Technische Hochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Einsatz von Data Mining zur Identifikation und Schätzung der Anzahl von Bananenpflanzen in einem Luftbild

Christoph Gresch

Masterarbeit • Studiengang Informatik • Fachbereich Informatik und Medien • 27.02.2017

Aufgabenstellung

Ziel der Arbeit ist die automatische Detektion und Zählung von Bananenpflanzen in einem Luftbild. Für die Identifizierung der Pflanzen wird ein vollständiger Data Mining-Prozess durchlaufen und abschließend durch ein Clustering ergänzt.

Datengrundlage

Für die Generierung der Wissensbasis wird das Luftbild einer Bananenplantage genutzt. Mithilfe der Geoinformationssysteme QGIS und GDAL werden 1000 Templates von 40x40 Pixeln extrahiert. Dabei werden zwei Klassen berücksichtigt: Bananen und Nicht-Bananen. Ein Template gilt genau dann als Bananen-Template, wenn der Mittelpunkt der Pflanze den Mittelpunkt des Templates bildet. Um eine ausgewogene Datenbasis zu gewährleisten werden verschiedene Szenarien abgebildet (z.B. Überlappungen von Pflanzen). Insgesamt enthält die Wissensbasis 200 Bananen-Templates.

Abb. 1: Beispiel-Templates aus der Datenbasis

Merkmalsextraktion und Merkmalsselektion

Auf den Templates wird eine Reihe von Merkmalen berechnet. Die Merkmale werden in drei Kategorien unterteilt: Farbe, Kontur und Textur. Insgesamt werden 268 Merkmale pro Template extrahiert. Die Farbmerkmale werden über den Histogrammen aller einzelnen RGB- und HSV-Kanäle, sowie dem Grauwertkanal erhoben. Aus allen Kanälen ergeben sich insgesamt 112 Farbmerkmale.

Da neben den Bananenpflanzen keine größeren geschlossenen Konturen in der Plantage existieren, dienen die Konturmerkmale hauptsächlich der Erfassung der Bananenflächen. Weiterhin werden verschiedene Bildmomente betrachtet. Für die Datenbasis werden 84 Konturmerkmale genutzt.

Für die Berechnung der Texturmerkmale wird eine Grauwertematrix genutzt. Diese erfasst vor allem den Kontrast zwischen Pflanzen und Untergrund. Insgesamt werden 72 Texturmerkmale erhoben.

Um irrelevante Merkmale aus der Datenbasis zu entfernen, werden vier Merkmalsselektionsverfahren getestet. Das beste Verfahren, das RFECV, wird für die finale Reduktion des Datenfeldes auf 25 Merkmale genutzt.

Modellauswahl und Prognose

Um die ungleichmäßige Klassenverteilung zu reduzieren, werden 300 Negativ-Beispiele vor der Klassifikation aus der Datenmenge entfernt. Die übrigen 700 Datensätze werden in 70% Trainingsdaten und 30% Testdaten aufgeteilt.

Insgesamt werden 34 Klassifikatoren verschiedener Arten und Settings getestet. Die Bewertung der einzelnen Klassifikatoren erfolgt mithilfe einer fünffachen Kreuzvalidierung. Über alle Durchläufe wird der mittlere CV-Score berechnet. Der beste Klassifikator wird für die Erstellung des 70%-Modells genutzt, das auf die Testdaten angewandt wird. Anschließend wird das 100%-Modell für die Prognose erstellt.

Für das Zielbild wird ein Fenster in Template-Größe pixelweise über das gesamte Bild geführt. Dabei werden für alle Koordinatenpaare die Merkmalsreihen extrahiert.

Clustering

Durch die pixelweise Verschiebung der Templates im Zielbild, die teilweise keine Unterschiede in den Merkmalen zwischen benachbarten Templates verursacht, entstehen um das Bananenzentrum herum größere Punktwolken. Mithilfe des Birch-Clustering, dass die Cluster mittels Distanzfunktion bildet, werden diese Wolken zu einzelnen Bananenpunkten zusammengefasst. Dabei steht jedes Clusterzentrum für einen Bananenpunkt und somit für die Gesamtzahl der Pflanzen im Bild.

Ergebnisse

Eine Support Vector Machine bildet den besten Klassifikator der Kreuzvalidierung und erzielt einen mittleren CV-Score von 82,24%. Das 70%-Modell erzeugt eine allgemeine Erfolgsrate von 88,57% mit 82% true positives und 91% true negatives. Nach dem Clustering werden 234 Bananen gezählt. Das Testbild enthält 242 Bananen. Trotz sehr guter Klassifikationsergebnisse von 88,57% und einer sehr guten Zählung, werden die Pflanzen durch das 100%-Modell schlecht getroffen und erzielen ein visuell unzureichendes Endergebnis.

Abb. 2: Ausschnitt aus dem Ergebnisbild

Fazit und Ausblick

In dieser Arbeit wurde eine automatische Identifikation von Bananenpflanzen in einem Luftbild mithilfe eines Klassifikationsprozesses inklusive Clustering vorgestellt. Die guten Ergebnisse des 70%-Modelles werden durch das 100%-Modell nicht bestätigt. Das Hauptproblem bildet dabei das Clustering, dass zukünftig durch eine Wahrscheinlichkeitsanalyse ersetzt werden soll.

Betreuung Prof. Dr.-Ing. Sven Buchholz • Technische Hochschule Brandenburg • Dr. Frederik Jung-Rothenhäusler • ORCA Geo Services • Dipl.-Inform. Ingo Boersch • Technische Hochschule Brandenburg