Innenraumrekonstruktion aus semantisch angereicherten 3D Punkten und Linien

DOROTA IWASZCZUK[1], TOBIAS KOCH[1] & UWE STILLA[1]

Zusammenfassung: In diesem Artikel präsentieren wir ein Konzept zur semantikbasierten Innenraumrekonstruktion aus überlappenden Bildserien. Herkömmliche Verfahren zur Erzeugung von bildbasierten 3D Punkten und 3D Linien erschweren besonders in schwach texturierten Innenräumen die Detektion von Begrenzungsflächen und resultieren meist in unvollständige Gebäudemodelle. Durch eine semantische Interpretation dieser 3D Strukturen können relevante 3D Objekte identifiziert werden um ein vollständiges 3D Raummodell zu erzeugen. Unsere Methode basiert auf einer pixelweisen semantischen Segmentierung der Bilder deren Klasseninformationen auf 3D Strukturen, wie 3D Punkte und 3D Linien, übertragen werden. Dabei werden den 3D Punkten und 3D Linien die Signaturen aus der Bildklassifikation zugewiesen. Aus diesen signierten Merkmalen werden die Innenräume eines Gebäudes rekonstruiert. Dabei werden die Hypothesen für die zu rekonstruierende Ebenen aus 3D Linien einer Klasse gebildet und mit den 3D Punkten derselben Klasse unterstützt. Die optimale Konfiguration aus den Ebenen wird unter Berücksichtigung der Orthogonalität und Parallelität iterativ gesucht.

1 Einleitung

Aufgrund der Urbanisierung und der damit einhergehenden stetigen Zunahme der Einwohneranzahl in städtischen Bereichen steigt die Nachfrage städtische Gebiete mit hohem Detaillierungsgrad zu modellieren. Von besonderem Interesse sind dabei Gebäude, da sich Menschen überwiegend in Gebäuden aufhalten und deren Aktivitäten dort konzentrieren.
In den letzten Dekaden sind zahlreiche Methoden entstanden die eine automatische 3D Gebäuderekonstruktion ermöglichen. Die Detaillierungsgrade erstrecken dabei von einfachen Klötzchenmodellen (LOD 1), über erweiterte Modelle mit zusätzlicher Modellierung von Dachstrukturen und einfachen Texturierungen (LOD 2), bis hin zu komplexen 3D Modellen mit vollständigen und texturierten Fassadenstrukturen (LOD 3). Ein aktueller Trend besteht in der Gebäudeinnenraumrekonstruktion, die zu einer Generierung von LOD 4 Gebäudemodellen dienen können. Besonders in den letzten Jahren wird diesem Bereich größere Aufmerksamkeit in der Forschung und Praxis gewidmet (COHEN et al. 2016).
Moderne Methoden aus dem Bereich der Computer Vision (SNAVELY et al. 2007; ENGEL et al. 2014; ENGEL et al. 2016) verfügen über robuste Verfahren um 3D Geometrien als fotorealistische 3D Punktwolken zu rekonstruieren, und die Objekte effektiv in 3D darzustellen. Bisher steht allerdings meist die visuelle Betrachtung von solchen Datensätzen im Vordergrund, während automatische Analysen dieser Daten oft vernachlässigt werden. Erschwerend für die Rekonstruktion von Innenräumen ist die oftmals fehlende oder schwach ausgeprägte Textur

[1] Technische Universität München, Ingenieurfakultät Bau Geo Umwelt, Arcisstr. 21, D-80333 München, E-Mail: [Dorota.Iwaszczuk, Tobias.Koch, Stilla]@tum.de

besonders bei Wänden und Decken. Dies resultiert meist in unvollständige Innenraummodelle, die eine geometrische als auch semantische Interpretation erschwert.
Innenraumrekonstruktion ist ein großer Bestandteil der Forschung mit mobilen RGBD-Kameras (WANG et al. 2016; NEVEROVA et al. 2013; CHOI et al. 2015). Die Verwendung dieser aktiven Sensoren ermöglicht auch bei texturlosen Oberflächen das Erzeugen von sehr dichten 3D Punktwolken, wodurch die Schätzung der Raumbegrenzungsflächen einfacher ist. In diesem Artikel beschränken wir uns jedoch auf die Verwendung von herkömmlichen RGB Kameras, die einen größeren Einsatzbereich unserer Methode abdecken.
Ein Grund für die meist unzufrieden stellenden Ergebnisse einer Innenraumrekonstruktion auf Basis von RGB Bildern ist, dass meist keine semantische Information verwendet wird und die als Punktwolken gespeicherte Geometrien keine Zusammenhänge darstellen. Um solche Analysen zu ermöglichen, sollen die im Innenraumbereich erzeugten Geometrien mit existierenden 3D Gebäudemodellen kombiniert werden. Dafür wird typischerweise eine Repräsentation als Polygone benötigt. Semantische Informationen, die aus den Bildern gewonnen werden, sollen nun zur besseren und vollständigeren Repräsentation von Innenraummodellen verwendet werden.

2 Konzept

In diesem Beitrag präsentieren wir ein Konzept zur 3D Polygonrekonstruktion von Gebäudeelementen aus Bilderserien von Innenräumen. Eine Übersicht über den Workflow der Methode ist in Abb. 1 dargestellt. Aus einer Vielzahl von Bildern des Innenraumes werden zum einen 3D Strukturen berechnet und zum anderen eine pixelweise semantische Segmentierung der Bilder vorgenommen. Die 2D Klassifikationsergebnisse werden anschließend auf die 3D Strukturen übertragen. Von besonderem Interesse sind hierbei Elemente, wie Wände, Decken, Böden, Türen und Fenster, während Mobiliar, Personen und andere Störobjekte vom Rekonstruktionsprozess ausgeschlossen werden. Aus den verbleibenden 3D Strukturen werden mehrere Ebenenhypothesen geschätzt und mithilfe zusätzlicher Bedingungen eine sinnvolle Konfiguration dieser Ebenen bestimmt.

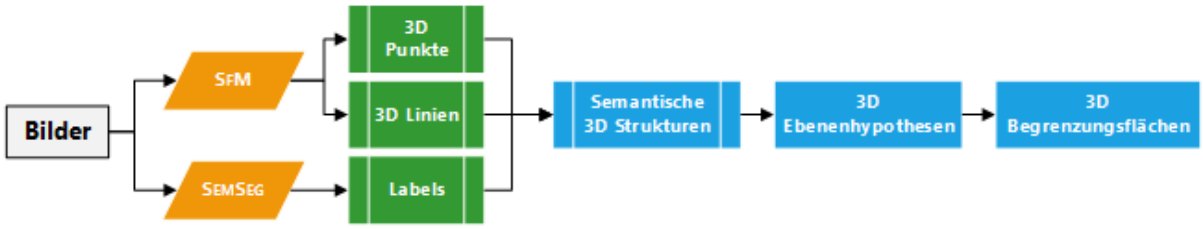

Abb. 1: Workflow zur semantischen Innenraumrekonstruktion. Aus einem Set von überlappenden Bildern werden Kamerapositionen und anschließend 3D Punkte und 3D Linien berechnet. Die Klassifikationsergebnisse einer semantischen Segmentierung aller Bilder in Bezug auf Raumbegrenzungsflächen (Wand, Boden, Decke, Tür, Fenster) werden auf die 3D Strukturen übertragen. Nach der Eliminierung irrelevanter Strukturen wird eine Vielzahl von 3D Ebenenhypothesen aufgestellt. Eine ML-Methode unter Berücksichtigung geometrischer Bedingungen wird zur Bestimmung einer sinnvollen Ebenenkonfiguration angewendet.

2.1 2D semantische Segmentierung

Um eine semantische Segmentierung der Bilder durchzuführen, stehen heutzutage zahlreiche Methoden zur Verfügung (CIREŞAN et al. 2011; GERKE & XIAO 2013; ARMENI et al. 2016). Großes Potential birgt hierfür ein überwachter Klassifikator in Form eines konvolutionalen neuronalen Netzes, um eine pixelweise Klassifikation aller Eingabebilder mit den gewünschten Klassen zu ermöglichen. Als Trainingsdaten für unsere Problemstellung stehen öffentlich zugängliche Datensätze zur Verfügung, wie z.B. in (LIU et al. 2015) mit über 1500 annotierten Bildern der Klassen Fenster und Türen von unterschiedlichen Raum- und Bebauungstypen. Da insgesamt jedoch nur eine geringe Anzahl von Datensätze für die Innenraumklassfikationen zur Verfügung steht, müssen eigens annotierte Bilder zur Verbesserung der Klassifikationsergebnisse hinzugefügt werden.

Eine vielversprechende Methode zur pixelweisen semantischen Segmentierung stellt *SegNet* dar (KENDALL et al. 2015). Neben einer pixelweisen Klassifikation von unbekannten Eingabebildern werden durch die Verwendung eines Softmax-Klassifikators in der letzten Schicht des konvolutionalen neuronalen Netzwerkes ebenfalls Wahrscheinlichkeitsverteilungen aller Klassen pro Pixel ausgegeben. Dies ermöglicht die Ausgabe von Unsicherheiten für alle klassifizierten Pixel und führt im nächsten Schritt zu robusteren Klassifikationsergebnissen der 3D Strukturen. Für unsere Anwendung müsste dieses Netz nochmals mit einer Vielzahl neuer Daten und Klassen antrainiert werden. Abb. 2 stellt beispielhaft die Ergebnisse der semantischen Bildsegmentierung für zwei Bilder dar.

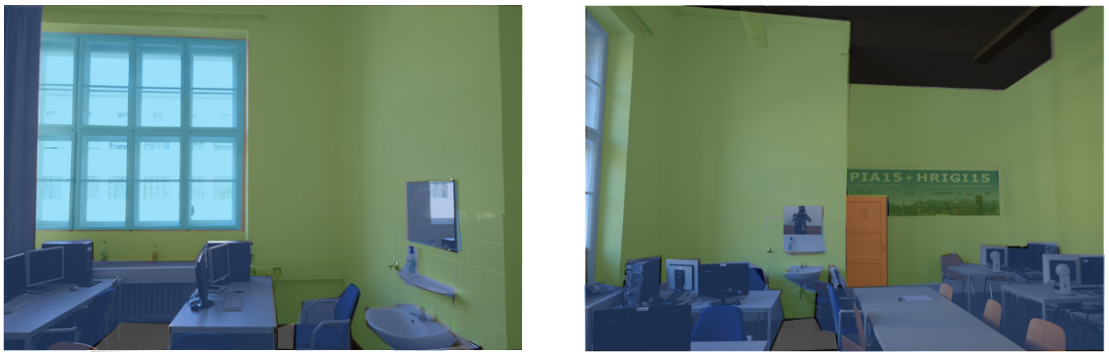

Abb. 2: Beispielergebnisse der 2D semantischen Segmentierung. Jedes Bild wird mit einem vorher antrainierten Klassifikator pixelweise segmentiert. Die Klassen Wand (grün), Boden (grau), Decke (schwarz), Tür (rot) und Fenster (cyan) werden für die Bestimmung der Raumbegrenzungsflächen verwendet während die letzte Klasse Objekt (blau) zur Eliminierung unbedeutender Störobjekte dient.

2.2 Initiale 3D Modellierung der Innenräume

Für die 3D Rekonstruktion von Innenräumen werden zunächst herkömmliche Verfahren aus den Bereichen Photogrammmetrie und Computer Vision verwendet. Aus einer Reihe von überlappenden Bildern des zu modellierenden Raumes werden zunächst mithilfe von *Structure-from-Motion* (SfM) Verfahren interne und externe Orientierungen der einzelnen Bilder berechnet die im nächsten Schritt zur Generierung von 3D Strukturen in Form von 3D Punkten und 3D Linien dienen. Aus einer Vielzahl bestehender SfM Implementierungen wird hier *VisualSfM* (WU 2011) verwendet.

2.2.1 3D Punkte

Die 3D Rekonstruktion erfolgt zunächst durch eine dichte 3D Punktwolke. Hierbei steht mit *Semi-Global-Matching* (SGM) weiterhin eine aktuelle Methode für das Erzeugen einer dichten 3D Punktwolke zur Verfügung. Hierfür werden mithilfe der geschätzten internen und externen Kameraparametern aus dem SfM Schritt und den Eingabebildern pixelweise Korrespondenzen in Bildpaaren gesucht, die anschließend trianguliert werden um eine dichte Punktwolke zu erzeugen. Dies setzt jedoch gut texturierte Oberflächen in den Bildern voraus, welches für Innenräume durch oftmals weiße Wände nicht immer gegeben ist. Als Folge entstehen oft große Lücken in den 3D Punktwolken an den Wänden und Decken des Raumes, welches das Bestimmen von 3D Begrenzungsflächen der Räume erschwert. Eine dichte Punktwolke eines Büros ist in Abbildung 3 dargestellt. Für die Generierung der dichten Punktwolke wird die Methode von (ROTHERMEL et al. 2012) verwendet.

2.2.2 3D Linien

Zur Unterstützung der 3D Strukturen werden neben einer dichten 3D Punktwolke ebenfalls 3D Linien rekonstruiert. Auch hierfür werden die Kameraparameter und Ursprungsbilder verwendet um korrespondierende 2D Linien in Bildpaaren zu identifizieren und anschließend zu triangulieren. Obwohl auch hier an schwach texturierten Flächen nur wenige Linienkorrespondenzen gefunden werden können, ermöglichen wenige 3D Linien eine robustere Schätzung von Begrenzungsflächen wie Wände oder Decken als die verrauschten 3D Punkte (Abb. 3). Für die Erstellung der Linien wird die Methode von (HOFER et al. 2016) verwendet.

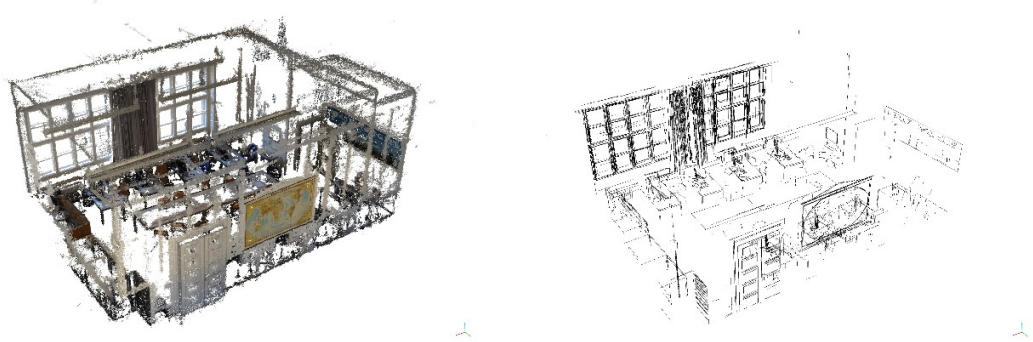

Abb. 3 Erzeugte 3D Strukturen eines Büros. Links: dichte 3D Punktwolke und rechts: 3D Linien

2.3 Verwendung der semantischen Signaturen in 3D Strukturen

Die im vorherigen Schritt erzeugten 3D Punkte und Linien enthalten neben den für die Ermittlung von Begrenzungsflächen relevanten Strukturen (wie Wände, Türen oder Fenster) vorwiegend Strukturen von zunächst unbedeutenden Objekten wie Mobiliar oder anderen Gegenständen. Da diese für die Bestimmung der Raumstruktur hinderlich sind kann die semantische Information aus dem ersten Schritt herangezogen werden.

Da sowohl die 3D Punkte als auch 3D Linien aus korrespondierenden 2D Punkten bzw. 2D Linien in Bildpaaren generiert wurden und für jedes Bild eine pixelweise Klassifikation vorliegt, können

den 3D Strukturen diese semantischen Informationen aus den Segmentierungsergebnissen zugewiesen und als Label gespeichert werden.
Für jeden 3D Punkt wird zunächst die jeweilige Klasse und Klassifikationsgenauigkeit des korrespondierenden 2D Punktes aus allen Bildern in denen dieser 3D Punkt gesehen worden ist gespeichert. Schließlich wird aus der Liste aller Klassifikationsergebnissen inklusive Genauigkeiten eine finales Label geschätzt. Für die Klassenzuordnung der 3D Linien wird ähnlich verfahren, indem jede 3D Linie in diejenigen Bilder, in denen diese 3D Linie sichtbar ist, projiziert wird. Diese 2D Linie gilt oftmals als Übergang zweier Klassen und somit werden für jede 3D Linie zwei Klassen gespeichert, und zwar diejenigen, die in der lokalen Nachbarschaft auf beiden Seiten der 2D Linie vorkommt.
Ein Beispiel dieser Klassenzuordnung ist in Abbildung 4 dargestellt.

Abb. 4: Projektion (gelb) eines 3D Punktes (links) und einer 3D Linie (rechts) in ein Bild. Links: Klassenzuordnung des 3D Punkts entsprechend des Klassifikationsergebnisses an dieser Bildstelle (hier: blau für Objekt). Rechts: Zwei Zuordnungen entsprechend beider Seiten 2D Linie (hier: rot für Tür und grün für Wand).

Da für die Bestimmung der Begrenzungsflächen von Innenräumen lediglich 3D Strukturen mit den Klassen Wand, Boden, Decke, Fenster oder Tür von Bedeutung sind, werden zunächst alle 3D Strukturen, die keine dieser Klassen beinhalten, entfernt.

2.4 Rekonstruktion der Innenräume

Aus der dichten 3D Punktwolke und der 3D Linien sollen nun Polygone geschätzt werden, die einen geschlossenen Raum bilden. Es wird angenommen, dass die Wände, Decken und Böden orthogonal und parallel zueinander stehen und sich in den Raumecken schneiden. Der Vorteil der gemeinsamen Verwendung von 3D Punkten und 3D Linien ist, dass die Punkte eine dichtere Rekonstruktion der Flächen liefern, die Linien dagegen häufig in Raumecken extrahiert und rekonstruiert werden. Dies erlaubt es die Begrenzungsebenen robuster und akkurater zu schätzen. Zusätzlich kann in Innenräumen häufig von einer Manhattan-World Annahme ausgegangen werden in der die Begrenzungsflächen entlang der drei Hauptrichtungen verlaufen. Diese Annahme wird auch hier für die Schätzung der Raumbegrenzungsflächen verwendet.

2.4.1 Aufstellen von Ebenenhypothesen

Um die Ebenenhypothesen zu ermitteln werden zunächst nur diejenigen 3D Punkte und 3D Linien betrachtet, die auf Basis der Signaturen in der Wand-, Decke- und Bodenebene zu liegen (3D Strukturen mit den Signaturen "Wand", "Decke", "Boden", „Fenster“ oder „Tür“). Anschließend werden die drei Hauptrichtungen der 3D Linien ermittelt um die zu erwartete Ausrichtung der rekonstruierten Ebenen zu bestimmen. Hierfür werden zunächst die Richtungsvektoren aller 3D Linien berechnet und mit einem k-Means Verfahren in drei Gruppen aufgeteilt. Mithilfe eines RANSAC Verfahrens können die drei Cluster der Hauptrichtungen robust geschätzt werden. Die Mittelwerte der Richtungsvektoren der drei Cluster ergeben deren Hauptrichtungen. Abb. 5 und Abb. 6 verdeutlichen dieses Verfahren.

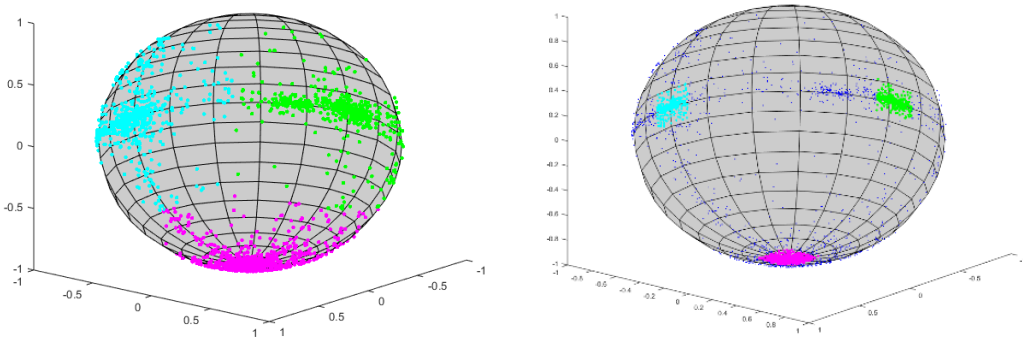

Abb. 5: Bestimmung der Hauptrichtungsvektoren (cyan, magenta, grün) aus den 3D Linien. Hauptrichtungen dargestellt auf einer Sphäre nach (links) k-Means Clustering und (Mitte) RANSAC Verfahren.

Anschließend werden die 3D Linien jeder Hauptrichtung in Ebenen gruppiert. Für diese Gruppierung werden die Mittelpunkte der 3D Linien und die Richtungsvektoren verwendet und ebenfalls ein k-Means Verfahren zur Gruppierung verwendet. Das Ergebnis der Liniengruppierung zeigt Abb. 7. Für jedes gruppierten Liniencluster werden 3D Punkte in derselben Ebene gesucht. Aus den Linien und Punkten eines Clusters werden anschließend die Parameter einer 3D Ebene geschätzt, die als Ebenenhypothese weiter bearbeitet werden.

2.4.2 Festlegung der Raumbegrenzungsflächen

In diesem Schritt wird jede Ebenenhypothese mit allen anderen Hypothesen auf Identität geprüft und bei Bedarf zusammengefügt. Die Prüfung erfolgt mit einem Schwellwertverfahren auf Basis des Winkels zwischen beiden Ebenennormalen und dem Unterschied im vierten Ebenenparameter. Für jede Hypothese wird die Anzahl der zugehörigen 3D Punkte und 3D Linien mitgespeichert, sowie auf Übereinstimmung mit den Hauptrichtungen geprüft. Aus diesen drei Größen wird pro Ebenenhypothese ein Faktor berechnet, der die Qualität dieser Hypothese bestimmt. Die Suche nach der besten Konfiguration aus allen Ebenen erfolgt iterativ. Für jede Ebenenhypothese werden weitere orthogonale und parallele Ebenen gesucht und der Qualitätsfaktor aller diesen Ebenen summiert. Als Endergebnis wird eine Konfiguration gewählt, die die den höchsten Qualitätsfaktor aufweist.

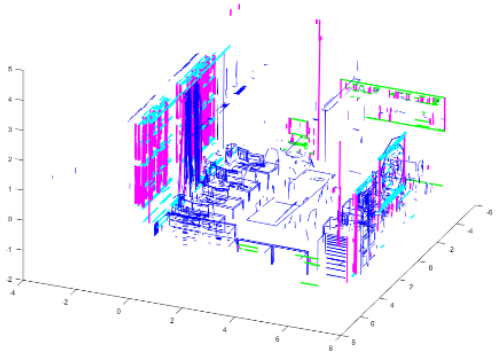

Abb. 6: 3D Linien mit Farbkodierung der drei Hauptrichtungen. In blau werden Linien dargestellt, die zu keiner der drei Richtungen gehören

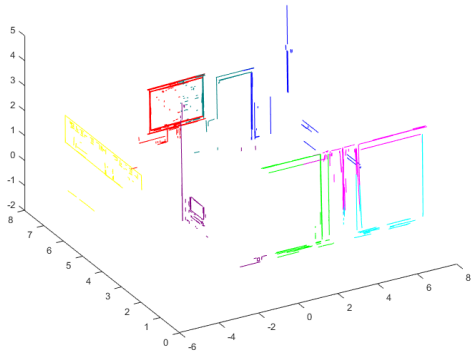

Abb. 7: Gruppierte 3D Linien für die drei Hauptrichtungen

3 Erste Ergebnisse

Um die vorgestellten Methoden zu evaluieren, wurden Bilder einer kalibrierten RGB-Kamera von einem Innenbereich eines Bürogebäudes verwendet (Abb. 8).

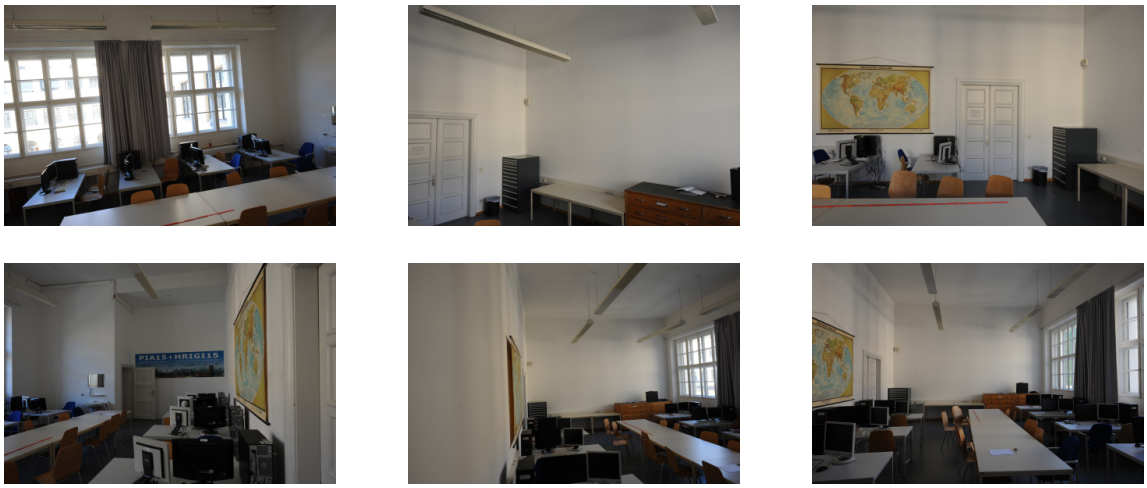

Abb. 8: Beispielbilder des Innenraumes

Aus den in Abbildung 8 präsentierten Bildern wurden 3D Punkte und 3D Linien rekonstruiert und bereits in Abbildung 3 vorgestellt. Die Klassifikation erfolgte zunächst manuell. Abbildung 9 stellt die klassifizierte 3D Punktwolke (links) und die klassifizierten 3D Linien (rechts) dar. Danach wurden die Hauptrichtungen mittels k-Means-Clustering und RANSAC Verfahren bestimmt. Die Ergebnisse dieser Suche wurden bereits in Abb. 5 und Abb. 6 gezeigt. Im Anschluss wurden alle 3D Linien, die den Hauptrichtungen entsprechen, in Cluster gruppiert und 3D Punkte gesucht, die in diesen Ebenen liegen. Der Schwellwert bei dem ein Punkt zu einer Ebene gehört wurde auf 10 cm festgelegt.

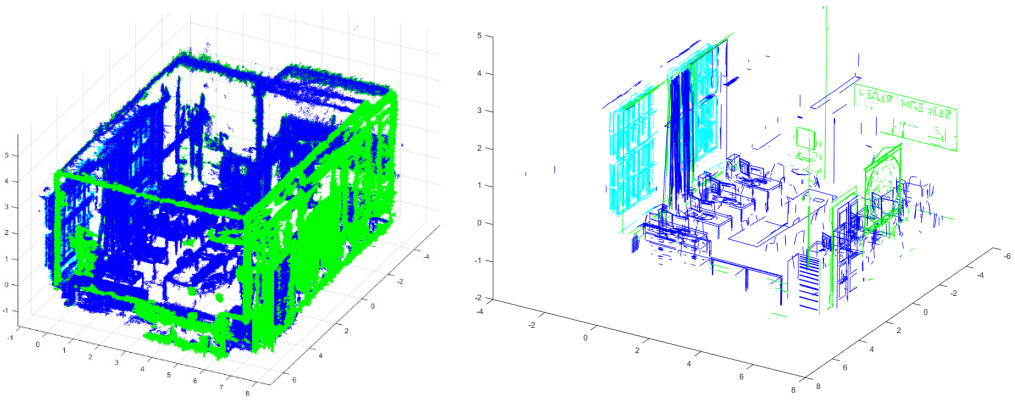

Abb. 9: Klassifizierte 3D Merkmale. Links: 3D Punkte; rechts: 3D Linien. Farbkodierung: grün – „Wand“, cyan – „Fenster“, blau – „Objekt“

Aus den Ebenenhypothesen wurden Ebenenkonfigurationen gebildet und die beste Konfiguration ausgewählt. Das Endergebnis ist in der Abbildung 10 zu sehen.

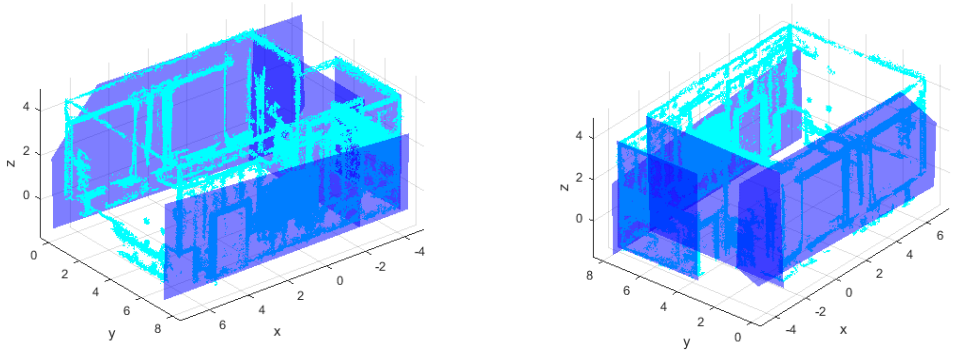

Abb. 10: Rekonstruierte Raumbegrenzungsflächen (blau) mit abgedichtete Punktwolke (cyan).

4 Fazit & Ausblick

In diesem Artikel präsentierten wir ein Konzept für eine Methode zur Innenraumrekonstruktion auf Basis von 3D Strukturen, die aus Bilderserien generiert wurden. Mithilfe einer semantischer Segmentierung der Bilder können die relevanten 3D Strukturen aus den rein geometrischen 3D Primitiven extrahiert werden und zu einer robusteren Bestimmung von Raumbegrenzungsflächen verwendet werden. In Zukunft soll neben der Realisierung der semantischen Segmentierung mittels eines konvolutionalen neuronalen Netzes die geometrischen Unsicherheiten der 3D Linien und 3D Punkte mitberücksichtigt werden, wodurch die verwendeten Schwellwerte vermieden werden können. Eine Erweiterung dieser Methode um eine nachträgliche Suche nach fehlenden Ebenen wird angestrebt und kann über eine Bedingung eines geschlossenen Raumes erreicht werden. Zusätzlich können auch Linien verwendet werden, um die Suche nach Raumecken zu unterstützen. Um das Verfahren zu evaluieren sollen in Zukunft die Eckpunkte des Raumes geodätisch aufgenommen und zur Erstellung eines Referenzmodells verwendet werden. Die Parameter der rekonstruierten Wandebenen werden dann mit den Referenzebenen verglichen und die Genauigkeit der Rekonstruktion bestimmt.

5 Literaturverzeichnis

ARMENI, I., SENER O., ZAMIR, A.R., JIANG, H., BRILAKIS, I., FISCHER. M. & SAVARESE, S., 2016: 3D Semantic Parsing of Large-Scale Indoor Spaces. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1534-1543.

CIREŞAN, D.C., MEIER, U., MASCI, J., GAMBARDELLA, L.M. & SCHMIDHUBER, J., 2011: Flexible, High Performance Convolutional Neural Networks for Image Classification. Proceedings of International Joint Conference on Artificial Intelligence **22**(1), 1237.

COHEN, A., SCHÖNBERGER, J.L., SPECIALE, P., SATTLER, T., FRAHM, J.M. & POLLEFEYS, M., 2016: Indoor-Outdoor 3D Reconstruction Alignment. In: European Conference on Computer Vision 2016, Springer International Publishing, 285-300.

ENGEL J., SCHÖPS, T. & CREMERS D., 2014: LSD-SLAM: Large-Scale Direct Monocular SLAM. European Conference on Computer Vision, Springer International Publishing, 834-849.

ENGEL, J., KOLTUN, V. & CREMERS, D., 2016: Direct Sparse Odometry. arXiv:1607.02565

GERKE, M. & XIAO, J., 2013: Supervised and unsupervised MRF based 3d scene classification in multiple view airborne oblique images. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **II-3**/W3, 25-30.

HOFER, M., MAURER, M. & BISCHOF, H., 2016: Efficient 3D Scene Abstraction Using Line Segments. Computer Vision and Image Understanding (CVIU), Elsevier B.V., http://dx.doi.org/10.1016/j.cviu.2016.03.017

SNAVELY, N., SEITZ, S.M. & SZELISKI, R., 2007: Modeling the World from Internet Photo Collections. International Journal of Computer Vision **80**(2), 189-210.

KENDALL, A., BADRINARAYANAN, V. & CIPOLLA, R., 2015: Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. arXiv preprint arXiv:1511.00561

LIU, C., SCHWING, A.G., KUNDU, K., URTASUN, R. & FIDLER, S., 2015: Rent3D: Floor-Plan Priors for Monocular Layout Estimation. Conference on Computer Vision and Pattern Recognition, 3413-3421.

ROTHERMEL, M., WENZEL, K., FRITSCH, D. & HAALA, N., 2012: Sure: Photogrammetric surface reconstruction from imagery. Proceedings LC3D Workshop, Berlin, http://www.ifp.uni-stuttgart.de/publications/2012/Rothermel_etal_lc3d.pdf.

WU, C. 2011: VisualSFM: A visual structure from motion system. http://www.cs.washington.edu/homes/ccwu/vsfm.

WANG, H., WANG, J. & LIANG, W., 2016: Online reconstruction of indoor scenes from rgb-d streams. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3271-3279.

NEVEROVA, N., MUSELET, D. & TRÉMEAU, A., 2013: 21/2 d scene reconstruction of indoor scenes from single rgb-d images. Computational Color Imaging, Springer Berlin Heidelberg, 281-295.

CHOI, S., ZHOU, Q. Y. & KOLTUN, V., 2015: Robust reconstruction of indoor scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5556-5565.