**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 môže byť vyjadrenie poďakovania napr. školiteľovi resp. konzultantovi za pripomienky a odbornú pomoc pri vypracovaní práce. Nie je zvykom ďakovať za rutinnú kontrolu, menšiu spoluprácu, alebo všeobecné rady.

*NEPOVINNÉ*

# ABSTRAKT

PRIEZVISKO, Meno autora: Názov práce. [Bakalárska/Diplomová práca]. Univerzita Konštantína Filozofa v Nitre. Fakulta prírodných vied a informatiky. Školiteľ: Titul. Meno Priezvisko, Titul. Stupeň odbornej kvalifikácie: Bakalár odboru Aplikovaná informatika. Nitra: FPVaI, 202x. .... 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

[DIPLOMOVÁ práca 2](#_Toc152870130)

[Zadanie ZP 3](#_Toc152870131)

[Poďakovanie 4](#_Toc152870132)

[ABSTRAKT 5](#_Toc152870133)

[ABSTRACT 6](#_Toc152870134)

[Obsah 7](#_Toc152870135)

[Úvod ..................................................................................................................................9](#_Toc152870136)

[1 Analýza súčasného stavu 10](#_Toc152870137)

[1.1 Computer Vision 10](#_Toc152870138)

[1.2 Základné metódy computer vision 11](#_Toc152870139)

[1.3. Modely klasifikujúce vo videách 12](#_Toc152870140)

[1.3.1 CNN 13](#_Toc152870141)

[1.3.2 RNN 14](#_Toc152870142)

[1.3.3 SVM 14](#_Toc152870143)

[1.3.4 3D Konvolunčné modely 16](#_Toc152870144)

[1.4 Modely využité pri analýze ERA datasetu 17](#_Toc152870145)

[**C3D Model** 17](#_Toc152870146)

[**P3D Model** 19](#_Toc152870147)

[**I3D Model** 20](#_Toc152870148)

[**TRN Model** 21](#_Toc152870149)

[1.3.5 Ďalšie modely 22](#_Toc152870150)

[1.4. Porovnanie modelov 23](#_Toc152870151)

[2 Ciele záverečnej práce 25](#_Toc152870152)

[3 Metodika výskumu 26](#_Toc152870153)

[3.1. Porozumenie problematike 26](#_Toc152870154)

[3.1.1. Klasifikácia vo videách 26](#_Toc152870155)

[3.2. Porozumenie dátam 26](#_Toc152870156)

[3.2.1 Dataset ERA 27](#_Toc152870157)

[3.3. Spracovanie dát 27](#_Toc152870158)

[3.4. Výber modelov 27](#_Toc152870159)

[3.5. Analýza dát 27](#_Toc152870160)

[3.6. Evalvácia výsledkov 28](#_Toc152870161)

[3.7. Vyhodnotenie a interpetácia 28](#_Toc152870162)

[Zoznam bibliografických odkazov 29](#_Toc152870163)

[Zoznam príloh 33](#_Toc152870164)

# 

# 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 (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 postupne samé. Taktiež aby sa nemuseli vytvárať modely nové od základu. Počítačové videnie pridáva k základom neurónových sietí učenie sa postupne v čase a tým idú do popredia. (Qu et al. 2023)

Počítačové videnie spracováva jednoduché popisy a teda čo vidí na scéne. Vzory obsahujú pochopenie scény a tie sa využívajú v modeloch. Algoritmy modelov vieme rozdeliť do kategórií (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) a to v roku 1966. V posledných desaťročiach počítačové videnie napreduje hlavne vďaka posunu v technológiách, ktoré umožňujú vyššiu výpočtovú náročnosť. (Sharma, Gupta, a Jaiswal 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. (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, na ktorom je text, snímka obrazovky, pestrofarebnosť, grafický dizajn

Automaticky generovaný popis

## 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. (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. (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)

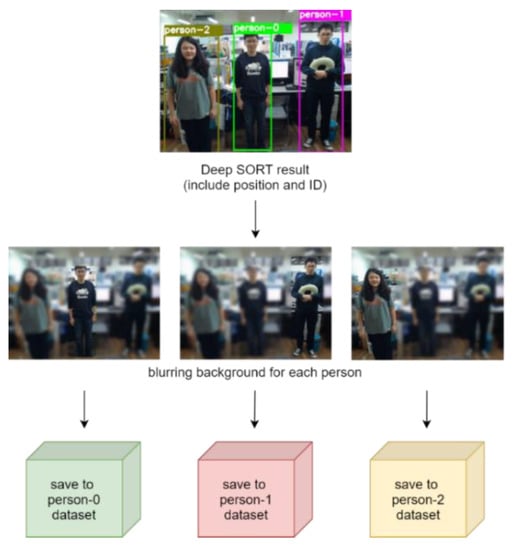
V športoch sa napríklad využívajú na rozoznanie typu športu a aktivity v reálnom čase(obrázok). Na základe monitorovania ľudí, konkrétne v tomto prípade ich stredu tela a štvorcového mapovania, sa vykonávajú pozorovania ich správania. Takéto modelovanie sa nevyužíva iba v analýze športovcov a ich výkonov,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

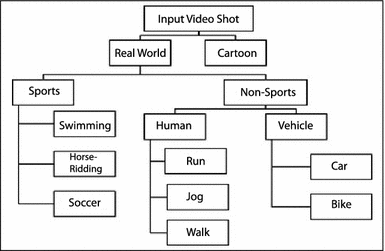
SVM alebo Support Vector Machine je algoritmus strojového učenia, ktorý sa často využíva pre klasifikačné a regresné úlohy. Pri videách sa momentálne skúma využitie SVM na kategorizáciu druhov videí a to napríklad na reklamy, kreslené rozprávky, hudobné videá, správy a športové videá (podľa výskumu Zhang 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.

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)

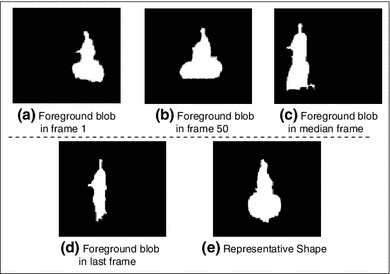
Obrázok, na ktorom je diagram, rad, náčrt, kresba

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 obrovské využitie v rozpoznávaní kontextu videa avšak do budúcna odporúčajú pridať ďalšie média ako zvuk a text. (Chattopadhyay a Maurya 2013)

### 1.3.4 3D Konvolunčné modely

CNN (convolutional neural network) je dopredná neurónová sieť, ktorá berie vstupné dáta a spracúva ich cez niekoľko vrstiev neurónovej siete, pričom každá z nich vykonáva tri operácie: konvolúcia, bodová nelinearita a zhlukovanie. Na základe tejto architektúry dokáže model extrahovať užitočné informácie. (Gama et al. 2018)

Z obrázku(obrázok) vidíme, že sa CNN skladá z dvoch hlavných časti a to extrakcia vlastností a klasifikácia. Pri extrakcii vlastností sa konvolunčná vrstva snaží na základe filtrov(mriežkové matice), ktoré sa posúvajú(konvulujú), získať číslo pre pozíciu prvku za pomoci násobenia. Toto číslo hovorí o sile prvku. CNN majú dôležitú vlastnosť a to zdieľanie váh, čo znamená, že sa každý filter požíva na každý vstupný prvok. Pri časti pooling ide o väčšinové hlasovanie. Tu sa vyfiltrujú jednotlivé prvky, ktoré následne putujú do neurónových sietí. (Mesuga a Bayanay 2021)

Obrázok, na ktorom je diagram, rad, text, plán

Automaticky generovaný popis

## 1.4 Modely využité pri analýze ERA datasetu

Autori Mou, L., Hua, Y., Jin, P. a Zhu, X. X., 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 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. (Wang et al. 2016)

Podľa výskumu, ktorý vykonal Du Tran a jeho kolegovia na učení modelu C3D, 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. (Tran et al. 2015)

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 (Wang a Schmid 2013). Shang 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. (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. (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. (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. (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 a jeho kolegov 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. (Full article: 3D genome organization in health and disease: emerging opportunities in cancer translational medicine n.d.)

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 konvolunčné 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)

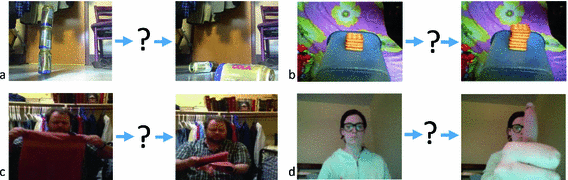
Focal Loss je špeciálny druh loss funkcie, ktorá sa zaoberá problémom nerovnováhy tried pri úlohách detekcie objektov. Keďže pri tejto problematike väčšina snímok pozostáva z pozadia, zatiaľ čo pre nás je dôležitá detekcia objektu záujmu, môže nerovnováha spôsobovať malú úspešnosť detegovania objektov 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. (Zhou et al. 2018)



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

Vo výskumoch neurovedy bol tento model skúmaný pri regulácii senzorického spracovania, pozornosti a kognície. (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

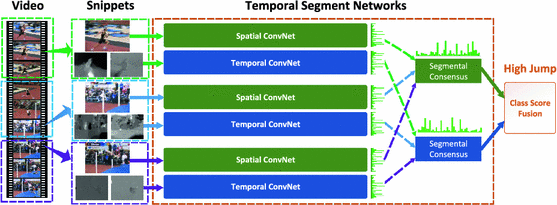
V dnešnej dobe sa výskumníci snažia modely zlepšovať a za týmto cieľom vznikli aj modely TSN a LSTM.

**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.

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. (Wang et al. 2016)

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



**LSTM**

Ďalším rozšíreným modelom je 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. (Zhang et al. 2016)

## 1.4. Porovnanie modelov

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

Automaticky generovaný popis

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 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 | 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 Konvolunčné 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 |

# 2 Ciele záverečnej práce

Hlavným cieľom práce je porovnať kvalitu modelov na tému klasifikácie vo videách. Tieto poznatky neskôr využijeme v praxi, kde porovnáme modely na datasete ERA.

Naše čiastkové ciele sú:

* Porovnať modely medzi sebou a nájsť ich najlepšie a najhoršie využitia na konkrétnych druhoch videí.
* Implementovať modely použité na datasete ERA a overiť ich výsledky.
* Zhodnotiť modely.
* Použiť ďalšie modely, ktoré sú momentálne využívané, na datasete ERA.
* Porovnať a interpretovať výsledky, ktoré nadobudneme pri práci s modelmi.
* Vizualizovať výsledky a použiť štatistické metódy na ich porovnanie.

# 3 Metodika výskumu

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

## 3.1. Porozumenie problematike

V prvom rade sme potrebovali porozumieť ako funguje klasifikácia vo videách a v čom sa líši od klasifikácie napríklad obrázkov a prečo existuje potreba využívať klasifikáciu pre videá.

### 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[[3]](#footnote-3) (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[[4]](#footnote-4), 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[[5]](#footnote-5) | Zamerané na človeka | Získané osobne autormi | áno | 12 | 104 | 2015 |
| Okutama[[6]](#footnote-6) | Ľudská činnosť | Získané osobne autormi | áno | 12 | - | 2017 |
| AIDER[[7]](#footnote-7) | 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.

### Videá v datasete

Dataset ERA pozostáva z 25 kategórií a 2,864 videí. Kategórie 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. Spracovanie dát

- Bolo potrebné videá nejako spracovať? Napríklad spraviť nejaký umelý šum, rozdeliť ich, skrátiť, pomenovať,...

## 3.4. Výber modelov

Pri modeloch sme si spravili prieskum aké sa využívajú v dnešnej dobe plus sme využili modely, ktoré aj samotní autori datasetu ERA[[8]](#footnote-8) používali na modelovanie.

- Popísať v skratke prečo ešte

## 3.5. Analýza dát

- Aké formy analýzy zvolíme a prečo.

## 3.6. Evalvácia výsledkov, Vyhodnotenie a interpetácia

- Aké štatistické metódy si zvolíme, ako budeme hodnotiť modely a prečo sú to vhodné spôsoby hodnotenia.

- Aké predpoklady si zvolíme a ako vieme na ich základe interpretovať výsledky, ktoré dosiahneme.

## 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 Excel a jeho funkcie, hlavne na urýchlenie výpočtov a dynamické analýzy. Pre ukladanie práce sme si zvolili GitHub[[9]](#footnote-9).

# 4 Výsledky

todo

## 4.2. 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. (metodika/diskusia)

### 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[[10]](#footnote-10). Dataset obsahoval viac ako jeden milión videí, ku ktorým sa pristupovalo cez URL adresu. Videá boli rozdelených do 467 športových aktivít. Keďže videá zachytávali športové udalosti tak sme očakávali, že dataset bude lepšie kategorizovať športy.

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

Náš postup pri testovaní modelu bol nasledovný:

- Zo štruktúry parametrov a vrstiev modelu sme si odvodili 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.

- 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 trinásť tisíc videí kategorizovaných do dvadsaťpäť kategórií, ktoré sa delili do celkov:

- interakcia človeka s objektami,

- pohyb a aktivity ľudí,

- interakcia medzi ľuďmi,

- a športy.

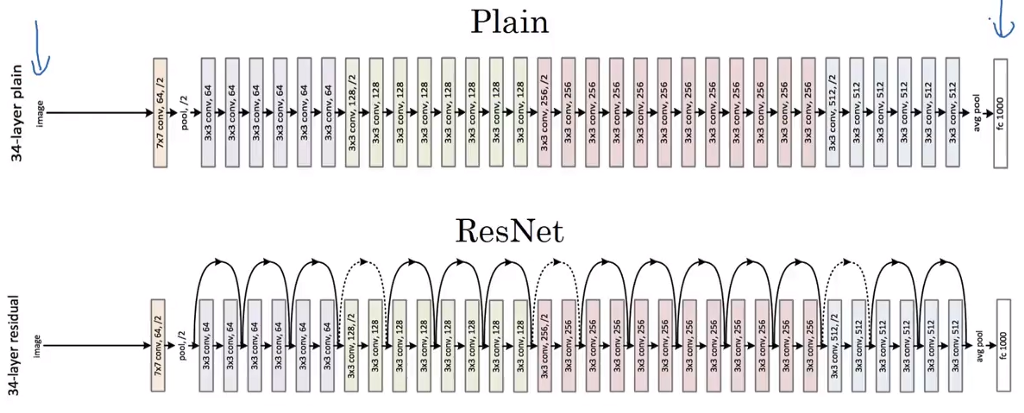
Dataset predstavoval komplexný súbor videí, z ktorého sme predpokladali dobré výsledky pri testovaní modelu.

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.

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. (https://arxiv.org/pdf/1611.01186.pdf)



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

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

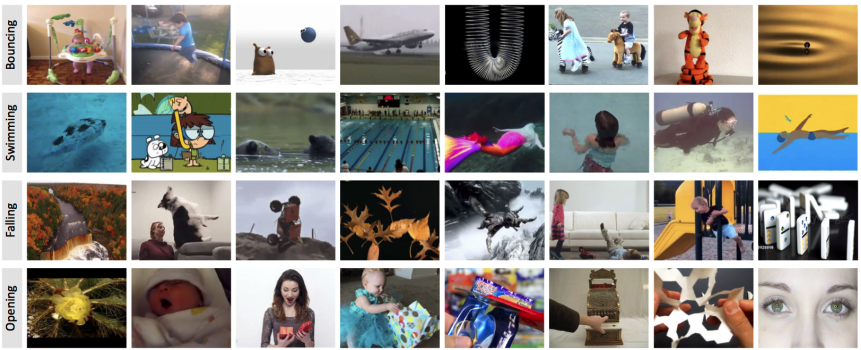
Oba modely sú vybudované na TRN architektúre, ale natrénované na iných datasetoch. Dataset Soemthing-Something V2[[11]](#footnote-11) predstavuje okolo dvestodvadsať tisíc videí, ktoré sa zameriavajú na gestá rúk, vykonávanie bežných vecí pomocou rúk. Takže na rozdiel od datasetu ERA videá neobsahujú 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 1 Dataset Something-Something V2

Druhý dataset, Moments in Time[[12]](#footnote-12), predstavuje milión zozbieraných videí, ktoré sa zameriavajú na dynamické scény. Predstavuje veľkú škálu videí od zvieracích, hýbajúcich sa objektov až po ľudí pri pohybe.



Obrázok 2 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>.)

Keďźe 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.

## 4.3. Analýza výsledkov

todo

### C3D-Sport1M

Náš predpoklad, že model bude lepšie klasifikovať videá zo športovej kategórie, nebol naplnený. 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 percentách |
| 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 |

Dôvodom 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 predikcia v % | C3DII – ERA predikcia v % |
| Basebal | 26,67 | 40,9 | 45,7 |
| Basketbal | 26,67 | 37,0 | 48,9 |
| Člnkovanie | 46,67 | 47,5 | 42,9 |
| 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 | 1,24 |
| 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.ň,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%.

### C3D-UCF101

Tento model bol natrénovaný pomocou datasetu zameraného na ľudí a športy (UCF). Predpokladali sme, že vo výsledkoch bude model horší v oblastiach detekcie áut, poľnohospodárstva a pohrôm (kategórie: 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 Porovnanie výsledkov nášho testovania s ERA výsledkami pre C3D-UCF101

|  |  |  |  |
| --- | --- | --- | --- |
| Názov kategórie | Priemerná správna predikcia v % | C3DI – ERA predikcia v % | C3DII – ERA predikcia v % |
| Basebal | 40,00 | 40,90 | 0,90 |
| Basketbal | 46,67 | 37,00 | 9,67 |
| Člnkovanie | 46,67 | 47,50 | 0,83 |
| Závody áut | 20,00 | 16,70 | 3,30 |
| Koncert | 6,67 | 28,20 | 21,53 |
| Konflikt | 6,67 | 18,20 | 11,53 |
| Výstavba | 46,67 | 45,50 | 1,17 |
| Cyklistika | 6,67 | 20,60 | 13,93 |
| Požiar | 26,67 | 30,90 | 4,23 |
| Potopy | 40,00 | 24,30 | 15,70 |
| Zber | 33,33 | 27,50 | 5,83 |
| Zosun pôdy | 6,67 | 19,50 | 12,83 |
| Zosun blata | 26,67 | 32,90 | 6,23 |
| Bez udalosti | 26,67 | 29,60 | 2,93 |
| Prehliadka protest | 20,00 | 37,80 | 17,80 |
| Párty | 66,67 | 25,80 | 40,87 |
| Orba | 33,33 | 36,10 | 2,77 |
| Policajná naháňačka | 33,33 | 50,00 | 16,67 |
| Po zemetrasení | 26,67 | 23,10 | 3,57 |
| Náboženská udalosť | 33,33 | 27,50 | 5,83 |
| Beh | 6,67 | 12,00 | 5,33 |
| Futbal | 40,00 | 58,30 | 18,30 |
| Plávanie | 33,33 | 36,20 | 2,87 |
| Dopravná zrážka | 6,67 | 7,00 | 0,33 |
| Dopravná zápcha | 33,33 | 15,50 | 17,83 |
| Priemer v % | 28,53 | 30,40 | 1,87 |

Podľa Tabuľky 4 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.

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.*

Táto metóda nám zlepšila výsledky aspoň z pohľadu, že sme model nepredikoval všetky videá na jednu kategóriu. Teraz sme mali predikcie naprieč všetkými kategóriami a zo tristosedemdesiatpäť predikcií sme dostali pri

### 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 vždy správna len v pätnástich prípadoch.

Pre zlepšenie kategorizácií sme zvolili transformácie a metódu vzorkovanie (sampling method). Ako transformácie sme skúšali normalizáciu a pomocné pre prácu so snímkami. Pre vzorkovanie sme skúšali štandardných desať snímok.

Vo výsledku sme dostali lepšie kategorizácie, modely začali kategorizovať do všetkých kategórií a nedostali sme len jednu výslednú kategóriu. Avšak výsledným počtom správnych kategorizácií sme neprišli k rapídnemu 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

## 4.4. Odporúčania

- pracovať s datasetom z jednej oblasti => autori možno chceli univerzálnosť, ale vyšlo to proti nim

- pridať šum do videí

- optimalizačné techniky ako Asymptoticka notacia

- Party dať preč

## 4.4. Vyhodnotenie

# Zoznam bibliografických odkazov

Kortylewski, Adam et al. 2020. “Compositional Convolutional Neural Networks: A Robust and Interpretable Model for Object Recognition under Occlusion”. <http://arxiv.org/abs/2006.15538> (02. december 2023).

Zhang, Bokai et al. 2021. “Surgical Workflow Recognition with 3DCNN for Sleeve Gastrectomy”. *International Journal of Computer Assisted Radiology and Surgery* 16(11): 2029–36.

Qu, Haoxuan et al. 2023. “Recent Advances of Continual Learning in Computer Vision: An Overview”. <http://arxiv.org/abs/2109.11369> (02. december 2023).

Zhou, Longfei, Lin Zhang, a Nicholas Konz. 2022. “Computer Vision Techniques in Manufacturing”. <https://www.techrxiv.org/articles/preprint/Computer_Vision_Techniques_in_Manufacturing/17125652/2> (02. december 2023).

Sharma, Ajay, Ankit Gupta, a Varun Jaiswal. 2021. “Solving Image Processing Critical Problems Using Machine Learning”. V *Machine Learning for Intelligent Multimedia Analytics: Techniques and Applications*, Studies in Big Data, ed. Pardeep Kumar a Amit Kumar Singh. Singapore: Springer, 213–48. <https://doi.org/10.1007/978-981-15-9492-2_11> (02. december 2023).

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.

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

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

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. <https://ieeexplore.ieee.org/document/7966210> (02. december 2023).

Zhengfeng, Huang. 2022. “Accurate Recognition Method of Continuous Sports Action Based on Deep Learning Algorithm”. *Wireless Communications and Mobile Computing* 2022: e3407935.

“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).

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

Mueller, Jonas, a Aditya Thyagarajan. 2016. “Siamese Recurrent Architectures for Learning Sentence Similarity”. *Proceedings of the AAAI Conference on Artificial Intelligence* 30(1). <https://ojs.aaai.org/index.php/AAAI/article/view/10350> (02. december 2023).

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

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.

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).

Mesuga, Reymond, a Brian Bayanay. 2021. *A Deep Transfer Learning Approach to Identifying Glitch Wave-form in Gravitational Wave Data*.

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.

Wang, Limin et al. 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.

Tran, Du et al. 2015. “Learning Spatiotemporal Features with 3D Convolutional Networks”. V *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile: IEEE, 4489–97. <http://ieeexplore.ieee.org/document/7410867/> (02. december 2023).

Wang, Heng, a Cordelia Schmid. 2013. “Action Recognition with Improved Trajectories”. V *2013 IEEE International Conference on Computer Vision*, , 3551–58. <https://ieeexplore.ieee.org/document/6751553> (02. december 2023).

“A Multimodal Pairwise Discrimination Network for Cross-Domain Action Recognition | IEEE Journals & Magazine | IEEE Xplore”. <https://ieeexplore.ieee.org/document/9160941> (02. december 2023).

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

Comoli, P. et al. 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* 16(7): 2106–16.

“MV-C3D: A Spatial Correlated Multi-View 3D Convolutional Neural Networks | IEEE Journals & Magazine | IEEE Xplore”. <https://ieeexplore.ieee.org/document/8736713> (02. december 2023).

Zhang, Zhiwu et al. 2010. “Mixed Linear Model Approach Adapted for Genome-Wide Association Studies”. *Nature Genetics* 42(4): 355–60.

Zhang, Qi et al. 2020. “Spatio-Temporal Graph Structure Learning for Traffic Forecasting”. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(01): 1177–85.

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

“3D genome organization in health and disease: emerging opportunities in cancer translational medicine”. <https://www.tandfonline.com/doi/full/10.1080/19491034.2015.1106676> (02. december 2023).

Wei, Pengbo et al. 2022. “Vision-Based Activity Recognition in Children with Autism-Related Behaviors”. <http://arxiv.org/abs/2208.04206> (02. december 2023).

Lin, Tsung-Yi et al. 2017. “Focal Loss for Dense Object Detection”. V *2017 IEEE International Conference on Computer Vision (ICCV)*, , 2999–3007. <https://ieeexplore.ieee.org/document/8237586> (02. december 2023).

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. <https://ieeexplore.ieee.org/document/8099985> (02. december 2023).

Shi, Zhensheng et al. 2020. “Learning Attention-Enhanced Spatiotemporal Representation for Action Recognition”. *IEEE Access* 8: 16785–94.

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.

Li, Yinqing et al. 2020. “Distinct Subnetworks of the Thalamic Reticular Nucleus”. *Nature* 583(7818): 819–24.

Terney, Daniella et al. 2008. “Increasing Human Brain Excitability by Transcranial High-Frequency Random Noise Stimulation”. *Journal of Neuroscience* 28(52): 14147–55.

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.

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.

Saddam Bekhet, 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* 56(4). <http://jsju.org/index.php/journal/article/view/994> (02. december 2023).

Karpathy, Andrej et al. 2014. “Large-Scale Video Classification with Convolutional Neural Networks”. V *2014 IEEE Conference on Computer Vision and Pattern Recognition*, , 1725–32. <https://ieeexplore.ieee.org/document/6909619> (02. december 2023).

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

Mou, Lichao, Yuansheng Hua, Pu Jin, a Xiao Xiang Zhu. 2020. “ERA: A Dataset and Deep Learning Benchmark for Event Recognition in Aerial Videos”. <http://arxiv.org/abs/2001.11394> (02. december 2023).

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

Ng, Joe Yue-Hei et al. 2015. “Beyond short snippets: Deep networks for video classification”. V *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, , 4694–4702. <https://ieeexplore.ieee.org/document/7299101> (03. december 2023).

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.

# 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://lcmou.github.io/ERA\_Dataset/ [↑](#footnote-ref-3)
4. https://elib.dlr.de/140908/1/2001.11394.pdf [↑](#footnote-ref-4)
5. <http://www.stat.ucla.edu/~tianmin.shu/AerialVideo/AerialVideo.html> [↑](#footnote-ref-5)
6. http://okutama-action.org [↑](#footnote-ref-6)
7. https://github.com/ckyrkou/AIDER [↑](#footnote-ref-7)
8. <https://lcmou.github.io/ERA_Dataset/> [↑](#footnote-ref-8)
9. https://github.com/batsysk/DiplomovaPraca [↑](#footnote-ref-9)
10. https://cs.stanford.edu/people/karpathy/deepvideo/ [↑](#footnote-ref-10)
11. https://developer.qualcomm.com/software/ai-datasets/something-something [↑](#footnote-ref-11)
12. http://moments.csail.mit.edu [↑](#footnote-ref-12)