Kollaborative 3D Rekonstruktion von Urbanen Gebieten

Arnold IRSCHARA, Christopher ZACH, Horst BISCHOF und Franz LEBERL

1 Einleitung

In den letzten Jahren sind in der digitale Bildverarbeitung „Computer Vision“ große Fortschritte im Bereich „Wide Baseline Matching“ und „Structure from Motion“ erzielt worden. Robuste Features wie SIFT ermöglichen eine vollautomatisch Berechnung von Punktkorrespondenzen zwischen Bildern von weit separierten Blickwinkeln. Aus der Korrespondenzinformation können anschließend Kameraorientierungen und 3D Struktur abgeleitet werden (Structure from Motion). In diesem Artikel beschreiben wir ein solches bildbasiertes Rekonstruktionssystem das im Rahmen des WikiVienna[1] Projekts entwickelt wurde. Ziel dieses Projekts ist es ein 3-dimensionales Modell der Wiener Innenstadt automatisch zu generieren. Gemäß dem Wiki-Prinzip sollen interessierte Benutzer mit Bilddaten dazu beitragen können, kollaborativ ein 3D Modell der Stadt zu erstellen und zu erweitern. Im Folgenden werden kurz die einzelnen Bildverarbeitungs und Photogrammetrie Schritte beschrieben und erste Resultate präsentiert.

2 Rekonstruktions Pipeline

Die Eingabedaten an das Rekonstruktionssystem sind Bilder von Digitalkameras bzw. Handy-Kameras. Die Bilder können von unterschiedlichen Benutzern / Kameras stammen und unter verschiedenen Beleuchtungsbedingungen aufgenommen werden. Eine Voraussetzung für unser System ist jedoch eine intrinsische Kalibration der Kameras, diese ist notwendig um qualitativ hochwertige Rekonstruktionen zu erstellen und degenerierte geometrische Konfigurationen auszuschließen. Um den Kalibrationsaufwand für den Endnutzer zu minimieren, verwenden wir eine Kalibrationsmethode basierend auf planaren, kodierten Markern. Der Kalibrationsprozess für den Benutzer wird dadurch trivial: die Marker (auf Papier aufgedruckte Kreis-Muster) werden auf einer ebenen Fläche ausgelegt und von verschiedenen Positionen abfotografiert. Aus diesen Bildern werden anschließend die genauen Kameraparameter (Brennweite und Hauptpunkt) bestimmt, sowie die radialen Verzeichnungsparameter abgeleitet. Das Verfahren ist in (IRSCHARA, ZACH & BISCHOF 2007) detailliert beschrieben. Anschließen werden alle Eingabebilder gemäß der Verzeichnungsparameter entzerrt, somit kann für die weitere Berechnung ein einfaches Lochkameramodell angenommen werden. Abbildung 1 gibt einen Überblick über die einzelnen Komponenten des Rekonstruktionsverfahrens.

[1] http://www.vrvis.at/research/projects/wikivienna/

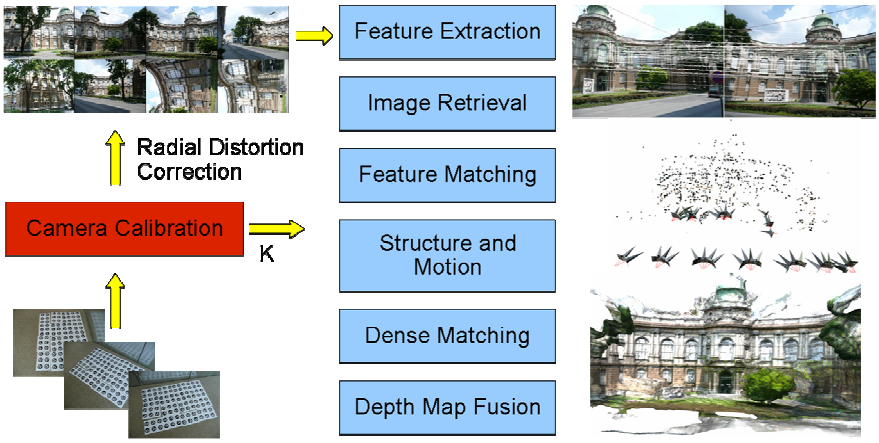

Abb. 1 *Links unten:* Bilder von Kalibrationsmarker. *Mitte:* Komponente des Systems. *Rechts:* Korrespondenzen von SIFT Features, triangulierte Punktwolke nach „Structure from Motion“ und anschließendes dichtes 3D Modell der Szene.

2.1 Berechnung von Korrespondenzen

Im ersten Verarbeitungsschritt werden SIFT Features (LOWE 2004) aus den Eingabebildern extrahiert. SIFT Features sind lokale Bild Deskriptoren auf ausgewählte Positionen im Bild und zeichnen sich dadurch aus dass sie invariant gegenüber Skalierung und Rotation (innerhalb der Bildebene) und auch relative robust gegenüber Beleuchtungsänderungen und geringen Blickwinkeländerungen sind. Um Punktkorrespondenzen zwischen zwei Bildern herzustellen werden wechselseitig die inneren Produkte der 128-dimensionalen normalisierten SIFT Deskriptoren berechnet. Diejenigen Punktpaare werden als Korrespondenzen akzeptiert, die sich wechselseitig als Maxima finden und das Skalarprodukt über einen Schwellwert von 0.95 liegt. Bei einer kontrollierten sequenziellen Aufnahmestrategie (z. B. Videosequenz) könnte die Korrespondenzsuche auf zeitlich aufeinander folgende Nachbar-Bilder beschränkt werden. In unserem Fall werden Bilder jedoch von verschiedenen Benutzern aufgenommen, deshalb müsste eine Korrespondenzsuche zwischen allen $n(n-1)/2$ Bildpaaren wie in (SNAVELY, SEITZ & SZELISKI 2006) vorgenommen werden. Dies ist jedoch mit sehr großem Rechenaufwand verbunden. In unserem System verwenden wir daher eine Technik die ursprünglich aus der Objekterkennung kommt, nämlich eine „Vocabulary Tree“ (Vokabelbaum) Datenstruktur (NISTER & STEWENIUS 2006) um relevante Bilder für die Korrespondenzsuche vorzuselektieren. Für jedes Eingabebild beschränkt sich dadurch das paarweise Bild-zu-Bild Matching auf die ersten k-Bilder die vom Vokabelbaum zurückgeliefert werden. Dadurch reduziert sich die Laufzeit Komplexität auf O(n k) mit (k<< n) und die Skalierbarkeit auf große Bilddatenbanken bleibt gewährleistet.

2.3 Struktur Berechnung

Unsere Eingangsbilder sind kalibriert, das ermöglicht eine direkte und effiziente euklidische Rekonstruktion. Wir verwenden den Fünf-Punkt Algorithmus (NISTER 2004) zusam-

men mit RANSAC um die relative Orientierung zwischen Bildpaaren zu berechnen. Eine robuste Schätzung der Orientierung ist essenziell, da die Korrespondenzberechnung meist viele Ausreißer zurückliefert. Mittels einer linearen Methode werden die Korrespondenzen zu 3D Punkte trianguliert. Weitere Kameras werden durch des Drei-Punkt Verfahren und RANSAC eingefügt. Ein globaler Bündelblockausgleich optimiert anschließen die Kamera Orientierungen P_j und die 3D Punkte X_{ij},

$$\min_{P_j X_{ij}} \sum_{i,j} d(P_j X_{ij}, x_{ij})$$

wobei wir für $d(x, y)$ die robuste Huber-Kostenfunktion verwenden. Unser System arbeitet inkrementell, d. h. Bilder können nacheinander hinzugefügt werden und müssen nicht von vornherein bekannt sein. Wir unterscheiden zwischen drei verschiedenen Möglichkeiten die sich für ein aktuelles Eingabebild ergeben:

1. **Registrierung:** Das Eingabebild kann auf ein bekanntes 3D Modell registriert werden, d. h. es gibt 3D zu 2D Punkt Korrespondenzen. Die aktuelle Kameraorientierung wird in diesem Fall direkt durch den Drei-Punkt Algorithmus und RANSAC bestimmt.
2. **Merge Operation:** Das Eingabebild kann auf mehrere 3D Modelle registriert werden. In diesem Fall transformieren wir die Rekonstruktionen über gemeinsame 3D Punkte in das gleiche Koordinatensystem (mittels robuster 3D Transformation).
3. **Initialisierung:** Es gibt (noch) kein passendes 3D Modell in der Datenbank mit Korrespondenzen zu dem Eingangsbild. Mittels relative Pose (Fünf-Punkt) und absolut Pose (Drei-Punkt) wird ein neues Bild-Tripel initialisiert.

2.4 Dense Matching und Fusionierung

Zum Erstellen von dichten 3D Modellen verwenden wir ein effizientes Verfahren zur Berechnung von Tiefenbildern wie es in (ZACH, SORMANN & KARNER 2006) beschrieben wird. Moderne Grafikkarten ermöglichen eine schnelle und effiziente Verarbeitung. Anschließend werden die einzelnen Tiefenbilder in einem robusten, global optimalen Fusionierungsprozess (ZACH, POCK & BISCHOF 2007) volumetrisch integriert und die Isooberfläche extrahiert.

3 Resultate

Um das System zu testen wurden 2.700 Bilder von interessanten Sehenswürdigkeiten von Wien aufgenommen. Dabei wurden zwei unterschiedliche Kameras verwendet, die Test-Aufnahmen sind an zwei verschiedenen Tagen durchgeführt worden. Über 80% der Bilder wurden erfolgreich registriert, 8 disjunkte 3D Modelle sind dabei entstanden. Abbildung 2 zeigt 3D Punkte und Kamera Orientierung von drei verschiedenen Szenen. Ein dichtes 3D Model des Michaelerplatz ist in Abbildung 3 abgebildet.

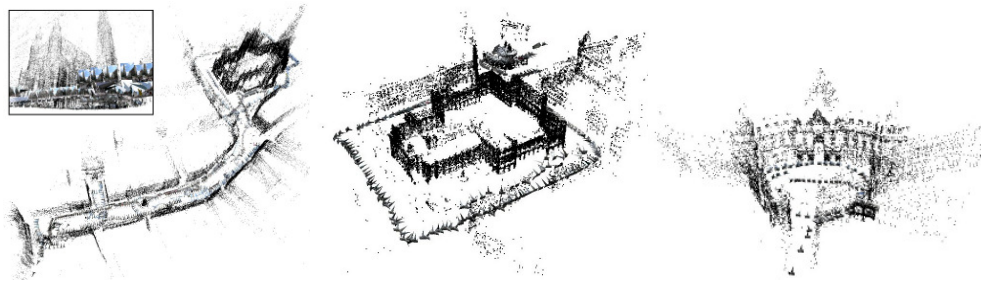

Abb. 2 : *Von Links nach Rechts*: Rekonstruktion von Graben mit Stephansdom (1300 Bilder), Staatsoper (330 Bilder) und Michaelerplatz (112 Bilder) .

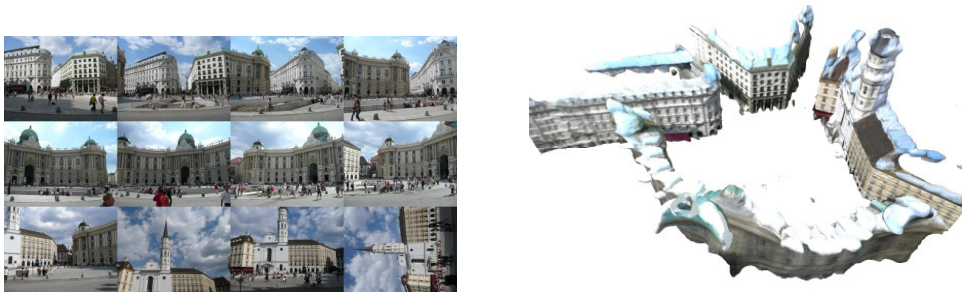

Abb. 3: *Links:* 12 von insgesamt 112 Bildern des Michaelerplatz. *Rechts:* Fusioniertes 3D Modell nach robuster Integration der einzelnen Stereo Tiefenbilder.

Literatur

Irschara, A, C. Zach & H. Bischof (2007): Towards Wiki-based Dense City Modeling. ICCV Workshop on Virtual Rep. and Mod. of Large-scale environments (VRML) 2007.

Lowe, D, (2004): Distinctive Image Features from Scale-Invariant Keypoints. IJCV 2004.

Nister, D, (2004): An Efficient Solution to the Five-Point Relative Pose Problem. PAMI 2004.

Nister, D, & H. Stewenius (2006): Scalable Recognition with a Vocabulary Tree. CVPR 2006.

Snavely, N, S. Seitz & R. Szeliski (2006): Photo Tourism: Exploring image collections in 3D. SigGraph 2006.

Zach, C, M. Sormann & K. Karner (2006): Scanline Optimization for Stereo on Graphics Hardware. 3DPVT 2006.

Zach, C, T. Pock & H. Bischof (2007): A Globally Optimal Algorithm for Robust TV-L1 Range Image Integration. ICCV 2007.