UO‘K 004.8

MATNDAN O‘ZBEK IMO-ISHORA TILIGA TARJIMA QILISH UCHUN MEDIAPIPE ASOSIDA MA’LUMOTLAR TO‘PLAMINI TAYYORLASH TEXNOLOGIYASI

Jurayev D.B.[1]

[1] Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Toshkent, O‘zbekiston

dilsamtuit@gmail.com

Annotatsiya. Maqolada o‘zbek imo-ishora tilini avtomatik tarzda matn va nutqqa tarjima qilish uchun MediaPipe texnologiyasiga asoslangan multimodal ma’lumotlar to‘plamini shakllantirish masalasi ko‘rib chiqiladi. Ishda imo-ishora tilining ko‘p qirrali xususiyatlari, qo‘l harakatlari, yuz ifodalari, tana pozitsiyasi va ko‘z yo‘nalishlari sababli dataset tayyorlashning texnik, lingvistik va etik jihatlari tahlil qilinadi. MediaPipe orqali qo‘l, yuz va tana nuqtalari aniqlab olinib, ular asosida GRU, LSTM va BiLSTM modellarida tajribalar o‘tkaziladi. Qiyosiy tahlil natijalari GRU modelining samaradorligi va resurs tejamkorligini, LSTM va BiLSTMning esa murakkab ketma-ketliklarni qayta ishlashdagi afzalliklarini ko‘rsatadi. Shunday qilib, tadqiqotda shakllantirilgan multimodal dataset real vaqt rejimida ishlaydigan imo-ishora tanish tizimlarini ishlab chiqish uchun ishonchli asos bo‘lib xizmat qiladi.

Kalit so‘zlar: imo-ishora tili, MediaPipe, multimodal dataset, RNN (GRU, LSTM, BiLSTM), kompyuter ko‘rish, tabiiy tilni qayta ishlash (NLP), sun’iy intellekt.

1 KIRISH

Imo-ishora tili eshitish yoki nutq qobiliyati cheklangan shaxslar uchun asosiy kommunikatsiya vositasi bo‘lib, u qo‘l va barmoqlar harakatlari bilan bir qatorda tana holati, yuz ifodalari va ko‘z ishoralari kabi noverbal elementlarni qamrab oladi. So‘nggi yillarda sun’iy intellekt va tabiiy tilni qayta ishlash sohalaridagi erishilgan yutuqlar imo-ishora harakatlarini matn yoki nutqqa, aksincha matn yoki nutqni imo-ishora harakatlariga avtomatik tarjima qilishga qiziqishni sezilarli darajada kuchayishiga olib keldi. Bunday tizimlarni yaratish uchun yuqori sifatli, kontekstga boy va multimodal o‘quv ma’lumotlar to‘plami (dataset) hal qiluvchi ahamiyat kasb etadi.

Mavjud tizimlar ko‘pincha cheklangan hajmdagi ma’lumotlarga asoslanganligi sababli, ularning amaliy qo‘llanishi chegaralangan. Shuningdek, har bir imo-ishora tili mintaqaviy lingvistik xususiyatlarga ega bo‘lgani uchun universal model yaratish murakkab. Imo-ishoralar vizual shaklda ifodalanadi, shu bois datasetlar video yozuvlar, uch o‘lchamli (3D) sensor ma’lumotlari va ularning yorliqlangan shakllaridan tashkil topishi lozim.

Tadqiqotlarda uchraydigan asosiy muammolar aniqlik, kontekstga bog‘liqlik, o‘xshash imo-ishoralarni farqlash hamda sintaktik struktura bilan bog‘liq. Ularni bartaraf etish uchun inson tomonidan yorliqlangan, ekspertlar tasdiqlagan keng qamrovli korpuslar va ba’zan real vaqt rejimida ishlovchi algoritmlar zarur. Shu jarayon texnik yechimlar bilan birga ilmiy va etik jihatdan ham puxta rejalashtirilishi kerak. Ayniqsa, ma’lumotlarni yig‘ishda shaffoflik, ishtirokchilarning roziligi va maxfiyligini ta’minlash muhimdir. Imo-ishora tilidan foydalanuvchi shaxslar ko‘pincha ijtimoiy jihatdan zaif guruh vakillari bo‘lgani bois, ularning roziligisiz video yoki boshqa turdagi ma’lumotlarni yig‘ish inson huquqlarining buzilishi hisoblanadi.

So‘nggi yillarda imo-ishora tilini avtomatik tarjima qilish bo‘yicha tadqiqotlar jadal rivojlanib, nafaqat lingvistik, balki keng ijtimoiy-texnologik ahamiyatga ega yo‘nalishga aylandi. Ushbu yutuqlar jamiyatda inklyuzivlikni oshirish, eshitish imkoniyati cheklangan shaxslarning faol ishtirokini ta’minlash hamda, yangi texnologiyalar orqali teng imkoniyatlar yaratishga xizmat qilmoqda. Bunday tadqiqotlar asosan tabiiy tilni qayta ishlash (NLP), kompyuter ko‘rish, mashinali o‘qitish va kompyuter lingvistikasi sohalarining kesishgan nuqtasida olib borilmoqda. Ularni ikki asosiy yo‘nalishga bo‘lish mumkin:

- imo-ishora tanib olish - imo-ishora harakatlarini matn yoki nutqqa aylantirish;
- imo-ishoraga tarjima - matn yoki nutq asosida imo-ishora harakatlarini yaratish.

Ushbu maqolada imo-ishora tilini matn va nutqqa tarjima qilish uchun zarur bo‘lgan ma’lumotlar to‘plamlarini shakllantirish, ularning tarkibi, yig‘ish usullari, etik talablari hamda texnik yechimlari tahlil

qilinadi. Tadqiqotimiz esa o‘zbek imo-ishora tili uchun MediaPipe asosida multimodal o‘quv ma’lumotlar to‘plamini yaratishga qaratilgan.

2 ASOSIY QISM

Imo-ishora tilini avtomatik tarjima qilish texnologiyasi eshituvchi shaxslar va eshitish imkoniyati cheklangan insonlar o‘rtasida samarali kommunikatsiyani ta’minlashga xizmat qiladi hamda ularning jamiyatga integratsiyasini qo‘llab-quvvatlaydi. Bunday tizimlarda muhim elementlardan biri bo‘lgan dinamik imo-ishoralar esa vaqt davomida o‘zgarib boruvchi harakatlar ketma-ketligiga asoslangan belgilar bo‘lib, ularni tahlil qilish murakkab jarayon sanalsa-da, real muloqot jarayoniga yanada yaqin turadi. Ushbu turdagi imo-ishoralarni aniqlash bo‘yicha yondashuvlar odatda ikki asosiy metodologiyaga ajratiladi: harakat traektoriyasi va qo‘l shakli asosidagi yondashuvlar, hamda video ketma-ketligiga asoslangan yondashuvlar [1].

Dinamik imo-ishoralar bu vaqt davomida o‘zgarib boradigan, harakatlar ketma-ketligiga asoslangan belgilar bo‘lib, ularning tahlili murakkab, biroq real muloqotga yaqinroq hisoblanadi. Ushbu turdagi imo-ishoralarni aniqlashga qaratilgan yondashuvlar odatda ikki asosiy metodologiyaga harakat traektoriyasi va qo‘l shakliga asoslangan yondashuvlar hamda video ketma-ketliklariga asoslangan yondashuvlarga bo‘linadi.

Harakat traektoriyasi va qo‘l shakli asosidagi yondashuvda imo-ishorani ifodalovchi harakatning yo‘nalishi, tezligi va shakli aniqlanadi. Ko‘p hollarda bu usullar sensorga asoslangan ma’lumotlardan - masalan, giroskop, akselerometr yoki chuqurlik kameralari orqali olingan signallardan foydalanadi. Kim va boshqalar [2] tomonidan taklif etilgan CNN modeli barmoqdagi imo-ishorasini aniqlashda qo‘l shakli xususiyatlariga tayangan bo‘lib, murakkab strukturalarni oddiy belgilar yordamida samarali ajrata olgan. Biroq, bu model harakat kontekstidan mustaqil ishlaydi, shu sababli murakkab imo-ishoralarda samaradorlik kamayib ketadi.

Mohandes va boshqalar [3] esa LSTM arxitekturasi yordamida faqat harakat traektoriyasini tahlil qilish orqali imo-ishoralarni farqlashga uringan. Bu kabi yondashuvlarda harakat davomiyligi va yo‘nalishi asosiy belgilar sifatida ko‘rilgan, biroq barmoq shakli va yuz ifodalari hisobga olinmagan.

Ding va Martinez [4] barmoqlarning 3D shaklini va ularning harakatlarini fazoviy traektoriya sifatida modellashtirib, har bir imo-ishoraning fazoviy tavsifini aniqlashga erishgan. Dogra va boshqalar [5] esa bir nechta sensorlardan (Leap Motion va Kinect) foydalanib, multisensorli imo-ishora tanib olish tizimini yaratgan. Biroq, ushbu yondashuvlar maxsus uskunalarni talab qilgani sababli amaliyotda keng qo‘llanilmaydi.

Zadgorban va Nahviy [6] Fors imo-ishora tilida so‘z chegaralarini harakat va qo‘l shakli xususiyatlari asosida aniqlagan. Ularning yondashuvi uzluksiz imo-ishoralarni alohida birliklarga ajratishga xizmat qiladi. Bu harakat traektoriyasi va shakliga asoslangan yondashuvlar faqatgina ma’lum kontekstdagi oddiy imo-ishoralar uchun yuqori aniqlik beradi, biroq ularni murakkab va uzluksiz imo-ishora ketma-ketliklariga qo‘llash imkoniyati cheklangan.

Video ketma-ketliklariga asoslangan yondashuvlar esa zamonaviy kompyuter ko‘rish texnologiyalariga tayanadi. Masalan, MediaPipe real vaqtda qo‘l, tana va yuzning asosiy nuqtalarini aniqlash orqali yuqori samaradorlikni ta’minlaydi. Bu yondashuv, an’anaviy CNN va RNN kombinatsiyalaridan farqli o‘laroq, maxsus uskunalarni talab qilmasdan, oddiy qurilmalarda ham ishlashi bilan ajralib turadi [12]. Masalan, Zhang va boshqalar [7] CNN modelidan foydalangan holda video ketma-ketliklarini tahlil qilgan bo‘lsa, MediaPipe engil algoritmlari orqali shunga o‘xshash natijalarni oddiy mobil qurilmalarda ham ta’minlay oladi.

Kishore va boshqalar [8] hind imo-ishora tilini CNN modeli yordamida tahlil qilgan hamda video selfi rejimida yozilgan tasvirlar asosida modelni o‘rgatgan. Manikanta [9] esa harakat va shakl xususiyatlarini birlashtirgan holda, dinamik imo-ishora videolarni noravshan tasniflash orqali tahlil qilgan. Boshqa tadqiqotlarda esa fazoviy va vaqtinchalik xususiyatlarni integratsiyalash orqali chuqur o‘qitish samaradorligini oshirishni taklif qilgan [10,11]. Bunday yondashuvlarda CNN yordamida qo‘l, tana va yuzning vizual xususiyatlari aniqlanadi, vaqtinchalik ketma-ketliklar esa RNN orqali tahlil qilinadi. Fazoviy-vaqtinchalikni modellashtirish imo-ishora ketma-ketliklarining semantik mazmunini tushunishga yordam beradi.

Statik imo-ishoralar qo‘lning harakatsiz, aniq bir holatdagi pozitsiyasini anglatadi va odatda tasvirlar asosida tahlil qilinadi. Bunday usullarda vizual xususiyatlarni tavsiflovchi texnikalar, jumladan HOG (Histogram of Oriented Gradients) va Zernike Invariant Moments (ZIM) keng qo‘llaniladi [11]. Allam va Hemayed [12] arab imo-ishora tilining alifbosini tasvirlar orqali avtomatik tarzda nutqqa aylantiruvchi tizim yaratgan. Altagafi va boshqalar [13] esa CNN asosida RGB tasvirlar yordamida 28 ta belgi turini 92,9% aniqlikda tanib olishga erishgan. Bunday yondashuvlar statik imo-ishoralarni, ya’ni qisqa muddatli

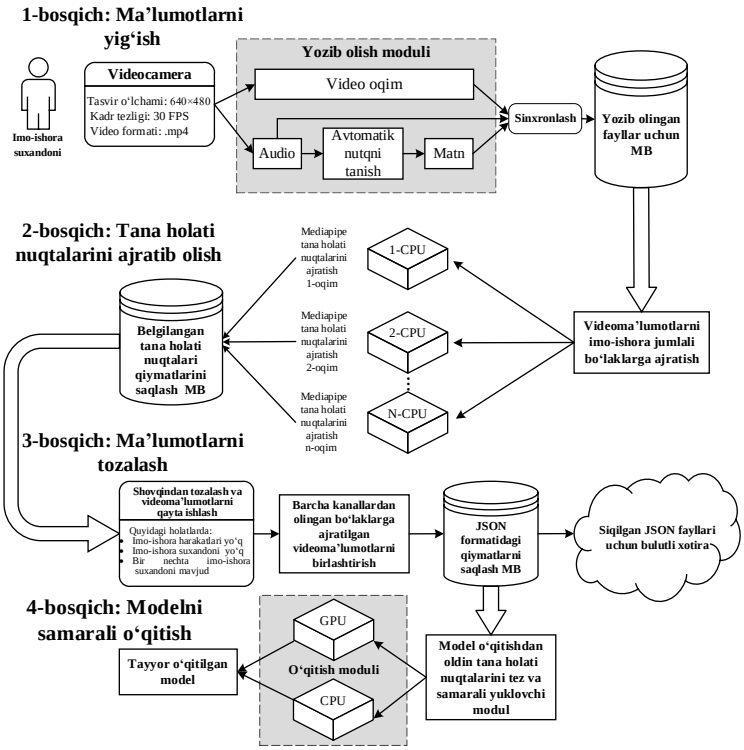
yoki bir holatdagi belgilarni tahlil qilishda samarali bo‘lsa-da, ularni murakkab va uzluksiz imo-ishora ketma-ketliklariga qo‘llash imkoniyati cheklangan [18].

Yuqoridagi ilmiy manbalar tahlilidan kelib chiqib aytish mumkinki, nutq yoki matndan imo-ishora tiliga tarjima qilish jarayonida dinamik imo-ishoralar muhim o‘rin tutadi. Chunki ular vaqt davomida o‘zgarib boruvchi harakatlar ketma-ketligiga asoslanib, real muloqotning semantik mazmunini yanada to‘liq ifodalash imkonini beradi. Shu bois samarali tarjima tizimlarini ishlab chiqishda dinamik imo-ishoralarni aks ettiruvchi ma’lumotlar to‘plamini yaratish va ularni tahlil qilish eng asosiy yo‘nalishlardan biri hisoblanadi.

Ushbu tadqiqotning asosiy maqsadi nutqdan imo-ishoraga tarjima qilish imkonini beruvchi tizimni yaratish uchun zarur bo‘lgan yangi, video asosidagi multimodal ma’lumotlar to‘plamini ishlab chiqishdan iborat. Shuningdek, ushbu to‘plam asosida uzluksiz harakatlarga asoslangan imo-ishora tanib olish modeli yaratiladi. Bu yondashuv statik imo-ishoralarni tahlil qilishdagi mavjud cheklovlarni bartaraf etishga qaratilgan muhim qadam hisoblanadi.

Tadqiqot metodologiyasi to‘rt bosqichli jarayon asosida qurilgan bo‘lib, uning umumiy arxitekturasi quyidagi 1-rasmda keltirib o‘tilgan.

Har bir bosqich tizimning ma’lumot yig‘ish bosqichidan tortib, modelni o‘qitishgacha bo‘lgan to‘liq jarayonni o‘z ichiga oladi va ular bir-biri bilan uzviy bog‘liqdir.



1-rasm. Imo-ishora ma’lumotlarni yig‘ish va o‘qitish moduli arxitekturasi

1-bosqich – Ma’lumotlarni yig‘ish. Ushbu bosqichda multimodal ma’lumotlar to‘plami shakllantirildi. To‘plam tarkibiga kundalik muloqotda eng ko‘p uchraydigan 200 ta asosiy imo-ishora iboralarini (“Salom”, “Rahmat”, “Ha”, “Yo‘q”, “Kechirasiz” va boshqalar) o‘z ichiga olgan. Yuqoridagi ibora uchun jami 1800 ta video yozib olingan. Ushbu iboralar nutqdan imo-ishoraga tarjima qilish tizimi uchun boshlang‘ich lug‘aviy bazani shakllantirishda muhim rol o‘ynaydi. Videolar yopiq va yoritilishi barqaror muhitda, mobil video yozish qurilmasi yordamida 30 FPS (kadr/sekund) tezlikda yozib olingan. Har bir video va uning transkripti bir xil davomiylikka keltirilib, barcha videokadrlarni standart o‘lchamga keltirish orqali modelni o‘qitishda izchillik va yaxlitlik ta’minlangan. Yig‘ilgan videolar va ularning transkriptlari ma’lumotlar bazasida tizimli ravishda saqlanadi.

2-bosqich – Tana holati nuqtalarini ajratib olish. Imo-ishora tili, asosan, qo‘llarning harakati va tana pozasini aniqlashga asoslanadi. Harakatdagi imo-ishoralarni tahlil qilishda quyidagi muammolar yuzaga kelishi mumkin:

- qo‘l joylashuvini aniqlash, qo‘lning fazodagi o‘rni va tana bilan bog‘liq joylashuvi;
- qo‘l shaklini aniqlash, barmoqlarning egilishi va kaftning holati;
- qo‘l yo‘nalishini belgilash, qo‘lning qaysi tomonga yo‘nalganligi.

Ushbu muammolarga yechim sifatida Google tomonidan ishlab chiqilgan MediaPipe tizimidan foydalanilgan [14,16]. MediaPipe har bir video kadr (frame) uchun qo‘l, yuz va tananing asosiy nuqtalarini (keypoints) X, Y va Z koordinatalarida aniqlaydi. Ushbu jarayon bir nechta markaziy protsessor (CPU)da parallel ravishda amalga oshiriladi. Natijada, har bir video uchun vaqt bo‘yicha ketma-ket joylashgan nuqtalar koordinatalarining to‘plami hosil bo‘ladi. Bu ma’lumotlar imo-ishoralarning dinamik harakatlari haqida boy kontekstual axborot beradi.

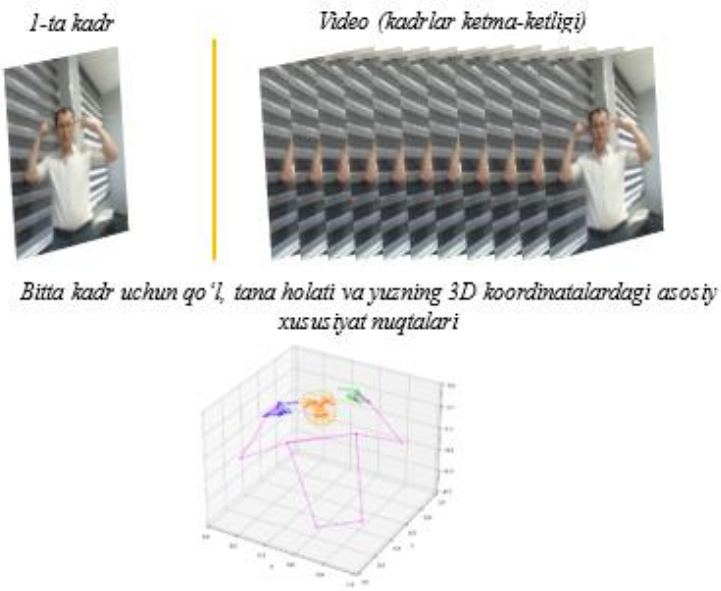
3-bosqich – Ma’lumotlarni tozalash. Dastlabki yig‘ilgan videolarni oldindan qayta ishlash va tozalash talab etiladi. Bu bosqichda sifatsiz, imo-ishorasi noto‘g‘ri bajarilgan yoki bir nechta imo-ishorani o‘z ichiga olgan kadrlar aniqlanib, olib tashlanadi. Tozalashdan so‘ng, qayta ishlangan ma’lumotlar JSON formatida saqlanadi. Bu format ma’lumotlarni tizimli va ixcham shaklda saqlash imkonini beradi hamda keyingi bosqichlarda qayta ishlashni soddalashtiradi.

4-bosqich – Modelni samarali o‘qitish. Tozalangan va qayta ishlangan JSON ma’lumotlari asosida imo-ishora tanib olish modeli o‘qitiladi. Modelni o‘qitish uchun maxsus o‘qitish moduli ishlab chiqilgan bo‘lib, u grafik protsessor (GPU) lar va CPU resurslaridan samarali foydalanadi. Modelning maqsadi – ketma-ket keladigan tana va qo‘l nuqtalari koordinatalari asosida imo-ishora iborasini aniq tasniflashdan iborat. O‘qitilgan model keyinchalik nutqdan imo-ishoraga avtomatik tarjima qiluvchi tizimning markaziy komponenti sifatida xizmat qiladi.

MediaPipe asosida xususiyatlarni ajratish - imo-ishora tili, asosan, qo‘llarning harakati va tana pozasini aniqlashga asoslanadi [15]. Biroq, imo-ishoralarning doimiy harakatda bo‘lishi sababli bir qator muammolar yuzaga keladi. Bu muammolarga quyidagilar kiradi:

- qo‘l joylashuvini aniqlash;
- qo‘l shaklini aniqlash;
- qo‘l yo‘nalishini belgilash.

MediaPipe ushbu muammolarga yechim sifatida qo‘llanilgan. Bu tizim har bir video kadr uchun qo‘llar va tananing asosiy nuqtalarini X, Y va Z koordinatalarida aniqlaydi va 2-rasm berilganidek imo-ishoraning fazodagi holati va harakati haqida to‘liq ma’lumot berib, modelning ishlashini sezilarli darajada yaxshilaydi.

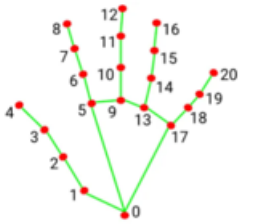


2-rasm. MediaPipe asosida imo-ishora harakatlarini yaratish uchun tana, qo‘l va yuzning uch o‘lchovli fazodagi joylashuvini modellashtirish

Tana holatini aniqlash (pose estimation) texnikasi qo‘lning tanaga nisbatan joylashuvini bashorat qilish va kuzatish uchun ishlatilgan. MediaPipe tizimining chiqish natijasi 3-rasmda ifodalanganidek qo‘llar va tana pozasiga oid asosiy nuqtalar ro‘yxatidir. Har bir qo‘l uchun MediaPipe 21 ta asosiy nuqta (keypoint) ni aniqlaydi. Bu nuqtalar uch o‘lchamli fazoda – X, Y va Z koordinatalarida hisoblanadi. Shuning uchun, qo‘llarga oid jami asosiy nuqtalar soni quyidagicha hisoblanadi:

$$Qo‘l nuqtalar\ soni = 21 \times 3 \times 2 = 126\ ta\ asosiy\ nuqta.$$

Bu qiymatlar ikkala qo‘l uchun ajratib olingan, uch o‘lchovli (3D) koordinatalarga asoslangan asosiy nuqtalardir.



0. Bilak
1. Bosh barmoq (Kaft suyaklari va bilak suyaklari orasidagi bo‘g‘im)
2. Bosh barmoq (Kaft va barmoq orasidagi bo‘g‘im)
3. Bosh barmoq (Bosh barmoqning yagona bo‘g‘imi)
4. Bosh barmoq uchi
5. Ko‘rsatkich barmoq (Kaft va barmoq orasidagi bo‘g‘im)
6. Ko‘rsatkich barmoq (Barmoqlarning o‘rta bo‘g‘imi)
7. Ko‘rsatkich barmoq (Barmoq-larning uchiga yaqin bo‘g‘imi)
8. Ko‘rsatkich barmoq uchi
9. O‘rta barmoq (Kaft va barmoq orasidagi bo‘g‘im)
10. O‘rta barmoq (Barmoqlarning o‘rta bo‘g‘imi)
11. O‘rta barmoq (Barmoqlarning uchiga yaqin bo‘g‘imi)
12. O‘rta barmoq uchi
13. Nomsiz barmoq (Kaft va barmoq orasidagi bo‘g‘im)
14. Nomsiz barmoq (Barmoqlarning o‘rta bo‘g‘imi)
15. Nomsiz barmoq (Barmoq-larning uchiga yaqin bo‘g‘imi)
16. Nomsiz barmoq uchi
17. Kichikkina barmoq (Kaft va barmoq orasidagi bo‘g‘im)
18. Kichikkina barmoq (Barmoqlarning o‘rta bo‘g‘imi)
19. Kichikkina barmoq (Barmoq-larning uchiga yaqin bo‘g‘imi)
20. Kichikkina barmoq uchi

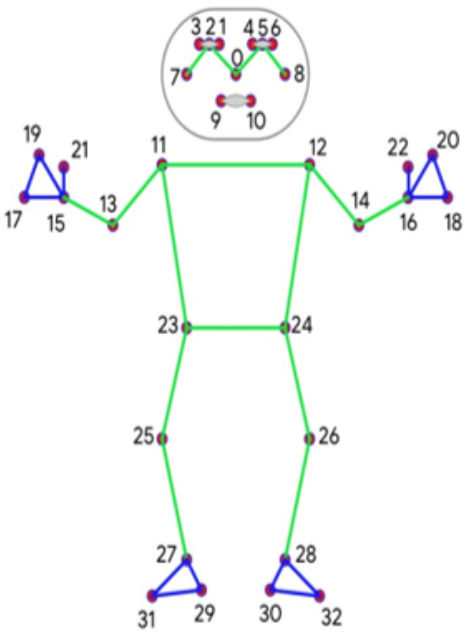
3-rasm. Qo‘lning asosiy nuqtalari va barmoqlar bo‘g‘imlarining indekslanishi (mediapipe asosida)

MediaPipe tizimi tana holatini baholash uchun 4-rasmda berilganidek har bir ramkada 33 ta asosiy nuqtalarni aniqlaydi [16]. Ushbu nuqtalar uch o‘lchovli fazoda – ya’ni X, Y va Z koordinatalarida, shuningdek ko‘rinish qiymati bilan birga hisoblanadi.

Shunday qilib, tana holatini baholash natijasida ajratib olingan umumiy nuqtalar soni quyidagicha hisoblanadi:

$$Pozadagi\ asosiy\ nuqtalar\ soni = 33 \times (3+1) = 132\, ta\ qiymat.$$

Bu qiymatlar har bir ramka uchun 33 ta nuqtaning X, Y, Z koordinatalari va ko‘rinish qiymatini o‘z ichiga oladi.



0. Burun
1. O‘ng ko‘z (ichki qismi)
2. O‘ng ko‘z
3. O‘ng ko‘z (tashqi qismi)
4. Chap ko‘z (ichki qismi)
5. Chap ko‘z
6. Chap ko‘z (tashqi qismi)
7. O‘ng quloq
8. Chap quloq
9. Og‘izning o‘ng burchagi
10. Og‘izning chap burchagi
11. O‘ng yelka
12. Chap yelka
13. O‘ng tirsak
14. Chap tirsak
15. O‘ng bilak
16. Chap bilak
17. O‘ng kichik barmoq (1-bo‘g‘im)
18. Chap kichik barmoq (1-bo‘g‘im)
19. O‘ng ko‘rsatkich barmoq (1-bo‘g‘im)
20. Chap ko‘rsatkich barmoq (1-bo‘g‘im)
21. O‘ng bosh barmoq (2-bo‘g‘im)
22. Chap bosh barmoq (2-bo‘g‘im)
23. O‘ng son (bel bo‘g‘imi)
24. Chap son (bel bo‘g‘imi)
25. O‘ng tizza
26. Chap tizza
27. O‘ng tovon (oyoqqa tutash joyi)
28. Chap tovon
29. O‘ng poshna
30. Chap poshna
31. O‘ng oyoqning oldingi uchi
32. Chap oyoqning oldingi uchi

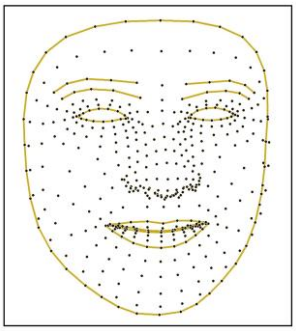
4-rasm. Inson tanasining MediaPipe tana holati (poza)ni aniqlash modeli asosidagi 33 nuqtali skelet tuzilishi va ularning indekslanishi

MediaPipe tizimi yuzni aniqlash jarayonida har bir ramkada 468 ta asosiy nuqtani ajratib oladi (5-rasm). Ushbu nuqtalar yuz, ko‘zlar, lablar va qoshlar atrofida joylashgan bo‘lib, ular orasidagi konturlar chiziqlar bilan, nuqtalar esa nishon belgilar (nuqtalar) orqali ifodalanadi.

Har bir nuqta uch o‘lchamli fazoda – ya’ni X, Y va Z koordinatalari asosida hisoblanadi. Shu sababli, yuzdan ajratib olingan umumiy asosiy nuqtalar soni quyidagicha hisoblanadi:

$$Asosiy\ nuqtalar\ soni = 468 \times 3 = 1404 ta\ qiymat.$$

Bu qiymatlar har bir ramka uchun yuzning uch o‘lchovli joylashuvi haqida batafsil ma’lumot beradi.



5-rasm. MediaPipeda yuz belgilari

Yuz asosiy nuqtalarini hisobga olmagan holda, har bir video kadr (frame) uchun ajratib olingan umumiy asosiy nuqtalar soni quyidagicha hisoblanadi:

$$Qo‘l\ nuqtalari + Poza\ nuqtalari = 126 + 132 = 258 ta\ nuqta.$$

Yuz asosiy nuqtalarini ham hisobga olgan holda, har bir kadr uchun ajratilgan umumiy asosiy nuqtalar soni quyidagicha bo‘ladi:

$$Qo‘l + Poza + Yuz = 126 + 132 + 1404 = 1662 ta\ nuqta.$$

Hisob-kitoblar har bir ramkada ajratib olingan xususiyatlar sonini ifodalaydi va ular keyinchalik mashinali o‘qitish jarayonida kirish ma’lumotlari sifatida qo‘llaniladi. Ushbu operatsiya butun video davomida takrorlanib, har bir kadr (frame) uchun asosiy nuqtalar (keypoints) aniqlanadi. Natijada imo-ishora ma’lumotlar to‘plamidagi barcha videolardan qo‘l, tana va yuzning joylashuvi, shakli hamda harakat yo‘nalishlari qayd etiladi.

Mazkur jarayonda MediaPipe ramkasi muhim ahamiyatga ega bo‘lib, u Face Detection (yuzni aniqlash), Face Mesh (yuz to‘r modeli), Hands (qo‘l harakatlari) va Pose (tana pozitsiyasi) kabi komponentlarni o‘z ichiga oladi[14]. Ushbu komponentlar vaqt ketma-ketligiga asoslangan ma’lumotlarni samarali qayta ishlash imkonini beradi.

Shu bilan birga, imo-ishora tilini avtomatik tanib olishda qator muammolar saqlanib qolmoqda. Jumladan, mavjud tizimlarda imo-ishoralarni aniqlash aniqligi ko‘p hollarda past bo‘lib, imo-harakatlarning murakkabligi va xilma-xilligi ularni to‘g‘ri tahlil qilishda qo‘shimcha qiyinchiliklarni yuzaga keltiradi.

Mazkur muammolarga yechim sifatida ushbu tadqiqotda RNN asosidagi quyidagi uchta ilg‘or model arxitekturasi tanlanib, ular o‘quv ma’lumotlar to‘plamida o‘qitiladi va natijalari o‘zaro taqqoslanadi:

- Long Short-Term Memory (LSTM);
- Bi-directional LSTM (BiLSTM);
- Gated Recurrent Unit (GRU).

RNN vaqt bo‘yicha ketma-ketlikda o‘zgaruvchi ma’lumotlarni qayta ishlash va vaqtinchalik bog‘lanishlarni xotirada saqlashga ixtisoslashgan sun’iy neyron tarmoq arxitekturasi hisoblanadi. Model oldingi bosqichdagi hisob-kitob natijalarini saqlab qolib, keyingi bosqichlarda ulardan foydalanadi. Aynan shu xususiyati uni, imo-ishora kabi ketma-ketlikda o‘zgaruvchi signallarni qayta ishlashda samarali vositaga aylantiradi.

Mazkur tadqiqot doirasida qo‘llanilgan RNN asosidagi modellarning har biri o‘ziga xos ustunliklarga ega. Chunki LSTM uzoq muddatli bog‘lanishlarni saqlab qolish qobiliyatiga ega bo‘lib, oddiy RNNlarda uchraydigan gradient yo‘qolishi muammosini samarali hal qiladi [19]. BiLSTM esa LSTM va ikki yo‘nalishli RNN kombinatsiyasi bo‘lib, ketma-ketlikni har ikki yo‘nalishda o‘qitish orqali imo-harakatlarning semantik mazmunini chuqurroq anglash imkonini beradi. Shuningdek GRU arxitekturasi ham LSTM bilan o‘xshash bo‘lsa-da, parametrlar soni kamroq bo‘lgani sababli hisoblash samaradorligi yuqori va yengilroq model hisoblanadi.

Ushbu modellar kirishdagi vaqt bo‘yicha ketma-ket o‘zgaruvchi xususiyatlar asosida chiqish natijalarini generatsiya qiladi va shu orqali imo-ishora ma’lumotlar to‘plamidagi harakatlarni samarali tarzda tanib olish imkonini beradi. Har bir model arxitekturasi o‘ziga xos bo‘lib, 6–8-jadvallarda ularning qatlam parametrlari umumiy ko‘rinishda keltirilgan. Strukturaviy jihatdan, model quyidagicha tashkil topadi: dastlabki uchta qatlam tanlangan RNN modelining asosiy bloklarini ifodalaydi, so‘nggi uchta

qatlam esa to‘liq bog‘lanishli qatlamlardan iborat bo‘lib, imo-ishoralarni aniqlash va tasniflash (klassifikatsiya) vazifasini bajaradi.

1-jadval. LSTM modelining arxitekturasi va parametrlar jadvali

Qatlam (turi)	Chiqish shakli	Parametrlar soni
Kirish qatlami (InputLayer)	(None, 30, 258)	0
1-LSTM qatlami (LSTM)	(None, 30, 64)	82,688
2-LSTM qatlami (LSTM)	(None, 30, 128)	98,816
3-LSTM qatlami (LSTM)	(None, 64)	49,408
1-To‘liq bog‘lanishli (Dense)	(None, 64)	4,160
2-To‘liq bog‘lanishli (Dense)	(None, 32)	2,080
Chiqish qatlami (Dense)	(None, 200)	6,600

2-jadval. Bi -LSTM modelining arxitekturasi va parametrlar jadvali

Qatlam (turi)	Chiqish shakli	Parametrlar soni
Kirish qatlami (InputLayer)	(None, 30, 258)	0
1-bidirectional qatlami (Bidirectional)	(None, 30, 256)	396,288
2-bidirectional qatlami (Bidirectional)	(None, 30, 512)	1,050,624
3-bidirectional qatlami (Bidirectional)	(None, 256)	656,384
1-To‘liq bog‘lanishli (Dense)	(None, 64)	16,448
2-To‘liq bog‘lanishli (Dense)	(None, 32)	2,080
Chiqish qatlami (Dense)	(None, 200)	6,600

3-jadval. GRU modelining arxitekturasi va parametrlar jadvali

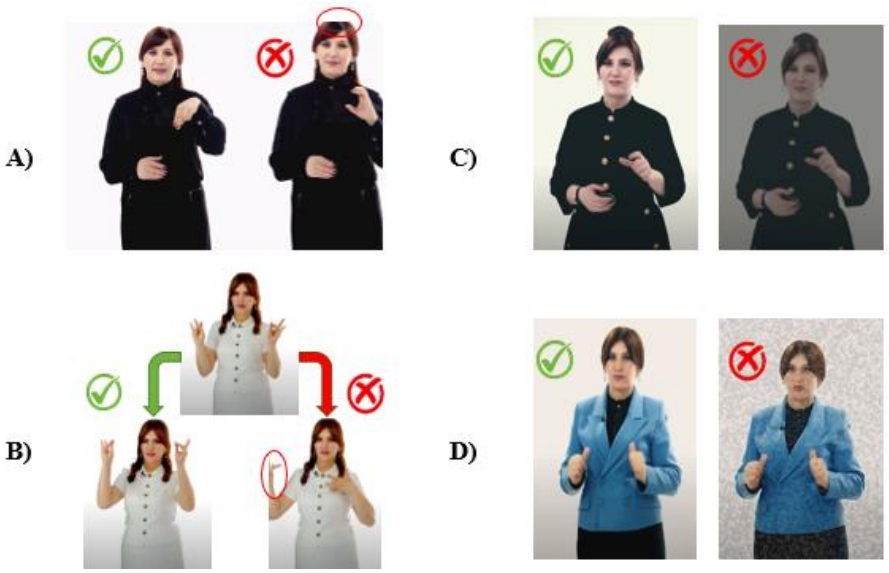
Qatlam (turi)	Chiqish shakli	Parametrlar soni
Kirish qatlami (InputLayer)	(None, 30, 258)	0
1-GRU qatlami (GRU)	(None, 30, 64)	62,288
2-GRU qatlami (GRU)	(None, 30, 128)	74,496
3-GRU qatlami (GRU)	(None, 64)	37,248
1-To‘liq bog‘lanishli (Dense)	(None, 64)	4,160
2-To‘liq bog‘lanishli (Dense)	(None, 32)	2,080
Chiqish qatlami (Dense)	(None, 200)	6,600

Modelni o‘qitish jarayonida optimallashtiruvchi parametrlar tanlanadi va har bir model uchun qatlam parametrlarining mos qiymatlari oldindan belgilab olinadi. So‘ngra ular keyingi o‘qitish (training) bosqichiga tayyorlanadi. Model uchun kirish qiymatlari sifatida ketma-ketlik uzunligi hamda asosiy nuqtalarning umumiy soni qabul qilinadi. Bu yerda ketma-ketlik uzunligi har bir videoklipdagi kadrlar sonini bildiradi, asosiy nuqtalarning umumiy soni esa yuz asosiy nuqtalarisiz 258 ta, yuz asosiy nuqtalari bilan esa 1662 tani tashkil etadi.

Shunday qilib, model ma’lumotlar to‘plamini qabul qilishga tayyorlanadi va videolardan ajratib olingan asosiy nuqtalar ketma-ketligiga asoslangan holda o‘qitish jarayonini boshlaydi. Ushbu jarayon asosida imo-harakatlar tahlil qilinib, mos qo‘l ishorasi aniqlanadi. Natijada, imo-ishora tili yuqori aniqlik va samaradorlik bilan tanib olinadi.

Ma’lumotlar to‘plami (DATASET). Mazkur tadqiqot doirasida ishlab chiqilgan ma’lumotlar to‘plami O‘zbek imo-ishora tiliga asoslangan bo‘lib, dinamik imo-ishoralarni avtomatik tanib olish va ilmiy-tadqiqot ishlari uchun foydalanishga mo‘ljallangan. Ushbu dataset yuqori aniqlikdagi video ma’lumotlarga ehtiyoj sezmaydigan, MediaPipe platformasi asosidagi yengil va hisoblash jihatidan samarali metodologiyaga moslashtirilgan. Ma’lumotlar to‘plamning yaratilishidan ko‘zlangan asosiy maqsad imo-ishora tilini soddalashtirilgan, biroq real sharoitga maksimal darajada yaqin holda aniqlash imkonini beruvchi modelni ta’minlashdir. To‘plam tarkibida kundalik muloqotda keng qo‘llaniladigan asosiy imo-ishoralar majmuasi jamlangan.

Mazkur tadqiqot uchun tuzilgan ma’lumotlar to‘plami qat’iy talablarga asoslangan bo‘lib, har bir imo-ishora uch nafar suxandon tomonidan ijro etilgan va ular tomonidan har bir ishora uchun uchta video yozib olingan. Shu tariqa, umumiy hajm 200 ishora × 3 suxandon × 3 video formulasi asosida 1800 ta video namunani tashkil etdi. Videolar HD Pro Webcam C920 qurilmasi yordamida USB kabel orqali kompyuterga ulanib, OpenCV kutubxonasining VideoCapture funksiyasi orqali yozib olindi. Har bir video 1 soniya davomiylikka, 30 FPS (kadr/sekund) tezlikka va 640×480 piksel o‘lchamga ega bo‘ldi. Yozib olish jarayonida imo-ishoralarni aniq va bir xil ifodalashga e’tibor qaratildi, hamda, bu jarayon maxsus ishlab chiqilgan soddalashtirilgan yozib olish modeli asosida amalga oshirildi. Bunday yondashuv murakkab texnik vositalardan foydalanmagan holda modelni o‘qitish va baholash imkonini beradi.



6-rasm. Toza va aniq ma’lumotlar to‘plamini yozib olish bo‘yicha ko‘rsatmalar: (A) Imo-ishora bajaruvchining tanasi, (B) Imo-ishora bajaruvchining harakati, (C) Yoritish, (D) Kamera sifati

Ma’lumotlarni yig‘ish jarayonida bir qator tavsiyalarga qat’iy amal qilindi. Avvalo, har bir kadrlarda imo-ishora bajaruvchining butun tanasi to‘liq ko‘rinib turishi ta’minlandi (6A-rasm). Imo-ishoraning barcha harakatlari kamera kadri doirasida aniq va to‘liq ifodalanishi kerakligi inobatga olindi (6B-rasm). Shuningdek, fon imkon qadar toza va barqaror bo‘lishi, kadrda boshqa qo‘l yoki yuz elementlari ishtirok etmasligi talab etildi. Yoritish sharoitlari asosiy nuqtalar to‘liq aniqlanishi uchun yetarli darajada bo‘lishi ta’minlandi (6C-rasm). Kamera tasvirni maksimal darajada barqaror va fokuslangan holatda yozib olish uchun statik stendga o‘rnatildi (6D-rasm).

Videolarni yig‘ishda davomiylik va umumiy kadrlar soni oldindan belgilab qo‘yildi. Optimal texnik sifatni ta’minlash maqsadida 640×480 o‘lchamdagi sensorli kameradan foydalanish tanlandi, chunki ushbu format zamonaviy kameralar tomonidan keng qo‘llab-quvvatlanadi va mashinali o‘qitish jarayonlari uchun yetarli aniqlikni ta’minlaydi.

3 TAJRIBA NATIJALARI

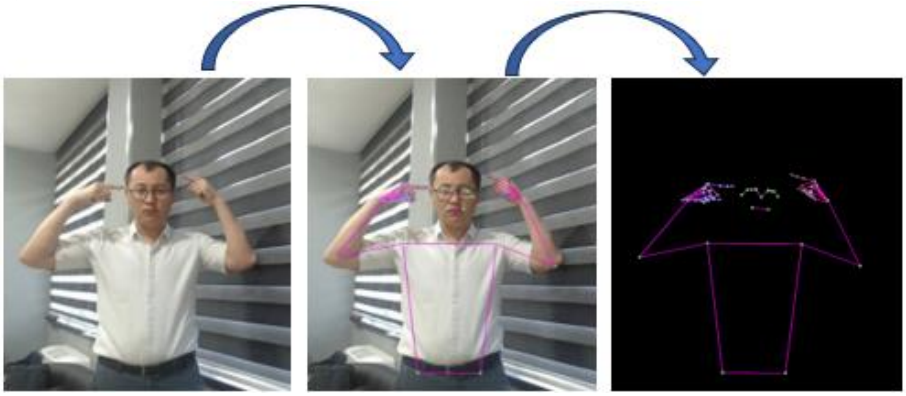
Taklif etilgan tizimning samaradorligini baholash maqsadida imo-ishora ma’lumotlar to‘plamidan tanlab olingan 20 ta imo-ishora belgisi asosida tajribalar tashkil etildi. Ma’lumotlar tasodifiy tarzda ajratilib, 70 foizi o‘qitish (training) va 30 foizi sinov (testing) uchun mo‘ljallandi. Natijada o‘qitish jarayoniga 1260 ta video klip, sinov jarayoniga esa 540 ta video klip ajratildi. Bunday yondashuv tajribalardagi tasodifiylikni kamaytirib, modellarni umumlashtirish qobiliyatini oshiradi.

Tajribalar yuqori samaradorlikka ega shaxsiy kompyuterda amalga oshirildi. Texnik konfiguratsiya quyidagicha: markaziy protsessor - AMD Ryzen 9 7950X (4.5 GHz chastotali), operativ xotira - 128 GB RAM, grafik protsessor - NVIDIA GeForce RTX 3090 (24 GB xotirali), xotira qurilmasi esa - Samsung SSD 990 EVO 1TB NVMe SSD. Ushbu hisoblash muhiti katta hajmdagi multimodal ma’lumotlarni qayta ishlash va chuqur o‘qitish modellarini samarali o‘qitish imkonini berdi.

Yuzning asosiy nuqtalarisiz MediaPipedan foydalanish bosqichda MediaPipe framework yordamida faqat qo‘l va tana holatiga oid 258 ta asosiy nuqta ajratib olindi. Muhim jihati shuki, bu nuqtalarning ayrimlari yuz yaqinida joylashgan bo‘lishi mumkin, biroq ular aslida yuz ifodalarini emas, balki tana pozitsiyasini aniqlashga xizmat qiladi. Ajratilgan asosiy nuqtalar GRU, LSTM va Bi-LSTM arxitekturalariga ega RNN modellariga uzatildi. Tajriba jarayonida kuzatilishicha, har bir modelning o‘qitish vaqti 20 daqiqadan 45 daqiqagacha davom etdi. 4-jadvalda esa barcha modellarning o‘qitish va sinov bo‘yicha F1-bali aniqligi hamda qo‘llanilgan epochlar soni keltirilgan.

4-jadval. Yuz asosiy nuqtalarisiz modellarning F1-bali aniqligi

Model	O‘qitish (train) F1-bali	Sinov (test) F1-bali	O‘qitish davrlari (epoch)
GRU	1.00	1.00	241
LSTM	0.999	0.996	65
BiLSTM	0.999	0.993	75



7-rasm. Kadrlardan asosiy nuqtalarni ajratib olish bosqichlari (yuzning asosiy nuqtalarisiz)

Natijalar tahlili shuni ko‘rsatdiki, GRU, LSTM va Bi-LSTM modellarining aniqlik ko‘rsatkichlari o‘zaro yaqin bo‘lib, sezilarli tafovut kuzatilmadi. Shunga qaramay, GRU arxitekturasi eng samarali va hisoblash nuqtai nazaridan yengil model sifatida ajralib chiqdi. U kamroq resurs talab qiladi hamda o‘qitish jarayonida har bir epochni LSTM va Bi-LSTM modellari bilan solishtirganda tezroq yakunlaydi. GRU modeli ko‘proq epoch talab etgan bo‘lsada, bashorat qilish bosqichida yuqori tezlikni ta’minladi.

Mazkur tajriba natijalari 7-rasmda ko‘rsatilganidek, yuz asosiy nuqtalari hisobga olinmagan holda amalga oshirildi. Keyingi bo‘limda esa yuz asosiy nuqtalari qo‘shilgan holatdagi o‘qitish natijalari keltirilib, ularning solishtirma tahlili beriladi.

Yuzning asosiy nuqtalari bilan MediaPipe dan foydalanish tajribasidan asosiy maqsadi MediaPipe framework yordamida qo‘l va tana pozitsiyasiga oid asosiy nuqtalarga yuz nuqtalarini ham qo‘shgan holda, ularning umumiy model samaradorligiga ta’sirini baholashdan iborat.

Qo‘l va tana pozitsiyasiga oid 258 ta asosiy nuqtaga qo‘shimcha ravishda, MediaPipe vositasi yordamida yuzning 1404 ta nuqtasi aniqlab olindi. Natijada jami 1662 ta asosiy nuqta shakllantirildi. Ushbu nuqtalar funksional jihatdan uch guruhga bo‘linadi: qo‘l harakatlariga oid nuqtalar, tana pozitsiyasini ifodalovchi nuqtalar hamda yuz ifodasi va tuzilishiga tegishli nuqtalar. Ajratilgan 1662 ta asosiy nuqta GRU, LSTM va BiLSTM arxitekturalariga ega RNN modellariga uzatildi. Bunday yondashuv xususiyatlar fazosining hajmini taxminan olti barobarga oshiradi. Bu esa hisoblash resurslari va vaqtiga sezilarli ta’sir ko‘rsatdi.

Natijalar tahlili shuni ko‘rsatdiki, yuz nuqtalarini ajratib olish bosqichi MediaPipe platformasi uchun ko‘proq vaqt talab qiladi. Shuningdek, modellar tomonidan ushbu xususiyatlarni qayta ishlash va bashorat qilish jarayonlari ham hisoblash resurslarini sezilarli darajada ko‘proq talab etdi. Yuz nuqtalari qo‘shilgan 8-rasmdagi holatda, har bir model uchun o‘qitish vaqti 30 daqiqadan 75 daqiqagacha davom etdi. Shunga qaramay, o‘qitish va sinov bosqichlarida yuqori aniqlik ko‘rsatkichlari saqlanib qoldi. Natijalariga tahliliga ko‘ra, turli model arxitekturalari orasidagi natijalar bir-biriga juda yaqin bo‘lib, farqlar statistik jihatdan sezilarli emasligi kuzatildi. Quyidagi 5-jadvalda GRU, LSTM va BiLSTM modellarning o‘qitish va sinov bo‘yicha F1-bali aniqligi hamda qo‘llanilgan epochlar soni keltirilgan.

5-jadval. Yuz asosiy nuqtalari bilan modellar aniqligi

Model	O‘qitish (train) F1-bali	Sinov (test) F1-bali	O‘qitish davrlari (epoch)
GRU	1.00	1.00	250
LSTM	0.99	0.996	36
BiLSTM	0.999	0.99	49

Biroq, avvalgi tajribada qayd etilganidek, GRU modeli ushbu sinovlarda ham eng samarali deb topildi. Chunki u resurs tejamkorligi, tezkorligi hamda ko‘plab xususiyatlar bilan ishlashda optimal yechim taklif qila olishi bilan ajralib turadi. Ayniqsa, xususiyatlar sonining keskin oshishi fonida GRU modelining hisoblash samaradorligi ayniqsa muhim omil bo‘lib xizmat qiladi.

Tajribalar natijalariga ko‘ra, yuz asosiy nuqtalari qo‘shilgan va qo‘shilmagan holatlarda model aniqligi deyarli teng bo‘lib chiqdi. Bu esa shuni ko‘rsatadiki, model konfiguratsiyasini tanlash foydalanish kontekstiga bog‘liq. Agar model imo-ishora tilida yuz ifodalari muhim rol o‘ynaydigan mintaqada qo‘llanilsa, yuz nuqtalarini kiritish maqsadga muvofiq bo‘ladi. Aksincha, yuz ifodalariga tayanilmaydigan hollarda bu zarurat bo‘lmasligi mumkin.Shunga qaramay, real vaqt rejimida ishlaydigan tizimlar uchun yuz nuqtalaridan foydalanish maqsadga muvofiq emas. Buning sababi shundaki, yuzga oid asosiy nuqtalar soni qo‘l va tana nuqtalariga nisbatan qariyb olti barobar ko‘p bo‘lib, mos ravishda 1662 ta va 258 ta ni tashkil

etadi. Shu bois modelning o‘qitish hamda bashorat qilish jarayonlarida sezilarli darajada ko‘proq vaqt va resurs talab qiladi.

8-rasm. Kadrlardan asosiy nuqtalarni ajratib olish bosqichlari (yuzning asosiy nuqtalari bilan)

Bundan tashqari, yuz asosiy nuqtalaridan foydalanish eshitish imkoniyati cheklangan foydalanuvchilardan har bir so‘z uchun aniq yuz ifodasi ko‘rsatishni talab qiladi. Bu holat foydalanuvchi uchun noqulaylik tug‘dirishi bilan birga, modelning to‘g‘ri aniqlash imkoniyatini ham pasaytirishi mumkin. Biroq oldindan belgilangan holatlarda, ayniqsa yuz ifodalari muhim bo‘lgan imo-ishoralar toifalarida yuz nuqtalarini kiritish foydali bo‘lishi ehtimoldan xoli emas.

Tajribalarda GRU modeli eng yuqori aniqlik ko‘rsatkichlariga erishgan bo‘lsa-da, bu natija uni har doim eng yaxshi tanlov sifatida qabul qilish mumkinligini anglatmaydi. Model tanlovi ma’lumotlar hajmi, murakkabligi va hisoblash resurslariga bog‘liq bo‘ladi. GRU arxitekturasi oddiy va kam parametrli ketma-ketliklarda LSTM hamda BiLSTM modellariga nisbatan samaraliroq natija beradi. Aksincha, LSTM va BiLSTM modellar murakkab, uzun yoki ko‘p qatlamli ma’lumotlarni qayta ishlashda ustunlikka ega[19].

Bu holat, o‘zbek imo-ishora tiliga oid yaratilgan datasetning nisbatan sodda va kichik hajmga egaligi bilan izohlanadi. Har bir klipdagi kadrlar sonining cheklanganligi va imo-ishoralar hajmining kichikligi GRU modelining samarali ishlashiga sharoit yaratgan. Shu bilan birga, murakkab va katta hajmdagi datasetlar uchun LSTM va BiLSTM modellar ko‘proq parametrlar va tugunlarni talab qiladi hamda hisoblash quvvatiga yuqori ehtiyoj sezadi. GRU modeli esa kamroq parametrlar bilan ishlaydi, resurs jihatidan tejamkor hisoblanadi va real vaqt rejimida bashorat qilish uchun qulay yechim sifatida qaraladi.

4 XULOSA

Mazkur tadqiqot o‘zbek imo-ishora tiliga asoslangan dinamik imo-ishoralarni avtomatik aniqlashga qaratilgan bo‘lib, GRU, LSTM va BiLSTM arxitekturalaridan foydalangan chuqur o‘qitish yondashuvlari ishlab chiqildi. Xususiyatlarni ajratish jarayoni MediaPipe platformasi yordamida amalga oshirildi, bunda qo‘l, yuz va tana pozitsiyalarining asosiy nuqtalari aniqlanib, modellarga kiritildi. O‘tkazilgan tajribalar ikki xil konfiguratsiyada sinovdan o‘tkazildi va har ikkisi ham yuqori aniqlik ko‘rsatkichlarini namoyon qildi.

Tahlillar shuni ko‘rsatdiki, GRU modeli resurs tejamkorligi bilan ajralib turib, mobil qurilmalar va real vaqt rejimida ishlovchi tizimlar uchun qulay hisoblanadi. LSTM va BiLSTM modellar esa murakkab va uzun ketma-ketliklarni o‘qitishda samaraliroq natija beradi. Yuz asosiy nuqtalarini qo‘shish aniqlikka sezilarli ta’sir ko‘rsatmagan bo‘lsa-da, ularning ko‘pligi hisoblash jarayonlarining sezilarli darajada sekinlashishiga olib keldi.

Bu kabi tadqiqotlar uchun turli yorug‘lik, fon va harakat tezligiga ega ishoralarni o‘z ichiga olgan balansli va keng qamrovli dataset yaratish zarur. Shuningdek, ma’lumotlar sifatini oshirish maqsadida shovqinli yoki noto‘g‘ri belgilangan kadrlarni avtomatik aniqlab tekshiruvchi algoritmlarini ishlab chiqish talab etiladi. Tabiiy sharoitda real vaqt rejimida imo-ishora jumlalarni tanib olish imkoniyatini ta’minlash istiqbolli yo‘nalishlardan biri hisoblanadi. Bundan tashqari, turli standartdagi videolarni normallashtirishga mo‘ljallangan oldindan qayta ishlash texnikalarini ishlab chiqish lozim. Kelgusida modellarni yanada umumlashtirish va barqaror ishlashini ta’minlash uchun keng hajmdagi, sifatli yorliqlangan ma’lumotlar to‘plamini shakllantirish muhim ahamiyat kasb etadi.

ADABIYOTLAR

[1] *Abdalla, M.S.; Hemayed, E.E.* Dynamic hand gesture recognition of Arabic sign language using hand motion trajectory features. Glob. J. Comput. Sci. Technol. 2013, 13, 27–33.

[2] *Kim, T.; Keane, J.; Wang, W.; Tang, H.; Riggle, J.; Shakhnarovich, G.; Brentari, D.; Livescu, K.* Lexicon-free fingerspelling recognition from video: Data, models, and signer adaptation. Comput. Speech Lang. 2017, 46, 209–232.
[3] *Mohandes, M.; Deriche, M.; Liu, J.* Image-based and sensor-based approaches to Arabic sign language recognition. IEEE Trans. Hum.-Mach. Syst. 2014, 44, 551–557.
[4] *Ding, L.; Martinez, A.* Three-dimensional shape and motion reconstruction for the analysis of American Sign Language. In Proc. 2006 IEEE CVPR Workshop, New York, NY, USA, 17–22 June 2006; pp. 146–146.
[5] Dogra, *D.P.; Kumar, P.; Gauba, H.; Roy, P.P.* A multimodal framework for sensor-based sign language recognition. Neurocomputing 2017, 259, 21–38.
[6] *Zadgorban, M.; Nahvi, M.* Continuous Persian sign language recognition using trajectory-based features. Pattern Recognit. Lett. 2017, 90, 36–43.
[7] *Zhang, J.; Zhou, W.; Pu, J.; Li, H.* Continuous sign language recognition with joint CNN-HMM model. In Proc. 2016 Int. Conf. on Pattern Recognition (ICPR); pp. 1838–1843.
[8] *Kishore, K.; Kumar, P.; Yadav, D.* Indian sign language recognition using CNN and transfer learning. Procedia Comput. Sci. 2020, 167, 2141–2149.
[9] *Manikanta, S.; Ramesh, M.* Dynamic sign language recognition using hybrid CNN-RNN architecture. Int. J. Recent Technol. Eng. 2019, 8(2), 210–216.
[10] *Camgoz, N.C.; Hadfield, S.; Koller, O.; Bowden, R.* SubUNets: End-to-end hand shape and continuous sign language recognition. In Proc. IEEE Int. Conf. Comput. Vision (ICCV), 2017; pp. 3056–3065.
[11] *Huang, J.; Zhou, W.; Zhang, H.; Li, H.* Video-based sign language recognition without temporal segmentation. In Proc. 2018 AAAI Conf. on Artificial Intelligence; pp. 2257–2264.
[12] *Allam, M.; Hemayed, E.* Arabic sign language recognition system for alphabets using CNN. J. King Saud Univ.-Comput. Inf. Sci. 2019, 31, 381–388.
[13] *Altagafi, S.; et al.* CNN-based recognition of Arabic sign language letters using RGB images. J. Comput. Sci. 2019.
[14] *Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Lee, J.; Grundmann, M.* MediaPipe: A Framework for Building Perception Pipelines. arXiv preprint arXiv:1906.08172, 2019.
[15] *Grishchenko, I., Bazarevsky, V., Zhang, F., & Grundmann, M.* (2020). Attention Mesh: High-fidelity face mesh prediction in real-time. arXiv preprint arXiv:2006.10962.
[16] *Zhang F., Bazarevsky V., Vakunov A., others.* MediaPipe Hands: On-device real-time hand tracking. CVPR Workshops, 2020, 6683–6692.
[17] *Li, K.; Zhou, Z.; Lee, C.H.* Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications. ACM Trans. Access. Comput. (TACCESS) 2016, 8, 1–23.
[18] *Yang, X.; Chen, X.; Cao, X.; Wei, S.; Zhang, X.* Chinese sign language recognition based on an optimized tree-structure framework. IEEE J. Biomed. Health Inform. 2016, 21, 994–1004.
[19] *Graves, A., & Schmidhuber, J.* (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5–6), 602–610.

Поступила в редакцию 30.04.2025

Citation: *Jurayev D.B.* (2025). Matndan o‘zbek imo-ishora tiliga tarjima qilish uchun mediapipe asosida ma’lumotlar to‘plamini tayyorlash texnologiyasi. Raqamli texnologiyalarning nazariy va amaliy masalalari xalqaro jurnali. 8(3). –B. 82-93. https://doi.org/10.62132/ijdt.v8i3.290.

TECHNOLOGY FOR PREPARING A DATASET BASED ON MEDIAPIPE FOR TRANSLATING TEXT TO UZBEK SIGN LANGUAGE

Juraev D.B.[1]

[1] Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

dilsamtuit@gmail.com

Abstract. This paper discusses the problem of creating a multimodal dataset based on MediaPipe technology for automatic translation of Uzbek sign language into text and speech.

Due to the multifaceted nature of sign language hand movements, facial expressions, body posture, and gaze direction the technical, linguistic, and ethical aspects of dataset preparation are analyzed. Using MediaPipe, key points of the hands, face, and body are extracted, and experiments are conducted with GRU, LSTM, and BiLSTM models. Comparative analysis demonstrates the efficiency and resource-saving capability of the GRU model, while LSTM and BiLSTM show advantages in processing complex sequences. Thus, the multimodal dataset developed in this study provides a reliable foundation for the development of real-time sign language recognition systems.

Keywords: sign language, MediaPipe, multimodal dataset, RNN (GRU, LSTM, BiLSTM), computer vision, natural language processing (NLP), artificial intelligence.

ТЕХНОЛОГИЯ ПОДГОТОВКИ НАБОРА ДАННЫХ НА ОСНОВЕ MEDIAPIPE ДЛЯ ПЕРЕВОДА ТЕКСТА НА УЗБЕКСКИЙ ЖЕСТОВЫЙ ЯЗЫК

Джураев Д.Б.[1]

[1] Ташкентский университет информационных технологий имени Мухаммада аль-Хорезми, Ташкент, Узбекистан

dilsamtuit@gmail.com

Аннотация. В статье рассматривается проблема формирования мультимодального набора данных на основе технологии MediaPipe для автоматического перевода узбекского жестового языка в текст и речь. Из-за многоаспектного характера жестового языка движений рук, мимики лица, позиций тела и направления взгляда анализируются технические, лингвистические и этические аспекты подготовки датасета. С помощью MediaPipe выделяются ключевые точки рук, лица и тела, на основе которых проводятся эксперименты с моделями GRU, LSTM и BiLSTM. Сравнительный анализ показал эффективность и ресурсосберегающий характер модели GRU, а также преимущества LSTM и BiLSTM при обработке сложных последовательностей. Таким образом, созданный в исследовании мультимодальный датасет служит надежной основой для разработки систем распознавания жестов, работающих в режиме реального времени.

Ключевые слова: жестовый язык, MediaPipe, мультимодальный датасет, RNN (GRU, LSTM, BiLSTM), компьютерное зрение, обработка естественного языка (NLP), искусственный интеллект.