Fachhochschule Vorarlberg

University of Applied Sciences

Masterthesis

Fachhochschul-Studiengang

Master Informatik

Implementierung eines Verfahrens zur automatisierten Verkehrsmittelerkennung

Transportation Mode Detection

ausgeführt von

Michael Zangerle, BSc

1310249004

zur Erlangung des akademischen Grades

Master of Science in Engineering, MSc

Dornbirn, im Juli 2015

Betreuer: Prof. (FH) DI Thomas Feilhauer

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich vorliegende Masterthesis selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Stellen sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder in gleicher noch in ähnlicher Form einer anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Dornbirn, am 29. Juli 2015 Michael Zangerle, BSc

Zusammenfassung

Zusammen
ergänzen

Abstract

Abstract e
zen

Inhaltsverzeichnis

1 Einleitung **1**

1.1 Ziele dieser Arbeit 2

1.2 Motivation und Nutzen 4

1.3 Weiterer Aufbau der Arbeit 5

2 State of the Art **7**

2.1 Daten 8

2.2 Verkehrsmittel 8

2.3 Analyse der Publikationen von Yu Zheng 8

2.3.1 Geolife 9

2.3.2 Segmentierung 10

2.3.3 Schlussfolgerungsmodelle 11

2.4 Analyse der Publikationen von Leon Stenneth 12

2.4.1 GIS-Informtionen 12

2.4.2 Schlussfolgerungsmodelle 13

2.5 Analyse der Publikationen von Filip Biljecki 14

2.5.1 Segmentierung 14

2.5.2 Schlussfolgerung 15

2.6 Analyse der Publikationen von Sasank Reddy 16

2.6.1 Sensoren 17

2.6.2 Schlussfolgerungsmodelle 17
2.7 Zusammenfassung 18

3 Modellbildung und Einbindung der GPS- und GIS-Daten 21
3.1 Trainingsdaten 22
3.1.1 Struktur der GPS-Daten 23
3.1.2 Entscheidungsbaum 23
3.2 Neue Aufzeichnungen 28
3.3 GIS-Daten 28
3.4 Weitere Daten 30

4 Der Prototyp 31
4.1 Funktionalitäten 32
4.2 Aufbau und Architektur 32

5 Processing? 33
5.1 Aufbereitung ohne GIS-Daten 33
5.1.1 Auswahl 33
5.2 Aufbereitung mit GIS-Daten 34
5.2.1 Auswahl 34
5.2.2 Abstand zu Bushaltestellen 34
5.2.3 Abstand zu Gleisen 34
5.3 Segmentierung 35
5.3.1 Terminologie 35

6 Analyse 37
6.1 Entscheidungsbaum 38
6.2 Nachbearbeitung 38

7 Auswertung **39**

7.1 Genauigkeit ohne GIS-Daten 40

7.2 Genauigkeit mit GIS-Daten 40

8 Ausblick **41**

Literaturverzeichnis **43**

Quellcodeverzeichnis **47**

Tabellenverzeichnis **49**

Abbilungsverzeichnis **51**

Anhang 1 **53**

Todo list

Zusammenfassung ergänzen i

Abstract ergänzen iii

entscheidungsbaum mit gis daten 28

einbindung der gis daten ins modell 30

verweis auf vorheriges filtern in anhang 32

Einleitung

Der Mensch produziert täglich eine extrem große Menge an Daten. Allein auf Youtube werden pro Minute 300 Stunden Video-Material veröffentlicht und täglich hunderte Millionen Stunden von Videos konsumiert. [Youtube, 2015] Hochgerechnet auf das gesamte Internet und die gesamte Bevölkerung ergibt dies eine unvorstellbar große Menge an Daten, die bewusst oder auch unbewusst generiert werden.

Sehr viel an Daten wird auch durch diverse Fitnessgadgets, Smartwatches, Smartphones sowie Navigations-Geräten und Ähnlichem generiert. Abseits von Fitnesswerten sind viele dieser Geräte im Regelfall in der Lage GPS-Spuren aufzuzeichnen. Dies bedeutet, dass man genau nachvollziehen kann, wann man wo unterwegs war. Mit ein wenig Rechenarbeit kann man auch die Geschwindigkeit und viele andere Werte berechnen, sofern dies die Geräte nicht schon selbst machen.

Genau auf diesen GPS-Daten basiert diese Arbeit. Dabei ist es nicht wichtig, von welcher Person diese Daten stammen, sondern dass sich mit Hilfe dieser aufgezeichneten Daten feststellen lässt, wann ein Individuum sich auf welcher Strecke mit welchem Verkehrsmittel fortbewegt hat.

Analysiert man viele dieser Daten, so lassen sich viele Erkenntnisse daraus gewinnen. Unter anderem lassen sich zum Beispiel viel frequentierte Strecken in der Infrastruktur

finden und mögliche Engstellen erkennen. Neben Engstellen lassen sich damit auch mögliche Verbesserungen, Einsparungen oder auch verstecktes Potential für zukünftige Projekte entdecken.

1.1 Ziele dieser Arbeit

Das Hauptziel dieser Arbeit ist es, einen Prototypen zu erstellen, welcher anhand von aufgezeichneten GPS-Spuren und in Kombination mit verschieden Methodiken das benutzte Verkehrsmittel mit einer möglichst hohen Wahrscheinlichkeit bestimmt. Diese aufgezeichneten GPS-Spuren enthalten dabei keinerlei Informationen über die jeweilige Person. Deshalb erfolgt die Auswertung ausschließlich über die jeweilige GPS-Spur sowie über öffentlich zugängliche Daten, wie zum Beispiel Busstationen und Gleise. Jede dieser Aufzeichnung kann mehrere Verkehrsmittel beinhalten und von unterschiedlicher Länge und Dauer sein.

Die in dieser Arbeit berücksichtigten Verkehrsmittel sind in folgender Liste ersichtlich. Besonders interessant ist hierbei die Unterscheidung von Bus und Auto in verkehrsabhängigen Situationen bzw. in der Stadt, da diese Transporttypen in diesen Situationen sehr ähnliche Verhalten und Werte aufweisen.

- Fußgänger
- Fahrrad
- Bus
- Auto (stellvertretend für Motorrad, Taxi, PKW etc)
- Zug

Es soll weiters, für jede Person, die ein Gerät besitzt, das in der Lage ist, eine GPS-Spur im GPX-Format aufzuzeichnen, möglich sein, diese Spur analysieren zu lassen. Dies bedeutet, dass keine speziellen Geräte oder andere Sensoren benötigt werden und dass sich dieser Prototyp mit möglichst geringen Anpassungen (Trainingsdaten, GIS-Daten

und Grenzwerte in der Konfiguration) auch auf andere Regionen anwenden lässt. Zum Aufzeichnen der in dieser Arbeit verwendeten GPS-Spuren, wurden mehrere Smartphones mit der App "MyTrack" sowie mehrere GPS-Geräte benutzt.

Im Zuge dieser Arbeit wird auch untersucht, welcher Grad an Genauigkeit sowohl mit als auch ohne geografische Zusatzinformationen im Raum Vorarlberg erreicht werden kann. Dabei soll es nach der Analyse die Möglichkeit geben, die automatisch bestimmten Verkehrsmittel manuell zu korrigieren, sollten diese nicht mit der Realität übereinstimmen. Weiters sollen diese manuellen Änderungen in die Auswertung mit einfließen.

Schlussendlich soll mit Hilfe der Auswertungen auch eine Aussage über die erzielte Genauigkeit mit und ohne den verwendeten Zusatzinformationen gemacht werden. Außerdem soll eine Aussage darüber gemacht werden, ob weitere Zusatzdaten für eine noch genauere Bestimmung benötigt werden und welche Daten dies sein könnten.

Nichtziele

In dieser Arbeit werden die diversen Schlussfolgerungsmodelle (neuronales Netz, Bayessches Netz, Random Forest, Support Vector Machine, ...) nicht betrachtet oder verglichen, sondern es wird auf ein Modell gesetzt, das bereits bei anderen Arbeiten wie z.B. [Stenneth et al., 2011], [Reddy et al., 2010], [Sebastian Nagel, 2011] und [Zheng et al., 2008b] vielversprechende Ergebnisse erreicht hat. Dieses Modell ist der Entscheidungsbaum. Außerdem werden die Daten zur Analyse bzw. Auswertung nicht in Echtzeit betrachtet sondern in Form einer GPS-Spur an den Prototypen übergeben.

Die Frage, an welcher Position man das Gerät zur Aufzeichnung am besten trägt, um möglichst genaue GPS-Daten zu erhalten, wird nicht weiter verfolgt, da die Daten möglichst realistisch sein sollen. Auch bleibt die Frage nach dem Energierverbrauch der App bzw. wie eine möglichst energieschonende App und die dazugehörige Kommunikation aufgebaut sein könnte, unberücksichtigt. Schlussendlich ist eine umfassende Behandlung der Themen Sicherheit und Privatsphäre im Rahmen der vorliegenden Arbeit nicht möglich

und daher bleiben auch diese Themen unberührt. An dieser Stelle soll allerdings erwähnt werden, dass keine benutzerspezifischen Information in den GPS-Spuren benötigt oder vom Prototypen generiert oder gespeichert werden.

1.2 Motivation und Nutzen

Schon früher wurde versucht, Aufzeichnungen über die Verkehrswege von verschiedenen Menschen zu sammeln. Aber die Protokolle in Papierform sowie die Telefonbefragungen waren zu aufwändig und die Menschen nicht zuverlässig genug. Darum ist es von entscheidendem Vorteil, eine App oder ein Gerät zur Verfügung zu haben, welches die Vorgänge des Aufzeichnens möglichst genau für einen übernimmt. [Zheng et al., 2010]

Wird eine Auswertung mit einer für das Zielgebiet aussagekräftigen Anzahl an Personen durchgeführt, so kann das Resultat für verschiedenste Zwecke verwendet werden. Auch ohne spezielle Analyse kann rein durch die Betrachtung der gesammelten GPS-Spuren festgestellt werden, welche Routen besonders häufig benutzt werden.

Zieht man nun verschiedene Werte aus der Auswertung hinzu, kann auch festgestellt werden, wo sich zum Beispiel verkehrstechnische Engstellen befinden und welche Routen sehr populär sind oder es können Aussagen über die allgemeine Verkehrssituation gemacht werden. Durch die gesammelten Daten könnten sich auch Simulationen für anstehende Bauvorhaben machen lassen und auch versucht werden, eine Vorhersage für bestimmte Situationen zu tätigen. Diese Aspekte können unter anderem für das Verkehrsministerium, den öffentlichen Personennahverkehr oder auch für die Stadtplanung sehr interessant sein (Optimierung von Auslastung, Einsparungspotentiale, ...).

Eine ganze Reihe von Apps lässt sich mit den Auswertungen erstellen. Unter anderem könnten diese Apps die Auswertungen in soziale Medien integrieren oder für Fitnessanalysen verwendet werden. Einen einfachen Rückblick über die eigene Fortbewegung kann man damit genauso ermöglichen wie für Umweltbewusste errechnen, wie viel CO2 sie

produziert oder gespart haben. Ein Reisetagebuch könnte daraus genau so Nutzen ziehen wie eine App, die beim Autofahren Auskunft über die aktuell billigste Tankstelle in näherer Umgebung gibt oder eine App, die einfach nur Vorschläge für alternative, schnellere Routen zu einem bekannten Ziel anbietet.

Zusammenfassend kann man sagen, dass die Verwendungsmöglichkeiten für solche Daten umfangreich sind und sich am besten unter den Begriffen kontextorientierte, geographische Apps zusammenfassen lassen. Nicht zuletzt öffnen sich mit solchen Daten aber auch umfangreiche Möglichkeiten für die Werbebranche.

1.3 Weiterer Aufbau der Arbeit

Der Hauptteil der vorliegenden Arbeit gliedert sich in fünf große Abschnitte:

Im ersten Abschnitt wird auf die Akquirierung der GPS-Daten eingegangen. Dies umfasst sowohl die gesammelt GPS-Spuren und deren Struktur, sowie die verwendeten GIS-Daten. Dabei geht es einerseits um deren Herkunft als auch darum, wie diese extrahiert wurden und wie auch diese Daten aufgebaut sind. Außerdem werden auch andere Daten, wie zum Beispiel GPS-Daten von Bussen des ÖPNV in Betracht gezogen.

Der zweite Abschnitt behandelt den entwickelten Prototypen, der die übergebenen GPS-Spuren analysiert. Dabei werden einerseits dessen Funktionalitäten erklärt sowie der grundlegende Ablauf für den Benutzer dargelegt. Weiters wird auch auf die Architektur des Prototyps sowie auf dessen Konfigurationsmöglichkeiten eingegangen.

Der dritte Abschnitt befasst sich sowohl mit dem Einbinden der GIS-Daten als auch dem Aufbereiten der GPS-Daten. Mit Aufbereiten der Daten ist gemeint, dass zusätzliche Werte zu GPS-Punkten für die spätere Analyse berechnet werden. Beispiele für diese Werte sind sowohl die Geschwindigkeit, Beschleunigung und Distanz als auch der Abstand zu der nächsten Bushaltestelle. Weiters wird in diesem Abschnitt auch der Prozess des Aufteilens von GPS-Spuren in Teile (Segmente), in denen nur ein Verkehrsmittel verwendet

wird, beschrieben. Schlussendlich sollen die eingebunden und erweiterten Daten verwendet werden, um eine einfache und sichere Bestimmung des Verkehrsmittels pro Segment zu ermöglichen. Diesem Schritt vorangegangen ist das in Anhang 1 beschriebene Filtern der GPS-Daten. Dies bedeutet in diesem Zusammenhang, dass ein Teil der fehlerhaften Ausreißer aus den GSP-Spuren entfernt wurden. Diese Ausreißer können sowohl die Geschwindigkeit betreffen, als auch unwahrscheinlich große Sprünge im dreidimensionalen Raum sein.

Aufbauend auf den Resultaten aus dem dritten Abschnitt befasst sich der vierte Abschnitt mit der tatsächlichen Erkennung der Verkehrsmittel anhand der berechneten Werte und den einzelnen Abschnitten der GPS-Spur. Die aus dem Entscheidungsbaum gewonnenen Erkenntnisse werden schlussendlich ein letztes Mal überprüft, um sinnfreie bzw. sehr unwahrscheinliche Wechsel zwischen Verkehrsmitteln zu verhindern.

Der fünfte und letzte Abschnitt des Hauptteils befasst sich mit der Auswertung der gewonnenen Erkenntnisse aus den vorhergehenden Abschnitten sowie den Testläufen mit neuen GPS-Spuren mit und ohne zusätzliche GIS-Informationen.

State of the Art

Zu den Meilensteinen auf dem Forschungsgebiet der Verkehrsmittelerkennung (Transport Mode Detection) zählt die Arbeit von Yu Zheng in welcher er unter anderem auf die Erkennung von den Abschnitten mit nur einem Verkehrsmittel eingeht. Weiters verwendete er in seiner Arbeit auch einen mit dem von den GPS-Spuren gesammelten geographischen Wissen aufgebauten Graphen, welcher zur abschließenden Auswertung verwendet wurde. [Zheng et al., 2010] [Zheng et al., 2008a] [Zheng et al., 2008b]

Mit der Frage, wie geographische Daten in eine solche Analyse miteinbezogen werden können hat sich unter anderem Leon Stenneth beschäftigt. Dabei hat er nicht nur fixe Daten wie Gleise und Busstationen, sondern auch aktuelle Buspositionen miteinbezogen. [Stenneth et al., 2011]

Sowohl Stenneth als auch Zheng haben in ihren Arbeiten detailliert erklärt, wieso sie welche Attribute (Geschwindigkeit, Beschleunigung, ...) für die Bestimmung des Verkehrsmittels verwendet haben und sie haben diese auch durch Versuche nach ihrer Wichtigkeit gereiht. Außerdem haben beide und auch Sasank Reddy [Reddy et al., 2008] mehrere Schlussfolgerungsmodelle (Entscheidungsbaum, Bayessches Netz, Markov Modelle, Random Forest, ...) betrachtet und miteinander verglichen.

Wie man mit Verbindungsabbrüchen umgeht und zwischen ähnlichen Verkehrsmitteln unterscheiden kann hat unter anderem auch Filip Biljecki beschäftigt. Des Weiteren baut er für die unterschiedlichen Kategorien von Verkehrsmitteln ein hierarchisches Modell auf, welches ihm helfen soll, bessere Entscheidungen zu treffen. [Biljecki et al., 2013]

2.1 Daten

Ein wesentlicher Unterschied zwischen all den betrachteten Publikationen sind die verwendeten Daten. Nur die GPS-Spuren bilden eine gemeinsame Basis. Manche untersuchten die Verwendung von GSM- und Wifi-Informationen [Reddy et al., 2010], stützten sich auf Zusatzinformationen durch weitere Sensoren wie zum Beispiel einem Beschleunigungssensor [Reddy et al., 2010] [Nadine Schüssler et al., 2011]. Andere, wie Leon Stenneth, verwendeten Live-Informationen von den öffentlichen Verkehrsmitteln und kombinierten diese mit GIS-Informationen, seien es Busstationen, Bahnstrecken, das Straßennetz oder Parkplätze [Stenneth et al., 2011].

2.2 Verkehrsmittel

Ein weiterer Unterschied zwischen den Publikationen sind die betrachteten Verkehrsmittel. Hierbei reicht die Spanne der unterschiedenen Fortbewegungsmöglichkeiten von "gehen, laufen, Fahrrad und motorisiert" [Reddy et al., 2010] bis hin zu "gehen, Zug, U-Bahn, Rad, Auto, Straßenbahn, Bus, Fähre, Segelboot und Flugzeug" [Biljecki et al., 2013].

2.3 Analyse der Publikationen von Yu Zheng

Zu den Meilensteinen auf dem Gebiet der Transport Mode Detection zählen die Publikationen von Yu Zheng und seinem Team: "Understanding Mobility Based on GPS

Data“ [Zheng et al., 2008a], “Understanding Transportation Modes Based on GPS Data for Web Applications“ [Zheng et al., 2010] und “Learning Transportation Mode from Raw GPS Data for Geographic Application on the Web“ [Zheng et al., 2008b]. In diesen Artikeln wird unter anderem auf die Erkennung von den Abschnitten mit nur einem Verkehrsmittel eingegangen. Weiters wird erklärt wie in den Projekten ein aus analysierten GPS-Spuren erstellten Graph mit geographischen Wissen zur Verbesserung der Erkennungsrate beiträgt und mit welchen Schlussfolgerungsmodellen welche Ergebnisse erzielt werden konnten.

2.3.1 Geolife

Die Applikation, welche im Zusammenhang mit den Arbeiten entstanden ist, nennt sich Geolife und ist eine Webapplikation mit den Ansätzen eines soziales Netzwerks. Dabei ging es darum, dass Benutzer GPS-Spuren in Form einer Aufzeichnung auf die Webseite laden konnten, diese Spur vom Algorithmus analysiert und die Verkehrsmittel bestimmt wurden. Dargestellt wurden die Resultate auf einer Karte und die Benutzer konnten diese mit anderen teilen.

Mit Hilfe der so gesammelten Daten konnte unter anderem Datamining betrieben und populäre Strecken festgestellt sowie Verkehrssituationen beurteilt werden. Außerdem konnten auch aktuelle Positionswerte mit einem GPS-fähigen Smartphone/Handy ausgewertet werden und Informationen wie z.B. Abfahrtszeiten der öffentlichen Verkehrsmittel angeboten werden. Wurde aber z.B. eine Strecke mit dem Auto in die Stadt gesucht so konnte auf Grund der bereits analysierten Fahrten über durchschnittliche Geschwindigkeit festgestellt werden, welche Strecke am schnellsten ans Ziel führt.

In diesem von Microsoft Research geführten Projekt standen umfangreiche Test- und Trainingsdaten (1,2 Millionen Kilometer und 48.000 Stunden) zur Verfügung. [Microsoft Research, 2015] Ein Bildschirmfoto der Applikation ist in Abbildung 2.1 zu sehen.

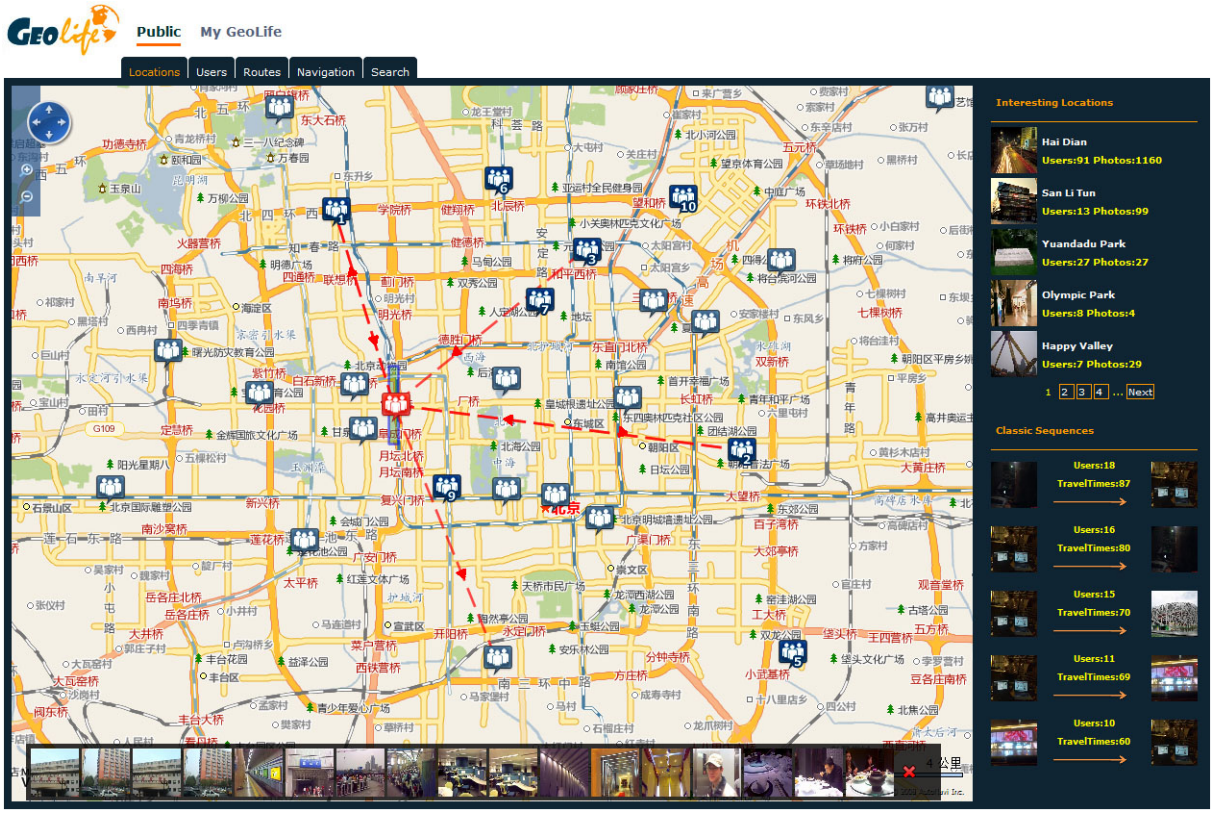

Abbildung 2.1: Geolife (Quelle: research.microsoft.com)

2.3.2 Segmentierung

Die Publikationen von Zheng beinhalten auch eine der detailliersten Beschreibungen des Segmentierungsvorgangs. Dabei werden GPS-Spuren in Abschnitte, in welchen nur ein Verkehrsmittel verwendet wird, unterteilt. Diese Abschnitte können dann in weiterer Folge genauer analysiert und das benutzte Verkehrsmittel bestimmt werden.

Beim Segmentieren stützt sich Zheng darauf, dass Personen bei einem Wechsel des Verkehrsmittels stehenbleiben und sich ein Stück zu Fuß bewegen. Dies bedeutet, dass es einige GPS-Punkte mit einer Geschwindigkeit von genau oder beinahe 0 km/h gibt. Diese aufeinander folgenden Punkte können dann in ein “Gehen“-Segment zusammengefasst werden. Alle anderen Segmente werden vorläufig als “Nicht-Geh“-Segmente klassifiziert. Zheng

sagt außerdem, dass der Beginn und das Ende eines Segments mit dem Typ "Gehen" ein wichtiger Indikator für einen Wechsel des Fortbewegungsmittels ist. Diese Aussage stützt sich auf die Erkenntnisse aus GPS-Daten, die von 65 Personen über 10 Monate gesammelt wurden.

2.3.3 Schlussfolgerungsmodelle

Um die Typen der "Nicht-Geh-Segmente" bestimmten zu können, hat sich Zheng auf verschiedene Zusatzinformationen zu den Segmenten gestützt, darunter:

- Distanz
- Maximale Geschwindigkeit
- Maximale Beschleunigung
- Durchschnittliche Beschleunigung
- Richtungswechsel
- Stopprate
- Erwartete Geschwindigkeit
- Geschwindigkeitsänderungsrate

In weiterer Folge stellte er fest, dass die Stopprate, Richtungswechselrate und die Geschwindigkeitsänderungsrate am effektivsten und stabilsten gegenüber den verschiedenen Verkehrssituationen sind. Diese 3 Eigenschaften können auch mit ein paar der anderen kombiniert werden um noch weitere Verbesserungen zu erhalten. Werden allerdings zu viele Eigenschaften miteinbezogen, so konnte er eine Verringerung der Genauigkeit beim Bestimmen des Typs feststellen.

Für die tatsächliche Bestimmung des Typs führte Zheng verschiedene Experimente mit dem Entscheidungsbaum, der Support Vector Machine, dem Bayesschen Netz und dem Conditional Random Field durch. Dabei stellte er fest, dass der Entscheidungsbaum die

besten Ergebnisse im Zusammenspiel mit der in Abschnitt 2.3.3 beschriebenen Segmentierungsmethode liefert.

Während dem Analysieren der Trainingsdaten wurde im Hintergrund ein Graph aufgebaut, welcher Informationen über die jeweils betrachtete geografische Region widerspiegelt. In diesen geografischen Regionen waren 28 große Städte in China, mehrere Städte in den USA sowie Südkorea und Japan enthalten. Mit diesem Graph wurden die Schlussfolgerungen nochmals überprüft und es konnte eine weitere Verbesserung erzielt werden. Schlussendlich konnte der von Zheng entwickelte Algorithmus 76,2% der Testfälle ohne jegliche Zusatzinformationen im Sinne von GIS-Daten oder weitere Sensordaten korrekt identifizieren.

2.4 Analyse der Publikationen von Leon Stenneth

Leon Stenneth vergleicht in der Publikation "Transportation Mode Detection usign Mobile Phones and GIS Information" [Stenneth et al., 2011], ähnlich wie Zheng auch, verschiede Schlussfolgerungsmodelle. Allerdings verwendete er auch zusätzliche GIS-Informationen, um eine genauere Bestimmung zu ermöglichen. Außerdem arbeitet die entwickelte Applikation nicht mit ganzen GPS-Spuren, sondern analysiert immer 2 Punkte innerhalb eines 30 Sekunden Intervalls, wodurch auch das eigentliche Segmentieren entfällt.

2.4.1 GIS-Informtionen

Die von Stenneth verwendeten GIS-Informationen beinhalten sowohl die Gleise von Zügen als auch Bushaltestellen sowie die aktuelle Position von allen Bussen in Chicago. Daraus berechnete er verschiedene Zusatzinformationen für die Bestimmung des Transporttyps:

- Durchschnittliche Nähe zu den Gleisen
- Durchschnittliche Nähe zu Bushaltestellen
- Durchschnittliche Nähe zu dem nächsten Bus

2.4.2 Schlussfolgerungsmodelle

Um den Typ des jeweils betrachteten Abschnitts zu bestimmen, zieht auch Stenneth mehrere Zusatzwerte als Kriterien zu Hilfe:

- Durchschnittliche Geschwindigkeit
- Durchschnittliche Nähe zu den Gleisen
- Durchschnittlicher Abstand zu einem Bus
- Durchschnittliche Beschleunigung
- Durchschnittlicher Richtungswechsel
- Durchschnittliche Bushaltestellennähe
- Durchschnittliche Genauigkeit der Koordinaten
- Distanz zum nächsten Bus

Wie auch Zheng stellt auch Stenneth fest, dass nur ein paar der Zusatzwerte wirklich effektiv sind. In seinem Fall waren das die durchschnittliche Geschwindigkeit, Beschleunigung und die Nähe zu den Gleisen sowie der Abstand zu anderen Bussen sowie die Distanz zum nächsten Bus.

Die von Stenneth betrachteten Modelle sind Naive Bayes, Bayessches Netz, Entscheidungsbaum, Random Forest und Multilayer Perceptron. In den Experimenten stellte er fest, dass der Random Forest mit den GIS-Informationen das vielversprechendste Modell mit mehr als 93% richtig erkannten Verkehrsmitteln ist. Ohne GIS-Informationen schneidet auch der Random Forest ähnlich wie bei Zheng der Entscheidungsbaum mit 76% ab. Mit GIS-Informationen erreicht auch der Entscheidungsbaum eine Erkennungsrate von mehr als 92%.

2.5 Analyse der Publikationen von Filip Biljecki

Filip Biljecki veröffentlichte eine Arbeit mit dem Titel "Transportation mode-based segmentation and classification of movment trajectories" [Biljecki et al., 2013]. Darin vergleicht er nicht nur eine Vielzahl von Publikationen zu dem Thema Verkehrsmittelerkennung (unterer anderem [Schuessler and Axhausen, 2009], [Zheng et al., 2010], [Reddy et al., 2010], [Gonzalez et al., 2010]) anhand der Fortbewegungsmittel, den Zusatzwerten, ob GIS-Daten verwendet wurden und welches Resultat erzielt worden ist, sondern führt auch ein hierarchisches Modell für die Erkennung der Transportmittel ein. Weiters verwendet er zur Bestimmung des Transportmittels auch GIS-Daten wie Bus- und Straßenbahnhaltestellen. Ein Kapitel ist in der Arbeit auch der Behandlung von Störungen oder Unterbrechungen des GPS-Signals gewidmet. Darin wird aufgeführt, welche Fälle weiterhin behandelt werden können und welche sich mit diesem Ansatz nicht beheben lassen. In dieser Arbeit konnte eine Genauigkeit von 95% erreicht werden.

2.5.1 Segmentierung

Für die Segmentierung verwendet Biljecki dieselbe Methodik wie Zheng aber erweitert diese dahingehend, dass auch dann segmentiert wird, wenn ein Signalverlust (30 Sekunden keine neuen Werte) feststellt werden kann. Diese Entscheidung wird damit begründet, dass es bei einem Verkehrsmittelwechsel oft zu einem Signalverlust kommt. Überschreitet ein Stopp eine bestimmte Dauer (12 Sekunden) so wird auch segmentiert, denn dies könnte auch auf einen Wechsel hindeuten. Da die Track-Punkte nicht 100% genau sind, wird alles unter 2km/h als Stopp eingestuft. Weil nach der Erkennung alle aufeinander folgenden Segmente mit dem selben Typ zusammengefasst werden, ist eine Übersegmentierung kein Problem.

2.5.2 Schlussfolgerung

Um das Fortbewegungsmittel feststellen zu können wird ein Expertensystem verwendet. Dieses System basiert auf einer Fuzzy-Logic und klassifiziert die Segmente mit Wahrscheinlichkeitswerten. Dazu verwendet dieses System auch die hierarchische Gliederung der verschiedenen Verkehrsmittel welche in Abbildung 2.2 ersichtlich ist. Diese Gliederung soll verhindern, dass sich das System zu früh auf einen Typ festlegen muss - das System kann den Typ dadurch Schrittweise bestimmen.

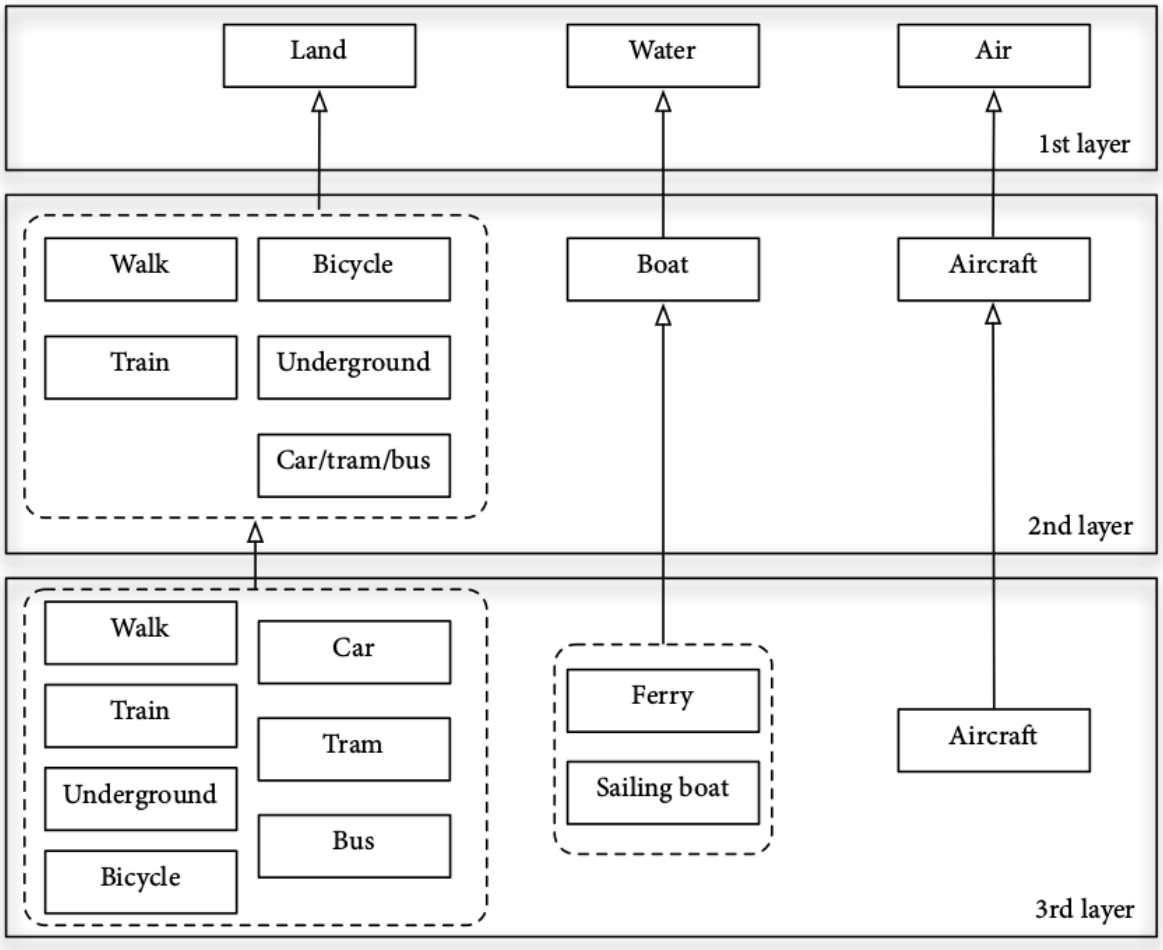

Abbildung 2.2: Fortbewegungsmittel-Hierarchie

Zusammenfassend sagt Biljecki, dass es sehr schwierig ist, alle möglichen Fälle der Realität abzubilden und wirklich zufriedenstellende Resultate zu erhalten. Dies betrifft vor allem den Segmentierungs- und Bestimmungsprozess von Daten mit sehr viel fehlerhaften Ausreißern (Rauschen) oder jene Fällen in denen wenig Beispieldaten vorhanden sind. Weiters meint er, dass Fehler nicht zwangsläufig auf das System zurückzuführen sind, sondern dass, es sich bei diesen Fehlern oft um spezielle Situationen handelt, welche sehr kompliziert zu modellieren sind oder drastische Auswirkungen auf die allgemeine Performanz haben.

Alles in allem kann aber gesagt werden, dass das von Biljecki vorgestellte Modell zehn verschiedene Verkehrsmittel unterscheiden kann. Dies sind mehr als in allen anderen vom Autoren betrachteten Publikationen. Außerdem konnte durch die Verwendung von GIS-Daten bessere Resultate (95% bis 100% je nach Daten) als in vielen anderen Publikationen im Bereich der Verkehrsmittelerkennung erzielt werden.

2.6 Analyse der Publikationen von Sasank Reddy

Sasank Reddy hat sich in den Publikationen "Using Mobile Phones to Determine Transportation Modes" [Reddy et al., 2010] und "Determining Transportation Mode on Mobile Phones" [Reddy et al., 2008] wie auch Stenneth und Zheng mit verschiedenen Schlussfolgerungsmodellen befasst. Bevor diese Modelle aber zum Einsatz kamen, wurde evaluiert, mit welchen weiteren Sensoren man sinnvolle Daten aufzeichnen könnte. Im Zuge dessen wurde auch überprüft, an welche Stelle am Körper das Smartphone die genauesten Daten liefert und wie möglichst Energieeffizient Daten aufgezeichnet und übermittelt werden können. Im Gegensatz zu anderen Publikationen wurde in diesen nicht genauer zwischen den motorisierten Verkehrsmitteln unterschieden.

2.6.1 Sensoren

Dadurch, dass die meisten Smartphones nicht nur über ein GPS-Modul sondern auch über WIFI, einen Beschleunigungssensor sowie Bluetooth und natürlich über ein GSM-Modul verfügen, wurde in diesen Arbeiten auch evaluiert, ob es möglich und sinnvoll ist Daten von diesen Geräten und Sensoren miteinzubeziehen.

Bluetooth wäre zwar interessant, aber es kommt hauptsächlich innerhalb von Gebäuden zu Einsatz (TV, Radio, Computer, etc). Darum konnte dieser Sensor nicht für die Bestimmung des Verkehrsmittels eingesetzt werden.

Wifi und GSM in den Erkennungsprozess miteinzubeziehen konnte nach einer Reihe von Versuchen ausgeschlossen werden, da der Grad der Verbesserung nur 0,6% betrug und ein wesentlich höherer Energiebedarf gegeben war. Außerdem war die Abhängigkeit von Wifi und GSM in ländlichen Gegenden mit schlechtem Empfang und / oder wenig Wifis ein weiterer Grund gegen diese Sensoren.

Die vielversprechendste Kombination war jene aus GPS-Daten und den Daten des Beschleunigungssensors, welche auch für die weiteren Experimente verwendet wurde.

2.6.2 Schlussfolgerungsmodelle

Um das Fortbewegungsmittel des jeweiligen Abschnittes zu erkennen, zieht Reddy folgende Zusatzinformationen heran:

- Geschwindigkeit
- Varianz des Beschleunigungssensorsignals
- 3 Beschleunigungssensorwerte (3Hz, 2Hz, 1Hz)

Die in dieser Publikation betrachteten Modelle sind der Entscheidungsbaum, K-Means Clustering, Naive Bayes, Nearest Neighbor, Support Vector Machine sowie ein Continuous Hidden Markov Model und ein Eintscheidungsbaum in Kombination mit einem

Discrete Hidden Markov Model. Dabei konnte festgestellt werden, dass die letzte Variante die besten Resultate mit 93,6% erzielt. Wie aber auch in den anderen Publikationen war der Entscheidungsbaum mit 91,3% nicht weit abgeschlagen hinter dem kombinierten Ansatz.

2.7 Zusammenfassung

Neben den vorgestellten Publikationen zum Thema Verkehrsmittelerkennung gibt es noch weitere interessante Publikationen deren Erfahrungen zum Teil in diese Arbeit eingeflossen sind. Unter diesen Publikationen sind unter anderem "Processing raw data from global positioning systems without additional information" [Schuessler and Axhausen, 2009] und "Improving post-processing routines for gps oversavations using propted-recall data" [Nadine Schüssler et al., 2011] von Nadine Schüssler sowie "Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks" [Gonzalez et al., 2010] von Paola Gonzalez.

Für diese Arbeit wurde für den Segementierungsprozess die Vorgehensweise von Yu Zheng verwendet. Darin sind aber auch die von Filip Biljecki vorgeschlagenen Änderungen eingeflossen, welche eine weitere Segmentierungsregel bei Empfangsverlust beinhaltet.

Für die weitere Verarbeitung der GPS-Daten wurde der Entscheidungsbaum aufgrund des guten Abschneidens in den erwähnten Publikationen verwendet. Der Entscheidungsbaum wurde zur nachfolgenden Analyse und Auswertung der verbesserten Verkehrsmittelerkennung in zwei Varianten integriert. In der einen Variante verwendet er nur aus den GPS-Daten berechnete Werte (Geschwindigkeit, Beschleunigung, ...) wie es auch in vielen anderen Arbeiten gemacht wurde. In der zweiten Variante greift der Entscheidungsbaum aber auch auf GIS-Informationen zurück wie es zum Beispiel Leon Stenneth in seiner Publikation [Stenneth et al., 2011] beschrieben hat.

Ein weiterer Grund neben dem guten Abschneiden bei der GIS-Daten-gestützten Verkehrsmittelerkennung war die eigen Zielsetzung, dass keine zusätzlichen Sensoren oder Geräte neben dem Gerät für die Aufzeichnung der GPS-Daten verwendet werden sollen.

Modellbildung und Einbindung der GPS- und GIS-Daten

Dieser Abschnitt behandelt die Akquirierung und die Struktur der verwendeten GPS-Daten für das Training des Entscheidungsbaums. Außerdem wird erläutert, wieso der Entscheidungsbaum als Schlussfolgerungsmodell ausgewählt und wie er erstellt wurde. Aufgrund der unterschiedlichen Attribute der beiden Fälle (mit und ohne GIS-Daten), sehen die zwei Entscheidungsbäume sehr unterschiedlich aus und werden daher nur oberflächlich miteinander verglichen.

Sowohl für die Trainingsdaten für den Entscheidungsbaum als auch für die Testdaten wird erläutert, wie diese aufgezeichnet wurden. Dies inkludiert sowohl die Geräte als auch die Software, welche dazu verwendet worden ist. Weiters wird auf die Struktur der GPS-Spuren und der GIS-Daten eingegangen.

Außerdem wird dargelegt, woher die verwendeten GIS-Daten stammen, wie diese extrahiert wurden und welche Rolle sie im weiteren Prozess spielen. Abschließend wird auch auf die Positionsdaten der verschiedenen Verkehrsmittel des ÖPNV eingegangen und in welcher Weise diese hätten eingesetzt werden können.

Typ	Anzahl	Distanz (km)
Auto	59	752,26
Fußgänger	58	129,60
Fahrrad	43	866,48
Zug	11	377,54
Bus	9	46,16
Gesamt	**180**	**2.172,04**

Tabelle 3.1: Trainingsdatenübersicht

3.1 Trainingsdaten

Für das Training der Entscheidungsbäume konnten GPS-Aufzeichnungen aus einem Projekt von Sebastian Nagel "Möglichkeitsstudie zum Projekt: Mobilitäts-Tracker"verwendet werden. Diese Daten wurden mit verschiedenen GPS-Geräten (Wintec WTB-202, Columbus V-900, photoMate 887 Lite, qStarz BT-Q1300, xaiox) und der Hilfe von mehreren Personen aufgezeichnet und beinhalten alle in dieser Arbeit betrachteten Verkehrsmittel. [Sebastian Nagel, 2011]

Weiters wurden zum Training auch neue Datensätze verwendet, die mit zwei verschiedenen Smartphones und mit Hilfe der App “MyTrack“ aufgezeichnet wurden. Diese App wurde ausgewählt, da sie sich sehr einfach handhaben lässt und die einzelnen Aufzeichnungen komfortabel exportiert werden können.

Einen groben Überblick über die gesammelten Daten bietet die Tabelle 3.1. Die erste Spalte enthält den Verkehrsmitteltyp, die Zweite die Anzahl der Segmente und die dritte Spalte enthält die Gesamtdistanz in Kilometern von dem jeweiligen Typ. Insgesamt beinhalten die Trainingsdaten 180 Segmente der verschiedenen Transportmittel und erstrecken sich über 2000 Kilometer.

3.1.1 Struktur der GPS-Daten

Diese Trainingsdaten sind in den einzelnen Dateien als XML abgelegt und entsprechen dem gängigen GPX-Format, wie es im Listing 3.1 ersichtlich ist. Die Spezifikation für GPX kann auf der Webseite von Topografix unter `http://www.topografix.com/GPX/1/1/` gefunden werden. Ein zugehöriges XML-Schema findet man hier `http://www.topografix.com/GPX/1/1/gpx.xsd`. [Topografix, 2004]

Track (trk) Üblicherweise beginnt eine solche Datei mit einem gpx-Element, welches wiederum einen Track enthält. Ein Track repräsentiert eine Aufzeichnung oder Spur und enthält eine Folge aller aufgezeichneten Trackpoints. Diese sind aber wiederum in ein oder mehrere Tracksegmente gegliedert.

Tracksegment (trkseg) Ein Track kann aus einem oder mehreren Tracksegmenten bestehen. Die Tracksegmente enthalten wiederum beliebig vielen aufeinander folgenden Trackpoints. Mit diesen Segmenten kann ein Track in logische Abschnitte unterteilt werden. Außerdem kann ein neues Segment kann begonnen werden, wenn zum Beispiel die Verbindung verloren oder der GPS-Empfänger aus- und wieder eingeschaltet wurde.

Trackpoint (trkpt) Ein Trackpoint entspricht einem Punkt des aufgezeichneten Tracks und enthält Koordinate (Längen- und Breitengrad) sowie einen Zeitstempel und die Höhenmeter. Es gibt aber auch Fälle, in welchen bei einem Trackpoint auch die Geschwindigkeits- oder Beschleunigungsdaten abgelegt wurden.

3.1.2 Entscheidungsbaum

Aufgrund des guten Abschneidens des Entscheidungsbaums in verschiedenen Arbeiten [Stenneth et al., 2011], [Reddy et al., 2010], [Sebastian Nagel, 2011]und [Zheng et al., 2008b] wurde auch in dieser Arbeit ein Entscheidungsbaum als Schlussfolgerungsmodell

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<gpx xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http:
    //www.topografix.com/GPX/1/1" ...>
    <metadata>
        <name>Badgasse - FH</name>
        <desc></desc>
    </metadata>
    <trk>
        <name>Badgasse - FH</name>
        <trkseg>
            <trkpt lat="47.39786" lon="9.735109">
                <ele>475.0</ele>
                <time>2015-02-19T07:20:18.156Z</time>
            </trkpt>
            ...
            <trkpt lat="47.405439" lon="9.744841">
                <ele>492.0</ele>
                <time>2015-02-19T07:24:35.160Z</time>
            </trkpt>
        </trkseg>
    </trk>
</gpx>
```

Listing 3.1: GPX-Datei

verwendet. Um einen Entscheidungsbaum erstellen zu können werden Trainingsdaten benötigt. Daraus werden dann die Regeln für den Entscheidungsbaum bzw. die jeweiligen Entscheidungen des Baums abgeleitet.

Deshalb wurde für diese Trainingsdaten ein Teil der gesammelten GPS-Daten manuell segmentiert und mit dem benutzten Verkehrsmittel ergänzt. Mit Hilfe des Prototyps wurden die Trainigsdaten eingelesen, gefiltert und mit zusätzlichen Informationen bereichert. Diese zusätzlichen Informationen sind bei der Variante ohne GIS-Daten zum Beispiel Geschwindigkeit und Beschleunigung. Bei der Variante mit GIS-Daten wird unter anderem die Nähe zu Bushaltestellen oder zu Schienen ergänzt. Danach wurde für alle eingelesenen Segmente die berechneten Werte und das dazugehörige Verkehrsmittel in einer Datei abgelegt.

Für die konkrete Generierung der Entscheidungsbäume wurde schließlich das Tool Rapidminer verwendet. Dabei handelt es sich um eine Open-Source Datamining Software welche sowohl verschiedenste Datenquellen unterstützt (Datenbanken und auch einzelne Dateien) als auch die Generierung von Entscheidungsbäumen über eine komfortable Benutzeroberfläche erlaubt. Die vom Prototypen genierte CSV-Datei konnte dadurch einfach eingebunden von RapidMiner eingelesen und ausgewertet und schlussendlich die Entscheindungsbäume generiert werden. Diese können dann sowohl als Bilder als auch in Textform exportiert werden.

Überanpassung (Overfitting)

Beim Erstellen von Entscheidungsbäumen wie auch bei anderen von Trainingsdaten lernenden Algorithmen muss beachtet werden, dass das Ergebnis nicht zu sehr auf die Trainingsdaten zugeschnitten ist. Dies bedeutet man möchte, dass mit Hilfe der Trainingsdaten ein Modell generiert wird, welches auch für Nichttrainingsdaten ein akzeptables Resultat liefert und nicht zu sehr auf die Gegebenheiten / Ungenauigkeiten in den Trainingsdaten

spezialisiert ist. Ist dies jedoch der Fall so spricht man von einer Überanpassung (Overfitting) des Modells auf die Daten. [Tom Dietterich, 1995]

Zurückschneiden (Pruning)

Ist ein Entscheidungsbaum zu sehr angepasst auf die Trainingsdaten so muss dieser wieder zurückgeschnitten werden. Dies bedeutet, dass einzelne Blätter oder auch Teilbäume wieder entfernt werden um ein bestmögliches Resultat (geringe Anzahl an Fehlern) für Nichttrainingsdaten zu erhalten. Für diese Aufgabe wurden verschiedene Algorithmen Entwickelt, welche unter anderem in der Publikation "Pruning Decision Trees with Misclassification Costs" von Jeffrey Bradford beschrieben werden. [Jeffrey P. Bradford et al., 1998]

Entscheidungsbaum ohne GIS-Daten

Der Entscheidungsbaum ohne Berücksichtigung von GIS-Daten ist in Abbildung 3.1 zu sehen. Bei diesem Entscheidungsbaum wurden mittlere und maximale Geschwindigkeit, Distanz sowie mittlere und maximale Beschleunigung als Indikatoren für den Entscheidungsprozess gewählt. Wie diese Werte berechnet und ergänzt werden wird im Abschnitt über den Prototypen selbst, erklärt.

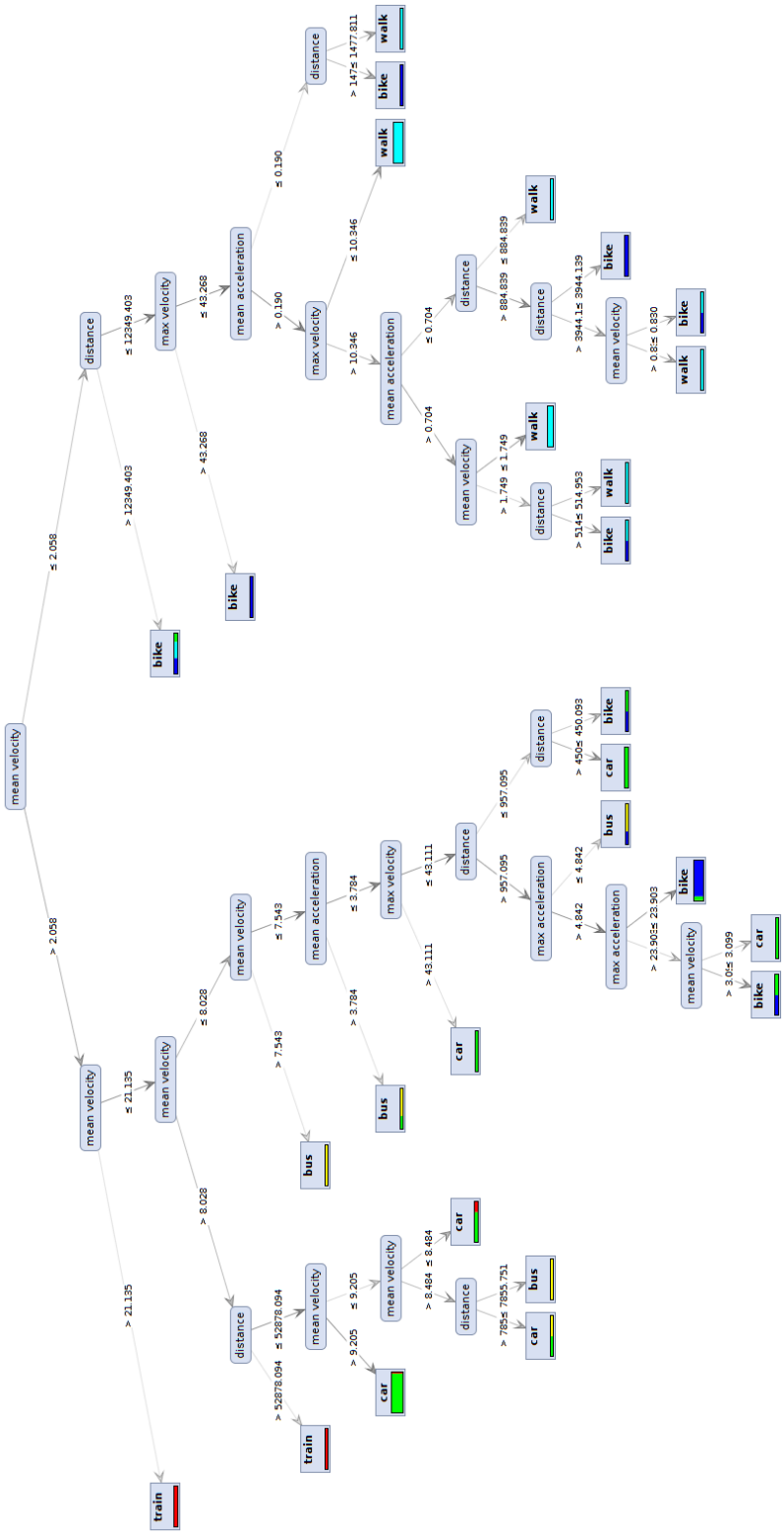

Abbildung 3.1: Entscheidungsbaum ohne GIS-Daten

Entscheidungsbaum mit GIS-Daten

idungsbaum
daten

3.2 Neue Aufzeichnungen

Wie bereits im Abschnitt 3.1 erwähnt, wurden die neuen Daten mit zwei Smartphones (Samsung Galaxy S und einem LG Nexus 5) und der App "MyTrack" (sieh Abbildung 3.2) aufgezeichnet. Neben der einfachen Handhabung bietet diese App auch an nur Punkte mit einer Mindestgenauigkeit aufzuzeichnen was in Folge beim Filtern ein wenig Arbeit abnimmt. Diese neuen Daten werden hauptsächlich zum Testen verwendet werden und nur ein Teil davon ist in die Trainingsdaten eingeflossen. Weiters kann man zwar ein Fortbewegungsmittel pro Aufzeichnung angeben, aber dies hat keinerlei Einfluss auf die Aufzeichnung selbst oder die Daten - es dient lediglich der visuellen Darstellung/Unterscheidung der einzelnen Aufzeichnungen.

3.3 GIS-Daten

Als relevante GIS-Daten kommt laut den verschiedensten Publikationen zum Thema Verkehrsmittelerkennung einiges in Frage wie z.B. Parkplätze, Busstationen, Gleise, Bahnhöfe und das gesamte Straßennetz. All diese Daten mögen zwar relevant sein, aber es handelt sich auch um sehr viele Daten was wiederum bedeutet, dass die Bearbeitungszeit einer Aufzeichnung rasch ansteigt. Da der Prototyp auch für konkrete Endbenutzer interessant sein soll, wird auf die Verwendung des Straßennetzes und der Parkplätze verzichtet, um die Bearbeitungszeit möglichst gering zu halten. Deshalb wird auf die Verwendung von Busstationen und der Gleise gesetzt da dies schon in der Arbeit von Stenneth [Stenneth et al., 2011] zu guten Resultaten geführt hat. Dieser Ansatz wird ergänzt mit den GIS-

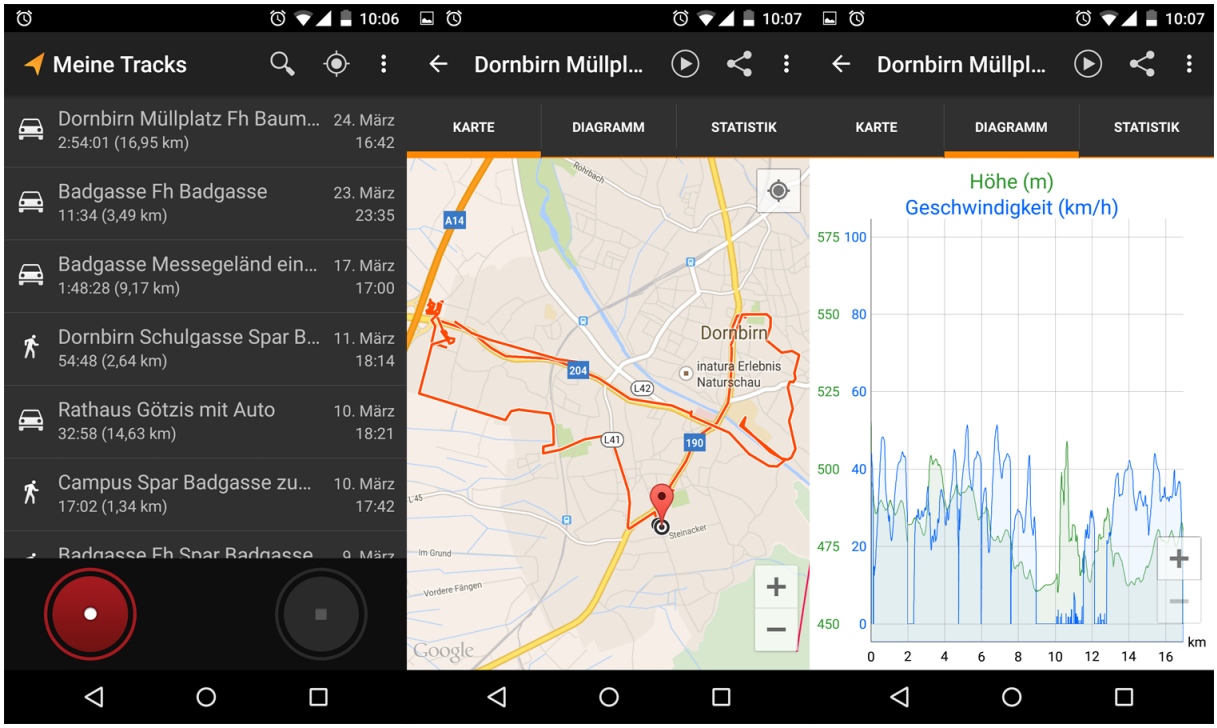

Abbildung 3.2: Die App myTrack

Daten des Autobahnnetzes da in diesen viel Potential vermutet wird und es sich um eine überschaubare Menge an Daten handelt.

Akquirierung

Allgemein sind GIS-Daten via OpenStreetMaps oder Google-Maps verfügbar, aber in einzelnen Stichproben hat sich herausgestellt, dass die Zusatzinformation in OpenStreetMaps wesentlich detaillierter und einfacher zum Extrahieren sind. Dafür wurde in Kauf genommen, dass diese Daten nicht standardisiert eingetragen wurden.

Die Österreich-Daten von OpenStreetMaps wurden als Archiv heruntergeladen und mit Hilfen von JOSM auf den relevanten Bereich eingegrenzt. JOSM ist ein Tool mit welchem die Daten von OpenStreetMaps gepflegt werden können. Nachdem der Bereich auf Vorarlberg eingegrenzt worden ist, konnte dieser mit Hilfe von osmosis (weiteres Tool von der OpenStreetMaps-Community) auf bestimmte Punkte und Verbindungen gefiltert werden.

Dadurch war es möglich, das Schienennetz von Vorarlberg sowie die Busstationen von Vorarlberg zu exportieren.

ung der
en ins mo-

3.4 Weitere Daten

Neben den zusätzlichen Werten, die aus den GPS-Spuren berechnet werden können und den GIS-Daten, wurde auch überlegt, Daten des öffentlichen Personennahverkehrs einzubinden, da diese in Vorarlberg über eine GPS-Position eines jeden Busses verfügen würden. Die Verwendung dieser Daten wäre insofern vielversprechend gewesen, als dass man einen ähnlichen Ansatz wie Stenneth verfolgen hätte können. Man hätte dadurch überprüfen können ob an der jeweiligen Stelle gerade ein Bus steht und darüber Rückschlüsse treffen können. Da diese Daten aber zum Zeitpunkt dieser Arbeit weder für diese Arbeit noch für die Öffentlichkeit verfügbar sind, scheidet diese Möglichkeit aus.

Der Prototyp

auf vor-
filtern in

4.1 Funktionalitäten

4.2 Aufbau und Architektur

Processing?

5.1 Aufbereitung ohne GIS-Daten

5.1.1 Auswahl

Durchschnittliche Geschwindigkeit

Maximale Geschwindigkeit

Durchschnittliche Beschleunigung

Maximale Beschleunigung

Höhenmeter

Distanz

5.2 Aufbereitung mit GIS-Daten

5.2.1 Auswahl

Durchschnittliche Geschwindigkeit

Maximale Geschwindigkeit

Durchschnittliche Beschleunigung

Maximale Beschleunigung

5.2.2 Abstand zu Bushaltestellen

5.2.3 Abstand zu Gleisen

5.3 Segmentierung

5.3.1 Terminologie

Analyse

6.1 Entscheidungsbaum

6.2 Nachbearbeitung

Auswertung

7.1 Genauigkeit ohne GIS-Daten

7.2 Genauigkeit mit GIS-Daten

Ausblick

Literaturverzeichnis

[Biljecki et al., 2013] Biljecki, F., Ledoux, H., and van Oosterom, P. (2013). Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science*, 27(2).

[Caron et al., 2006] Caron, F., Duflos, E., Pomorski, D., and Vanheeghe, P. (2006). GPS/IMU data fusion using multisensor Kalman filtering: introduction of contextual aspects. *Information Fusion*, 7(2):221–230.

[Gonzalez et al., 2010] Gonzalez, P. A., Weinstein, J. S., Barbeau, S. J., Labrador, M. A., Winters, P. L., Georggi, N. L., and Perez, R. (2010). Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. *IET intelligent transport systems*, 4(1):37–49.

[Jeffrey P. Bradford et al., 1998] Jeffrey P. Bradford, Clayton Kunz, Ron Kohavi, Cliff Brunk, and Carla E. Brodley (1998). Pruning Decision Trees with Misclassification Costs. Technical Report 51, Purdue University.

[Jun et al., 2006] Jun, J., Guensler, R., and Ogle, J. H. (2006). Smoothing methods to minimize impact of Global Positioning System random error on travel distance, speed, and acceleration profile estimates. *Transportation Research Record: Journal of the Transportation Research Board*, 1972(1):141–150.

[Microsoft Research, 2015] Microsoft Research (2015). GeoLife GPS Trajectories - Microsoft Research. `http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/`.

[Nadine Schüssler et al., 2011] Nadine Schüssler, Lara Montini, and Christoph Dobler (2011). Improving post-processing routines for gps oversavations using propted-recall data. In *9th International conference on survey methods in transport*.

[Reddy et al., 2008] Reddy, S., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. (2008). Determining transportation mode on mobile phones. In *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on*, pages 25–28. IEEE.

[Reddy et al., 2010] Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2).

[Schuessler and Axhausen, 2009] Schuessler, N. and Axhausen, K. W. (2009). Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board*, 2105(1):28–36.

[Sebastian Nagel, 2011] Sebastian Nagel (2011). Möglichkeitsstudie zum Projekt: Mobilitäts-Tracker.

[Stenneth et al., 2011] Stenneth, L., Wolfson, O., Yu, P. S., and Xu, B. (2011). Transportation mode detection using mobile phones and GIS information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 54–63. ACM.

[Tom Dietterich, 1995] Tom Dietterich (1995). Overfitting and Undercomputing in Machine Learning. 27(3):326–327.

[Topografix, 2004] Topografix (2004). GPX 1.1 Schema Documentation. `http://www.topografix.com/GPX/1/1/`.

[Youtube, 2015] Youtube (2015). Youtube statistics. `http://www.youtube.com/yt/press/statistics.html`.

[Zheng et al., 2010] Zheng, Y., Chen, Y., Li, Q., Xie, X., and Ma, W.-Y. (2010). Understanding transportation modes based on GPS data for web applications. *ACM Transactions on the Web (TWEB)*, 4(1).

[Zheng et al., 2008a] Zheng, Y., Li, Q., Chen, Y., Xie, X., and Ma, W.-Y. (2008a). Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321. ACM.

[Zheng et al., 2008b] Zheng, Y., Liu, L., Wang, L., and Xie, X. (2008b). Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web*, pages 247–256. ACM.

Listings

3.1 GPX-Datei 24

Tabellenverzeichnis

3.1 Trainingsdatenübersicht 22

Abbildungsverzeichnis

2.1 Geolife (Quelle: research.microsoft.com) 10
2.2 Fortbewegungsmittel-Hierarchie 15

3.1 Entscheidungsbaum ohne GIS-Daten 27
3.2 Die App myTrack 29

8.1 Filtern - 1. Fall 56
8.2 Filtern - 2. Fall 56
8.3 Filtern - 3. Fall 57

Anhang 1

Wie auch bei vielen anderen Publikation die sich mit GPS-Daten beschäftigen, konnte auch bei dieser Arbeit festgestellt werden, dass sich in den GPS-Spuren einige Ausreißer befanden. Dies konnte vor allem dann beobachtet werden, wenn man sich in einem Zug befand, durch einen Tunnel fuhr oder auch wenn man sich auf einem überdachten Bahnsteig befand. Die Ausreißer werden durch unrealistisch große Distanzabstände oder Sprünge in den Höhenwerten bemerkt. Da diese Werte das Ergebnis verfälschen würden, mussten verschiedene Filter implementiert werden.

In verschiedensten Publikation zum Thema GPS wird beim Filtern von fehlerhaften Außreisern auf den Kalman-Filter gesetzt (z.B. [Caron et al., 2006] und [Jun et al., 2006]). Da das Filtern an sich aber nicht im Fokus dieser Arbeit stand und es zur Zeit dieser Arbeit keine Implementation für die gewählte Programmiersprache PHP gab wurde auf einfachere Filtermöglichkeiten gesetzt.

Konkret wurden Filter für Zeit, Distanz und Höhenmeter implementiert. Die somit bereinigten Resultate hatten bereits auf den Entscheidungsbaum ohne GIS-Daten große Auswirkungen. Die Grenzwerte für diese Filter können in einer Konfigurationsdatei festgelegt und je nach geografischer Region und Testdaten angepasst werden. Diese Filter können auch als konfigurierbares Regelwerk verstanden und wie folgt beeinflusst werden:

- Minimale Zeitspanne zwischen 2 Punkten
- Maximale Distanz pro Sekunde

- Minimale Distanz pro Sekunde
- Maximale Änderung der Höhenmeter zwischen 2 Punkten pro Sekunde
- Minimaler Wert für Verhältnis zwischen gültigen und aufgezeichneten Punkten
- Anzahl der zu überspringenden Punkte am Start
- Minimale Anzahl von Punkten pro Segment

GPS-Punkte, welche nicht den definierten Grenzwerten entsprechen werden im der weiteren Verarbeitung nicht betrachtet. Abgesehen vom Zeitfilter betrachten alle anderen Filter die gemessenen Werte in Relation zur gemessenen Zeit, was bedeutet, dass der Zeitwert größer als 0 sein muss und von mehren GPS-Punkten mit der selben Zeit nur einer betrachtet wird. In den nächsten Absätzen findet sich eine genauere Beschreibung der einzelnen Parameter und ihrer Auswirkungen.

Minimale Zeitspanne zwischen 2 Punkten Mit Hilfe dieses Konfigurationsparameters kann zum einen sichergestellt werden, dass sich nicht mehrere Trackpunkte mit dem selben Zeitstempel eingeschlichen haben und zum anderen kann man damit auch die Genauigkeit bzw. die Anzahl der zu verarbeitenden Punkte steuern.

Maximale/minimale Distanz pro Sekunde Durch diesen Parameter kann sichergestellt werden, dass man sich innerhalb einer gewissen Zeitspanne nur eine gewisse Strecke zurücklegen kann beziehungsweise auch Punkte filtern in denen man sich nicht wirklich bewegt hat. Dadurch können die richtig großen Sprünge gefiltert werden und Punkte ohne Bewegung herausgefiltert werden..

Maximale Änderung der Höhenmeter zwischen 2 Punkten pro Sekunde Hierbei wird überprüft ob es große Sprünge im Bereich der Höhenmeter gibt.

Minimaler Wert für Verhältnis zwischen gültigen und aufgezeichneten Punkten Durch dieses Verhältnis kann entschieden werden ob sich überhaupt noch genug gültige GPS-Informationen für den weiteren Bestimmungsprozess in einem Track.

Anzahl der zu überspringenden Punkte am Start Durch diesen Konfigurationsparameter kann gesteuert werden wie viele Punkte am Beginn einer Aufzeichnung übersprungen werden. Dies rührt daher, dass sich beim Start einer Aufzeichnung gerne Ausreiser einschleichen bis sich der Positionsbestimmungsvorgang eingependelt hat. Durch diesen Parameter können einige dieser Trackpoints übersprungen werden.

Minimale Anzahl von Punkten pro Segment Über diesen Parameter kann gesteuert werden, wie viele Punkte sich in einem Tracksegment befinden, damit es überhaupt weiterverarbeitet wird. Der minimale Wert für diesen Parameter ist 2.

Verschiedene Fälle beim Filtern

Es gibt drei verschiedene Fälle, welche beim Filtern von Ausreißern abgedeckt werden sollten. Der grundlegende Algorithmus, welcher vom aktuellen Punkt ausgehend einen neuen gültigen Punkt sucht und alle ungültigen überspringt, funktioniert in den ersten zwei Fällen. Im dritten Fall muss noch eine zusätzliche Überprüfung stattfinden.

1. Fall

Beim ersten Fall befinden sich ein oder mehrere Ausreißer am Ende der GPS-Spur wie es in Abbildung 8.1 bei dem letzten Punkt der Fall ist. Dies bedeutet, dass ab einem gewissen Punkt keine weiteren validen Punkte gefunden und alle folgenden Punkte übersprungen werden.

2. Fall

Beim zweiten Fall befinden sich ein oder mehrere Ausreißer zwischen validen vorange-

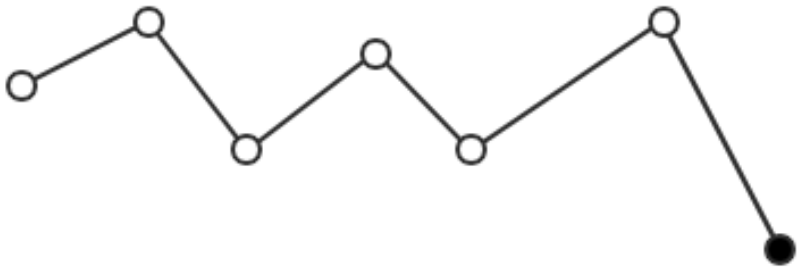

Abbildung 8.1: Filtern - 1. Fall

gangenen und nachfolgenden Punkten. Ein Beispiel ist in Abbildung 8.2 mit dem vierten Punkt als Ausreißer abgebildet. Dies bedeutet, dass ein oder mehrere Punkte übersprungen werden und danach mit den gültigen Punkten weitergearbeitet werden kann.

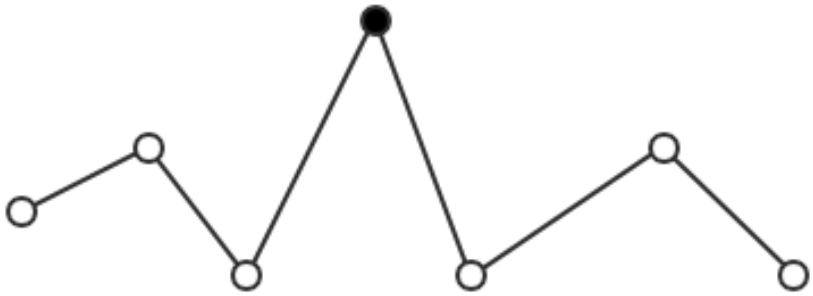

Abbildung 8.2: Filtern - 2. Fall

3. Fall

Im dritten Fall befinden sich ein oder mehrere Ausreißer am Beginn der GPS-Spur. Damit ist gemeint, dass vom Start weg keine gültigen Punkte vorhanden sind und erst im Laufe der Aufzeichnung gültige Punkte aufgezeichnet werden. Dies kann vorkommen, wenn die Aufzeichnung der GPS-Spur sofort nach dem Aktivieren des GPS-Moduls startet. Die Position konnte noch nicht mit ausreichender Genauigkeit bestimmt werden und es wird mit einer niederen Genauigkeit gestartet. Im Laufe der Aufzeichnung steigt die Genauig-

keit und es kann zu einem Sprung von ungenauen zu den genauen Punkten kommen. Ein Beispiel hierfür ist in Abbildung 8.3 mit den ersten 3 Punkten als Ausreißern ersichtlich.

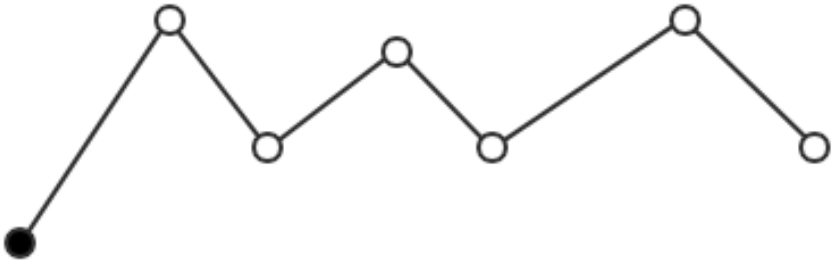

Abbildung 8.3: Filtern - 3. Fall

Da keinerlei Information über die Genauigkeit der aufgezeichneten GPS-Punkte gesammelt wurde, ist es eine komplexe Aufgabe einen gültigen Startpunkt zu finden ohne unnötig viele GPS-Daten zu überspringen. In einzelnen Testfällen kam es vor, dass ein großer Teil der Strecke einfach weggelassen wurde, weil kein gültiger Startpunkt bzw. nicht genügend valide und aufeinander folgende Punkte gefunden werden konnten. Deshalb wurde eine konstanter Parameter für die Anzahl der zu überspringenden GPS-Punkte am Anfang eines Tracks festgelegt.

Zeitfilter

Der Zeitfilter überprüf,t ob der Abstand zwischen zwei GPS-Punkten größer gleich einem minimalen Wert (in diesem Fall 0) ist. Dadurch wird verhindert, dass zwei Punkte mit demselben Zeitstempel verarbeitet werden und bei den zeitabhängigen Berechnungen durch 0 dividiert wird. Außerdem kann man dadurch auch steuern, wie viele Punkte pro GPS-Spur überprüft werden (z.B. nur jeder 2. Punkt), beziehungsweise welche Punkte ausgelassen werden sollen um den Prozess zu beschleunigen oder weil sich der Grad an Genauigkeit nicht wesentlich verbessert.

Distanzfilter

Der Distanzfilter kontrolliert, ob sich der Abstand zwischen zwei Punkten im Verhältnis zur Zeit in einem gewissen Bereich befindet. In dieser Arbeit wurde größer 0 m pro Zeiteinheit als minimale und kleiner 50 m pro Zeiteinheit als maximale Distanz festgelegt. Liegt ein Punkt nicht innerhalb dieser Grenzen so wird der aktuelle Punkt mit dem Punkt nach dem Ausreißer verglichen. Dies wird solange gemacht bis wieder ein Punkt mit valider Distanz gefunden wird oder keine GPS-Punkte mehr vorhanden sind.

Höhenfilter

Der Höhenfilter filtert, ähnlich wie der Geschwindigkeitsfilter, jene GPS-Punkte, bei welchen die Differenz der Höhenwerte zu groß ist. Im Fall der hier verwendeten Trainingsdaten wurde 25 m/s für diesen Filter festgelegt und alle Punkte mit einen größeren Differenz werden herausgefiltert.