Emberi és számítógépes mozgásfelismerés összefüggéseinek vizsgálata

Assessing connections between human and computer-based motion recognition

SAMU Krisztián PhD[1], HAJNAL Alen PhD[2], FABRINCZIUS Zorán György[3], HUNYADY Botond[4], GÖRCS András[5], SZABÓ András[6]

[1,3,4,5,6]Budapesti Műszaki és Gazdaságtudományi Egyetem, Gépészmérnöki Kar, Mechatronika, Optika és Gépészeti Informatika Tanszék, H-1111 Budapest, Műegyetem rkp. 3., +36 1 463 2602, www.mogi.bme.hu
[2]School of Psychology, The University of Southern Mississippi
118 College Drive, Hattiesburg, MS 39406-0001, USA, +1 601 266 4177, https://www.usm.edu/psychology
[1]samuk@mogi.bme.hu, [2]alen.hajnal@usm.edu, [3]fabri.zorcsi@gmail.com,
[4]hunyady.botond@gmail.com, [5]andras.gorcs2@gmail.com, [6]szaboandras2112@gmail.com

Abstract

Although vision is an integral part of human motion recognition, in itself it's not always sufficient for identifying movements. Our research is based on a human study conducted by Masoner in 2022. While Masoner identified which movements are more easily discernible to humans, we aimed to measure the movements' said distinctiveness using computer based video analysis. Our results underscore the previous study's findings, therefore we can infer how humans recognize movements through our algorithm's inner workings.

Kivonat

Az ember mozgásérzékelésének jelentős része a látáson - vizuális információn alapszik, azonban vannak helyzetek, amikben nem triviális pusztán ennek segítségével a mozgásfelismerés. A munkánk Masoner 2022-ben végzett humán kutatásán alapul. Míg Masoner mozgások ember általi megkülönböztethetőségét vizsgálta, mi ugyanennek az összetéveszthetőségnek a vizsgálatára számítógépes videóelemzési módszert hoztunk létre. Az eredményeink alátámasztják az előzmény kutatás értékeit, így következtethetünk az emberi mozgásfelismerés működésére az algoritmusunk felépítéséből.

Kulcsszavak: gépi látás, látás, mozgásfelismerés, optic flow, pszichológia, videóelemzés

1. Bevezetés

Az emberi érzékelést és ezen belül a vizuális érzékelést régóta kutatják annak érdekében, hogy sikeresen feltárják azokat az elveket, amelyek alapján a legjelentősebb információszerző módszerünk működik. A Gibson [1] által megalkotott modell szerint a környezeti fényviszonyok egy optikai eloszlást hoznak létre, amelynek változása az optic flow. Johannson [2] és Robert [3] két különböző módszerrel vizsgálták és bizonyították, hogy pusztán ezen optic flow megfigyelése elegendő az észleléshez.

A jelen kutatás egy hasonló azonban jelentősen eltérő célkitűzésű dolgozat folytatásaként készült el. Masoner [4] kísérletében ugyanis nem egy a felvételen, szem előtt megjelenő alany mozgását kellett a résztvevőknek felismerniük, hanem belső (egocentrikus) nézetből, vagyis testre erősített kamerával felvett videókon kellett megállapítaniuk a viselő által végzett mozgást. Több kutatás is foglalkozott vele és bizonyította azt a tényt, hogy az emberek esetében a vizuális információ felül tud írni más forrásokat a saját mozgásunk érzékelése során, azonban azt nem jelentették ki önmagában elég-e a felismeréshez [5][6][7]. Masonernél az eredmény alátámasztja a hipotézist, miszerint az emberek a saját mozgásuk felismerésére is képesek pusztán vizuális információ alapján.

A mi feladatunk ezen eredmény validálása videóelemzéssel és a vizuális információ az ember számára releváns részének elkülönítése. Ezen célokat figyelembe véve Masoner három különböző felméréséből az utolsót választottuk alapnak, ugyanis a résztvevők itt hagyatkoztak pusztán vizuális információkra. Az online tesztet kitöltők először egy videót láttak belső vagy külső perspektívából, majd ennek eltűnése után egyszerre két új felvételt a másik perspektívából. A későbbi videók közül az egyiken végzett mozgás egyezett a már

eltűnt videón történővel, míg a másik felvételen ezen mozgás párja volt látható. Ezen párok nem triviális mozgásokból álltak, ami jelen esetben a fő mozgásirány egyezését jelenti pl.: leülés és guggolás, előre haladó szökdelés és kocogás, egy helyben ugrálás és jumping jack. Jelen kutatásunkban azt a vizuális mozgásmintázatot szerettük volna vizsgálni amely alapján a résztvevők a másodszorra megjelenő videók közül kiválasztották melyik mozgás egyezik az elsővel.

2. Módszerek és megoldás

A feladatunk tehát a Masoner kísérletében szereplő mozdulatsorok közötti összefüggések megtalálása volt. A felvételeket egy GoPro kamerával 1080p, 30 FPS minőségben készítették, fejre, mellkasra rögzítve, illetve külső nézetből.

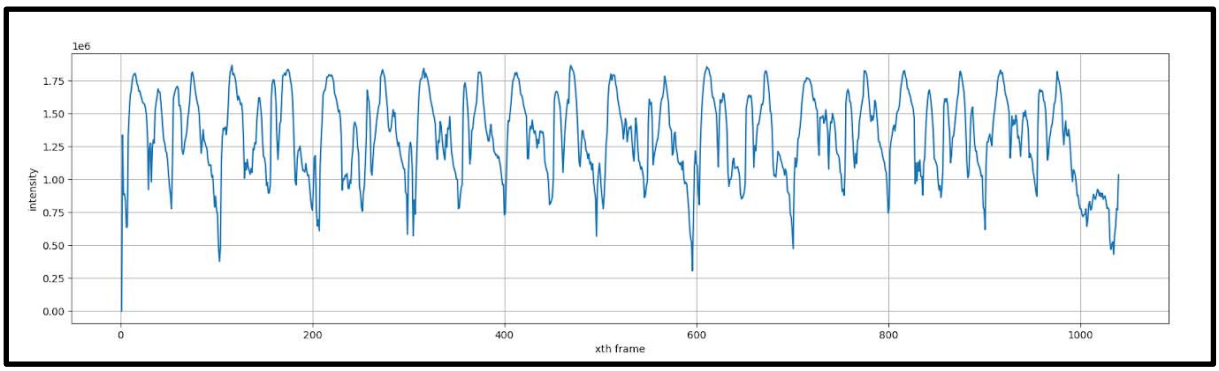
A videók tartalmaztak releváns, illetve irreleváns információkat melyek elkülönítésére az emberi agy képes, a résztvevők is szűrik és tömörítik a látottakat, feltevésünk szerint bizonyos mintázatokra figyelnek. Ehhez hasonlóan a számítógépes elemzés első lépéseként nekünk is tömörítenünk kell a videókat felépítő információhalmazt: csak azt megtartani, ami alapján az alanyok a mozdulatokat felismerik és megkülönböztetik. Ehhez a frame differenciálás módszerét [8] alkalmaztuk.

A mozdulatsorokról kapott felvételeket kézzel felvágtuk mozdulatokra és szürkeárnyalatossá alakítottuk a könnyebb feldolgozhatóság érdekében. Ezután frame differenciálás segítségével kezdtük el elemezni a videókat: minden képkocka felfogható egy-egy mátrixként, ahol a mezőkben az adott pixel árnyalata található, számértékben. A két képkockán az összetartozó mezőket vonjuk ki egymásból majd vesszük az eredmény abszolút értékét. Ahol a különbség 0, tehát fekete pixel szerepel, ott nem történt mozgás. Meghatároztuk a nem fekete pixelek számát, ami a videó mozgalmasságát írja le, ezt neveztük el intenzitásnak.

Az algoritmus egyszerűségéből eredően fellépnek különböző hiányosságok: a mozdulatok irányát nem veszi figyelembe hiszen csak a mozgó pixelek számát vizsgáljuk. Ebből kifolyólag önmagában nem alkalmas mozgások beazonosítására. Itt felvetődhet, hogy miért hanyagoljuk el ezen információkat a videókban. Mint említettük Masoner harmadik kísérletében szereplő párok (guggolás - felülés, szökdelés - kocogás, ugrálás - jumping jack) voltak kísérletünk fókuszában. A mozgás iránya ezen párok esetén megegyezik vagy nagyon hasonló így a közös tulajdonság elhanyagolásával egyszerűsödött a feladatunk és jobban tudtunk a mozgásmintázatbeli különbségekre koncentrálni.

2.1 Alkalmazása:

A bemutatott frame differenciálás algoritmust futtattuk a videókon és ábrázoltuk az intenzitást az idő (képkockák múlása) függvényében (1. ábra).



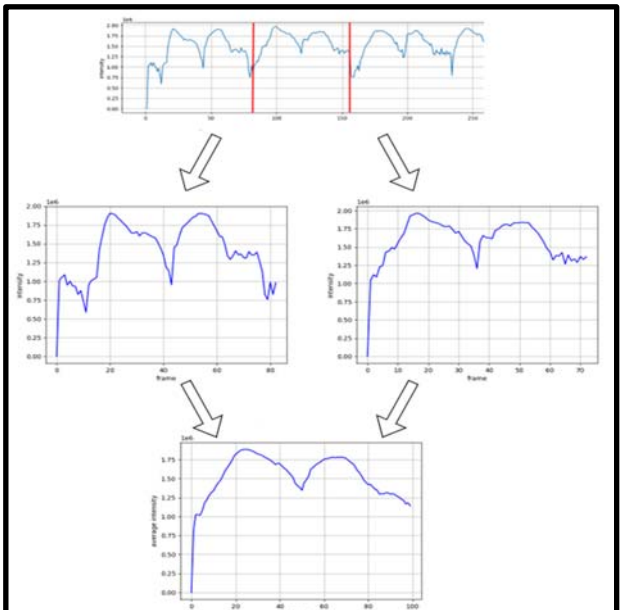
34. ábra: leülés teljes ciklus (függőleges tengelyen intenzitás, vízszintes tengelyen idő (képkockák))

Az így kapott intenzitás-idő függvényeken jól kivehetőek a mozgások karakterisztikái, pl. az ábrán látható leülés-felállás grafikonján a leülésnél keletkező nagyobb intenzitáscsúcs, székben hátra- majd előredőlésnél kisebb csúcsok, majd felállásnál ismét nagyobb csúcs ismétlődik periodikusan.

A felvételekről viszont a mozgások közti apró eltérések, anomáliák és zaj is átadódik a grafikonokra. Ezeket legegyszerűbben a mozdulatsor mozdulatokra vágásával, ezek átlagolásával lehet elsimítani. Az előbb említett leülés-felállás egyenes testtartásból kiindulva, leülésen keresztül, ismételt teljes felegyenesedésig tartó szakaszát értjük egy teljes mozdulatnak. A mozdulatokra vágás korántsem triviális feladat. A felvételek zajossága miatt nem keletkeztek jól beazonosítható, domináns szélsőértékek, ezért a kézi videóvágás módszerét alkalmaztuk.

Továbbá a mozdulatokhoz tartozó adatsorok hossza is különbözött, hiszen a felvételen szereplő alany mindig pár képkockával gyorsabban-lassabban végezte el a mozdulatokat. Ebből kifolyólag normalizálnunk kellett idő szerint, ezt lineáris interpolációval végeztük el, minden mozdulatot 100 adatpont hosszúságúra nyújtottunk.

Az immáron azonos hosszúságú adatsorokat pontonként átlagoltuk, ezzel minden mozgásfajtához kaptunk egy átlagos intenzitás-időfüggvényt, vagyis egy átlagos mozdulatot, amely a mozgást jellemző karakterisztikus mintázat. Ezeket a függvényeket már tudjuk hasonlítani, erre két módszert alkalmaztunk.



35. ábra: átlagos mozgáskarakterisztika függvény meghatározása

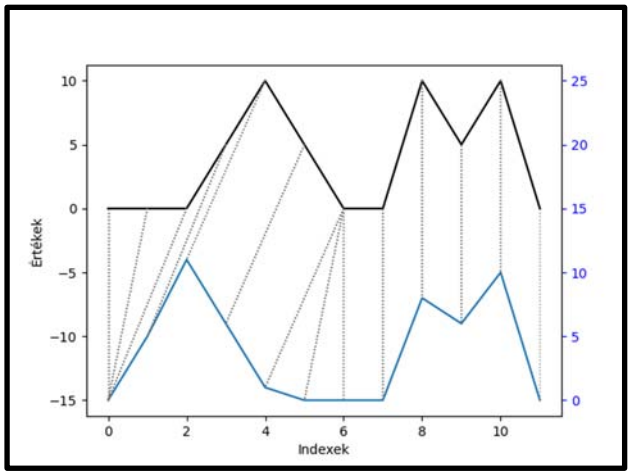
2.2 Lag-korreláció (keresztkorreláció)[9]

A lag-korreláció csúsztatásokat végez az adatsorok között, majd Pearson-féle korrelációt számol közöttük. Ennek a módszernek viszont vannak hibái. Ha guggolásnál az alany egy tized másodperccel többet időz a szokásosnál, mielőtt ismét elkezdene felegyenesedni, akkor a két adatsort semmilyen csúsztatással nem fogjuk tudni jól fedésbe hozni. Hosszabb adatsoroknál az ilyen apró eltérések nagyon gyorsan halmozódnak. Ezek a hibák alkalmazás során is jelentkeztek, emiatt a lag-korrelációt elvetettük. Helyette a DTW algoritmust alkalmaztuk.

2.3 DTW (Dynamic Time Warping – Dinamikus Idővetemítés)[10]

A DTW lokálisan lassítja vagy gyorsítja az időt, ezzel legjobban fedésbe hozva a két adatsort (3. ábra), természetesen ez az időmanipuláció növeli a végső távolságot, vagyis DTW értékét (a hasonlóság mértéke, minél nagyobb, annál jobban különböznek az adatsorok).

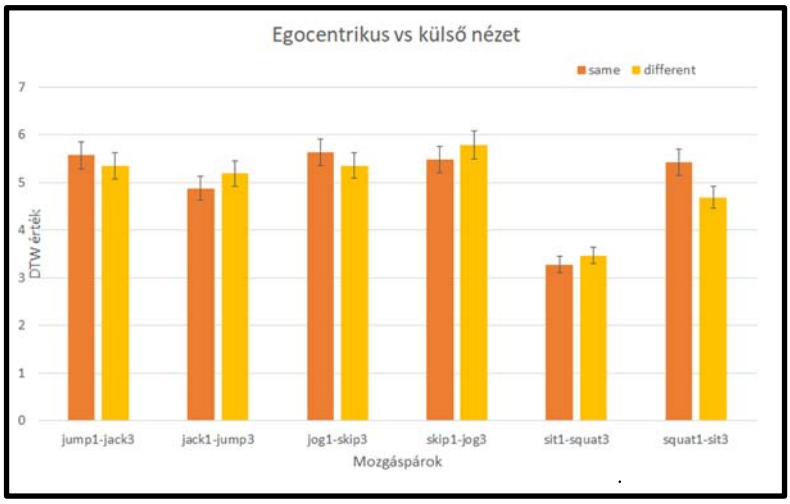
Ezzel megtalálja a jellegre hasonló, de időben eltolt részeit a hasonlított adatsoroknak. Ez mozgások elemzésénél előnyös, hiszen kissé eltérő szüneteket tartalmazó és sebességgel végzett mozdulatokra kis távolságot, DTW értéket kapunk. A DTW értékek normalizált változatával számoltunk, ami a szokásos DTW távolságot elosztja a két hasonlított adatsor hosszának összegével.



36. ábra: DTW működése

3. EREDMÉNYEK:

Tehát az eredményeket a mozgáspárok között mérhető DTW korrelációs szám formájában kaptuk meg. Két videó annál jobban hasonlít egymásra minél kisebb a számolt DTW érték.

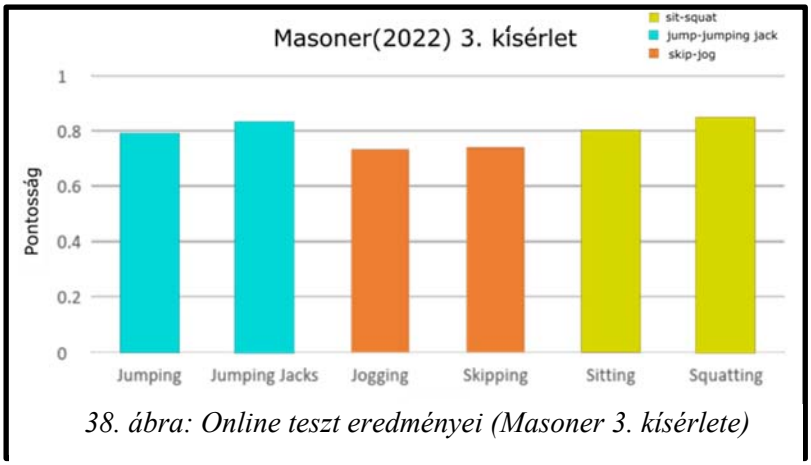


37. ábra: Egocentrikus és külső nézetes videók összehasonlítása

Az eredmények értelmezéséhez, vissza kell tekintenünk a humán kísérletekhez és ugyanazokat az összehasonlításokat kell megismételni, vagyis bizonyos mozgáspárok belső és külső nézetes felvételei között mért DTW értékekre vagyunk kíváncsiak.

A 4. ábrán tulajdonképpen azt látjuk, hogy a belső nézetes videók hogyan viszonyulnak mind saját és mind nem triviális mozgáspárjuk külső nézetes megfelelőjéhez.

Az 5. ábrán pedig az eredeti kísérletek eredményeit látjuk: az online kérdőívet kitöltő résztvevők képesek voltak a nézeteket összekapcsolni, mert a helyes találatok aránya lényegesen nagyobb mint 50 % tehát az eredmények nem csupán szerencse miatt születtek.



38. ábra: Online teszt eredményei (Masoner 3. kísérlete)

Az emberek a leülés-guggolás páros tagjait azonosították a legnagyobb pontossággal, ezért ezen párosításnál vizsgáljuk meg algoritmusunk eredményeit. Az értékek a diagramon láthatóak (4. ábra). Innen leolvashatjuk, hogy a leülés belső nézetes videója valóban a saját külső nézetes megfelelőjére hasonlít jobban, viszont a guggolás belső nézetből jobban hasonlít a leülés külső nézetére, mint a sajátjára.

Ennek magyarázatához finomítanunk kell a korábban bemutatott metódust. Az online kísérletekben résztvevőknek először egy vagy belső vagy külső nézetből felvett videót mutattak ezután a másik perspektívából kellett eldönteni a mozgáspár kettő tagja közül, hogy melyik volt az eredeti videón látható. Emiatt feltételezhetjük, hogy nem csak az számít, mennyire hasonlít két videó egymásra, hanem az is, hogy a videópár melyik tagja mutat nagyobb hasonlóságot a másik perspektívával. Ezeket a tényezőket a $DTW_{compound}$ értékkel vettük figyelembe.

$$DTW_{compound} = DTW_{same} + (DTW_{same} - DTW_{different})$$

Ahol:

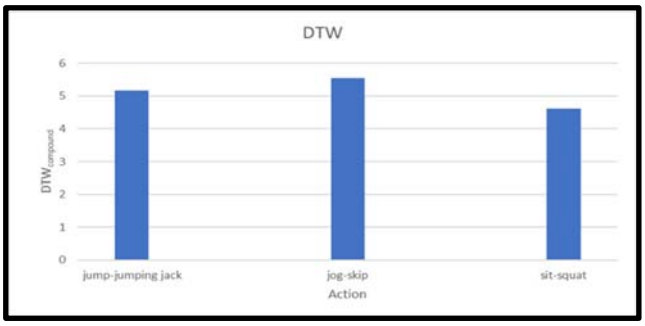
- DTW_{same}: az eredeti mozgás külső és belső nézetes felvételének hasonlósága
- $DTW_{different}$: az eredeti mozgás egyik és mozgáspárja másik perspektívából felvett videójának hasonlósága
- $DTW_{compound}$ a fenti egyenletből kapott mérőszám

Így a $DTW_{compound}$ változó magába foglalja azt, hogy az eredeti mozgás két perspektívájának mekkora a hasonlósága és kiegészíti azzal

39. ábra: Masoner 3. kísérletében elért pontosság

mennyivel nagyobb ez, mint az eredeti mozgás és a mozgáspárja között ugyanezen érték. A kiegészítés az online kísérletben történő két videó közötti választásból ered.

Ezzel a célunk az volt, hogy bebizonyítsuk, hogy az emberek döntésében befolyásoló tényező a pixel intenzitás időbeli változása. A résztvevők pontossága a videópárok azonosításában a 6. ábrán látható. Az ezek között mérhető $DTW_{coumpound}$ hasonlósági értékek átlaga a 7. ábráról olvasható le. A kettő grafikont összehasonlítva megfigyelhetjük, minél pontosabban beazonosítható az algoritmusunk szerint egy mozgáspár annál nagyobb pontossággal tudták az emberek is felismerni a videókat. Ez azt jelenti, hogy a videók hasonlósága korrelációt mutat a humán kísérlet eredményeivel.



40. ábra: DTWcompound értékek átlaga a kiemelt mozgáspároknál

4. Diszkusszió

Az eredményeink fontos bizonyítékkal szolgálnak a humán észlelés megismeréshez. Ugyanis statisztikai módszerekkel bizonyítják be a Masoner (2022) kísérlet eredményeit. A módszerünk emberi tevékenységek összehasonlítására azok felismerésére nélkül is alkalmazható, ami fontos első lépésnek számít a gépi látás szakterületén.

A kutatást számos irányba lehetne folytatni, akár mesterséges intelligenciát vagy kifinomultabb videóelemzési módszert alkalmazva. Esetleg megismételni több és jobb minőségű mozgásvideóval melyekben különböző emberek szerepelnek.

Irodalmi hivatkozások

[1] J. J. Gibson, The ecological approach to visual perception. Boston: Houghton Mifflin,PP 332, 1979.
[2] Johansson, G., Visual perception of biological motion and a model for its analysis.Perception & Psychophysics, 1973, 14(2), PP 201-211
[3]Sophia Robert, Leslie G. Ungerleider, and Maryam Vaziri-Pashkam, Disentangling Object Category Representations Driven by Dynamic and Static Visual Input, The Journal of Neuroscience, 2023, 43(4), PP 621–634
[4] Masoner, H., Does Optic Flow Provide Information about Actions? Doctoral dissertation. 2022, https://aquila.usm.edu/dissertations/1972. (2023.02.28.)
[5] Lishman, J. R., & Lee, D. N., The autonomy of visual kinaesthesis. Perception, 1973, 2(3), PP 287-294
[6] Rogers, B. J., Optic flow: Perceiving and acting in a 3-D world, i-Perception, 2021, 12(1), PP 1–25, https://doi.org/10.1177/2041669520987257) (2023.02.28.)
[7] Rogers, B. Young, K. Tootell, S., Optic flow and the maintenance of balance., Journal of Vision, 2007, 7(9), PP 1023, 1023a, http://journalofvision.org/7/9/1023/ (2023.02.28.)
[8] Cross Correlation, United States Naval Academy, https://docs.opencv.org/3.4/d4/dee/tutorial_optical_flow.html (2023.02.28.)
[9] Optical Flow, OpenCV, https://www.usna.edu/Users/oceano/pguth/md_help/html/time0alq.htm (2023.02.28.)
[10] Sakoe, H., & Chiba, S., Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(1), PP 43-49.