting » [CPT03], mais restent limités en termes de possibilités. Les dernières approches commencent en revanche à utiliser des informations comme la position de la caméra pour améliorer le rendu final et diversifier les scénarios possibles [Sil15, KSY15]. Cependant, ces travaux récents restent de la suppression et non de l'édition car les informations 3D sont insuffisantes. Parallèlement, l'arrivée de capteurs couplant couleur et profondeur (on parle de capteurs RGB-D) de faible coût comme la Kinect permet d'obtenir une meilleure connaissance de l'environnement 3D. Plusieurs travaux utilisent les possibilités offertes par ces capteurs RGB-D afin d'améliorer la compréhension de la scène [NSF12, WCM15].
Nous proposons dans cet article de montrer les limites des approches actuelles d'édition de scène et d'utiliser ensuite les différents travaux de recherche sur la compréhension de scène 3D pour améliorer la chaîne de traitement via plusieurs contributions (rectification globale et processus d'édition).
Dans cet article nous présentons un premier processus global ainsi que les techniques de l'état de l'art qui permettent de mettre en œuvre cette chaîne de traitement, les limitations des approches de l'état de l'art, nos contributions notamment grâce à des informations 3D, le scénario amélioré envisagé et enfin les perspectives.

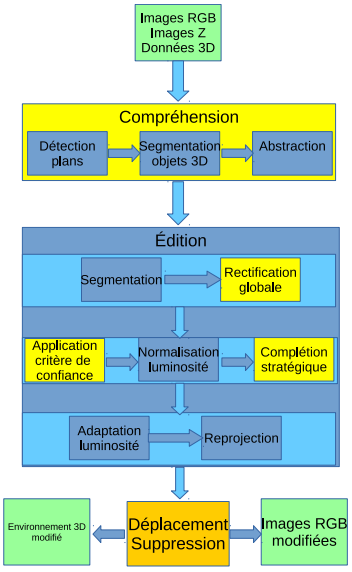

Figure 1: Chaîne globale de traitement proposée (les contributions sont encadrés en jaune).

2. Vue générale

La Figure 1 présente la chaîne globale de traitement proposée avec quatre étapes principales. Elle prend comme données d'entrée N points de vue ayant chacun une image de couleur, une image de profondeur et la pose de la caméra. La première étape (premier bloc) consiste à effectuer un travail de compréhension de la scène pour en extraire notamment la position des plans et des meubles. La deuxième étape (du deuxième au quatrième bloc) consiste à effectuer une transformation d'un objet dans l'environnement 3D (translation, rotation, suppression) et de répercuter celle-ci dans les

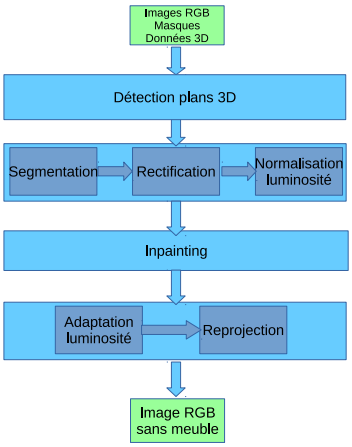

Figure 2: Chaîne globale de processus des approches actuelles comme Kawai et al. [KSY15] et Siltanen et al. [Sil15].

images des différents points de vue. Généralement, l'objectif est de combler les zones de l'image laissées vacantes suite au déplacement ou à la suppression d'un objet tout en respectant les contraintes de l'environnement de la scène (par exemple la cohérence spatiale des textures ou la luminosité).

2.1. Analyse et compréhension de la scène

Dans une première étape, l'ensemble des données d'entrée est traité afin d'obtenir un premier modèle de la scène d'intérieur : positions des plans, positions et catégories des meubles, relations de support entre les meubles.

Détection des plans Détecter les plans, notamment ceux du sol et des murs, permet d'identifier la transformation pour passer du repère caméra à celui de la pièce ; cela contribue également à faciliter la détection des objets. Pour cela, nous utilisons les méthodes proposées par Silberman et al. [NSF12] et Deng et al. [DTL16] qui calculent les normales en chaque pixel de l'image RGBD. Ils ré-orientent ensuite le repère en se basant soit sur l'hypothèse du "monde de Manhattan", soit sur un calcul du vecteur gravité. La détection se fait ensuite en effectuant une procédure robuste basée sur RANSAC.

Segmentation Deng et al. [DTL16] proposent de travailler sur un nuage de points 3D généré à partir de l'image de profondeur et des paramètres intrinsèques du capteur calibré. Ils retirent les plans dominants (représentant le sol et les murs) en effectuant un regroupement euclidien via un arbre Kd. Nous utilisons cette approche plus rapide en temps d'exécution pour des scènes où les plans principaux sont bien visibles (peu de meubles). Pour des scènes riches où les principaux plans sont peu visibles, nous utilisons une approche plus coûteuse [NSF12], qui combine une segmentation 2D basée contours [Arb06] avec un classifieur entraîné sur des caractéristiques basées sur la couleur, la position dans l'image et la profondeur.

Abstraction des objets et des relations Afin de procéder à l'étape de labellisation, il est nécessaire de trouver des primitives pour abstraire l'espace 3D. Si la représentation des plans reste aisée, le cas des objets reste non résolue. Actuellement, les objets sont généralement représentés soit par un cuboïde (généralement, la boîte englobante) [WCM15] [DTL16], soit par un ensemble de cuboïdes [RZS*16]. De plus, des relations dites de "support" sont créés entre les objets de la scène comme connaissances a priori sur les relations d'objets [NSF12, WCM15]. Ces relations sont liées à la position relative d'un objet par rapport à un autre (ex : une lampe de chevet sera sur une commode, un lit sera accolé au mur), et permettent de prédire le label de chaque partie segmentée. Dans notre chaîne de processus, nous utilisons l'approche proposée dans [WCM15] pour calculer les cuboides, les labels ainsi que les relations de support.

Figure 3: Résultats de l'approche de Kawai et al. [KSY15].

2.2. Édition de la scène

Si les techniques d'inpainting [RPMK13, CPT03, CS01] sont généralement utilisées pour reconstruire la zone occultée, elles restent limitées par leur application à l'image 2D à cause de leur manque de respect des contraintes 3D. De plus, elles ne peuvent gérer à elles seules le problème des données intermédiaires : les contraintes environnementales (perspective, luminosité) d'un patch P_m de la zone de l'image à compléter ne peuvent pas être retrouvées par simple combinaison de patchs de la zone connue comme dans le cas d'un carrelage où l'on ne pourra pas le reconstruire en se servant directement des patchs voisins à moins de se trouver "face" au plan. Il est donc proposé généralement de se servir de données supplémentaires (notamment 3D) pour améliorer le résultat des techniques de recouvrement. La Figure 2 schématise la chaîne de processus des approches ci-dessous.
Kawai et al. [KSY15] proposent l'effacement des objets en incorporant des données 3D issues de leur pipeline SLAM qui leur permet d'estimer les plans présents dans la scène. Les plans permettent d'effectuer une segmentation entre les masques des plans de chaque point de vue et ensuite de calculer, pour chaque plan, l'homographie de rectification perspective. On peut ensuite effectuer un procédé d'inpainting [KY12] et re-projeter le résultat dans l'espace de départ. La Figure 3 montre un résultat de l'approche tout à fait convenable pour un scénario de base.

Siltanen et al. [Sil15] utilisent un marqueur pour récupérer l'équation du plan du sol, alors que les murs de la pièce sont indiqués par l'utilisateur. Une segmentation et une rectification sont ensuite effectuées. Ils effectuent ensuite une normalisation de l'illumination afin d'éviter le problème des données intermédiaires concernant la luminosité. Un procédé d'inpainting est ensuite effectué sur chaque texture normalisé. Ils réadaptent ensuite le résultat à la luminosité ambiante [KAS10] qui est ensuite re-projetée.

2.3. Limitations des approches d'édition

Si les approches ci-dessus ont des résultats satisfaisants, elles restent cependant limitées. Les scènes utilisées sont généralement assez simples : peu de meubles présents, hypothèses fortes (plans des murs toujours deux à deux orthogonaux). Il devient difficile d'avoir un résultat acceptable avec des scènes d'intérieur réalistes (beaucoup de meubles, murs non perpendiculaires).
Le processus d'inpainting s'effectue sur l'image rectifiée d'une seule prise de vue (notamment [Sil15]) ce qui limite l'apport d'informations sur la texture et la luminosité. Kawai et al. [KSY15] propage en effet le résultat d'une trame clef vers les trames suivantes mais l'apport reste limité par le fait que le masque de l'objet indésirable généré par les

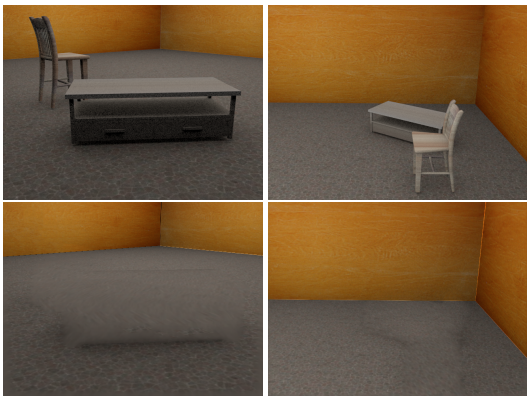

Figure 4: Rendu d'un point de vue virtuel avec le sol soumis à une perspective forte (à gauche) ou faible (à droite). Haut : images de départ ; bas : images après effacement (inpainting effectué avec G'MIC [Tsc16]).

trames suivantes empêche l'utilisation de données cachées à la trame n mais visibles à la trame $n+1$. Cela se justifie par le fait que les objets ne sont considérés que comme des masques 2D projetés sur des plans 3D et non des objets 3D en tant que tels. À la fin, il est suggéré de considérer que l'apport des capteurs RGB-D peuvent effectivement fournir une meilleure segmentation des objets dans l'image.

Problème de résolution des données Quand nous effectuons une rectification, la position des plans par rapport au point de vue influence la qualité du résultat. Par exemple, dans la Figure 4, on constate que le rendu de la zone masquée du mur (en position fronto-parallele par rapport à la caméra) est de meilleure qualité que celle du sol (forte perspective).
Aussi, les techniques de complétion ne tiennent pas compte de ces variations de résolution. On se retrouve avec un résultat, comme dans la Figure 5, où l'inpainting a complété la zone masquée sans prendre en considération la résolution des données. Ceci pose problème une fois l'image rectifiée re-projetée dans l'espace de départ. En effet, des données de basse résolution peuvent être copiées dans une zone masquée qui devrait être de plus forte résolution.
De plus, on peut avoir le même problème qu'avec la luminosité ou la perspective ; nous ne disposons pas forcément dans l'image rectifiée de données de même résolution que la zone masquée. Ceci entraîne des artefacts correspondant aux frontières entre les données de faible résolution et les données de forte résolution.

3. Contributions

La présence de données 3D ainsi que le travail de compréhension fait en amont permet d'améliorer les scénarios de Réalité Diminuée. Nous allons aborder dans cette partie les apports développés.

3.1. Image rectifiée globale

Notons P le plan d'un mur d'une pièce visible sur plusieurs points de vue W_i et I_i l'image rectifiée du masque recouvrant P de l'image résultant de W_i. Au lieu d'effectuer les traitements de complétion sur chaque image I_i, nous profitons du fait d'avoir des données sur les plans (normale du plan mais aussi délimitations) et les poses des caméras pour créer une image rectifiée de dimensions proportionnelles à celles du plan qui contient toutes les projections de chaque masque par l'homographie H_i associée. En plus d'éviter les

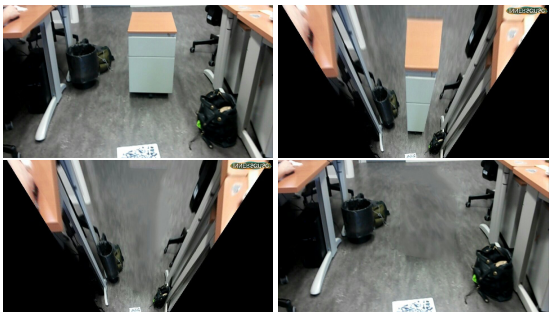

Figure 5: Rendu d'un point de vue réel avec le sol soumis à une forte perspective : en haut, à gauche : image de départ ; à droite : image rectifiée ; en bas, à gauche : image rectifiée complétée ; à droite : résultat (inpainting effectué avec une approche PatchMatch [BSFG09]).

incohérences spatiales, cela permet de profiter des données des autres points de vue pour réduire le temps de calcul et améliorer le rendu d'inpainting.
Nous ne sommes pas gênés par les objets présents dans l'image 2D. En effet, grâce au travail de compréhension, nous disposons de primitives pour situer chaque objet dans l'environnement 3D. Nous pouvons ainsi générer des masques 2D d'objets automatiquement pour chaque prise de vue sans demander l'aide d'un utilisateur. Seule la partie correspondant au plan considéré est alors projetée. La Figure 6 montre un exemple d'image rectifiée globale appliquée au sol en considérant quatre points de vue.

3.1.1. Extensions possibles

Critère de confiance 3D Il peut arriver que dans certaines situations (généralement lorsque le point de vue provoque forte perspective par rapport à un plan), les données rectifiées n'aient pas une résolution suffisante pour pouvoir être utilise comme source lors du processus de complétion. Nous proposons donc d'attribuer à chaque pixel de l'image rectifiée un critère de confiance basé sur les caractéristiques comme la distance par rapport à la position de la caméra ou le produit scalaire entre les normales du plan considéré et du plan image.
Grâce à ce critère de confiance, nous pouvons créer une relation d'ordre entre deux données qui seraient projetées sur le même pixel de l'image rectifiée. De plus, nous pouvons définir un seuil pour éliminer les données trop approximées. Nous pouvons également utiliser ce critère de confiance dans le cas de techniques de recouvrement comme l'inpainting basé patch [BDTL15]. Généralement, un calcul de priorité P_p est effectué pour tous les pixels présents sur la frontière entre le masque et la source qui est basé sur la confiance C et le terme aux données D. Nous proposons d'y rajouter notre critère de confiance C^{3D}.

Approche stratégique de complétion La complétion d'une texture est soumise à deux contraintes : le type de la texture ainsi que la quantité de données utilisables pour compléter. Dans le premier cas, si la texture est presque régulière, voire régulière (ex : carrelage), il est préférable d'effectuer une synthèse de texture par coupure de graphe [KSE*03] pour remplir l'image rectifiée. Cette dernière approche sera également préférée si nous disposons de peu d'informations. À l'inverse, nous utiliserons une approche d'inpainting si la texture est stochastique.
Une manière de détecter si une texture est régulière ou non peut consister à regarder la distribution statistique [HS12]

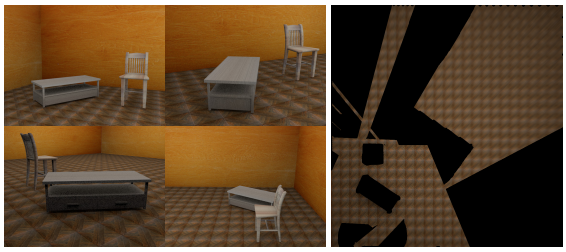

Figure 6: Exemple d'image rectifiée globale (ici avec le sol).

3.2. Édition

Une fois l'étape de compréhension de la scène effectuée, nous générons l'image rectifiée globale pour chaque plan majeur de la scène (sol, murs, plafond). Pour chaque image, nous complétons les zones masquées par les différents objets de la scène. La méthode de complétion est déterminée selon la stratégie expliquée au-dessus. Pour déplacer un objet, nous récupérons ensuite ses points 3D ainsi que les masques 2D de chaque point de vue. Nous effectuons ensuite une transformation T correspondant au déplacement. Pour chaque point de vue, la zone masquée dans l'image 2D avant transformation est complétée avec les informations de chaque image rectifiée de chaque plan visible. Nous profitons du caractère dense du nuage de points pour projeter dans chaque image les points de l'objet après transformation. Cependant, si nous rapprochons l'objet du point de la caméra, les projetés des points 3D seront plus dispersés. Pour gérer cela, nous effectuons sur le rendu un inpainting basé diffusion [ESQD05] pour compléter les pixels non colorés. La Figure 7 illustre le déplacement d'un meuble sur des données réelles.

4. Perspectives et conclusion

Bien que la chaîne globale du processus d'édition fonctionne, certains points restent à améliorer. Si l'ensemble a été testé sur des scène virtuelles complètes et des scènes réelles simples, il reste à tester la robustesse sur des scènes réelles complètes notamment concernant la création de l'image rectifiée globale (calage des masques rectifiés de plusieurs points de vue d'un même plan).
Certains objets n'auront pas de représentation 3D complète (notamment ceux accolés aux murs), une rotation d'un angle élevé ne pourra se faire car la partie masquée quelque soit le point de vue sera alors visible. Une idée consisterait à générer un modèle complet [LSWZ16] grâce à une base de données de meubles classés par catégories. Il est possible également d'améliorer au préalable les cuboïdes représentant le meuble par une approche comme celle de Shao at al. [SMZ*14] qui proposent d'étendre certains cuboïdes de la scène qui représentent des objets incomplets car occultés en partie par d'autres en devinant les contacts entre eux. Un autre type de relation qui représente la stabilité physique de la scène ("touche", "fixé", "sans extension") est alors utilisé.
Un autre problème qui n'est pas mis en évidence concerne la quantité de données que nous possédons pour reconstruire une image rectifiée. Dans une scène réaliste, nous risquons de ne pas disposer de suffisamment d'informations pour effectuer une technique d'inpainting optimale. Une idée serait de considérer cette quantité dans la stratégie de complétion pour favoriser une re-synthèse de textures même si celle-ci n'est pas régulière comme suggéré en 3.1.1. Une dernière perspective concerne également la gestion de la luminosité notamment en calibrant l'exposition de chaque prise de vue.

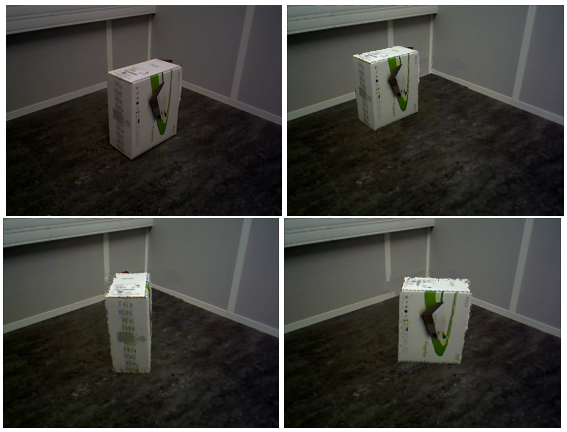

Figure 7: Déplacement d'un meuble. En haut, à gauche : position d'origine ; en haut, à droite : translation. En bas : rotation de $\pm 45°$.

Remerciements Ce travail a été mené dans le cadre de la subvention CIFRE ANRT 2016/0139 et du projet Region CNRS REALISM 15056689

Références

[Arb06] ARBELAEZ P. : Boundary extraction in natural images using ultrametric contour maps. In *Proc. CVPRW 2016* (Washington, DC, USA, 2006), CVPRW '06, IEEE Computer Society, pp. 182–.

[BDTL15] BUYSSENS P., DAISY M., TSCHUMPERLE D., LEZORAY O. : Exemplar-based inpainting : Technical review and new heuristics for better geometric reconstructions. *IEEE TIP. Vol. 24*, Num. 6 (June 2015), 1809–1824.

[BSFG09] BARNES C., SHECHTMAN E., FINKELSTEIN A., GOLDMAN D. B. : PatchMatch : A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH). Vol. 28*, Num. 3 (août 2009).

[CPT03] CRIMINISI A., PEREZ P., TOYAMA K. : Object removal by exemplar-based inpainting. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on* (June 2003), vol. 2, pp. II–721–II–728 vol.2.

[CS01] CHAN T. F., SHEN J. : Nontexture inpainting by curvature-driven diffusions. *Journal of Visual Communication and Image Representation. Vol. 12*, Num. 4 (2001), 436 – 449.

[DTL16] DENG Z., TODOROVIC S., LATECKI L. J. : Unsupervised object region proposals for rgb-d indoor scenes. *CVIU* (2016), In Press.

[ESQD05] ELAD M., STARCK J.-L., QUERRE P., DONOHO D. : Simultaneous cartoon and texture image inpainting using morphological component analysis (mca). *Applied and Computational Harmonic Analysis. Vol. 19*, Num. 3 (2005), 340 – 358.

[HS12] HE K., SUN J. : Statistics of patch offsets for image completion. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part II* (Berlin, Heidelberg, 2012), ECCV'12, Springer-Verlag, pp. 16–29.

[KAS10] KORKALO O., AITTALA M., SILTANEN S. : Light-weight marker hiding for augmented reality. In *Mixed and Augmented Reality (ISMAR), 2010 9th IEEE International Symposium on* (Oct 2010), pp. 247–248.

[KSE*03] KWATRA V., SCHÖDL A., ESSA I., TURK G., BOBICK A. : Graphcut textures : Image and video synthesis using graph cuts. In *ACM SIGGRAPH 2003 Papers* (New York, NY, USA, 2003), SIGGRAPH '03, ACM, pp. 277–286.

[KSY15] KAWAI N., SATO T., YOKOYA N. : Diminished reality based on image inpainting considering background geometry. *Visualization and Computer Graphics, IEEE Transactions on. Vol. PP*, Num. 99 (2015), 1–1.

[KY12] KAWAI N., YOKOYA N. : Image inpainting considering symmetric patterns. In *Proc. ICPR 2012* (2012), IEEE, pp. 2744–2747.

[LSWZ16] LI D., SHAO T., WU H., ZHOU K. : Shape completion from a single rgbd image. *IEEE Transactions on Visualization and Computer Graphics. Vol. PP*, Num. 99 (2016), 1–1.

[NSF12] NATHAN SILBERMAN DEREK HOIEM P. K., FERGUS R. : Indoor segmentation and support inference from rgbd images. In *ECCV* (2012).

[RPMK13] RAVI S., PASUPATHI P., MUTHUKUMAR S., KRISHNAN N. : Image inpainting techniques - a survey and analysis. In *Innovations in Information Technology (IIT), 2013 9th International Conference on* (March 2013), pp. 36–41.

[RZS*16] RONG Y., ZHENG Y., SHAO T., YANG Y., ZHOU K. : An interactive approach for functional prototype recovery from a single rgbd image. *Computational Visual Media* (2016), 1–10.

[Sil15] SILTANEN S. : Diminished reality for augmented reality interior design. *The Visual Computer* (2015), 1–16.

[SMZ*14] SHAO T., MONSZPART A., ZHENG Y., KOO B., XU W., ZHOU K., MITRA N. J. : Imagining the unseen : Stability-based cuboid arrangements for scene understanding. *ACM Transactions on Graphics. Vol. 33*, Num. 6 (2014), 209 :1–209 :11.

[Tsc16] TSCHUMPERLE D. : G'mic - greyc's magic for image computing : A full-featured open-source framework for image processing, 2016.

[WCM15] WONG Y.-S., CHU H.-K., MITRA N. J. : Smartannotator an interactive tool for annotating indoor rgbd images. *Computer Graphics Forum. Vol. 34*, Num. 2 (2015), 447–457.