# ZADANIE

Rozpoznávanie objektov vo videách

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

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

Charakter práce:

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

Predmetové prerekvizity:

Úvod do strojového učenia (1., mgr);

Neurónové siete (1., mgr);

Hĺbková analýza dát (2., mgr).

Najdôležitejšie kompetentnosti získané spracovaním témy:

vykonávať vedecký výskum;

princípy umelej inteligencie;

poskytnúť vizuálnu prezentáciu údajov;

referovať o výsledkoch analýzy;

vykonať analýzu údajov.

# Úvod

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

# Popis riešeného problému / Analýza súčasného stavu

Súčasná doba digitálneho pokroku prináša interakcie medzi človekom a strojom na vyššej úrovni. Jedným z najúžasnejším aspektom tejto postupnej revolúcie je oblasť umelého videnia, taktiež známa ako computer vision. Táto disciplína umelej inteligencie sa stala nepostrádateľnou súčasťou nášho každodenného života a to najmä v oblasti spracovania obrazov a analýzou videí. 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 a to každý deň. To prinieslo nový dopyt pre modely aby sa učili a vylepšovali postupne samé, aby sa nemuseli vytvárať nové od základu. Na rozdiel od neurónových sietí, ktoré nezvládajú jednoduchý prechod na učenie sa postupne v čase, počítačové videnie, ktoré to zvláda, ide do popredia. (<https://arxiv.org/pdf/2109.11369.pdf>)

Počítačové videnie spracováva jednoduché popisy: čo vidí na scéne. Popisy obsahujú pochopenie scény a tie sa využívajú v modeloch. Algoritmy modelov vieme rozdeliť do kategórií: detekcia, rozpoznanie, segmentácia a 3D rekonštrukcia. (<https://www.techrxiv.org/articles/preprint/Computer_Vision_Techniques_in_Manufacturing/17125652>)

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(<https://people.csail.mit.edu/brooks/idocs/AIM-100.pdf>) a to v roku 1966. Počítačové videnie napreduje hlavne vďaka posunu v technológiách, ktoré umožňujú vyššiu výpočtovú náročnosť. (<https://link.springer.com/chapter/10.1007/978-981-15-9492-2_11>)

Rozvoj počítačového videnia v posledných rokoch 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. (<https://www.hindawi.com/journals/mpe/2022/9422584/>)

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. (<http://revistas.uni.edu.pe/index.php/tecnia/article/view/801>)

Taktiež sa zvyšuje výskumná oblasť

## 1.2 Základné metody computer vision

Pri počítačovom videní vo videách je dôležité rozoznávať kontext, čiže akúkoľvek informáciu, ktorá sa nielen týka vzhľadu cieľového objektu alebo udalosti samotnej, ale zahŕňa aj iné objekty alebo udalosti v scéne a to vizuálne alebo nevizuálne. (<https://www.sciencedirect.com/science/article/pii/S1077314223000267?via%3Dihub>)

### 1.2.1 Rozdelenie

Podľa Xuan Wanga a Zhigang Zhua je práve dôležité deliť počítačové videnie podľa kontextu a ten na priestorový, časový a iný.

Potom sa vieme pozrieť na využitie kontextu na rôznych úrovniach: úroveň predchádzajúcich znalostí, úroveň globálnych znalostí a úroveň lokálnych znalostí.

Následne vieme deliť jednotlivé modely podľa týchto kategórií.

(<https://www.sciencedirect.com/science/article/pii/S1077314223000267?via%3Dihub>) OBRÁZOK

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

Automaticky generovaný popis

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

Automaticky generovaný popis

### 1.2.2 Porovnanie medzi modelmi

## 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.

### 1.3.1 Hĺbkové učenie v kontexte modelov pre klasifikáciu videí

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 je rozpoznávane akcií vo videách. (<https://ieeexplore.ieee.org/document/7966210>)

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.

(<https://www.hindawi.com/journals/wcmc/2022/3407935/>)

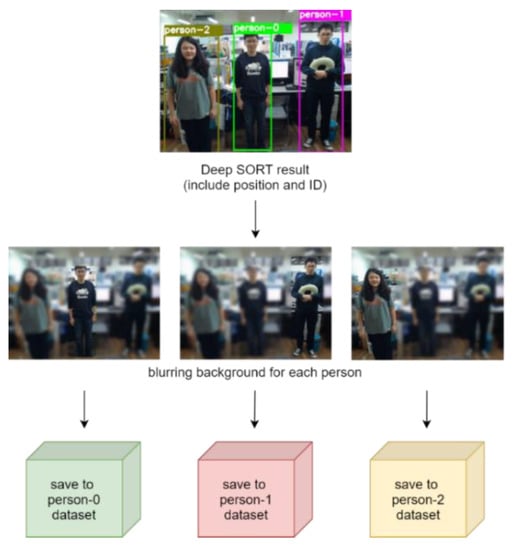
Obrázok, na ktorom je text, snímka obrazovky, fyzická kondícia, tanec

Automaticky generovaný popis

(<https://www.sciencedirect.com/science/article/pii/S1077314223000267?via%3Dihub>)

Výkony basketbalistov napríklad skúmal Peng Yao, ktorý vo svojom výskume popisuje jednoznačné využitie do budúcnosti na personifikáciu stratégie pre jednotlivých hráčov. Vďaka tomuto poznatku vie 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.

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 jeho kolegovia riešia túto problematiku a výsledok ich výskumu popisujú v článku, kde vidia využitie modelu I3D. (<https://www.mdpi.com/1424-8220/20/17/4758>)



RNN(Recurrent Neural Network) využívajú výskumníci napríklad v detekovaní 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 siamese based RNN popri modelu I3D a na dvoch datestoch 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 duplikovaných snímok. (<https://www.techscience.com/iasc/v29n3/43054/html>)

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. (<https://ojs.aaai.org/index.php/AAAI/article/view/10350>)

### 1.3.2 SVM a klasifikácia vo videách

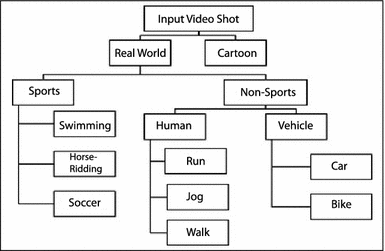
SVM alebo Support Vector Machine je algoritmus strojového učenia, ktorý sa často využíva pre klasifikačné a regresné úlohy. Pri videách sa momentálne skúma využitie SVM na kategorizáciu druhov videí a to napríklad na reklamy, kreslené rozprávky, hudobné videá, správy a športové videá (podľa výskumu Zhang Min-qinga a Li Wen-pinga). 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). (<https://www.hindawi.com/journals/sp/2021/4728584/>)

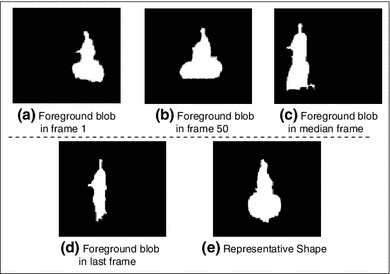
Obrázok, na ktorom je diagram, rad, náčrt, kresba

Automaticky generovaný popis

Podobne využili SVM aj Chattopadhyay a Maurya aby preskúmali možnosť rýchleho označovania videí do kategórií pomocou kľúčových slov (tagov). (OBRÁZOK) (<https://link.springer.com/article/10.1007/s13735-013-0034-8>)



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. (<https://link.springer.com/article/10.1007/s13735-013-0034-8>)



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. (<https://link.springer.com/article/10.1007/s13735-013-0034-8>)

### 1.3.3 3D Konvolunčné modely

CNN 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. (<https://arxiv.org/pdf/1803.02247.pdf>)

Z obrázku 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.

(<https://www.researchgate.net/publication/353049916_A_Deep_Transfer_Learning_Approach_to_Identifying_Glitch_Wave-form_in_Gravitational_Wave_Data>)

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

Automaticky generovaný popis

### 1.3.4 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)[[1]](#footnote-1) popisujú využitie na modeloch, ktoré vybrali.

#### 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. (<https://ieeexplore.ieee.org/document/8985265>)

OBRAZOK

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

Automaticky generovaný popis

(<https://openaccess.thecvf.com/content_iccv_2015/papers/Tran_Learning_Spatiotemporal_Features_ICCV_2015_paper.pdf>)

Vďaka stratégii, kedy nespracováva každý snímok, ale dané číslo snímok v intervale, sa zvyšuje jeho efektivita a znižuje výpočtová náročnosť. (<https://link.springer.com/chapter/10.1007/978-3-319-46484-8_2>)

Podľa výskumu, ktorý vykonal Du Tran a jeho kolegovia na učení 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. (https://openaccess.thecvf.com/content\_iccv\_2015/papers/Tran\_Learning\_Spatiotemporal\_Features\_ICCV\_2015\_paper.pdf)

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.

OBRAZOK

Obrázok, na ktorom je text, snímka obrazovky

Automaticky generovaný popis

(https://arxiv.org/pdf/1906.02182.pdf)

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 (<https://ieeexplore.ieee.org/document/6751553/>). Shang taktiež popisuje zlepšenie výkonov s použitím modelu ActionVLAD (<https://ieeexplore.ieee.org/document/9160941>).

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. (<https://www.hindawi.com/journals/cin/2016/6750459/>)

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. (<https://www.hindawi.com/journals/cin/2016/6750459/>)

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.

(<https://ieeexplore.ieee.org/document/7410867>)

V oblasti medicínskeho výskumu sa C3D využíva pri pacientoch, ktorým bola transplantovaná oblička. Podľa výskumu (<https://linkinghub.elsevier.com/retrieve/pii/S1600613522009261>) by mal byť tento model schopný predikovať rizika odmietnutie 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.

(<https://ieeexplore.ieee.org/document/8736713>)

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.

(<https://ieeexplore.ieee.org/document/8736713>)

#### 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. (<https://www.nature.com/articles/ng.546>)

P3D model pol porovnávaný s C3D a predstavuje menšiu výpočtovú náročnosť a predstavuje menšiu obtiažnosť pri tréningu a nižšiu možnosť preučenia.

(<https://ojs.aaai.org/index.php/AAAI/article/view/5470>)

P3D 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. (<https://www.sciencedirect.com/science/article/abs/pii/S0378775319307116?via%3Dihub)>

OBRÁZOK

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.

(<https://www.tandfonline.com/doi/full/10.1080/19491034.2015.1106676>)

V medicíne má P3D model úžitok pri modelovaní geometrie genómov, kde sa pozorujú bunky a ich dopad na zdravie ľudí. (<https://www.tandfonline.com/doi/full/10.1080/19491034.2015.1106676>)

#### 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. (<https://arxiv.org/pdf/2208.04206.pdf>)

V oblasti rozpoznávania aktivít vo videách sa tento model využíva v chirurgii, kde sa rozpoznávajú postupy pri práci. V štúdii (<https://link.springer.com/article/10.1007/s11548-021-02473-3>) bol model I3D, konkrétne architektúra Inflated 3D ConvNet (na základe ktorej je vybudovaný model I3D) (<https://ieeexplore.ieee.org/document/8099985>), využitá na rozpoznávanie chirurgických postupov v rámci operácií Tubulizácie žalúdka. Model bol trénovaný pomocou metódy Focal loss a dosiahol presné rozpoznávanie rôznych chirurgických krokov.

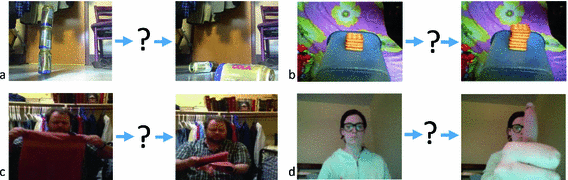
Focal Loss je špeciálny druh loss funkcie, ktorá sa zaoberá problémom nerovnováhy tried pri úlohách detekcie objektov. Keďže pri tejto problematike väčšina snímok pozostáva z pozadia, zatiaľ čo pre nás je dôležitá detekcia objektu záujmu, môže nerovnováha spôsobovať malú úspešnosť detekovania objektov 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. (<https://ieeexplore.ieee.org/document/8237586>)

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. (<https://ieeexplore.ieee.org/document/8963915>)

I3D model sa využíva pri bezpečnostných aplikáciách v reálnom svete. Napríklad v autonómnych autách a pri detekcii osôb na bezpečnostných záznamoch. (<https://arxiv.org/abs/2006.15538>)

#### 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.

(https://link.springer.com/chapter/10.1007/978-3-030-01246-5\_49) 

Veľkou výhodou TRN modelu oproti iným je jeho ľahšie pochopenie pre vyvodenie výsledkov. (<https://link.springer.com/chapter/10.1007/978-3-030-01246-5_49>)

Vo výskumoch neurovedy bol tento model skúmaný pri regulácii senzorického spracovania, pozornosti a kognície. (<https://www.nature.com/articles/s41586-020-2504-5>) 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.

(<https://www.jneurosci.org/content/28/52/14147>)

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. (<https://www.sciencedirect.com/science/article/pii/S0378778816313937?via%3Dihub>)

### 1.3.5 Ďalšie modely

TODO

Obrázok, na ktorom je text, kompaktný disk, kruh, zariadenie na ukladanie údajov

Automaticky generovaný popis

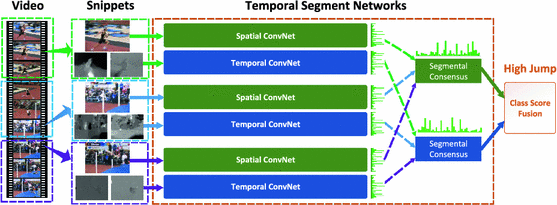
#### TSN model

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 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.

(<https://link.springer.com/chapter/10.1007/978-3-319-46484-8_2>)

Využite sa momentálne najviac priraďuje k zlepšeniu výkonov pri rozpoznávaní aktivít vo videách. (<https://link.springer.com/chapter/10.1007/978-3-319-46484-8_2>)



#### 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

(<https://www.isca-speech.org/archive/pdfs/interspeech_2016/chen16n_interspeech.pdf>)

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. (<https://link.springer.com/chapter/10.1007/978-3-319-46478-7_47>)

## 1.4. Porovnanie modelov

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

Automaticky generovaný popis

Tabuľka popisuje modely, ktoré využívame v tejto práci. Jej stĺpce sú:

* 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) context 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 vypracovať prehľad modelov na tému klasifikácie vo videách. Tieto poznatky neskôr využijeme v praxi, kde porovnáme modely na datasete ERA.

Naše čiastkové ciele sú:

* Overiť, že modely, ktoré boli použité už na datasete ERA majú v našej práci podobné výsledky.
* Ak by sa výsledky priveľmi odlišovali, tak chceme prísť na možné dôvody.
* 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. Pridržiavali sme sa metodiky CRISP-DM, ktorú sme si upravili pre naše potreby.

## 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 je užitočná. Využitie sme popísali v kapitole číslo jeden.

### 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. (<http://jsju.org/index.php/journal/article/view/994>) 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. (<https://ieeexplore.ieee.org/document/6909619>) Na obrázku môžeme vidieť, že sa videá musia brať ako viac dimenzionálne média, pretože v nich ide o kontext v čase. (<https://books.google.sk/books?hl=sk&lr=&id=wXmSPPB_c_0C&oi=fnd&pg=PP1&dq=video+processing&ots=UTTsyHktdL&sig=4d68oqpxfK3aolkMjUVwRhvJe0E&redir_esc=y#v=onepage&q=video%20processing&f=false>)

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

Automaticky generovaný popis

## 3.2. Výber datasetu

Pri výbere datasetu bolo pre nás kľúčové myslieť na tieto parametre:

- Aby dataset mal dostatočný počet samostatných videí a tie boli rôznorodé.

- Aby sme nemuseli zostavovať dataset manuálne, čiže zbierať videá po

internete.

- Aby sme mohli využiť dataset v našej práci a nevzťahovali sa na videá autorské

práva.

Za týchto podmienok sme si vybrali voľne dostupný dataset ERA, čiže dataset na rozpoznávanie udalostí v leteckých videách.

### 3.2.1 Dataset ERA

Ako popísal Mou, Hua, Jin a Zhu v ich práci: Dataset ERA[[2]](#footnote-2) (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. (<https://arxiv.org/abs/2001.11394>)

### 3.2.2 Využitie datasetu ERA

## 3.3. Porozumenie dátam v datasete

Našou ďalšou úlohou bolo porozumieť aké dáta sú v datasete a aký majú formát.

## 3.4. Spracovanie dát

## 3.5. Výber modelov

Pri modeloch sme si spravili prieskum aké sa využívajú v dnešnej dobe plus sme využili modely, ktoré aj samotní autori datasetu používali na modelovanie.

## 3.6. Analýza dát

## 3.7. Evalvácia výsledkov

## 3.8. Vyhodnotenie a interpetácia

# Zoznam bibliografických odkazov

1. <https://lcmou.github.io/ERA_Dataset/> [↑](#footnote-ref-1)
2. https://lcmou.github.io/ERA\_Dataset/ [↑](#footnote-ref-2)