**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

**Zadanie ZP**

# Poďakovanie

Na tomto mieste by som sa chcela poďakovať Mgr. Ľubomír Benko, PhD za cenné rady, konzultácie a spätnú väzbu. Taktiež mojej rodine 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: Bakalár odboru Aplikovaná informatika. Nitra: FPVaI, 2024. .... s. (uviesť počet strán ZP).

Abstrakt musí byť písaný ako **jeden odsek**. **Abstrakt by mal mať rozsah maximálne jednej strany.** Abstrakt musí byť pre čitateľa zrozumiteľný, aj bez prečítania textu celej práce. **Abstrakt nie je len úvod práce, obsahuje zmienku o cieľoch, postup riešenia, výsledky a prínos.** Je potrebné zachovať základné informácie a súlad abstraktu s textom práce. Treba byť čo najstručnejší, dodržiavať požiadavky na obsahovú stránku abstraktu, ale vyhnúť sa nepresnostiam. Nie je vhodné uvádzať informácie alebo tvrdenia, ktoré sa nevyskytujú v samotnom texte práce. V abstrakte sa používa tretia osoba jednotného alebo množného čísla. Vždy, keď je to možné, treba použiť slovesá v prítomnom čase a v činnom rode. Prispieva to k jasnému, stručnému a pôsobivému písaniu. Trpný rod možno použiť na indikatívne konštatovania a na informatívne konštatovania, ak treba súčasne zdôrazniť osobu, na ktorú sa činnosť zameriava. *Terminológia:* V abstrakte sa používajú plnovýznamové slová z textu, ktoré pomôžu aj pri automatickom vyhľadávaní v textoch. Treba sa vyhnúť neznámym termínom, skratkám, akronymom alebo symbolom alebo ich treba v odôvodených prípadoch definovať hneď, keď sa prvý raz vyskytnú v abstrakte. Používajú sa jednotky, symboly, terminológia ISO vždy, keď je to možné. Ak nie sú k dispozícii, používajú sa národné normy.

Kľúčové slová: (napr. Informatika. Prírodné vedy.) Minimálne 3 – 5 výrazov. Jednotlivé slová sa oddeľujú bodkou a začínajú sa veľkým písmenom.

*POVINNÉ*

# ABSTRACT

LAST NAME, First Name: Title of the thesis. [Bachelor/Master Thesis]. Constantine the Philosopher University in Nitra. Faculty of Natural Sciences and Informatics. Supervisor: Title. First name Last name, title. Degree of Qualification: Bachelor of Applied Informatics. Nitra: FNSaI, 202x. .... p.

The abstract must be written as a single paragraph. The abstract should be a maximum of one page. It must be comprehensible to the reader, even without reading the full text of the paper. The abstract is not just an introduction to the thesis, but it includes a mention of the objectives, the solution procedure, the results and the contribution. It is necessary to keep the background information and consistency of the abstract with the text of the thesis. It is necessary to be as concise as possible, adhering to the content requirements of the abstract but avoiding inaccuracies. It is not appropriate to include information or statements that do not appear in the text of the thesis itself. The abstract should be written in passive voice. The imperfect gender can be used for indicative statements and for informative statements when the person on whom the action is directed needs to be emphasised at the same time. Terminology: the abstract uses full-meaning words from the text, which will also help with automatic text searches. Unfamiliar terms, abbreviations, acronyms or symbols should be avoided or, where justified, defined as soon as they first appear in the abstract. Units, symbols, ISO terminology shall be used whenever possible. Where they are not available, national standards shall be used.

Keywords: Informatics. Key. MS Word.

*POVINNÉ*

# Obsah

[Rozpoznávanie objektov vo videách 1](#_Toc162903179)

[Rozpoznávanie objektov vo videách 2](#_Toc162903180)

[DIPLOMOVÁ práca 2](#_Toc162903181)

[Zadanie ZP 3](#_Toc162903182)

[Poďakovanie 4](#_Toc162903183)

[ABSTRAKT 5](#_Toc162903184)

[ABSTRACT 6](#_Toc162903185)

[Obsah 7](#_Toc162903186)

[Úvod 1](#_Toc162903187)

[1 Analýza súčasného stavu 2](#_Toc162903188)

[1.1 Computer Vision 2](#_Toc162903189)

[1.2 Základné metódy computer vision 3](#_Toc162903190)

[1.3. Modely klasifikujúce vo videách 4](#_Toc162903191)

[1.3.1 CNN 4](#_Toc162903192)

[1.3.2 RNN 6](#_Toc162903193)

[1.3.3 SVM 6](#_Toc162903194)

[1.3.4 3D Konvolučné modely 8](#_Toc162903195)

[1.4 Modely využité pri analýze ERA datasetu 9](#_Toc162903196)

[**C3D Model** 9](#_Toc162903197)

[**P3D Model** 12](#_Toc162903198)

[**I3D Model** 13](#_Toc162903199)

[**TRN Model** 13](#_Toc162903200)

[1.3.5 Ďalšie modely 14](#_Toc162903201)

[1.5. Porovnanie modelov 15](#_Toc162903202)

[1.6. Optimalizačné metódy 16](#_Toc162903203)

[Temporal pooling 16](#_Toc162903204)

[Vzorkovacia metóda 17](#_Toc162903205)

[Temporal aggregation 17](#_Toc162903206)

[Transformácie 17](#_Toc162903207)

[1.7. Architektúry použité v modeloch ERA 18](#_Toc162903208)

[ResNet 18](#_Toc162903209)

[Inception 19](#_Toc162903210)

[2 Ciele záverečnej práce 20](#_Toc162903211)

[3 Metodika výskumu 21](#_Toc162903212)

[3.1. Porozumenie problematike 21](#_Toc162903213)

[Klasifikácia vo videách 21](#_Toc162903214)

[3.2. Porozumenie dátam 22](#_Toc162903215)

[Dataset ERA 22](#_Toc162903216)

[Videá v datasete 23](#_Toc162903217)

[Pred trénované modely 24](#_Toc162903218)

[3.3. Práca s modelmi 24](#_Toc162903219)

[Práca s pred trénovanými modelmi 24](#_Toc162903220)

[C3D-Sport1M 25](#_Toc162903221)

[C3D-UCF101 26](#_Toc162903222)

[P3D-ResNet-199\_Kinetics a P3D-ResNet-199\_Kinetics-600 27](#_Toc162903223)

[TRN-Something-Something-V2 a TRN-Moments-In-Time 28](#_Toc162903224)

[I3D-Kinetics a I3D-Kinetics+ImageNet 30](#_Toc162903225)

[3.4. Analýza dát 31](#_Toc162903226)

[3.7. Využité technológie 32](#_Toc162903227)

[3.8. Hypotézy 32](#_Toc162903228)

[4 Výsledky 33](#_Toc162903229)

[4.1. Analýza výsledkov 33](#_Toc162903230)

[C3D-Sport1M 33](#_Toc162903231)

[C3D-UCF101 35](#_Toc162903232)

[P3D-ResNet-199\_Kinetics a P3D-ResNet-199\_Kinetics-600 38](#_Toc162903233)

[TRN-Something-Something-V2 a TRN-Moments-In-Time 38](#_Toc162903234)

[I3D-Kinetics a I3D-Kinetics+ImageNet 39](#_Toc162903235)

[4.2. Vyhodnotenie a diskusia 41](#_Toc162903236)

[4.3. Odporúčania 42](#_Toc162903237)

[Záver 44](#_Toc162903238)

[Zoznam bibliografických odkazov 46](#_Toc162903239)

[Zoznam príloh 50](#_Toc162903240)

# 

# Úvod

V úvode autor záverečnej práce (ďalej ZP) stručne a výstižne charakterizuje stav poznania alebo praxe v oblasti, ktorá je predmetom ZP a oboznamuje čitateľa s významom ZP. Autor v úvode zdôrazňuje, prečo je práca dôležitá a prečo sa rozhodol spracovať danú tému. Autor v krátkosti uvedie ciele a zámery práce. V úvode autor stručne a jasne:

* charakterizuje súčasný stav poznania problematiky, ktorá je predmetom ZP,
* zdôvodní aktuálnosť témy,
* nastolí problémy, ktoré chce riešiť,
* uvedie v krátkosti ciele, ktoré sa majú v práci dosiahnuť,
* opíše použité metódy a postup riešenia,
* spresní, ktoré hlavné informačné zdroje použil,
* zdôvodní význam riešenia problematiky,
* načrtne stručný obsah jednotlivých kapitol.

V prípade výskumnej práce autor uvedie ako jeho výskum môže prispieť k riešeniu v úvode nastoleného problému.

*ÚVOD PREDSTAVUJE CCA 1 STRANU. ČÍSLOVANIE STRÁN ZAČÍNA OD ÚVODU.*

# 1 Analýza súčasného stavu

Súčasná doba digitálneho pokroku prináša interakcie medzi človekom a strojom na vyššej úrovni. Jedným z najviac viditeľných aspektov tejto postupnej revolúcie je oblasť umelého videnia, taktiež známa ako computer vision. Táto disciplína umelej inteligencie sa stala súčasťou nášho každodenného života a to najmä v oblasti spracovania obrazov a analýzy videí. Spomenieme napríklad autonómne autá, kde sa detegujú objekty a ich trajektórie (Kortylewski et al., 2020) alebo využitie v chirurgii pri rozpoznávaní jednotlivých úkonov chirurga. (B. Zhang et al., 2021)

V rámci tejto širokej oblasti sa sústredíme na schopnosť počítačov a technológií „vidieť“ a teda porozumieť digitálnym dátam. Zameriame sa na modely, ktoré sa rozvinuli vďaka tomuto pokroku a prinášajú automatické spracovanie s neuveriteľnou presnosťou a efektivitou.

## 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, aby sa učili a vylepšovali automaticky a boli by znovu použiteľné. Taktiež aby sa nemuseli vytvárať modely nové od základu. Počítačové videnie pridáva k základom neurónových sietí toto učenie sa postupne v čase a tým idú do popredia v mnohých odvetviach. (Qu et al., 2023)

Počítačové videnie spracováva jednoduché popisy a teda čo vidí na scéne. Hľadá vzory, aby zistilo kontext jednotlivých snímok. Vzory obsahujú pochopenie scény a tie sa využívajú v modeloch. Algoritmy modelov vieme rozdeliť do kategórií (L. Zhou, Zhang, a Konz, 2022):

1. detekcia,
2. rozpoznanie,
3. segmentácia,
4. 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. (Yan Li, 2022)

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

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

Obrázok, na ktorom je text, snímka obrazovky, pestrofarebnosť, grafický dizajn

Automaticky generovaný popis

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

## 1.2 Základné metódy computer vision

Pri vyhodnocovaní 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. (X. Wang a Zhu, 2023)

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

1. priestorový kontext,
2. časový kontext,
3. alebo iný kontext.

Na kontext sa potom dá podľa ich názoru dá pozerať na rôznych úrovniach a to menovito na:

1. úroveň predchádzajúcich znalostí
2. úroveň globálnych znalostí
3. alebo taktiež úroveň lokálnych znalostí

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

## 1.3. Modely klasifikujúce vo videách

Modely na klasifikáciu videí predstavujú 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. (Ng et al., 2015)

### 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 akcií vo videách. (Wu, Sharma, a Blumenstein, 2017)

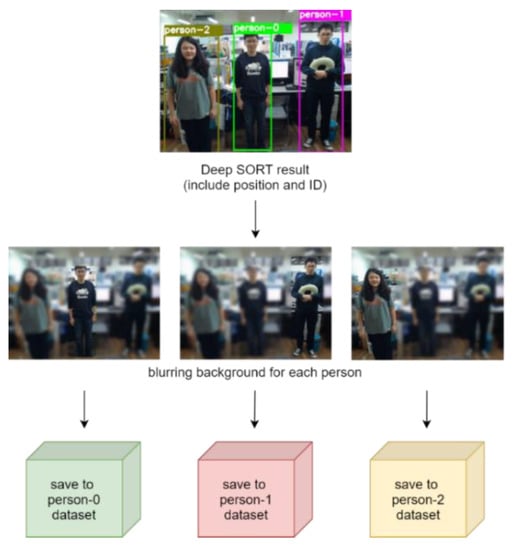
V športoch sa napríklad využívajú na rozoznanie typu športu a aktivity v reálnom čase(obrázok). Na základe monitorovania ľudí, konkrétne v tomto prípade ich stredu tela a štvorcového mapovania, sa vykonávajú pozorovania ich správania. Takéto modelovanie sa nevyužíva iba v analýze športovcov a ich výkonov,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

Obrázok, na ktorom je text, snímka obrazovky, fyzická kondícia, tanec

Automaticky generovaný popis

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

Ďalšie využitie na základe pozorovania ľudí sa využíva v bezpečnostnom monitorovaní. Vďaka modelom sa znižujú náklady na obstarávanie týchto systémov a ich monitorovanie. Najväčšou prekážkou je momentálne problém, že sa modely zameriavajú na aktivity jednej osoby, avšak v tomto odvetí je potrebné monitorovať všetky osoby na danej snímke videa (obrázok). Jen-Kai Tsai a kol. (2020) riešia túto problematiku a výsledok ich výskumu popisujú v článku, kde vidia využitie modelu I3D (Sensors | Free Full-Text | Deep Learning-Based Real-Time Multiple-Person Action Recognition System, n.d.)



### 1.3.2 RNN

RNN (Recurrent Neural Network) využívajú výskumníci napríklad v detegovaní podvodných videí. Keďže sa videá využívajú v mnohých oblastiach ako dôkazový materiál: právo, forenzné štúdie, novinárstvo a iné, je nesmierne dôležité overiť ich autenticitu. Autori popisujú využitie architektúry siamských RNN v modeli I3D a na dvoch datasetoch dosahujú až 86.6% (dataset VIRAT - Image Retrieval and Analysis Tool) a 93% (dataset MFC - Media Forensic Challenge) presnosti. V princípe modely využívajú na hľadanie duplikovanie snímok (Duplicate Frame Video Forgery Detection Using Siamese-based RNN, n.d.).

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

### 1.3.3 SVM

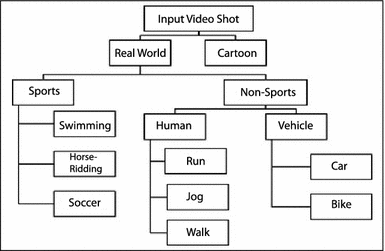
SVM alebo Support Vector Machine je algoritmus strojového učenia, ktorý sa často využíva pre klasifikačné a regresné úlohy. Pri videách sa momentálne skúma využitie SVM na kategorizáciu druhov videí a to napríklad na reklamy, kreslené rozprávky, hudobné videá, správy a športové videá (podľa výskumu Zhang Min-qinga a Li Wen-pinga, 2021). Pri detekcii využívajú aj zvukovú zložku priloženú k videu. Autori sa dodatočne snažia odhaliť aj teroristický alebo násilný obsah. (Min-qing a Wen-ping, 2021)

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). (Min-qing a Wen-ping, 2021)

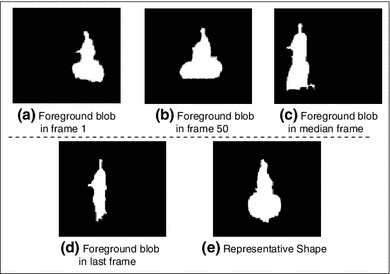
Obrázok, na ktorom je diagram, rad, náčrt, kresba

Automaticky generovaný popis

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



Pri týchto výskumoch sa videá spracovávajú tak, že sa najskôr dajú do unitárneho farebného spektra a následne sa snímky konvertujú do šedých farieb kvôli výpočtovej náročnosti. Následne sa pracuje s tvarmi. Pri tomto kroku sa rozlišuje pozadie a objekty záujmu, tie sa segmentujú pričom vzniká akýsi útvar.(obrázok) Tento útvar sa oddeľuje od pozadia a neskôr sa, na základe ich relatívneho centra, útvary prekrývajú a pozorujú zmeny. Veľkou pomocou sa ukázalo porovnávať a zaznamenávať aj textúry objektov. (Chattopadhyay a Maurya, 2013)



Po tomto spracovaní sa hľadajú rozdiely medzi kategóriami pričom vo výsledku autori poukazujú na rozsiahle využitie v rozpoznávaní kontextu videa avšak do budúcna odporúčajú pridať ďalšie média ako napríklad zvuk a text. (Chattopadhyay a Maurya, 2013)

### 1.3.4 3D Konvolučné modely

CNN (convolutional neural network) je dopredná neurónová 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 ). (Gama et al., 2018)

Vstupná vrstva predstavuje hodnoty pixelov jednotlivých snímok. (https://arxiv.org/pdf/1511.08458.pdf)

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. (https://arxiv.org/pdf/1511.08458.pdf)

Pooling alebo vrstva zlučovania následne vykonáva zmenšenie vzorkovania podľa priestorovej dimenzionality zo vstupu. Týmto znižuje počet parametrov. (https://arxiv.org/pdf/1511.08458.pdf)

Plne prepojená vrstva vytvára klasifikácie do jednotlivých tried. (https://arxiv.org/pdf/1511.08458.pdf)

Výstupná vrstva spája výsledky do dimenzie podľa počtu kategórií. (https://arxiv.org/pdf/1511.08458.pdf)

A diagram of different types of squares

Description automatically generated

Obrázok 2 Architektúra CNN

## 1.4 Modely využité pri analýze ERA datasetu

Autori Mou, L., Hua, Y., Jin, P. a Zhu, X. X. (2020), ktorí spracovali videá do datasetu ERA Event Recognition in Aerial Videos - Rozpoznávanie udalostí v leteckých videách)[[2]](#footnote-2) otestovali dataset na štyroch modeloch:

* C3D,
* P3D,
* I3D
* a TRN modely.

**C3D Model**

C3D (Convolutional 3D model - Konvolučný 3D model) 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. (Hadidi et al., 2020)

Obrázok, na ktorom je text, snímka obrazovky, písmo, rad

Automaticky generovaný popis

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

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), tak C3D prakticky vždy bol 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 akcií a klasifikácie videí. Aplikuje sa v rôznych oblastiach vrátane rozpoznávania športových gest, detekcie udalostí vo videách, detekcie depresie z tvárových výrazov a monitorovania spánku.

Obrázok, na ktorom je text, snímka obrazovky

Automaticky generovaný popis

Najrozšírenejšie využitie má pri rozpoznávaní akcií, kde tento model ukázal svoju schopnosť prekonať iné často využívané modely ako sú HOF (Histogram of Optical Flow) a MBH (Motion Boundary Histogram) a to z hľadiska rozpoznávania pohybov. (H. Wang a Schmid, 2013)

Shang (rok todo) taktiež popisuje zlepšenie výkonov s použitím modelu ActionVLAD (A Multimodal Pairwise Discrimination Network for Cross-Domain Action Recognition | IEEE Journals & Magazine | IEEE Xplore n.d.)

HOF a MBH používajú histogramy na rozpoznávanie 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 binov. Výsledný histogram reprezentuje distribúciu smerov pohybu vo videu. (Q. Li et al., 2016)

MBH naopak zachytáva hranice pohybu vo videu. Počíta gradient optického toku a kvantifikuje orientácie gradientu do rôznych binov. V tomto prípade histogram reprezentuje distribúciu hraníc pohybu vo videu. (Q. Li et al., 2016)

V praxi sa využíva aj kombinácia HOF alebo MBH s C3D modelom a tu dodávajú doplnkové vlastnosti k priestorovým informáciám zachyteným modelom C3D. Avšak sa ukazuje, že C3D model jednoducho prekonáva tieto modely aj ak pracuje osamote a to pri viacerých testoch výkonu. (Tran et al., 2015)

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 sa vyvíja variácia na model C3D: MV-C3D pre 3D konvolučné neurónové siete. Tento nový model preukazuje svoju prispôsobivosť v reálnych situáciách, najmä pri analýze 3D rotovaných reálnych obrázkoch. Čiže sa dá tento model využiť v počítačovom videní, najmä pri rozpoznávaní objektov a pochopeniu scény. (MV-C3D: A Spatial Correlated Multi-View 3D Convolutional Neural Networks | IEEE Journals & Magazine | IEEE Xplore n.d.)

Taktiež nájdeme využitie aj v imunológii, kde C3D zlepšuje vytváranie protilátok na vírusy prvého typu. Čiže sa dá potencionálne vyžiť na vytváranie stratégií pri tvorbe vakcín a imunoterapie pri vírusových infekciách. (MV-C3D: A Spatial Correlated Multi-View 3D Convolutional Neural Networks | IEEE Journals & Magazine | IEEE Xplore n.d.)

**P3D Model**

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

P3D model pol porovnávaný s C3D a predstavuje menšiu výpočtovú náročnosť ako aj menšiu obťažnosť pri tréningu a nižšie percento preučenia. (Q. 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) pomocou tohto pseudo trojrozmerného modelu – P3D. Model bol použitý na simuláciu difúzie látok, ktoré prezentovali jediný transportný mechanizmus v aktívnych častiach SSFBs. P3D umožnil znázorniť časovo závislé profily napätia, distribúciu prúdu a distribúciu stavu napätia v aktívnych častiach. (Chayambuka, Fransaer, a Dominguez-Benetton, 2019)

Obrázok, na ktorom je text, snímka obrazovky, diagram, rad

Automaticky generovaný popis

Využitie modelu nájdeme aj v biológii, pri génových štúdiách, kde sa využíva napríklad na pochopenie organizácie a funkcionality genómov ohľadom zdravia a chorôb populácie. V medicíne má P3D model úžitok pri modelovaní geometrie genómov, kde sa pozorujú bunky a ich dopad na zdravie ľudí. (Full article: 3D genome organization in health and disease: emerging opportunities in cancer translational medicine n.d.)

**I3D Model**

I3D model predstavuje architektúru, ktorá zväčšuje dvojrozmerné konvolučné siete aby 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. (Wei et al., 2022)

V oblasti rozpoznávania aktivít vo videách sa tento model využíva v chirurgii, kde sa rozpoznávajú 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 v rámci operácií Tubulizácie žalúdka. Model bol trénovaný pomocou metódy ohnisková stráta (Focal loss) a dosiahol presné rozpoznávanie rôznych chirurgických krokov. (Carreira a Zisserman, 2017)

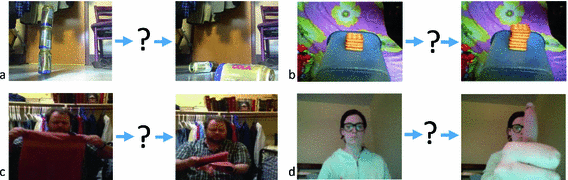
Focal Loss je špeciálny druh loss funkcie, ktorá sa zaoberá problémom nerovnováhy tried pri úlohách detekcie objektov. Keďže pri tejto problematike väčšina snímok pozostáva z pozadia, zatiaľ čo pre nás je dôležitá detekcia objektu záujmu, môže nerovnováha spôsobovať malú úspešnosť detegovania objektov popredia pri modeloch. Na klasifikáciu používa špeciálnu váhu, ktorá znižuje hodnotu pri ľahko odhadnuteľných prípadoch a teda sa trénuje na ťažkých a viac ojedinelých príkladoch tried. Týmto zlepšuje výkony v triedach menšín. (Lin et al., 2017)

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

I3D model sa využíva pri bezpečnostných aplikáciách v reálnom svete. Napríklad v autonómnych autách a pri detekcii osôb na bezpečnostných záznamoch. (Kortylewski et al., 2020)

**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. 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 vidíme ako si ľudský mozog ľahko doplní kontext medzi obrázkami, ale pre modely je to náročné. Tento problém sa snaží vyriešiť model TRN za pomoci zisťovania vzťahov medzi snímkami. (B. Zhou et al., 2018)



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

Vo výskumoch neurovedy bol tento model skúmaný pri regulácii senzorického spracovania, pozornosti a kognície. (Yinqing Li et al., 2020)

Taktiež, na základe dostupných referencií, sa využíva podobne ako P3D model v génovej vede. Tam pomáha skúmať komplexné interakcie medzi samotnými génmi a ich regulačnými prvkami. Týmto nám odkrývajú pohľad na mechanizmy, ktoré súvisia s prejavovaním génov a bunkovými procesmi. (Terney et al., 2008)

V iných odvetviach TRN nachádza využitie napríklad pri počítaní energetickej náročnosti budov. Model využíva náhodné lesy a neurónové siete na to, aby spravil predikcie s vysokým rozlíšením budov. Pracuje s historickými dátami a hľadá vzory a vzťahy ktoré pridávajú k vysokej spotrebe budov. (Ahmad, Mourshed, a Rezgui, 2017)

### 1.3.5 Ďalšie modely

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

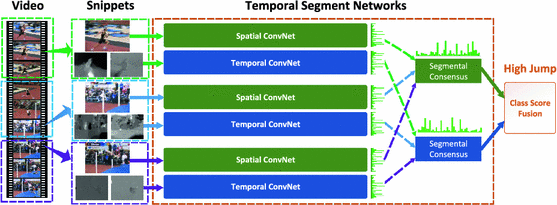
* TSN
* a LSTM.

**TSN model**

TSN model bol vyvinutý na základe potreby rozoznávať dlhodobé aktivity a naplniť potrebu trénovať modely na veľkom množstve videí. Autori videli veľký potenciál na zlepšenie a vyvinuli model TSN alebo Temporal Segment Network (Sieť využívajúca časových segmentov). Tento rámec(framework) extrahuje krátke úryvky z dlhej sekvencie pôvodného videa. Využíva sa na to vzorkovacia stratégia a to nie s hustým vzorkovaním. Vďaka tomuto je model schopný pracovať s dlhými videami a zachovať si relevantné informácie za menšej výpočtovej záťaže. (L. Wang et al., 2016)

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

Využite sa momentálne najviac priraďuje k zlepšeniu výkonov pri rozpoznávaní aktivít vo videách. (L. Wang et al., 2016)



**LSTM**

LSTM- Long Short-Therm Memory model alebo dlhodobá krátkodobá pamäť. Ide o rekurentnú neurónovú sieť, ktorá ale zachytáva a modeluje dlhodobé súvislosti v sekvenčných dátach.

Obrázok, na ktorom je text, snímka obrazovky, diagram, číslo

Automaticky generovaný popis

LSTM má 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. (K. Zhang et al., 2016)

## 1.5. Porovnanie modelov

Modely okrem architektúry majú odlišné výsledky učenia a efektívnej aplikácie. Tabuľka zhromažďuje základné údaje o modeloch, ktoré využívame v tejto práci. Jej stĺpce 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.
* Úroveň kontextu: Či model zohľadňuje globálny (kontext celej scény), lokálny (kontext menšej časti scény) kontext alebo ich kombináciu.
* Mechanizmus: Popis aký mechanizmus alebo techniky využíva model na vykonanie úlohy počítačového videnia.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Úloha, využitie | DNN model | Typ kontextu | Úroveň kontextu | Mechanizmus |
| C3D | Rozpoznávanie vo videách | 3D Konvolučné Siete | Temporálny | Lokálny, globálny | Temporálna pamäťová banka |
| P3D | Rozpoznávanie vo videách | 3D Konvolučné Siete | Temporálny | Lokálny, globálny | Temporálna pamäťová banka |
| I3D | Rozpoznávanie vo videách | Nafúknuté 3D Konvolučné Siete | Temporálny | Lokálny, globálny | Samopozornosť (Self Attention) |
| TRN | Rozpoznávanie vo videách | Siete s temporálnymi vzťahmi | Temporálny | Lokálny, globálny | Grafické konvolučné siete (GCN) |
| TSN | Rozpoznávanie vo videách | Siete s temporálnymi vzťahmi | Temporálny | Lokálny, globálny | Temporálny hierarchický kontextový model |
| LSTM | Sekvenčné modelovanie | Rekurentné neurónové siete | Temporálny | Lokálny, globálny | Pamäťové bunky a brány |

## 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.

### Temporal pooling

Video môže byť vnímané 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. (https://link.springer.com/article/10.1007/s11263-016-0957-7#ref-CR14)

Algoritmus 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). (https://www.isca-archive.org/interspeech\_2018/lu18\_interspeech.pdf)

### Vzorkovacia metóda

Metód vzorkovania existuje mnoho, medzi základné patria napríklad náhodné vzorkovanie alebo replikácia. (https://link-1springer-1com-1it0563by0116.erproxy.cvtisr.sk/chapter/10.1007/978-3-319-19219-2\_1)

Na predchádzanie nedostatku flexibility a výpočtovej náročnosti je ideálne náhodné vzorkovanie. Táto metóda vyberá náhodný počet snímok, ktoré reprezentujú dané video. (https://openaccess.thecvf.com/content/ICCV2021/papers/Zhi\_MGSampler\_An\_Explainable\_Sampling\_Strategy\_for\_Video\_Action\_Recognition\_ICCV\_2021\_paper.pdf)

### Temporal aggregation

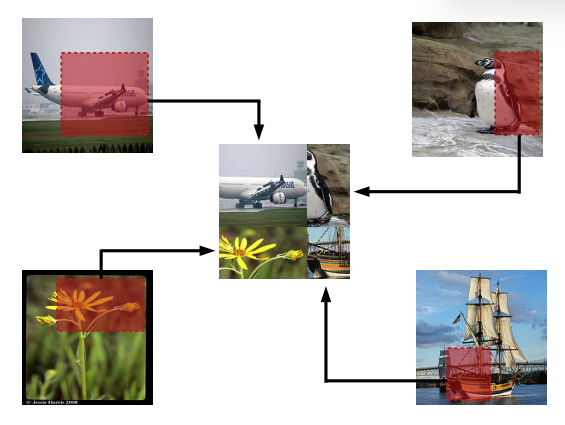
Metóda časovej agregácie predstavuje zlepšenie odhadu modelu, kedy sa komponenty v časových radoch stanú viac alebo menej významnými. (https://www-1sciencedirect-1com-1s3x6wrby0108.erproxy.cvtisr.sk/science/article/pii/S0925527315003382?via%3Dihub)

Môže napríklad zvýšiť frekventovanosť komponentov s malým výskytom pričom sú dôležité v kontexte videa. (https://www.sciencedirect.com/science/article/pii/S0377221717301911?casa\_token=g3t9kCil4lYAAAAA:1x\_wOFWncwNw6TezBag8lBVKrGKkbsEIOxT0Uwfx6ZATYNsGV8jP6jPLccXM3YEfgSqLz9H5-ew)

### Transformácie

Podľa autorky Shasha Li a spol. je pri video klasifikácií vo veľa prípadoch lepšie využiť transformácie ako zlepšovať vyhľadávanie v temporálnom priestore. Vedie to k lepšej štruktúre a zmenšeniu priestoru na vyhľadávanie. (https://www.isca-archive.org/interspeech\_2018/lu18\_interspeech.pdf)

Pretrénovanie je stály problém pri trénovaní modelov. Technika transformácie sa snaží tomu zabrániť popri obohacovaniu datasetov. Jednou z metód je náhodné orezanie snímok a následné prepojenie týchto orezaných kúskov (Obrázok 1). Popri tejto metóde existujú aj bežne používané iné metódy ako normalizácia, otáčanie, zmena rozmerov, zmena farieb, kontrastu alebo saturácie a iné. Taktiež metóda odstraňovania pixelov vytvára šumové snímky a tým vie predchádzať pretrénovaniu. (https://da.lib.kobe-u.ac.jp/da/kernel/90008115/90008115.pdf)



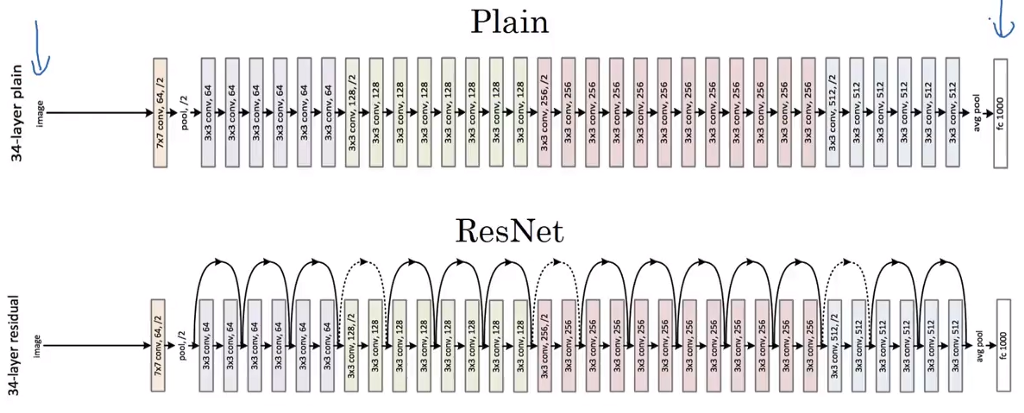
Obrázok 3 Orezávanie snímok ako forma transformácie.

## 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.

### ResNet

Neurónová sieť ResNet (Residual Network) predstavuje architektúru, ktorá vytvára skratku medzi spojeniami aby znížila náročnosť trénovania. Výsledkom je rýchlejšie trénovanie a lepšia generalizačná chyba. Ide o vrstvenie reziduálnych blokov (vrstvy, ktorých výsledky idú hlbšie do iných vrstiev v bloku) na seba a prepájanie dvoch výsledkov s cieľom znížiť trénovaciu chybu (Obrázok 1). (https://arxiv.org/pdf/1611.01186.pdf)



Obrázok 4 Architektúra ResNet

(https://medium.com/analytics-vidhya/opening-resnets-46bb28f43b25)s

### 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 za pomoci znižovania počtu parametrov, bez znižovania presnosti. (https://www.xenonstack.com/blog/inception-architecture-computer-vision)

Architektúra kombinuje 1x1, 3x3 a 5x5 paralelných konvolučných filtrov na extrakciu vlastností zo vstupu. (https://www.xenonstack.com/blog/inception-architecture-computer-vision)

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 1). (https://www.xenonstack.com/blog/inception-architecture-computer-vision)

A diagram of a filter concat

Description automatically generated

Obrázok 5 Architektúra Inception-v4 (11231-Article Text-14759-1-2-20201228.pdf)

# 2 Ciele záverečnej práce

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

Naše čiastkové ciele sú:

* Pochopiť architektúru pred trénovaných modelov a vytvoriť kód v jazyku Python na ich otestovanie.
* Otestovať pred trénované modely autorov datasetu ERA a porovnať ich výsledky s oficiálnymi výsledkami autorov.
* Porovnať modely medzi sebou a nájsť ich najlepšie a najhoršie využitia na konkrétnych druhoch videí.
* Natrénovať a otestovať ďalšie modely na datasete ERA.
* 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[[3]](#footnote-3).

## 3.1. Porozumenie problematike

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

### Klasifikácia vo videách

Pri klasifikácií vo videách ide o proces kategorizácie do rôznych tried alebo kategórií na základe obsahu a charakteristík. Cieľom je automaticky analyzovať a pochopiť informácie, ktoré sa nachádzajú vo videu. (Saddam Bekhet 2021) Oproti klasifikácií v obrázkoch obsahujú videá viaceré snímky, ktoré majú medzi sebou informačné prepojenie. Obsahujú taktiež aj viac informácií, ktoré vieme využiť ako napríklad zvuk. Tieto dodatočné informácie umožňujú komplexnejšie porozumenie obsahu a môžu zlepšiť presnosť klasifikácie. (Karpathy et al. 2014) Na obrázku môžeme vidieť, že sa videá musia brať ako viac dimenzionálne média, pretože v nich ide o kontext v čase. (Bovik, 2009)

Obrázok, na ktorom je diagram, text, rad, snímka obrazovky

Automaticky generovaný popis

## 3.2. Porozumenie dátam

Dataset ERA sme použili v tejto práci z dôvodu jeho konzistentného obsahu, to znamená, že videá mali rovnakú dĺžku, obsahovo sa pridržiavali kategórie, rovnakú kvalitu a počet videí bol dostatočný na ďalšie využitie.

### Dataset ERA

Ako popísal Mou, Hua, Jin a Zhu v ich práci: Dataset ERA[[4]](#footnote-4) (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 YouTube, pričom každé video je označené jednou z 25 rôznych tried udalostí. Dataset predstavuje benchmark pre modely na rozpoznávanie udalostí. (Mou et al., 2020)

Jedna z úloh na porozumenie datasetu a modelom bolo nájsť dokumentáciu od autorov. Pri hľadaní sme našli dostupný iba na jeden článok[[5]](#footnote-5), kde autori popisovali ako postupovali pri zbieraní videí, aké technológie použili (drony) a porovnávajú svoj dataset s datasetmi, kde sú videá získané prostredníctvom satelitných snímok. Poukazovali taktiež na horšiu kvalitu a vyššie náklady ktoré prinášajú satelitné snímky oproti dronovým snímkam.

V článku popisovali podrobnejšie postup kategorizácie videí a ich voľbu formátu. Taktiež spravili podrobnejší prieskum ďalších využívaných datasetov, ktorý zhrnuli do tabuľky 1.

Tabuľka 1 Porovnanie datasetu autormi

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Typ úlohy | Zdroj dát | Video | Počet tried | Počet vzoriek | Rok |
| UCLA[[6]](#footnote-6) | Zamerané na človeka | Získané osobne autormi | áno | 12 | 104 | 2015 |
| Okutama[[7]](#footnote-7) | Ľudská činnosť | Získané osobne autormi | áno | 12 | - | 2017 |
| AIDER[[8]](#footnote-8) | Pohromy | Internet | nie | 5 | 2,545 | 2019 |
| ERA | Všeobecné | YouTube | áno | 25 | 2,864 | 2019 |

Pri výbere pre našu prácu sme porovnávali napríklad aj datasety: UCF101, HMDB51, Kinetics, Charades, YouTube-8M, Sports-1M, Something-Something a iné, ktoré autori síce neporovnávajú, ale priamo ich využili pri trénovaní vlastných modelov.

Významným problémom pri práci s datasetom bola absencia hlbšieho popisu modelov. Nie je nám známa žiadna oficiálna dokumentácia, okrem článku, ktorý sme popisovali na začiatku tejto kapitoly. Článok sa zameriava na motiváciu a zber dát, nepopisuje samotné trénovanie a testovanie modelov. Autori v závere zhrňujú svoje výsledky z testovaní v tabuľkách.

Dataset ERA pozostáva z dostatočného počtu videí na to, aby sa dali na ňom natrénovať modely a následne ich otestovať. Autori rozdelili dataset na testovacie a trénovacie videá a tie následne ja jednotlivé kategórie. Prínos s využitím datasetu vidíme v tom, že dronové datasety nie sú úplne využívané na klasifikáciu a taktiež k datasetu nie je oficiálna dokumentácia.

### Videá v datasete

Kategórie datasetu sú členené do hlavných 7 kategórií nasledovne:

- Šport: Basketbal, Basebal, Člnkovanie, Cyklistika, Beh, Futbal, Plávanie, Závody áut

- Bezpečnosť: Policajná naháňačka, 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.

Dataset je rozdelený na testovacie a trénovacie dáta v každej kategórii.  Každé video má vždy 5 sekúnd pri rozlíšení 640x640 pixelov a 24 snímok za sekundu.

Pri výbere datasetu sme sa upriamovali na:

* 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 unitárny vzhľad, čiže rovnakú stopáž, rozmery videí a počet snímok za sekundu.

Kvalitu datasetu autori overili testovaním modelov, ktoré natrénovali.

### Pred trénované modely

Dataset ERA mal dostupné pred trénované modely, ktoré boli stiahnuteľné priamo na stránke s datasetom. Autori datasetu natrénovali osem modelov pre klasifikáciu videí a jedenásť modelov pre klasifikáciu snímok, ktoré brali z videí. Predmetom našej diplomovej práce nie je klasifikácia pre snímky, takže sme tieto modely netestovali a neskúmali bližšie.

Modely pre klasifikáciu môžeme rozdeliť podľa ich architektúry na:

* C3D,
* P3D,
* I3D,
* a TRN.

Každá architektúra bola natrénovaná dvoma spôsobmi podľa rôznych datasetov (Kinetics, Sport1M, UCF a iné), čiže sme otestovali osem modelov celkovo.

## 3.3. Práca s modelmi

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

### Práca s pred trénovanými modelmi

Pred trénované modely sme v prvom rade chceli otestovať 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 a po prvých testoch sme znížili počet testovaných videí vzhľadom na výpočtovú náročnosť.

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

Pre výber konkrétnych videí, ktoré sme použili na testovanie, sme vytvorili jednoduchý Python program na výber pätnástich jedinečných videí. Tie sme potom opakovane používali na testovanie, aby naše výsledky neboli skreslené rôznymi testovacími podmienkami. Zároveň výberom 15x25 videí sme testovaciu dobu znížili na 4-6 hodín.

### C3D-Sport1M

Model autori natrénovali okrem datasetu ERA aj na datasete Sports-1M[[9]](#footnote-9). Dataset obsahoval viac ako jeden milión videí, ku ktorým sa pristupovalo cez URL adresu. Videá boli rozdelených do 467 športových aktivít (Obrázok 1). Keďže videá zachytávali športové udalosti tak sme očakávali, že dataset bude lepšie kategorizovať športy.



Obrázok 6 Dataset Sport1M

Samotný pred trénovaný model obsahoval všetky potrebné časti ako súbor s jednotlivými vrstvami, architektúru a parametre.

Pri práci s modelom sme najskôr odvodili zo štruktúry parametrov a vrstiev modelu: vstupný (input\_x), výstupný (logits) a trénovací (trainings) tenzor. Pracovali sme s knižnicou TensorFlow, takže na testovanie sme otvorili sedenie (session) a pracovali v ňom. TensorFlow bola naša voľba kvôli štruktúre modelu, ktorý obsahoval checkpoint (uložené pred trénované data modelu) a uložený graf. Na uloženie výsledkov sme použili knižnicu csv, ktorá vytvára, otvára a zapisuje údaje do csv súborov. Pre načítanie videí a prístup k ich snímkam sme použili knižnicu cv2, ktorá nám taktiež pomohla upraviť snímky na správne rozmery. Taktiež sme potrebovali vytvoriť umelé snímky, aby sme dostali správny rozmer pre vstupný tenzor. Po predikcii jednotlivých snímok sme spriemerovali výsledky pre jednotlivé videá a tieto údaje sme ukladali do csv súboru.

### C3D-UCF101

Ako aj pri modeli C3D-Sport1M sme pozerali na čo je zameraný dataset na ktorom natrénovali 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 2).

A collage of different activities

Description automatically generated

Obrázok 7 Dataset UCF101

Štruktúra pred trénovaného modelu bola rovnaká ako pri C3D-Sport1M, model taktiež obsahoval štyri súbory, ktoré určovali jeho architektúru, váhy a vrstvy. Pri práci sme teda zvolili rovnaký postup ako pri predošlom modeli.

### P3D-ResNet-199\_Kinetics a P3D-ResNet-199\_Kinetics-600

Oba modely P3D boli vybudované na architektúre ResNet a trénované na datasete Kinetics. Tento dataset obsahuje 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 v celku, ruky vykonávajúce aktivitu, športové aktivity, tváre ľudí a aj zábery rúk, ktoré manipulujú s objektami (Obrázok 3).

A collage of images of different people

Description automatically generated

Obrázok 8 Dataset Kinetics https://paperswithcode.com/dataset/kinetics

Dataset DeepMind Kinetics obsahuje štyristo videí pre každú udalosť. Každé video má okolo desať sekúnd a zdrojom je platforma YouTube. Videá sú zamerané na na ľudské aktivity a zahŕňa kategórie od infekcií ľudí s objektami (hranie na hudobné nástroje) ako aj interakcie medzi ľuďmi (padávanie rúk). Autori popisujú taktiež štatistické výsledky datastu a výkonnosť. Taktiež vykonali predbežnú analýzu toho, či nerovnováha v datasete vedie k skresleniu klasifikácie. (https://arxiv.org/pdf/1705.06950v1.pdf)

Architektúra na ktorej sú natrénované modely P3D vyžadovala menej priamočiary prístup ako pri testovaní C3D modelov. Pri C3D nám stačilo 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 a teda neurónovú sieť so 152 vrstvami. Keďže sme mali pred trénovaný model, tak tento výber bol pre nás najvhodnejší aj keď sme očakávali zhoršenie predikcií.

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

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

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

Oba modely sú vybudované na TRN architektúre, ale natrénované na iných datasetoch. Dataset Something-Something V2[[10]](#footnote-10) predstavuje okolo dvestodvadsať tisíc videí, ktoré sa zameriavajú na gestá rúk, vykonávanie bežných vecí pomocou rúk. Takže na rozdiel od datasetu ERA videá neobsahujú ľudí, stroje a ani autá. Medzi jeho kategórie paria napríklad: pokladať niečo na podklad, posúvať niečo, hádzať niečo, pokladať niečo vedľa niečoho a iné (Obrázok 1).

A collage of many images of different objects

Description automatically generated

Obrázok 9 Dataset Something-Something V2

Druhý dataset, Moments in Time[[11]](#footnote-11), predstavuje milión zozbieraných videí, ktoré sa zameriavajú na dynamické scény. Predstavuje veľkú škálu videí od zvieracích, hýbajúcich sa objektov až po ľudí pri pohybe.

A collage of images of various things

Description automatically generated

Obrázok 10 Dataset Moments in Time (<https://paperswithcode.com/dataset/moments-in-time#:~:text=for%20event%20understanding-,Moments%20in%20Time%20is%20a%20large-scale%20dataset%20for%20recognizing,gist%20of%20a%20dynamic%20scene>.)

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 u ostatných modelov, čiže po načítaní sme vytvorili csv súbor, spravili predikcie na vopred určených 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 model autori datasetu ERA taktiež natrénovali na datasete ImageNet. Dataset predstavuje viac ako štrnásť miliónov obrázkov, ktoré sú všetky popísané (Obrázok 6). Príkladom môže byť obrázok s popisom „na obrázku sú autá“, ale „obrázok neobsahuje tigre“. Taktiež majú anotáciu „na obrázku je skrutkovač umiestnený na pozícii (20,25) so šírkou 50 pixelov a výškou 30 pixelov“. (https://paperswithcode.com/dataset/imagenet)

A collage of many images

Description automatically generated

Obrázok 11 ImageNet dataset

Modely I3D mali oba dostupnú architektúru a postup na testovanie mali podobný modelom C3D. Najskôr sme načítali 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 požadovaných šestnásť do dimenzie tenzora a vytváraním nového rozmeru dávky (batch dimension).

## 3.4. Analýza dát

Ako hlavnú analýzu sme zvolili správnosť kategorizácie modelov. Všetky výsledky sme exportovali do csv súborov, ktoré sme následne spracovali v jednotnom súbore v programe Excel. Ako formu vizualizácie dát sme si zvolili tabuľky, do ktorých sme dali názvy videí a ich kategorizácie podľa modelov. Následné výpočty sme zautomatizovali pomocou formúl a vstavaných funkcií Excelu.

Popri tabuľkách sme zvolili vizualizáciu v grafoch.

## 3.7. Využité technológie

V tejto práci sme najmä využívali programovací jazyk Python (verzia Python 3.10.11). Pre načítanie a prácu s modelmi sme použili 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 Miscrosoft Excel a jeho funkcie, hlavne na urýchlenie výpočtov a dynamické analýzy. Pre ukladanie práce sme si zvolili GitHub[[12]](#footnote-12).

## 3.8. Hypotézy

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

- H1: Predpoklad, že model C3D-Sport1M bude lepšie klasifikovať videá zo športovej kategórie vzhľadom na to, že bol natrénovaný na športovom datasete.

- H2: Predpoklad, že model C3D-UCF101 bude horšie kategorizovať videá v oblasti detekcie áut, poľnohospodárstva a pohrôm vzhľadom na natrénovaný datset.

- H3: Predpokladáme, že naše testovacie výsledky jednotlivých modelov sa budú približovať výsledkom autorov datsetu ERA.

# 4 Výsledky

Pri práci s modelmi sme potrebovali často krát hľadať riešenia k problémom a optimalizácie našich riešení. Výsledky sme potrebovali analyzovať, interpretovať a porovnávať s výsledkami autorov ERA datasetu.

## 4.1. Analýza výsledkov

Po načítaní pred trénovaných modelov sme získali jednotlivé kategorizácie videí, ktoré sme následne vyhodnotili podľa ich presnosti (accuracy), odôvodnili výsledky a zlepšili predikcie.

### C3D-Sport1M

Naša hypotéza H1 nebola naplnená. Pri analýze sme porovnali priemerné percento správnych predikcií pre všetky kategórie: 25,07% a percentá priemerných predikcií pre športové kategórie (Basketbal, Basebal, Člnkovanie, Cyklistika, Beh, Futbal, Plávanie a Závody áut). V tabuľke 2 môžeme vidieť, že oproti priemeru zo všetkých kategórií sa predikcie výrazne nezlepšili pri športových videách.

Tabuľka 2 C3D-Sport1M predikcie pre športy

|  |  |
| --- | --- |
| Názov kategórie | Priemerná správna predikcia v % |
| Basebal | 26,67 |
| Basketbal | 26,67 |
| Člnkovanie | 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, ktoré sú unitárne v parametroch a hlavne sú snímané z pohľadu z vrchu.

Pri porovnaní našich výsledkov s výsledkami autorov datasetu ERA, sme nemali veľké odchýlky (Tabuľka 3)

Tabuľka 3 Porovnanie výsledkov nášho testovania s ERA výsledkami pre C3D-Sport1M

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |
| Člnkovanie | 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 |
| Policajná naháňačka | 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 3 obsahuje názov kategórie, naše výsledky z testovaní a dva stĺpce údajov z testovaní od autorov. Naše priemery sa zhodovali s výsledkami prvého modelu C3D. 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.

Najlepšie predikované kategórie boli: Člnkovanie a Futbal (obe 46,67%), čo sa plne nezhoduje s najčastejšie predikovanou Párty v celkovom počte (správna predikcia pre párty bola iba 40%).

Najväčší rozdiel v testovaných dátach autorov a našich bol 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 H3 za splnenú.

### C3D-UCF101

Tento model bol natrénovaný pomocou datasetu zameraného na ľudí a športy (UCF). Podľa hypotézy H2 sme predpokladali horšie kategorizovanie v kategóriách: závody áut, výstavba, požiar, záplavy, zber, zosun pôdy, zosun bahna, orba, po zemetrasení, dopravná zrážka a dopravná zápcha.

Náš predpoklad sa čiastočne splnil, medzi najhoršie predikované kategória patrili: koncert, konflikt, cyklistika, zosun pôdy, beh a zrážka áut (Tabuľka 4). 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 4).

Tabuľka 4 Porovnanie kľúčových kategórií

|  |  |
| --- | --- |
| Názov kategórie | Accuracy v % |
| Člnkovanie | 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 |

Tabuľka 5 Porovnanie výsledkov nášho testovania s ERA výsledkami pre C3D-UCF101

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |
| Člnkovanie | 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 |
| Policajná naháňačka | 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 |

Podľa Tabuľky 5 vidíme, že oproti predošlému modelu C3D sa nám priemerné percento predikcie zvýšilo o 3,47% (28,53%). Je to vidieť aj na jednotlivých predikciách, napríklad všetky kategórie boli predikované aspoň v jednom prípade z pätnástich videí v kategórii boli správne. Hypotézu H3 môžeme pokladať, pre tento model, za splnenú.

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 (rozdiel o 40,87%) a koncert (rozdiel o -21,53%). Model skôr v týchto kategóriách predikoval kategórie párty a výstavbu. Keďže ani jedna z týchto kategórií nebola medzi kategóriami datasetu UCF, tak sme nevedeli s určitosťou odôvodniť toto kategorizovanie.

Náš predpoklad, že sa predikovanie kategórie party rapídne zníži, sa nesplnil. Predpoklad vznikol pri skúmaní datasetu UCF, ktorý obsahoval videá zosnímané z boku 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).

### P3D-ResNet-199\_Kinetics a P3D-ResNet-199\_Kinetics-600

Modely nekategorizovali správne. Zakaždým priraďovali každé video do rovnakej kategórie, ktorá sa menila pri znovu pustení programu. Dôvod vidíme v zle nastavenej architektúre, kedy sme nemali 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, tak sme modely skúšali na podobných a teda ResNet18, ResNet34, ResNet101 a alternatívach k nim ResNeXt (architektúry vybudované na zlepšenie výkonnosti, predstavujú parameter kardinality). Naše výsledky boli zhodné vo všetkých architektúrach.

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 dostupné. Implementovali sme výpočet, ktorý zobral set všetkých predikcií, vybral maximum a priradil ho ku kategórii. (Kód 1)

predicted\_category\_index = max(set(all\_predictions), key=all\_predictions.count)

*Kód 1 Výpočet dočasného zoskupenia.*

Metóda zoskupenia nezlepšila kategorizovanie a vo výsledku tieto modely nesplnili predpoklad H3.

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

Modely 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.

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

Vo výsledku sme dostali lepšie kategorizácie, modely začali kategorizovať do všetkých kategórií a nedostali sme iba jednu výslednú kategóriu. Avšak celkovým počtom správnych kategorizácií sme neprišli k veľkému zlepšeniu. 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 2).

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 2 Časový kontext implementovaný pre TRN modely.*

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 do tried. Taktiež sa zachovalo správanie modelov, ktoré kategorizovali do všetkých tried a nie iba pre do jednej.

Vo výsledku sa ale kategorizovanie nepriblížilo percentu správnych zaradení do tried autorom ERA datasetu a teda hypotézu H3 zamietame.

### I3D-Kinetics a I3D-Kinetics+ImageNet

Pri testovaní sme pri základnom načítaní a otestovaní dostali kategorizáciu všetkých videí do jednej kategórie (pri IC3-Kinetics kategória „Party“ a pri I3D-Kinetics+ImageNet kategória „ReligiousActivity“), to znamená, že správne kategorizovali pätnásť videí.

Pre 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.

Ako druhú sme implementovali náhodnú augmentáciu snímok (random augmentation). Metódu 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 3 Implementácie náhodnej augmentácie.*

Výsledky s augmentáciou sa nezlepšili, zostala kategorizácia do skoro všetkých kategórií, ale správne sme dostali pri I3D-Kinetics iba osem videí a pri I3D-Kinetics+ImageNet deväť videí.

Zamietame hypotézu H3 pre tieto modely z dôvodu malého percenta správnych kategorizácií, ktoré sa nepribližuje autorom datasetu.

## 4.2. Vyhodnotenie a diskusia

V našej práci sme porovnávali presnosť pred trénovaných modelov na datasete ERA. Modely okrem C3D-Sport1M a C3D-UCF101 nemali podobné výsledky s výsledkami autorov datasetu ERA (Graf 1).

Graf 1 Porovnanie výsledkov

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 v rovnakom postupe. Bolo potrebné vybrať správne parametre a navoliť tenzory. Pri rovnakých videách sme pre C3D modely dostali veľmi príbuzné výsledky ako prezentovali autori datasetu, avšak pre I3D naše výsledky boli úplne iné. Pri druhom spôsobe exportovania súborov (modely TRN a P3D) sme mali podobne nesúhlasné výsledné dáta.

Našim ďalším pokusom o priblíženie sa k výsledkom autorov boli rôzne optimalizácie. Tie vo všeobecnosti pomohli, ale nie dostatočne aby sa dali považovať za úplne úspešné.

Ako ďalší dôvod odlišnosti vo výsledkoch môže byť veľkosť testovacích dát. Autori rozdelili dataset na trénovacie a testovacie videá, s jednotlivými kategóriami, avšak pri dostupnom technickom vybavení bolo potrebné testovací dataset zmenšiť. To mohlo skresliť výsledné dáta, keďže naša vzorka neobsahovala úplný počet videí.

V rámci optimalizácie výsledkov sme skúmali štyri metódy, ktoré nám pri viacerých modeloch priniesli zlepšenie. V rámci ďalšieho výskumu by sme navrhovali vyskúšanie iných metód poprípade natrénovanie modelov od začiatku.

Pri vytváraní práce sme pre testovanie modelov využívali programovací jazyk Python. Tento jazyk sme si vybrali z dôvodu veľkého katalógu knižníc, ktoré sa priamo zameriavajú na prácu s neurónovými sieťami a jeho predošlému využívaniu počas štúdia na vysokej škole.

Samotné programovanie sme vykonávali v prostredí PyCharm, s ktorým sme mali taktiež predošlé skúsenosti a bolo vhodné vďaka rýchlej inštalácii samotných knižníc a nástroju na detekciu chýb (debugger). Ten nám urýchlil opravovanie syntaktických a logických chýb v kóde. Taktiež sme využili aj rýchle prehľadávanie funkcií knižníc.

Na vizualizáciu a spracovanie výsledkov sme využili program Microsoft Excel. Vďaka vbudovaným funkciám sme vedeli na rôznych hárkoch oddeliť výsledky modelov a vzájomne ich prepojiť do jedného na porovnanie výsledkov. Taktiež sme využili mnohé funkcie ( COUNTIFS, SUM, ABS, TRIM, LEFT) a kontingenčné tabuľky. Na vizualizáciu sme využili podmienené formátovanie a grafy.

## 4.3. Odporúčania

Dataset ERA sám o sebe je zostavený veľmi dobre a odporúčame jeho využitie najmä pri modeloch s menším dôrazom na presnosť kategorizácií, poprípade na využitie vo vzdelávacej sfére. Vidíme jeho prínos vzhľadom na celistvú štruktúru, dobré delenie, označenie videí a nie veľký priveľký počet videí. Študenti zameriavajúci sa na klasifikáciu vo videách by vedeli využiť tento dataset na naučenie sa princípov fungovania klasifikácie a využitie v praxi.

Taktiež, vďaka dobrému členeniu videí do kategórií a oblastí, sa dá ľahko využiť časť videí a zlúčiť viaceré datasety pre komplexné natrénovanie modelu. Ako prínos vidíme to, že ide o dronové snímky, ktoré nebývajú zastúpené bežne vo veľkých datasetoch.

Pri analýze datasetu sme taktiež pozorovali, že autori sa snažili obsiahnuť širší záber oblastí. Dataset obsahuje od športových videí cez videá z dopravy až po prírodné katastrofy. Toto môže byť aj dôvod prečo presnosť nie je úplne najvyššia. Z nášho pohľadu by bolo lepšie zamerať sa na konkrétnu oblasť a natrénovať modely len na predikciu napríklad druhov športov alebo pohrôm. To by sme videli aj ako prínos do reálneho života, kedy sa takéto systémy na detekciu využívajú cielene a tým vyžadujú aj vysokú presnosť.

Ako ďalší prínos vidíme vo vložení šumových snímok alebo videí do datasetu, ktoré by predstavovali priblíženie k realite a predchádzali by preučeniu.

Vzhľadom na trénovanie a testovanie modelov, by sme odporúčali natrénovať vlastné modely, poprípade použiť našu prácu k zlepšeniu práce s modelmi s cieľom vylepšiť presnosť jednotlivých modelov.

# Záver

Cieľom diplomovej práce bolo vytvoriť výskum na tému rozpoznávania objektov vo videách. Pre tento cieľ sme si vybrali prácu s datasetom ERA. Dataset obsahoval pred trénované modely, ktoré sme mali otestovať.

Výsledky nami otestovaných modelov sa až na modely C3D nepribližovali výsledkom autorov. Pri testovaní sme skúmali architektúru a obmedzené dostupné informácie k pred trénovaným modelom, na základe ktorých sme vytvorili kódy v programovacom jazyku Python. Pre testovanie sme vybrali náhodne videá a netestovali sme celý dostupný dataset určený na testovanie z dôvodu technického obmedzenia, čo mohlo viesť k zhoršeniu výsledkov. Taktiež nezhodné výsledky sa dajú odôvodniť absenciou dokumentácie o trénovaní modelov. Pre zlepšenie výsledkov sme skúmali a otestovali možné optimalizačné metódy. Tie vo výsledku priniesli malé zlepšenie.

Na porovnávanie modelov sme používali metriku presnosti, ktorá určovala s akou percentuálnou hodnotou model predikoval na daných videách kategórie.

Pre analýzu a porovnanie výsledkov sme využili program Microsoft Excel, kde sme prehľadne zdokumentovali jednotlivé kategorizácie a vytvorili formuly na ľahšiu manipuláciu. Na vizualizáciu sme vytvorili grafy, ktoré sme pridali do práce.

Prácu pokladáme za dobrý podklad pre ďalšie výskumy. Medzi potencionálne oblasti by sme zaradili v prvom bode zlepšenie kategorizácie modelov a tým optimalizovanie výsledkov. Taktiež sme navrhli nové implementácie, ktoré by pomohli využiť modely v reálnej praxi. Napríklad pridať šumové snímky alebo videá, implementovať nové optimalizácie alebo vylepšiť testovanie modelov.

V rámci výskumu k diplomovej práci v oblasti klasifikácie vo videách sme preskúmali viaceré využívané datasety a zhodnotili sme kvalitu videí v ERA datasete ako dobrú na trénovanie nových datasetov. Tu vidíme aj potencionálne využitie tohto datasetu na nové výskumné práce, ktorých cieľom by bolo zdokumentovať trénovanie neurónových sietí.

Naša práca prináša základ pre prácu s pred trénovanými modelmi, ktorá je rozšíriteľná, ale zároveň komplexná a vidíme jej potencionálne využitie aj v edukačnej sfére. Študenti by mohli prácu využiť ako podklad pre testovanie pred trénovaných modelov alebo pre prácu s datasetom ERA, ktorý je voľne dostupný.

Naša diplomová práca sa skladá:

- z výskumnej časti, ktorá je popísaná v tomto dokumente,

- z kódovej časti, ktorá je taktiež zdokumentovaná a je voľne dostupná na platforme GitHub[[13]](#footnote-13),

- z výsledkov jednotlivých modelov, ktoré sú taktiež na platforme GitHub

- a zo súboru xlsx, v ktorom sú zozbierané, spracované a vizualizované výsledky.

# Zoznam bibliografických odkazov

Ahmad, Muhammad Waseem, Monjur Mourshed, a Yacine Rezgui. 2017. “Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption”. *Energy and Buildings* 147: 77–89. doi:10.1016/j.enbuild.2017.04.038.

Bovik, Alan C. 2009. *The Essential Guide to Video Processing*. Academic Press.

Carreira, João, a Andrew Zisserman. 2017. “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”. V *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, , 4724–33. doi:10.1109/CVPR.2017.502.

Delextrat, Anne, a Daniel Cohen. 2009. “Strength, Power, Speed, and Agility of Women Basketball Players According to Playing Position”. *The Journal of Strength & Conditioning Research* 23(7): 1974. doi:10.1519/JSC.0b013e3181b86a7e.

Díaz, Aldo André, a Paulo R. G. Kurka. 2020. “Computer Vision Methods for Automotive Applications”. *TECNIA* 30(2): 74–81. doi:10.21754/tecnia.v30i2.801.

“Duplicate Frame Video Forgery Detection Using Siamese-based RNN”. https://www.techscience.com/iasc/v29n3/43054/html (02. december 2023).

Gama, Fernando, Antonio G. Marques, Alejandro Ribeiro, a Geert Leus. 2018. “MIMO Graph Filters for Convolutional Neural Networks”. http://arxiv.org/abs/1803.02247 (02. december 2023).

Hadidi, Ramyad, Jiashen Cao, Michael S. Ryoo, a Hyesoon Kim. 2020. “Toward Collaborative Inferencing of Deep Neural Networks on Internet-of-Things Devices”. *IEEE Internet of Things Journal* 7(6): 4950–60. doi:10.1109/JIOT.2020.2972000.

Chattopadhyay, Chiranjoy, a 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* 2(4): 289–97. doi:10.1007/s13735-013-0034-8.

Chayambuka, Kudakwashe, Jan Fransaer, a Xochitl Dominguez-Benetton. 2019. “Modeling and design of semi-solid flow batteries”. *Journal of Power Sources* 434: 226740. doi:10.1016/j.jpowsour.2019.226740.

Kortylewski, Adam, Qing Liu, Angtian Wang, Yihong Sun, a Alan Yuille. 2020. “Compositional Convolutional Neural Networks: A Robust and Interpretable Model for Object Recognition under Occlusion”. doi:10.48550/arXiv.2006.15538.

Li, Qingwu, Haisu Cheng, Yan Zhou, a Guanying Huo. 2016. “Human Action Recognition Using Improved Salient Dense Trajectories”. *Computational Intelligence and Neuroscience* 2016: e6750459. doi:10.1155/2016/6750459.

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* 2022: e9422584. doi:10.1155/2022/9422584.

Li, Yinqing, Violeta G. Lopez-Huerta, Xian Adiconis, Kirsten Levandowski, Soonwook Choi, Sean K. Simmons, Mario A. Arias-Garcia, et al. 2020. “Distinct Subnetworks of the Thalamic Reticular Nucleus”. *Nature* 583(7818): 819–24. doi:10.1038/s41586-020-2504-5.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, a Piotr Dollár. 2017. “Focal Loss for Dense Object Detection”. V *2017 IEEE International Conference on Computer Vision (ICCV)*, , 2999–3007. doi:10.1109/ICCV.2017.324.

Min-qing, Zhang, a Li Wen-ping. 2021. “An Automatic Classification Method of Sports Teaching Video Using Support Vector Machine”. *Scientific Programming* 2021: e4728584. doi:10.1155/2021/4728584.

Mou, Lichao, Yuansheng Hua, Pu Jin, a Xiao Xiang Zhu. 2020. “ERA: A Dataset and Deep Learning Benchmark for Event Recognition in Aerial Videos”. doi:10.48550/arXiv.2001.11394.

Mueller, Jonas, a Aditya Thyagarajan. 2016. “Siamese Recurrent Architectures for Learning Sentence Similarity”. *Proceedings of the AAAI Conference on Artificial Intelligence* 30(1). doi:10.1609/aaai.v30i1.10350.

Ng, Joe Yue-Hei, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, a George Toderici. 2015. “Beyond short snippets: Deep networks for video classification”. V *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, , 4694–4702. doi:10.1109/CVPR.2015.7299101.

Papert, Seymour A. 1966. “The Summer Vision Project”. https://dspace.mit.edu/handle/1721.1/6125 (02. december 2023).

Qu, Haoxuan, Hossein Rahmani, Li Xu, Bryan Williams, a Jun Liu. 2023. “Recent Advances of Continual Learning in Computer Vision: An Overview”. http://arxiv.org/abs/2109.11369 (02. december 2023).

“Sensors | Free Full-Text | Deep Learning-Based Real-Time Multiple-Person Action Recognition System”. https://www.mdpi.com/1424-8220/20/17/4758 (02. december 2023).

Sharma, Vijeta, Manjari Gupta, Ajai Kumar, a Deepti Mishra. 2021. “Video Processing Using Deep Learning Techniques: A Systematic Literature Review”. *IEEE Access* 9: 139489–507. doi:10.1109/ACCESS.2021.3118541.

Shi, Zhensheng, Liangjie Cao, Cheng Guan, Haiyong Zheng, Zhaorui Gu, Zhibin Yu, a Bing Zheng. 2020. “Learning Attention-Enhanced Spatiotemporal Representation for Action Recognition”. *IEEE Access* 8: 16785–94. doi:10.1109/ACCESS.2020.2968024.

Terney, Daniella, Leila Chaieb, Vera Moliadze, Andrea Antal, a Walter Paulus. 2008. “Increasing Human Brain Excitability by Transcranial High-Frequency Random Noise Stimulation”. *Journal of Neuroscience* 28(52): 14147–55. doi:10.1523/JNEUROSCI.4248-08.2008.

Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, a Manohar Paluri. 2015. “Learning Spatiotemporal Features with 3D Convolutional Networks”. V *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile: IEEE, 4489–97. doi:10.1109/ICCV.2015.510.

Wang, Heng, a Cordelia Schmid. 2013. “Action Recognition with Improved Trajectories”. V *2013 IEEE International Conference on Computer Vision*, , 3551–58. doi:10.1109/ICCV.2013.441.

Wang, Limin, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, a Luc Van Gool. 2016. “Temporal Segment Networks: Towards Good Practices for Deep Action Recognition”. V *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, ed. Bastian Leibe, Jiri Matas, Nicu Sebe, a Max Welling. Cham: Springer International Publishing, 20–36. doi:10.1007/978-3-319-46484-8\_2.

Wang, Xuan, a Zhigang Zhu. 2023. “Context understanding in computer vision: A survey”. *Computer Vision and Image Understanding* 229: 103646. doi:10.1016/j.cviu.2023.103646.

Wei, Pengbo, David Ahmedt-Aristizabal, Harshala Gammulle, Simon Denman, a Mohammad Ali Armin. 2022. “Vision-Based Activity Recognition in Children with Autism-Related Behaviors”. http://arxiv.org/abs/2208.04206 (02. december 2023).

Wu, Di, Nabin Sharma, a Michael Blumenstein. 2017. “Recent advances in video-based human action recognition using deep learning: A review”. V *2017 International Joint Conference on Neural Networks (IJCNN)*, , 2865–72. doi:10.1109/IJCNN.2017.7966210.

Zhang, Bokai, Amer Ghanem, Alexander Simes, Henry Choi, a Andrew Yoo. 2021. “Surgical Workflow Recognition with 3DCNN for Sleeve Gastrectomy”. *International Journal of Computer Assisted Radiology and Surgery* 16(11): 2029–36. doi:10.1007/s11548-021-02473-3.

Zhang, Ke, Wei-Lun Chao, Fei Sha, a Kristen Grauman. 2016. “Video Summarization with Long Short-Term Memory”. V *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, ed. Bastian Leibe, Jiri Matas, Nicu Sebe, a Max Welling. Cham: Springer International Publishing, 766–82. doi:10.1007/978-3-319-46478-7\_47.

Zhang, Qi, Jianlong Chang, Gaofeng Meng, Shiming Xiang, a Chunhong Pan. 2020. “Spatio-Temporal Graph Structure Learning for Traffic Forecasting”. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(01): 1177–85. doi:10.1609/aaai.v34i01.5470.

Zhang, Zhiwu, Elhan Ersoz, Chao-Qiang Lai, Rory J. Todhunter, Hemant K. Tiwari, Michael A. Gore, Peter J. Bradbury, et al. 2010. “Mixed Linear Model Approach Adapted for Genome-Wide Association Studies”. *Nature Genetics* 42(4): 355–60. 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* 2022: e3407935. doi:10.1155/2022/3407935.

Zhou, Bolei, Alex Andonian, Aude Oliva, a Antonio Torralba. 2018. “Temporal Relational Reasoning in Videos”. V *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, ed. Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, a Yair Weiss. Cham: Springer International Publishing, 831–46. doi:10.1007/978-3-030-01246-5\_49.

Zhou, Longfei, Lin Zhang, a Nicholas Konz. 2022. “Computer Vision Techniques in Manufacturing”. doi:10.36227/techrxiv.17125652.v2.

# Zoznam príloh

Príloha A – dátový súbor + zdrojový kód + odkaz na GIT + výsledky výskumu, rozsiahle tabuľky a pod.

Príloha B – Fotodokumentácia k prípravnej fáze záverečnej práce ........

PRÍLOHA A

**Príloha A**

* A1 Odkaz na GIT: https://github.com/
* A2 ...
* A3 ...
* A4 ...

1. https://www.webofscience.com/wos/woscc/basic-search [↑](#footnote-ref-1)
2. <https://lcmou.github.io/ERA_Dataset/> [↑](#footnote-ref-2)
3. https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview [↑](#footnote-ref-3)
4. https://lcmou.github.io/ERA\_Dataset/ [↑](#footnote-ref-4)
5. https://elib.dlr.de/140908/1/2001.11394.pdf [↑](#footnote-ref-5)
6. <http://www.stat.ucla.edu/~tianmin.shu/AerialVideo/AerialVideo.html> [↑](#footnote-ref-6)
7. http://okutama-action.org [↑](#footnote-ref-7)
8. https://github.com/ckyrkou/AIDER [↑](#footnote-ref-8)
9. https://cs.stanford.edu/people/karpathy/deepvideo/ [↑](#footnote-ref-9)
10. https://developer.qualcomm.com/software/ai-datasets/something-something [↑](#footnote-ref-10)
11. http://moments.csail.mit.edu [↑](#footnote-ref-11)
12. https://github.com/batsysk/DiplomovaPraca [↑](#footnote-ref-12)
13. https://github.com/batsysk/DiplomovaPraca [↑](#footnote-ref-13)