

Budapesti Műszaki és Gazdaságtudományi Egyetem

Villamosmérnöki és Informatikai Kar

Dancsó Marcell

Jelnyelv fordítás

TDK dolgozat

Konzulens

BUDAPEST, 2023

Tartalomjegyzék

[Összefoglaló 3](#_Toc148516031)

[Abstract 5](#_Toc148516032)

[1 Bevezetés 7](#_Toc148516033)

[1.1 A dolgozat felépítése 7](#_Toc148516034)

[1.2 Jelnyelv alapok 8](#_Toc148516035)

[1.3 Amerikai jelnyelv (ASL) 9](#_Toc148516036)

[1.3.1 Ujjbetűzés az amerikai jelnyelvben 9](#_Toc148516037)

[1.3.2 Általános ASL 10](#_Toc148516038)

[2 Kapcsolódó kutatások 11](#_Toc148516039)

[2.1 Hagyományos képfeldolgozáson alapuló módszerek 11](#_Toc148516040)

[2.2 Segédeszközt használó megoldások 11](#_Toc148516041)

[2.3 A dolgozat célja 12](#_Toc148516042)

[3 Megközelítés 13](#_Toc148516043)

[3.1 Póz approximáció 13](#_Toc148516044)

[4 Statikus ujjbetűzés 14](#_Toc148516045)

[5 Ujjbetűzés szekvenciális bemenetből 15](#_Toc148516046)

[6 Hagyomás jelelés 16](#_Toc148516047)

[7 Általános jelnyelv fordítása 17](#_Toc148516048)

[8 Összegzés 18](#_Toc148516049)

[8.1 Formázási tudnivalók 18](#_Toc148516050)

[8.1.1 Címsorok 18](#_Toc148516051)

[8.1.2 Képek 18](#_Toc148516052)

[8.1.3 Kódrészletek 18](#_Toc148516053)

[8.1.4 Irodalomjegyzék 19](#_Toc148516054)

[9 Utolsó simítások 20](#_Toc148516055)

[Irodalomjegyzék 21](#_Toc148516056)

[Függelék 22](#_Toc148516057)

Összefoglaló

A nyelvfeldolgozás terén elért sikerek rengetek kaput nyitottak ki előttünk. A hangalapú asszisztensek, alap esetben is hasznos, de legtöbbször forradalmi funkciókat hoznak életünkbe. A mesterséges intelligencián alapuló „voice to text” modellek manapság gyakorlatilag tetszőleges nyelvről képesek felismerni szöveget, valamint a közelmúltban nyelvfeldolgozás területén elért eredményeknek köszönhetően pedig nincsenek határok, a felhasználók, és fejlesztők előtt. Illetve előbbi állítás sajnos csak egy nagy ferdítéssel igaz. Ugyanis világszerte körülbelül 1,5 milliárd ember hallássérült, és több mint 70 millió ember használja a jelnyelvet, mint elsődleges kommunikációs formát. Számukra „saját nyelvükön” ezek a funkciók nem, vagy csak korlátozottan érhetőek el. A jelnyelv egy speciális fajtája az ujjbetűzés. Gyakran használják nevek, címek, telefonszámok, valamint olyan fogalmak közvetítése során, melyekre nincs bevett gesztus. Egy tapasztalt jelelő képes közel kétszer olyan sebesen betűzni, mint egy virtuális billentyűzeten pötyögő egyén, nem beszélve arról ha minden jelet alkalmazva kommunikál. Így égető egy számukra is kényelmesen használható interfész kialakítása.

Még jelentősebb problémát vet fel a kommunikáció kérdése. Manapság a nyelvek közötti gépi fordítás széles körben és formátumban elérhető az internet kapcsolattal rendelkezőknek. Régen túl vagyunk már az egyszerű szótár alapú fordítókon, a különböző transzformer, és nagy nyelvi modellek térhódításával gyakorlatilag tetszőleges nyelvek között megoldható a kommunikáció. Ez alól kivétel azonban a jelnyelv. Nem létezik olyan megoldás a piacon, ami segíti a jelelő kommunikációját a jelelni nem tudó személy felé. A terület előrehaladásán nem segít, hogy a jelnyelv csakúgy mint a verbális párjai, nem standardizált. Szinte minden nemzetnek saját jelnyelve van, nem beszélve a helyi sajátosságokról, és dialektusokról. Tovább nehezíti a feladatot, hogy a rendelkezésre álló, nagy méretű, és jó minőségű adatbázisok száma eltörpül a klasszikus nyelvfeldolgozásban megszokottaktól.

Munkámmal a széles tömegek számára elérhető jelnyelv fordító rendszerek fejlesztéséhez járulok hozzá. Megvizsgálom többek között a kesztyűvel történő felismerés, valamint hagyományos képfeldogozás eredményeit, továbbá részletesen foglalkozom a póz approximáción alapuló algoritmusokkal. A nyílt kérdés utóbbival kapcsolatban, hogy bár rendkívül jó arányban tömöríti a képi adatot, ami lehetővé tenné nagy mennyiségű, egységes adatbázis konstruálását, de vajon tart-e ott a technológia, hogy megbízható módon kódoljon minden fordításhoz szükséges információt?

Az amerikai jelnyelven belül külön-külön vizsgálom az ujjbetűzés, és általános jelbeszéd lehetőségeit, kitérve a pillanatképből, valamint mozgásszekvenciából dolgozó megoldásokra. A teljesség igénye nélkül összehasonlítom a feladatra adaptált konvolúciós, rekurrens, LSTM, Transzformer háló architektúrákat. A jelelés nyelvtani adottságai következtében, hiányoznak a segédigék, valamint gyakran más a szórend a hagyományos angolhoz képest. A primitív fordítás eredményeit ezért kontextus függő módon transzformálom generatív nagy nyelvi modellekkel, valamint az ujjbetűzés esetében a megbízhatóság növelésére is felhasználom őket.

A dolgozatomban bemutatom az adatgyűjtés és a tervezés lépéseit, az alkalmazott mesterséges intelligencia algoritmusok részleteit, valamint elemzem a fejlesztés során felmerült tervezői döntéseket, és alternatíváikat. Ezen felül bemutatásra kerül az általam készített teljes megoldás, mely nagymértékben segíthet a hallássérültekkel való kommunikáció során.

Abstract

The successes achieved in the field of natural language processing have opened up numerous doors before us. Voice-based assistants are not only useful, but they often introduce revolutionary functionalities into our lives. Nowadays, artificial intelligence-based voice-to-text models are capable of recognizing text from virtually any language, and thanks to recent advancements in natural language processing, there are no boundaries for users and developers. However, the previous statement is a significant distortion of the truth. Globally, approximately 1.5 billion people are hearing impaired, and more than 70 million people use sign language as their primary means of communication. For them, these functionalities are either not available in their „own language”, or only in a limited manner. Sign language, particularly finger-spelling, is a special form of communication often used for communicating names, addresses, and phone numbers, or conveying concepts without standard gestures. An experienced signer can spell almost twice as fast as an individual typing on a virtual keyboard, let alone when using all available signs for communication. Hence, there is an urgent need to create an interface that is comfortable for them to use.

An even more significant issue arises in the realm of communication. Machine translation between spoken languages is widely available to those with internet access in various formats. We have long surpassed simple dictionary-based translators, and with the rise of different transformers and large language models, communication between virtually any language has long become a reality. However, sign language even today remains an exception. No solution on the market assists signers in communicating with non-signing individuals. The progress in this area is hindered by the fact that sign language, like its verbal counterparts, is not standardized. Almost every nation has its own sign language, not to mention regional variations and dialects. Furthermore, the task is complicated by the fact that the available large and high-quality datasets are minuscule compared to what is customary in classical natural language processing.

My work contributes to the development of sign language translation systems accessible to the broader masses. I have conducted research encompassing various aspects, including glove-based recognition and traditional image processing methods. However, my primary focus was directed towards the comprehensive exploration of using pose approximation algorithms. While the latter is highly promising as it compresses visual data effectively, enabling the construction of large, homogeneous databases, the question of whether current technology is capable of reliably encoding all the information needed for translation has so far remained open.

Within American Sign Language, I separately explored finger-spelling and general sign language possibilities, paying close attention to solutions working from both snapshots and motion sequences. Without aiming for completeness, I compared convolutional, recurrent, LSTM, and Transformer network architectures adapted for the task. Due to the grammatical characteristics of sign language, auxiliary verbs are missing, and the word order is often different from traditional English when translated word by word. Therefore, I processed the results contextually using generative large language models and employed them to enhance the quality and reliability of the translation.

In my thesis, I walk through the steps of data collection, provide details about the applied artificial intelligence algorithms, and analyze the design decisions and their alternatives that emerged during development. Furthermore, I present the comprehensive solution I've developed, offering significant support for communication with the hearing impaired.

# Bevezetés

A nyelvfeldolgozás területén hatalmas áttöréseket értünk el az elmúlt években. A mai „voice to text” modellekkel lehetőségünk van rendkívül sokszínű és természetes módon interakcióba lépni az eszközeinkkel. Kiegészülve a világot alapjaiban megrengető intelligens nagy nyelvi modellekkel, saját virtuális asszisztenst alkalmazhatunk, chat formában érhetjük el a világ tudását és nem utolsó sorban kinyílik a kapu a szabad kommunikáció előtt tetszőleges nyelvek között. Az hasonló rendszerek egyik sajnálatos velejárója, hogy csupán hagyományos nyelveken érhetőek el. Bár látszólag a technikai tudás rendelkezésre áll, még sem jelent meg még számottevő jelnyelvet támogató rendszer a piacon. A jelnyelv szöveggé formálásával közel 70 millió embernek nyílna lehetősége saját nyelvén interakcióba lépni okos eszközökkel, nem beszélve a jelelni nem tudók felé irányuló kommunikációról. Nagy segítséget nyújthatna továbbá a tanulásban azoknak, akik annak ellenére, hogy hallássérüléssel kénytelenek élni, nincs lehetőségük megtanulni, mind pénzügyi, mind tanulást segítő eszközök híján. Ezen csoport mérete meglepően még jelentősebb, mint a jeleni tudók száma, bizonyítva a nyelv komplexitását, és ezzel a feladat nehézségét.

## A dolgozat felépítése

A bevezető után, rövid leírást adok a jelnyelvek felépítéséről, illetve eltéréseiről a hagyományos nyelvekhez képest, részletesebben foglalkozva az amerikai jelnyelv sajátosságaival.

Megvizsgálom a téma irodalmát, kitérve a segédeszközöket felhasználó, valamint hagyományos képfeldolgozáson alapuló algoritmusokra. Összehasonlítom más kutatások eredményeit, valamint az alkalmazott módszerek hátrányait. Utóbbiból kiindulva szemléltetem a dolgozat motivációját, célkitűzéseit.

Ezt követően nagy léptékben haladva, ismertetem a kutatás menetét, alkalmazott módszertanokat, valamint a kiindulási alapként szolgáló megoldások algoritmikus hátterét.

A következő fejezetekben külön-külön vizsgálom a póz approximáción alapuló ujjbetűzés lehetőségeit mind pillanatképekből, valamint szekvenciális képkockákból, a hagyományos jelelésre kialakított megoldásom, valamint részletezem a kettő ötvözésének lehetőségét. Részletes leírást adok az adatgyűjtés folyamatáról, egyes adathalmazok leírásáról, illetve feldolgozásukról. Ismertetem a kialakított modelleket, nagy hangsúlyt fektetve a tervezői döntések indoklására, alternatívák feltárására, továbbá számos metrika szerint kiértékelem, és összehasonlítom őket. A legjobb modellekhez különböző „usecase”-ek szerint csoportosítva kiegészítő algoritmusokat tervezek.

Végül összefoglalom az eredményeket, további javaslatokat teszek a fejlesztésre, és megválaszolom a kérdést: Vajon a póz approximáció alkalmas-e jelen formájában komplex rendszerek kialakítására.

## Jelnyelv alapok

A jelnyelv egy vizuális gesztusokra nyelv, amelyet a hallássérült és a halló közösségek egyaránt használnak a kommunikációhoz. A beszélt nyelvektől eltérően a jelnyelvben a jelentést a kézmozdulatok, az arc kifejezései és a testtartás hordozzák.

Míg sokan tévesen úgy gondolják, hogy a jelnyelv egyetemes, valójában számos különböző jelnyelv létezik világszerte. Minden ország saját jelnyelvvel rendelkezik. Érdekes például, hogy az amerikai jelnyelv (ASL) és a brit jelnyelv (BSL) annyira különbözőek, hogy nem érthetőek egymás számára, annak ellenére, hogy mindkét ország angolul beszél. A két nyelvnek mellesleg teljesen különböző fejlődése, míg az ASL-nek francia jelnyelvi gyökerei vannak, addig a BSL teljesen külön fejlődött. Továbbá hasonlóan a verbális nyelvekhez, itt is léteznek dialektusok, helyi sajátosságok, ami további komplexitáshoz vezet intelligens megoldások kialakításakor.

A legtöbb jelnyelv egy speciális részhalmaza az ujjbetűzés, amely kézzel formázott betűk sorozata. Néhány jelnyelvben az ujjbetűzést gyakran használják nevek, idegen szavak vagy specifikus terminológiák kifejezésére, amelyeknek nincs saját "jelük". Más jelnyelvekben az ujjbetűzést csak ritkán használják, és az emberek inkább a teljes kifejezések és mondatok használatát részesítik előnyben. Számos kisebb jelnyelv létezik világszerte, amelyek különleges közösségekben alakultak ki és nincsenek befolyásolva a nagyobb jelnyelvektől. A Kolok Kata, más néven balinéz jelnyelv, egy falusi jelnyelv, amely két szomszédos faluban őshonos Bali északi részén, Indonéziában. A Kata Kolokban nincs hivatalos ujjbetűzés, ami rámutat arra, hogy az ujjbetűzés nem szükséges egy jelnyelv működéséhez. Ezek a kisebb jelnyelvek gyakran egyedülállóak és tükrözik az adott közösség kultúráját és történelmét.

Mint minden nyelvnek, a jelnyelveknek is van saját nyelvtana és szókincse. A szavak és mondatok jelentését a kézformák, helyzetek, mozgások és az arc kifejezései együttesen hordozzák. A nyelvtan komplex és nem feltétlenül követi a hallók által beszélt nyelv szerkezetét.

## Amerikai jelnyelv (ASL)

Az amerikai jelnyelv (ASL) az Egyesült Államokban és Kanadában élő hallássérült közösségek által használt nyelv. Az ASL története a 19. századig nyúlik vissza, amikor Thomas Gallaudet és Laurent Clerc megalapították az Egyesült Államok első hallássérült iskoláját. Az itt használt jelnyelv kombinálta az amerikai helyi jelnyelveket a francia jelnyelvvel, ami az ASL alapját képezte.

Közel félmillió ember beszéli, ami meglepő, hiszen ez a szám körülbelül 10%-a a 40 millió hallássérültnek az országban. Az ASL a 7. legtöbbet használt, viszont az interneten szabadon hozzáférhető források által legjobban dokumentált jelnyelv a világon, így a kutatás is ezzel foglalkozik. Fontos kiemelni azonban, hogy hasonlóan rögzített adathalmazok esetén a modellek, és algoritmusok univerzálisak.

### Ujjbetűzés az amerikai jelnyelvben

Mint sok más jelnyelvben, az amerikai jelnyelvben is található ujjbetűzés. A brit jelnyelvvel ellentétben itt csupán egy kézre van szükség, a számok, illetve angol abc betűinek kommunikációjára. Ez persze nem azt jelenti, hogy az egyéb gesztusoknak nincs jelentése ebben a kontextusban, viszont a karakterek teljes mértékben beazonosíthatóak csupán a kezek megfigyelésével. pl.: A fejmozgás, vagy előre dőlő testhelyzet ujjbetűzés során hangsúlyozó szereppel bírhat.

Az ujjbetűzés olyan kézformákat használ, amelyek egyes betűket képviselnek, hogy kifejezzenek szavakat. Bár az ujjbetűzés csak az ASL egy része, gyakran használják nevek, címek, telefonszámok és egyéb, mobiltelefonon gyakran beírt információk közlésére. Sok siket okostelefon-felhasználó gyorsabban tudja ujjbetűzéssel kifejezni a szavakat mintsem, hogy a mobilbillentyűzeten begépelje őket. Valójában az ASL ujjbetűzése jelentősen gyorsabb lehet, mint a tipikus okostelefonos virtuális billentyűzeten való gépelés (átlagosan 57 szó/perc az amerikai 36 szó/perc átlagához képest).

A legtöbb gesztus statikus, vagyis nem szükséges dinamikus mozdulat a megkülönböztetéshez. Ez alól kivétel a j és z betűk, melyeket ugyan azzal a kéz pózzal kell jelelni, mint az i és d betűt, csupán a mutató ujjal leírt alakzatban térnek el.

### Általános ASL

A legfontosabb különbség az ujjbetűzéshez képest, hogy sokkal fontosabb szerepet játszik a fej, és mimika, valamint egyéb gesztusok is. Egy jó példa erre a jelenségre az Anya, illetve Apa gesztus, melyek ugyanúgy széttárt tenyérrel viszont a fej különböző részét érintve fejezhetőek ki. Ez nem is egyedi eset, léteznek adatbázisok, amelyekben kéz alak alapján lehet gesztusokat keresni. Módosító tényező lehet még a másik kéz alakja, és a végrehajtott mozgáskombináció is.

A hagyományos angolhoz képest általában más a szórend. Tipikusan előre helyezi az idő vagy igeidő információt, majd a mondat témáját, és végül a megjegyzést. Például az "I am going to the store" (Én megyek a boltba) mondatot ASL-ben így jeleznénk: "NOW I STORE GO" (MOST ÉN BOLT MENNI), de az is gyakori, hogy az igét megismétlik a jelelés elején és végén. Ezen felül, mikor írásos formában akarjuk ábrázolni a jelnyelvet, még szokás vizuális annotációkkal is ellátni. Ezt nevezi az ASL „gloss” vagy „glossing” -nak. Ebben az olyan jelek is feltüntetésre kerülnek, melyeknek nincs angol megfelelője.

# Kapcsolódó kutatások

## Hagyományos képfeldolgozáson alapuló módszerek

Az ilyen módszeren alapuló megoldásokra igaz, hogy bár a sikeres projektek átalakítják a kamera képeit valamilyen magasabb szintű reprezentációba, de sosem lépnek ki a kép dimenzióból, nem modellezik a kéz fiziológiáját. Gyakori például a kezek szegmentálása. Vagy a dinamikus információ kinyerése az egyes képkockák pixeleinek különbségéből.

### Jelnyelv automatikus fordítása többfolyamos 3D CNN-nel és mesterséges mélységtérképek generálása

Ebben a kutatásban videó képkockáiból azonosítottak be jeleket. Első lépésként a képkockák számát csökkentették le. Egy előtanított VGG16 nevű modellt alkalmaztak a képkockák csoportosítására, majd K-means klaszterezés és Fő komponens analízis alkalmazásával 10 képet választanak ki a videóból. Erre a modellek és tanítási idő komplexitás csökkentése miatt van szükség. Ezt követően több érdekes algoritmust is kipróbáltak, egy újabb képi dimenzió bevezetésére a videó mellé.

* Optical flow, vagyis a tárgyak látszólagos mozgásának mintázatát közelítő módszert alkalmaztak az egyes képkockák között. Ehhez Gunnar Farneback 2003-ban publikált algoritmusát alkalmazták. Az így keletkező újabb képkockákat, hasonlóan a következő módszerhez, az RGB dimenzióhoz csatolták, egy újabb „optical flow” dimenzióként.
* A másik módszer mélység információt tartalmazó képek alkalmazása. Ehhez azonban nem egy második, mélységélességre is érzékeny kamerát alkalmaztak, hanem algoritmikusan generálták. A feladatot egy módosított GAN (Generative adversarial network) modellel végezték, melyben a szokásos zaj helyett, két bemenete van mind a generátornak, mind a diszkriminátornak, és melyen keresztül mindkét modell megkapja az eredeti képet is, így sikeres tanítás esetén lehetősége lesz a generátornak a bemenő képkockára illeszkedő mélység térképet generálni.

Az architektúrát tekintve a továbbiakban rétegekbe szervezett 3D konvolúciós blokkokat használtak, majd az eredményt 2 Dense réteggel osztályozzák.

A nehézséget, az ilyen jellegű megoldásoknál az okozza általában, hogy mivel képi adathalmazokról van szó, azok mérete és kezelése rendkívül költséges. Ez annak az eredménye, hogy a képek/videók sokkal több információt is hordoznak, mint amire szükség van a fordításhoz, és egyáltalán nem triviális olyan feldolgozó algoritmusokat találni, melyek optimálisak mind a tárol információ, mind az adatok tömörítése terén. A problémát fokozza, hogy a finom gesztusok, mint az arc mimikája, szemöldök mozgása, mivel minimális mozgással jár, megkívánja a nagy felbontású, jó minőségű képeket. Továbbá az is szembetűnő, hogy a kutatások jelentős része, kevés kategória közül képes gesztusokat azonosítani, ehhez persze a szűkös adathalmazok is hozzájárulnak, de az sem zárható ki, hogy a nagyobb kategória számhoz szükséges megnövekedett adathalmaz, és tanítási idők, korlátozzák a fejlesztőket.

## Segédeszközt használó megoldások

Segédeszköz használata a gesztusok felismerésére egy hatalmas előnnyel jár a hagyományos képfeldolgozást alkalmazó módszerekkel szemben, itt a gesztusok kategorizálásához különböző szenzorok alkalmazásával rendelkezésre áll valós térbeli információ a kezekről, valamint a mozgásokról. Ez hatalmas előnyt jelenthet a pontosság terén hiszen, kamerakép használatakor előfordulhat, hogy olyan szögben kerül rögzítésre a jelelő alany, amiből nem megállapítható a gesztus. Ez persze nem áll fent egy hétköznapi frontális kommunikáció, vagy az eszközeinkkel való interakció során, hisz ekkor egyéb iránt is a kamera felé orientálja magát a jelelő, de a segédeszközbe rejtett szenzorok minden esetben egy biztosabb, és pontosabb képed adnak a gesztusokról.

### Kesztyű alapú kutatások

Az egyik jelentős irány a kesztyű használata. Irodalma jelentős, néhány évente mindig érkezik egy-egy nagyobb áttörés a területen. A [korai prototípusok](https://ieeexplore.ieee.org/document/7785276) az egyszerűbb mozdulatok, például az ábécé betűinek és a számjegyeknek a felismerését célozzák általában, míg a [későbbi rendszerek](https://link.springer.com/article/10.1007/s40820-022-00887-5) bonyolultabb jelnyelvi struktúrákat és kifejezéseket is képesek kezelni.

A kesztyű valójában több szenzor összehangolt méréseit használja fel a gesztusok azonosítására. A hüvelykujj kivételével a maradék 4 ujj mozgatásának szabadságfokai limitáltak, elsődleges mozgási tartományuk a tenyér felé hajlítás. A dőlésszögek meghatározásának leggyakrabban használt érzékelője a hajlítás érzékelő szenzor, melynek ellenállása arányosan növekszik az ujjak mozgása során. Gyakran szerelnek továbbá gyorsulásmérőket az ujjak végére, valamint a tenyérre is. Nyomás érzékelőket az ujjbegyekre, valamint az egyik megoldás még a Hall effektust kihasználó távolság szenzor adatait is feldolgozza. Nyilvánvaló probléma viszont, hogy jeleléskor szükséges az eszközzel rendelkezni, ezáltal a természetességéből könnyen veszíthet a kommunikáció, valamint a minden kísérletben változó hardvereknek köszönhetően speciális adathalmazra van szükség, ami nem feltétlen robosztus.

### SignRing: Amerikai jelnyelv-felismerés IMU szenzorral ellátott gyűrűkkel

Egy újabb megközelítés a gyűrűbe helyezett inerciális mérőegység (IMU) használata. A hivatkozott kutatásban a két mutatóujjon elhelyezve 6 szabadsági fokos adatokból rekonstruálták a jelelt gesztusokat. Ez egy rendkívül friss, de ígéretesnek tűnő kutatás, mivel egy gyűrű viselete nem akadályozza az embereket a hétköznapokban, és javaslatot tesznek több szögből felvett videó esetén tanító adatok kinyerésére, mellyel a már meglévő videó alapú adathalmazokat is hasznosítani lehetne, viszont az általános felhasználhatósága egyelőre nem bizonyított. Ennek oka hogy mint sok más jelnyelvhez hasonlóan, az amerikai változatban is, nem csak a kezek hordoznak információt. pl.: Azok helyzete az archoz képest jelentősen módosíthatja a gesztus jelentését. Mivel az arc szenzorokkal való monitorozása, végképp egy természet ellenes lehetőség, esetleg előzetes kalibrációra, vagy kontextusfüggő javításra van szükség a fordításhoz. Ezen felül az arc mimikája nélkül elveszítjük a jövőbeli lehetőséget a „gloss” teljes értékű fordítására. Így az ilyen módszereket nem tartom alkalmasnak univerzális megoldásként a jelnyelv fordítására.

## A dolgozat célja

A szövegbevitelhez szükséges jelnyelv-felismerő mesterséges intelligencia, bár a hivatkozott kutatásokból látszik, hogy sok fejlődést, és nagy potenciált rejt magában, még messze elmarad a hangról szöveggé átalakítás mögött, vagy akár a gesztus alapú gépelés mögött is. A dolgozat ehhez az aktívan kutatott területhez járul hozzá, mind az ujjbetűzés, mind az általános jelnyelv fordítása terén. További célkitűzés, hogy a fejezetben feltárt nehézségekre megoldást adjon:

* Mindennemű segédeszköz használata nélkül, a kéz fiziológiai modelljével dolgozzon.
* Univerzális, könnyen kezelhető adatokhalmazokat lehessen létrehozni a tanításra.
* A már meglévő videó/képi formátumú adatbázisok felhasználhatóak.
* Mindezt valós időkorlátok között.

# Megközelítés

Az előző fejezetben kiemelt nehézségeket egy újszerű képfeldolgozási módszerrel oldom meg. Ennek lényege, hogy 3 csatornás képekből előállítható, a képen látható kéz ízületeinek 3 dimenziós modellje. A módszer alkalmazható a test többi részén elhelyezkedő jellegzetes pontok beazonosítására, beleértve a fejet, és az arcot. A megfelelő pontok koordinátáit felhasználva kialakított modell alkalmas lehet a kitűzött összes cél megvalósítására. Konvertálhatóak a meglévő képi adatbázisok, a modellek futtatására felhasznált információ univerzális, ha előre megállapodunk mely jellegzetes pontokat használjuk fel, valamint rendkívüli mértékben tömöríthetné a képi adatokat.

A dolgozatban pózbecslés alapon működő modelleken kísérletezve vizsgálom, hogy az adatok ilyen formában való kódolása alkalmas-e a jelnyelv általános fordítására, illetve, hogy a jelenleg elérhető póz felismerő algoritmusok milyen pontosságú megoldásokat eredményeznek.

## Póz felismerés

A póz detektálás a számítógépes látás egyik kulcsfontosságú területe, amely az emberi test alakjának és tartásának felismerésére és nyomon követésére összpontosít. Ezen technológia segítségével a gépek képesek azonosítani és elemzni az emberi test különböző részeit, mint például a kezek, lábak, fej és a test főbb ízületei.

A működése lényegében képfeldolgozási és mély tanulási technikákra épül. A mély neurális hálózatok, különösen a konvolúciós neurális hálózatok (CNN), forradalmasították ezt a területet, lehetővé téve a nagy pontosságú és valós idejű póz detektálást. A hálózatok képesek "tanulni" az emberi anatómia jellemzőiről és azok változatosságáról nagy adathalmazokon keresztül, így képesek azonosítani a testrészeket és azok relatív elhelyezkedését.

Az elmúlt években jelentős előre lepések történtek ezen a területen. A modellarchitektúrák és az adatmegoldási technikák fejlődésével a póz detektálási rendszerek gyorsabbak és pontosabbak lettek. Ezen kívül az átfogó adathalmazok elérhetősége és az erősebb számítási kapacitások is hozzájárultak a terület fejlődéséhez.

### Open pose

Az OpenPose kiemelkedik a póz detektálás világában, mint az első nyílt forráskódú, valós idejű 2D test-, kéz- és arc-póz detektálási rendszer. A Carnegie Mellon Egyetem által kifejlesztett projekt olyan technológiát hozott el a közönségnek, amely a korábbi megoldásoknál jobban skálázható és sokoldalúbb. Míg számos rendszer létezik a póz detektálásra, az OpenPose különösen azzal emelkedik ki, hogy egyetlen neurális hálózattal képes kezelni a test, kéz és arc pózait, így integráltabb megoldást kínál. Továbbá, a nyílt forráskódú természetének köszönhetően sok kutató és fejlesztő számára elérhetővé vált, ami gyors innovációt és széleskörű alkalmazást tett lehetővé. Ezen funkciók kombinációja teszi az OpenPoset az iparág egyik vezető eszközévé a póz detektálásban.

A bemeneti kép először egy előtanított képfeldolgozó modellen megy keresztül. A következő lépésben egy kétfázisú konvolúciós blokkokból álló modell alkalmazása. Az első fázis végén hibafüggvénnyel kényszerítik a modellt, hogy egy úgynevezett végtag „affinitás” mezőt generálon. Ez a mező minden képpontban egy 2 dimenziós vektort tartalmaz, ha adott pont rajta van a végtagon, akkor értéke a 2 legközelebbi felismerni kívánt kulcspont között feszülő vektor normálva, egyébiránt pedig 0. Természetesen a modell ezt csak megközelíti, de tanítás során egy olyan mezőt használnak „elvárt” adatként, ami az az összes emberre kiszámolt mező átalgolása. A pontossága ellenőrizhető, két kulcspont között, egy integrállal a két pontot összekötő szakasz mentén.

A képen szöveg, Betűtípus, képernyőkép, sor látható

Automatikusan generált leírás

A második iterációban az eredeti bemenet és a kiszámolt mező együttese halad tovább, egy hasonló konvolúciós architektúrába, ami pedig az egyes kulcspontok előfordulási valószínűségét adják. Több magas valószínűséggel rendelkező pozíció is lehetséges, hiszen a képen lévő emberek száma ismeretlen. Ekkor egy feladat maradt hátra, mégpedig az összetartozó kulcspontok azonosítása. A feladatot lépésekben oldja meg, egy gráfelméleti algoritmussal. Két kulcspont, ami között akarunk összeköttetést, vegyük például a vállat és a könyököt, képzeljünk el úgy, mint egy párosgráf csúcsai. (Egymással nem akarjuk különböző emberek vállát összekötni, így megvan a két csúcshalmaz.) Két csúcs közöti él súlya legyen a fenti integrállal kiszámolt érték. Ekkor a feladat enyhíthető egy maximális súlyú párosítás keresésével. Enyhíthető, mert általános esetben szétválogatni a kulcspontokat változó számú emberre NP nehéz feladat.

### Mediapipe holistic

A MediaPipe Holistic azon kevés modell közé tartozik, amely egyesíti a különböző emberi testrészek detektálását, és még tovább megy: képes az egymáshoz képesti térbeli elhelyezkedésük azonosítására is. A Google által kifejlesztett MediaPipe platformon belül a Holistic modell olyan innovatív technológiákat alkalmaz, amelyek lehetővé teszik az arc, kéz és test pózának egyidejű elemzését.

Ami igazán különlegessé teszi a Holistic modellt, az az integrált megközelítés: több modellt kombinál együtt a lehető legpontosabb eredmények eléréséhez. Például külön modell végzi a testpóz, kéz, fej kulcspontjainak meghatározását. Ezek az almodelljei külön-külön is kiemelkedő teljesítményt nyújtanak területükön, de együtt alkalmazva a Holistic modell egy integrált és átfogó képet ad a felhasználó pózáról és interakciójáról környezetével.

Sajnos nem elérhető olyan részletes leírás az algoritmikus hátteréről, mint az OpenPosenak, de teljesítménye magáért beszél. Csupán egy embert képes felismerni egy képkockán, de ez egyáltalán nem korlátozó jelenleg, hiszen a feladat egy időpillanatban csak egy jelelő fordítására terjed ki. Előny viszont, hogy nem csak 2D koordinátákat ad vissza, hanem térbeli fogalommal is rendelkezik az ember pozíciójáról. Ez a kezek, jelnyelv céljával való feldolgozása szempontjából hatalmas előnyt jelent. További pozitív tulajdonság, hogy az könyvtár, és vele együtt a modellek könnyen futtathatóak a különböző platformokon. Elérhető például Androidon, illetve böngészős környezetben is. A valós idejű futás hasonló erőforrás korlátozott környezetekben algoritmikusan támogatott. Ha folyamatos videóból hajt végre póz felismerést, abban az esetben az egyes képkockák között kihasználja a lokalitásokat, vagyis azt a tényt, hogy a végtagok nem tudnak hirtelen túl nagy távot megtenni fizikai korlátaik miatt, így egy kisebb komplexitású követő algoritmust alkalmaz.

Az imént felsorolt előnyök miatt a dolgozat további részében a mediapipe holistic, illetve a kézre dedikált verzióját (mediapipe hands) fogom használni.

# Statikus ujjbetűzés

Két alapvető módszer különíthető el a jelnyelv-felismerés terén: a dinamikus videóból és a statikus képkockákból történő fordítás. Míg a dinamikus videóból történő fordítás lehetővé teszi az egymást követő mozdulatok és azok közti átmenetek azonosítását, ami gyakran kulcsfontosságú a jelentés teljes megértéséhez, a statikus képkockákon alapuló fordítás kihívások elé néz. Ebben az esetben a rendszernek egyetlen pillanatfelvétel alapján kell meghatároznia a mozdulat jelentését, ami korlátozott információt kínál a mozdulat teljes kontextusáról. Ennek ellenére mindkét módszernek megvannak a maga előnyei és alkalmazási területei, amelyekről mélyebben is érdemes beszélni.

A statikus módszer nem véletlen az ujjbetűzés témakörében jön elő, ugyanis a j és z kivételével egyik gesztus sem kíván dinamikus mozdulatot, ez a kettő pedig a lentebb tárgyalt algoritmussal korrigálható. Egy másik probléma is fennáll, ha az 1-9-ig lévő számjegyeket is bevonjuk a kiterjesztett ujjbetűzésbe, ugyanis ezekkel együtt már vannak olyan kéztartások, melyek két kategóriába is tartoznak. Ilyen például a 2 és v, melyek gyakorlatilag identikusok kéztartás szempontjából. A gyakorlatban szövegkörnyezetből megkülönböztethetőek, de ha mégis szükség van izolált környezetben a megkülönböztetésükre, olyankor szám esetében a jelelő tenyerét maga felé fordíthatja.

## Adathalmaz

Mivel a kutatás időpontjában nem állt rendelkezésre statikus póz adathalmaz a feladatra, így először képi adatbázisokat kerestem. Mindegyikben közös volt, hogy a képeket erősen előfeldolgozták, és kis felbontásúvá konvertálták, a legtöbb esetben pedig még a szín információt is elhagyták. Pontos specifikáció nem érhető el, a minimális minőségi követelmények tekintetében, így több, nem csak ASL adathalmazon is validáltam az eredményeket:

1. Sign language MNIST: A klasszikus kézzel írt, 28x28 pixeles számjegyeket tartalmazó MNIST adathalmaz variánsa, amerikai jelnyelv feladatokra. A képek hasonlóan névrokonához fekete fehérek, és hasonló felbontásúak.
2. University of Exter ASL: Két féle verzióban is elérhető. Tartalmaz alacsony felbontású színes képeket, valamint mélység információt is. A képek minősége elég változatos, és majdnem mindegyik tartalmaz artifaktumokat.
3. Sign language for Alphabets Kaggle adathalmaz: Egy magasabb felbontású fekete fehér adathalmaz, mint az első pontban említett.

Nyers formában a képek nem voltak megfelelő minőségűek, a mediapipe hands api számára, így különböző minőség fokozó technikákat alkalmaztam a képekre:

1. Interpolációs algoritmusok (Nearest-neighbour interpoláció Bilineáris Bicubic Lanczos)
2. Deep learning alapú megoldás (A2N, DRLN)
3. A kettő kombinációja.

-research A2N and DRLN

-show results

-saját adathalmaz mp hands, principles

## Modellek

-Tensorflow bevezető, sequential api

-Dense, CNN

## Kiegészítő algoritmusok

-Többségi döntés

-signing detection model (LSTM, DENSE, data collection etc. Proof that it works instead of heuristic)

-llm correction (few shot learning, vs zero shot learning)

## Kiértékelés

-nem képes csak szavakat átadni. A szünetnek nincs konkrét jele, a testbeszédből időzítésből következik.

# Ujjbetűzés szekvenciális bemenetből

Asd

# Hagyomás jelelés

Asd

# Általános jelnyelv fordítása

Asd

# Összegzés

Asd

## Formázási tudnivalók

A dokumentum folyószövegéhez használjuk a **Normál** (angol Word esetén Normal) stílust.

### Címsorok

A fejezetcímek esetén a **Címsor 1-4** (Heading 1-4) stílusokat használjuk.

### Képek

A képhez használjuk a **Kép** stílust.

Képaláírást a képen jobb gombbal kattintva a Képaláírás beszúrása… opcióval adhatunk hozzá, így az automatikusan **Képaláírás** (Caption) stílusú lesz.



.. ábra: Példa képaláírásra

### Kódrészletek

Kódrészletek beillesztése esetén használjuk a **Kód** stílust.

using System;

namespace MyApp

{

class Program

{

static void Main( string[] args )

{

Console.WriteLine( "Szia Világ!" );

}

}

}

### Irodalomjegyzék

Az Irodalomjegyzékben szereplő hivatkozásokat **Irodalomjegyzék sor** stílussal formázzuk, a címüket pedig **Irodalomjegyzék forrás** stílussal emeljük ki.

A szövegbe a hivatkozásokat a Kereszthivatkozás beszúrása (Insert cross-reference) funkcióval helyezzük el (példa egy így beszúrt hivatkozásra: [1]), így azok automatikusan frissülnek a hivatkozások átrendezésekor.

# Utolsó simítások

Miután elkészültünk a dokumentációval, ne felejtsük el a következő lépéseket:

* Kereszthivatkozások frissítése: miután kijelöltük a teljes szöveget (Ctrl+A), nyomjuk meg az F9 billentyűt, és a Word frissíti az összes kereszthivatkozást. Ilyenkor ellenőrizzük, hogy nem jelent-e meg valahol a "Hiba! A könyvjelző nem létezik." szöveg.
* Dokumentum tulajdonságok megadása: a dokumentumhoz tartozó meta adatok kitöltése (szerző, cím, kulcsszavak stb.). Erre való a Dokumentum tulajdonságai panel, mely a Fájl / Információ / Tulajdonságok / Dokumentumpanel megjelenítése úton érhető el.
* Kinézet ellenőrzése PDF-ben: a legjobb teszt a végén, ha PDF-et készítünk a dokumentumból, és azt leellenőrizzük.

Irodalomjegyzék

1. Levendovszky, J., Jereb, L., Elek, Zs., Vesztergombi, Gy.: Adaptive statistical algorithms in network reliability analysis, Performance Evaluation - Elsevier, Vol. 48, 2002, pp. 225-236
2. National Istruments: LabVIEW grafikus fejlesztői környezet leírása, <http://www.ni.com/> (2010. nov.)
3. Fowler, M.: UML Distilled, 3rd edition, ISBN 0-321-19368-7, Addison-Wesley, 2004
4. Wikipedia: Evaluation strategy, <http://en.wikipedia.org/wiki/Evaluation_strategy> (revision 18:11, 31 July 2012)

Függelék