Objekterkennung in Luftbildern mit Methoden der Computer Vision durch kombinierte Verwendung von Redundanz, Farb- und Höheninformation

Stefan KLUCKNER, Georg PACHER, Horst BISCHOF und Franz LEBERL

Zusammenfassung

Die automatische 3D Modellierung von urbanen Lebensräumen für „Location Aware“ Applikationen, wie *Virtual Earth* und *Google Earth*, bietet für die Computer Vision eine enorme Menge an Forschungsthemen. Die Bandbreite reicht von detailgetreuen Rekonstruktionen ganzer Städte, über die Erkennung von gesamten Straßennetzen bis hin zur Detektion sehr kleiner Objekte, wie etwa Fahrzeugen oder Zebrastreifen in hochauflösenden *Microsoft UltraCamX* Luftbildern. Die enormen Datenmengen und die hohe Redundanz stellen dabei an die Computer Vision hohe Anforderungen. Für die Bearbeitung der Bilder werden ausgeklügelte und effiziente Algorithmen benötigt. Diese Arbeit gibt einen Einblick, wie Gebäude, Straßen, Bäume, Wasser- und Grünflächen durch Integration von Farb- und Höheninformation, mit schnellen Methoden der Vision unterschieden werden können. Weiters wird gezeigt, dass sich durch die Verwendung der Redundanz eine verbesserte Segmentierung ergibt. Der vorgestellte Ansatz wird mithilfe der Datensätze *Graz* und *San Francisco* auf Performance und Effizienz mit State-of-the-Art Methoden verglichen.

1 Einleitung

Diese Arbeit beschäftigt sich mit der Erkennung von Objekten zur vollständigen Segmentierung in Luftbildern, unter Einbeziehung verschiedener Modalitäten, wie Farb- und Höheninformation. Zusätzlich kann durch Verwendung der Redundanz die Erkennungsrate der Segmentierung erhöht werden. Im Folgenden werden die verwendeten Modalitäten erklärt, es wird auf die hier verwendete Methode der Objekterkennung eingegangen und die Kombination von Farb- und Höheninformation wird erläutert.

1.1 Luftbilder

Die *Microsoft UltracamX* Kamera liefert Bilder im RGB Farbraum mit einer Auflösung von 11430 x 9420 Pixel und einer „Ground Sampling Distance“ (GSD) von 8 cm bzw. 15 cm. Die Farbbilder werden mit 80 % along-track und 60 % across-track Überlappung aufgenommen. Durch die hohe Überlappung und die Kameradaten kann jeweils für ein Triple von Bildern (Bild plus zwei Nachbarbilder) ein 2.5D Höhenbild berechnet werden. Die Berechnung erfolgt nach einem flächenbasierten Matching Algorithmus und einem semi-globalen Optimierungsprozess (KLAUS, SORMANN & KARNER 2006). In dieser Arbeit werden die Datensätze *Graz* und *San Francisco* verwendet. Der Datensatz *Graz* umfasst ca. 3.000 Bilder und wurde mit einer GSD von 8 cm aufgenommen. Die verwendeten Bilder zeigen hauptsächlich representative Teile einer typischen europäischen

Stadt. *San Francisco* hingegen zeigt periphere Erscheinung mit vielen Vorstadthäusern und enthält ein anspruchvolles Untergrundterrain. Abbildung 1 zeigt einen Ausschnitt eines Farbbildes und eines korrespondierenden Höhenfeldes aus dem Datensatz *San Francisco*.

Abb. 1: Ausschnitt eines RGB Farbbildes und korrespondierender Höheninformation aus dem *San Francisco* Datensatz.

Die Objekterkennung in Luftbildern erfordert aufgrund der enormen Datenmengen (z. B. *Graz* mit 3.000 Bildern zu je 300 Mb) sehr effiziente Methoden. Für die Erkennung von Gebäuden und Straßenzügen werden sogenannte „Random Forest"-Klassifikatoren (RF) (BREIMAN 2001) eingesetzt. Diese Methode ermöglicht eine „Multi Class"-Klassifikation und basiert auf einer Kombination von mehreren Entscheidungsbäumen. Knoten (Nodes) in einzelnen Bäumen enthalten im einfachsten Fall binäre Entscheidungen, wie größer/kleiner Vergleiche (Split Criterions). Diese Vergleiche werden auf zufällig ausgewählte extrahierte Featurewerte angewandt und dienen der Wegweisung durch den Baum zu den Blättern (Leaf Nodes). Die Blätter der resultierenden Bäume enthalten die Wahrscheinlichkeiten der Klassenzugehörigkeit. Ein multiplikatives Zusammenfassen aller Bäume in einem „Forest" resultiert in einer Wahrscheinlichkeitsverteilung über alle vorkommenden Klassen. Eine Beschränkung der maximalen Tiefe der Bäume und die Einfachheit der Vergleiche ermöglichen ein schnelles Lernen von Daten und vor allem ein effizientes Testen auf unbekannte Featurevektoren.

1.2 Kombination von Farb- und Höheninformation

Die Extraktion der Featurevektoren erfolgt direkt auf den Pixel Werten, um die Komplexität der Berechnung und den Speicherbedarf gering zu halten. Da die Höhendaten aus dem Matching Prozess ein „Digital Surface Model" (DSM) darstellen, wird in einem Vorschritt das „Digital Terrain Model" (DTM) über lokale Minima erstellt. Die Differenz zwischen DTM und DSM wird hier als vierter Kanal representativ für die direkte Höheninformation neben den drei Farbkanälen gehandhabt. Hierbei wird, in einer kleinen Umgebung der Größe d x d, ein Featurevektor mit fixer Dimension generiert. Werte wie (a) direkter Pixel Wert $p_{x1,y1,c1}$, (b) Summe aus $p_{x1,y1,c1}+p_{x2,y2,c2}$, (c) Differenz aus $p_{x1,y1,c1}+p_{x2,y2,c2}$ und (d) Betrag der Differenz von $|\, p_{x1,y1,c1}+p_{x2,y2,c2}\,|$ stellen die Einträge des Featurevektors (SHOTTON, JOHNSON & CIPOLLA 2008). Die Pixel Positionen x1, x2, y1 und y2 innerhalb der Umgebung werden zufällig ausgewählt und müssen für den trainierten „Forest" immer bekannt sein. c1 und c2 geben hier den relevanten Kanal an.

2 Objekterkennung in Luftbildern

2.1 Lernen der Daten

Für das Lernen der RFs werden einzelne Luftbilder annotiert (nicht notwendigerweise vollständig), indem dargestellte Regionen den fünf Klassen Gebäude, Straßen, Bäume, Grün- und Wasserflächen zugewiesen werden. Ein extrahierter Featurevektor pro annotiertem Pixel erhält jeweils die Klasse der Region als Trainingslabel und fließt direkt in den Trainingsprozess der RF ein. Es ist anzumerken, dass ein Trainingsprozess im Bereich von nur einigen Minuten liegt.

2.2 Objekterkennung zur vollständigen Segmentierung

Nach dem Trainingsprozess kann der RF direkt auf Luftbilder zur vollständigen Segmentierung angewandt werden. Dazu wird für jeden Pixel ein Featurevektor nach den gleichen Vorschriften der Trainingsvektorenberechnung extrahiert und durch den RF klassifiziert. Diese pixelweise und vollständige Klassifikation erfolgt einzeln auf perspektivischen Bildern. Eine Bildsegmentierung durch Verwendung eines RF Klassifikators benötigt ca. 30 s mit einer unoptimierten Implementierung. Zum Vergleich: Bekannte State-of-the-Art Methoden wie z. B. Support Vector Machines (SVM) liefern eine qualitative Luftbildsegmentierung in einigen Minuten (ZEBEDIN, KLAUS, GRUBER & KARNER 2006). Diese Arbeit berichtet von Erkennungsraten im Bereich 85 bis 95 %. Tabelle 1 zeigt eine Confusion Matrix für die fünf Klassen auf einem manuell erstellten Testset für *San Francisco*. Die Anwendung von RFs als Klassifikatoren ergeben Raten über 92 %.

Tabelle 1: Klassifikationsergebnisse für die fünf Objektklassen auf dem Datensatz *San Francisco* in Prozent.

	Klassifikation				
	Gebäude	Wasserflächen	Grünflächen	Baum	Straße
Gebäude	97.020	0.509	0.028	0.752	1.691
Wasserflächen	1.185	94.948	0.199	1.276	2.391
Grünflächen	0.175	1.358	95.561	1.438	1.467
Baum	3.231	1.806	0.506	94.183	0.274
Straße	0.955	5.550	0.333	0.292	92.869

2.3 Verwendung der Redundanz

Durch die hohe Redundanz der *UltracamX* Bilder (jeder Punkt auf Grund wird aus ca. zehn verschiedenen Ansichten aufgenommen) liegen auch die RF Klassifikationsergebnisse durch die Einzelbildsegmentierung mehrfach vor. Eine Fusionierung der Klassifikationen mittels bekannter Kameradaten in der Orthoprojektionsansicht kann auf einfachste Weise,

über eine Erstellung eines Histogramms H in einer kleinen Umgebung, durchgeführt werden. Dazu wird für jede der fünf Klassen c ein Histogrammeintrag pro Pixel und der Umgebung erstellt und für jedes Auftreten aufakkumuliert. Die normalisierte Verteilung erlaubt eine Berechnung von argmax_k P(k=c|H) und ergibt ein konfidentes Maß für die Klassenzugehörigkeit. Durch die Fusionierung kann die Objekterkennungsrate zusätzlich im Mittel um ca. 3 % gesteigert werden. Abbildung 2 zeigt einen *Graz* Ausschnitt eines vollständig fusionierten Klassifikationsergebnisses (berechnet aus zehn *UltracamX* Einzelbildern) mit eingefärbter Klassenzugehörigkeit.

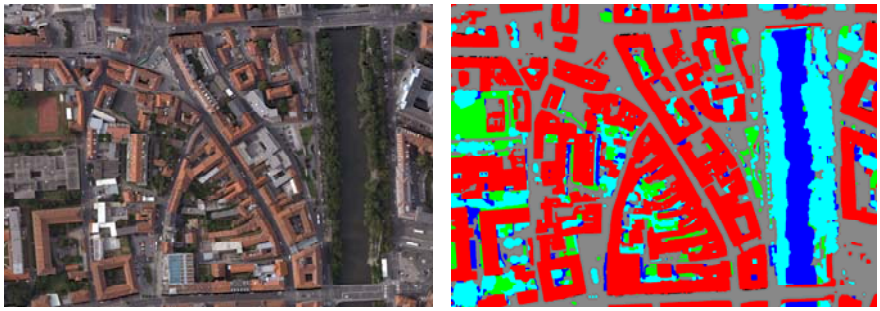

Abb. 2: Farbbild (links) und Klassifikationsergebnis (rechts) für *Graz:* Ausschnitt aus zehn fusionierten Bildern. Klassen: Gebäude (rot), Wasserflächen (blau), Grünflächen (green), Baum (cyan) und Straße (grau).

2.4 Erkenntnisse

Durch den Einsatz von einfachen RF Klassifikatoren und einer unkomplizierten Kombination von Farb- und Höheninformation können State-of-the-Art Erkennungsraten in sehr kurzer Berechnungszeit auf Luftbildern erreicht werden. Probleme bereiten Falsch-Detektionen in Regionen von Schattenbereichen einiger Straßen. Diese werden aufgrund der spektralen Ähnlichkeit als Wasser klassifiziert. Eine Erweiterung des vorgestellten Ansatzes um einen zusätzlichen Informationskanal, wie Infrarot, könnte hier Abhilfe schaffen.

Literatur

Shotton J., Johnson M. & Cipolla R. (2008): Semantic Texton Forests for Image Categorization and Segmentation. In Proc. IEEE CVPR08

Klaus A., Sormann M. & Karner K. (2006): Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In Proc. ICPR06

Breimann Leo (2001): Random Forests. In Machine Learning. Volume 45, pp 5-32

Zebedin L., Klaus A., Gruber-Geymayer B. & Karner K. (2006): Towards 3d map generation from digital aerial images. In Photogrammetry and Remote Sensing. Volume 60, pp 413-427