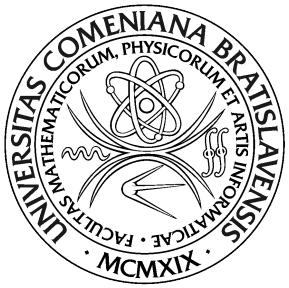
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

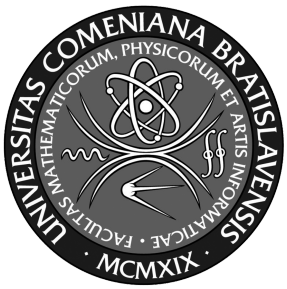


SEGMENTÁCIA OBJEKTOV VO VIDEU, ROZOSTRENÝCH RÝCHLYM POHYBOM

Diplomová práca

2022 Bc. Tomáš Maťko

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



SEGMENTÁCIA OBJEKTOV VO VIDEU, ROZOSTRENÝCH RÝCHLYM POHYBOM

Diplomová práca

Študijný program:	Aplikovaná informatika
Študijný odbor:	2511 Aplikovaná informatika
Školiace pracovisko:	Katedra aplikovanej informatiky
Školiteľ:	RNDr. Zuzana Černeková, PhD.

Bratislava, 2022 Bc. Tomáš Maťko

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

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

Názov: Segmentácia objektov vo videu, rozostrených rýchlym pohybom
Segmenting motion blurred objects in video

Anotácia: Naštudovať problematiku segmentovania objektov. Oboznámiť sa s metódami a aplikáciami postprocessingu videa. Analyzovať existujúce riešenia publikované v dostupnej odbornej literatúre. Navrhnúť metódu, ktorá umožní vysegmentovanie objektu s rozostrenými hranami spôsobené rýchlym pohybom. Vyhodnotiť dosiahnuté výsledky.

Cieľ: Naštudovať problematiku segmentovania objektov. Oboznámiť sa s metódami a aplikáciami postprocessingu videa. Analyzovať existujúce riešenia publikované v dostupnej odbornej literatúre. Navrhnúť metódu, ktorá umožní vysegmentovanie objektu s rozostrenými hranami spôsobené rýchlym pohybom. Vyhodnotiť dosiahnuté výsledky.

Vedúci: RNDr. Zuzana Černeková, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: prof. Ing. Igor Farkaš, Dr.

Dátum zadania: 23.09.2019

Dátum schválenia: 23.09.2019 prof. RNDr. Roman Ďurikovič, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Čestne prehlasujem, že túto diplomovú prácu som vypracoval samostatne len s použitím uvedenej literatúry a za pomoci konzultácií u môjho školiteľa.

. . . . . .

Bratislava, 2022 Bc. Tomáš Maťko

Poďakovanie

Touto cestou by som sa chcel v prvom rade poďakovať mojej školiteľke RNDr. Zuzane Černekovej, PhD. za jej cenné rady a usmernenia, ktoré mi veľmi pomohli pri riešení tejto diplomovej práce. Pokračovanie ——————————
———————————-

Abstrakt

Naštudovať problematiku segmentovania objektov. Oboznámiť sa s metódami a aplikáciami postprocessingu videa. Analyzovať existujúce riešenia publikované v dostupnej odbornej literatúre. Navrhnúť metódu, ktorá umožní vysegmentovanie objektu s rozostrenými hranami spôsobené rýchlym pohybom. Vyhodnotiť dosiahnuté výsledky.

Kľúčové slová: pohyb,kamera,segmentácia,video,objekt,rozostrenie

Abstract

Study the tracking and segmentation of objects. Learn about actual methods and applications of post-processing of videos. Analyze actual results and study state of the art of the task. To implement a method for detecting a object with motion blur with aim to precision. Show the results.

Keywords: movement,camera,blur,video, motion, segmentation

Obsah

Kapitola 1

Úvod

V modernom svete, v ktorom žijeme, sa vývoj spotrebiteľskej techniky natoľko urýchlil, že sa často každodenne spoliehame na technológie ktoré donedávna ešte neexistovali. Každým dňom je vyrábané enormné množstvo technických zariadení a takmer každý človek na svete vlastní nejakú formu mobilného zariadenia. Na týchto zariadeniach vzniká množstvo fotografii a videí, s ktorými pracujeme, editujeme, zdieľame na socialných sieťach alebo si len chceme vytvoriť spomienku na daný moment. My, ľudia, chápeme akú pointu má každé z týchto videí. Vidíme na nich objekty ktoré poznáme, vieme kam sa dané objekty hýbu, vieme akú činnosť vykonávajú, chápeme ich kontext. Tento proces ktorý ani nevnímame, pretože si nevyžaduje námahu a je nám prirodzený od detského veku, je ale marginálne zložitejší pre počítače. Pre počítače je ťažké pochopiť kontext videa, pretože už len oddelenie objektu záujmu od pozadia je algoritmicky náročná činnosť. A ak je video náhodou rozostrené, pre počítač ide o ďalšiu komplikáciu. Ak chceme aby si počítač vedel odvodiť aký je kontext daného videa, ako prvé je nutné dosiahnuť aby vedel rozpoznať jednotlivé objekty ktoré sú vo videu. Tento problém má ale aj iné súvislosti. Ak by sme sa snažili zdokonaliť určitý algoritmus

na ovládanie auta ktorý funguje na základe senzorov a kamier, je nutné aby počítač vedel správne klasifikovať všetky objekty (napríklad zviera na vozovke). Ak počítač nevie správne objekt segmentovať, nemusí ho ani správne klasifikovať, a presná segmentácia je obzvlášť náročná ak je video rozostrené, napríklad rýchlym pohybom auta snímajúceho vozovku pred ním. V mojej práci sa preto budem snažiť vyriešiť problém segmentácie objektov z videa ktoré sú rozostrené rýchlym pohybom.

Súčasťou tejto diplomovej práce je aj prehľad rôznych prístupov zaoberajúcich sa touto problematikou a analýza ich kladov a záporov. Výsledky analýzy dalších riešení nam môže pomocť pri vytváraní našej implementácie.

Kapitola 2

Prehľad problematiky

V mojej práci sa bude usilovať o segmentáciu objektu vo videu, ktoré je rozostrené rýchlym pohybom. Takúto úlohu je teda možné rozdeliť na dve časti:

- segmentácia objektu vo videu 2.2
- detekcia rozmazaných častí videá v okolí objektu záujmu 2.3

V mojej práci sa v súvislosti so segmentáciou objektov budem vo veľkej miere odkazovať na problematiku neurónových sietí, preto by som rád začal práve pri nich.

2.1 Neurónové siete

Neurónové siete sú čoraz viac využívané pre riešenie najrozličnejších typov úloh. Ich názov je odvodený kvôli ich podobnosti ku biologickým neurónom, vyskytujúcich sa v mozgu, ktorými bol vznik neurónových sietí sčasti inšpirovaný. V mozgu je rozsiahla sústava neurónov, ktoré prijímajú vstupné

impulzy a spracovávajú ich do výsledného impulzu ktorý je posunutý nasledovným neurónom. V našom kontexte je každý neurón zastúpený istou matematickou operáciou s vstupnými dátami ktorých výstup je vstupom do ďalších neurónov. Neuróny so spoločnou charakteristikou spájame do vrstiev, ktorých môžeme mať ľubovoľné potrebné množstvo, minimálne však dve. Takto môžeme jednotlivé neurónové vrstvy reťaziť ako stavebné bloky, čím vytvárame neurónovú sieť. To ako sú jednotlivé neuróny prepojené, koľko neurónov používame určuje topológia neurónovej siete, ktorá býva pre každé použitie rozličná. [Hay09]

Prvá vrstva neúronovej siete zvyčajne príjme vstupné dáta ktoré vieme interpretovať avšak následne jednotlivé úrovne neurónov môžu generovať dáta ktoré by pre nás nedávali zmysel. Tieto vrstvy sa volajú taktiež aj skryté vrstvy, keďže nezasahujú do vonkajšieho prostredia. V našej práci sa budeme zaoberať konvolučnými neurónovými sieťami, ktoré majú možnosť naučiť sa vlastnosti objektov a na ich základe vyriešiť zamýšľaný problém. Našim cieľom je aby posledná vrstva vygenerovala dáta očakávaného typu, ktoré vieme interpretovať, a to aj pre vstupné dáta bez priloženého očakávaného výstupu.

2.1.1 Trénovanie Neurónovej siete

Toto dosiahneme učením neurónovej siete. Ide o proces kedy dávame sieti k dispozícii veľké množstvo vstupných dáta za cieľom aby sme sieť natrénovali. Každá vrstva má určitý počet parametrov ktoré nazývame váhami, ktoré sa budú meniť počas trénovania, pričom sa budeme snažiť minimalizovať stratovú funkciu (loss function). Táto hodnotí celkovú úspešnosť siete, a je založená na porovnávaní rozdielu medzi požadovaným výstupom a dosiahnutým výstupom. Váhy sa budú aktualizovať na základe výpočtu gradientu, vďaka ktorému vieme zaistiť že budeme minimalizovali stratovú funkciu, a

teda aby sa naše výstupy líšili čo najmenej od požadovaných výstupov. Trénovanie neurónovej siete je kľúčovým štádiom pre jej správne fungovanie, a pri procese trénovania je nutné zvoliť správne hyperparametre, teda premenné ktoré majú veľký vplyv na funkčnosť siete.Na úspešné natrénovanie siete je potrebné velké množstvo dát, a teda je dôležitý výber správneho datasetu, na ktorom budeme sieť trénovať. Ďalšia možnosť je použiť model neurónovej siete ktorá je už predtrénovaná alebo takýto model mierne prispôsobiť na požadovaný účel.

2.2 Segmentácia objektu vo videu

Pre človeka je jednoduché na videu určiť ktorá časť videa zachytáva postavu. Počítač však vidí iba jednotlivé pixely, reprezentované číslami, na slede obrázkov a nevie ktoré pixely prislúchajú hľadanému objektu - postave. Segmentácia objektu je problém, kedy sa snažíme na základe vzťahov medzi pixelmi určiť kde je hranica objektu ktorý sledujeme, a teda presne vymedziť ktoré pixely patria ku popredíu alebo pozadiu daného objektu. Výstup segmentácie objektu je teda v ideálnom prípade maska objektu, ktorá určuje pre každý pixel, či patrí danému objektu alebo nie. Témou segmentácie objektu sa zaoberá odbor počítačového videnia pomerne dlhý čas, ide o jeden z fundamentálnych konceptov počítačového videnia, a aj vďaka tomu existuje mnoho spôsobov ktorými riešiť tento problém. Najnovšie algoritmy v tejto oblasti sa veľkými krokmi približujú k takej segmentácii objektu akú by sme intuitívne vedeli určiť aj my ľudia.

Práce zaoberajúce sa segmentáciou objektu môžeme rozdeliť na tie ktoré sú založené na princípe konvolučných neurónových sietí a na iné riešenia. V mojej práci sa budem zaoberať prístupmi ktoré sú postavené na metóde kon-

volučných neurónových sietí teda napríklad práce ako Osvos [CMPT+17], CFBI+ [YWY21], či Monet [XFL+18]. Neurónové siete sa čoraz častejšie uplatňujú vo viacerých oblastiach počítačového videnia. Zvyčajne dosahujú rádovo lepšie výsledky pri porovnaní s inými prístupmi. Medzi prístupy nepoužívajúce neurónové siete patria napríklad JumpCut [FZL+15] alebo Fusionseg [JXG17]. Tieto metódy sú založené na farebnom rozložení pixelov, na základe optického toku, detekcii pohybu, grafových modeloch či napríklad lineárnych klasifikátoroch. V článku [WZP+21] autori porovnávajú do hĺbky rozdiely vo výsledkoch medzi „tradičnými prístupmi“ a prístupmi s použitím konvolučných neurónových sietí (ďalej CNN - Convolutional neural network).

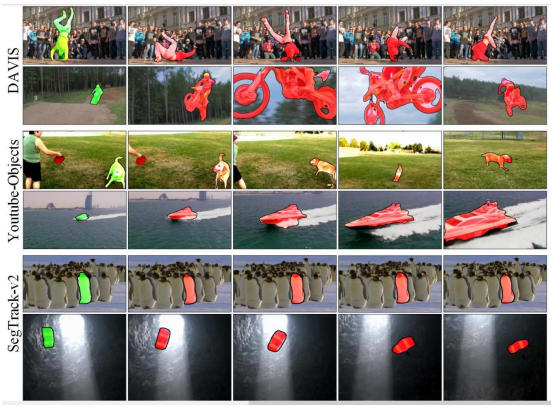
2.2.1 Rozdelenie segmentačných prác

V súvislosti s riešeniami na princípe cnn je taktiež viacero prístupov. Mohli by sme ich rozdeliť do dvoch hlavných podkategórii :

- Automatické práce (Unsupervised)
- Polo-automatické práce (Semi-supervised)

Hlavným rozdielom medzi týmito prístupmi je eliminácia počiatočného vstupu užívateľa pri automatických prácach. Tieto práce dostanú určité video-dáta bez iných anotácii a snažia sa vysegmentovať objekt(y) záujmu. Narozdiel od nich polo-automatické práce potrebujú nielen samotné video-dáta ale aj vysegmentovaný objekt záujmu, zvyčajne prvého snímku videa. Z tohto dôvodu je nutné pri snahe o segmentovanie neanotovaného videa túto masku pre prvý snímok získať segmentáciou pomocou inej formy. Objekt na nasledujúcich snímkach videa je segmentovaný aj na základe tejto prvej segmentácie, preto je nutné aby táto počiatočná segmentácia bola čo najpresnejšia. Jednou možnosťou je túto snímku manuálne segmentovať, čo je časovo-náročný

prístup, ktorý ale môže zlepšiť presnosť celkovej segmentácie. Zmysluplnejšie riešenie je ale použiť na túto počiatočnú segmentáciu existujúce riešenie samotného obrázku. O tomto probléme budem viac písať v sekcii 3



Obr. 2.1: Ilustrácia segmentácie objektov na rôznych videách z rôznych datasetov. Ground truth maska objektu je zaznačená na prvom snímku zelenou farbou, červenou farbou je znázornená segmentovaná maska objektov. Obrázok je prevzaný z [XFL+18].

V kontexte počítačového videnia sa môžeme stretnuť aj s tzv. offline a online učením. Offline učenie spočíva v získaní datasetu z ktorého sa bude naša sieť po malých dávkach učiť až do bodu, kedy sa jej účinnosť viac nezvyšuje. Online učenie môžeme taktiež inicializovať na datasete, avšak líši sa že takto navrhnutá sieť sa naďalej učí na všetkých dátach ktoré dáme na vstup tejto siete. Online učenie zvyčajne zvyšuje presnosť, no je výpočtovo náročnejšie, keďže sa sieť s každým prechodom neustále snaží učiť. Offline metódy sa preto snažia vyhnúť online učeniu pri zachovaní presnosti.

V našej práci sa budeme zaoberať primárne polo-automatickými prístupmi a tie ďalej môžeme rozdeliť podľa spôsobu samotného prístupu ku segmentácii:

- Metódy založené na propagácii

- spočívajú v trénovaní tzv. propagátora objektovej masky (object mask propagator). Ide o neurónovú sieť ktorá sa snaží napasovať nezrovnanú masku voči objektu záujmu. Aby bola sieť špecifická pre daný objekt, používa prvú snímku na generáciu online trénovacích dát pomocou deformácie masky objektu alebo spojením snímok na jemné doladenie. Teda neustále propaguje pôvodnú masku na dalšie snímky. Niektoré z týchto prác taktiež zahŕňajú do systému modul opätovnej identifikácie na získanie chýbajúcich objektov v dôsledku posunov. Ilustruje ich písmeno a) na obrázku 2.2.

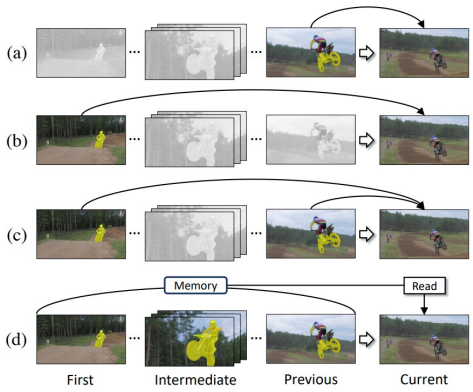
- Metódy založené na detekcii
  - fungujú tak, že učia objektový detektor využívať vzhľad objektu na prvej snímke, pričom sa na segmentáciu cieľového objektu použije špecifický objektový detektor, ktorý bol natrénovaný a jemné vyladený na trénovacej množine. Zjednodušene sa teda detektor snaží segmentovať objekt na základe podobnosti voči počiatočnému vzhľadu objektu. Obmena tohto prístupu je že sú pixely zo snímky vložené do príznakového priestoru a klasifikované podľa priradenia do šablón. Ilustruje ich písmeno b) na obrázku 2.2.

- Hybridné metódy
  - sú navrhnuté tak, aby využívali výhody detekčných aj propagačných prístupov. V niektorých prácach s týmto prístupom boli navrhnuté siete, ktoré využívajú vizuálne vedenie z prvej snímky aj priestorové stopy z predchádzajúcej snímky. Okrem toho sa niektoré metódy pokúsili využiť všetky predchádzajúce informácie. Napríklad bola snaha o sekvenčnú sieť využívajúcu myšlienku online adaptácie a priebežného aktualizovania detektora pomocou priebežných výstupov. Ilustruje ich

písmeno c) na obrázku 2.2.

- Metódy využívajúce prídavnú pamäť

  Nedávno vznikli práce ktoré uchovávajú priebežné výstupy v externej pamäti namiesto toho aby pomocou nich trénovali sieť, a adaptívne vyberajú potrebné informácie počas behu. Vďaka tomuto flexibilnému prístupu táto metóda prekonáva s rezervou vyššie uvedené metódy. Tento prístup je rýchly, keďže čítanie pamäte sa vykonáva ako súčasť dopredného chodu siete, a preto nie je potrebné žiadne online učenie. Zároveň sú ale využité predchádzajúce segmentácie na zvýšenie presnosti. Ilustruje ich písmeno d) na obrázku 2.2.



Obr. 2.2: Rozličné typy segmentácie pomocou cnn. a) Metódy využívajúce propagáciu, b) Metódy využívané detekciu, c) Hybridné metódy, d) Metódy využívajúce prídavnú pamäť. Obrázok je prevzaný z [OLXK19]

2.3 Detekcia rozmazaných častí v okolí objektu záujmu

Rozmazanie videa je komplikácia ktorá môže segmentáciu objektu značne stažiť. Čím viac je obraz rozmazaný, tým tažšie je detekovať alebo segmento-

vať sledovaný objekt. Väčšina prístupov ku segmentácii objektov z videa sa zameriava hlavne na ostré videá bez žiadnych typicky vyskytujúcich sa negatívnych vplyvov, ako napríklad rozmazanie. Poznáme dva hlavné dôvody rozmazania fotografie či videí.

1. pohyb kamery alebo objektu

2. zlé zaostrenie, zlá hĺbka ostrosti alebo iné parametre kamery.

Nás bude zaujímať hlavne prvý bod a teda rozmazanie ktoré je spôsobené pohybom kamery alebo objektu. V praxi je zvyčajne jednoduché ich rozoznať nakoľko ak sa hýbe objekt tak je zväčša ostrý zvyšok snímanej plochy, avšak, ak sa hýbe kamera, je rozmazaná celá snímaná plocha. Takýto efekt ľahko vznikne ak sa snažíme zachytiť rýchlo pohybujúci objekt alebo s kamerou pohneme v čase keď sníma. V oboch prípadoch je problém nedostatočná rýchlosť uzávierky fotoaparátu ktorá vo výsledku spôsobí deformáciu obrazu. Existujú riešenia ktoré sa snažia problém takejto deformácie opraviť, no najprv je potrebné dané rozmazanie na videu vôbec detegovať. Je viacero spôsobov akými vieme riešiť rozmazanie vo videu, je ich ale oveľa viac ako riešiť rozmazanie iba na jednom obrázku. Video má vlastnosť že vieme využiť aj snímky ktoré sú pred alebo po aktuálnom snímku, na ktorom sa snažíme segmentovať objekt. Väčšina rozmazaní je prítomná iba na niekoľkých snímkoch sekvencie, a preto ak vieme priradiť östré"pixely objektu z susedných snímkov, môžeme zrekonštruovať pozíciu a tvar objektu na rozmazaných snímkoch. Na tomto fakte sa zakladá množstvo riešení ktoré vezmú vzorku viacerých snímkov, a snažia sa z viacerých rozmazaných snímkov upraviť rozmazanie aktuálneho snímku.[JZW+20]. Pohybujúce sa objekty alebo pohybujúca sa kamera sú jedny z najťažších typov rozmazania, v oblasti obnovy ostrého snímku, pretože vplyv rozmazania sa mení vrámci celého

snímku. Takéto rozmazania môže byť globálne (teda na celom obraze), alebo lokálne (teda na určitej časti obrazu), radiálne alebo lineárne (teda sústredeného okolo bodu alebo jednosmerné), uniformné alebo neuniformné (teda v rovnakej či rôznej intenzite v pozícii rozmazania).

2.3.1 Riešenia založených na neurónových sieťach

Aj v tematike problému rozmazania rýchlym pohybom sú v tomto čase najviac rozšírené riešenia ktoré využívajú neurónové siete. V prvých prácach sa trénovanie CNN zameriavalo najmä na identifikáciu jadra rozmazania a jeho lokalizovanie po jednotlivých pixeloch. Spočiatku takéto riešenie pracovalo na princípe prikladania vzorov rozmazania na obrázky, a ďalší vývoj spočíval v nahradení tejto myšlienky na tradične plne konvolučné riešenie ktorým bol prehľadávaný obraz. Taktiež boli snahy o využitie optického toku na odhad smeru rozmazania. Veľa z metód ktoré využili cnn ich však používali iba na lokalizovanie jadra rozmazania a určenie smeru rozmazania, nie na samotné vyostrenie obrazu. Aj to sa zmenilo príchodom GAN sietí (Generative Adversial Networks).Tieto siete fungujú na princípe dvoch častí - generátora a diskriminátora. Počas trénovania sa generátor snaží generovať vizuálne vierohodný ostrý obraz, ktorý následne predá diskriminátoru, ktorého úlohou je naopak rozoznať či ide o generovaný obraz alebo autentický ostrý obraz. Týmto spôsobom sa navzájom snažia súťažiť do bodu kedy diskriminátor bude schopný vytvárať obraz ktorý je na nerozoznanie od pravého. Autori práce [GR19] ktorá ma ale zaujala pre účely mojej diplomovej práce najviac je na rozdiel od posledných zmienených prác relatívne jednoduchá a napriek tomu dosahuje porovnateľné výsledky. Viac o nej budem písať v 3.3

2.4 Datasety

Keďže moje riešenie bude založené na konvolučných neurónových sieťach, je kľúčové aby sme mali dostatok dát z dôvodov trénovania a vyhodnocovania práce. Preto popíšem populárne existujúce datasety pre oblasť objektovej segmentácie, ktoré budú spomínané v ďalších častiach práce. Je podstatné spomenúť že väčšina datasetov je určená predovšetkým na účely klasifikácie a získavanie ground-truth masky objektov je pomerne náročné, preto niektoré z nasledujúcich datasetov sú vhodné iba na tréning, kým iné vieme použiť aj na vyhodnocovaniu.

2.4.1 ImageNet

ImageNet je jeden z najväčších a najpopulárnejších datasetov ktoré sú využívané v oblasti počítačového videnia. Vznikol v roku 2009 a neustále sa rozrastá, už v roku 2019 tento dataset obsahoval viac než 14 miliónov obrázkov, ktoré sú priradené k viac ako 21 tisícom tried alebo skupín a viac než 1 milión z týchto obrázkov obsahuje aj bounding box (bodové ohraničenie) na základe ktorého vieme určiť kde sa na obrázku nachádza objekt. Takto rozsiahly dataset je skvelý nástroj pre trénovanie konvolučných neurónových sietí pre účely počítačového videnia. V praxi sa často ale využíva iba jeho podmnožina ktorá objekty rozdeľuje do 1000 tried a obsahuje ïba"1 431 167 obrázkov. [DDS+09]

2.4.2 DAVIS

V úlohe segmentácie objektu z videa v posledných rokoch dominujú práce ktoré sú založené na CNN. Na ilustráciu tohto trendu môžeme použiť aj výsledky z súťaže DAVIS [dav],kde najnovšie sú k dispozícii z roku 2020,

ktorá porovnáva úspešnosť segmentácie objektov rôznych prístupov na rovnakých dátach. Od prvého ročníku súťaže DAVIS v roku 2016 boli každý ročník riešenia umiestnené na prvých priečkach založené práve na princípe cnn. Súťaží sa v troch kategóriach a to automatické (Unsupervised) , poloautomatické (Semi-Supervised) alebo Interaktívne(Interactive). Rozdiel je v počiatočných dátach ktoré jednotlivé CNN dostanú. Automatické riešenia nedostanú na riešenie úlohy nič len samotné video dáta, polo-automatické dostanú segmentovanú masku objektu na prvom snímku videa a pri interaktívnych dostane sieť ako vstup zopár línii, ktoré naznačujú ktorý objekt chceme segmentovať, teda sieť musí byť pripravená na rôznorodé vstupy. Súčasťou tejto súťaže je aj dataset, ktorý je vhodný pre účely tejto práce a preto sa budem na neho odkazovať v ďalších častiach práce. Narozdiel od datasetu Image net /refImageNet ktorý obsahuje jednotlivé obrázky a zaradenie do tried, tento obsahuje videodáta, kde pre každý obrázok videa (frame) je k dispozícii aj ground truth segmentačná maska daného objektu.[dav]

2.4.3 Youtube-VOS

Tento dataset by sme mohli nazvať ako nástupcu predchádzajúceho datasetu Davis. Ide o dataset ktorý obsahuje rádovo viac videí, s manuálne anotovanými ground truth maskami pri 5 snímkoch každých 30 sekúnd. Takto dostaneme 6 anotovaných snímkov pre sekundu dlhú pasáž videa. Ako môžeme vidieť v obrázku 2.3 je oveľa komplexnejší ako všetky skôr vyvinuté datasety v počte videí aj rôznorodosti samotných videí. Ako naznačuje názov, tieto videá sú z portálu youtube, kde prešli manuálnou selekciou, kategorizáciou a anotáciou aby bola zabezpečená ich kvalita. [XYF+18]

Scale	JC [21]	ST [22]	YTO [16]	FBMS [24]	DAVIS [15]	DAVIS [20]	**YouTube-VOS (Ours)**
Videos	22	14	96	59	50	90	**4,453**
Categories	14	11	10	16	-	-	**94**
Objects	22	24	96	139	50	205	**7,755**
Annotations	6,331	1,475	1,692	1,465	3,440	13,543	**197,272**
Duration	3.52	0.59	9.01	7.70	2.88	5.17	**334.81**

Obr. 2.3: Porovnanie parametrov datasetov vytvorených pre účel segmentácie obrázkov.[XYF+18]

2.4.4 COCO

Coco je dalším z populárnych datasetov pre tréning neurónových sietí pre účely detekcie, segmentácie alebo popisu objektov. Obsahuje približne 330 tisíc obrázkov spadajúcich do 80 objektových kategórii. Pre účel vyhodnotenia našej segmentácie však nieje vhodný nakoľko neobsahuje ground truth masky pre objekty.

2.4.5 Segtrack

Segtrack je další dataset ktorý obsahuje videodáta s segmentačnou maskou pre každý snímok videí, avšak je oveľa menší ako DAVIS. Oproti datasetu DAVIS ktorý obsahuje väčšie časti videa, tento dataset obsahuje iba 12 rôznych scén, a navyše je pre každú iba zopár snímkov, avšak môže byť vhodný na vyhodnotenie nami dosiahnutého výsledku, nakoľko viacero videí obsahuje snímky z rozostrenými objektmi. [LKH+13a]

Kapitola 3

Návrh modelu

V tejto kapitole sa budem snažiť o navrhnutie riešenia ktoré sa budem snažiť v ďalšej kapitole implementovať. Vysvetlím tu práce ktoré spĺňajú čiastkové ciele mojej diplomovej práce a budú tvoriť celkové riešenie.

V mojej implementácii som sa rozhodol použiť konvolučnú neurónovú sieť z kategórie polo-automatických riešení. Ak by teda mala výsledná aplikácia pracovať s ľubovoľným videom, bude potrebné získať počiatočnú masku objektu záujmu z jedného zo snímkov videa (zvyčajne prvého snímku vo video-sekvencii). Preto sa budem v prvej časti tejto kapitoly budem venovať riešeniu segmentácie objektu z jedného snímku.

3.1 GrabCut

Za zlatý štandard v problematike segmentovania objektu z obrazu by sme mohli považovať metódu GraphCut a neskôr jej vylepšenú verziu GrabCut. Je

Táto metóda pre začiatok segmentácie vyžaduje manuálny vstup a to ohraničením približnej lokality objektu ktorý chceme segmentovať obdĺžni-

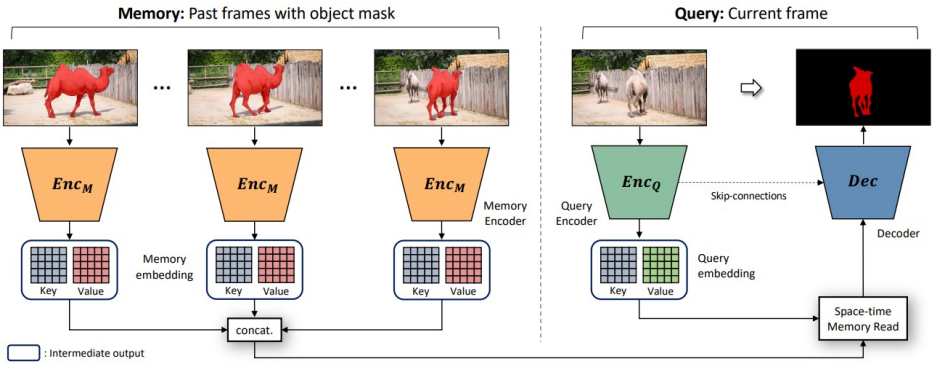
kom a metóda sama určí popredie a pozadie objektu. Následne je možné aj manuálne vyznačenie popredia a pozadia častí objektu, ak metóda na prvý krát určité časti objektu nevysegmentovala správne. Takéto vyznačenia môžeme pridávať iteratívne čím potencionálne zvyšujeme výslednú presnosť segmentácie. Princíp metódy GrabCut spočíva v prevedení všetkých pixelov vnútri obdĺžnika na vrcholy grafu, teda vznikne plne prepojený graf. Následne sú v grafe zavedené váhy priradené ktoré zodpovedajú pravdepodobnosti či daný pixel patrí popredíu alebo pozadiu. Susediacim pixelom je priradená váha podľa toho či medzi nimi je detegovaná hrana alebo na základe vzájomných podobnosti. Ak je vo vedľajších pixeloch napríklad príliš veľká zmena farebnosti, bude im priradená malá váha. Potom hľadáme minimálny rez grafom ktorý oddelí pozadie od popredia.

3.2 Neurónová sieť na segmentáciu objektu

V mojej práci som sa rozhodol použiť prácu [CTT21] ktorá je založená na pôvodnej práci [OLXK19]. Ide o neurónovú sieť s koncepciou ktorá bola vyvinutá iba v roku 2021, a následne dosiahla jednen z najlepších výsledkov na datasete YoutubeVoS [XYF+18], kde dosiahla v celkovom skóre 2.miesto v kategórii polo-automatickej objektovej segmentácie. K tejto práci je dostupný aj model siete s možnosťou dotrénovania.

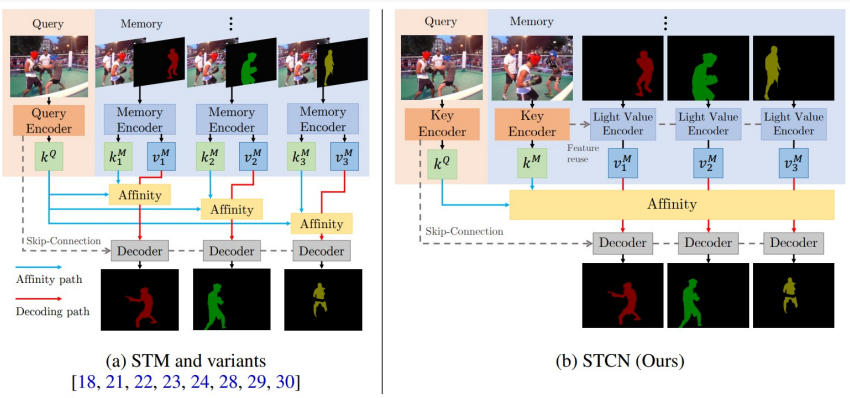
Hlavná myšlienka tejto siete je že si uchováva priebežné výstupy v externej pamäti a tým ich vie použiť na zvýšenie presnosti. Prístupy ktoré chceli v minulosti využiť predchádzajúce segmentácie sekvenice boli nútené tieto snímky zakomponovať do tréningu sieťe, čím sa zvyšovala presnosť segmentácie, no exponencialne sa zvyšovala výpočtová náročnosť pri dlhších videách. Namiesto toho táto architektúra utilizuje prídavnú pamäť do ktorej si ukladá

kľúč a hodnotu pre každý už segmentovaný obrázok v danej sekvencii.



Obr. 3.1: Diagram hlavnej myšlienky neurónovej siete STM. [OLXK19]

Dotaz (Query Encoder, EncQ) enkóder prijma ako vstup aktuálny obrázok, Pamäťový enkóder (Memory Encoder,EncM) príjma obrázok aj masku. Klúč sa používa na adresovanie. Konkrétne sa vypočítajú podobnosti medzi klúčom aktuálnej snímky a snímkami v pamäti, aby sa určilo ktoré hodnoty z pamäťe sú najbližšie a budú nápomocné pri segmentácii aktuálnej snímky. Klúč bude teda vektor popisujúci vizuálnu sémantiku na porovnávanie vzhladu objektov vo snímkach.



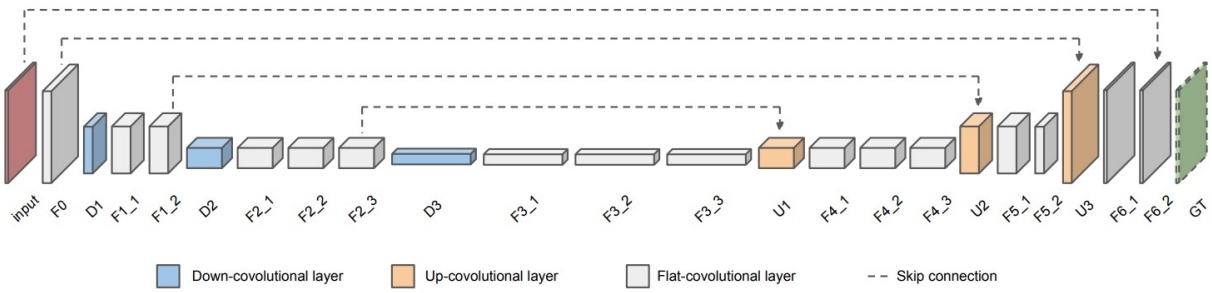
(a) STM and variants [18, 21, 22, 23, 24, 28, 29, 30]

(b) STCN (Ours)

Obr. 3.2: Diagram hlavných rozdielov medzi neurónovou sietou STCN(vpravo) [CTT21] v porovnaní s STM(vľavo) [OLXK19]. v STM sú jednotlivé objekty zakódované zvlášť a príbuznosti sú špecifické iba pre masku. V STCN sú zakódované pomocou siamského klúča na výpočet priamo z RGB obrázka, vďaka čomu sú robustnejšie a efektívnejšie.

3.3 Neurónová sieť na odstránenie rozmazania

Na odstránenie rozmazania by som chcel použiť prácu [GR19]. Ide o neurónovú sieť s pomerne klasickou koncepciou ktorá vychádza z [SDW+16]. Ide o sieť ktorá deteguje a následne aj odstraňuje rozmazanie pomocou okolitých snímok. Na jej trénovanie boli použité videa natočené frekvenciou 240fps (snímok za sekundu), kde každých sedem nasledujúcich snímkov vycentrujú podľa 4. v poradí a spoja do jedného snímku videa, čím vytvárajú syntetické rozmazanie pohybom. Takto vytvorené snímky potencionálne vytvárajú nežiadané artefakty a preto sú takto spojené iba tie snímky kde celkový priemer vzdialenosti pixelov je iba 1 pixel. Celkovo je nasnímaných 71 videí, každé s dlžkou 3-5 sekúnd, z ktorých je vygenerovaných 6708 rozmazaných snímkov s korešpondujúcimi ground truth snímkami.



Obr. 3.3: Koncepcia pôvodnej siete [SDW+16].

Kapitola 4

Implementácia

4.1 Softvérové požiadavky

- Python3
- Pytorch
- OpenCV
- Pillow-SMD
- Numpy, Matplotlib a Scipy:
- colorama

4.2 Grafické rozhranie

1. Užívateľ si spustí aplikáciu a nahrá video ktoré bude chcieť segmentovať.

2. Uživateľ na prvom snímku pomocou metódy grabcut označí objekt záujmu pomocou bounding boxu. Ak je to potrebné, vyznačí iteratívne popredie a pozadie ťahmi. Spustí segmentáciu

3. Video-sekvenciu prejde konvolučná neurónová sieť na detekciu a opravu rozmazania

4. Video-sekvenciu prejde konvolučná neurónová sieť na segmentáciu objektovej masky pre každú snímku.

4.3 Grabcut Metóda

Implementačné detaily Grabcut Metódy

4.4 STCN

Implementačné detaily STCN

4.5 Deblur

Implementačné detaily Deep Video Deblurring.

Kapitola 5

Výsledky

Literatúra

[CMPT+17] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation, 2017.

[CTT21] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation, 2021.

[dav] Davis : Densely annotated video segmentation. `https://davischallenge.org/index.html`. Accessed: 2021-11-30.

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[FZL+15] Qingnan Fan, Fan Zhong, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Jumpcut: Non-successive mask transfer and interpolation for video cutout. *ACM Trans. Graph.*, 34(6), oct 2015.

[GR19] Jochen Gast and Stefan Roth. Deep video deblurring: The devil is in the details, 2019.

[Hay09] Simon S. Haykin. *Neural networks and learning machines*. Pearson Education, Upper Saddle River, NJ, third edition, 2009.

[JXG17] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos, 2017.

[JZW+20] Runhua Jiang, Li Zhao, Tao Wang, Jinxin Wang, and Xiaoqin Zhang. Video deblurring via temporally and spatially variant recurrent neural network. *IEEE Access*, 8:7587–7597, 2020.

[KJK+01] Munchurl Kim, J.G. Jeon, J.S. Kwak, M.H. Lee, and C. Ahn. Moving object segmentation in video sequences by user interaction and automatic object tracking. *Image and Vision Computing*, 19(5):245–260, 2001.

[LKH+13a] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg. Video segmentation by tracking many figure-ground segments. In *2013 IEEE International Conference on Computer Vision*, pages 2192–2199, 2013.

[LKH+13b] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013.

[OLXK19] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks, 2019.

[RKB04] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23:309–314, 08 2004.

[SCXP15] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. *CoRR*, abs/1503.00593, 2015.

[SDW+16] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring, 2016.

[SLT11] Bolan Su, Shijian Lu, and Chew Lim Tan. Blurred image region detection and classification. pages 1397–1400, 11 2011.

[WJR+21] Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai. Swiftnet: Real-time video object segmentation. *CoRR*, abs/2102.04604, 2021.

[WZP+21] Wenguan Wang, Tianfei Zhou, Fatih Porikli, David Crandall, and Luc Van Gool. A survey on deep learning technique for video segmentation, 2021.

[XFL+18] Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. Monet: Deep motion exploitation for video object segmentation. 06 2018.

[XYF+18] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark, 2018.

[YWY21] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration, 2021.

Zoznam obrázkov