Meno kūrinių krypčių atpažinimas vaizdo įrašuose

Augustinas Jarockis
Informatikos institutas
Matematikos if informatikos fakultetas
Vilnius, Lietuva
augustinas.jarockis@mif.stud.vu.lt

Gita Gliaudytė
Informatikos institutas
Matematikos if informatikos fakultetas
Vilnius, Lietuva
gita.gliaudyte@mif.stud.vu.lt

***Santrauka*—Šiame darbe pristatome įrankį, skirtą automatiškai atpažinti meno kūrinių stilius vaizdo įrašuose. Panaudojome UNet modelį, kurį pasitelkėme vaizdų segmentavimui, ir ResNet50 modelį, apmokytą ArtBench duomenų rinkinyje, kurį pritaikėme iškarpytoms segmentų sritims klasifikuoti. Kūrimo metu gautas bendrinis įrankis, gebantis atpažinti meno kūrinių stilių pasiskirstymą laiko ašyje, prie paveikslo nurodant atitinkamą stiliaus pavadinimą. Siūlomas įrankis gali būti pritaikytas ne tik edukaciniams tikslams muziejuose ir galerijose, bet ir tolimesniems meno ir kultūros tyrimams dirbtinio intelekto kontekste.**

***Index Terms*—Video-Based Artwork Classification, Art Recognition, Dynamic Scene Analysis, Image Segmentation, Deep Learning, Neural Networks**

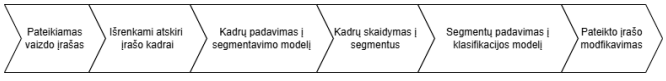
I. Įvadas

Pastaraisiais metais, dirbtinis intelektas vis dažniau taikomas meno analizės ir kultūrinio paveldo tyrimuose, tačiau iki šiol publikuotuose darbuose buvo orientuojamasi į statinių vaizdų analizę. Šiame darbe pristatome naują įrankį, leidžiantį identifikuoti meno kūrinių stilius vaizdo įrašuose. Tam panaudojome UNet modelį įrašo kadrų segmentavimui bei ResNet50 modelį paveikslų klasifikacijai. Sukurto įrankio išskirtinumas yra gebėjimas analizuoti dinamišką vaizdo turinį - automatiškai atrinkti reikšmingus kadrus, juose segmentuoti paveikslus ir priskirti jiems meno kryptį laiko ašyje, kas iki šiol buvo menkai ištirta.

Klasifikavimo modelio treniravimui pasitelktas ArtBench duomenų rinkinys [4]. Sukurto įrankio veiksmingumas buvo įvertintas naudojant FineVideo ir autorių sukurtą TenZorro Art vaizdo įrašų duomenų rinkinį. Įrankis buvo kuriamas Google Colab aplinkoje, naudojant Jupyter Notebook užrašines.

Taip pat pateikiamas galutinis įrankio variantas „Python“ aplinkoje su naudotojui patogia naudotojo sąsaja. Sukurtas kodas ir modeliai yra pasiekiami šioje nuorodoje.

II. Metodai



1 pav. – Įrankio veikimo procesas

Sukurtas meno kūrinių atpažinimo įrankis apdoroja jam pateiktą vaizdo įrašą ir pažymi jame matomus meno kūrinius bei identifikuoja jų meno kryptį. Visas procesas atliekamas automatiškai ir po kelių minučių įrankis grąžina vaizdo įrašo kopiją su jame pažymėtais meno kūriniais. Šis įrankis darbą atlieka 5 žingsniais (1 pav.): kadrų išrinkimas iš pateikto vaizdo įrašo, pasirinktų kadrų segmentavimas siekiant nustatyti paveikslų vietas, kadro sukarpymas pagal segmentuotas sritis, paveikslų klasifikavimas ir, galiausiai, paveikslo dalių ir nustatytų meno krypčių pažymėjimas originaliame įraše.

A. Kadrų išrinkimas

Vaizdo įrašai yra sudaryti iš daugybės kadrų, kurių dauguma yra panašūs vienas į kitą. Siekiant atpažinti meno kūrinius unikaliuose video įrašo kadruose, buvo reikalingas metodas jiems efektyviai atrinkti. Kadrų išrinkimas vykdomas šiuo algoritmu: paimamas pirmasis įrašo kadras, jis įtraukiamas į atrinktų kadrų rinkinį. Sekantys kadrai paimami fiksuotais laiko intervalais. Kiekvienas toks kadras yra lyginamas su jau anksčiau paimtu kadru, pasitelkiant struktūrinio panašumo (angl. *structural similarity*, toliau - SSIM) metriką (1), įgyvendintą *structural_similarity* Python funkcija iš *skimage.metrics* bibliotekos. Jei SSIM reikšmė yra mažesnė už nustatytą slenkstinę reikšmę, kadras laikomas pakankamai nepanašiu į prieš tai buvusį ir yra įtraukiamas į atrinktų kadrų rinkinį. Kitu atveju, kadras praleidžiamas. Šis procesas kartojamas tol, kol baigiasi vaizdo įrašas.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (1)$$

kur:

- μ_x, μ_y – vidutinės reikšmės paveikslėliuose x ir y;
- σ_x^2, σ_y^2 – dispersijos;
- σ_{xy} – kovariacija tarp x ir y;
- C_1, C_2 – stabilizavimo konstantos.

B. Kadrų segmentavimas naudojant UNet

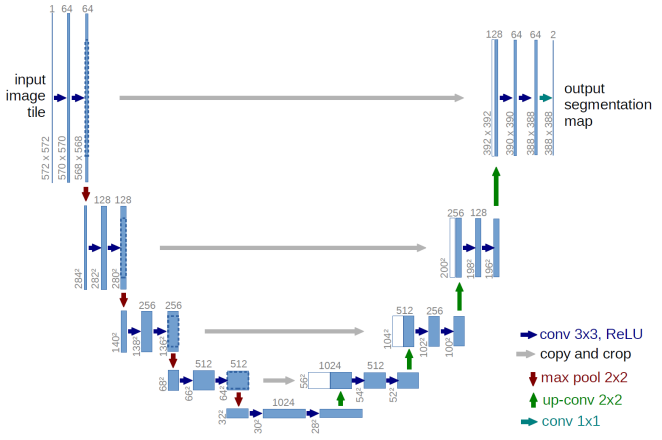
Norint tiksliau nustatyti, kur kadruose yra paveikslai, buvo pasitelktas segmentavimo neuroninių tinklų modelis. Naudojant jį, gauname segmentavimo kaukes, kuriose yra pažymėtos paveikslų vietos. Modeliui sukurti buvo panaudota UNet architektūra (2 pav.). UNet pasižymi savo U formos architektūra, sudarytą iš simetriško suspaudimo ir išplėtimo kelio bei praleidimo jungčių (angl. *skip connections*), kurios padeda išlaikyti svarbią erdvinę informaciją esant ribotam duomenų kiekiui [5].

Modelio optimizacija buvo įgyvendinta naudojant dvejetainio kryžminio entropijos nuostolio funkciją su logitais (2). Ši

funkcija leidžia dirbti su logitais dvejetainėje klasifikacijoje, todėl užtikrina stabilų mokymąsi.

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \qquad (2)$$

Modelis buvo treniruotas naudojant OpenImagesV7 duomenų rinkinio paveikslų rėmų (angl. *picture frame*) nuotraukas. Taip pat buvo naudojamos ir dažnai pasitaikančių su paveikslais objektų nuotraukos, siekiant sumažinti klaidingai teigiamų (angl. *false positive*, toliau - FP) reikšmių kiekį. Tokių paveikslėlių tikrosios reikšmės (angl. *ground truth*) kaukėms buvo naudojami tušti, visiškai juodi paveikslėliai.



2 pav. – UNet modelio architektūra

C. Kadrų išskaidymas į segmentus

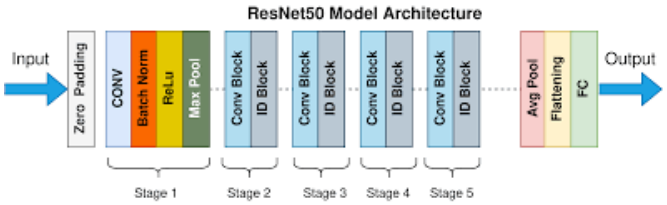
Po kadrų segmentavimo, paveikslėliai yra išskaidomi į atskirus segmentus, kuriuose yra po vieną paveikslą. Skaidymui naudojama OpenCV biblioteka [1], kuri identifikuoja atskirus segmentacijos kaukės segmentus, apskaičiuoja minimalų statmeną stačiakampį, į kurį telpa kiekvienas segmentas, ir iškerpa tą stačiakampį iš originalaus kadro. Taip pat buvo atmesti tie stačiakampiai, kurių plotas yra mažesnis už dinamiškai pagal kadro dydį nustatytą ribą, laikant, kad tokie segmentai greičiausiai atitinka FP reikšmes. Po šio etapo gaunamas didelis kiekis paveikslėlių, iškirptų iš originalių kadrų, kuriuose kiekviename yra po vieną paveikslą.

D. Segmentų klasifikavimas naudojant ResNet50

Suskaidyti paveikslo segmentai yra paduodami mūsų klasifikacijos modeliui. Klasifikacijos modelis priskiria kiekvienam segmentui vieną iš 10 klasių. Klasės yra paimtos pagal klases, pateiktas ArtBench duomenų rinkinyje [4]. Modeliui sukurti naudojome ResNet50 architektūrą (3 pav.). ResNet50 - tai gilus konvoliucinis neuroninis tinklas, sudarytas iš 50 sluoksnių, kuris naudoja likutines jungtis (angl. *residual connections*), kad užtikrintų mokymąsi be gradientų išnykimo [3]. Modelis buvo optimizuojamas naudojant kryžminės entropijos nuostolio funkciją (angl. *cross entropy loss*) (3). Tai yra standartinė funkcija sprendžiant kelių klasių klasifikavimo uždavinius.

$$\text{CrossEntropyLoss} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{z_{i,y_i}}}{\sum_{j=1}^{C} e^{z_{i,j}}}\right) \qquad (3)$$

Modelį treniravome pagal visus 50000 ArtBench duomenų rinkinyje esančių paveikslėlių.



3 pav. – ResNet50 modelio architektūra

E. Vaizdo įrašo modifikavimas

Galiausiai, atlikus pirmus keturis žingsnius, yra modifikuojamas vaizdo įrašas, jame pažymimi rasti paveikslai ir jiems priskirtos klasės. Einama per kiekvieną vaizdo įrašo kadrą ir tikrinama, tarp kurių dviejų atrinktų kadrų yra einamasis kadras ir tam kadrui pritaikoma pirmesniojo kadro segmentacijos kaukė. Taip prie kiekvieno tame išrinktame kadre surasto segmento, kuriame yra paveikslas, yra parašomas suklasifikuotos meno krypties pavadinimas. Šis procesas yra pritaikomas kiekvienam kadrui. Pasiekus vaizdo įrašo pabaigą, modifikuotas vaizdo įrašas yra pateikiamas įrankio naudotojui. Vaizdo įraše yra matomi pažymėti paveikslai bei parašyta kiekvieno paveikslo meno kryptis.

Pasirinkto metodo taikymas sujungia vaizdų segmentavimą, klasifikavimą bei demonstruoja jų taikymo galimybes sprendžiant kompleksines užduotis vaizdo įrašuose. Kompleksiškumas analizuojant vaizdo įrašus su meno kūriniais kyla dėl netolygaus apšvietimo, kintančios kūrinių sekos ir skirtingų kampų. Dėl to šis metodas prisideda prie gilesnio dinaminių vaizdų analizės sistemų supratimo, ypač meno duomenų kontekste.

III. Duomenys

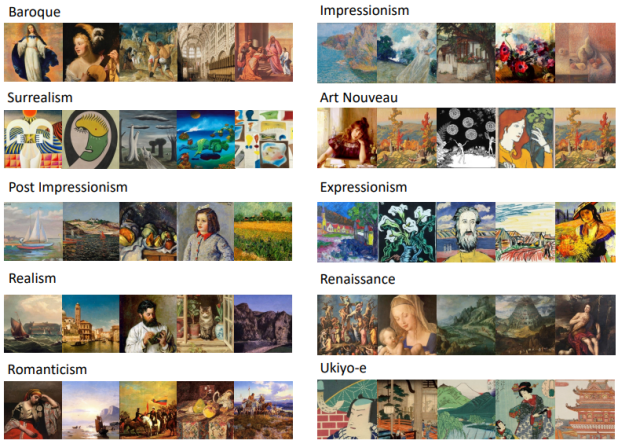
A. OpenImagesV7

OpenImagesV7 - tai Google sukurtas didelės apimties vaizdo duomenų rinkinys, skirtas objektų atpažinimui ir vaizdų klasifikacijai [6]. Jame pateikiamos aukštos kokybės anotacijos, pavyzdžiui, segmentavimo kaukės.

Šiame duomenų rinkinyje esančia paveikslo rėmo (angl. *picture frame*) klase buvo apmokytas segmentavimo modelis. Norint pagerinti tikslumą, iš to paties rinkinio buvo parsiųsta nuotraukų su dažnai nuotraukose su paveikslais pasitaikančiais objektais, jiems suteikiant visiškai tuščias kaukes. Tokių nuotraukų klasės buvo dėžė (angl. *box*) ir dokumentų spintelė (angl. filing cabinet).

B. ArtBench

Modelių treniravimui naudojome ArtBench-10 duomenų rinkinį [4]. Rinkinį sudaro 60000 nuotraukų, iš kurių 50000 yra skirtos mokymui, o likusios yra skirtos validacijai. Nuotraukos yra suskirstytos į 10 klasių, atitinkančių meno kryptis (4 pav.), o kiekviena klasė turi po lygiai 6000 nuotraukų. Šis subalansuotas duomenų rinkinys pasižymi aukšta duomenų kokybe, kiekvienoje nuotraukoje yra aiškiai matomas konkrečios klasės paveikslas.



4 pav. – 10 meno stilių ir jiems atitinkančių vaizdų pavyzdžiai iš ArtBench-10 duomenų rinkinio [4].

C. FineVideo

FineVideo duomenų rinkinys - tai didelės apimties, gausiai komentuotų vaizdo įrašų rinkinys [2]. Jį sudaro daugiau kaip 43000 YouTube vaizdo įrašų, apimančių 122 kategorijas (5 pav.).

Įrankio testavimui buvo pasirinktas meno parodų (angl. *art exhibitions*) duomenų poaibis, kuriame yra parodų apžvalgų, menininkų pristatymų ir ekskursijų filmuotos medžiagos, tinkančios tirti meną dinamiškoje aplinkoje.

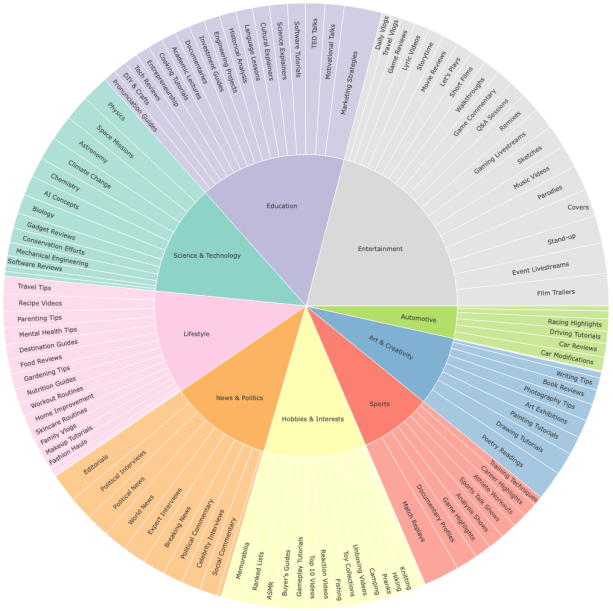
Šiame poaibyje taip pat buvo pastebėta neatitikimų - jį sudarė ir su parodomis nesusiję video įrašai, nesuteikiantys vertės tyrimui. Identifikavus šiuos vaizdus, jie buvo nebenaudojami įrankio testavimo procese.

D. TenZorro Art

Įvairiapusiškai įvertinti įrankio veikimą, testavimo procese buvo sukurtas TenZorro Art duomenų rinkinys. Šį rinkinį sudaro 16 vaizdo įrašų, kurie buvo filmuoti meno galerijose, muziejuose. Įrašų trukmė svyruoja nuo 10 sekundžių iki 7 minučių. Tai leidžia patikrinti įrankio veikimo spartą skirtingo duomenų kiekio atžvilgiu. Atrenkami vaizdo įrašai buvo viešai pasiekiami internete, tokiose platformose kaip „YouTube“ ir „TikTok“. Jie buvo panaudoti tik mokymosi tikslais. TenZorro Art duomenų rinkinys yra pasiekiamas šioje nuorodoje.

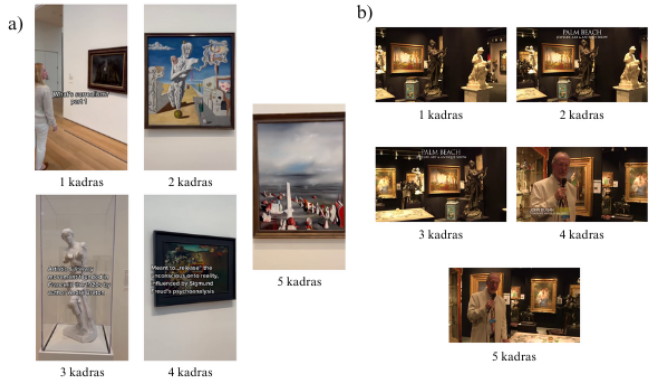
IV. Rezultatai

Panaudoję dviejų skirtingų architektūrų – UNet bei ResNet50 – modelius ir OpenCV biblioteką, sukūrėme įrankį,



5 pav. – FineVideo duomenų rinkinio kategorijos ir jų subkategorijos [2].

leidžiantį vaizdo įrašuose atpažinti, pažymėti bei klasifikuoti matomus paveikslus. Taip pat sukūrėme du neuroninių tinklų modelius: vieną paveikslų segmentavimo, o kitą – paveikslų klasifikavimo pagal meno kryptis.



6 pav. – 5 iškirpti kadrai, išrinkti naudojant SSIM: a) TenZorro Art duomenų rinkinio vaizdo įraše, b) FineVideo duomenų rinkinio vaizdo įraše.

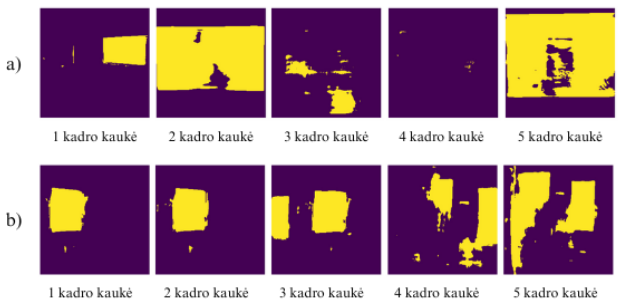
Galutinis įrankis geba išsirinkti svarbiausius kiekvieno vaizdo įrašo kadrus. TenZorro Art duomenų rinkinio vaizdo įraše buvo 555 kadrai, tačiau taikant panašumo paiešką (angl. *similarity search*) ir tikrinant kas 9 kadrus, buvo patikrinti 61 kadras ir atrinkti tik 5 esminiai - tai sudaro apie 0,9 proc. visų kadrų. Šis skaičius gali skirtis, priklausomai nuo vaizdo įrašo turinio – vaizdo įrašai, kuriuose matomas vaizdas dažnai ir greitai keičiasi, turės daugiau atrinktų kadrų. Pavyzdžiui, vaizdo įraše iš FineVideo duomenų rinkinio, turinčiame 1309 kadrus, buvo patikrinti 62 kadrai (kas 21-as kadras), o iš jų

atrinkti tik 29, t.y. apie 2,2 proc. visų kadrų (6 pav.).

7 pav. – Segmentacijos modelio treniravimo ir validacijos aibių nuostolių reikšmės treniravimo metu.

Pastebėta, kad toks metodas kadrams atrinkti padeda ženkliai sumažinti laiką, kurį įrankis užtrunka apdorodamas vaizdo įrašą. FineVideo vaizdo įrašo analizė su panašumo paieškos ir praleidžiamų kadrų metodu užtruko 72 sekundes, kai tuo tarpu bandant analizuoti įrašą tik su praleidžiamais kadrais (be panašumo paieškos), analizė užtruko 146 sekundes, t.y. apie 2 kartus ilgiau.

Sukurtasis segmentacijos modelis geba aptikti paveikslų rėmus nuotraukose. Modelis, kurtas pagal UNet architektūrą, buvo mokomas beveik 40 epochų, iki kol įvyko ankstyvasis stabdymas (angl. *early stopping*). Pasilikta naudoti buvo 30-osios epochos modelio versija, kuri grąžino tiksliausias kaukes (7 pav.). Šio modelio tikslumas (angl. *accuracy*) yra 82 proc., atkūrimas (angl. *recall*) – 81 proc., precizija (angl. *precision*) – 79 proc., F1 metrika – 79 proc., o IoU (*Intersection over Union*) metrika – 68 proc. Metrikos rodo, kad modelis patikimai aptinka paveikslus, tačiau maža IoU metrika gali reikšti, jog kaukės ne visada preciziškai atitinka tikrąsias objektų formas. Tai gali nutikti dėl skirtingo kampo nuotraukų, apšvietimo ir paveikslus užstojančių žmonių.

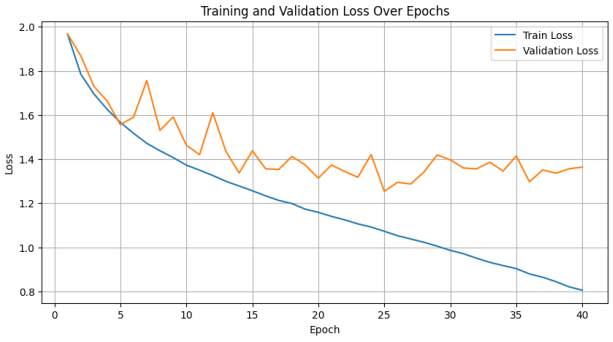


8 pav. – 5 kaukės, gautos po kadrų segmentavimo: a) TenZorro Art duomenų rinkinio vaizdo įraše, b) FineVideo duomenų rinkinio vaizdo įraše.

Sukurtas įrankis analizuoja atrinktus kadrus, naudodamas šį segmentacijos modelį, kuris kiekvienam kadrui sugeneruoja segmentacijos kaukę, pažyminčią modelio aptiktus paveikslus. (8 pav.). Remdamasis šiomis kaukėmis, pasitelkęs OpenCV biblioteką, įrankis suskaido kiekvieną kadrą į segmentus, kuriuose yra matomi paveikslai (paveikslų vietos yra nustatomos pagal segmentacijos kaukes) (9 pav.).

9 pav. – 5 kadrų segmentai, iškirpti pagal kaukes: a) TenZorro Art duomenų rinkinio vaizdo įraše, b) FineVideo duomenų rinkinio vaizdo įraše.

Klasifikacijos modelis, sukurtas pagal ResNet50 architektūrą, buvo treniruojamas 40 epochų. Šiam modeliui treniruoti taip pat buvo panaudotas ankstyvasis stabdymas (10 pav.). Mes naudojome 25-osios epochos modelį, nes šio modelio nuostolis buvo pats mažiausias. Šis modelis pasiekė vidutinį 91 proc. tikslumą (angl. *accuracy*), 56 proc. atkūrimą (angl. *recall*), 56 proc. preciziją (angl. *precision*) bei 56 proc. F1 metrikas (I lent.). Šie rezultatai rodo, kad modelis gerai atlieka bendrą paveikslų klasifikavimo uždavinį, tačiau vis dar kyla iššūkių tiksliai atpažinti tam tikrą meno kryptį.



10 pav. – Klasifikacijos modelio treniravimo ir validacijos aibių nuostolių reikšmės treniravimo metu.

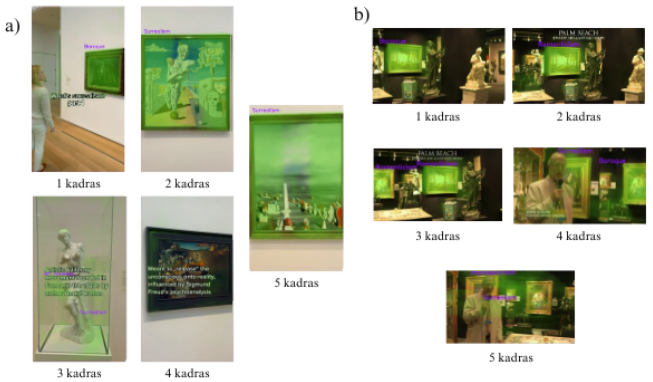
Rezultatų analizė parodė, kad modelio veikimas priklauso nuo konkretaus meno stiliaus. Pavyzdžiui, Ukijo-e stilius buvo aptiktas beveik tobulai (F1 metrika - 95 proc.), greičiausiai dėl unikalių šio meno bruožų. Tuo tarpu labiau vizualiai ir stilistiškai panašūs meno stiliai, kaip romantizmas (F1 - 41

proc.) ir realizmas (F1 - 35 proc.), kėlė modeliui iššūkį, klaidino. Tokius rezultatus galima paaiškinti tuo, kad kūrinių priskyrimas konkrečioms meno kryptims yra sudėtinga užduotis, nes pavieniai darbai dažnai turi kelioms kryptims būdingų bruožų, o pačios kryptys neretai persidengia ir neturi aiškių ribų. Net ir srities ekspertams gali būti sunku tiksliai nustatyti meno kryptį, todėl natūralu, kad šioje užduotyje iššūkių patiria ir neuroniniai tinklai.

Klasė	Tikslumas	Atkūrimas	Precizija	F1
Art Nouveau	0.91	0.45	0.56	0.50
Barokas	0.93	0.74	0.65	0.69
Ekspresionizmas	0.89	0.46	0.43	0.45
Impresionizmas	0.88	0.47	0.43	0.45
Postimpresionizmas	0.89	0.49	0.46	0.47
Realizmas	0.87	0.35	0.36	0.35
Renesansas	0.94	0.66	0.69	0.68
Romantizmas	0.90	0.35	0.50	0.41
Siurealizmas	0.92	0.72	0.56	0.63
Ukijo-e	0.99	0.92	0.97	0.95
Vidurkis	0.91	0.56	0.56	0.56

I lentelė – Klasifikacijos modelio metrikos atskiroms klasėms.

Nepaisant to, šis klasifikacijos modelis ganėtinai patikimai geba atpažinti meno kryptį tais atvejais, kai paveikslas yra gerai matomas, aiškiai priklauso kokiai nors krypčiai ir yra aiškiai matomas, neužstojamas, pavyzdžiui, vaikštančių žmonių.



11 pav. – 5 modifikuoti kadrai: a) TenZorro Art duomenų rinkinio vaizdo įraše, b) FineVideo duomenų rinkinio vaizdo įraše.

Galiausiai, pagal surinktą informaciją, mūsų įrankis modifikuoja originalų vaizdo įrašą, jame pažymėdamas aptiktus paveikslus (pagal segmentacijos kaukes) bei prie kiekvieno segmento parašydamas tekstu, kokiai meno srovei priklauso pažymėtasis paveikslas (11 pav.).

V. Išvados

Šio darbo rezultatas - automatinis įrankis, gebantis atskirti meno kūrinių kryptis vaizdo įrašuose. Jis buvo įgyvendintas remiantis dviejų giliųjų neuroninių tinklų architektūromis: UNet segmentacijai ir ResNet50 klasifikacijai. Galutinis įrankis atpažįsta meno kūrinius laiko ašyje ir priskiria jiems atitinkamus stilius.

Rezultatai parodė, kad:

- Segmentacijos modelis pasiekė 82 proc. tikslumą ir 79 proc. F1 metriką. Tai reiškia, jog paveikslai vaizdo įrašų kadruose yra aptinkami patikimai.
- Klasifikacijos modelis pasiekė 91 proc. tikslumą, bet vidutinė F1 metrikos reikšmė siekė tik 56 proc. Tai rodo, jog susidurta su sunkumais teisingai klasifikuoti meno kūrinius į meno kryptis.
- Panašumo paieškos naudojimas atrenkant kadrus segmentavimui, sutrumpina analizės laiką apie 2 kartus.

Nepaisant tam tikrų trūkumų, įrankis potencialiai gali prisidėti ne tik prie edukacijos ir apsilankymo muziejuose bei galerijose patirties gerinimo, bet ir prie platesnių kultūros analizės tyrimų. Dinaminio meno ir kultūrinio turinio atpažinimas vaizdo įrašuose tebėra siaurai išanalizuota sritis, todėl šis darbas gali tapti reikšmingu pagrindu tolimesniems darbams, kurie orientuotųsi į neuroninių tinklų taikymo plėtrą vizualiai sudėtinguose vaizdo analizės scenarijuose. Tolimesni tyrimai galėtų apimti klasifikuojamų stilių klasių praplėtimą, pažangesnių segmentavimo ir klasifikavimo modelių pritaikymą, procesų lygiagretinimą greitesnei analizei ir duomenų rinkinių balanso gerinimą.

Literatūra

[1] Ivan Culjak, David Abram, Tomislav Pribanic, Hrvoje Dzapo, and Mario Cifrek. A brief introduction to opencv. In *2012 Proceedings of the 35th International Convention MIPRO*, pages 1725–1730, 2012.

[2] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404*, 2022.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

[6] Farjana Yesmin. Bias detection and fairness analysis in object detection and image classification using open images v7, 2023.