

UDK 004.8:004.93:37.02

SUN'IY INTELLEKT YORDAMIDA MOBIL KAMERA TASVIRLARIDA OBYEKTЛАRNI ANIQLASH VA ULARNING INGLIZCHA NOMLARINI OVOZLI IFODALASH TIZIMINI ISHLAB CHIQISH

Husan Arziqulov Normurod o'g'li

Navoiy davlat universiteti dotsenti

arziqulovhusannormurodovich@gmail.com

<https://orcid.org/0009-0005-0970-3766>

Annotatsiya: Mazkur maqolada Android mobil qurilmasi kamerasi orqali real muhitdagi o'quv qurollarini aniqlash va aniqlangan obyektning inglizcha nomini ovozli ifodalashga mo'ljallangan tizimning konseptual-texnik arxitekturasini hamda statik tasvirlardagi pilot AI sinovi yoritilgan. Taklif etilgan arxitektura CameraX asosidagi kamera oqimi, server tomondagi open-vocabulary obyekt detektori, natijalarni filtrlash va Android Text-to-Speech modullarini birlashtiradi.

Kalit so'zlar: mobil kamera, obyektlarni aniqlash, open-vocabulary detection, OWLv2, Android, CameraX, Text-to-Speech, kompyuter ko'rishi, o'quv qurollari, ingliz tili, multimodal ta'lim, temporal filtrlash.

Аннотация: В данной статье представлена концептуально-техническая архитектура системы, предназначенной для распознавания учебных принадлежностей в реальной среде с использованием камеры мобильного устройства на базе Android и голосового воспроизведения английских названий распознанных объектов, а также рассмотрены результаты пилотного тестирования компонента искусственного интеллекта на статических изображениях. Предлагаемая архитектура объединяет поток изображений, формируемый с использованием CameraX, серверный модуль обнаружения объектов с открытым словарём (open-vocabulary object detection), модуль фильтрации результатов и технологию синтеза речи Android Text-to-Speech.

Ключевые слова: мобильная камера, обнаружение объектов, open-vocabulary detection, OWLv2, Android, CameraX, Text-to-Speech, компьютерное

зрение, учебные принадлежности, английский язык, мультимодальное обучение.

Abstract: This article presents the conceptual and technical architecture of a system designed to recognize school supplies in a real-world environment using an Android mobile device camera and to provide voice output of the English names of the recognized objects; it also discusses the results of pilot testing the artificial intelligence component on static images. The proposed architecture integrates an image stream generated via CameraX, a server-side open-vocabulary object detection module, a result filtering module, and Android Text-to-Speech technology.

Key words: mobile camera, object detection, open-vocabulary detection, OWLv2, Android, CameraX, Text-to-Speech, computer vision, school supplies, English vocabulary, multimodal learning, temporal filtering.

Kirish

Mobil qurilmalar kamerasi kompyuter ko‘rishi modellariga bevosita kirish nuqtasi bo‘lib, foydalanuvchining kundalik muhitidagi predmetlarni o‘quv kontekstiga aylantirish imkonini beradi. Bunda tasvirni shunchaki klassifikatsiya qilish emas, balki obyektning kadrdagi joylashuvini topish, unga semantik nom biriktirish va natijani foydalanuvchiga qulay modalitetda yetkazish talab qilinadi. Ingliz tili lug‘atini o‘rganishda bu yondashuv "ko‘rish - nomlash - eshitish" zanjirini shakllantiradi: o‘quvchi real predmetni kameraga ko‘rsatadi, tizim uning inglizcha nomini aniqlaydi va shu nomni talaffuz qiladi. Bunday jarayon abstrakt lug‘at ro‘yxatini real kontekst bilan bog‘laydi.

Zamonaviy vision-language modellarining rivojlanishi ushbu vazifani oldindan qat’iy belgilangan sinflar doirasidan chiqarishga imkon berdi. CLIP tabiiy til va tasvir reprezentatsiyalarini yagona semantik fazoda bog‘lash orqali zero-shot tanib olish imkoniyatini namoyish etdi [1]. Ushbu g‘oya obyekt deteksiyasiga ko‘chirilgach, OWL-ViT matn so‘rovi bilan boshqariladigan open-vocabulary detektor sifatida yangi sinflarni maxsus qayta o‘qitmasdan izlash imkonini berdi [2]. Keyingi OWLv2 modeli self-training va katta hajmdagi image-text ma’lumotlaridan foydalanib open-vocabulary deteksiyaning masshtablanuvchanligini kuchaytirdi [3]. Grounding DINO

[4] va YOLO-World [5] ishlari ham matn orqali berilgan obyekt tushunchalarini lokalizatsiya qilish open-set va real vaqt deteksiyasi uchun istiqbolli yo‘nalish ekanini tasdiqlaydi.

Ta’lim nuqtai nazaridan real predmetni so‘z, tasvir va ovoz bilan bog‘lash alohida ahamiyatga ega. ARbis Pictus tadqiqotida real obyektlarni dinamik yorliqlash asosidagi lug‘at o‘rganish an’anaviy kartochkalarga nisbatan eslab qolish natijalarini yaxshilagani ko‘rsatilgan [6]. VocabulARy tizimi atrofdagi obyektlarni yozma va ovozli lug‘at birliklari bilan boyitib, kontekstga bog‘langan o‘rganishning afzalligini ko‘rsatadi [7]. 2025-yilda taqdim etilgan LangEye mobil ilovasi esa foydalanuvchining o‘zi kamera orqali real predmetlarni tanlab, ulardan shaxsiy lug‘at materiallari yaratishi mumkinligini amaliy namoyish etdi [8]. 2026-yildagi AR va multimodal lug‘at o‘rganish bo‘yicha tajriba ham tasvir, ovoz va kontekstni birlashtirish lug‘atni o‘zlashtirish va saqlab qolishga ijobiy ta’sir ko‘rsatishi mumkinligini qayd etadi.

Shunga qaramay, oddiy “kamera - detektor – ovoz” sxemasi jonli muhitda yetarli emas. Kamera harakatlanganda kadrlar tez almashadi, obyekt qisman yopilishi, yoritilish o‘zgarishi yoki o‘xshash predmetlar bir-biriga juda yaqin turishi mumkin. Agar har bir kadr natijasi darhol Text-to-Speech moduliga uzatilsa, bir xil nomning takror-takror aytilishi yoki qisqa muddatli noto‘g‘ri deteksiya ovozga aylanishi ehtimoli ortadi. Shu sababli mobil o‘quv tizimida vizual aniqlash bilan bir qatorda temporal barqarorlashtirish, ishonchlilik chegarasi, ovozli javobni boshqarish va foydalanuvchi interfeysidagi qayta aloqa mexanizmlari ham loyihalanishi kerak.

Mazkur tadqiqotning maqsadi Android kamera oqimida o‘quv qurollarini open-vocabulary sun’iy intellekt modeli yordamida aniqlash, aniqlangan obyektning inglizcha nomini ekranda ko‘rsatish va ovozli talaffuz qilish tizimining konseptual-texnik arxitekturasini ishlab chiqish hamda foydalanuvchi taqdim etgan tasvirlar asosida pilot funksional tahlil o‘tkazishdan iborat. Tadqiqot vazifalari: o‘quv qurollariga mos detektor turini tanlash; Android va AI server o‘rtasidagi ma’lumot almashinuvini loyihalash; ovozli chiqish uchun temporal tasdiqlash qoidalarini shakllantirish; turli murakkablikdagi tasvirlarda tizimning kuchli va zaif tomonlarini

tahlil qilish; kengaytirilgan jonli sinov uchun o‘lchov mezonlarini belgilashdan iborat.

Mavzuga oid adabiyotlarning tahlili

Obyektlarni aniqlashning klassik yondashuvlarida model oldindan belgilangan yopiq sinflar to‘plamida o‘qitiladi. Bunday sxema umumiy benchmarklarda samarali bo‘lsa-da, ta’lim ilovasida predmetlar ro‘yxatini kengaytirish zarurati tez-tez paydo bo‘ladi. Masalan, pencil, pen, marker, compass, protractor, stapler yoki hole punch kabi tushunchalarning barchasi bitta standart deteksiya datasetida mavjud bo‘lmashligi mumkin. CLIP tomonidan ko‘rsatilgan tasvir-matn moslashtirish paradigmasi sinf nomini tabiiy til orqali berish imkonini yaratdi. Bu esa o‘quv ilovasida kategoriyalarni kod ichida qayta o‘qitish bilan emas, prompt ro‘yxatini o‘zgartirish bilan boshqarish g‘oyasini asoslaydi.

Minderer va hammualliflar OWL-ViT modelida Vision Transformer arxitekturasini image-text pretraining bilan birlashtirib, text-conditioned zero-shot deteksiyani amalga oshirdi [2]. Modelning muhim afzalligi - so‘rov sifatida tabiiy til birliklaridan foydalanishi. O‘quv qurollari uchun “a photo of a pencil” yoki “a photo of a hole punch” kabi promptlar shu imkoniyatdan foydalanadi. Keyingi OWLv2 [3] self-training yordamida deteksiya ma’lumotlarini masshtablashga qaratilgan. Mazkur ishning statik tasvirlar bo‘yicha pilot sinovida aynan google/owlv2-base-patch16-ensemble checkpointi tanlandi; tanlov modelning matn promptlari bilan ishlashi va ochiq lug‘atli deteksiya vazifasiga mosligi bilan asoslandi.

Open-vocabulary deteksiyada muqobil yondashuvlar ham mavjud. Grounding DINO [4] kategoriya nomlari yoki referring expression kabi matnli so‘rovlar bilan obyektlarni lokalizatsiya qilishga mo‘ljallangan open-set detektordir. YOLO-World [5] esa YOLO arxitekturasini vision-language modeling bilan birlashtirib, real vaqtga yo‘naltirilgan open-vocabulary deteksiyani taklif etadi. Ushbu modellar keyingi bosqichda mobil tizimning server qismida OWLv2 bilan qiyosiy benchmark o‘tkazish uchun nomzod bo‘la oladi. Mazkur pilotning vazifasi modellar o‘rtasida benchmark g‘olibini aniqlash emas, balki tanlangan modelning berilgan uchta rasmda xatti-

harakatini qayd etish va Android tizimi uchun tekshiriladigan arxitektura talablarini shakllantirishdir.

Kompyuter ko‘rishi komponentining pedagogik qiymati foydalanuvchi obyekt bilan qanday o‘zaro ta’sir qilishiga bog‘liq. Ibrahim va hammualliflarning ARbis Pictus tizimida real muhitdagi obyektlarni xorijiy til birliklari bilan bog‘lash produktiv eslab qolish ko‘rsatkichlariga ijobiy ta’sir ko‘rsatgan [6]. Weerasinghe va hammualliflar ishlab chiqqan VocabulARy tizimi ham atrofdagi obyektlar uchun yozma va audio belgilarni taqdim etish orqali real kontekstni lug‘at o‘rganishga ulaydi [7]. Ushbu ishlar tasvir va tovushning birlashtirilishi nafaqat interfeys funksiyasi, balki o‘quv mazmunini kontekstualizatsiya qilish mexanizmi ekanini ko‘rsatadi.

LangEye [8] mazkur tadqiqotga eng yaqin zamonaviy ishlardan biridir: foydalanuvchi smartfon kamerasi yordamida real predmetni suratga oladi, keyin obyektни tanish va generativ NLP komponentlari orqali talaffuz, ta’rif va misollar bilan boyitilgan shaxsiy lug‘at materialı yaratiladi. Farqli jihat shundaki, taklif qilinayotgan tizim o‘quv qurollariga yo‘naltirilgan real vaqtga yaqin aniqlash oqimini va noto‘g‘ri ovozli chiqishni kamaytirishga qaratilgan temporal tasdiqlashni markaziy texnik komponent sifatida ajratadi. Android CameraX hujjatlarida ImageAnalysis yordamida kamera kadrlarini kompyuter ko‘rishi yoki mashinaviy o‘qitish tahliliga uzatish imkoniyati ko‘rsatiladi. Android TextToSpeech API esa matnni qurilmadagi nutq sintezatori orqali ovozli ijro etish imkonini beradi [10].

1-jadval. Mavzuga yaqin tadqiqotlar va texnologiyalarning qiyosiy tahlili

Manba	Yondashuv	Kuchli tomoni	Mazkur ish uchun xulosa
[1] CLIP	Tasvir-matn kontrastiv pretraining	Zero-shot semantik moslik	Matn promptlari bilan sinf berish g‘oyasi
[2] OWL-ViT	Open-vocabulary object detection	Text-conditioned lokalizatsiya	OWLv2 uchun metodik asos

Manba	Yondashuv	Kuchli tomoni	Mazkur ish uchun xulosa
[3] OWLv2	Self-training asosidagi OVD	Deteksiya ma'lumotini masshtablash	Pilot sinovda tanlangan detektor
[4] Grounding DINO	Open-set grounded detection	Erkin matnli so'rovlarni lokalizatsiya qiladi	Murakkab promptlarda qiyoslash mumkin
[5] YOLO-World	Real-time open-vocabulary YOLO	Tezlikka yo'naltirilgan	Jonli oqim uchun muqobil server detektori
[6] ARbis Pictus	AR til o'rganish	Real obyekttni xorijiy so'z bilan bog'lash	Pedagogik kontekstni asoslaydi
[7] VocabulARY	AR + yozma/audio lug'at	Multimodal taqdimot	Tasvir + ovoz integratsiyasini asoslaydi
[8] LangEye	Kamera + object recognition + pronunciation	Real predmetdan shaxsiy lug'at yaratish	Mazkur tizimga eng yaqin ta'limiy precedent

Tadqiqot metodologiyasi

Tadqiqot amaliy prototiplash va tizimli loyihalash yondashuvlariga asoslandi. Taklif etilayotgan yechim ikki asosiy qatlamdan iborat: Android mijoz ilovasi va AI inference serveri. Android qatlami CameraX orqali Preview hamda ImageAnalysis oqimini boshqarishi, tahlil uchun tanlangan kadr esa ixcham tasvir formatida HTTPS orqali serverga yuborilishi ko'zda tutiladi. Server qatlamida OWLv2 matnli promptlar asosida obyektlarni lokalizatsiya qiladi va bounding box, label hamda confidence score qiymatlarini Android ilovasiga qaytaradi. Mazkur ishda Android jonli kamera oqimi to'liq empirik sinovdan o'tkazilmagan; bevosita tajriba AI detektorining statik rasmga qo'llanishi bilan cheklangan.

Pilot AI sinovida Hugging Face Transformers orqali google/owlv2-base-patch16-ensemble checkpointidan foydalanildi [12]. Barcha uch rasm uchun bir xil 11 ta kategoriya - pencil, pen, eraser, ruler, notebook, book, scissors, stapler, hole

punch, pencil case va marker - berildi; modelga ular “a photo of a {label}” shaklidagi inglizcha matnli soʻrovlar sifatida uzatildi. Post-processingda deteksiya chegarasi 0,10 etib belgilandi. Bu qiymat yakuniy optimal threshold sifatida emas, uchta rasmdagi model chiqishlarini koʻrish uchun pilot parametr sifatida ishlatildi. Open-vocabulary yondashuv yangi kategoriyani prompt orqali soʻrashga imkon beradi, biroq yangi prompt qoʻshish yuqori deteksiya sifatini kafolatlamaydi [2], [3].

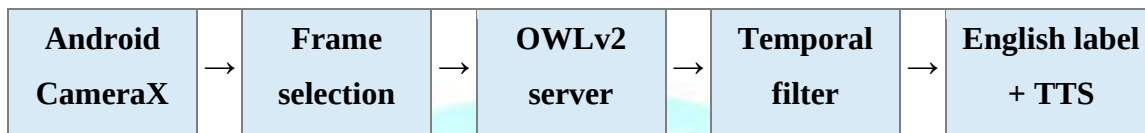
Jonli kamera bosqichi uchun CameraX ImageAnalysis modulining STRATEGY_KEEP_ONLY_LATEST backpressure strategiyasidan foydalanish maqsadga muvofiq: rasmiy Android hujjatlarida bu strategiya tahlilchi band boʻlganda eski kadr oʻrniga eng yangi mavjud kadrni saqlashi koʻrsatilgan. Server bilan ishlaydigan prototipda ushbu mexanizmga qoʻshimcha ravishda bir vaqtning oʻzida yuboriladigan inference soʻrovlari sonini cheklash tavsiya etiladi.

Ovozli chiqish uchun Android TextToSpeech API da ingliz tili lokali oʻrnatilib, aniqlangan label speak() metodi orqali talaffuz qilinishi mumkin [10]. Jonli oqimda qisqa muddatli label almashishlari yoki bir xil soʻzning takroriy aytilishini cheklash maqsadida uchta tahlil qilingan kadrda kamida ikkitasida bir xil labelni tasdiqlash va bir xil label uchun ikki soniyalik cooldown intervali boshlangʻich algoritmik parametr sifatida taklif etiladi. Bu parametrlar ushbu ishda tajriba orqali optimallashtirilmagan; ularning foydasi va optimal qiymati faqat jonli kamera tajribasida baholanishi mumkin.

Foydalanuvchi interfeysida kamera preview qatlami, bounding box overlay, inglizcha label va ishonchlilik qiymati, shuningdek oxirgi ovozli aytilgan soʻz koʻrsatiladi. Taʼlimiy foydalanish uchun labelni foydalanuvchi bosganda soʻzni yana bir marta talaffuz qilish funksiyasi foydali: bu avtomatik cooldown cheklovini buzmasdan oʻquvchiga ixtiyoriy takrorlash imkonini beradi. Shuningdek, server javobi kelmasa yoki internet mavjud boʻlmasa, kamera oynasini bloklamasdan “AI server bilan aloqa yoʻq” holatini koʻrsatish tavsiya etiladi. Bu ajratish tarmoq xatosini obyekt “topilmadi” holati bilan chalkashtirmaslikka xizmat qiladi.

Server orqali ishlov berish maxfiylik va trafik masalasini ham keltirib chiqaradi. Oʻquv ilovasi uchun faqat inference zarur boʻlgan kadr yuborilishi,

serverda tasvirlarni doimiy saqlamaslik va HTTPSdan foydalanish maqsadga muvofiq. Empirik ma'lumot yig'iladigan keyingi bosqichda foydalanuvchining roziligi, anonimlashtirish va loglarda faqat zarur texnik metrikalarni saqlash talab qilinadi. Bu maqoladagi pilotda inson yuzi yoki shaxsga doir ma'lumot tahlil qilinmagan; test obyektlari o'quv qurollari bilan cheklangan.



1-rasm. Mobil kamera asosidagi obyektни aniqlash va ovozli ifodalash tizimining arxitekturası

2-jadval. Tizimning funksional komponentlari

Komponent	Texnologiya / usul	Vazifasi
Kamera qatlami	Android CameraX Preview + ImageAnalysis	Jonli tasvirni ko'rsatish va AI uchun kadr olish
Tarmoq qatlami	HTTPS/REST, JSON javob	Kadrni serverga yuborish va deteksiyani qabul qilish
AI detektorı	OWLv2, text prompts	O'quv qurollarini open-vocabulary usulda lokalizatsiya qilish
Post-processing	Threshold + takroriy boxlarni filtrlash	Zaif va ortiqcha natijalarni kamaytirish
Temporal filtr	Taklif: 3 kadrđan kamida 2 tasida tasdiq	Jonli sinovda tekshiriladigan barqarorlashtirish qoidasi
Ovozli chiqish	Android TextToSpeech, English locale	Aniqlangan obyektning inglizcha nomini talaffuz qilish
Interfeys	Bounding box, label, score, repeat action	Vizual va audio qayta aloqani birlashtirish

Sinov metodikasi ikki darajaga ajratildi. Birinchi daraja - ushbu maqolada keltirilgan uchta real tasvir asosidagi pilot funksional tekshiruv. Bu bosqich modelning turli vizual sharoitlarda qanday muammolarga duch kelishini aniqlashga xizmat qiladi va statistik umumlashtirish uchun mo'ljallanmagan. Ikkinchi daraja -

Android qurilmasida jonli kamera orqali kengaytirilgan validatsiya. Unda har bir o'quv quroli uchun turli fon, masofa, yoritilish va qisman berkitish holatlarida belgilangan miqdorda kadr yig'ilib, ground-truth bounding boxlar bilan solishtirilishi kerak. Deteksiya uchun Precision, Recall va mAP@0.5, ovozli modul uchun spoken-label accuracy, butun pipeline uchun esa frame yuborilgan paytdan ovoz boshlanguncha bo'lgan end-to-end latency o'lchanadi.

Tahlil va natijalar

Pilot materiallar taqdim etgan uchta real fotosuratdan iborat. Birinchi rasmda yorug' fon ustida ikkita rangli qalam, ikkinchi rasmda yashil qog'oz teshgich va kadr chetida qisman ko'ringan yozuv qurollari, uchinchi rasmda esa qalamdon ichida bir-birini qisman berkitgan ko'plab yozuv qurollari tasvirlangan. Ushbu uch holat mos ravishda ajratilgan obyekt, dominant obyekt hamda zich va occlusion mavjud sahnani ifodalaydi.

Bir xil promptlar va 0,10 threshold bilan o'tkazilgan OWLv2 sinovida 1-rasmdagi ikki qalam uchun pencil labeli 0,6561 va 0,6207 confidence ballari bilan qayd etildi. Shu rasmda stapler labeli bilan 0,2934 balli qo'shimcha deteksiya ham hosil bo'ldi; kadr chetidagi obyektning object-level ground truth labeli mavjud bo'lmagani sababli ushbu deteksiya to'g'riligi hisoblanmadi. 2-rasmda hole punch deteksiyasi 0,8401 ball bilan qayd etildi; shu bilan birga ruler uchun 0,5405 balli deteksiya va kadr chetidagi yozuv qurollariga tegishli bir nechta past balli pencil deteksiyalari kuzatildi. Rasmning barcha obyektlari uchun ground-truth annotatsiya mavjud bo'lmagani sababli ruler deteksiyasi ham statistik to'g'ri yoki noto'g'ri sinfga kiritilmadi. 3-rasmda eng yuqori natijalar pen (0,5238) va marker (0,5202; 0,5065) labelariga tegishli bo'lib, ko'plab bounding boxlar bir-biri bilan ustma-ust tushdi; pencil uchun ham deteksiyalar mavjud bo'ldi. Uchinchi rasmda ham har bir buyum bo'yicha ground-truth bounding box tayyorlanmagani sababli bu deteksiyalarning barchasini to'g'ri yoki noto'g'ri deb son jihatdan tasniflash mumkin emas.

Natijalar threshold tanlashning muhimligini ko'rsatadi, biroq uchta rasm va to'liq object-level ground truth mavjud bo'lmagan sharoitda optimal thresholdni aniqlash mumkin emas. Faqat confidence ballining kattaligiga qarab deteksiyaning

semantik to‘g‘riligini tasdiqlab bo‘lmaydi. Threshold kelgusida annotatsiyalangan validatsiya to‘plamida Precision va Recall muvozanati asosida tanlanishi kerak. Jonli kamera uchun taklif etilgan temporal tasdiqlash ham barqaror label olishga qaratilgan, ammo uning xatolarni qanchalik kamaytirishi mazkur statik tajribada o‘lchanmagan.

Pilotning ilmiy talqini ehtiyotkor bo‘lishi kerak. Uchta tasvir asosida mAP, umumiy aniqlik yoki modelning barcha o‘quv qurollariga nisbatan ishonchliligi haqida statistik xulosa chiqarib bo‘lmaydi. Mazkur bosqichning natijasi - tanlangan modelning uchta rasm bo‘yicha chiqishlarini qayd etish, vizual murakkablik turlarini aniqlash va keyingi empirik tajriba dizaynini shakllantirishdir. Kengaytirilgan sinovda har bir predmet uchun annotatsiyalangan kadrlar, turli yoritilish va joylashuv sharoitlari hamda bir xil qurilmada vaqt o‘lchovlari to‘planishi talab etiladi.



a) rangli qalamlar



b) qog‘oz teshgich



c) zich qalam-ruchkalar

2-rasm. Pilot sinov uchun taqdim etgan turli murakkablikdagi o‘quv qurollari tasvirlari

3-jadval. Pilot tasvirlar bo‘yicha funksional tahlil

Holat	OWLv2 chiqishi	Confidence score	Ilmiy talqin
1-rasm	pencil; pencil; stapler	0,6561; 0,6207; 0,2934	Ikki qalam mos label bilan topildi; stapler chiqishi object-level ground truth bilan tasdiqlanmadi.

Holat	OWLv2 chiqishi	Confidence score	Ilmiy talqin
2-rasm	hole punch; ruler; pencil	0,8401; 0,5405; 0,2436 va past	Hole punch mos label bilan topildi; ruler chiqishi uchun ground truth yo‘q. Chetdagi yozuv qurollari past ball bilan qayd etildi.
3-rasm	pen; marker; pencil	eng yuqori: 0,5238; 0,5202; 0,4568	Ko‘plab ustma-ust boxlar; object-level ground truth bo‘lmagani uchun statistik to‘g‘rilik hisoblanmadi.

Kengaytirilgan tajriba uchun to‘rtta nazorat omili taklif etiladi. Birinchisi - yoritilish: yorqin, odatiy xona va nisbatan xira sharoit. Ikkinchisi - obyekt masofasi: yaqin, o‘rta va uzoq. Uchinchisi - fon murakkabligi: bir xil rangli fon, stol ustidagi odatiy fon va ko‘p obyektli fon. To‘rtinchisi - occlusion: ochiq, qisman yopilgan va zich joylashgan holat. Har bir kombinatsiyada ground-truth label va bounding box tayyorlansa, modelning qaysi sharoitda xato qilishi aniq ko‘rinadi. Bunda faqat umumiy accuracy emas, ayniqsa false positive va false negative turlari alohida tahlil qilinishi lozim.

Ovozli interfeys uchun alohida baholash protokoli talab etiladi. Detektor to‘g‘ri bo‘lsa ham, foydalanuvchi bir obyektни bir necha soniya kamerada ushlab turganda TTS nechta marta ishga tushishi, noto‘g‘ri qisqa deteksiya ovozga aylanadimi va obyekt almashganda yangi so‘z qancha kechikish bilan aytilishi o‘lchanadi. Shuning uchun spoken-label accuracy deteksiya aniqligiga teng emas: u temporal filtr va TTS boshqaruvining yakuniy natijasini o‘lchaydi. Ushbu ajratish mobil ta’lim ilovasini oddiy computer vision demo-dasturidan metodik jihatdan farqlaydi.

Taklif etilgan tizimning yana bir muhim xususiyati - lug‘atni bosqichma-bosqich kengaytirish imkoniyati. O‘qituvchi yoki ishlab chiquvchi dastlab 10-15 ta maktab buyumidan boshlashi, keyin classroom objects, laboratory tools yoki robotics components kabi tematik to‘plamlarni qo‘shishi mumkin. Open-vocabulary modelda bu kengayish promptlar bilan boshqariladi. Shunga qaramay, yangi prompt qo‘shilishi avtomatik ravishda yuqori aniqlik kafolati bermaydi; semantik jihatdan yaqin sinflar

birgalikda sinovdan o'tkazilishi va threshold qayta kalibrlanishi kerak. Shu sababli tizimning pedagogik moslashuvchanligi texnik validatsiya jarayoni bilan birga yurishi zarur.

Mazkur arxitektura ta'lim jarayonida bir nechta amaliy ssenariyni qo'llab-quvvatlaydi. Boshlang'ich ssenariyda o'quvchi predmetni ko'rsatadi va inglizcha nomini eshitadi. Keyingi ssenariyda ovoz avtomatik aytilmasdan avval o'quvchidan nomni taxmin qilish so'ralishi, so'ng tekshirish tugmasi orqali to'g'ri javob eshittirilishi mumkin. Yana bir bosqichda topilgan obyektlar sessiya davomida shaxsiy lug'atga saqlanib, keyinchalik flashcard yoki mini-quiz shaklida takrorlanishi mumkin. Bunday funksiyalar LangEye kabi learner-driven yondashuvlar bilan uyg'un [8], biroq mazkur maqolada asosiy e'tibor obyekttni tanish va ovozli ifodalashning ishonchli texnik konturiga qaratilgan.

Umuman olganda, pilot tahlil OWLv2 modelining berilgan uchta tasvirdagi konkret chiqishlarini qayd etdi: ayrim aniq obyektlar mos label bilan topildi, shu bilan birga ground truth bilan tasdiqlanmagan qo'shimcha va ustma-ust deteksiyalar ham kuzatildi. Bu natija modelni Android ilovasiga integratsiya qilish mumkinligini texnik jihatdan o'rganish uchun boshlang'ich material beradi, biroq deteksiya sifati yoki jonli ishlash samaradorligi bo'yicha umumiy xulosa bermaydi. Keyingi eksperimental bosqich Android qurilmada real kamera oqimi va oldindan belgilangan annotatsiyalangan sinov protokoli bilan o'tkazilishi kerak.

Xulosa va takliflar

Tadqiqotda mobil kamera tasvirlarida o'quv qurollarini aniqlash va ularning inglizcha nomlarini ovozli ifodalashga mo'ljallangan Android-AI tizimining konseptual-texnik arxitekturasi ishlab chiqildi. Taklif etilgan arxitektura CameraX orqali tasvir olish, serverdagi OWLv2 open-vocabulary detektori orqali matnli promptlar asosida obyektlarni lokalizatsiya qilish, natijani filtrlash va Android Text-to-Speech orqali inglizcha nomni talaffuz qilish bosqichlarini birlashtiradi. Ushbu ishda to'liq jonli Android pipeline empirik baholanmagan; sinov qismi uchta statik rasmda AI detektorining chiqishlarini tekshirish bilan chegaralandi.

Pilot tajribada 1-rasmdagi ikkita qalam pencil sifatida 0,6561 va 0,6207 confidence ballari bilan, 2-rasmdagi qog‘oz teshgich esa hole punch sifatida 0,8401 ball bilan qayd etildi. Shu bilan birga 1-rasmda stapler (0,2934), 2-rasmda ruler (0,5405) va boshqa qo‘shimcha deteksiyalar ham hosil bo‘ldi. 3-rasmda pen, marker va pencil bo‘yicha ko‘plab ustma-ust deteksiyalar kuzatildi.

Keyingi bosqichda Android qurilmada jonli eksperiment o‘tkazib, har bir sinfni turli fon, masofa, yoritilish va occlusion sharoitlarida annotatsiyalangan kadrlar bilan baholash zarur. Deteksiya sifati Precision, Recall va mAP@0.5 bilan, foydalanuvchiga yetib boradigan yakuniy javob esa spoken-label correctness hamda end-to-end latency bilan o‘lchanishi lozim. Taklif etilgan 2/3 temporal tasdiqlash va ikki soniyalik cooldown parametrlarining samarasi ham aynan shu tajribada tekshirilishi kerak. OWLv2 ni Grounding DINO va YOLO-World bilan bir xil test to‘plamida qiyoslash texnik tanlovni mustahkamlaydi; pedagogik samaradorlikni baholash esa alohida nazoratli ta’lim tajribasini talab qiladi.

Foydalanilgan adabiyotlar ro‘yxati

1. Radford, A., Kim, J. W., Hallacy, C., et al. Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021, pp. 8748-8763. <https://proceedings.mlr.press/v139/radford21a.html>.
2. Minderer, M., Gritsenko, A., Stone, A., et al. Simple Open-Vocabulary Object Detection with Vision Transformers. Computer Vision - ECCV 2022, Part X, 2022, pp. 728-755. https://doi.org/10.1007/978-3-031-20080-9_42.
3. Minderer, M., Gritsenko, A., Hounsby, N. Scaling Open-Vocabulary Object Detection. Advances in Neural Information Processing Systems, 36, 2023, pp. 72983-73007. https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6d58fc68c0f3c36ae6e0e64478a69c0-Abstract-Conference.html.
4. Liu, S., Zeng, Z., Ren, T., et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. Computer Vision - ECCV 2024, 2024. https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6319_ECCV_2024_paper.php.

5. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y. YOLO-World: Real-Time Open-Vocabulary Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16901-16911. <https://doi.org/10.1109/CVPR52733.2024.01599>.
6. Ibrahim, A., Huynh, B., Downey, J., Höllerer, T., Chun, D., O'Donovan, J. ARbis Pictus: A Study of Vocabulary Learning with Augmented Reality. IEEE Transactions on Visualization and Computer Graphics, 24(11), 2018, pp. 2867-2874. <https://doi.org/10.1109/TVCG.2018.2868568>.
7. Weerasinghe, M., Biener, V., Grubert, J., Quigley, A. J., Toniolo, A., Čopič Pucihar, K., Kljun, M. VocabulARy: Learning Vocabulary in AR Supported by Keyword Visualisations. IEEE Transactions on Visualization and Computer Graphics, 28(11), 2022, pp. 3748-3758. <https://doi.org/10.1109/TVCG.2022.3203116>.
8. Shimabukuro, M., Panchal, D., Collins, C. LangEye: Toward “Anytime” Learner-Driven Vocabulary Learning From Real-World Objects. Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025), 2025, pp. 446-459. <https://doi.org/10.18653/v1/2025.bea-1.33>.
9. Android Developers. Image analysis with CameraX. Android Developers Documentation. <https://developer.android.com/media/camera/camerax/analyze> (murojaat sanasi: 01.10.2026).
10. Android Developers. TextToSpeech API reference. Android Developers Documentation. <https://developer.android.com/reference/android/speech/tts/TextToSpeech> (murojaat sanasi: 01.10.2026).
11. Mohamed, R. A. A. Enhancing academic English vocabulary learning through augmented reality and multimodal input. International Journal of Adolescence and Youth, 31(1), 2026, Article 2659778. <https://doi.org/10.1080/02673843.2026.2659778>.
12. Google. OWLv2 base patch16 ensemble model card. Hugging Face. <https://huggingface.co/google/owlv2-base-patch16-ensemble> (murojaat sanasi: 01.10.2026).