**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

# Todo list

# 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 all diese 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, wurde mehre Smartphones mit der App “MyTrack“ sowie mehrere GPS-Geräte benutzt.

Im Zuge dieser Arbeit wird auch untersucht welches Level an Genauigkeit sowohl mit als auch ohne geografischen 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 getroffen werden können, ob noch weitere Zusatzdaten für eine noch genauere Bestimmung benötigt werden würden und welche Daten dies sein könnten.

### Nichtziele

In dieser Arbeit werden die diverse Schlussfolgerungsmodelle (neuronales Netz, Bayessches Netz, Random Forest, Support Vector Machine, ...) nicht betrachtet oder verglichen sondern es wird auf Modell gesetzt das bereits bei anderen Arbeiten wie z.B. [Stenneth et al., 2011,Reddy et al., 2010,Sebastian Nagel, 2011,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 Aussagen über die allgemeine Verkehrssituation machen. 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. Diese Apps könnten die Auswertungen in soziale Medien zu integrieren, für Fitnessanalysen verwendet werden, einen einfachen Rückblick über die eigene Fortbewegung ermöglichen oder 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 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 Filtern als auch dem Aufbereiten der Daten sowie das Aufteilen von GPS-Spuren in Teile in denen nur ein Verkehrsmittel verwendet wurde. Filtern der Daten bedeutet in diesem Zusammenhang, dass Ausreißer aus den GSP-Spuren entfernt werden. Diese Ausreißer können aufgrund von Geschwindigkeitssprüngen als auch unwahrscheinlich große Sprünge im dreidimensionalen Raum sein. Mit Aufbereiten der Daten ist gemeint, dass Werte zu GPS-Punkten für die spätere Analyse berechnet werden. Diese Werte können sowohl die Geschwindigkeit, Beschleunigung, Distanz als auch der Abstand zu der nächsten Bushaltestelle sein. Schlussendlich sollen die gefilterten und erweiterten Daten verwendet werden um eine einfache und sichere Aufteilung der GPS-Spuren anhand der Verkehrsmittel zu ermöglichen.

Aufbauend auf die Resultate 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.

Im fünften und letzten 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 diesem Gebiet der Forschung zählt sicher 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 weiteren 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 auch Leon Stenneth beschäftigt. Dabei hat er nicht nur fixe Daten wie Gleise und Busstationen sondern auch aktuelle Buspositionen miteinbezogen.

Sowohl Stenneth als auch Zheng haben in ihren Arbeiten detailiert 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 Verkehrsmittel 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. Einzig 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 ein Beschleunigungssensor [Reddy et al., 2010, Nadine Schüssler et al., 2011]. Andere wie Leon Stenneth verwendeten Live-Informationen von den öffentlichen Verkehrsmittel 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 Verkehrsmittel von "Gehen und Motorisiert"(siehe [Reddy et al., 2010]) bis hin zu "Gehen, Zug, U-Bahn, Rad, Auto, Straßenbahn, Bus, Fähre, Segelboot und Flugzeug" (siehe [Biljecki et al., 2013]).

## 2.3 Zheng

Zu den Meilensteinen auf diesem Gebiet der Forschung zählen die Publikationen von Yu Zheng und seinem Team (“Understanding Mobility Based on GPS Data“, “Understanding Transportation Modes Based on GPS Data for Web Applications“ und “Learning Transportation Mode from Raw GPS Data for Geographic Application on the Web“). 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

Das Projekt welches 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.

### 2.3.2 Segmentierung

Die Publikationen von Zheng beinhalten auch eine der detailiersten Beschreibungen des Segementierungsvorgangs. 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 Segementieren stützt sich Zheng darauf, dass Personen bei einem Wechsel des Verkehrsmittels stehenbleiben und sich dann ein Stück zu Fuß bewegen. Dies bedeutet, dass es einige GPS-Punkte mit einer Geschwindigkeit von oder beinahe 0 km/h gibt und, dass der Beginn und das Ende eines Segments mit dem Typ “Gehen“ ein wichtiger Indikator für einen Wechsel ist. Diese Aussage stützt sich auf die Daten die von 45 Personen über 6 Monate gesammelt wurden.

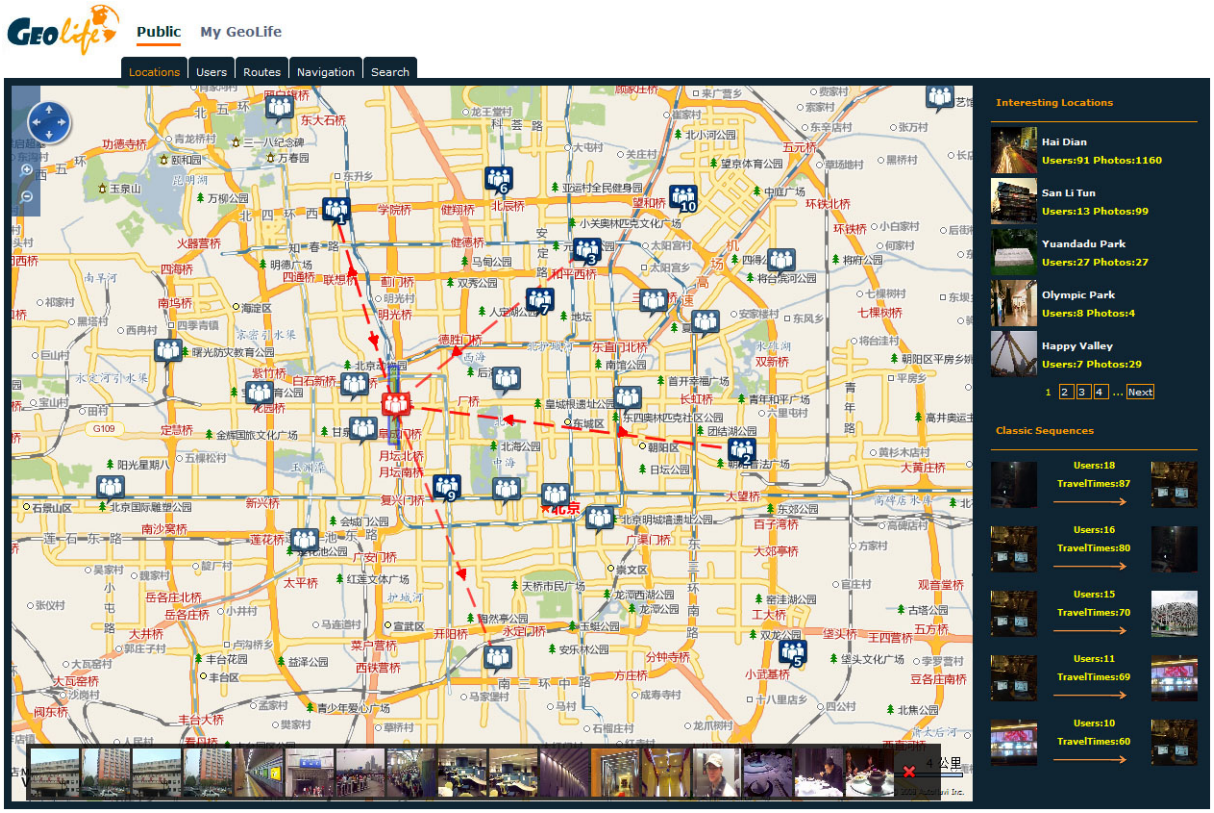


Abbildung 2.1: Geolife (Quelle: research.microsoft.com)

### 2.3.3 Schlussfolgerungsmodelle

Um die Typen der "Nicht-Geh-Segmente" bestimmten zu können hat sich Zheng sich 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 ist. 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 das betrachtete Gebiet widerspiegelt. Mit diesem Graph wurde die Schlussfolgerung nochmals überprüft und es konnte eine weitere Verbesserung erzielt werden. Schlussendlich konnte der von Zheng entwickelte Algorithmus 76,2% der Testfälle ohne jegliche Zusatzinformationen korrekt identifizieren.

## 2.4 Stenneth

Leon Stenneth vergleicht in der Publikation "Transportation Mode Detection usign Mobile Phones and GIS Information" 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. Daraus berechnete er verschiedene Zusatzinformationen für die Bestimmung des Transporttyps:

- Die durchschnittliche Nähe zu den Gleisen ist der euklidische Distanz von den übermittelten Koordinaten zu den nächsten Gleisen.
- Die durchschnittliche Nähe zu Bushaltestellen ist ebenfalls die durchschnittliche euklidische Distanz zur nächsten Bushaltestelle. Basierend auf Erfahrungswerten ist dieser Wert erhöht wenn eine Person mit dem Bus unterwegs ist.
- Die durchschnittliche Nähe zu einem Bus ist wiederum die durchschnittliche euklidische Distanz zu einem Bus. Dies wird verwendet um festzustellen ob ein Reisender mit dem Bus unterwegs ist.

### 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 die Nähe zu einem Bus sowie die Distanz zum nächsten Bus.

Die von Stenneth betrachteten Modelle sind Naive Bayes, Bayesches 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. 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 Biljecki

## 2.6 Reddy

[Reddy et al., 2010, Stenneth et al., 2011, Zheng et al., 2010]

# Daten

Dieser Abschnitt behandelt die Akquirierung und die Struktur der verwendeten GPS-Daten für das Training des Entscheidungsbaums. Außerdem wird erklärt 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 erklärt 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 erklärt woher die verwendeten GIS-Daten stammen, wie diese extrahiert wurden und welche Rolle sie im weiteren Prozess spielen. Abschließend wird auch auf weitere Daten 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: Datenübersicht

## 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, columbus, photomate, qstarz, 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

Diese Trainingsdaten sind in den einzelnen Dateien als XML abgelegt und entsprechen dem gängigen GPX-Format wie es im Listing 3.1 ersichtlich ist. Üblicherweise enthält

```
<?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

eine solche Datei einen Track (trk) welcher aus mehreren Tracksegmenten (trkseg) bestehen kann. Die Tracksegmente bestehen wiederum aus beliebig vielen Trackpoints (trkpt) welche eine genaue Koordinate sowie einen Zeitstempel und die Höhenmeter beinhalten.

### 3.1.2 Entscheidungsbaum

Aufgrund des guten Abschneidens des Entscheidungsbaums in verschiedenen Arbeiten [Stenneth et al., 2011,Reddy et al., 2010,Sebastian Nagel, 2011,Zheng et al., 2008b] wurde auch in dieser Arbeit ein Entscheidungsbaum verwendet. Dazu wurden die Trainingsdaten mit Hilfe des Prototyps analysiert und in die Ergebnisse als CSV abgelegt. Daraus konnte dann mit Hilfe des Tools RapidMiner ein Entscheidungsbaum generiert werden.

### Entscheidungsbaum ohne GIS-Daten

Der Entscheidungsbaum ohne Berücksichtigung von GIS-Daten ist in Abbildung 3.1 zu sehen. Bei Entscheidungsbäumen muss beachtet werden, dass sie nicht zu sehr auf die Trainingsdaten angepasst werden.

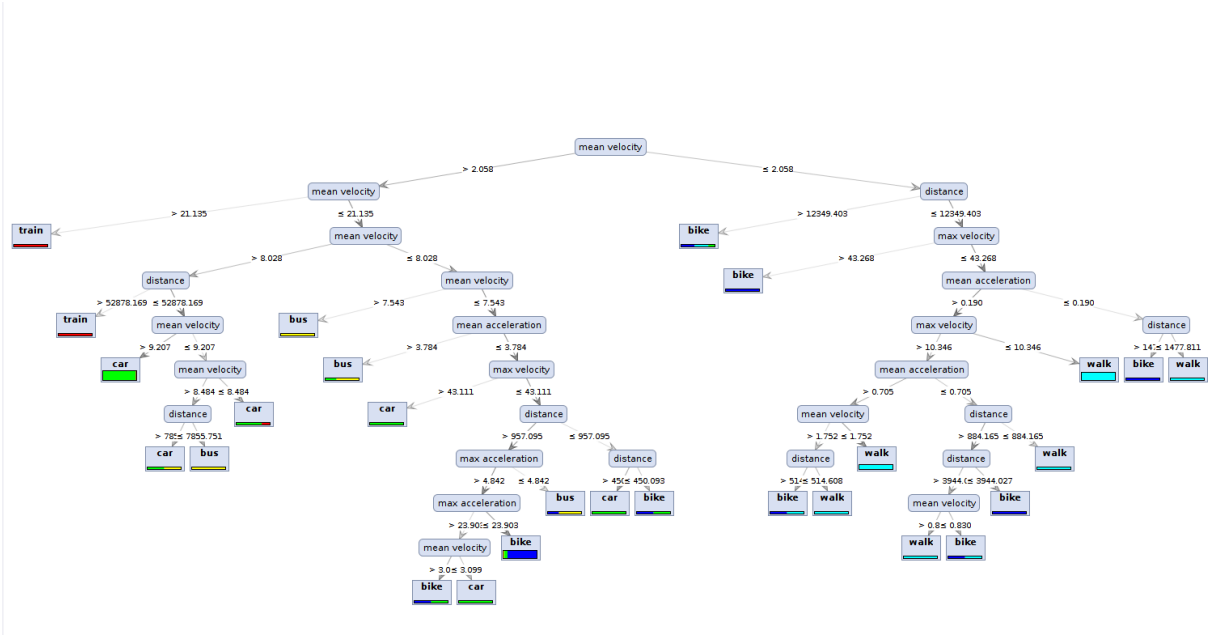


Abbildung 3.1: Entscheidungsbaum ohne GIS-Daten

### 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 und dient nur zur optischen Unterscheidung.

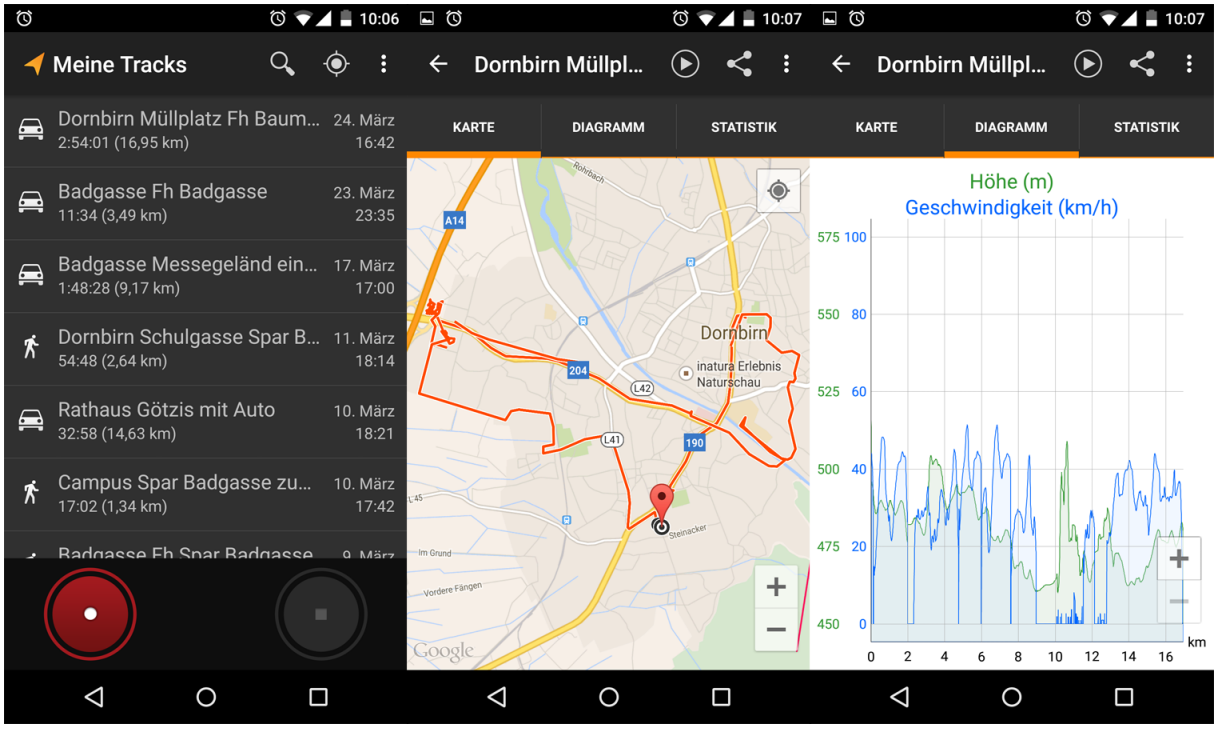


Abbildung 3.2: Die App myTrack

## 3.3 GIS-Daten

Als relevante GIS-Daten kommt in den verschiedensten Arbeiten viel in Frage, darunter Parkplätze, Busstationen, Gleise, Bahnhöfe und das gesamte Straßennetz. All diese Daten mögen zwar relevant sein, aber da der Prototyp auch für konkrete Benutzer relevant sein soll, wird auf die Verwendung des Straßennetzes und der Parkplätze verzichtet, da dies schlicht zu viele Daten sind. Darum 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.

**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 Relationen gefiltert werden. Dadurch war es möglich das Schienennetz von Vorarlberg sowie die Busstationen von Vorarlberg zu exportieren.

## 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

## 4.1 Funktionalitäten

## 4.2 Aufbau und Architektur

# Filter

Wie auch bei vielen anderen Arbeiten 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, zu kleine Zeitabstände oder Sprünge in den Höhenwerten bemerkt. Da diese Werte das Ergebnis verfälschen würden, mussten verschiedene Filter implementiert werden.

Konkret wurden Filter für Zeit, Distanz und Höhenmeter implementiert. Die bereinigten Resultate hatten bereits auf den Entscheidungsbaum große Auswirkungen. Die Grenzwerte sind in der Konfiguration festgelegt und können je nach Region und Testdaten angepasst werden.

Punkte werden insofern gefiltert, als dass sie übersprungen werden und der aktuelle Punkt mit dem nachfolgenden des Ausreißers verglichen wird. Die übersprungenen Punkte werden dann nicht für die weitere Verarbeitung berücksichtigt. Abgesehen von Zeitfilter betrachten alle anderen Filter die gemessen Werten in Relation zur gemessenen Zeit was bedeutet, dass der Zeitwert größer als 0 sein muss.

## 5.1 Verschiedenen 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 befindet sich ein oder mehrere Ausreißer am Ende der GPS-Spur wie es in Abbildung 5.1 bei dem letzten Punkt der Fall ist. Dies bedeutet, dass ab einem gewissen Punkt keine weiteren validen Punkte gefunden werden und alle folgende Punkte übersprungen werden.

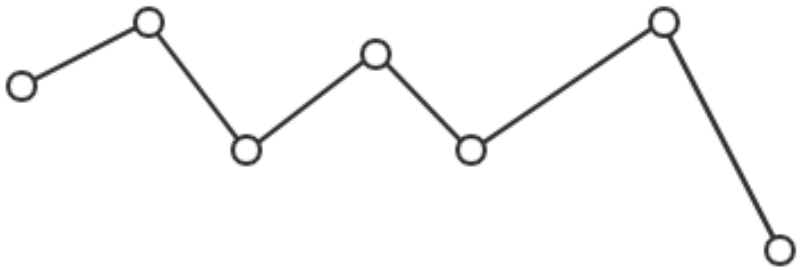

Abbildung 5.1: Filtern - 1. Fall

**2. Fall**
Beim zweiten Fall befinden sich ein oder mehrere Ausreißer zwischen validen vorangegangenen und nachfolgenden Punkten. Ein Beispiel ist in Abbildung 5.2 mit dem vierten Punkt aus Ausreißer abgebildet. Dies bedeutet, dass ein oder mehrere Punkte übersprungen werden und danach mit den gültigen Punkten weitergearbeitet werden kann.

**3. Fall**
Im dritten Fall befinden sich ein oder mehreren Ausreißern 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

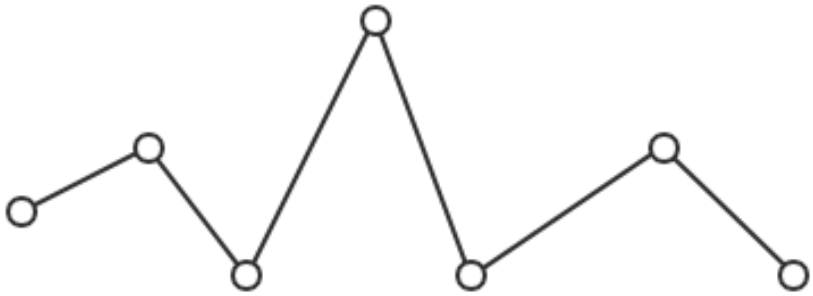

Abbildung 5.2: Filtern - 2. Fall

es wird mit einer niederen Genauigkeit gestartet. Im Laufe der Aufzeichnung steigt die Genauigkeit und es kann zu einem Sprung von ungenauen zu den genauen Punkten kommen. Ein Beispiel hierfür ist in Abbildung 5.3 mit dem ersten Punkt als Ausreißer ersichtlich.

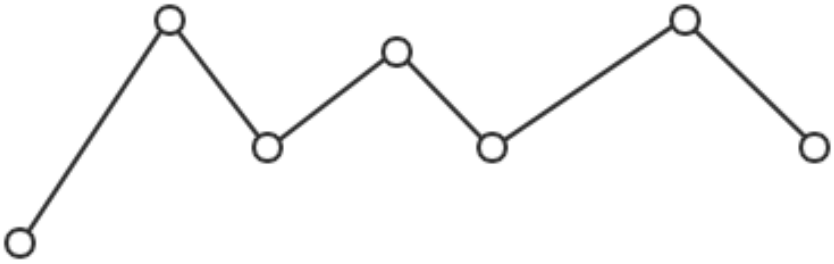

Abbildung 5.3: Filtern - 3. Fall

## 5.2 Zeitfilter

Der Zeitfilter überprüft ob der Abstand zwischen zwei GPS-Punkten größer gleich einem minimalen Wert ist. Dadurch wird verhindert, dass zwei Punkte mit demselben Zeitstempeln 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.

## 5.3 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/s als minimale und kleiner 50 m/s als maximale Distanz festgelegt. Liegt ein Punkt nicht innerhalb dieser Grenzen so wird der aktuelle Punkt mit den 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.

## 5.4 Höhenfilter

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

## 5.5 Aufbereitung ohne GIS-Daten

### 5.5.1 Auswahl

**Durchschnittliche Geschwindigkeit**

**Maximale Geschwindigkeit**

**Durchschnittliche Beschleunigung**

**Maximale Beschleunigung**

**Höhenmeter**

**Distanz**

## 5.6 Aufbereitung mit GIS-Daten

### 5.6.1 Auswahl

#### Durchschnittliche Geschwindigkeit

#### Maximale Geschwindigkeit

#### Durchschnittliche Beschleunigung

#### Maximale Beschleunigung

### 5.6.2 Abstand zu Bushaltestellen

### 5.6.3 Abstand zu Gleisen

## 5.7 Segmentierung

### 5.7.1 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):385–407.

[Microsoft Research, 2015] Microsoft Research (2015). GeoLife GPS Trajectories - Microsoft Research.

[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):13.

[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.

[Youtube, 2015] Youtube (2015). Youtube statistics.

[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):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

# Tabellenverzeichnis

# Abbildungsverzeichnis

# Anhang