UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE

FAKULTA PRÍRODNÝCH VIED A INFORMATIKY

ROZPOZNÁVANIE OBJEKTOV VO VIDEÁCH

DIPLOMOVÁ PRÁCA

2024 **Bc. Johana Heneková**

UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE

FAKULTA PRÍRODNÝCH VIED A INFORMATIKY

ROZPOZNÁVANIE OBJEKTOV VO VIDEÁCH

DIPLOMOVÁ PRÁCA

Študijný odbor:	18. Informatika
Študijný program:	Aplikovaná informatika
Školiace pracovisko:	Katedra informatiky
Školiteľ:	Mgr. Ľubomír Benko, PhD.

Nitra 2024 Bc. Johana Heneková

304504

Univerzita Konštantína Filozofa v Nitre
Fakulta prírodných vied a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Johana Heneková
Študijný program: aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: Diplomová práca
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Rozpoznávanie objektov vo videách

Anotácia: Výskumníci a vývojári v oblasti informačných technológií sa inšpirovali funkciou ľudského zraku a rozhodli sa vytvoriť aplikácie na detekciu videoobjektov, ktoré poskytujú strojom schopnosť analyzovať obrázky a zisťovať objekty, ktoré sa v nich nachádzajú. Najprv vyvinuli protokoly a postupy určené na fungovanie len na obrázkoch. Dnes sa však veci posunuli k video-obrazom. Cieľom takéhoto nástroja je umožniť stroju lokalizovať, identifikovať a klasifikovať objekty, ktoré možno vidieť na vstupných pohyblivých obrázkoch. Stroje a ich postupy neberú do úvahy obrazy ako celok. Na ich analýzu musia snímky rozdeliť a pracovať s pixelmi a ich vlastnosťami. Zvyčajne kombinujú detekciu obrazu a videostopy, aby prišli so svojimi výsledkami.

Cieľom záverečnej práce je predstaviť rôzne prístupy k rozpoznávaniu objektov vo videách a vyhodnotiť ich úspešnosť. V teoretickej časti je žiadúce sa zamerať na rôzne algoritmy rozpoznávania objektov vo videách. V praktickej časti je žiadúce vybrať vhodnú testovaciu sadu videí, na ktorých sa porovnajú a vyhodnotia rôzne prístupy k rozpoznávaniu objektov.

Charakter práce:
Výskumný – stanovenie predpokladov/hypotéz, metodika výskumu, výsledky výskumu (štatistická interpretácia), interpretácia výsledkov výskumu (vecná interpretácia).

Predmetové prerekvizity:
Úvod do strojového učenia (1., mgr);
Neurónové siete (1., mgr);
HĺBková analýza dát (2., mgr).

Najdôležitejšie kompetentnosti získané spracovaním témy:
vykonávať vedecký výskum;
princípy umelej inteligencie;
poskytnúť vizuálnu prezentáciu údajov;
referovať o výsledkoch analýzy;
vykonať analýzu údajov.

Školiteľ: Mgr. Ľubomír Benko, Ph.D.
Oponent: Mgr. Janka Pecuchová, PhD.
Katedra: KI - Katedra informatiky
Dátum zadania: 31.10.2022

Dátum schválenia: 07.03.2024

RNDr. Ján Skalka, PhD., v. r.
vedúci/a katedry

POĎAKOVANIE

Na tomto mieste by som sa chcela poďakovať školiteľovi mojej diplomovej práce Mgr. Ľubomírovi Benkovi, PhD za cenné rady, konzultácie a spätnú väzbu. Taktiež mojej rodine a kamarátom za nekonečnú podporu.

ABSTRAKT

HENEKOVÁ, Johana: Rozpoznávanie objektov vo videách. Diplomová práca. Univerzita Konštantína Filozofa v Nitre. Fakulta prírodných vied a informatiky. Školiteľ: Mgr. Ľubomír Benko, PhD. Stupeň odbornej kvalifikácie: Magister odboru Aplikovaná informatika. Nitra: FPVaI, 2024. 67 s.

Diplomová práca sa zaoberá problematikou klasifikácie vo videách. Hlavným cieľom práce je otestovanie predtrénovaných modelov na datasete ERA a analyzovať výsledky na základe ich správnosti kategorizácie. Klasifikáciu popisujeme na konkrétnom datasete ERA, ktorý obsahuje videá vo forme dronových záberov vo viacerých oblastiach ako športy, poľnohospodárstvo alebo pohromy. V metodike porovnávame aj iné datasety ako napríklad Sport1M, UCF101 alebo Kinetics, ktoré sa využívajú pri klasifikácii. V práci sa primárne zameriavame na modely C3D, I3D, P3D a TRN. Popri testovaní sme stanovili hypotézy, napríklad predpoklady pre lepšie alebo horšie klasifikovanie modelov na základe ich trénovania. Pre testovanie sme využili programovací jazyk Python a jeho knižnice. Výsledky testovania sa vo všeobecnosti nezhodujú s výsledkami autorov ERA, okrem modelu C3D, ktorý sa približuje najviac. Z tohto dôvodu v práci taktiež popisujeme optimalizačné metódy ako napríklad transformácie alebo vzorkovacia metóda. Popri datasetoch analyzujeme aj architektúry ResNet a Inception, ktoré využívajú predtrénované modely. Výsledky spracovávame v programe Excel, ktorý obsahuje porovnania, percentuálnu úspešnosť a grafické vizualizácie. Medzi výsledky práce paria okrem vyhodnotení hypotéz taktiež súbor odporúčaní a využitie našej práce ako podkladový materiál pre výučbu a prácu s predtrénovanými modelmi.

Kľúčové slová: Datasety. Klasifikácia vo videách. Neurónové siete. Optimalizačné metódy. Architektúry modelov.

ABSTRACT

HENEKOVÁ, Johana: Object recognition in videos. [Master Thesis]. Constantine the Philosopher University in Nitra. Faculty of Natural Sciences and Informatics. Supervisor: Mgr. Ľubomír Benko, PhD. Degree of Qualification: Master of Applied Informatics. Nitra: FNSaI, 2024. 67 p.

The diploma thesis deals with the issue of classification in videos. The main goal of the work is to test the pre-trained models on the ERA dataset and analyze the results based on their classification accuracy. We describe the classification on the specific ERA dataset, which contains videos in the form of drone footage in several areas such as sports, agriculture or disasters. In the methodology, we also compare other datasets such as Sport1M, UCF101 or Kinetics, which are used in the classification. In our work, we primarily focus on C3D, I3C, P3D and TRN models. In addition to testing, we establish hypotheses, for example assumptions for better or worse classification of models based on their training sources. For testing, we use the programming language Python and its libraries. The results of the testing generally do not agree with the ERA authors results, except for the C3D model, which came closest. For this reason, the work also describes optimalization methods such as transformations of the sampling method. In addition to the datasets, we also analyze the ResNet and Inception architectures, which use pre-trained models. We process the results in the Excel program, that includes comparisons, percentage of the success and graphic visualizations. In addition to the evaluation of the hypotheses, the results of the work include a set of recommendations and the use of our work as background material for teaching and working with pre-trained models.

Keywords: Datasets. Classification in videos. Neural networks. Optimalization methods. Architectures of models.

OBSAH

ÚVOD

Klasifikácia vo videách je v dnešnej dobe napredujúce odvetvie umelej inteligencie, ktoré nachádza využitie vo veľkom množstve oblastí nášho života. Úlohou tejto práce je porozumieť danej oblasti a priniesť príklady využitia.

V súčasnej dobe je dobre známe využitie automatizovanej kategorizácie v oblastiach ako autonómne autá, individuálne odporúčanie obsahu, identifikácia tvárí alebo v technológiách bezpečnostných kamier. Rozvíja sa avšak aj v mnohých iných sektoroch ako napríklad poľnohospodárstvo, kde sa najmä upriamuje na monitorovanie rastlín a prepája sa s monitorovaním počasia. (Khan a AlSuwaidan, 2022) V medicíne má za cieľ primárne zefektívniť a urýchliť prácu lekárov. (Zhang et al., 2021) Taktiež v našej práci spomíname využitie v športoch, kde sa za pomoci klasifikácie a detekcie objektov vo videách analyzujú športovci a predikujú ich nasledujúce kroky. (Delextrat a Cohen, 2009) Naprieč všetkými odvetviami, ktoré využívajú klasifikáciu, sa upriamuje vývoj na hlavný cieľ a teda priniesť riešenia, ktoré nám uľahčia každodenný život a vytvoria nám spoľahlivú pomoc.

V tejto práci sa zameriavame na oblasť počítačového videnia pri videách a teda porozumenia digitálnym dátam a kontextu. Popisujeme modely, ktoré sa rozvinuli vďaka pokroku v tejto sfére neurónových sietí a prinášajú automatické spracovanie s neuveriteľnou presnosťou a efektivitou.

Medzi hlavné oblasti, kde sa rozvíja klasifikácia vo videách patria hlavne konvolučné a rekurentné neurónové siete. Na konkrétnych modeloch popisujeme okrem ich architektúry a princípu fungovania aj ich využitie a testujeme ich na datasete.

Zameriavame sa na prácu s datasetom ERA, ktorý priniesol dronové videá ako primárny zdroj a tým ponúkol nové využitie a smer v klasifikácií.

Práca prináša prínos najmä z pohľadu absencie podrobného popisu postupu práce s predtrénovanými modelmi C3D, I3D, P3D a TRN, ktoré natrénovali autori Mou et al. (2020) a taktiež porovnania ich výsledkov a optimalizácie.

1 ANALÝZA SÚČASNÉHO STAVU

Postupná revolúcia technológií priniesla rozvoj oblasti umelého videnia, taktiež známeho ako computer vision. Táto disciplína umelej inteligencie sa stala súčasťou mnohých odvetví a to najmä v oblasti spracovania obrazov a analýzy videí. Z využitia v reálnom svete spomenieme napríklad prácu Kortylewski et al. (2020), ktorá popisuje implementovanie do autonómnych áut, ktoré potrebujú identifikovať objekty a ich trajektórie v reálnom čase. Ďalším využitím, priamo zameraným na videá ľudí, je napríklad rozpoznávanie jednotlivých úkonov chirurga, ktoré prináša veľký prínos v medicíne. (Zhang et al., 2021)

1.1 COMPUTER VISION

K postupne rozširujúcemu sa využívaniu a dostupnosti kamier a telefónov vznikajú aj veľké množstvá zdieľaných videí. To prinieslo nový dopyt pre modely hlavne s cieľom, aby sa učili a vylepšovali automaticky. Taktiež sa modely zameriavajú na znovu použiteľnosť a tým sa stávajú univerzálne, aby sa nemuseli budovať od začiatku. Dôležitým znakom počítačového videnia je časovo priestorový rámec, ktorý pridáva kontext neurónovým sieťam. (Qu et al., 2021)

Počítačové videnie spracováva jednoduché popisy a teda, čo vidí na scéne. Hľadá vzory, aby zistilo obsah jednotlivých snímok. Vzory obsahujú pochopenie scény a tie sa využívajú v modeloch. (Zhou et al., 2022)

Algoritmy modelov vieme rozdeliť podľa Zhou et al. (2022) do kategórií:

- detekcia,
- rozpoznanie,
- segmentácia,
- a 3D rekonštrukcia.

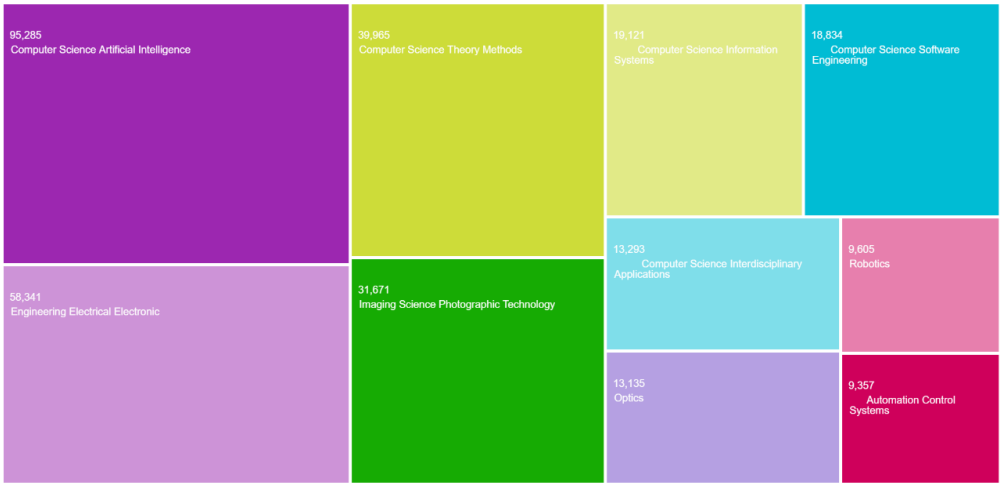
Počítačové videnie, ako obor umelej inteligencie, vieme datovať podľa prvých teoretických konceptov medzi roky 1943 a 1956, kedy cieľom výskumu bolo hlavne porozumieť strojovému učeniu. Prvá konkrétna práca vznikla na MIT s názvom: The Summer vision Project (Papert, 1966). V posledných desaťročiach počítačové videnie napreduje hlavne vďaka posunu v technológiách, ktoré umožňujú vyššiu výpočtovú náročnosť. (Sharma et al., 2021)

Rozvoj počítačového videnia už od 50tych rokov minulého storočia výrazne ovplyvnil rôzne oblasti. Jednou významnou oblasťou je priemyselný sektor, kde sa

počítačové videnie čoraz viac využíva pre inteligentnú výrobu v kontexte bezdrôtového pripojenia 5G a priemyslu 4.0 ako popísal Li (2022).

Táto technológia tiež zohrala kľúčovú úlohu v automobilovom priemysle, najmä v oblasti bezpečnosti a identifikácie vozidiel, ako aj pri rozpoznávaní a klasifikácii plodín v poľnohospodárstve. (Salazar a Kurka, 2020)

Z portálu Web of Science vidíme, že pojem „computer vision“ je veľmi aktuálnou témou s veľkým množstvom výskumných prác v rôznych odvetviach (Obrázok 1).



Obrázok 1 Web of Science diagram vyhľadávania pojmu "computer vision"[1]

1.2 ZÁKLADNÉ METÓDY COMPUTER VISION

Pri vyhodnocovaní scén sa nemôžu modely upriamovať iba na cieľovú udalosť alebo objekt. Je dôležité rozoznávať kontext a prepájať obsah medzi vizuálnymi ako aj nevizuálnymi objektami a udalosťami počas celého videa. (Wang a Zhu, 2023)

Wang a Zhu (2023) rozlišujú tri druhy kontextu v počítačovom videní a to:

- priestorový kontext,
- časový kontext,
- alebo iný kontext.

Na kontext, ako popísali Wang a Zhu (2023) vo svojom článku o kontexte pochopenia v počítačovom videní, sa potom dá pozerať na rôznych úrovniach a to menovito na:

- úroveň predchádzajúcich znalostí,
- úroveň globálnych znalostí

[1] Zdroj Obrázok 1: https://www.webofscience.com/wos/woscc/basic-search

- alebo taktiež úroveň lokálnych znalostí.

Následne vieme deliť jednotlivé modely podľa týchto kategórií. (Wang a Zhu, 2023)

1.3 MODELY KLASIFIKUJÚCE VO VIDEÁCH

Modely na klasifikáciu videí opisujú Ng et al. (2015) ako algoritmy alebo architektúry navrhnuté tak, aby kategorizovali a pochopili kontext videí. Tieto modely využívajú rôzne techniky ako napríklad hĺbkové učenie pre klasifikáciu. A to na základe vizuálnych vlastností, čiastočnej informácie alebo textového podkladu.

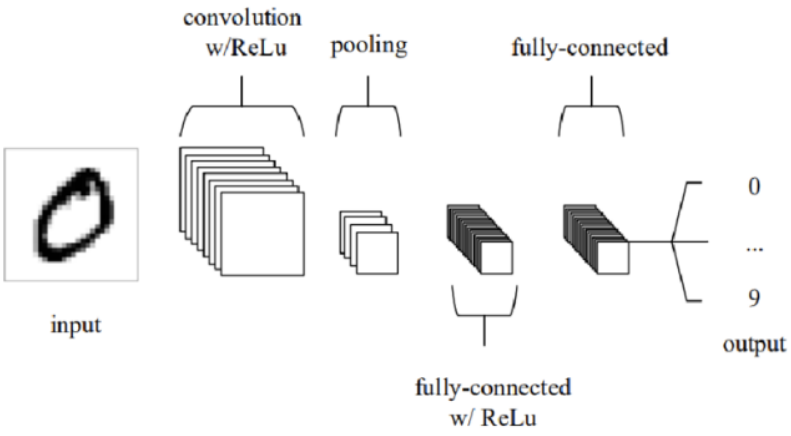
1.3.1 CNN

Hĺbkové učenie sa stalo silným nástrojom pre klasifikáciu vo videách. Využíva sa tu najmä CNN (Convolutional Neural Network) na extrakciu vysoko úrovňových vlastností zo snímok videí a vykonávanie predikcie. Momentálne najviac rozšíreným využitím CNN je rozpoznávane aktivít vo videách. (Alzubaidi et al., 2021)

CNN predstavuje doprednú neurónovú sieť, ktorá berie vstupné dáta a spracúva ich cez niekoľko vrstiev neurónovej siete, pričom sa skladá z vrstiev:

- vstupná vrstva,
- konvolučná vrstva,
- pooling vrstva,
- plne prepojená vrstva
- a výstupná vrstva.

Na základe tejto architektúry dokáže model extrahovať užitočné informácie (Obrázok 2). (Gama et al., 2018)

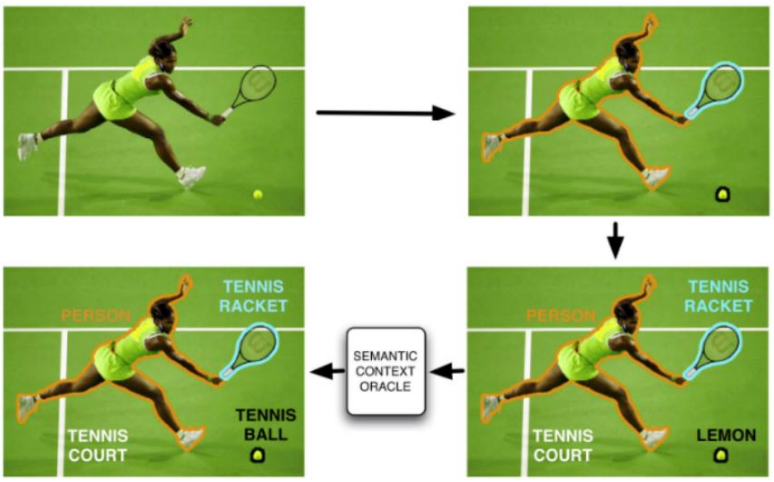


Obrázok 2 Architektúra CNN[2]

Konkrétnu funkcionalitu jednotlivých vrstiev popisujú O'Shea a Nash (2015). Vstupná vrstva predstavuje hodnoty pixelov jednotlivých snímok. Konvolučná vrstva vytvára výstup neurónov, ktoré sú prepojené s lokálnymi regiónmi vstupu. Výstup je kalkulovaný pomocou skaláru medzi váhami a regiónmi. Na tejto vrstve sa taktiež vykonáva lineárna operácia (ReLu), ktorá predstavuje aktivačnú funkciu ako napríklad sigmoid. Lineárna operácia sa vykonáva na výstupe konvolučnej vrstvy. Pooling alebo vrstva zlučovania následne vykonáva zmenšenie vzorkovania podľa priestorovej dimenzionality zo vstupu. Týmto znižuje počet parametrov. Plne prepojená vrstva vytvára klasifikácie do jednotlivých tried. Výstupná vrstva spája výsledky do dimenzie podľa počtu kategórií

V športoch sa napríklad využívajú CNN na rozoznanie typu športu a konkrétnej aktivity v reálnom čase (Obrázok 3). Na základe monitorovania ľudí, konkrétne v tomto prípade ich stredu tela a štvorcového mapovania, sa vykonávajú pozorovania ich správania. Takéto modelovanie sa nevyužíva iba v analýze športovcov a ich výkonov, ale aj vo virtuálnej realite a monitorovaní reakcií ľudí. (Zhengfeng, 2022)

[2] Zdroj Obrázok 2: (O'Shea a Nash, 2015)

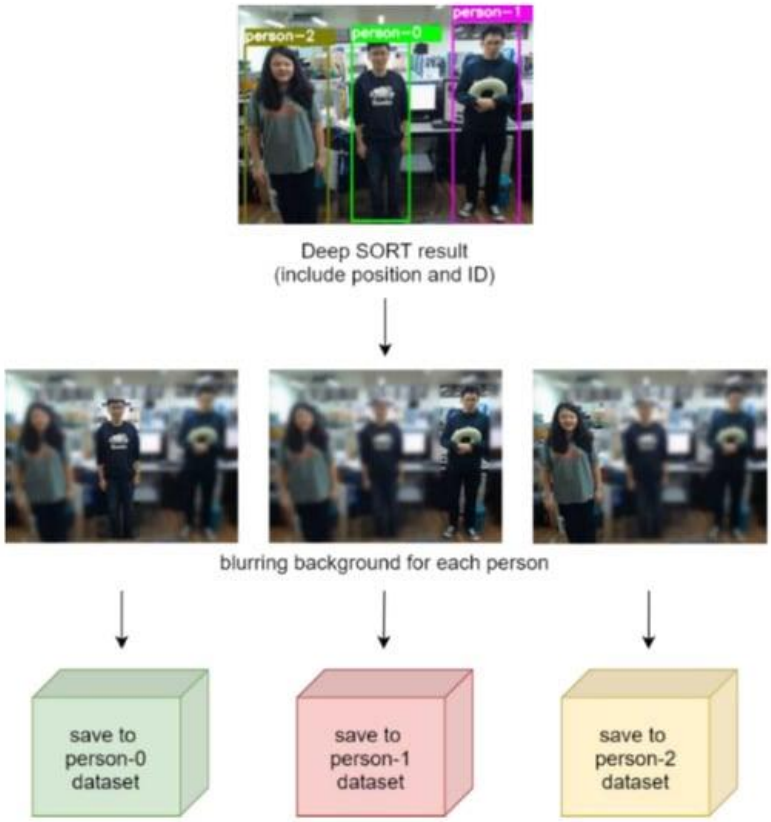


Obrázok 3 Rozpoznávanie aktivity a objektov[3]

Výkony basketbalistov napríklad skúmali Delextrat a Cohen (2009), ktorí vo svojom výskume popisujú jednoznačné využitie do budúcnosti hlavne na personifikáciu stratégie pre jednotlivých hráčov. Vďaka tomuto poznatku vedia dosiahnuť lepšiu ofenzívu ako aj defenzívu u hráčov, ako aj dodatočné predpovedanie dynamiky medzi spoluhráčmi v tíme. Kombinuje sa tu pri tom predpovedanie hráčovho pohybu, ako aj trajektória letu jeho hodov s loptou. Pri väčšom preskúmaní tejto problematiky by mala analýza videí veľký dopad na športový priemysel.

Ďalšie využitie, ktorého základom je pozorovania ľudí, sa využíva v bezpečnostnom monitorovaní. Vďaka modelom sa znižujú náklady na obstarávanie týchto systémov a ich údržba. Najväčšou prekážkou je momentálne problém, kde sa modely zameriavajú na aktivity jednej osoby. Avšak v tomto odvetí je potrebné monitorovať všetky osoby na danej snímke videa (Obrázok 4). Tsai et al. (2020) sa zamerali na túto problematiku a výsledok ich výskumu popisujú v článku, ktorý opisuje využitie hĺbkového učenia v reálnom čase, ktoré sa upriamuje na viaceré osoby. Využívajú pri tom model I3D.

[3] Zdroj Obrázok 3: (Wang a Zhu, 2023)



Obrázok 4 Rozpoznávanie osôb[4]

1.3.2 RNN

RNN (Recurrent Neural Network) využívajú výskumníci napríklad v detegovaní podvodných videí. Videá, ako médium, sa používajú v mnohých oblastiach ako dôkazový materiál: právo, forenzné štúdie, novinárstvo a mnohé iné. Z tohto dôvodu je nesmierne dôležité overiť ich autenticitu. Autori Munawar a Noreen (2021) popisujú využitie architektúry siamských RNN v modeli I3D a na dvoch datasetoch dosahujú presnosti až 86.6 % (dataset VIRAT - Image Retrieval and Analysis Tool) a 93 % (dataset MFC - Media Forensic Challenge). V princípe modely využívajú na vyhľadávanie podvodov alebo manipulácií vo videách ICD model so siamskými RNN (Siamese-based RNN).

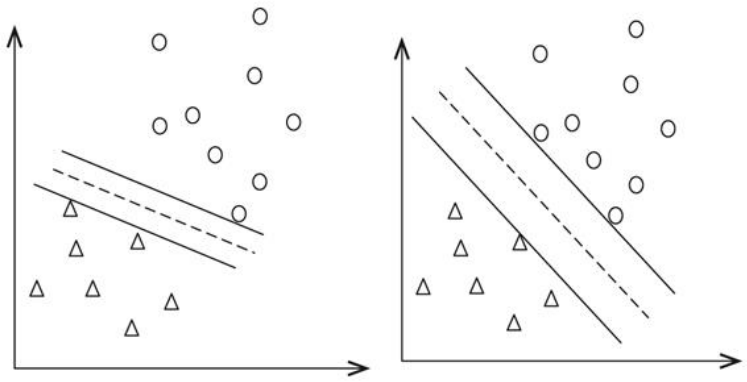
Siamese-based RNN odkazuje na architektúru rekurentnej neurónovej siete, ktorá zahŕňa koncept siamských sietí. Neurónové siete tu spolupracujú. Zdieľajú si váhy a architektúru, čo im umožňuje spracovávať viaceré vstupy paralelne a učiť sa podobnosti alebo rozoznávať rozdiely. (Mueller a Thyagarajan, 2016)

[4] Zdroj Obrázok 4: (Tsai et al., 2020)

1.3.3 SVM

SVM alebo Support Vector Machine je algoritmus strojového učenia, ktorý sa často využíva pre klasifikačné a regresné úlohy. V oblastí videí sa skúma využitie SVM na kategorizáciu druhov videí a to napríklad na reklamy, kreslené rozprávky, hudobné videá, správy a športové videá. Pri detekcii sa využíva taktiež zvuková zložka, ktorá sprevádza video. Autori Zhang a Li (2021) sa dodatočne snažia odhaliť aj teroristický alebo násilný obsah, tým by sa predišlo zobrazovaniu nevhodného obsahu.

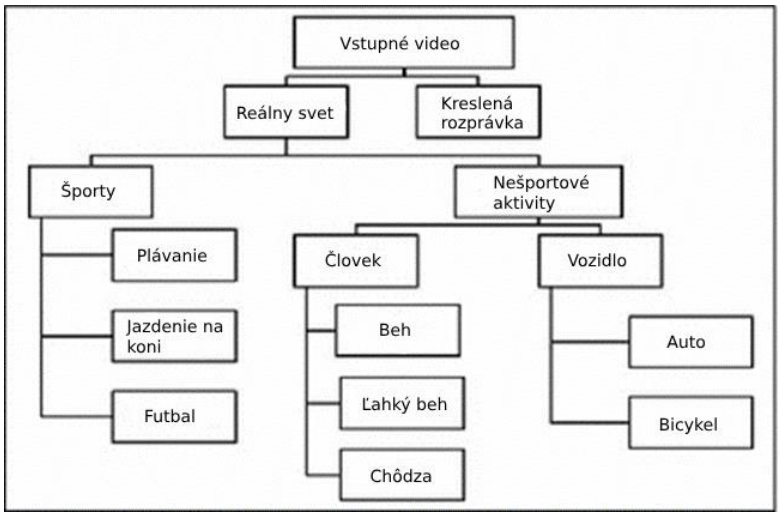
Princípom fungovania SVM je najmä štatistický prístup k nájdeniu najlepšej plochy. Plocha sa hľadá v pôvodnom priestore alebo vo vyššej dimenzii po projekcii. Cieľom je maximalizovať medzeru medzi dvoma kategóriami, čím sa znižuje interval neistoty pri generalizácii a tak sa minimalizujú skutočné riziká (Obrázok 5). (Zhang a Li, 2021)



Obrázok 5 Maximalizovanie medzere v SVM[5]

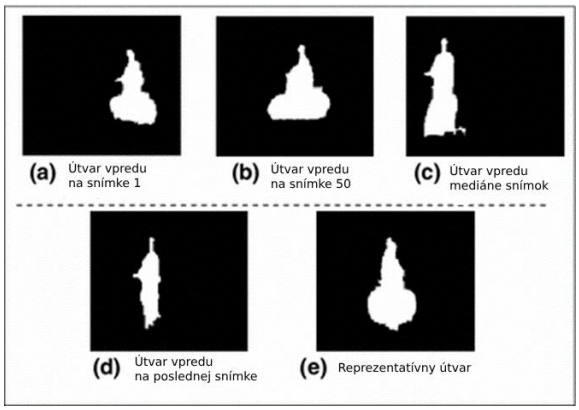
Chattopadhyay a Maurya (2013) využili SVM, aby preskúmali možnosť rýchleho označovania videí do kategórií pomocou kľúčových slov (tagov). (Obrázok 6).

[5] Zdroj Obrázok 5: (Zhang a Li, 2021)



Obrázok 6 Kategorizovanie videí[6]

SVM, v týchto prípadoch, spracováva videá tak, že sa najskôr dajú do unitárneho farebného spektra a následne sa snímky konvertujú do šedých farieb. Konverzia je odporúčaná z dôvodu výpočtovej náročnosti. Následne sa pracuje s útvarmi. Veľkou pomocou pri detegovaní útvarov sa ukázalo porovnávať a zaznamenávať aj textúry objektov. Pri tomto kroku sa rozlišuje pozadie a objekty záujmu, ktoré sa segmentujú. Pričom vzniká akýsi útvar (Obrázok 7), ktorý sa oddeľuje od pozadia. Útvary sa prekrývajú a na základe ich relatívneho centra sa pozorujú zmeny. (Chattopadhyay a Maurya, 2013)



Obrázok 7 Definovanie útvarov naprieč snímkami[7]

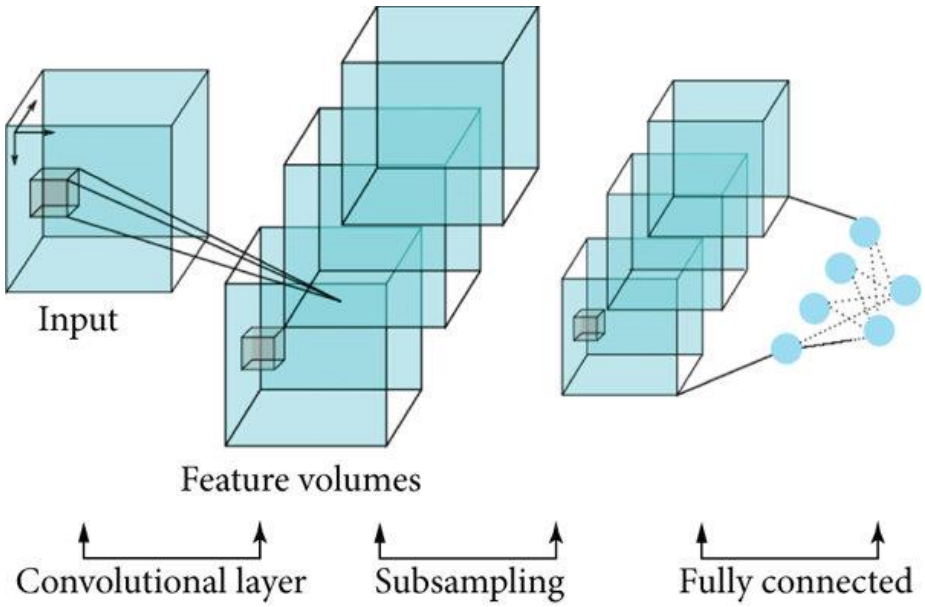
Po zozbieraní zmien sa nachádzajú rozdiely medzi kategóriami. Vo výsledku autori Chattopadhyay a Maurya (2013) poukazujú na rozsiahle využitie v rozpoznávaní

[6] Zdroj Obrázok 6: (Chattopadhyay a Maurya, 2013)
[7] Zdroj Obrázok 7: (Chattopadhyay a Maurya, 2013)

kontextu videa, avšak do budúcna odporúčajú pridať ďalšie média ako napríklad zvuk a text.

1.3.4 3D Konvolučné modely

3D CNN (3D Convolutional Neural Network) je dopredná neurónová sieť, ktorá na rozdiel od 2D CNN zaznamenáva pozície objektov v čase, čiže uchováva priestorový kontext. Architektúra 3D CNN vytvára aktivačnú mapu počas konvolučného kroku. Táto mapa uľahčuje vytváranie nelineárnosti a získavanie vzorov. Na získanie kategorizácie sa opakujúce konvolučné a zhlukové (pooling) vrstvy spájajú v plne prepojených vrstvách (fully connected layers) (Obrázok 8). (Karasawa et al., 2018)



Obrázok 8 Architektúra 3DCNN[8]

1.4 MODELY VYUŽITÉ PRI ANALÝZE ERA DATASETU

Autori Mou et al. (v tlači), ktorí spracovali videá do datasetu ERA: Event Recognition in Aerial Videos - Rozpoznávanie udalostí v leteckých videách, otestovali dataset na štyroch základných modeloch:

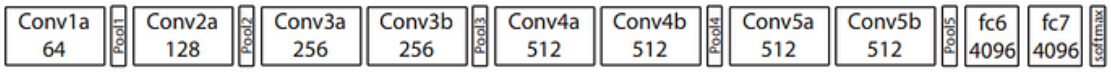
- C3D,
- P3D,
- I3D,
- a TRN modely.

[8] Zdroj Obrázok 8: (Hameed et al., 2022)

Tieto modely boli natrénované na rôznych datasetoch, napríklad Kinetics, UCF101 alebo Sport1M a každý má dve varianty natrénovania, ktoré sa okrem datasetu môžu líšiť aj pridanou architektúrou.

C3D Model

C3D (Convolutional 3D model - Konvolučný 3D model) popisujú Hadidi et al. (2020) ako model, ktorý pracuje s použitím 3D konvolúcií na extrakciu priestorovo časových prvkov z videí. Na rozdiel od dvojrozmerného priestoru snímok zohľadňuje dimenziu navyše a berie z nej údaje, čiže zachytáva informácie o pohybe a čase (Obrázok 9).



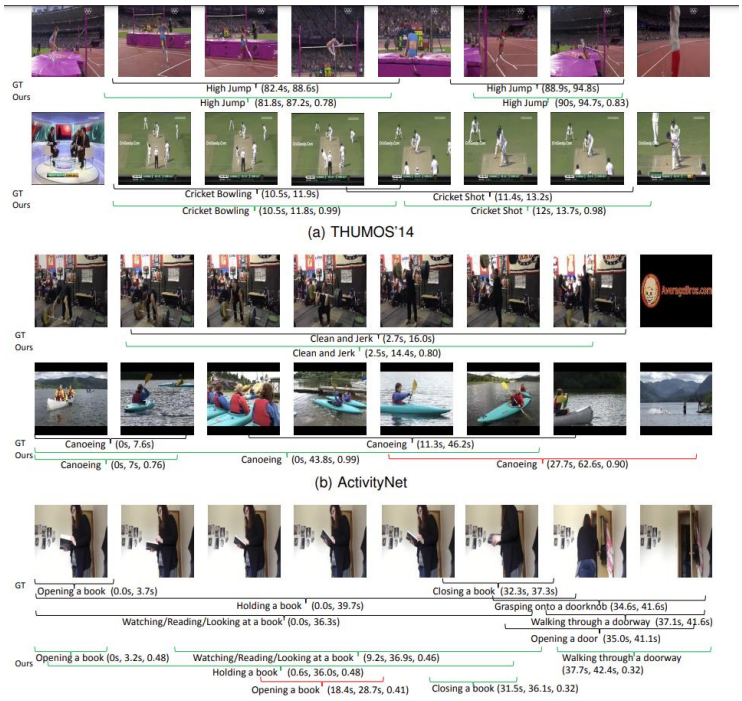
Obrázok 9 Architektúra C3D[9]

Wang et al. (2016) popisujú stratégiu, kedy nespracováva model každý snímok, ale dané číslo snímok v intervale, tým sa zvyšuje jeho efektivita a znižuje výpočtová náročnosť.

Podľa výskumu na učeniach modelu C3D, ktorý vykonali Tran et al. (2015), zistili, že ak zobrali dataset so zameraním (videá zo športu, rozpoznávanie aktivít vo videách, rozpoznávanie objektov a scény), v tom prípade bol C3D prakticky vždy najlepší alebo jeden z najlepších modelov. A to vo výskume použili špecializované modely na dané úlohy.

Model C3D sa využíva v oblasti počítačového videnia pre úlohy rozpoznávania udalostí s cieľom klasifikácie videí. Aplikuje sa v rôznych oblastiach vrátane identifikácie športových gest, detekcie aktivít vo videách ako aj napríklad depresie z tvárových výrazov alebo monitorovania spánku (Obrázok 10). (De Melo et al., 2019; Jazaery a Guo, 2021)

[9] Zdroj Obrázok 9: https://medium.com/analytics-vidhya/extracting-features-from-videos-using-c3d-v1-1-in-ubuntu-16-04-16f2b2990e3b



Obrázok 10 Predikovanie aktivít pomocou C3D[10]

Pri rozpoznávaní akcií tento model preukázal svoju schopnosť prekonať iné, často využívané modely ako sú HOF (Histogram of Optical Flow) a MBH (Motion Boundary Histogram) a to konkrétne z hľadiska klasifikácie pohybov. (Wang a Schmid, 2013)

Shang (2020) taktiež popisuje zlepšenie výkonov s použitím modelu ActionVLAD.

HOF a MBH používajú histogramy na rozpoznávanie najmä aktivít vo videách. HOF využíva výpočty optického toku medzi susednými snímkami. Konkrétne informácie o pohybe získavajú rozdelením obrazu snímky do priestorových buniek a kvantifikáciou optického toku do rôznych intervalov (binov). Výsledný histogram reprezentuje distribúciu smerov pohybu vo videu. (Li et al., 2016)

Li et al. (2016) ďalej popisujú, že MBH naopak zachytáva hranice pohybu vo videu. Počíta gradient optického toku a kvantifikuje orientáciu gradientu do rôznych intervalov. V tomto prípade histogram reprezentuje distribúciu hraníc pohybu vo videu.

V praxi sa využíva kombinácia HOF alebo MBH s C3D modelom a tu dodávajú doplnkové vlastnosti k priestorovým informáciám zachyteným modelom C3D. Ukazuje sa avšak, že C3D jednoducho prekonáva HOF a MBH modely aj v prípade, že pracuje osamote. Čo sa preukázalo pri viacerých testoch výkonu. (Tran et al., 2015)

[10] Zdroj Obrázok 10: (Xu et al., 2017)

V oblasti medicínskeho výskumu sa C3D využíva pri pacientoch, ktorým bola transplantovaná oblička. Podľa výskumu Comoli et al., (2016) by mal byť tento model schopný predikovať riziká odmietnutia transplantátu.

Pri výskumoch v oblasti počítačovej vedy popisujú Xuan et al. (2019) vývoj variácie pre model C3D: MV-C3D pre 3D konvolučné neurónové siete. Tento nový model preukazuje svoju prispôsobivosť v reálnych situáciách a to najmä pri analýze rotovaných reálnych obrázkov v 3D priestore. Z toho vyplýva, že sa tento model dá využiť v počítačovom videní, najmä pri rozpoznávaní objektov a pochopenia scény.

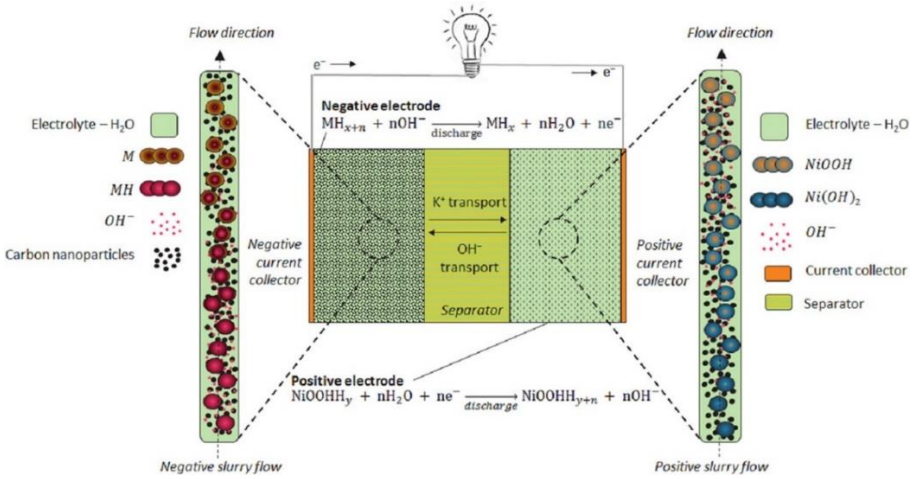
Taktiež Ross et al. (2000) popisujú využitie aj v imunológii, kde C3D zefektívňuje vytváranie protilátok na vírusy prvého typu. Tým pádom sa dá potencionálne vyžiť na vytváranie stratégií pri tvorbe vakcín a imunoterapie pri vírusových infekciách.

P3D Model

P3D (Pseudo 3D model) model pozostáva z trojdimenzionálnych blokov, ktoré nahrádzajú konvolučné jadrá a tým uľahčujú výpočtový proces. Ide teda o nový výpočtový prístup, ktorý eliminuje potrebu opätovného výpočtu variačných komponentov. (Zhang et al., 2010)

P3D model pol porovnávaný s C3D a predstavuje okrem menšej výpočtovej náročnosti aj zníženie obťažnosti pri tréningu, taktiež znižuje percento potencionálneho preučenia. (Zhang et al., 2020)

Model nachádza využitie v mnohých sférach. Za jednu z najviac zaujímavých považujeme výskum v oblasti batérií. Jedna zo štúdií od Chayambuka et al. (2019) sa zameriava na modelovanie a návrh polotuhých článkových batérií (SSFBs) (Obrázok 11). Model P3D bol použitý na simuláciu difúzie látok, ktoré prezentovali jediný transportný mechanizmus v aktívnych častiach SSFBs. Model umožnil znázorniť časovo závislé profily napätia, distribúciu prúdu a distribúciu stavu napätia v aktívnych častiach.



Obrázok 11 Schéma polotuhej článkovej batérie[11]

Využitie P3D popisujú v biológii Babu a Fullwood (2015), konkrétne v oblasti génových štúdií, kde sa používa napríklad na pochopenie organizácie a funkcionality genómov pri zdraví a populačných chorobách. V medicíne nachádza P3D využitie pri modelovaní geometrie genómov, kde sa pozorujú bunky a ich dopad na zdravie ľudí.

I3D Model

Architektúru modelu I3D (Inflated 3D) predstavujú Wei et al. (2023) na princípe, kedy sa zväčšujú dvojrozmerné konvolučné siete, aby model spracovával informácie času a priestoru z videí. V podstate využíva 2D siete, ktorým pridáva filtre a presúva ich do 3D. Čiže napríklad štvorcový filter sa stáva kubický.

V oblasti rozpoznávania aktivít vo videách sa tento model využíva v chirurgii, kde sa pozorujú postupy pri práci. V štúdii Zhang et al., (2021) bol model I3D, konkrétne architektúra Inflated 3D ConvNet (na základe ktorej je vybudovaný model I3D), využitá na rozpoznávanie chirurgických postupov pri operácií tubulizácie žalúdka. Model bol trénovaný pomocou metódy ohniskovej straty (Focal loss) a dosiahol presné rozpoznávanie rôznych chirurgických krokov.

Focal Loss je špeciálny druh loss funkcie, ktorá sa zaoberá problémom nerovnováhy tried pri úlohách detekcie objektov. Keďže pri tejto problematike väčšina snímok pozostáva z pozadia, zatiaľ čo pre nás je dôležitá detekcia objektu záujmu, môže nerovnováha spôsobovať malú úspešnosť detegovania objektov v popredí. Na klasifikáciu používa špeciálnu váhu, ktorá znižuje hodnotu pri ľahko odhadnuteľných

[11] Zdroj Obrázok 11: (Chayambuka et al., 2019)

prípadoch a teda sa trénuje na ťažkých a viac ojedinelých príkladoch tried. Týmto zlepšuje výkony v triedach menšín. (Lin et al., 2017)

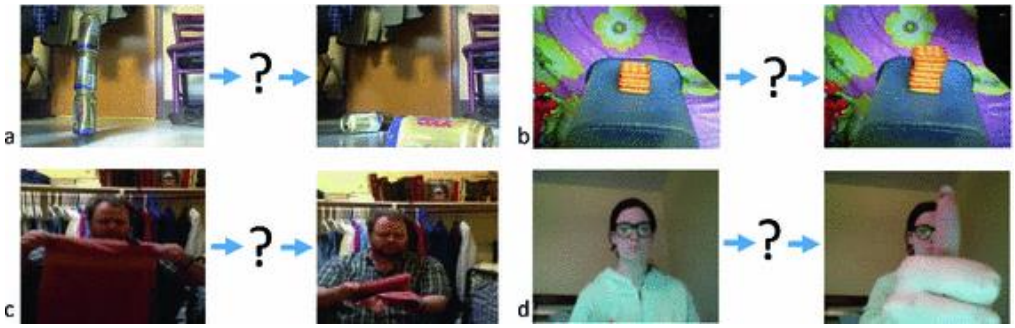
Shi et al. (2020) poukazujú na tri kľúčové výhody využitia I3D v praxi: využíva menej parametrov, znižuje výpočtové náklady a dokáže naučiť priestorovo-časové funkcie pre využitie v datasetoch.

I3D model sa využíva taktiež pri bezpečnostných videách v reálnom svete. Napríklad v autonómnych autách alebo pri detekcii osôb na záznamoch. (Liu et al., 2022)

TRN Model

TRN model (Temporal Relation Network – dočasne relačný model) vznikol na princípe relačných posudkov v čase (Temporal relational reasoning). Ide o schopnosť prepojiť si zmysluplné transformácie objektov alebo entít v čase. Je to vlastnosť, ktorá sa prisudzuje inteligentným bytostiam. (Zhou et al., 2018)

Ako autori Zhou et al. (2018) popisujú, model TRN je vytvorený tak, aby sa vedel efektívne učiť a interpretovať vzťahy medzi snímkami vo videu a to vo viacerých časových vrstvách. Na obrázku 12 vidíme ako si ľudský mozog ľahko doplní kontext medzi obrázkami, ale pre modely je to náročné. Tento problém sa snaží vyriešiť model v architektúre TRN za pomoci zisťovania vzťahov medzi snímkami.



Obrázok 12 Dopĺňanie kontextu videa[12]

Veľkou výhodou TRN modelu oproti iným je jeho ľahšie pochopenie pre vyvodenie výsledkov. (Zhou et al., 2018)

Vo výskumoch neurovedy skúmali Wolff et al. (2021) tento model pri regulácii senzorického spracovania, pozornosti a kognície.

[12] Zdroj Obrázok 12: https://neurohive.io/en/state-of-the-art/temporal-relational-reasoning-in-videos/

Xu et al. (2019) popisujú problém ukladania indícií ako napríklad vzhľadu, polohy alebo topológie do samostatných sietí. Ako riešenie predstavili nadstavbu na TRN, ktorá pridáva modul priestorovo temporálnej relácie do modelu STRN. Nový model testovali na datasetoch, ktoré sa zmeriavajú na sledovanie viacerých objektov. Úspech dosahovali aj v monitorovaní v reálnom čase.

1.4.1 Ďalšie modely

V dnešnej dobe sa výskumníci snažia modely zlepšovať a za týmto cieľom vznikli aj modely:

- TSN
- a LSTM.

Oba modely pristupujú ku klasifikácii novým spôsobom a tým sa snažia minimalizovať čas na trénovanie a zlepšovať detekciu kontextu.

TSN model

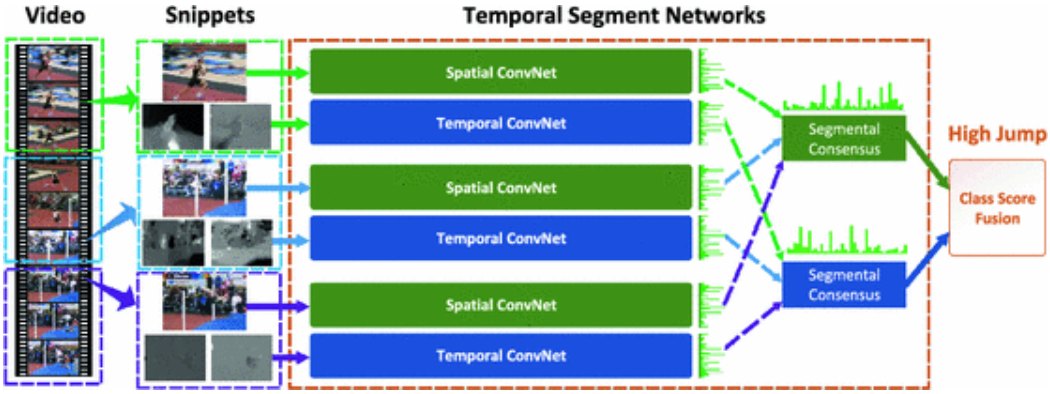
TSN model bol vyvinutý na základe potreby rozoznávať dlhodobé aktivity a naplniť potrebu trénovať modely na veľkom množstve videí. Tým zlepšiť výkonnosť modelov oproti architektúram postaveným iba na konvolučných sieťach. Autori videli veľký potenciál na zlepšenie a vyvinuli model TSN alebo Temporal Segment Network (Sieť využívajúca časových segmentov). Rámec (framework) v architektúre modelu extrahuje krátke úryvky z dlhej sekvencie pôvodného videa. Využíva sa na to vzorkovacia stratégia, ktorá nemá charakter hustého vzorkovania. Vďaka tomu je model schopný pracovať s dlhými videami a zachovať si relevantné informácie za menšej výpočtovej záťaže. (Wang et al., 2016)

Na obrázku 13 je popísaná segmentácia videa a náhodné vybranie úryvku, na ktorom sa vykonáva následne modelovanie. Na to sa tu využíva Spatial and Temporal ConvNet čiže konvolučné siete vybudované na základe priestoru a času. (Wang et al., 2016)

Prínos TSN autori Wang et al. (2016) priraďujú momentálne najviac k zlepšeniu výkonov pri rozpoznávaní aktivít vo videách.

Autori Yang et al. (2022) sa zaoberajú taktiež problematikou vynechávania nepodstatných snímok alebo oblastí vo videách, ktoré ale môžu ovplyvniť rozpoznávanie akcií. Preto vo svojom výskume navrhujú nadstavať TSN na architektúru STA-TSN (Spatial-Temporal Attention Temporal Segment Networks), ktorá zachováva schopnosť

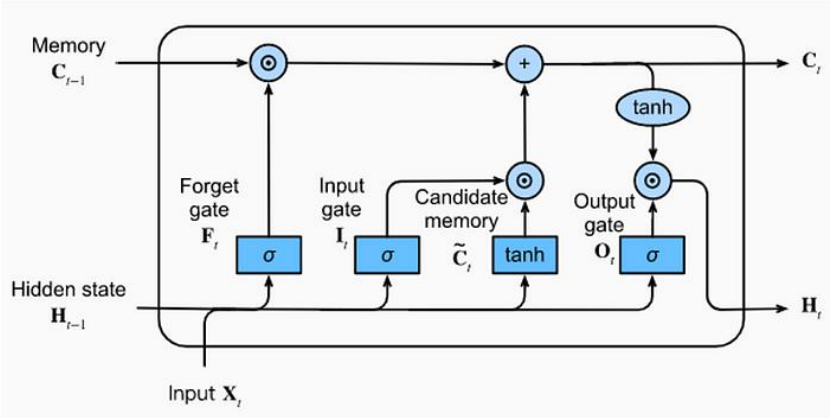
zachytiť dlhodobé informácie a umožňuje adaptívne sústredenie na kľúčové časti v priestore aj čase.



Obrázok 13 Schéma TSN modelu a segmentácia[13]

LSTM

LSTM- Long Short-Therm Memory model alebo model s dlhodobou krátkodobou pamäťou. Ide o rekurentnú neurónovú sieť, ktorá ale zachytáva a modeluje dlhodobé súvislosti v sekvenčných dátach (Obrázok 14). (Zhang et al., 2016)



Obrázok 14 Architektúra LSTM[14]

Model má taktiež využitie v rozpoznávaní obsahu videa a jeho popísaní. Zaznamenáva najviac dôležité momenty a na ich základe hodnotí obsah. LSTM sa využíva najmä pri dlhých videách. (Zhang et al., 2016)

V práci autorov Yao et al. (2018) popisovali ako sa využíva LSTM na optimalizovanie dopytu taxi služieb. Predstavujú avšak vylepšenie modelu (DMVST-Net, Deep Multi-View Spatial-Temporal Network) za pomoci ktorého by sa priblížila

[13] Zdroj Obrázok 13: (Wang et al., 2016)

[14] Zdroj Obrázok 14: https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c

spoločnosť k dosiahnutiu inteligentného transportného systému, ktorý by prispel k dosiahnutiu konceptu inteligentných miest.

1.5 POROVNANIE MODELOV

Modely okrem architektúry majú taktiež odlišné spôsoby učenia, ktoré vedie k rozličnej aplikácii. Tabuľka 1 zhromažďuje základné údaje o modeloch, ktoré využívame v tejto práci.

Tabuľka 1 Porovnanie modelov[15]

Model	Úloha, využitie	DNN model	Typ kontextu	Úroveň kontextu	Mechanizmus
C3D	Rozpoznávanie vo videách	3D Konvolučné Siete	Temporálny	Lokálny, globálny	3D konvolúcie
P3D	Rozpoznávanie vo videách	3D Konvolučné Siete	Temporálny	Lokálny, globálny	Pseudoprechody, 3D bloky
I3D	Rozpoznávanie vo videách	Nafúknuté 3D Konvolučné Siete	Temporálny	Lokálny, globálny	Nafúknuie 2D konvolúcie
TRN	Rozpoznávanie vo videách	Siete s temporálnymi vzťahmi	Temporálny	Lokálny, globálny	Modelovanie dlhodobých vzťahov
TSN	Rozpoznávanie vo videách	Siete s temporálnymi vzťahmi	Temporálny	Lokálny, globálny	Segmentovanie videa
LSTM	Sekvenčné modelovanie	Rekurentné neurónové siete	Temporálny	Lokálny, globálny	Pamäťové brány a bunky na kontrolovanie toku a ukladanie informácií

Stĺpce tabuľky popisujú:

- Model: Predstavuje názov daného modelu.
- Úloha: Akú základnú úlohu spracováva model.
- DNN (Deep Neural Network) model: Na akom type neurónovej siete alebo architektúry je model postavený.
- Typ kontextu: Ak sa model zameriava na priestorový, temporálny kontext, poprípade ich kombináciou.

[15] Zdroj Tabuľka 1: autor

- Úroveň kontextu: Či model zohľadňuje globálny kontext (celá scéna), lokálny kontext (menšia časť scény) alebo ich kombináciu.
- Mechanizmus: Popisuje aký mechanizmus alebo techniky využíva model na vykonanie úlohy počítačového videnia.

1.6 OPTIMALIZAČNÉ METÓDY

Pre zlepšenie kategorizácií videí existujú rôzne metódy. Vo veľa prípadoch chceme predchádzať pretrénovaniu modelov, zlepšiť ich výsledky alebo znížiť technické zaťaženie. Z týchto dôvodov existujú optimalizačné metódy.

Temporal pooling

Video sa dá vnímať ako usporiadaná kolekcia snímok. Klasifikovanie videa po snímkach s CNN sa viaže na ignorovanie charakteristík pohybu, keďže sa zanedbáva temporálna informácia. V závislosti na úlohe, zlučovanie priestorových funkcií, ktoré vyprodukovala CNN s temporálnym zlučovaním, môže byť dobrou stratégiou časové zhlukovanie (temporal pooling) podľa Pigou et al. (2016).

Algoritmus, ktorý popisujú autori Lu at et. (2018), sa zakladá na predpoklade, že cieľová kategória je priradená každej temporálnej lokácií (priemerný pooling) alebo je priradená iba jednej temporálnej lokácii s maximálnou odozvou (maximálny pooling).

Vzorkovacia metóda

Metód vzorkovania existuje mnoho, medzi základné patria napríklad náhodné vzorkovanie alebo replikácia. (Buckland et al., 2015)

Na predchádzanie nedostatku flexibility a výpočtovej náročnosti modelov je vhodné použiť náhodné vzorkovanie. Táto metóda vyberá náhodný počet snímok, ktoré reprezentujú dané video. (Zhi et al., 2021)

Temporal aggregation

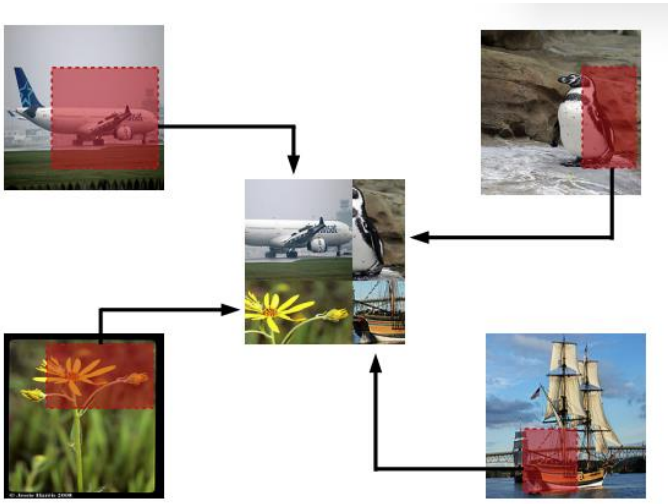
Kourentzes a Petropoulos (2016) popisujú metódu časovej agregácie (temporal aggregation). Podľa nich predstavuje zlepšenie odhadu modelu, v procese, kedy sa komponenty v časových radoch stanú viac alebo menej významnými.

Môže napríklad zvýšiť frekventovanosť komponentov s malým výskytom pričom sú dôležité v kontexte videa. (Athanasopoulos et al., 2017)

Transformácie

Podľa autorov Li et al. (2018) je pri klasifikácií videí vo veľa prípadoch lepšie využiť transformácie ako zlepšovať vyhľadávanie v časovom priestore. Vedie to k neporušovaniu štruktúry a zmenšeniu priestoru na vyhľadávanie.

Pretrénovanie je stály problém pri práci s modelmi. Technika transformácie sa snaží tomu zabrániť a popri tom obohatiť datasety. Jednou z metód je náhodné orezanie snímok a následné prepojenie týchto orezaných kúskov (Obrázok 15). Medzi iné bežne používané metódy patria aj normalizácia, otáčanie, zmena rozmerov, zmena farieb, kontrastu alebo saturácie a iné. Taktiež metóda odstraňovania pixelov vytvára šumové snímky a tým vie predchádzať pretrénovaniu. (Takahashi et al., 2020)



Obrázok 15 Orezávanie snímok ako forma transformácie[16]

1.7 ARCHITEKTÚRY POUŽITÉ V MODELOCH ERA

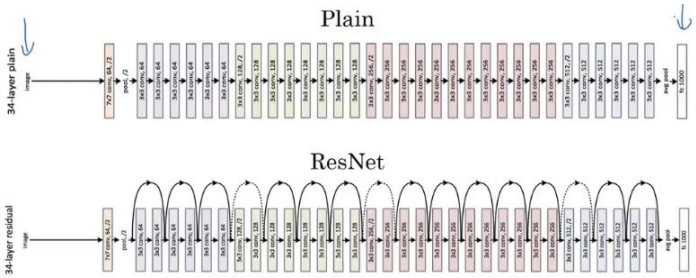
Autori datasetu ERA použili pri trénovaní modelov architektúry ResNet a Inception, aby modely mali lepšie kategorizácie a zároveň optimalizovali náročnosť trénovania.

ResNet

Neurónová sieť ResNet (Residual Network) predstavuje architektúru, ktorá vytvára skratku medzi spojeniami s cieľom znížiť náročnosť trénovania. Výsledkom je rýchlejšie trénovanie a nižšia generalizačná chyba. ResNet v princípe vrství reziduálne bloky (vrstvy, ktorých výsledky idú hlbšie do iných vrstiev v bloku) na seba.

[16] Zdroj Obrázok 15: (Shorten a Khoshgoftaar, 2019)

Následne prepája dva výsledky s cieľom znížiť trénovaciu chybu (Obrázok 16). (Li et al., 2017)



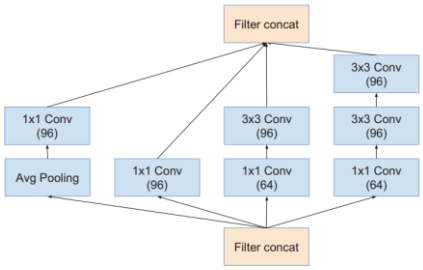
Obrázok 16 Architektúra ResNet[17]

Inception

Spoločnosť Google predstavila v roku 2014 novú architektúru s názvom GoogLeNet, ktorá sa neskôr premenovala na Inception. Jej cieľom bolo zlepšiť efektivitu hĺbkových neurónových sietí (deep neural networks) za pomoci znižovania počtu parametrov. Avšak chceli zachovať cieľ predchádzaniu znižovania presnosti. (Gill, 2023)

Autor Gill (2023) popisuje taktiež aj pôvodnú architektúru, ktorá kombinuje 1x1, 3x3 a 5x5 paralelných konvolučných filtrov na extrakciu vlastností zo vstupu. Vstupy sú následne spojené a poslané nasledujúcej vrstve, kde sa proces opakuje. Tento prístup umožňoval zachytiť vstupné lokálne a globálne vlastnosti, za minimalizovania počtu parametrov.

V dnešnej dobe existuje viacero verzií, najnovšia je Inception-v4, ktorá vylepšuje predošlú verziu a pridáva reziduálne prepojenia. Okrem prepojení obsahuje normalizáciu dávok (batch normalization) a architektúru s 5x5 konvolučnými filtrami s dvoma 3x3 filtrami v sérii (Obrázok 17). (Gill, 2023)



Obrázok 17 Architektúra Inception-v4[18]

[17] Zdroj Obrázok 16: https://medium.com/analytics-vidhya/opening-resnets-46bb28f43b25
[18] Zdroj Obrázok 17: (Li et al., 2018)

2 CIELE ZÁVEREČNEJ PRÁCE

Hlavným cieľom diplomovej práce je porovnať kvalitu modelov na základe presnosti kategorizácií videí do tried datasetu ERA.

Naše čiastkové ciele sú:

- Pochopiť architektúru predtrénovaných modelov a vytvoriť kód v jazyku Python na ich otestovanie.
- Otestovať predtrénované modely autorov datasetu ERA a porovnať ich výsledky s oficiálnymi výsledkami autorov uvedenými v článku o datasete.
- Porovnať modely medzi sebou a nájsť ich najlepšie a najhoršie využitia na konkrétnych druhoch videí.
- Vizualizovať a vyhodnotiť výsledky, ktoré dostaneme z testovania modelov.

3 METODIKA VÝSKUMU

V tejto kapitole predstavíme metódy a postupy, ktorými sme sa riadili počas nášho výskumu. Vychádzali sme z metodiky CRISP-DM. (IBM, 2021)

3.1 POROZUMENIE PROBLEMATIKE

Po dôkladnom rešerši teoretických východísk klasifikácie vo videách pomocou neurónových sietí, boli zostavené metódy a postupy pre túto prácu.

Klasifikácia vo videách

Klasifikáciu videí popisujú Bekhet a Alghamdi (2021) ako proces kategorizácie do rôznych tried alebo kategórií na základe obsahu a charakteristík. Cieľom je automaticky analyzovať a pochopiť informácie, ktoré sa nachádzajú vo videu.

Oproti klasifikácií v obrázkoch obsahujú videá viaceré snímky, ktoré majú medzi sebou informačné prepojenie (kontext). Obsahujú taktiež aj viac informácií, ktoré vieme využiť ako napríklad zvuk. Tieto dodatočné informácie umožňujú komplexnejšie porozumenie obsahu a pri ich využití môžu zlepšiť presnosť klasifikácie, ako popisujú Karpathy et al. (2014).

3.2 POROZUMENIE DÁTAM

Dataset ERA sme zvolili pre túto prácu z dôvodu jeho konzistentného obsahu. To znamená, že videá mali rovnakú dĺžku, obsahovo sa pridržiavali kategórie, ich kvalita bola rovnaká a počet videí bol dostatočný na ďalšie využitie pre testovanie alebo trénovanie modelov. Taktiež bol dataset voľne stiahnuteľný.

Dataset ERA

Ako popísali Mou et al. (2020) v ich práci: Dataset ERA (A dataset and deep learning benchmark for event recognition in aerial videos) predstavuje významný prínos v oblasti rozpoznávania udalostí vo videách natočených zo vzduchu. Pozostáva z 2864 videí, ktoré zaznamenávajú udalosti, získaných z platformy YouTube, pričom každé video je označené jednou z dvadsaťpäť rôznych tried udalostí. Hlavným znakom datasetu je forma akou boli videá natočené, ide o dronové zábery. Tým sa odlišuje od iných často využívaných datasetov pre klasifikáciu, ako napríklad UCF, Kinetics alebo Sport1M.

Dataset predstavuje benchmark (porovnanie výkonnosti) pre modely na rozpoznávanie udalostí.

Jedna z prvých úloh na porozumenie datasetu a modelom bolo nájsť dokumentáciu alebo podklady pre testovanie od autorov. Pri hľadaní sme našli dostupný iba jeden článok, kde autori popisovali ich postup pri zbieraní videí, akou technológiu boli vytvorené videá (drony) a porovnávajú svoj dataset s inými datasetmi. Najmä vyzdvihujú dostupnosť dronových záberov naproti ostatným datasetom, kde sú videá získané prostredníctvom satelitných snímok. Poukazovali taktiež na horšiu kvalitu a vyššie náklady na tvorbu videí pri iných datasetoch. (Mou et al., 2020)

V článku popisovali Mou et al. (2020) podrobnejšie postup kategorizácie videí a ich voľbu formátu. Taktiež porovnávali datasety medzi sebou a svoje výsledky zhrnuli do tabuľky 2.

Tabuľka 2 Porovnanie datasetu autormi[19]

Dataset	Typ úlohy	Zdroj dát	Video	Počet tried	Počet vzoriek	Rok	Zdroj
UCLA	Zamerané na človeka	Získané osobne autormi	áno	12	104	2015	(Shu et al., 2015)
Okutama	Ľudská činnosť	Získané osobne autormi	áno	12	-	2017	(Barekatain et al., 2017)
AIDER	Pohromy	Internet	nie	5	2,545	2019	(Kyrkou a Theocharides. 2020)
ERA	Všeobecné	YouTube	áno	25	2,864	2019	(Mou et al, v tlači)

Pri výbere datasetu pre našu prácu sme porovnávali napríklad aj datasety spomenuté v tabuľke 2. Taktiež aj iné, častokrát využívané datasety: UCF101 (Soomro et al., 2012), HMDB51 (Kuehne et al., 2011), Kinetics (Kay et al., 2017), Charades (Sigurdsson et al., 2018), YouTube-8M (Abu-El-Haija et al., 2016), Sports-1M (Karpathy et al., 2014) a Something-Something (Goyal et al., 2017). Mnohé z nich autori využili na trénovanie vlastných modelov.

Významným problémom pri práci s datasetom bola absencia hlbšieho popisu modelov. Článok, ktorý popisujeme v tejto kapitole, sa zameriava na motiváciu a zber

[19] Zdroj Tabuľka 2: (Mou et al., 2020)

dát, avšak nepopisuje samotné trénovanie a testovanie modelov. Autori Mou et al. (2020) poskytli v článku iba výsledky z ich testovaní vo forme tabuliek, ale podrobnejší popis neposkytli.

Dataset ERA pozostával z dostatočného počtu videí na to, aby sa dali na ňom natrénovať modely a následne ich otestovať. Stiahnutý dataset sa delí na testovacie a trénovacie dáta, ktoré sú delené na samotné kategórie.

Videá v datasete

Dataset sa delí do sedem hlavných oblastí, ktoré sa členia podrobnejšie do jednotlivých kategórií nasledovne:

- Šport: basketbal, basebal, kanoistika, cyklistika, beh, futbal, plávanie, závody áut.
- Bezpečnosť: policajné prenasledovanie, konflikt.
- Pohromy: po zemetrasení, potopy, požiar, zosun pôdy, zosun bahna.
- Doprava: zrážka, zápchy.
- Produktívna činnosť: zber, orba, stavba.
- Bez udalosti.
- Sociálne aktivity: párty, koncert, prehliadka/protest, náboženská udalosť.

Každé video má vždy 5 sekúnd pri rozlíšení 640x640 pixelov a 24 snímok za sekundu. Pri výbere datasetu boli pre nás kľúčové nasledovné parametre:

- jeho dostupnosť, aby sme mali jednoduchý prístup k videám a boli zozbierané na jednom mieste,
- aby bol dostatočný počet vzoriek a boli správne označené a kategorizované,
- a aby mali jednotný vzhľad, čiže rovnakú dĺžku, rozmery videí a počet snímok za sekundu.

Aj keď kvalitu datasetu autori overili testovaním modelov, ktoré natrénovali, dataset sa nenachádza v tejto dobe vo veľa výskumoch a priame využite modelov sme nenašli. Tento poznatok sme brali ako ďalší potencionálny faktor pri výbere, kedy by sa naša práca dala použiť ako podklad pre ďalšie výskumy a prácu s datasetom ERA.

Predtrénované modely

Dataset ERA mal dostupné predtrénované modely, ktoré boli stiahnuteľné priamo na stránke, kde sa nachádzal aj odkaz na samotný dataset. Autori natrénovali osem modelov pre klasifikáciu videí a jedenásť modelov pre klasifikáciu snímok, ktoré brali

priamo z videí. Predmetom našej diplomovej práce nie je klasifikácia pre snímky, takže sme tieto modely netestovali a neskúmali bližšie.

Modely pre klasifikáciu videí môžeme rozdeliť podľa ich hlavnej architektúry na:

- C3D,
- P3D,
- I3D,
- a TRN.

Každá architektúra bola natrénovaná dvoma spôsobmi za pomoci rôznych datasetov (Kinetics, Sport1M, UCF a iné) a v niektorých prípadoch aj dodatočných architektúr (ResNet a Inception).

3.3 PRÁCA S MODELMI

Pri testovaní modelov sme narazili na viaceré problémy, ktoré by sme zosumarizovali do dvoch hlavných častí: problém s obmedzenými možnosťami výpočtovej techniky a neexistujúca dokumentácia k modelom.

Práca s predtrénovanými modelmi

Za cieľ v práci sme si stanovili otestovanie predtrénovaných modelov v našich podmienkach. Skúšali sme testovať na celej vzorke, ktorú poskytli autori datasetu. Vzorka na testovanie obsahovala v priemere 50 videí v každej z 25 kategórií. Avšak po prvých testoch sme znížili počet testovaných videí vzhľadom na časovú náročnosť.

Po priebežnom testovaní a ladení sme následne vybrali pre testovanie pätnásť videí z každej kategórie. Tento počet vyhovoval aj najmenej početnej kategórii „závodov áut“ (CarRacing), ktorá obsahovala iba 19 videí.

Pre vytvorenie zoznamu konkrétnych videí, sme naprogramovali Python skript na náhodný výber pätnástich unikátnych videí (Kód 1). Tie sme následne opakovane používali na testovanie, aby naše výsledky neboli skreslené rôznymi testovacími podmienkami. Zároveň výberom 15x25 videí sme výrazne znížili časovú náročnosť na testovaciu dobu.

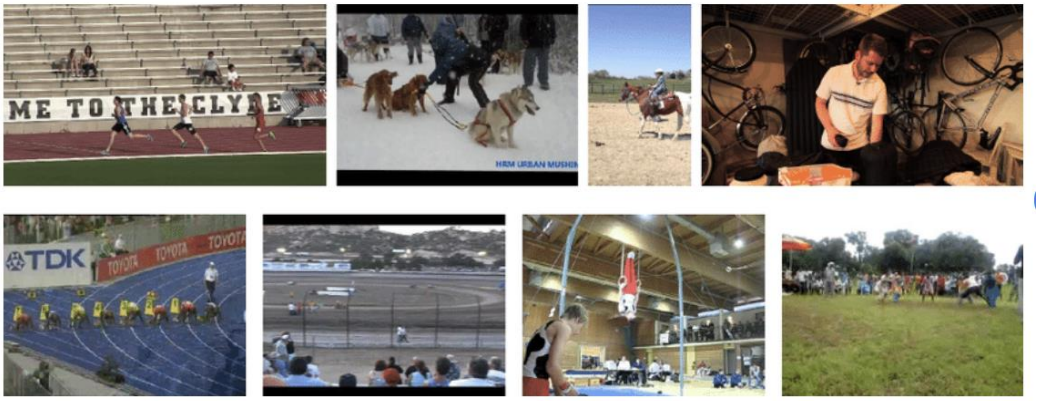
```
selected_videos = random.sample(video_files, 15)
        selected_video_paths = [os.path.join(category_path, video) for video in selected_videos]
```

Kód 1: Náhodný výber pätnásť videí z testovacích dát[20]

[20] Zdroj Kód 1: autor

C3D-Sport1M

Model autori natrénovali okrem datasetu ERA aj na datasete Sports-1M. Dataset obsahoval viac ako jeden milión videí, ku ktorým sa pristupovalo cez URL adresu. Videá boli rozdelené do 467 športových aktivít (Obrázok 18). Keďže sa dataset upriamuje na športové aktivity, tak sme očakávali, že dataset bude lepšie kategorizovať športy. (Karpathy et al., 2014)



Obrázok 18 Dataset Sport1M[21]

Samotný predtrénovaný model obsahoval všetky potrebné časti ako napríklad uloženú architektúru s jednotlivými vrstvami a parametre.

Pri práci s modelom sme najskôr potrebovali odvodiť jednotlivé parametre pre inicializáciu modelu: vstupný parameter pre tenzor (input_x), výstupný tenzor (logits) a trénovací tenzor (trainings). Keďže sme pracovali s knižnicou TensorFlow, tak bolo potrebné na testovanie otvoriť sedenie (session). TensorFlow bola naša voľba kvôli štruktúre modelu, ktorý obsahoval checkpoint (uložené predtrénované data modelu) a uložený graf vo formáte tenzorov.

Na uloženie výsledkov sme použili knižnicu csv, ktorá vytvára, otvára a zapisuje údaje do csv súborov. Pre načítanie videí a prístup k ich jednotlivým snímkam sme použili knižnicu cv2, za pomoci ktorej sme taktiež upravili snímky na správne rozmery pre potreby modelu. Pri vytváraní umelých snímok z dôvodu, aby sme dostali správny rozmer pre vstupný tenzor sme využili opätovne knižnicu cv2. Po kategorizácii jednotlivých snímok sme spriemerovali výsledky pre jednotlivé videá a tieto údaje sme ukladali do csv súboru.

[21] Zdroj Obrázok 18: https://paperswithcode.com/dataset/sports-1m

C3D-UCF101

Ako aj pri modeli C3D-Sport1M sme pozerali na čo je zameraný dataset na ktorom natrénovali Mou et al. (2020) tento model.

Dataset UCF obsahoval viac než 13 000 videí kategorizovaných do dvadsaťpäť kategórií, ktoré sa delili do celkov:

- interakcia človeka s objektami,
- pohyb a aktivity ľudí,
- interakcia medzi ľuďmi,
- a športy.

Dataset predstavoval komplexný súbor videí, z ktorého sme predpokladali dobré výsledky pri testovaní modelu (Obrázok 19). (Soomro et al., 2012)

Obrázok 19 Dataset UCF101[22]

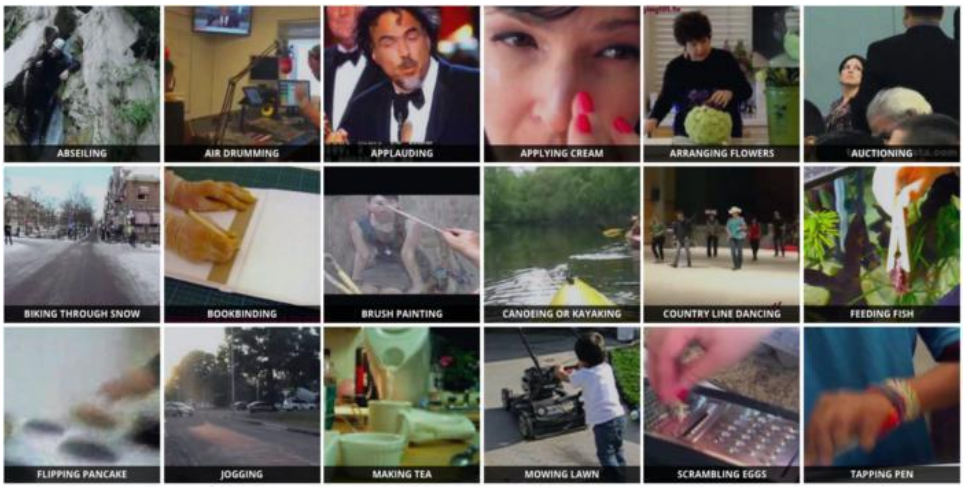
Štruktúra predtrénovaného modelu bola rovnaká ako pri C3D-Sport1M, model taktiež obsahoval štyri súbory, ktoré obsahovali jeho architektúru, váhy a vrstvy. Na otestovanie modelu sme postupovali rovnako ako u druhého C3D modelu.

[22] Zdroj Obrázok 19: https://www.crcv.ucf.edu/data/UCF101.php

P3D-ResNet-199_Kinetics a P3D-ResNet-199_Kinetics-600

Oba modely P3D vybudovali Mou et al. (2020) na architektúre ResNet a natrénovali na datasete Kinetics.

Tento dataset obsahoval okolo 500 000 videí zameraných na činnosti ľudí. Každé video malo okolo desať sekúnd a kategorizovalo sa do šesťsto tried (staršia forma datasetu mala štyristo kategórií). Videá v datasete zobrazovali napríklad ľudí vykonávajúcich aktivitu, ruky vykonávajúce aktivitu, športové aktivity, tváre ľudí a aj zábery rúk, ktoré manipulujú s objektami (Obrázok 20). (Carreira a Zisserman, 2017)



Obrázok 20 Dataset Kinetics[23]

Dataset Kinetics popisujú Kay et al. (2017) ako súbor videí, ktorých zdrojom bola platforma YouTube. Autori zahrnuli taktiež štatistické výsledky datasetu a výkonnosť. V práci popisujú aj ako vykonali predbežnú analýzu kde sa zameriavali na koreláciu medzi videami. Konkrétne ich cieľom bolo zistiť či nerovnováha v datasete vedie k skresleniu klasifikácie.

Architektúra na ktorej boli natrénované modely P3D vyžadovala menej priamočiary prístup ako pri testovaní C3D modelov. Pri C3D bolo potrebné pozrieť premenné a vybrať správne tenzory, ktoré sme načítali do prístupnej architektúry. Pri P3D sme si museli vytvoriť vlastnú architektúru na základe ResNet.

Keďže konkrétna architektúra ResNet199 nebola dostupná v knižnici TorchVision využili sme najbližšiu a teda ResNet152, ktorá obsahovala neurónovú sieť so 152 vrstvami. Predtrénovaný model sme týmto načítali do architektúry s váhami a checkpointom, avšak očakávali sme zhoršenie klasifikácie.

[23] Zdroj Obrázok 20: https://paperswithcode.com/dataset/kinetics

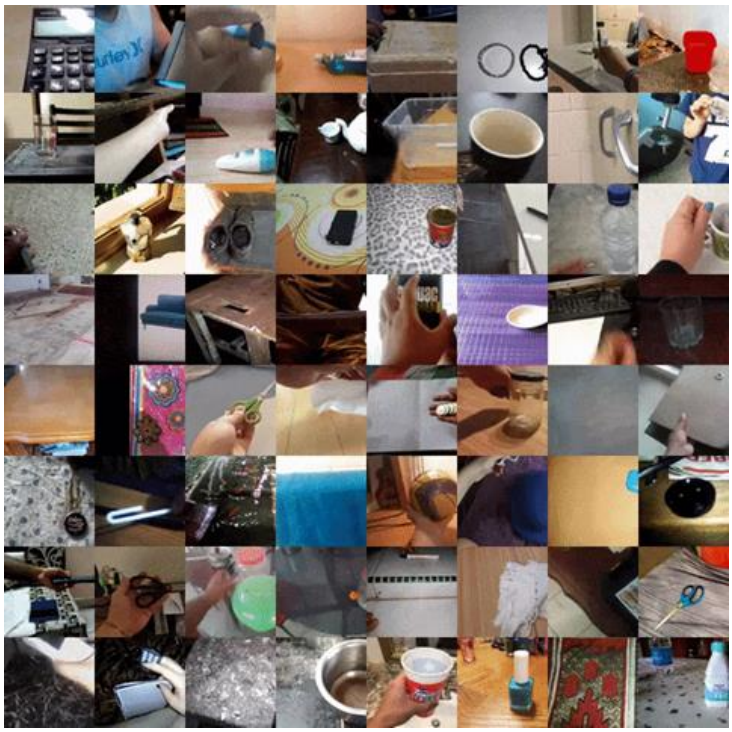
V prispôsobovaní modelovej architektúry sme taktiež riešili problém s počtom vstupných a výstupných nezhodujúcich sa kategórií, kedy sme pracovali s dvadsaťpäť, ale ResNet152 bol nastavený na dvetisícštyridsaťosem tried.

Okrem vyvárania triedy pre nastavenie ResNet152 sme postupovali rovnako ako pri ostatných modeloch, vybrané videá sme načítavali po snímkach, tie sme predikovali za pomoci modelu a v závere sme výsledky ukladali a zapisovali do csv súboru.

TRN-Something-Something-V2 a TRN-Moments-In-Time

Oba modely boli vybudované na TRN architektúre, ale natrénované na rôznych datasetoch. Dataset Something-Something V2 (Niečo-Niečo V2) predstavoval súbor okolo dvestodvadsať tisíc videí, ktoré sa zameriavali na gestá rúk a vykonávanie bežných vecí pomocou rúk. (Goyal et al., 2017)

Z toho sme vyvodili, že na rozdiel od datasetu ERA videá neobsahovali ľudí, stroje a ani autá. Medzi jeho kategórie patrili napríklad: pokladať niečo na podklad, posúvať niečo, hádzať niečo, pokladať niečo vedľa niečoho iného a mnoho ďalších (Obrázok 21).

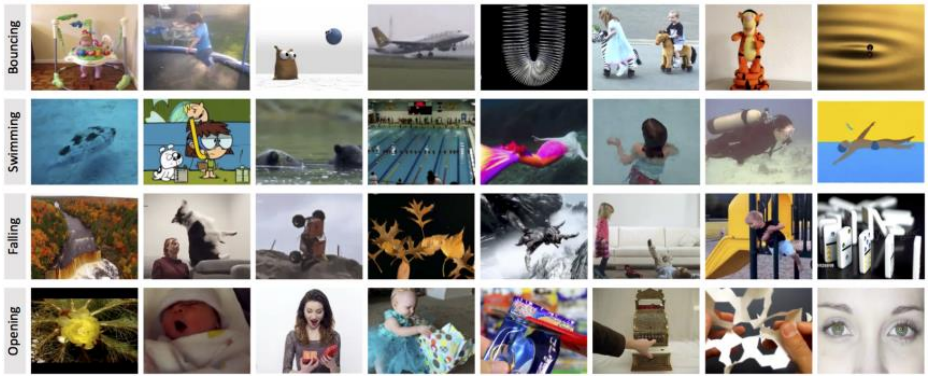


Obrázok 21 Dataset Something-Something V2[24]

Druhý dataset, Moments in Time (Momenty v čase), popísali autori Monfort et al. (2020) ako súbor milión zozbieraných videí, ktoré sa zameriavali na dynamické scény.

[24] Zdroj Obrázok 21: https://developer.qualcomm.com/software/ai-datasets/something-something

Predstavoval veľkú škálu videí od zvieracích, hýbajúcich sa objektov až po ľudí v pohybe (Obrázok 22).



Obrázok 22 Dataset Moments in Time[25]

Keďže modely, rovnako ako pri P3D, nemali dostupnú svoju architektúru, tak sme si na ich testovanie potrebovali vytvoriť vlastnú. Zo špecifikácií sme zistili, že oba sú vybudované na Inception architektúre.

Za pomoci TorchVision knižnice sme stiahli základnú kostru, ktorú sme obohatili o vrstvy a váhy modelov. Ďalší postup sme dodržali ako pri ostatných modelov, čiže po načítaní sme vytvorili csv súbor, spravili predikcie na vopred vybraných videách a zapísali naše predikcie na ďalšiu analýzu.

I3D-Kinetics a I3D-Kinetics+ImageNet

Modely I3D boli oba natrénované na datasete Kinetics, rovnako ako modely P3D. Jeden z modelov natrénovali Mou et al., (2020) taktiež na datasete ImageNet. Tento dataset predstavuje viac ako štrnásť miliónov obrázkov a každý má svoj minimálne jeden vlastný popis (Obrázok 23). Príkladom môže byť obrázok s popismi „na obrázku sú autá“ a „obrázok neobsahuje tigre“. Taktiež majú anotáciu „na obrázku je skrutkovač umiestnený na pozícii (20,25) so šírkou 50 pixelov a výškou 30 pixelov“. (Deng et al., 2009)

[25] Zdroj Obrázok 22: https://paperswithcode.com/dataset/moments-in-time

Obrázok 23 ImageNet dataset[26]

Modely I3D mali oba dostupnú architektúru a postup na testovanie mali podobný k modelom C3D. Najskôr sme načítali samotný model, tenzory a checkpoint. Následne sme kategorizovali pomocou modelu vybrané videá a výsledky sme zapisovali do csv súboru.

Pri načítavaní snímok videa sme narazili na komplikáciu, kedy sa rozmery vstupného tenzora nezhodovali s rozmermi snímok. Tento problém sme vyriešili zopakovaním snímok, aby sme mali požadovaný počet (šestnásť) do dimenzie tenzora na vytváranie nového rozmeru dávky (batch dimension).

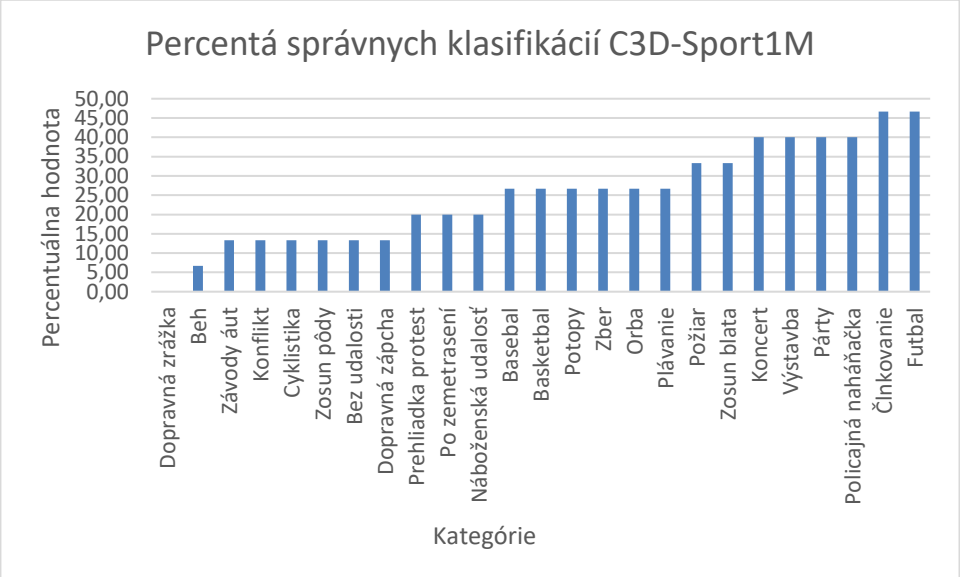
3.4 ANALÝZA DÁT

Ako hlavný parameter pre analýzu sme zvolili správnosť kategorizácie modelov (accuracy). Počas testovania sme pre každý model exportovali všetky výsledky do csv súboru. Tie sme následne spojili do jedného súboru a postupovali sme s analýzou

[26] Zdroj Obrázok 23: https://paperswithcode.com/dataset/imagenet

a vizualizáciou dát v programe Excel. Ako formu vizualizácie dát sme si zvolili tabuľky, do ktorých sme pridali názvy videí a ich kategorizácie z modelov. Následné výpočty sme zautomatizovali pomocou formúl a vstavaných funkcií Excelu.

Pre zvýšenie prehľadnosti naprieč výsledkami sme vytvorili grafy. Napríklad graf 1 zobrazuje percentuálnu hodnotu správnych kategorizácií pre model C3D-Sport1M. Taktiež sme vygenerovali grafy pre správnosť kategorizácie iných grafov, počet jednotlivých zaradení do kategórií ako aj grafy, ktoré sumarizujú výsledky. Všetky výsledky sú dostupné na našom GitHube[27].



Graf 1 Príklad vizualizácie pre C3D-Sport1M model[28]

Hypotézy

Pri príprave modelov na testovanie, sme si stanovili nasledovne hypotézy, ktoré sme vyhodnotili vo výsledkoch:

- H1: Predpokladáme, že model C3D-Sport1M bude lepšie klasifikovať videá zo športovej kategórie vzhľadom na to, že bol natrénovaný na športovom datasete.
- H2: Predpokladáme, že model C3D-UCF101 bude horšie kategorizovať videá v oblasti detekcie áut, poľnohospodárstva a pohrôm vzhľadom na natrénovaný dataset.
- H3: Predpokladáme, že sa zníži klasifikácia videí do kategórie párty pri modeli C3D-UCF101.

[27] https://github.com/batsysk/DiplomovaPraca
[28] Zdroj Graf 1: autor

- H4: Predpokladáme, že naše testovacie výsledky jednotlivých modelov sa budú približovať výsledkom autorov datasetu ERA.

3.5 VYUŽITÉ TECHNOLÓGIE

V tejto práci sme najmä využívali programovací jazyk Python (verzia Python 3.10.11). Pre načítanie a prácu s modelmi sme použili nasledovné knižnice:

- PyTorch (práca s modelmi TRN a P3D),
- TensorFlow (práca s modelmi C3D a I3D),
- NumPy (matematické operácie),
- cv2 (práca s videami),
- csv (zapisovanie a vytváranie Excel súborov)
- a os (práca so súborovým systémom a cestami).

Kód v jazyku Python sme písali v programovacom prostredí PyCharm (PyCharm Community Edition 2023.3.3) od spoločnosti JetBrains. Pre analýzu výsledkov sme využívali program Microsoft Excel a jeho funkcie, hlavne na urýchlenie výpočtov a dynamické analýzy. Pre ukladanie práce sme si zvolili platformu GitHub[29].

[29] https://github.com/batsysk/DiplomovaPraca

4 VÝSLEDKY

Pri práci s modelmi sme potrebovali často krát hľadať riešenia k problémom a optimalizácie našich riešení. Výsledky sme analyzovali, interpretovali a porovnali s výsledkami autorov ERA datasetu.

4.1 ANALÝZA VÝSLEDKOV

Po načítaní predtrénovaných modelov sme získali jednotlivé kategorizácie videí, ktoré sme následne vyhodnotili podľa ich správnosti (accuracy) klasifikácie, odôvodnili výsledky a zlepšili predikcie.

C3D-Sport1M

Pre výsledky modelu C3D sa naša hypotéza H1 nepreukázala. Pri analýze sme porovnali priemerné percento správnych predikcií pre všetky kategórie: 25,07 % a percentá priemerných predikcií pre športové kategórie (basketbal, basebal, kanoistika, cyklistika, beh, futbal, plávanie a závody áut). V tabuľke 3 môžeme vidieť, že oproti priemeru 25,07 % zo všetkých kategórií sa predikcie výrazne nezlepšili pri športových videách.

Tabuľka 3 C3D-Sport1M predikcie pre športy[30]

Názov kategórie	Priemerná správna predikcia v %
Basebal	26,67
Basketbal	26,67
Kanoistika	46,67
Cyklistika	13,33
Beh	6,67
Futbal	46,67
Plávanie	26,67
Závody áut	13,33

Dôvod vidíme ten, že model bol síce natrénovaný na komplexnom datasete športov (Sport1M), ale dataset obsahuje videá ktoré sú v rôznej kvalite, veľkosti a dĺžke, čiže môže byť natrénovaný lepšie vo všeobecnosti, ale nie na videách z datasetu ERA. Tie sú jednotné v parametroch a hlavne sú snímané z pohľadu z vrchu (dronové zábery).

[30] Zdroj Tabuľka 3: autor

Pri porovnaní našich výsledkov s výsledkami autorov datasetu ERA, sme neidentifikovali veľké odchýlky (Tabuľka 4).

Tabuľka 4 Porovnanie výsledkov nášho testovania s ERA výsledkami pre C3D-Sport1M[31]

Názov kategórie	Priemerná správna predikcia v %	C3DI – ERA accuracy v %	C3DII – ERA accuracy v %
Basebal	26,67	40,9	45,7
Basketbal	26,67	37,0	48,9
Kanoistika	46,67	47,5	41,9
Závody áut	13,33	16,7	18,2
Koncert	40,00	38,2	32,0
Konflikt	13,33	18,2	11,1
Výstavba	40,00	45,5	40,0
Cyklistika	13,33	20,6	13,6
Požiar	33,33	30,9	32,7
Potopy	26,67	24,3	56,5
Zber	26,67	27,5	42,3
Zosun pôdy	13,33	19,5	10,2
Zosun blata	33,33	32,9	23,9
Bez udalosti	13,33	29,6	28,5
Prehliadka protest	20,00	37,8	28,1
Párty	40,00	25,8	17,4
Orba	26,67	36,1	31,1
Policajné prenasledovanie	40,00	50,0	51,9
Po zemetrasení	20,00	23,1	27,9
Náboženská udalosť	20,00	27,5	35,8
Beh	6,67	12,0	9,3
Futbal	46,67	58,3	41,9
Plávanie	26,67	36,2	38,2
Dopravná zrážka	0,00	7,0	8,3
Dopravná zápcha	13,33	15,5	38,5
Priemer v %	25,07	30,4	31,1

Tabuľka 4 obsahuje názov kategórie, naše výsledky z testovaní a dva stĺpce údajov z testovaní od autorov. Naše priemery sa zhodovali s výsledkami prvého modelu C3D.

[31] Zdroj Tabuľka 4: autor

V priemere naše testovanie malo odchýlku od C3DI – ERA o 7,38 %. Dôvodom odchýlky mohlo byť množstvo testovacích dát, ktoré je iba pätnásť videí z každej kategórie.

Kategória párty (party) bola najviac predikovaná a to s počtom dvestosedem krát. Čo bolo 55,2 % z celkového počtu prípadov. Dôvodom mohlo byť, že väčšina datasetu obsahovala ľudí a kategória party obsahovala videá s ľuďmi nahromadenými na jednom mieste vo väčšom počte. Túto vlastnosť spĺňali videá aj v iných kategóriách.

Kategória dopravných zrážok nemala ani jednu správnu predikciu, avšak v prípade výsledkov autorov datasetu išlo tiež o malé percento a teda iba 7 %. Prvý dôvod mohol byť malý testovací počet videí, druhým dôvodom mohlo byť slabé odhadovanie udalostí s autami vo všeobecnosti (pri zápche áut to bolo 13,33 %, u autorov 15,5 %). Dôvodom môže byť horšia predikcia modelu na videá s autami. Namiesto správnej kategorizácie sme dostávali najčastejšie predikciu pre párty.

Zaujímavé pozorovanie bolo medzi najviac predikovanou kategóriou party naprieč všetkými videami s počtom 207 a nízkou správnosťou kategorizácie pre party, iba 40 % prípadov. Najlepšie percento pre správnu kategorizáciu mali kanoistika a futbal s 46,67%.

Najväčší rozdiel v testovaných dátach autorov a našich sme identifikovali pri kategórii prehliadka/protest, kde bol rozdiel 17,80 %.

V prípade tohto modelu, vzhľadom na malý rozdiel v percente (7,82 %) správne kategorizovaných videí, pokladáme hypotézu H4 za splnenú.

C3D-UCF101

Tento model bol natrénovaný pomocou datasetu zameraného na ľudí a športy (UCF101). Podľa hypotézy H2 sme predpokladali nižšie kategorizovanie v kategóriách: závody áut, výstavba, požiar, záplavy, zber, zosun pôdy, zosun bahna, orba, po zemetrasení, dopravná zrážka a dopravná zápcha. Dôvodom je zameranie datasetu na ľudí a aktivity zahrňujúce snímanie rúk.

Náš predpoklad sa čiastočne splnil, medzi najhoršie predikované kategória patrili: koncert, konflikt, cyklistika, zosun pôdy, beh a zrážka áut (Tabuľka 6). Tieto kategórie môžeme porovnať s priemernou správnou kategorizáciou naprieč všetkými videami: 28,53 %.

Tabuľka 5 Porovnanie kľúčových kategórií[32]

Názov kategórie	Accuracy v %
Kanoistika	46,67
Závody áut	20,00
Výstavba	46,67
Požiar	26,67
Záplavy	40,00
Zber	33,33
Zosun pôdy	6,67
Zosun bahna	26,67
Orba	33,33
Po zemetrasení	26,67
Dopravná zrážka	6,67
Dopravná zápcha	33,33

Všetky tieto kategórie sa nevyskytovali v kategorizácii datasetu UCF. Avšak ostatné nešportové, poľnohospodárske kategórie a kategórie s detekciou áut a pohrôm nemali výrazne nízku predikciu (Tabuľka 5).

Podľa Tabuľky 6 vidíme, že oproti predošlému modelu C3D sa nám priemerné percento predikcie zvýšilo o 3,47 % (pri C3D-Sport1M to bolo 28,53 %). Je to vidieť aj na jednotlivých predikciách, napríklad všetky kategórie boli predikované aspoň v jednom prípade v danej kategórii správne (v C3D-Sport1M sme nemali zastúpenú kategóriu dopravnej zrážky). Hypotézu H4 môžeme pokladať, pre tento model, za preukázanú.

Kategória párty mala správnu predikciu až v 66,67 %, čo bolo zlepšenie o 26,67 % oproti modelu C3D-Sport1M. Ale taktiež to bola najčastejšie predikovaná kategória a to stosedemdesiatdva krát.

Najväčší rozdiel medzi našimi výsledkami a výsledkami autorov ERA datasetu sme pozorovali v kategóriách: párty (stúpnutie o 40,87 %) a koncert (pokles o -21,53 %). Model, v týchto kategóriách, predikoval častejšie kategórie párty a výstavbu. Keďže ani jedna z týchto kategórií nebola obsiahnutá v datasete UCF, tak sme nevedeli s určitosťou odôvodniť toto kategorizovanie.

Hypotéza H3 sa nesplnila, keďže sa počet kategorizácie párty výrazne neznížil. Predpoklad vznikol pri skúmaní datasetu UCF, ktorý obsahoval videá zosnímané z boku

[32] Zdroj Tabuľka 5: autor

osoby a bolo na nich vidieť jednotlivých ľudí. Predikcie pre kategóriu párty boli iba o niečo nižšie ako pri C3D-Sport1M (UCF101: 172 krát, Sport1M: 207 krát).

Tabuľka 6 Porovnanie výsledkov nášho testovania s ERA výsledkami pre C3D-UCF101[33]

Názov kategórie	Priemerná správna predikcia v %	C3DI – ERA accuracy v %	C3DII – ERA accuracy v %
Basebal	40,00	40,90	45,70
Basketbal	46,67	37,00	48,90
Kanoistika	46,67	47,50	41,90
Závody áut	20,00	16,70	18,20
Koncert	6,67	38,20	32,00
Konflikt	6,67	18,20	11,10
Výstavba	46,67	45,50	40,00
Cyklistika	6,67	20,60	13,60
Požiar	26,67	30,90	32,70
Potopy	40,00	24,30	56,50
Zber	33,33	27,50	42,30
Zosun pôdy	6,67	19,50	10,20
Zosun blata	26,67	32,90	23,90
Bez udalosti	26,67	29,60	28,50
Prehliadka protest	20,00	37,80	28,10
Párty	66,67	25,80	17,40
Orba	33,33	36,10	31,10
Policajné prenasledovanie	33,33	50,00	51,90
Po zemetrasení	26,67	23,10	27,90
Náboženská udalosť	33,33	27,50	35,80
Beh	6,67	12,00	9,30
Futbal	40,00	58,30	41,90
Plávanie	33,33	36,20	38,20
Dopravná zrážka	6,67	7,00	8,30
Dopravná zápcha	33,33	15,50	38,50
Priemer v %	28,53	30,40	31,10

P3D-ResNet-199_Kinetics a P3D-ResNet-199_Kinetics-600

Oba P3D modely nekategorizovali správne. Zakaždým priraďovali každé video do rovnakej jednej kategórie. Dôvod vidíme v zle nastavenej architektúre, kedy sme nemali

[33] Zdroj Tabuľka 6: autor

možnosť načítať buď ResNet199 alebo použiť architektúru od autorov. Modely neobsahovali túto architektúru ako tomu bolo pri C3D modeloch a teda sme použili najbližšiu možnú alternatívu. Pri práci s neurónovými sieťami je to avšak nesprávna cesta a teda naše výsledky boli očakávané.

Keďže ResNet199 nie je dostupná architektúra v knižnici TorchVision, tak sme modely skúšali na viacerých podobných, menovito: ResNet18, ResNet34, ResNet101 a alternatívach k nim ResNeXt (architektúry vybudované na zlepšenie výkonnosti, predstavujú parameter kardinality). Naše výsledky boli zhodné vo všetkých architektúrach a teda vo výsledku modely nekategorizovali správne.

V ďalšom kroku sme skúsili vylepšiť predikcie modelu s architektúrou ResNet199. Ako prvé sme skúsili metódu dočasného zoskupovania (temporal pooling). Zoskupovanie sme pridali pri určovaní kategórie na konci, keď už sme mali všetky predikcie uložené v listovej štruktúre. Implementovali sme výpočet, ktorý zobral set všetkých predikcií, vybral maximum a priradil ho ku kategórii. (Kód 2)

```
predicted_category_index = max(set(all_predictions), key=all_predictions.count)
```

Kód 2 Výpočet dočasného zoskupenia[34]

Metóda zoskupenia vo výsledku nezlepšila kategorizovanie a tieto modely nesplnili predpoklad H4.

TRN-Something-Something-V2 a TRN-Moments-In-Time

Modely TRN pri základnom načítaní a otestovaní nekategorizovali správne. Pri každom z videí priraďovali rovnakú kategóriu (Something kategorizoval „Po Zemetrasení“ a Moments kategorizoval každé video ako „Zosun pôdy“). Takže ich kategorizácia bola správna len v pätnástich prípadoch pri každom modeli.

Pre zlepšenie kategorizácií sme zvolili transformácie a metódu vzorkovania (sampling method). Ako transformácie sme skúsili normalizáciu a pomocné metódy pre prácu so snímkami (PILImage a zmena rozmeru snímok). Pre vzorkovanie sme vybrali štandardných desať snímok ako podklad pre jednotlivé vzorky.

Vo výsledku sme dostali lepšie kategorizácie, modely začali kategorizovať do všetkých kategórií a nedostali sme iba jednu výslednú kategóriu. Avšak vo výsledku

[34] Zdroj Kód 2: autor

kategorizácií sme neprišli k veľkému zlepšeniu percenta správnosti. TRN-Something kategorizoval správne iba osemnásť videí a TRN-Moments pätnásť. Aj keď počet nebol lepší, tak predikcie boli aspoň rôznorodé.

V ďalšom kroku zlepšení predikcií sme implementovali časovú agregáciu (temporal aggregation). Agregáciu sme implementovali ako zoznam výstupov kategorizácií z modelu, ktoré sme spriemerovali po časovom úseku a tým vytvorili kontext v určitom čase. Následne sme spravili priemer týchto časových kontextov (Kód 3).

```
temporal_sequence.append(outputs.cpu().numpy())
        if len(temporal_sequence) == sequence_length:
            aggregated_predictions = np.mean(temporal_sequence, axis=0)
            all_predictions.append(aggregated_predictions)
            temporal_sequence = []
```

Kód 3 Časový kontext implementovaný pre TRN modely[35]

Implementácia časového kontextu nám zlepšila kategorizovanie o niekoľko správnych zaradení do tried. Konkrétne pre TRN-Something sme mali dvadsaťšesť a pre TRN-Moments devätnásť správnych kategorizácií videí. Taktiež sa zachovalo správanie modelov, kedy kategorizovali do všetkých tried a nie iba pre do jednej.

[35] Zdroj Kód 3: autor

Vo výsledku sa ale kategorizovanie nepriblížilo percentu správnych zaradení do tried autorom ERA datasetu a teda hypotéza H4 sa nepreukázala (Tabuľka 7).

Tabuľka 7 Porovnanie výsledkov naprieč modelmi[36]

Verzia a názov modelu	Počet správnych kategorizácií	Accuracy v %	ERA accuracy v %
TRN-Something-Something-V2 základ	15	4,00	62,00
Something-Something-V2 metóda vzorkovania	18	4,80	-
TRN-Something-Something-V2 časový kontext	26	6,93	-
TRN-Moments-In-Time základ	15	4,00	64,30
TRN-Moments-In-Time metóda vzorkovania	15	4,00	-
TRN-Moments-In-Time časový kontext	19	5,07	-

I3D-Kinetics a I3D-Kinetics+ImageNet

Pri testovaní modelov I3D sme popri základnom načítaní a otestovaní dostali kategorizáciu všetkých videí do jednej kategórie (pri IC3-Kinetics kategória „Párty“ a pri I3D-Kinetics+ImageNet kategória „Náboženská aktivita“), to znamená, že správne kategorizovali v pätnásť prípadoch.

Ako prvé zlepšenie kategorizácií sme skúsili metódu vzorkovania na šestnástich snímkach. Po implementovaní tejto metódy sme dostali predikcie do všetkých kategórií a to pri oboch modeloch. Avšak celkové správne kategorizovanie sa nezlepšilo, pri I3D-Kinetics bolo správne kategorizovaných iba štrnásť videí a pri I3D-Kinetics+ImageNet to bolo desať videí.

[36] Zdroj Tabuľka 7: autor

Ako druhú sme implementovali náhodnú augmentáciu snímok (random augmentation). Táto metóda využíva náhodné transformácie snímok, konkrétne otočenie o stoosemdesiat stupňov a orezanie (Kód 3)

```
def apply_augmentation(frame):
    if random.choice([True, False]):
        frame = cv2.flip(frame, 1)
    x1, y1 = random.randint(0, 200), random.randint(0, 200)
    x2, y2 = random.randint(440, 640), random.randint(440, 640)
    frame = frame[y1:y2, x1:x2]
    return frame
```

Kód 4 Implementácie náhodnej augmentácie[37]

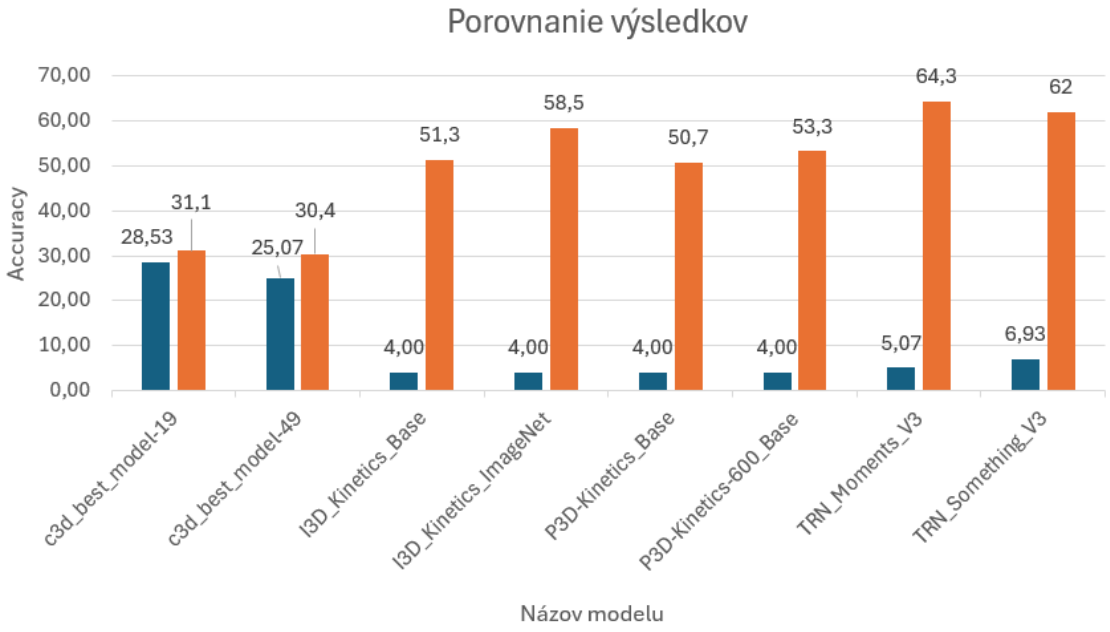
Augmentácia nepriniesla zlepšenie správnosti kategorizácie. Výsledky sme mali v skoro všetkých kategóriách, ale správne sme dostali pri I3D-Kinetics iba osem videí a pri I3D-Kinetics+ImageNet deväť videí.

Hypotéza H4 sa nepreukázala pre tieto modely z dôvodu malého percenta správnych kategorizácií, ktoré sa nepribližuje výsledkom autorov datasetu.

[37] Zdroj Kód 4: autor

4.2 VYHODNOTENIE, DISKUSIA A ODPORÚČANIA

V našej práci sme porovnávali správnosť predtrénovaných modelov na datasete ERA. Modely okrem C3D-Sport1M a C3D-UCF101 nemali podobné výsledky s výsledkami autorov datasetu ERA (Graf 2).



Graf 2 Porovnanie výsledkov[38]

Hlavným dôvodom pre tento rozdiel vidíme možné zlé načítanie a otestovanie modelov. Pri práci s modelmi sme vystupovali z informácií dostupných v dokumentáciách knižníc jazyka Python a iných výskumov, ktoré riešili podobnú problematiku. Keďže sme nemali dostupne informácie ako prebehlo trénovanie modelov, akú architektúru presne zvolili autori alebo akým spôsobom testovali tieto modely, tak vidíme tento nedostatok informácií ako hlavný dôvod rozdielov.

Pri testovaní sme dostali pri jednom spôsobe exportu modelov úplne odlišné výsledky. Modely C3D a I3D mali oba rovnakú štruktúru exportovaných súborov, takže ich testovanie bolo realizované na rovnakom základe s podobným postupom. Bolo potrebné vybrať správne parametre a zvoliť tenzory. Pri testovaní rovnakých vidí sme pre C3D modely dostali veľmi podobné výsledky ako prezentovali Mou et al. (2020). Avšak pre I3D naše výsledky boli úplne odlišné.

[38] Zdroj Graf 2: autor

Pri druhom spôsobe exportovania súborov (modely TRN a P3D) sme mali výsledky podobné, ale oba modely nekorešpondovali s výsledkami autorov Mou et al. (2020).

Našim ďalším pokusom o priblíženie sa k výsledkom autorov boli rôzne optimalizácie. S ich implementáciou sme dosiahli zlepšenie správnosti kategorizácie videí, avšak nejednalo sa o dostatočne štatisticky významné navýšenie, aby bolo považované za úspešné.

Ako ďalší dôvod odlišnosti vo výsledkoch mohla byť veľkosť testovacích dát. Autori rozdelili dataset na trénovacie a testovacie videá, s jednotlivými kategóriami, avšak z dôvodu obmedzenia našich technických prostriedkov bolo potrebné testovací dataset zmenšiť. To mohlo skresliť výsledné dáta, keďže naša testovacia vzorka neobsahovala úplný počet videí.

V rámci optimalizácie výsledkov sme skúmali štyri metódy, ktoré nám pri viacerých modeloch priniesli zlepšenie. Pre ďalší výskum navrhujeme vyskúšanie iných metód poprípade natrénovanie modelov od začiatku.

Dataset ERA sám o sebe je zostavený veľmi dobre a odporúčame jeho využitie najmä pri modeloch s menším dôrazom na presnosť kategorizácií. Poprípade taktiež na využitie vo sfére vzdelávania. Vidíme jeho prínos vzhľadom na celistvú štruktúru, dobré delenie, označenie videí a nie priveľký počet videí.

Na základe dobrého členenia videí do kategórií a oblastí, sa dá ľahko využiť menšia časť videí a zlúčiť viaceré datasety pre komplexné natrénovanie modelu, podobne ako to spravili Mou et al. (2020). Dataset ERA sa najmä odlišuje od ostatných tým, že videá zbiera ako dronové snímky, ktoré nebývajú zastúpené bežne vo veľkých datasetoch.

Pri analýze datasetu sme taktiež pozorovali, že autori sa snažili obsiahnuť širší záber oblastí. Dataset obsahoval od športových videí cez videá z dopravy až po prírodné katastrofy. Toto môže byť aj dôvod prečo presnosť nebola úplne najvyššia. Z nášho pohľadu by bolo lepšie zamerať sa na konkrétnu oblasť a natrénovať modely len na predikciu napríklad druhov športov alebo pohrôm. Podobne sa zameriavajú aj iné datasety a vidíme v tom prínos aj pri trénovaní modelov, ktoré by sa mohli užšie špecializovať. Tým by mohli mať modely lepšie predikcie a skôr využitie v reálnom svete, kde sa netrénujú modely aby boli univerzálne naprieč rôznymi oblasťami.

Ako ďalší prínos vidíme vo vložení snímok obsahujúcich šum alebo videí do datasetu, ktoré by predstavovali priblíženie k realite a predchádzali by preučeniu. To by sa dalo uskutočniť či už viacerými transformáciami alebo prelínaním viacerých datasetov.

Vzhľadom na trénovanie a testovanie modelov, by sme odporúčali natrénovať vlastné modely, poprípade použiť našu prácu ako podkladový materiál.

ZÁVER

Hlavným cieľom diplomovej práce bolo popísať tému rozpoznávania objektov vo videách, ktorú sme naplnili. Pre tento cieľ sme si vybrali dataset ERA, ku ktorému natrénovali autori taktiež modely C3D, I3D, P3D a TRN. K nim sme pridali aj teoretický popis ďalších dvoch a teda TSN a LSTM. Tieto modely sme porovnali medzi sebou v analýze a na testovanie sme identifikovali štyri hypotézy. Okrem samotného testovania sme sa v práci venovali taktiež analýzou iných datasetov ako napríklad Sport1M, UCF101, Kinetics alebo Something-Something-V2. Popri nich sme popísali dve architektúry, ktoré využívali samotné modely: ResNet a Inception.

Dataset ERA hodnotíme ako využiteľný na prácu s klasifikáciou, keďže videá sú jednotné v ich dĺžke, rozmeroch a frekvencii. Ako veľký potenciál pre ERA vidíme to, že pozostáva z dronových záberov. Všeobecne sa na internete zvyšuje množstvo tohto ruhu snímania videí každým dňom a datasety ich v dnešnej dobe nevyužívajú plne.

Predtrénované modely sme testovali na videách, ktoré rozdelili a dali do kategórií Mou et al. (2020), autori datasetu ERA. Tento súbor dát sme avšak z dôvodu technickej náročnosti skrátili na podmnožinu 15x25 videí (v každej z 25 kategórií, sme vybrali 15 videí). Toto obmedzenie mohlo viesť k zhoršeniu výsledkov. Taktiež nezhodné výsledky z testovaní sa dajú odôvodniť absenciou dokumentácie, ktorá by popisovala ako autori trénovali modely, ako vyzerali architektúry a aké nastavenia použili pre neurónové siete.

Na to, aby sme zlepšili výsledky, sme pridali variácie k testovaným modelom v podobe optimalizačných metód. Tie sme najskôr popísali v analytickej časti práce a následne pridali do Python kódov, kde sme ich následne otestovali na videách. Na uskutočnenie testovania sme v skriptoch jazyku Python využili knižnice používané pri práci s neurónovými sieťami ako napríklad PyTorch alebo Tensorflow.

Vo fáze porovnávaní a analýz výsledkov sme použili metriku správnosti (accuracy), ktorá vyjadrovala s akou percentuálnou presnosťou model predikoval správne na daných videách kategórie.

Na zosumarizovanie výsledkov sme využili tabuľky v súbore csv, kde sme prehľadne zdokumentovali jednotlivé kategorizácie a vytvorili formuly na rýchlejšie vyhodnotenie. Grafy nám poslúžili na sprehľadnenie výsledkov a poukázanie na rozdiely, ktoré sme následne použili aj v tejto práci.

Práca sa dá využiť ako podkladový materiál pre ďalšie výskumy a to vzhľadom na komplexný popis problematiky, fungovania modelov, optimalizačných metód a architektúr v kapitole s analýzou súčasného stavu. Taktiež sme vytvorili programy, ktoré testovali predtrénované modely, kedy podobné zdokumentovanie nie je oficiálne dostupné od autorov. Medzi potencionálne oblasti pre ďalšie výskumy by sme zaradili v prvom rade zlepšenie kategorizácie modelov a optimalizovanie výsledkov. Taktiež sme v kapitole s diskusiou navrhli nové implementácie, ktoré by modely pomohli využiť pri ďalšej úprave. Okrem dodatočných výskumov vidíme využiteľnosť práce aj v edukačnej sfére, kde by sa dala využiť ako podklad pre prácu s predtrénovanými modelmi alebo s datasetom ERA.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

ABU-EL-HAIJA, Sami, Nisarg KOTHARI, Joonseok LEE, Paul NATSEV, George TODERICI, Balakrishnan VARADARAJAN and Sudheendra VIJAYANARASIMHAN, 2016. YouTube-8M: a Large-Scale Video Classification Benchmark. arXiv.org [online]. Dostupné na: https://arxiv.org/abs/1609.08675

ALZUBAIDI, Laith, Jinglan ZHANG, Amjad J. HUMAIDI, Ayad Q. AL-DUJAILI, Ye DUAN, Omran AL-SHAMMA, José SANTAMARÍA, Mohammed A. FADHEL, Muthana AL-AMIDIE and Laith FARHAN, 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data [online]. 2021, vol. 8, no. 1. Dostupné na: doi:10.1186/s40537-021-00444-8

BEKHET, Saddam and Abdullah M. ALGHAMDI, 2021. A COMPARATIVE STUDY FOR VIDEO CLASSIFICATION TECHNIQUES USING DIRECT FEATURES MATCHING, MACHINE LEARNING, AND DEEP LEARNING. Journal of Southwest Jiaotong University [online]. 2021. Dostupné na: http://jsju.org/index.php/journal/article/view/994

ATHANASOPOULOS, George, Nikolaos KOURENTZES and Fotios PETROPOULOS, 2017. Forecasting with temporal hierarchies. *European Journal of Operational Research* [online]. 2017, vol. 262, no. 1, pp. 60–74. Dostupné na: doi:10.1016/j.ejor.2017.02.046

BABU, Deepak and Melissa J. FULLWOOD, 2015. 3D genome organization in health and disease: emerging opportunities in cancer translational medicine. *Nucleus (Austin, Tex. Online)* [online]. 2015, vol. 6, no. 5, pp. 382–393. Dostupné na: doi:10.1080/19491034.2015.1106676

BAREKATAIN, Mohammadamin, Miquel MARTI, Hsueh-Fu SHIH, Samuel MURRAY, Kotaro NAKAYAMA, Yutaka MATSUO and Helmut PRENDINGER, 2017. Okutama-Action: an aerial view video dataset for concurrent human action detection [online]. Dostupné na: https://openaccess.thecvf.com/content_cvpr_2017_workshops/w34/html/Barekatain_Okutama-Action_An_Aerial_CVPR_2017_paper.html

BUCKLAND, Stephen T., Eric A. REXSTAD, Tiago A. MARQUES and C. S. OEDEKOVEN, 2015. The basic methods. In: *Methods in statistical ecology* [online]. p. 3–13. Dostupné na: doi:10.1007/978-3-319-19219-2_1

CARREIRA, João and Andrew ZISSERMAN, 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *IEEE Conference on Computer Vision and Pattern Recognition* [online]. 2017. Dostupné na: doi:10.1109/cvpr.2017.502

COMOLI, Patrizia, Michela CIONI, Augusto TAGLIAMACCO, Giuseppe QUARTUCCIO, Annalisa INNOCENTE, I. FONTANA, Antonella TRIVELLI, Alberto MAGNASCO, Angela NOCCO, Catherine KLERSY, Laura RUBERT, Miriam RAMONDETTA, Marco ZECCA, Giacomo GARIBOTTO, Gian Marco GHIGGERI, Massimo CARDILLO, Arcangelo NOCERA and Fabrizio GINEVRI, 2016. Acquisition of C3D-Binding activity by de novo Donor-Specific HLA antibodies correlates with graft loss in nonsensitized pediatric kidney recipients. *American Journal of Transplantation (Print)* [online]. 2016, vol. 16, no. 7, pp. 2106–2116. Dostupné na: doi:10.1111/ajt.13700

DE MELO, Wheidima Carneiro, Éric GRANGER and Abdenour HADID, 2019. Combining Global and Local Convolutional 3D Networks for Detecting Depression from Facial Expressions. *IEEE International Conference on Automatic Face & Gesture Recognition* [online]. 2019. Dostupné na: doi:10.1109/fg.2019.8756568

DELEXTRAT, Anne and Daniel D. COHEN, 2009. Strength, power, speed, and agility of women basketball players according to playing position. *Journal of Strength and Conditioning Research* [online]. 2009, vol. 23, no. 7, pp. 1974–1981. Dostupné na: doi:10.1519/jsc.0b013e3181b86a7e

DENG, Jia, Wei DONG, Richard SOCHER, Lijia LI, Kai LI and Feifei LI, 2009. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition [online]. 2009. Dostupné na: doi:10.1109/cvpr.2009.5206848

GAMA, Fernando, Antonio G. MARQUES, Alejandro RIBEIRO and Geert LEUS, 2018. MIMO graph filters for convolutional neural networks. *arXiv.org* [online]. Dostupné na: http://arxiv.org/abs/1803.02247

GILL, Jagreet Kaur, 2023. Inception Architecture for Computer Vision and its Future. *XenonStack* [online]. Dostupné na: https://www.xenonstack.com/blog/inception-architecture-computer-vision

GOYAL, Raghav, Samira Ebrahimi KAHOU, Vincent MICHALSKI, Joanna MATERZYŃSKA, Susanne WESTPHAL, Heuna KIM, Valentin HAENEL, Ingo FRUEND, Peter YIANILOS, Moritz MUELLER-FREITAG, Florian HOPPE, Christian THURAU, Ingo BAX and Roland MEMISEVIC, 2017. The "something something" video database for learning and evaluating visual common sense. arXiv.org [online]. Dostupné na: https://arxiv.org/abs/1706.04261

HADIDI, Ramyad, Jiashen CAO, Michael S. RYOO and Hyesoon KIM, 2020. Toward collaborative inferencing of deep neural networks on Internet-of-Things devices. *IEEE Internet of Things Journal (Online)* [online]. 2020, vol. 7, no. 6, pp. 4950–4960. Dostupné na: doi:10.1109/jiot.2020.2972000

HAMEED, Mazhar, Fengbao YANG, Sibghat Ullah BAZAI, Muhammad Imran GHAFOOR, Ali ALSHEHRI, Ilyas KHAN, Mehmood BARYALAI, Mulugeta ANDUALEM and Fawwad Hassan JASKANI, 2022. Urbanization detection using LIDAR-Based remote sensing images of Azad Kashmir using novel 3D CNNs. Journal of Sensors (Print) [online]. 2022, vol. 2022, pp. 1–9. Dostupné na: doi:10.1155/2022/6430120

CHATTOPADHYAY, Chiranjoy and Amit Kumar MAURYA, 2013. Genre-specific modeling of visual features for efficient content based video shot classification and retrieval. *International Journal of Multimedia Information Retrieval* [online]. 2013, vol. 2, no. 4, pp. 289–297. Dostupné na: doi:10.1007/s13735-013-0034-8

CHAYAMBUKA, Kudakwashe, Jan FRANSAER and Xochitl DOMÍNGUEZ-BENETTON, 2019. Modeling and design of semi-solid flow batteries. *Journal of Power Sources (Print)* [online]. 2019, vol. 434, p. 226740. Dostupné na: doi:10.1016/j.jpowsour.2019.226740

IBM, 2021. *IBM documentation* [online]. Dostupné na: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

JAZAERY, Mohamad Al and Guodong GUO, 2021. Video-Based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing* [online]. 2021, vol. 12, no. 1, pp. 262–268. Dostupné na: doi:10.1109/taffc.2018.2870884

KARASAWA, Hiroki, Chien-Liang LIU and Hayato OHWADA, 2018. Deep 3D convolutional neural network architectures for Alzheimer's disease diagnosis. In: Lecture notes in computer science [online]. p. 287–296. Dostupné na: doi:10.1007/978-3-319-75417-8_27

KARPATHY, Andrej, George TODERICI, Sanketh SHETTY, Thomas LEUNG, Rahul SUKTHANKAR and Feifei LI, 2014. Large-Scale Video Classification with Convolutional Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition* [online]. 2014. Dostupné na: doi:10.1109/cvpr.2014.223

KAY, Will, Joao CARREIRA, Karen SIMONYAN, Brian ZHANG, Chloe HILLIER, Sudheendra VIJAYANARASIMHAN, Fabio VIOLA, Tim GREEN, Trevor BACK, Paul NATSEV, Mustafa SULEYMAN and Andrew ZISSERMAN, 2017. The Kinetics Human Action Video Dataset. arXiv.org [online]. Dostupné na: https://arxiv.org/abs/1705.06950

KHAN, Shakir and Lulwah ALSUWAIDAN, 2022. Agricultural monitoring system in video surveillance object detection using feature extraction and classification by deep learning techniques. Computers & Electrical Engineering [online]. 2022, vol. 102, p. 108201. Dostupné na: doi:10.1016/j.compeleceng.2022.108201

KORTYLEWSKI, Adam, Qing LIU, Angtian WANG, Yihong SUN and Alan YUILLE, 2020. Compositional Convolutional Neural Networks: A Robust and Interpretable Model for Object Recognition under Occlusion. *arXiv.org* [online]. Dostupné na: http://arxiv.org/abs/2006.15538

KOURENTZES, Nikolaos and Fotios PETROPOULOS, 2016. Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics* [online]. 2016, vol. 181, pp. 145–153. Dostupné na: doi:10.1016/j.ijpe.2015.09.011

KUEHNE, Hilde, Hueihan JHUANG, Estíbaliz GARROTE, Tomaso POGGIO and T. SERRE, 2011. HMDB: A large video database for human motion recognition. 2011 International Conference on Computer Vision [online]. 2011. Dostupné na: doi:10.1109/iccv.2011.6126543

KYRKOU, Christos and Theocharis THEOCHARIDES, 2020. EmergencyNet: Efficient aerial image classification for Drone-Based emergency monitoring using Atrous convolutional Feature Fusion. *IEEE Journal of Selected Topics in Applied Earth*

Observations and Remote Sensing (Print) [online]. 2020, vol. 13, pp. 1687–1699. Dostupné na: doi:10.1109/jstars.2020.2969809

LI, Meng, Yan ZHANG, Haicheng SHE and Li ZHANG, 2018. Automated segmentaiton and classification of arterioles and venules using Cascading Dilated Convolutional... ResearchGate [online]. 2018. Dostupné na: https://www.researchgate.net/publication/329388333_Automated_segmentaiton_and_classification_of_arterioles_and_venules_using_Cascading_Dilated_Convolutional_Neural_Networks

LI, Qingwu, Haisu CHENG, Yan ZHOU and Guanying HUO, 2016. Human action recognition using improved salient dense trajectories. *Computational Intelligence and Neuroscience (Print)* [online]. 2016, vol. 2016, pp. 1–11. Dostupné na: doi:10.1155/2016/6750459

LI, Sihan, Jiantao JIAO, Yanjun HAN and Tsachy WEISSMAN, 2017. Demystifying ResNet. *arXiv (Cornell University)* [online]. 2017. Dostupné na: https://openreview.net/pdf?id=SJAr0QFxe

LI, Yan, 2022. Application of Computer Vision in Intelligent Manufacturing under the Background of 5G Wireless Communication and Industry 4.0. *Mathematical Problems in Engineering (Print)* [online]. 2022, vol. 2022, pp. 1–9. Dostupné na: doi:10.1155/2022/9422584

LIN, Tsung-Yi, Priya GOYAL, Ross GIRSHICK, Kai HE and Piotr DOLLÁR, 2017. Focal Loss for Dense Object Detection. *IEEE International Conference on Computer Vision* [online]. 2017. Dostupné na: doi:10.1109/iccv.2017.324

LIU, Daizong, Xiaoye QU, Pan ZHOU and Yang LIU, 2022. Exploring motion and appearance information for temporal sentence grounding. Proceedings of the ... AAAI Conference on Artificial Intelligence [online]. 2022, vol. 36, no. 2, pp. 1674–1682. Dostupné na: doi:10.1609/aaai.v36i2.20059

LU, Xugang, Peng SHEN, Sheng LI, Yu TSAO and Hisashi KAWAI, 2018. Temporal Attentive Pooling for Acoustic Event Detection. *Interspeech* [online]. 2018. Dostupné na: doi:10.21437/interspeech.2018-1552

MONFORT, Mathew, Carl VONDRICK, Aude OLIVA, Alex ANDONIAN, Bolei ZHOU, K. R. RAMAKRISHNAN, Sarah Adel BARGAL, Tom YAN, Lisa M. BROWN, Quanfu FAN and Dan GUTFREUND, 2020. Moments in Time Dataset: One million videos for event understanding. IEEE Transactions on Pattern

Analysis and Machine Intelligence [online]. 2020, vol. 42, no. 2, pp. 502–508. Dostupné na: doi:10.1109/tpami.2019.2901464

MOU, Lichao, Yuansheng HUA, Pu JIN and Xiao Xiang ZHU, 2020. ERA: a dataset and deep learning benchmark for event recognition in aerial videos. *arXiv.org* [online]. Dostupné na: http://arxiv.org/abs/2001.11394

MOU, Lichao, Yuansheng HUA, Pu JIN and Xiao Xiang ZHU, v tlači. ERA Dataset: A Dataset and Deep Learning Benchmark for Event Recognition in Aerial Videos [online]. Dostupné na: https://lcmou.github.io/ERA_Dataset/

MUELLER, Jonas and Aditya THYAGARAJAN, 2016. Siamese recurrent architectures for learning sentence similarity. Proceedings of the AAAI Conference on Artificial Intelligence [online]. 2016, vol. 30, no. 1. Dostupné na: doi:10.1609/aaai.v30i1.10350

MUNAWAR, Maryam and Iram NOREEN, 2021. Duplicate frame video forgery detection using Siamese-based RNN. *Intelligent Automation & Soft Computing (Print)* [online]. 2021, vol. 29, no. 3, pp. 927–937. Dostupné na: doi:10.32604/iasc.2021.018854

NG, Joe Yue-Hei, Matthew HAUSKNECHT, Sudheendra VIJAYANARASIMHAN, Oriol VINYALS, Rajat MONGA and George TODERICI, 2015. Beyond short snippets: Deep networks for video classification. *IEEE Conference on Computer Vision and Pattern Recognition* [online]. 2015. Dostupné na: doi:10.1109/cvpr.2015.7299101

O'SHEA, Keiron and Ryan NASH, 2015. An introduction to convolutional neural networks. *arXiv.org* [online]. Dostupné na: https://arxiv.org/abs/1511.08458

PAPERT, Seymour A., 1966. *The Summer Vision Project* [online]. Dostupné na: https://dspace.mit.edu/handle/1721.1/6125

PIGOU, Lionel, Aäron VAN DEN OORD, Sander DIELEMAN, Mieke VAN HERREWEGHE and Joni DAMBRE, 2016. Beyond Temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision* [online]. 2016, vol. 126, nos. 2–4, pp. 430–439. Dostupné na: doi:10.1007/s11263-016-0957-7

QU, Haoxuan, Hossein RAHMANI, Li XU, Bryan WILLIAMS and Jun LIU, 2021. Recent Advances of Continual Learning in Computer Vision: An Overview. *arXiv.org* [online]. Dostupné na: http://arxiv.org/abs/2109.11369

ROSS, Ted M., Yixin XU, Rick A. BRIGHT and Harriet L. ROBINSON, 2000. C3d enhancement of antibodies to hemagglutinin accelerates protection against influenza virus challenge. Nature Immunology [online]. 2000, vol. 1, no. 2, pp. 127–131. Dostupné na: doi:10.1038/77802

SALAZAR, Aldo André Díaz and Paulo Roberto Gardel KURKA, 2020. Computer Vision methods for automotive applications. *Tecnia (Lima)* [online]. 2020, vol. 30, no. 2, pp. 74–81. Dostupné na: doi:10.21754/tecnia.v30i2.801

SHANG, Fanhua, Tao HAN, Feng TIAN, Jun TAO and Zan GAO, 2020. A multimodal pairwise discrimination network for Cross-Domain action recognition. *IEEE Access* [online]. 2020, vol. 8, pp. 143545–143557. Dostupné na: doi:10.1109/access.2020.3014691

SHARMA, Ajay, Ankit GUPTA and Varun JAISWAL, 2021. Solving image processing critical problems using machine learning. In: *Studies in big data* [online]. p. 213–248. Dostupné na: doi:10.1007/978-981-15-9492-2_11

SHI, Zhibin, Liangjie CAO, Cheng GUAN, Haiyong ZHENG, Zhaorui GU, Zhibin YU and Bing ZHENG, 2020. Learning Attention-Enhanced spatiotemporal representation for action recognition. *IEEE Access* [online]. 2020, vol. 8, pp. 16785–16794. Dostupné na: doi:10.1109/access.2020.2968024

SHORTEN, Connor and Taghi M. KHOSHGOFTAAR, 2019. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data [online]. 2019, vol. 6, no. 1. Dostupné na: doi:10.1186/s40537-019-0197-0

SHU, Tianmin, Dan XIE, Brandon ROTHROCK, Sinisa TODOROVIC and Song-Chun ZHU, 2015. Joint inference of groups, events and human roles in aerial videos. *IEEE Conference on Computer Vision and Pattern Recognition* [online]. 2015. Dostupné na: doi:10.1109/cvpr.2015.7299088

SIGURDSSON, Gunnar A., Abhinav GUPTA, Cordelia SCHMID, Ali FARHADI and Karteek ALAHARI, 2018. Charades-Ego: a Large-Scale dataset of paired third and first person videos. arXiv.org [online]. Dostupné na: https://arxiv.org/abs/1804.09626

SOOMRO, Khurram, Amir Roshan ZAMIR, Zamir SHAH and Mubarak SHAH, 2012. UCF101 - Action Recognition Data Set [online]. Dostupné na: https://www.crcv.ucf.edu/data/UCF101.php

TAKAHASHI, Ryo, Takashi MATSUBARA and Kuniaki UEHARA, 2020. Data augmentation using random image cropping and patching for deep CNNs. *IEEE Transactions on Circuits and Systems for Video Technology (Print)* [online]. 2020, vol. 30, no. 9, pp. 2917–2931. Dostupné na: doi:10.1109/tcsvt.2019.2935128

TRAN, Du, Lubomir BOURDEV, Rob FERGUS, Lorenzo TORRESANI and Manohar PALURI, 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. *IEEE International Conference on Computer Vision* [online]. 2015. Dostupné na: doi:10.1109/iccv.2015.510

TSAI, Jen-Kai, Chen-Chien HSU, Wei-Yen WANG and Shao-Kang HUANG, 2020. Deep Learning-Based Real-Time Multiple-Person Action Recognition System. *Sensors* [online]. 2020, vol. 20, no. 17, p. 4758. Dostupné na: doi:10.3390/s20174758

WANG, Heng and Cordelia SCHMID, 2013. Action Recognition with Improved Trajectories. *IEEE International Conference on Computer Vision* [online]. 2013. Dostupné na: doi:10.1109/iccv.2013.441

WANG, Limin, Yuanjun XIONG, Zhe WANG, Yu QIAO, Dahua LIN, Xiaoou TANG and Luc VAN GOOL, 2016. Temporal Segment Networks: towards good practices for deep action recognition. In: *Lecture Notes in Computer Science* [online]. p. 20–36. Dostupné na: doi:10.1007/978-3-319-46484-8_2

WANG, Xuan and Zhigang ZHU, 2023. Context understanding in computer vision: A survey. *Computer Vision and Image Understanding* [online]. 2023, vol. 229, p. 103646. Dostupné na: doi:10.1016/j.cviu.2023.103646

WEI, Pengbo, David AHMEDT-ARISTIZABAL, Harshala GAMMULLE, Simon DENMAN and Mohammad Ali ARMIN, 2023. Vision-based activity recognition in children with autism-related behaviors. *Heliyon* [online]. 2023, vol. 9, no. 6, p. e16763. Dostupné na: doi:10.1016/j.heliyon.2023.e16763

WOLFF, Mathieu, Sarah MORCEAU, Ross FOLKARD, Jesús MARTÍN-CORTECERO and Alexander GROH, 2021. A thalamic bridge from sensory perception to cognition. Neuroscience & Biobehavioral Reviews/Neuroscience and Biobehavioral Reviews [online]. 2021, vol. 120, pp. 222–235. Dostupné na: doi:10.1016/j.neubiorev.2020.11.013

XU, Huijuan, Abir DAS and Kate SAENKO, 2017. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection [online]. Dostupné na: https://openaccess.thecvf.com/content_iccv_2017/html/Xu_R-C3D_Region_Convolutional_ICCV_2017_paper.html

XU, Jiarui, Yue CAO, Zheng ZHANG and Han HU, 2019. Spatial-Temporal relation networks for Multi-Object tracking [online]. Dostupné na: https://openaccess.thecvf.com/content_ICCV_2019/html/Xu_Spatial-Temporal_Relation_Networks_for_Multi-Object_Tracking_ICCV_2019_paper.html

XUAN, Qi, Fuxian LI, Yi LIU and Yun XIANG, 2019. MV-C3D: a spatial correlated Multi-View 3D convolutional neural networks. *IEEE Access* [online]. 2019, vol. 7, pp. 92528–92538. Dostupné na: doi:10.1109/access.2019.2923022

YANG, Yongqing, Yong YANG, Zhengzhi LU, Junjie YANG, Deyang LIU, Chuanbo ZHOU and Zhibin FAN, 2022. STA-TSN: Spatial-Temporal Attention Temporal Segment Network for action recognition in video. PloS One [online]. 2022, vol. 17, no. 3, p. e0265115. Dostupné na: doi:10.1371/journal.pone.0265115

YAO, Huaxiu, Fei WU, Jintao KE, Xianfeng TANG, Yitian JIA, Siyu LU, Pinghua GONG, Jieping YE and Zhenhui LI, 2018. Deep Multi-View Spatial-Temporal Network for taxi demand prediction. Proceedings of the ... AAAI Conference on Artificial Intelligence [online]. 2018, vol. 32, no. 1. Dostupné na: doi:10.1609/aaai.v32i1.11836

ZHANG, Bokai, Amer GHANEM, Alexander SIMES, Henry CHOI and Andrew YOO, 2021. Surgical workflow recognition with 3DCNN for Sleeve Gastrectomy. *International Journal of Computer Assisted Radiology and Surgery (Print)* [online]. 2021, vol. 16, no. 11, pp. 2029–2036. Dostupné na: doi:10.1007/s11548-021-02473-3

ZHANG, Ke, Weilun CHAO, Fei SHA and Kristen GRAUMAN, 2016. Video Summarization with Long Short-Term Memory. In: *Lecture notes in computer science* [online]. p. 766–782. Dostupné na: doi:10.1007/978-3-319-46478-7_47

ZHANG, Minqing and Wenping LI, 2021. An automatic classification method of sports teaching video using support vector machine. *Scientific Programming* [online]. 2021, vol. 2021, pp. 1–8. Dostupné na: doi:10.1155/2021/4728584

ZHANG, Qi, Jianlong CHANG, Guang MENG, Shiming XIANG and Pan CHEN, 2020. Spatio-Temporal graph Structure learning for traffic forecasting. *Proceedings of the ... AAAI Conference on Artificial Intelligence* [online]. 2020, vol. 34, no. 01, pp. 1177–1185. Dostupné na: doi:10.1609/aaai.v34i01.5470

ZHANG, Zhiwu, Elhan S. ERSOZ, Chao-Qiang LAI, Rory J. TODHUNTER, Hemant K. TIWARI, Michael A. GORE, Peter J. BRADBURY, Jianming YU, Donna K. ARNETT, José M. ORDOVÁS and Edward S. BUCKLER, 2010. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics (Print)* [online]. 2010, vol. 42, no. 4, pp. 355–360. Dostupné na: doi:10.1038/ng.546

ZHENGFENG, Huang, 2022. Accurate recognition method of continuous sports action based on deep learning algorithm. *Wireless Communications and Mobile Computing (Print)* [online]. 2022, vol. 2022, pp. 1–10. Dostupné na: doi:10.1155/2022/3407935

ZHI, Yuan, Zhan TONG, Limin WANG and Gangshan WU, 2021. *MGSampler: An Explainable sampling strategy for video action Recognition* [online]. Dostupné na: https://openaccess.thecvf.com/content/ICCV2021/html/Zhi_MGSampler_An_Explainable_Sampling_Strategy_for_Video_Action_Recognition_ICCV_2021_paper.html

ZHOU, Bolei, Alex ANDONIAN, Aude OLIVA and Antonio TORRALBA, 2018. Temporal relational reasoning in videos. In: *Lecture Notes in Computer Science* [online]. p. 831–846. Dostupné na: doi:10.1007/978-3-030-01246-5_49

ZHOU, Lei, Lin ZHANG and Nicholas KONZ, 2022. Computer Vision Techniques in Manufacturing. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* [online]. 2022. Dostupné na: doi:10.36227/techrxiv.17125652.v2

ZOZNAM PRÍLOH

Príloha A – Odkaz na GitHub repozitár.

PRÍLOHA A

Príloha A

- A1 Odkaz na GitHub repozitár: https://github.com/batsysk/DiplomovaPraca