Gebärdenspracherkennungsmodelle: Datenaufnahme, Entwicklung und Leistungsvergleich

Florian Brunner, Ludwig Breu, Sebastian Forster, Regina Oßner

Betreut von: Prof. Dr. Sandra Eisenreich und Prof. Dr. Eduard Kromer

Ziel

Ziel des Projekts ist es ein Modell zu entwickeln, das zur Laufzeit aus Videos Gebärdensprache in Text übersetzen kann. Langfristig soll eine Mobile-App entstehen, die zur Echtzeit Gebärden in Text übersetzt, um Probleme im Alltag zu lösen.

Einleitung

- In Deutschland gibt es etwa 300.000 Menschen mit Hörbeeinträchtigungen [1] die sich mit Gebärdensprache verständigen
- Es gibt derzeit keine leistungsfähigen Sprachmodelle für die Übersetzung in oder aus Gebärdensprache wie sie es für gesprochene Sprachen gibt
- Projekte wie SignON und EASIER kooperieren mit Universitäten und der European Union of the Deaf, um den Fortschritt in der Gebärdensprachübersetzung voranzutreiben [2][3]
- Ziel ist es, Bewegungen aus Gebärdensprachaufnahmen zu erkennen und Algorithmen des maschinellen Lernens zu trainieren
- Herausforderung besteht darin, ausreichend große und vielfältige annotierte Korpora zu generieren
- Gebärdensprachen variieren stark und die Erfassung der Gebärden im dreidimensionalen Raum ist komplex
- Es existieren keine festen Regeln für die Verwendung bestimmter Gebärden [4]

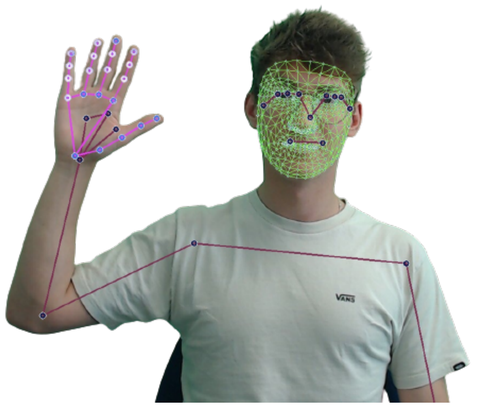

Abbildung 1. Gebärde Hallo mit Keypoints Overlay (Eigene Erstellung)

Methode

Erster Ansatz

- Aufnahme von Videos einzelner Gebärden
- Erkennen der Gebärden anhand von Keypoints auf den Händen und Gesicht, siehe Abb. 1
- Anschließendes LSTM trainiert mit den Daten aus den Keypoints, entnehme Ablauf aus Abbildung 2
- Performance Metrik: **Categorical Accuracy**

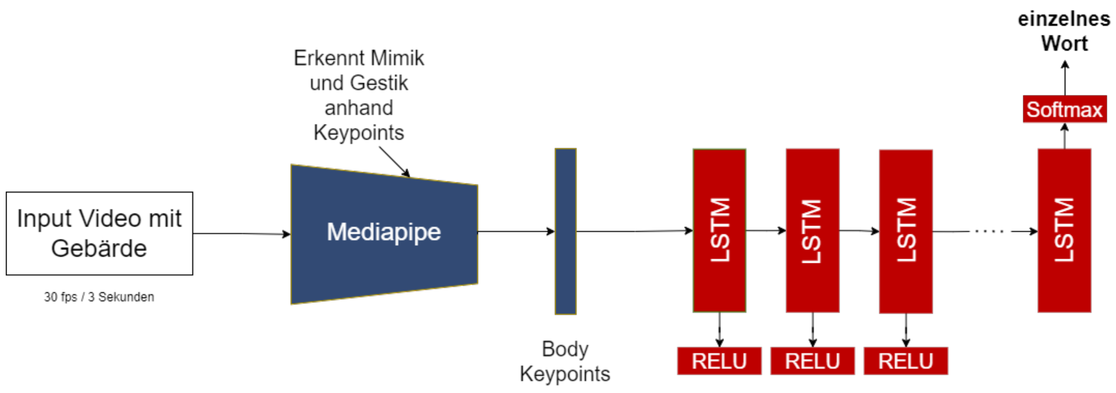

Abbildung 2. Grafische Darstellung der Architektur des ersten Modell Ansatzes (Eigene Erstellung)

Zweiter Ansatz

- Verwenden eines vorhandenen Datensatzes bestehend aus Videos längerer Sequenzen von Gebärden
- CNN zur Erkennung der Bewegungen und anschließendes LSTM trainiert mit Features aus dem CNN, entnehme aus Abbildung 3
- Correlation und Identification Module im CNN [5], visualisiert in Abb. 5.
- Performance Metrik: **WER Score** (misst den Prozentsatz der Wörter, die in der erkannten Ausgabe im Vergleich zur Referenzausgabe falsch erkannt oder hinzugefügt wurden oder fehlen) [5]

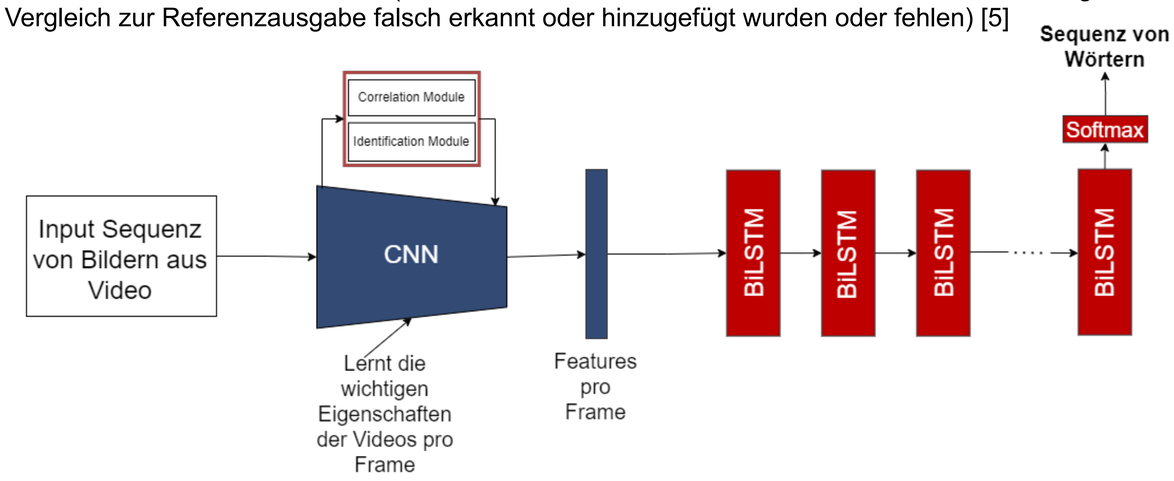

Abbildung 3. Grafische Darstellung der Architektur des zweiten Modell Ansatzes (Eigene Erstellung)

Ergebnis

Erster Ansatz

- Modell weist eine Genauigkeit von 0,9 auf Testdaten auf, wie zu sehen auf Abb. 4
- Aufnahme von Testdaten gestaltet sich zeitaufwendig und anspruchsvoll
- Echtzeitvorhersagen des Modells für Gebärdenerkennung sind unzureichend
- Modell kann nur einzelne Gesten bestimmen

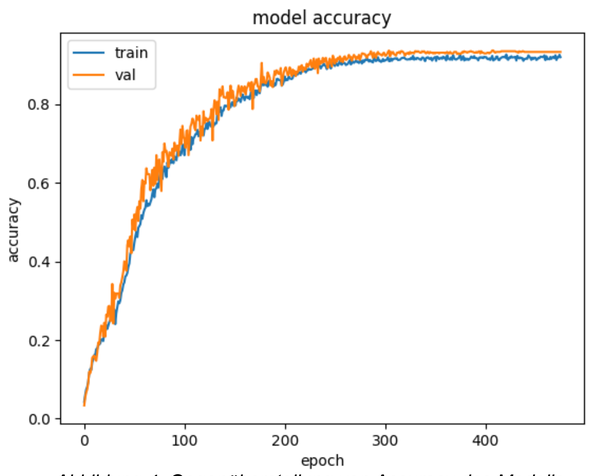

Abbildung 4. Gegenüberstellung von Accuracy des Modells auf Trainings- und Validierungsdaten über trainierte Epochen hinweg (Eigene Erstellung)

Zweiter Ansatz

- Keine Aufnahme von Testdaten nötig
- Vorhandene Datensatz enthält 8247 Sätze mit einem Wortschatz von 1085 Zeichen [5]
- Modell kann ganze Sequenzen von Gesten bestimmen
- Vermutlich bessere Performance bei Echtzeitvorhersagen
- Datensatz auf deutsche Gebärden
- WER Score: 20,5% (Je niedriger der WER desto besser die Accuracy) [5]

Tabelle 1. Vergleich der WER Scores von state-of-the-art Methoden die auf demselben Datenset getestet wurden

Methode	WER Score
SMKD	22.4%
TLP	21.2%
SEN	20.7%
SLT	24.6%
STMC	21.0%
C2 SLCR	20.4%
Projektmodell	**20.5%**

Correlation und Identification Module

- Berechnen die Korrelationen basierend auf informativen Regionen in benachbarten Frames [5]

Abbildung 5. Visualisierung von Korrelationen aus dem Korrelationsmodul des CNNs

Diskussion

1. Diskussion der Ergebnisse
   - Hohe Genauigkeit des ersten Modells (0,9) durch Verwendung großer Video-Datenmenge pro Geste
   - Aufnahme der Testdaten zeitaufwendig und anspruchsvoll
   - Unzureichende Echtzeitvorhersagen des ersten Modells (80% Fehlerrate)
   - Erkennt nur einzelne Gesten, keine Sequenzen
   - Zweiter Ansatz erfordert keine Testdatenaufnahme und kann ganze Sequenzen bestimmen
2. Vergleich mit Literatur
   - Literaturvergleich zeigt realistische Ergebnisse beider Modelle unter Beachtung des kleinen Testdatensatzes des ersten Modells
3. Konsequenzen für weitere Schritte und entwickelte Technologie
   - Erster Ansatz ungeeignet auch für mobile App, zweiter Ansatz vielversprechender
4. Diskussion der eigenen Vorgehensweise und Limitationen
   - Begrenztes Wissen über Gebärdensprache führte zu mangelhaften Datenaufnahmen
   - Zweiter Ansatz mit öffentlich verfügbarem Datensatz als bessere Alternative
5. Ausblick
   - Möglicher nächster Schritt ist Feintuning des zweiten Modells mit Testdaten, Entwicklung einer Echtzeit-Gebärdenerkennungs-App
6. Fazit
   - Notwendigkeit von ausreichend großen und vielfältigen Daten für optimale Modellleistung

Literatur

[1] Statistisches Bundesamt, Statistik der schwerbehinderten Menschen, Kurzbericht 2021
[2] SignON Project. „SignON Project - Sign Language Translation Mobile Application“. Zugegriffen 21. Mai 2023. https://signon-project.eu.
[3] EASIER – Intelligent Automatic Sign Language Translation. „About EASIER“, 2. März 2021. https://www.project-easier.eu/about-easier/.
[4] Sisto, Mirella De, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, und Horacio Saggion. „Challenges with Sign Language Datasets for Sign Language Recognition and Translation“, o. J.
[5] Hu, Lianyu, Liqing Gao, Zekang Liu, und Wei Feng. „Continuous Sign Language Recognition with Correlation Network“. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.