UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

DETEKCIA VÝZNAMNÝCH OBLASTÍ VO VIDEU

Diplomová práca

2016 **Bc. Martin Kuchyňár**

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

DETEKCIA VÝZNAMNÝCH OBLASTÍ VO VIDEU

Diplomová práca

Študijný program: Aplikovaná informatika
Študijný odbor: 9.2.9. aplikovaná informatika
Školiace pracovisko: Katedra Aplikovanej Informatiky
Školiteľ: RNDr. Elena Šikudová, PhD.

Bratislava 2016 **Bc. Martin Kuchyňár**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Martin Kuchyňár
Študijný program: aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.2.9. aplikovaná informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Detekcia významných oblastí vo videu
Spatio-temporal salient object detection

Cieľ: Metódy na detekciu významných oblastí vo videu:
1. Naštudovanie
2. Návrh zlepšenia
3. Implementácia
4. Porovnanie výsledkov

Vedúci: RNDr. Elena Šikudová, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: doc. PhDr. Ján Rybár, PhD.

Dátum zadania: 20.10.2014

Dátum schválenia: 24.10.2014

prof. RNDr. Roman Ďurikovič, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Čestné vyhlásenie

Čestne prehlasujem, že som túto diplomovú prácu vypracoval samostatne s použitím uvedených zdrojov.

V Bratislave .....

Poďakovanie

Ďakujem mojej vedúcej práce RNDr. Elene Šikudovej, PhD. za odborné vedenie, profesionálne usmerňovanie a cenné rady, ktoré mi spokytla pri vypracovávaní diplomovej práce.

Abstrakt

V diplomovej práci sú navrhnuté a zvalidované metódy pre spracovanie videa a detekciu významných oblastí vo videu. Navrhovaná metóda sa zameriava na získanie čisto dynamických príznakov, ktoré je možné následne kombinovať s klasickými príznakmi alebo použiť ako samostaný príznak pri vytváraní významných oblastí obrazu. Pre uľahčenie prototypovania podobných modelov bola vytvorená ucelená aplikácia v prostredí Matlab, poskytujúca automatickú validáciu modelu pomocou štandardných datasetov a jednoduché vkladanie konkurenčných modelov pre okažité porovnávanie. Práca je rozdelená do piatich kapitol a zaoberá sa výskumom alternatívnych spôsobov detekcie významných oblastí pre potreby ďalšieho spracovania obrazu.

Kľúčové slová: *významné oblasti, video, Matlab*

Abstract

In the diploma thesis, we propose and validate methods for processing video and detection of salient areas in it. Suggested method are focuses on acquirement of purely dynamic attributes, which can be subsequently combined with classic attributes or can be used as separate feature for generating salient maps. To simplify prototyping of similar models, a comprehensive application was created in Matlab environment, which is providing automatic validation of model using standard datasets and simple input of competitive models for immediate comparison. In five chapters, we deal with research of alternative approaches of detection of salient areas for requisites of further image processing.
Keywords: *saliency, video, Matlab*

Obsah

1. Úvod

Ľudské oko je schopné spracovať 10^8 až 10^9 bitov obrazových dát za sekundu. Ľudský mozog nie je schopný spracovať také množstvo dát naraz, preto sa získané informácie filtrujú pomocou ľudského vizuálneho systému[26]. Ľudský vizuálny systém je pravdepodobne najzložitejším mechanizmom akým človek disponuje. Je natoľko kľúčovým pre fungovanie spoločnosti či jedinca, že psychológovia sa zaoberajú jeho výskumom. Už viacero dekád študujú vlastnosti tohoto mechanizmu z pohľadu psychológie, fyziológie alebo neurobiológie.

Vyfiltrované oblasti obrazu je možné spracovať vo výrazne rýchlejšom čase ako nefiltrované, ideálne v reálnom čase. Takéto oblasti sa nazývajú významné alebo charakteristické (v literatúre salient - prebraté z angličtiny). Významné oblasti sú vyberané pomocou mnohých faktorov. Najznámejšími sú prechody vo farbe, intenzite alebo orientácii. Ľudský vizuálny systém taktiež využíva skúsenosti pri pozorovaní. Oblasti takto vyfiltrované nesú pre pozorovateľa viac potencionálnych informácií ako ostatné oblasti obrazu a preto sa stávajú salientnými.

V systémoch počítačového videnia sa snažíme využívať primárne tieto oblasti pre pridelenie väčšej časti zdrojov. Z tochto dôvodu je zistenie významných oblastí častým prvým krokom mnohých algoritmov v oblasti počítačového videnia.

Algoritmy na detekciu významných oblastí sa delia do troch skupín podľa princípu akým spracovávajú dáta[11]

1. Zdola-nahor: Prístup je cielený na nezávislosť od používateľa. Zameriava sa na fyziologicky významné oblasti vizuálneho systému ako výrazné zmeny v tvare, jase alebo farbe.

2. Zhora-nadol: Prístup je založený na čiastočnom riadení zo strany používateľa (konanie je podmienené úlohou). Riadenie je prínosom, pretože obsahuje aj informáciu používateľa a jeho predchádzajúcich vedomostí či skúseností, ktoré ovplyvňujú vnímanie.

3. Algoritmy využívajúce neurónové siete.

Cieľom práce je štúdium a výskum nových metód na detekciu významných oblastí vo videu a ich porovnanie s existujúcimi v rôznych štandardných oblastiach.

V prvej časti sa nachádza prehľad metód na detekciu významných oblastí vo videu alebo metód na detekciu v statických obrazoch, ktoré majú potenciál pre použitie aj vo videu. Ďalej obsahuje detailné vysvetlenie fungovania metód, ktoré budú použité v implementácii zlepšenia.
V druhej časti je popísaný postup, princíp zlepšenia a implementácie novej metódy.
Tretiu časť tvorí validácia výsledkov a porovnanie výsledkov s inými modelmi, ktoré poskytujú lepšiu predstavu o efektivite algoritmu. A následne aj diskusiu obsahujúcu analýzu výsledkov validácie a z nej vyplývajúce možnosti na zlepšenie narvhnutého algoritmu.
Záver obsahuje zhrnutie vytýčených cieľov práce a ich objektívne zhodnotenie.

2. Prehľad literatúry

2.1 Úvod do problematiky

Saliency a teda detekcia významných oblastí je využívaná v rôznych oblastiach. Automatizáciou sú modely významných oblastí (anglicky saliency modelov) ťažiskom pri segmentácii obrazu alebo detekcii špecifických objektov. Od saliency modelov sú taktiež závislé aj programy ovládajúce zabezpečovacie zariadenia. Pomocou saliency modelov zužujú možnosti a proaktívne upozorňujú na podozrivé situácie pomocou zúženia obrazu na zopár oblastí. Saliency model používa aj oblasť reklamy, kde je vizuálna pozornosť kľúčovým parametrom, čo môže rozhodnúť o úspechu produktu. Veď aký význam by mala reklama, kde si nevšimnete prezentovaný produkt alebo si všimnete iba jeho 'menej' dokonalé časti.

2.2 Metódy pre statické obrazy

Algoritmy pre statické obrazy tvoria základ všetkých saliency modelov a sú najstaršou oblasťou výskumu. V tejto časti uvediem prehľad algoritmov pre výpočet saliency modelov od najjednoduchších cez najznámejšie až po najefektívnejšie. Na záver uvediem porovnanie všetkých metód pomocou všeobecne uznávaných metrík a dát získaných zo zariadení merajúcich pohyb očí používateľa (eyetrackera).

2.2.1 Baseline Center

Baseline center je triviálny model, ktorý sa vypočítava pomocou Gaussovej krivky vzhľadom na pomer strán, čím predpokladá salientné oblasti presne v strede obrazu. Nezachytáva však žiadne sémantické aspekty videa, ako ani podvedomé informácie vnímanania obrazu, iba rozlíšenie dané optikou skenujúcou scénu.

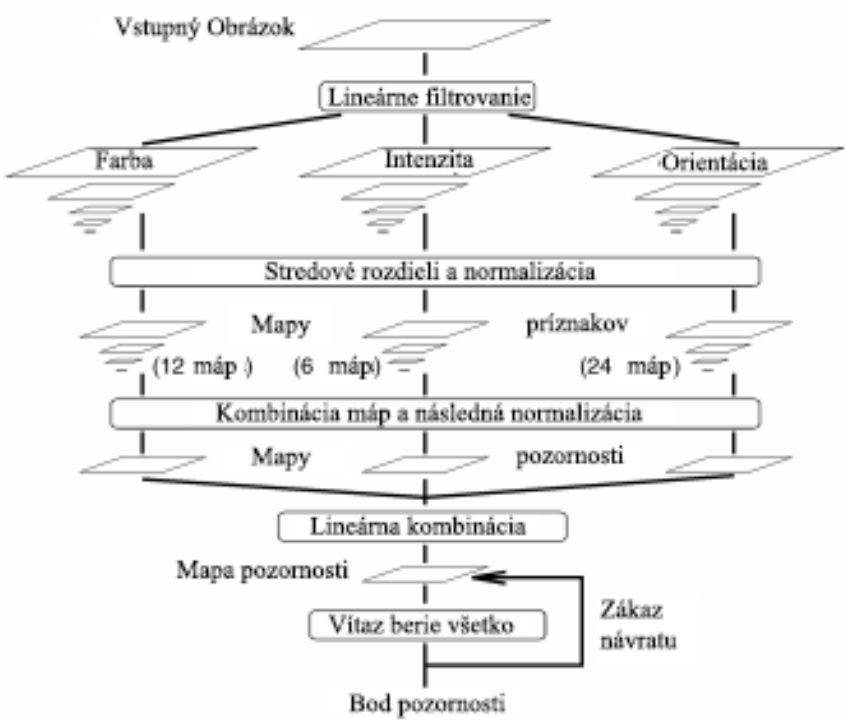
2.2.2 Hrany

Skupinu algoritmov využívajúcu význačné prechody v obraze inak nazývame hrany. Metódy tohoto typu sú vyžívané hlavne v prírodných scénach, kde nie je (sémanticky) význačný objekt. Takéto metódy sa zakladajú priamo na štúdiu fyziologických vlastností ľudského

vizuálneho systému. Následná imitácia procesov odohrávajúcich sa na sietnici viedla ku vzniku saliency modelov generujúcich plausibilné výsledky[3].

2.2.3 Ittiho model

Najznámejším modelom pre výpočet významných oblastí pre statické farebné obrazy je ittiho model navrhnutý v roku 1998. Model zakladá na rozložení obrazu na tri základné charakteristiky obrazu a to na farbu, intenzitu a orientáciu.

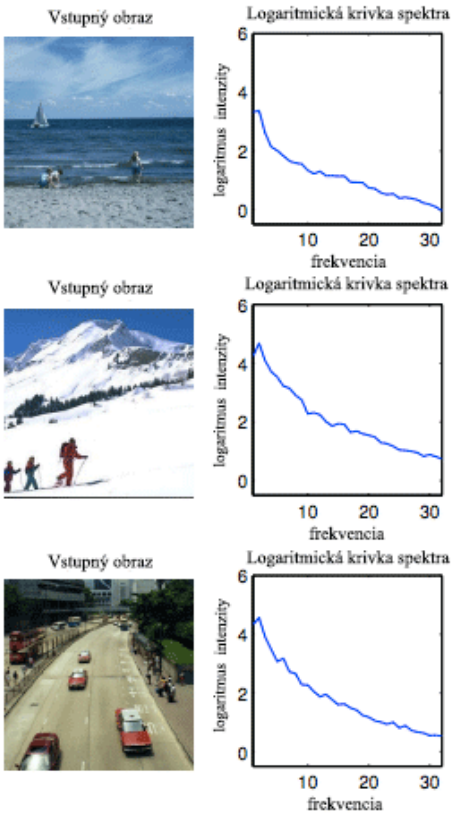


Obr. 2.1: Ucelená vizualizácia ittiho modelu[15]

Chrakteristika farby obsahuje 12 máp (šedotónové obrazy), pričom model používa farebný model RGB. Na začiatku sa vypočíta intenzita podľa vzťahu $I = (R + G + B)/3$. Pomocou mapy I sa následne normalizujú všetky farebné kanály modelu RGB. Model extrahuje štyri farebné kanály červený (r), zelený (g), modrý (b), žltý (y) a pomocou Gausvých pyramíd vytvorí tri rôzne mapy každej farebnej zložky separátne. Červená zložka sa počíta difenčným spôsobom ako $R = r-(g+b)/2$, zelená ako $G = g-(r+b)/2$, modrá ako $B = b-(r+g)/2$ a žltá ako $Y = (r+g)/2-|r-g|/2-b$. Chrakteristika intenzity obsahuje šesť máp. Získaná je pomocou orientovaných gáborových filtrov s orientáciou 0°, 45°, 90°, 135°. Spolu 42 máp charakteristík je následne lineárne skombinovaných do jednej saliency mapy[15].

2.2.4 Spektrálne reziduá

Metóda využíva princíp, že potláča štatisticky často opakujúce sa časti obrazu a do popredia stavia časti obrazu, ktoré sa štatisticky odlišujú od ostatných. Na detekciu používa rýchlu fourierovu transformáciu. Pomocou nej rozdelí obrázok na amplitúdovú časť a fázovú časť.



Obr. 2.2: Príklad rozloženia typovo rôznych obrazov[14]

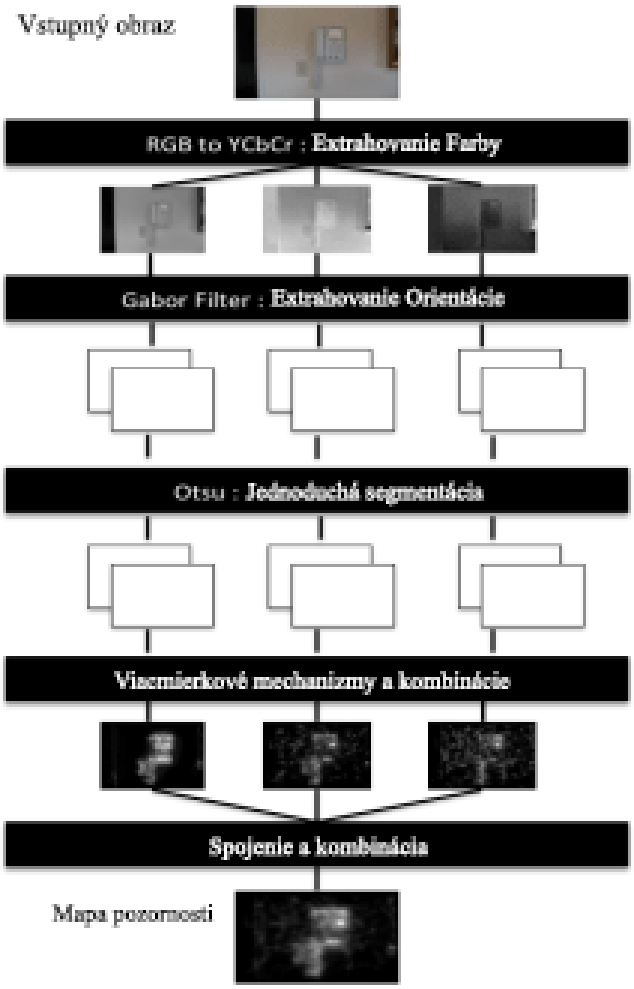
Amplitúdová zložka sa následne vyhladí, čím sa do popredia dostanú iba informácie, ktoré sa vymykajú z priemeru. Odčítaním od pôvodnej amplitúdovej zložky dostaneme iba časti obrazu, ktoré sú významné [14].

2.2.5 Sun Model

Sun model (Saliency Using Natural statistics) sa snaží simulovať potencionálne ciele sledovania ľudského vizuálneho systému. Model aktívne hodnotí tieto ciele odhadom pravdepodobnosti vzhľadom na všetky pozorované charakteristiky. Charakteristiky sú spracovávané separátne a teda model nepočíta s chrakteristikami navzájom sa ovplyvňujúcimi. Údaje získané zo všetkých charakteristík následne spracuje štatisticky. Model zakladá hlavne na Bayesovom pravidle. Ako výsledok hľadania potom udáva asymetrie v týchto štatistických štruktúrach[27].

2.2.6 Rare Model

Výrazná väčšina modelov pozornosti typu bottom-up funguje ustáleným postupom, kde sa z pôvodného obrazu extrahuje definovaná množina chrakteristík paralelne a tie sa následne kombinujú alebo inak použijú na výpočet výslednej mapy pozornosti. Rare model narvhuje sekvenčnú architektúru, kde z pôvodného obrazu extrahuje nízko úrovňové príznaky. Následne na výsledkoch sériovo vykonáva extrakciu ďalších príznakov (v literatúre nazývané mid-level). Nakoniec ako posledný krok spojí a normalizuje výsledné chrakteristiky do konečnej mapy významných oblastí. Rare model ako nízkoúrovňové chrakteristiky používa jas a colorimetrické rozdiely (ako farebný model používa YCbCr) a následne na mapách rozložených zložiek farebného modelu detekuje orientáciu pomocou gáborových filtrov[23]. Po extrakcii všetkých charakteristík použije iteratívnu metódu pre optimálne kvantovanie, založenú na metóde Otsu[1]. Na takto upravenom vstupe sa následne vyhľadávajú vzácne (z angl. rare) oblasti obrazu. Metóda preskúmala možnosti nesekvenčnej extrakcie príznakov z obrazu, čo bolo novým prístupom v oblasti modelov pozornosti.



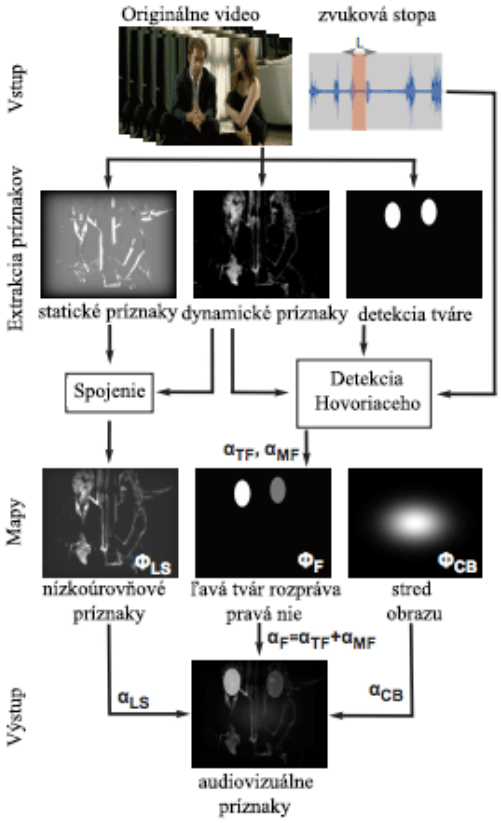
Obr. 2.3: Rare model workflow[23]

2.3 Metódy pre videá

Video obsahuje rozsiahlejšie možnosti ako obrazová informácia, pribúdajú ďalšie rozmery ako je pohyb objektov na obraze alebo vplyv zvuku na ľudské vnímanie. Avšak oproti obrazu je potrebné spracovávať väčšie množstvo dát. Navyše vo väčšine algoritmov využívajúcich saliency modely je potrebné, aby model dával výsledky v reálnom čase hlavne v oblasti zabezpečovacej techniky.

2.3.1 Zohľadnenie audio informácie

Saliency modely využívajú rôznorodé druhy príznakov a to od geneticky zakorenených, ako sú prechody farieb alebo intenzít, až po sémantické príznaky, ako je detekcia tváre [25]. Majoritná väčšina saliency modelov využíva iba obrazovú zložku, pričom zvuková stopa býva nevyužívaná alebo úplne zanedbaná. Použitie zvuku je známym trikom filmovej scény už desatročia. Režiséri posilňujú kontrolu nad diváckou pozornosťou práve pomocou zvukového doprovodu. Prvé štúdie sa zaoberali detekciou reči a tváre, kde je spojitosť jednoznačná [9]. Neskoršie štúdie dokazujú korelácie aj na všeobecnejšej úrovni a pokusy o extrakciu samotnej charakteristiky zo zvukovej stopy[10]. Tieto pokusy viedli aj k zostaveniu modelov zohľadňujúc zvukovú stopu ako samostatnú charakteristiku spolu s kombináciou nízkoúrovňových príznakov obrazu [6].



Obr. 2.4: Vizualizácia audiovizuálneho modulu[6]

Model extrahuje video na sekvenciu obrazov (framy) a audio stopu v tvare grafu vlnovej dĺžky. Potom extrahuje tri typy rôznych chrakteristík. Nízkoúrovňové príznaky založené na biologicky inšpirovaných saliency modeloch rozdelených na dynamickú časť a statickú časť. Statická časť sa zameriava na najasnejšie a najkontrastnejšie časti obrazu. Dynamická časť sa zameriava na relatívny pohyb objektov vzhľadom na pozadie (eliminácia pohybu kamery). Tieto dve časti sa nakoniec spoja. Ďalšou charakteritikou použitou v tomto modeli je detekcia tváre. Každý objekt klasifikovaný ako tvár je v saliency mape nahradený oválnym objektom. Intenzita daných objektov je daná pomocou metódy Speaker Diarization[6], ktorá detekuje podľa zvukovej stopy objekt, ktorý generuje zvuk. Metóda predpokladá striedavú konverzáciu viacerých objektov oddelenú pauzou. Následne spojí vyššie spomínané charakteristiky do jednej výslednej mapy. Ako posledný krok preloží cez celú mapu baseline center model popísaný v časti 2.2.1.

2.3.2 Detekcia pohybu

Táto časť je zameraná na segmentáciu objektov, ktoré sa na scéne pohybujú. Metódy tohoto typu sa snažia vizualizovať 3D prostredia (v našom prípade disponujeme výškou, šírkou a časom) na 2D výstup (obrazový výstup). Takáto informácia dokáže priblížiť výpočtové modely bližšie k realite. Ľudský vizuálny systém totiž nepoužíva iba 2D vstup (ako to prebieha v drvivej väčšine metód na výpočet významných oblastí). Takéto obrazy sú v ľudskom vizuálnom systéme vysoko hodnotené. Dôvody, prečo takto ľudský vizuálny systém zvyšuje prioritu práve takýmto oblastiam môžeme nájsť v antropológii. Vysvetlenie je jednoduché a to snaha zabezpečiť bezpečné prostredie okolo seba a všetko pohybujúce sa narušuje pocit bezpečnosti. V nasledujúcom texte rozoberieme dva najpožívanejšie algoritmy na detekciu oblastí pohybu v obraze a výpočet vektoru posunu. Výpočet vektoru posunu je však iba projekcia 3D vstupných dát do 2D obrazu a nemusí vždy reprezentovať iba pohyb. Prvým z nich bude Lucas Kanade[4] a druhým Horn Schunck[13]. Obidva tieto algoritmy používajú jeden spoločný predpoklad a to, že jas daného objektu sa časom nemení. To znamená, že objekt sa na scéne môže presunúť, ale svoj jas nemôže zmeniť. Matematicky vyjadrené I(x(t),y(t),t) je obrazová dvojrozmerná funkcia, ktorá sa mení vzhľadom na čas. Keďže sa jas obrazu nemení môžeme povedať, že platí:

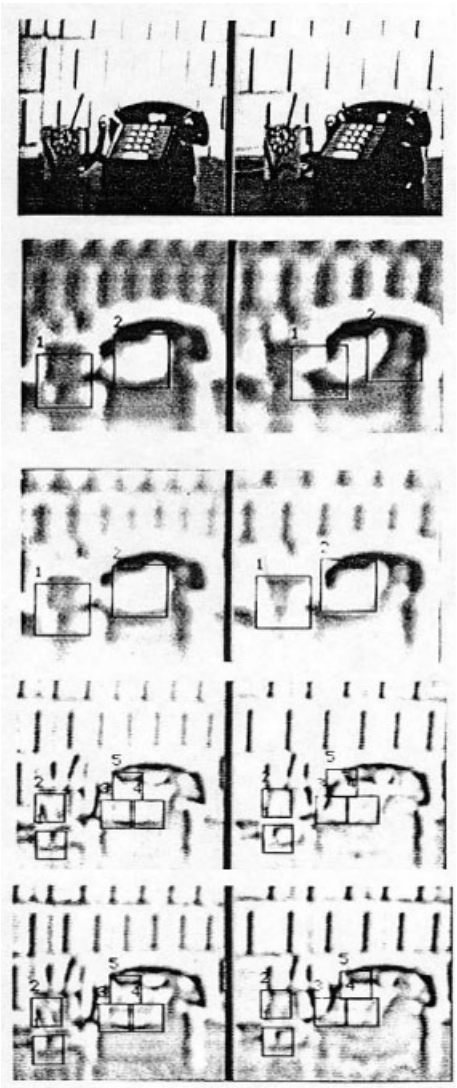
$$I(x + dx/dt, y + dy/dt, t+) = I(x, y, t) \tag{2.1}$$

Z čoho je ľahko odvoditeľné, že:

$$dI/dt = /dx/dt + /dy/dt + dI/dt = 0 \tag{2.2}$$

2.3.3 Lucas Kanande

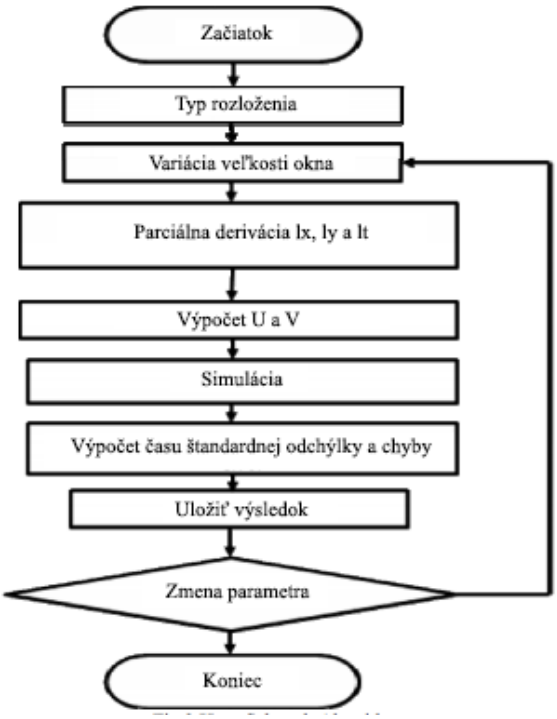
Algoritmus Lucas Kanande prvotne vznikol ako návrh pre časovú optimalizáciu problému výpočtu vektora posunu medzi dvomi krivkami. Pôvodné primitívne riešenie vyžadovalo $O(M^2 * N^2)$ času pre výpočet daného vektora, ak M,N bolo rozlíšenie daného obrazového vzoru. Vtedy navrhovaná optimalizácia vyžadovala zadanie rozsahu hľadania, pomocou ktorého sa vypočítali diferencie pre celý obraz a pre ďalšiu iteráciu sa rozsah vypočítal pomocou horolezeckého algoritmu. Metóda Lucas Kanade využíva priestorový gradient pre výpočet nových hodnôt a zároveň upravuje hodnotu rozsahu pri výpočte každého obrazového pixelu v obraze a nie iba po výpočte celého obrazu. Pomocou takejto úpravy naivného algortimu sa časová zložitosť zlepšila na $O(M^2 logN)$[4].



Obr. 2.5: Vizualizácie výsledkov algoritmu Lucas-Kanade vždy po 1 iterácii[4]

2.3.4 Horn-Schunck

Metóda Horn-Schunck[13] bola prvá, kde bola použitá metóda variácie na výpočet optického toku. Táto globálna metóda priniesla výpočet konštanty pre obmedzenie plynulosti optického toku. Algoritmus používa dva základné parametre: počet iterácií a vyhladzovaciu konštantu. Počet iterácií určuje dĺžku (počet cyklov) simulácie, vyhladzovacia konštanta je použitá po každom cykle simulácie kvôli zjemneniu prechodov a na výpočet optimálneho optického toku.



Obr. 2.6: Vizualizácia pracovného postupu metódy Horn-Schunck[2]

2.4 Metriky úspešnosti

Metriky úspešnosti sú algoritmy na čo najpresnejšie vyjadrenie presnosti modelov v merateľných jednotkách. Takýto algoritmus dostáva na vstupe čisté dáta z eye trackera. Tieto je potrebné predspracovať z dôvodu, že každý výrobca poskytuje iné zariadenia na hardwarovej úrovni a výrobcovia neštandardizujú výstup do jednotnej formy. Následne je potrebné vytvoriť mapu fixácií, ktorá sa používa ako jeden zo vstupných parametrov v algoritmoch počítajúcich metriky úspešnosti.

Metriky úspešnosti možno rozdeliť do troch štandardných skupín podľa druhu hodnôt, na ktoré porovnávajú reálne dáta (v literatúre nazývané ground truth) s vygenerovanými mapami významných oblastí[22].

1. **Založené na porovnávaní hodnôt** - NSS, Percentile, Pf
2. **Založené na vyhodnocovaní vzdialeností** - AUC-Judd, AUC-Zhao, AUC Borji, AUC-Li
3. **Založené na distribúcii** - KL-Div, EMD, CC, SRCC

2.4.1 NSS

NSS (Normalized Scanpath Saliency) je metrika narvrhnutá v roku 2005, ktorej autormi sú R. J. Peters a L. Itti. Metrika zakladá na ohodnotení salientných oblastí vzhľadom na pozíciu fixácií samostatne a následne hodnotu normalizuje vzhľadom na počet fixácií v celom obraze[22].
Pre každú fixáciu používa vzťah:

$$NSS(p) = (SM(p) - \mu_{SM})/\sigma_{SM} \tag{2.3}$$

Kde SM je mapa význačných oblastí, p je bod danej fixácie, pre ktorú sa hodnota vypočítava. Mapa fixácií SM je normalizovaná tak, aby nadobúdala nulovú strednú hodnotu a zároveň jednotkovú štandardnú dochýlku. Metrika NSS nadhodnocuje, ak je na výslednej saliency mape minimálna rozmanitosť hodnôt (malý rozdiel medzi hodnotami fixácií a strednou hodnotou), pretože v takomto prípade nebude model dostatočne ohodnotený. Ak hodnotený model nájde presné pozície a odchýlka je malá, alebo rozdiel medzi hodnotami fixácie a strednou hodnotou je vysoký. Potom finálna hodnota NSS metriky je určená priemerom hodnôt pre všetky fixácie[22]. Vzťah možno vyjadriť nasledovne:

$$NSS = 1/N * \sum_{p=1}^{N} NSS(p) \tag{2.4}$$

Kde p je počet fixácií a NSS(p) je výsledok vzťahu 2.3.

2.4.2 AUC-Judd

Metrika je clasická AUC, ktorú navrhol Judd [16]. Ako prvé sa pixely označené ako fixácie spočítajú s rovnakým počtom náhodných pixelov vybraných z mapy významných oblastí a nakoniec sú považované za klasifikátor úspešnosti. Nasleduje prahovanie zvolenou hodnotou. Pixely, ktoré sú menšie ako prahovacia hodnota sú pokladané za pozadie obrazu a pixely, ktoré majú hodnotu vyššiu sú pokladané za fixácie. Pre ľubovolne zvolenú prahovaciu hodnotu sú niektoré výsledné oblasti manuálne označené ako pozitívne (True Positives). Pobobne niektoré oblasti, ktoré nie sú označené ako fixácie, sú manuálne označené ako falošne pozitívne (False Positive). Tieto operácie sú zopakované tisíckrát. Nakoniec sa

vizualizuje pomocou ROC krivky a plocha pod krivkou (Area Under the Curve preto AUC) je výsledným klasifikátorom, ktorého ideálna hodnota je 1. Hodnota náhodného výberu je 0.5.

2.4.3 KL-Div

KL-Div v literatúre nazývaná aj Kullback-Leiblerova divergencia[17] je bežne používaná aj mimo oblasti počítačového videnia, ako metóda pre odhad celkovej rozdielnosti medzi dvoma distribúciami. Mnoho autorov saliency modelov používa túto metriku ako hodnotu straty informácie, tj. koľko informácií sa stratí po vypočítaní daného saliency modelu voči mape fixácií.

Každý projekt vytvárajúci model významných oblastí si volí vlastné metriky úspešnosti, podľa ktorých sa určuje úspešnosť daného modelu. Pre meranie úspešnosti modelov je okrem samotných algoritmov potrebné zabezpečiť dostatočne rôznorodú skupinu testovacích dát tzv. datasetov.

2.5 Referenčné datasety

Dataset je testovacia množina vstupov, ktorá sa snaží obsiahnuť dostatočne rôznorodé vzorky vhodné pre komplexné testovanie. Pri zostavovaní datasetov sú dôležité nielen videá ale aj eyetracker data alebo nejakým spôsobom zverejnené fixácie (získané napríklad aj ručným označovaním významných oblastí), aby bolo možné výsledky validovať pomocou vyššie uvedených metrík. Ďalšou charakteristikou datasetu je množsto ľudí, na ktorých boli dané viedeá nahrávané.
Príklady datasetov:

- **RSD**[18]
- **SAVAM**[12]
- **Coutrot datasets**[**courot-database**]
- **ASCMN**[24]

2.5.1 RSD

Regional Saliency Dataset sa snaží o čo najobšírnejšie testovanie a o rôznorodosť videií. Je rozdelený do 4 hlavných kategórií:

- **bezpečnostné záznamy** - Štadardné záznamy z bezpečnostných kamier obsahujú statické pozadie a salientné pohybujúce sa objekty. Pre túto časť datasetu boli využité záznamy z projektu CAVIAR[20].
- **Grafika** - Použité animované filmy/seriály ktoré obsahujú 2D aj 3D grafiku.
- **Prirodzené videá s prvkami grafiky** - Videá podobné bezpečnostným videám ale s prvkami umelo vloženými priamo do obrazového kanálu.
- **Prirodzené videá** - Videá bez pridaných grafických prvkov, tak ako boli nasnímané kamerou.

Na vyznačenie významných oblastí nebola zvolená technika (eyetracker) ale manuálne vyznačovanie zaujímavých oblastí pomocou používateľov. Výskumu sa zúčastnilo 17 mužov a 6 žien medzi 10-23 rokov veku, na označení každého z videa sa podielalo 10-23 ludí.

Obr. 2.7: Ukážka z každej kategórie videa s označenými významnými oblasťami

2.5.2 SAVAM

SAVAM (Semiautomatic Visual-Attention Modeling) je dataset nahrávaný priamo pomocou eyetrackera pri sledovaní videí v HD rozlíšení, pričom každému nahrávanému používateľovi sú pridelené dáta separátne pre kažké oko. Spolu obsahuje 13 minút videa, ktoré bolo otestované na 50-tich používateľoch rôzneho veku. Dataset je rozdelený na videá z filmov, ukážky z komerčných videí a na stereoskopické videá. SAVAM taktiež poskytuje všetky raw dáta z eyetrackere ako aj vizualizácie daných dát[12].

2.5.3 ASCMN dataset

ASCMN nazvaný podľa rozčlenenia do piatich skupín videa: abnormálne, bezpečnostné, videá s davom, videá s pohybom a videá s chybami v obraze (z anglických názvov: Abnormal, Surveillance, Crowd, Moving, Noise). Spolu obsahuje 24 videí. Každé video bolo namerané na 10-tich rôznych používateľoch. K datasetu je taktiež dostupný validačný kód[24] počítavajúci hodnotiace metriky na ľubovolnom modeli pozornosti.

2.5.4 Coutrot dataset

Ide o dva rôzne datasety, obidva sú nazývané podľa authora Antoine Coutrota. **Prvý dataset[9]** obsahuje videá s dynamickou povahou scény. Je rozčlenený do štyroch vizuálne rozličných kategórií:

- Jeden pohybujúci sa objekt
- Viacej pohybujúcich sa objektov
- Prírodné scény
- Konverzačné scény

Tento dataset obsahuje spolu 60 videí, ktoré sledovali vždy 18-tich rôznych používateľov. Všetky videá boli zaznamenávané v štyroch rôznych zvukových podmienkach (využívať budeme iba dáta s pôvodnou zvukovou stopou).
Druhý dataset [7] obsahuje 15 videí. Všetky videá obsahujú nahraté stretnutie štyroch konverzujúcich ľudí so statickou kamerou. Dataset nie je členený do žiadnych kategórií. Dáta oboch datasetov boli nahrávané pomocou eyetrackera EyeLink 1000 pri 1000Hz, pričom používatelia sedeli 57 cm od monitoru. Eyetracker nahrával iba dáta z dominantného oka pre daného používateľa.

2.6 Porovnanie štandardných metód

Porovnávanie metód je štandardne publikované formou ucelených benchmarkov. Príkladom takéhoto banchmarku je mit saliency benchmark[5], ktorý sa snaží zgrupovať a porovnávať obrazové modely pozornosti a zverejňovať referencie na ďalšie podobné projekty. Na účely validácie bude vypracovaný podobný benchmark určený pre porovnanie rôznych modelov pozornosti na typovo rôznych datasetoch.

3. Špecifikácia

3.1 Platforma pre riešenie

Ako platfomu pre implementáciu bude použitý programovací jazyk Matlab®. Všetky zdrojové kódy budú poskytnuté výhradne v tomto vývojárskom jazyku.

3.2 Očakávané výsledky

Výsledkom práce bude model pozornosti, ktorý zohľadňuje príznaky extrahovateľné iba z videa a nie z čisto obrazovej informácie. Pôjde o pohyb objektov na scéne a iné sémantické informácie, ktorými sa video odlišuje od statickej scény. Sekundárnym prínosom práce bude vytvorenie jednotnej aplikácie pre vizuálne porovnávanie modelov, kde používateľ bude môcť jednoducho pridávať modely. Ideálne priamo použiť ukážkové zdrojové kódy zverejnené autormi jednotlivých modelov alebo úpravou, ktorá nevyžaduje znalosť logiky stojacej za daným modelom. Následne automatický výpočet štadardných metrík na implementovanom datasete, pre jednoduchú validáciu výsledkov na rovnakých dátach, spolu s konkurečnými modelmi z dôvodu jednoduchého ladenia počas vývoja modelu.

3.3 Ideálne prípady

Ideálne prípady sa očakávajú v prípade použitia záznamov z bezpečnostných kamier, z dôvodu statickej kamery. Vďaka statickému pozadiu sú výsledky detekcie optického toku objektov najrelevantnejšie, čo predurčuje takéto videá k najlepším výsledkom.

3.4 Problémové prípady

Najproblémovejšími vstupmi sú očakávané videá s dynamickým pohybom kamery kombinovaným s pohybom objektov, alebo videá s fixovanou kamerou na objekt dynamicky sa pohybúci po scéne. Vo videách takéhoto charakteru sa predpokladá chybné označovanie oblastí a z toho vyplývajúce nepresnosti v mapách pozornosti. Preto v týchto prípadoch budú utlmované dynamické príznaky videa a budú sa používať iba statické príznaky aktuálneho obrazu.

4. Implementácia

4.1 Návrh metódy

Navrhovaná metóda zohľadnuje vlastnosti, ktoré nie je možné získať iba zo statického obrazu. Budem ich nazývať dynamické príznaky videa. Avšak metóda stále zohľadňuje v pozorovanom videu aj aspekty statického obrazu, ktoré sa budú nazývať statickými príznakmi videa. Tieto príznaky sú vypočítavané separátne a nakoniec ich metóda spája do jednej výslednej mapy pozornosti. Výsledkom je postupnosť máp pozornosti pre každý frame videa (podľa vstupnej konfigurácie), ktorý možno spojiť do videa pozornosti pre ľubovolné vstupné video.

4.1.1 Dynamické príznaky videa

Dynamické príznaky videa metóda najskôr extrahuje pomocou metódy Horn-Schunck[13], ktorá vypočíta optický tok na každých dvoch po sebe idúcich framoch videa, čím vzniká sémantický príznak pohybu rôznych objektov po scéne spolu so smerovými vektormi pohybu. Smerové vektory spočítavame, aby sme získali celkový obraz optického toku pre danú dvojicu obrazov. Obraz sa následne prahuje statickou konštantou kvôli ostráneniu šumu. Prahovanie prebieha dynamicky vzhľadom na počet nájdených 8-spojitých regiónov, tj. na výstup optikého toku. V našej implementácii je obmedzený počet regiónov na maximálnu hodnotu 200 regiónov. Prahovanie začne s konštantou, ktorá je určená pomocou algoritmu Otsu[1] a následne sa určí počet 8-spojitých regiónov. Ak je počet regiónov väčší ako maximálna hodnota, zvýši sa konštanta o 10% z pôvodnej hodnoty. Tento proces sa opakuje pokiaľ sa v obraze vyskytuje viac ako maximálny počet regiónov. Takéto prahovanie je nutné pre optimalizáciu výkonu algoritmu, pretože v prípadoch keď obraz obsahuje veľké množstvo regiónov, takýto proces znižuje výpočtovú náročnosť algoritmu. Pixely s valídnou honotou sa rozdelia na regióny podľa spojitosti a podobnosti štandardným spôsobom. Pripomeňme, že v tomto obraze sa spočítali hodnoty posunu v oboch smeroch aritmeticky do jednej hodnotiacej konštanty (pre každý pixel obrazu), ktorá už nereprezentuje smer posunu daného obrazového pixelu, ale iba hodnotí celkový posun obrazového pixelu. Takto získané regióny budeme vyhodnocovať a spájať podľa pôvodných výsledkov metódy Horn-Schunck. Vďaka využitiu pôvodných vektorov z výsledku metódy Horn-Schunck, vieme rozlíšiť

pohyb horizontálny aj vertikálny separátne. Pre všetky dvojice regiónov v obraze zisťujeme nasledovné charakteristiky:

1. **Rozdiel smerových vektorov v horizontálnom smere**
2. **Rozdiel smerových vektorov vo vertikálnom smere**
3. **Rozdiel vo vzdialenosti**

4.1.1.1 Rozdiel smerových vektorov v horizontálnom smere

Charakteristika sa vypočítava zo smerových horizontálnych vektorov metódy Horn-Schunck. Pre každý región sa vypočíta maximálna hodnota z indexov daného regiónu. Následne sa za hodnotu chrakteristiky považuje absolútna hodnota rozdielu týchto hodnôt pre každý región.

$$H_A = max(HS(i_A)) \tag{4.1}$$

$$H_B = max(HS(i_B)) \tag{4.2}$$

$$R_H = abs(H_A - H_B) \tag{4.3}$$

Kde A, B reprezentujú všetky dvojice regiónov, ktoré sa nachádzajú v obraze. V_a, V_b je maximálna hodnota horizontálnych smerových vektorov z výsledku Horn-Schunck algoritmu pre všetky oblasti patriace danému regiónu. R_H je výsledná hodnota charakteristiky.

4.1.1.2 Rozdiel smerových vektorov vo vertikálnom smere

Charakteristika sa vypočítava zo smerových vertikálnych vektorov metódy Horn-Schunck. Pre každý región sa vypočíta maximálna hodnota z indexov daného regiónu. Následne sa za hodnotu chrakteristiky považuje absolútna hodnota rozdielu týchto hodnôt.

$$H_A = max(HS(i_A)) \tag{4.4}$$

$$H_B = max(HS(i_B)) \tag{4.5}$$

$$R_V = abs(H_A - H_B) \tag{4.6}$$

Kde A, B reprezentujú všetky dvojice regiónov, ktoré sa nachádzajú v obraze. V_a, V_b je maximálna hodnota vertikálnych smerových vektorov z výsledku Horn-Schunck algoritmu pre všetky oblasti patriace danému regiónu. R_V je výsledná hodnota charakteristiky.

4.1.1.3 Rozdiel vo vzdialenosti

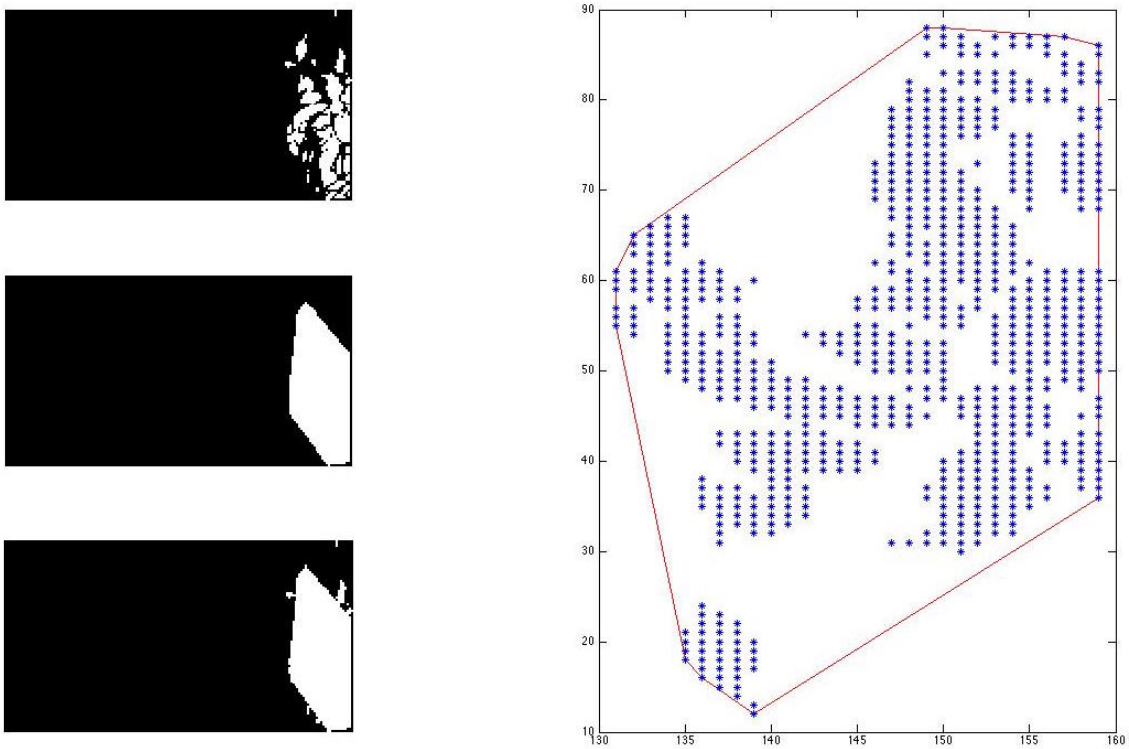
Chrakteristika sa vypočítava ako minimálna hodnota vzdialenosti medzi dvojicou regiónov. Hodnota je počítaná euklidovskou metódou.

forall the *rohA ako každý extrém regiónu A* **do**
  forall the *rohB ako každý extrém regiónu b* **do**
    vzdialenosť = sqrt((corner2(1,1)-rohB(1,1))^2 + $(rohB(1,2) - rohA(1,2))^2$)
  end
end

Algorithm 1: Výpočet minimálnej vzdialenosti euklidovskou metódou

4.1.1.4 Spájanie regiónov

Po výpočte všetkých troch charakteristík spojíme všetky dvojice regiónov, pre ktoré sú všetky chrakteristiky nižšie ako zadefinovaná konštanta. Regióny spájame pomocou konvexného obalu zjednotenia bodov ležiacich v oboch regiónoch.



Obr. 4.1: Vizualizácia spojenia regiónov pomocou konvexného obalu

4.1.2 Statické príznaky videa

Pri videách, kde sa pohybuje celá scéna (kamera je v pohybe) nedávajú dynamické príznaky dobré výsledky, keďže logicky označia celú scénu alebo väčšinovú časť scény za výrazne salientnú. Preto je vhodné dynamické príznaky kombinovať s klasickými modelmi pozornosti, ktoré síce zanedbajú postupnosť obrazov, ale nezlyhajú ako dynamické príznaky. Pre extrakciu statických obrazov sme zvolili metódu založnú na spektrálnych reziduách[14]. Vďaka svojmu princípu potláčania štatisticky opakujúcich sa predmetov na scéne, sa dá predpokladať vhodné doplnenie statických objektov, ktoré môžu zaujať pozornosť na videu, ak zlyhávajú dynamické príznaky.

4.1.3 Výsledné spojenie príznakov

Spájanie dynamických a statických príznakov bude prebiehať pomocou sčítania oboch máp, pričom vždy sa použijú v určitom pomere. Výpočet pomeru bude určovať pomer výskytu salientných pixelov v mape dynamických príznakov. Vzťah možno vyjadriť nasledovne:

$$pomer = \sum_{n=0}^{Pix_{count}} P_D(n) > 0 / Pix_{count} \tag{4.7}$$

Kde P_D reprezentuje mapu dynamických príznakov a Pix_{count} je počet všetkých pixelov, ktoré obraz obsahuje.
Ak je vysoký výskyt salientých pixelov, potrebujeme utlmiť zobrazovanie tejto časti príznakov a prioritizovať zobrazovanie statických príznakov, preto zmiešavacia funkcia vyzerá nasledovne:

$$Výsledok = (P_D * (1 - pomer)) + (P_S * pomer) \tag{4.8}$$

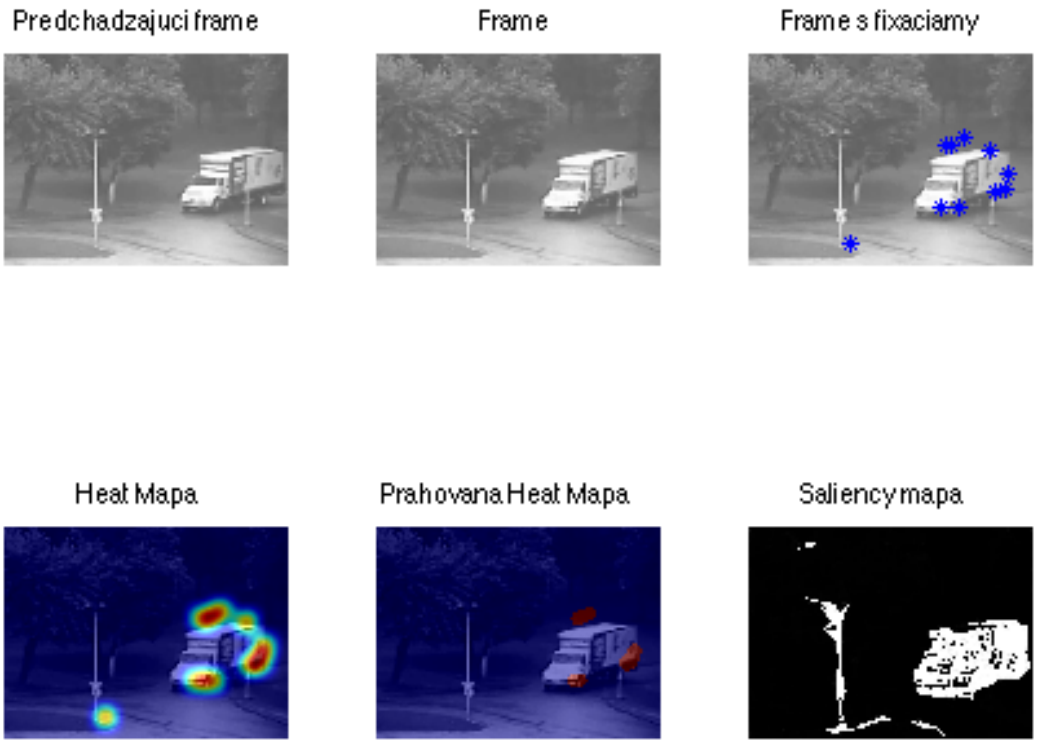
Kde P_D reprezentuje mapu dynamických príznakov a P_S mapu statických príznakov.
V prípade, že algoritmus nedokáže detekovať žiadny pohyb na scéne, bol by model pozornosti prázdny. Preto v prípade, keď je vyššie spomínaný pomer dynamických pixelov extrémne nízky, použijeme ako výstup algoritmu iba statické príznaky. Naopak v prípade, ak kamera je v pohybe Horn-Schunck algoritmus označí ako pohybujúcu sa väčšinovú oblasť obrazu, v tom prípade je potrebné utlmiť dynamické príznaky obrazu a do popredia vystupujú statické príznaky.

4.1.4 Zdrojové kódy modelu

Zdrojový kód obsahuje jednu metódu, ktorá prijíma na vstupe vždy dva parametre. Prvý parameter je aktuálny frame videa a druhý parameter je frame videa určený na extrakciu dynamických príznakov videa pomocou diferencie vzhľadom na prvý obrazový frame. Tieto dva obrazové vstupy nemusia byť nutne po sebe idúce. Je na používateľovi, či použije model serializovane na každý frame videa, alebo zvolí vlastnú implementáciu keyframingu (napríklad kvôli časovej náročnosti algoritmu). Algoritmus je schopný procesovať farebné aj čiernobiele obrazové vstupy. V prílohe je možné nájsť dve implementácie a to implementáciu modelu pre aplikáciu na porovnávanie modelov načítavajúcu kažké dva po sebe idúce obrazové framy. Druhá implementácia načítava na vstupe priamo video a na výstup dá video s korešpondujúcim videom významných oblastí. Táto implementácia je určená na použitie mimo aplikácie na testovanie. Obidve implementácie sú dostupné v prílohe na CD alebo voľne dostupné na internete.

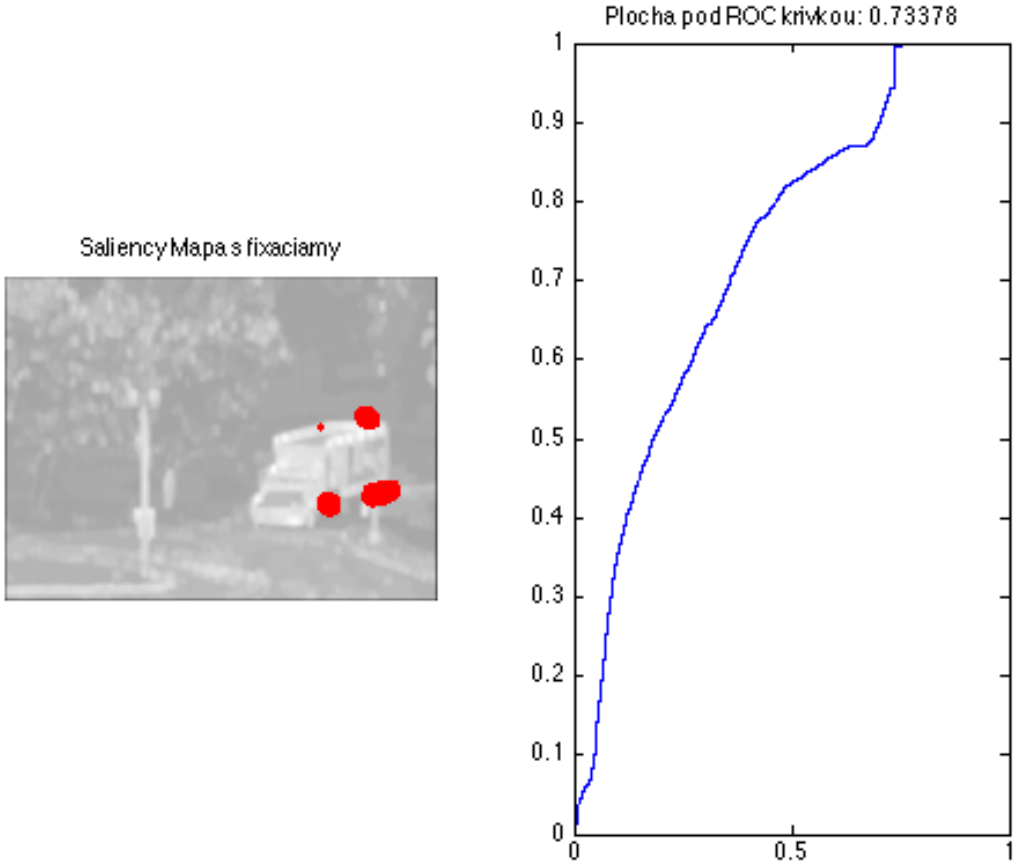
4.1.5 Ukážky výsledkov

V tejto sekcii budem prezentovať výsledky konkrétnych prípadov videí (framov) zachytené počas testovania a validácie. Uvediem po sebe nasledujúce originálne framy videa a vizualizáciu fixácií. Ako príklad uvediem frame 53 z videa č. 23 z datasetu ASCMN[24].



Obr. 4.2: Porovnanie výstupu mapy pozornosti a reálnych dát

Grafy sú generované pomocou pozmeneného ukážkového scriptu distribuovaného spolu s datasetom ASCMN[24] a porovnávané s výsledkom navrhovaného modelu pomocou grafu 4.2. V prvom riadku vidíme ako prvý pôvodný frame (n-1) nasledovaný testovaným framom a posledný obrázok zobrazuje fixácie. V druhom riadku uvádzam postupne heat mapu pre daný frame, prahovanú heat mapu a ako poslednú výslednú saliency mapu.



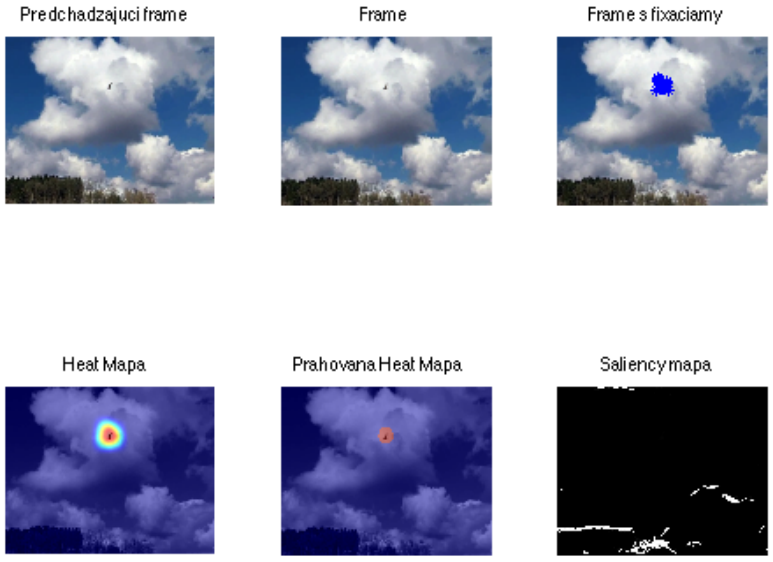
Obr. 4.3: Vizualizácie metriky AUC-Judd pomocou kódu zverejneného v mit saliency benchmark[5]

Z grafov 4.2 vidno koreláciu dát, čo taktiež potvrdzuje metrika vypočítaná na danom frame na grafe 4.3 reprezentujúca AUC-Judd[16] krivku.

4.1.5.1 Problémové typy videí

V tejto sekcii uvediem typové video s rovnakou analýzou ako je uvedená vyššie, iba na typ videa bude nevhodný navhovaný model. Typovo možno videá označiť ako videá s fixovanou kamerou v pohybe. Kamera je fixovaná k sledovanému objektu na scéne z čoho vyplýva, že pozadie obrazu (okolie sledovaného objektu) je pre náš algoritmus v pohybe aj keď reálne sa hýbe kamera a preto je pozadie obrazu považované za významný objekt. Zároveň sledovaný objekt (často aj objekt v pozornosti používateľov) sa vizuálne nepohybuje z dôvodu, že jeho

pohyb je kompenzovaný fixáciou kamery a preto ho navrhovaný algoritmus považuje za vizuálne nevýznamný. Ako príklad uvediem frame č. 80 z videa č. 10 z datasetu coutrot 1.

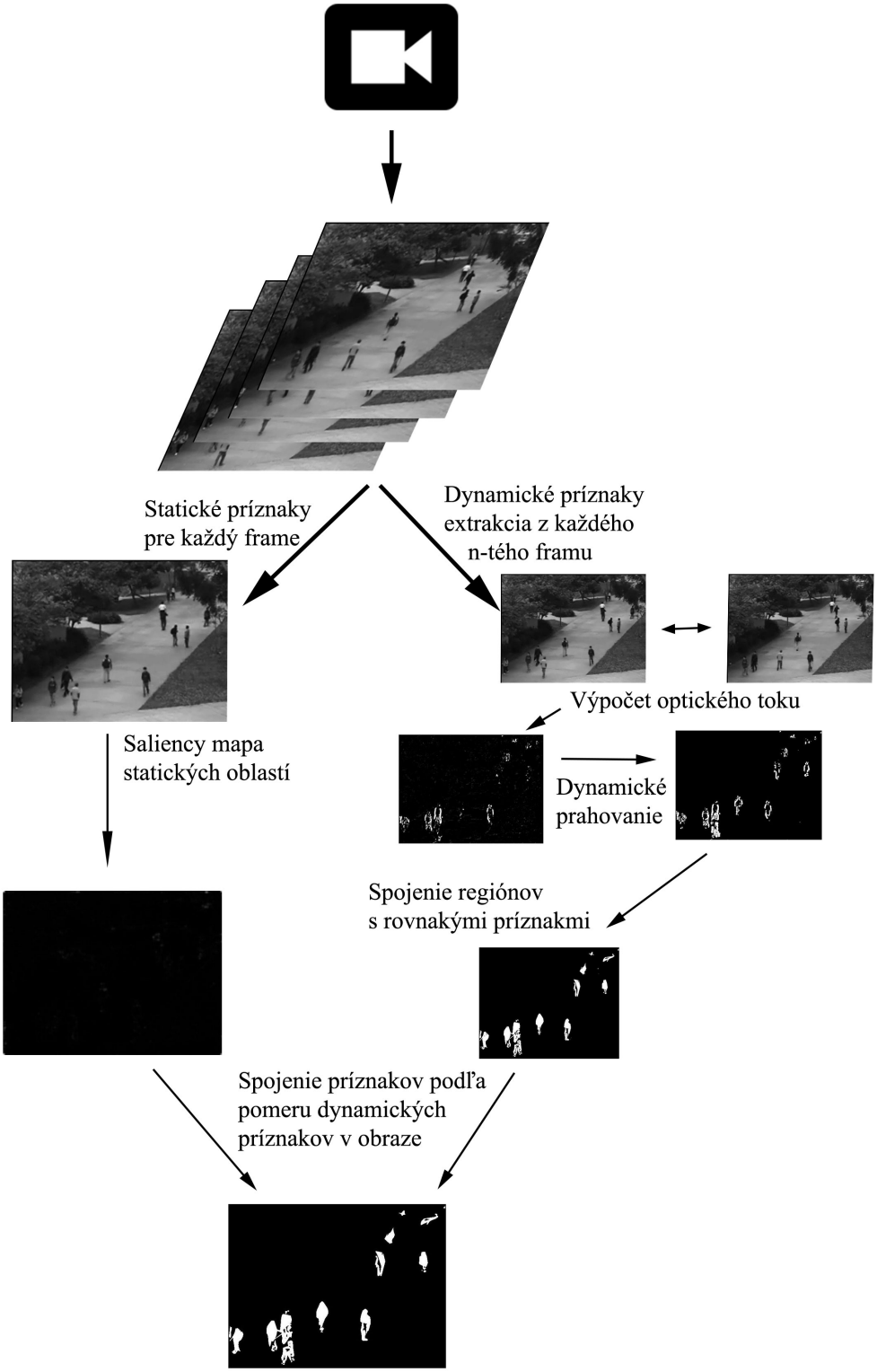


Obr. 4.4: Porovnanie výstupu mapy pozornosti a reálnych dát

Už podľa grafu 4.1.5.1 nemá výsledok navrhovaného modelu žiadnu koreláciu s reálnymi dátami. Preto nebudeme ďalej uvádzať žiadnu metriku. Návrh riešenia pre tieto prípady bude analyzovaný v sekcii 4.4.

4.1.6 Pipeline metódy

Grafický popis metódy obsahujúci graf zostavenia mapy pozornosti.



Obr. 4.5: Ucelená vizualizácia algoritmu

4.2 Implementácia riešenia

Implementácia vyššie uvedeného algoritmu je implementovaná ako modul pre aplikáciu na porovnávanie a automatickú validáciu výsledkov. Aplikácia na porovnávanie je tiež implementovaná v prostredí matlab.

4.2.1 Aplikácia na porovnávanie a automatickú validáciu

Sekundárnym prínosom práce je vytvorenie aplikácie pre zjednodušenie budúcej práce pri prototypovaní nových modelov pozornosti a následné uľahčenie validačného procesu pre potencionálnych vývojárov.
Základná functionalita:

1. **Oddelenie logiky testovania a logiky samotného modelu**
2. **Simultálne sledovanie videa z viacerých modelov**
3. **Automatická validácia modelu**
4. **Vizualizácia výsledkov validácie**

4.2.1.1 Oddelenie logiky testovnia a logiky samotného modelu

V aplikácii na testovanie je možné pridávať ľubovolné modely, pre ktoré je dostupná implmentácia v jazyku matlab. Pre iné jazyky je potrebné doprogramovať wrapper, ktorý spustí daný jazyk a vypočíta mapu pozornosti. Ukážkový wrapper je súčasťou aplikácie. Na pridanie nového modelu je potrebné pridať wrapper do zložky 'models'. Knižnice vyžadované modelom je potrebné skopírovať do ľubovolnej podzložky tohoto priečinka. Pri spustení apikácie sa načítajú všetky modely aj knične uložené v príslušných podzložkách.

4.2.1.2 Simultálne sledovanie videa z viacerých modelov

Pre rýchle prototypovanie je vhodné pozorovať rovnaké video pri rôznych úpravách. Táto funkcionalita je dostupná pre každý model s vygenerovanými mapami pozornosti na zvolenom videu.

4.2.1.3 Automatická validácia modelu

Validovanie výsledkov je nutnou súčasťou každého modelu pozornosti, preto aplikácia ponúka automatizovaný spôsob ako zvalidovať výsledky na vybraných referenčných datasetoch. Validácia tvorí pre kažké video perzistentný súbor obsahujúci tri metriky: AUC-Judd, KL-Div, a NSS. Vyššie spomenuté metriky sa vypočítavajú pre každý

frame videa. Validácia prebieha paralelne pre všetky videá zvoleného datasetu. Vytvorené súbory sú perzistentné z dôvodu dlhého výpočtového času a ukladajú sa do priečinku results a podložky podľa názvu testovaného datasetu v tvare $názovModelu.názovDatasetuČísloVidea.mat$. Formát súboru obsahuje tri premenné s názvami: AUROC_score, KLDIV_score, NSS_score. Každá premenná obsahuje pole podľa dĺžky videa (počet frameov) a s hodnotami danej metriky. Aplikácia aktuálne podporuje dva datasety a to: ASCMN[24], coutrotove testovacie datasety 1[8] a 2[7]. Tieto datasety sú voľne dostupné a súčasťou aplikácie je programový kód slúžiaci na načítanie a validovanie výsledkov (samotné vstupné videá a fixácie je potrebné stiahnuť zo stránky autorov). Dataset ASCMN[24] je poskytovaný autormi aj s testovacím algoritmom na výpočet vyššie uvedených metrík. Testovací algoritmus je do aplikácie na testovanie iba pozmenený pre načítanie ľubovolného modelu a prispôsobený na paralelný výpočet všetkých videí naraz. Pre Coutrot datasety aplikácia na testovanie obsahuje upravenú verziu validačného algoritmu z datasetu ASCMN.

4.2.1.4 Vizualizácia výsledkov validácie

Pre analýzu výsledkov validácie dokáže aplikácia prehľadne vizualizovať všetky dáta získané testovaním. Vizualizácie sú súčasťou validácie v ďalších kapitolách.

4.2.2 Implementácia modelu

Implementácia nového modelu pozornosti je jednoduchá. Pre integrovanie ľubovolného modelu je možné použiť vzorovú implementáciu, ktorá je dostupná v prílohách.

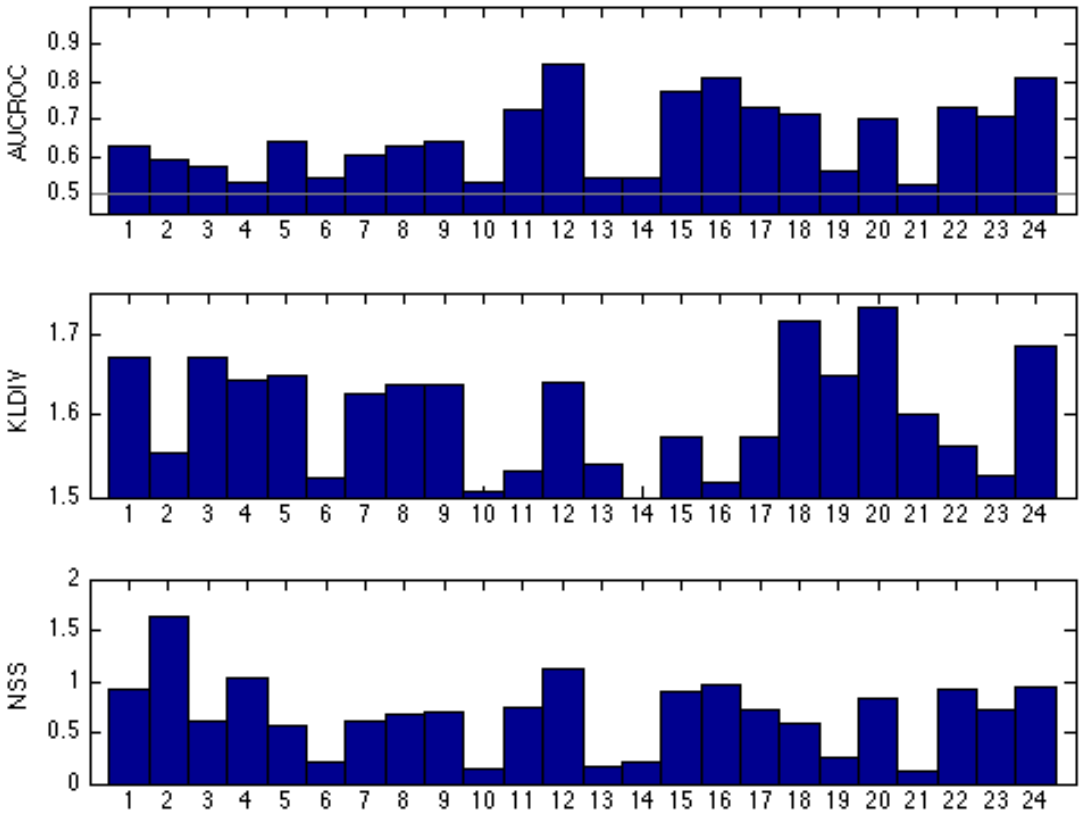
4.3 Validácia výsledkov

Validácia vyššie spomínaného modelu prebiehala pomocou automatického testovania v aplikácii na testovanie. Prezentovať budem výsledky z nasledujúch datasetov: **ASCMN[24]**, **Coutrot 1[8]**, **Coutrot 2[7]**. Výsledky budem hodnotiť pomocou nasledujúcich metrík: **AUCROC[22]**, **KLDIV[22]**, **NSS[22]**. V nasledujúcich sekciách budem prezentovať výsledky validácie pre navrhovaný model a ďalej vyhodnocovať vytvorený benchmark.

4.3.1 Analýza výsledkov

V tejto sekcii prezentujem výsledky všetkých datasetov vzhľadom na navrhovaný model. Výsledky sú vizualizované pomocou charakteristiky vzniknutej zo strednej hodnoty framov jednotlivých videí. Každá metrika bude vyhodnocovaná samostane. Ako prvý budeme analyzovať dataset ASCMN[24] a následne oba Coutrotove datasety[8] [7].

4.3.1.1 ASCMN



Obr. 4.6: Vizualizácia všetkých testovaných metrík pre dataset ASCMN[24] pre jednotlivé videá

4.3.1.2 ASCMN - AUCROC

Ideálna hodnota tejto metriky je 1, čo reálne značí 100% úspešnosť. Náhodná mapa pozornosti má hodnotu 0.5, z toho dôvodu aby sme dokázali správnosť nového modelu potrebujeme dokázať, že model má hodnotu AUC v intervale [0.5,1]. Z grafu 4.6 vyplýva, že všetky videá spĺňajú vyššie uvedenú podmienku. Výsledná priemerná hodnota pre AUC je 0.6515. Táto hodnota už priamo dokazuje hypotézu, keďže ho tvorí stredná hodnota všekých testovaných videí. Hodnota vyššia ako 0.5 vyhodnocuje náš model ako korelujúci s reálne nameranými dátami na používateľoch zúčastnených sa na tvorbe tohoto datasetu.

4.3.1.3 ASCMN - KLDIV

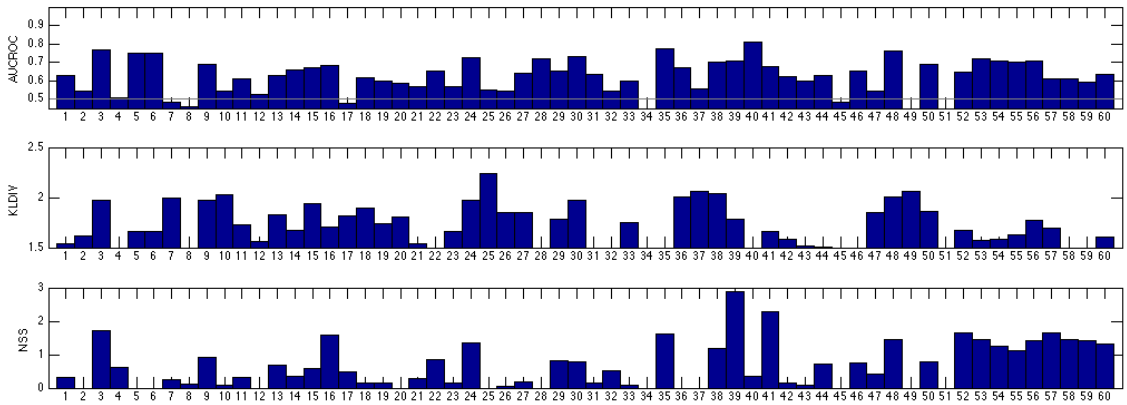
Ideálna hodnota tejto metriky je 0, čo reálne značí že saliency mapa je totožná s ground truth mapou. Výsledné hodnoty pre všetky videá sa podľa grafu 4.6 pohybujú v intervale [1.5,1.7]. Hodnoty menšie ako 2 znamenajú tiež koreláciu s reálnymi dátami. Výsledná hodnota KLDIV je 1.6027 a je nižsia ako 2, preto aj metrika KLDIV úspešne validuje náš model v rámci datasetu.

4.3.1.4 ASCMN - NSS

Výsledná hodnota NSS je 0.6809. Hodnota NSS metriky indentifikuje, že hodnoty v oblastiach fixácií sú minimálne nepresné oproti normalizovaným fixáciám a model by mal zmeniť hodnoty v oblastiach fixácií.

4.3.1.5 Coutrot 1

V datasete Coutrot 1 boli z testovania vyňaté videá číslo: 34, 39, 51, 5 z dôvodu chybného zdrojového videa alebo fixácií zverejnených autormi, alebo nevalídnymi výsledkami v niektorej z vypočítavaných metrík.



Obr. 4.7: Vizualizácia všetkých testovaných metrík pre dataset Coutrot 1[8] pre jednotlivé videá

4.3.1.6 Coutrot 1 - AUCROC

Výsledná priemerná hodnota pre AUC je 0.6086. Takáto hodnota vyhodnocuje náš model ako korelujúci s reálne nameranými dátami na používateľoch zúčastnených sa na tvorbe tohoto datasetu.

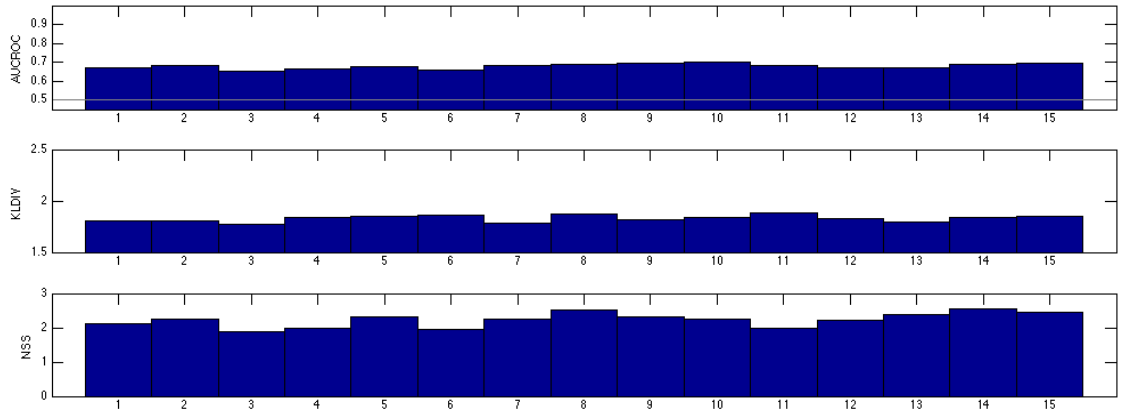
4.3.1.7 Coutrot 1 - KLDIV

Výsledná hodnota KLDIV je 1.6280. Hodnota KLDIV je nižsia ako 2, preto aj metrika KLDIV úspešne validuje náš model v rámci datasetu.

4.3.1.8 Coutrot 1 - NSS

Výsledná hodnota NSS je 0.6277. Hodnota NSS metriky identifikuje, že hodnoty v oblastiach fixácií korelujú s normalizovanými fixáciami používateľov, ale na tomto type videí má model ešte rezervu.

4.3.1.9 Coutrot 2



Obr. 4.8: Vizualizácia všetkých testovaných metrík pre dataset Coutrot 2[7] pre jednotlivé videá

4.3.1.10 Coutrot 2 - AUCROC

Výsledná priemerná hodnota pre AUC je 0.6782.

4.3.1.11 Coutrot 2 - KLDIV

Výsledná hodnota KLDIV je 1.8288 a je nižšia ako 2, preto aj metrika KLDIV úspešne validuje náš model v rámci datasetu.

4.3.1.12 Coutrot 2 - NSS

Výsledná hodnota NSS je 2.2313. Vysoká hodnota tejto metriky na tomto type videí je logickým dôvodom typu videa. V tomto datasete bola pozornosť podľa výstupných meraní používateľov upriamená na málo oblastí v obraze naraz (v tomto prípade tvár alebo ruky účinkujúcich) a týmto oblastiam pridáva vysokú hodnotu. Navrhovaná métoda postupuje podobne a to tak, že vyčlení pohybujúce sa oblasti (pre takýto typ videa je pohybúcich sa oblastí málo) a keď že netvoria majoritnú oblasť videa utlačí statické príznaky, čoho dôsledkom je vysoké hodnotenie týchto obastí, čo zodpovedá vysokej hodnote porovnávajúcej sa v metrike.

4.3.1.13 Zhrnutie hodnotenia

Pri celkovom hodnotení bude pre nás smerodajná hlavne metrika AUCROC. Priemer pre všetky tri testované datasety má hodnotu 0.6277, z čoho vyplýva úspešná validácia korelácie s nameranými dátami. Metrika KLDIV dosahuje priemernú hodnotu pre všeky datasety 1.6438. Metrika KLDIV poukazje na fakt, že metóda označuje viacej oblastí ako významných, aj keď neboli namerané. Metrika NSS dosahuje priemernú hodnotu pre všeky

datasety 0.5487. Táto metrika validuje, že aj hodnoty v oblastich fixácií korelujú s dátami. V tejto metrike však záleží od typu videa viacej ako u iných metrík.

4.3.2 Porovnávanie s konkurenčnými modelmi pozornosti

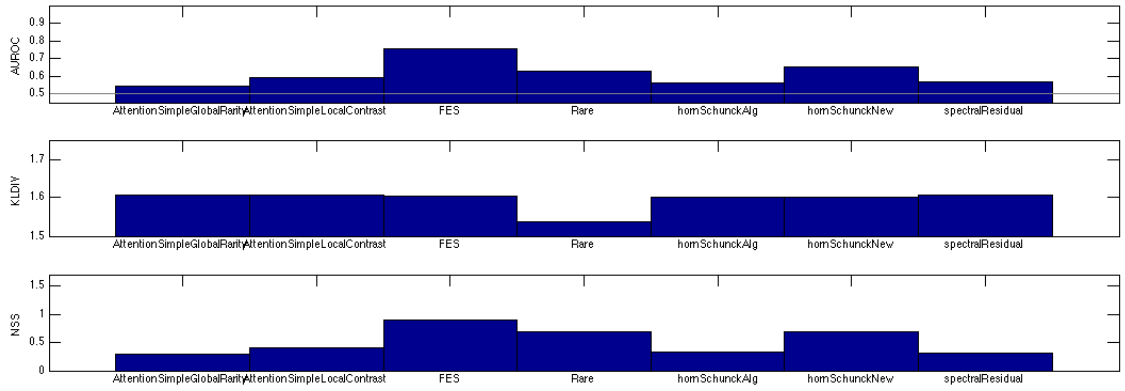
Porovnanie s konkurenciou nám poskytne ďalšie relevantné poznatky. V tejto časti sa budem snažiť dokázať, že efektivita nášho nového modelu je vyššia ako pôvodného horn-struck algoritmu[2] používaného na extrakciu dynamickej zložky príznakov. Zároveň sa budem snažiť o dôkaz, že algoritmus je efektívnejší aj ako samotná statická zložka, ktorá je vypočítavaná pomocou modelu Spektralnych rezidual[14]. Pre tento účel som uskutočnil benchmark obsahujúci nasledovné algoritmy:

1. **AttentionSimpleGlobalRarity[19]**
2. **AttentionSimpleLocalContrast[19]**
3. **FES[21]**
4. **RARE[23]**
5. **Horn-struck[2]**
6. **Lucas-Kanade[4]**
7. **Spektralne rezidua[14]**

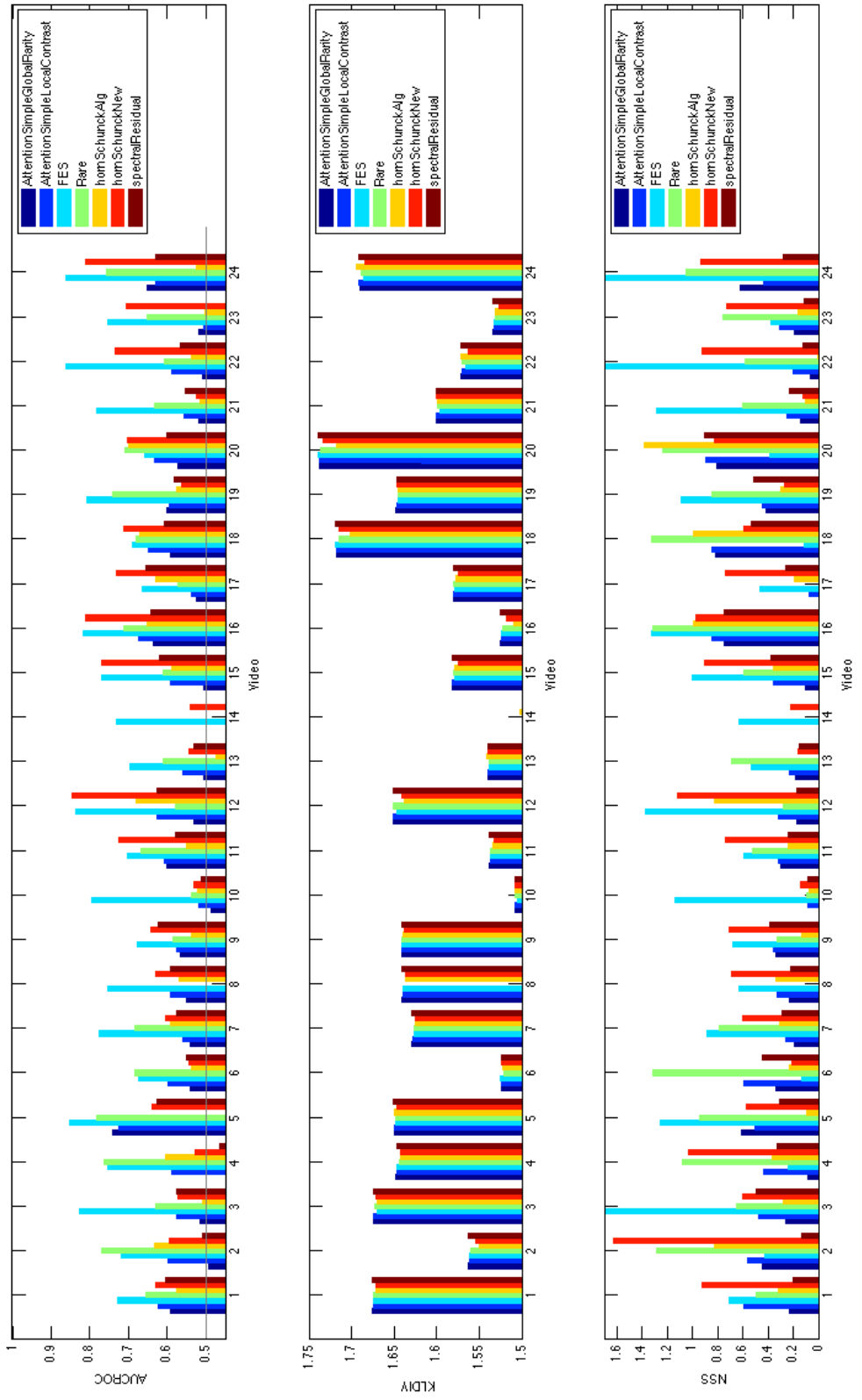
na všetkých datasetoch. Ku každému datasetu uvediem dve súhrnné štatistiky. Prvá vyjadruje porovnanie priemernej hodnoty každej z metrík pre kažkdé video samostatne. Druhá obsahuje priemerné hodnoty všetkých videí, pre všetky testované medódy.

4.3.2.1 ASCMN

V ASCMN datasete bolo testovanie uskutočnené na všetkých videách a na všetkých vyššie uvedených modeloch.



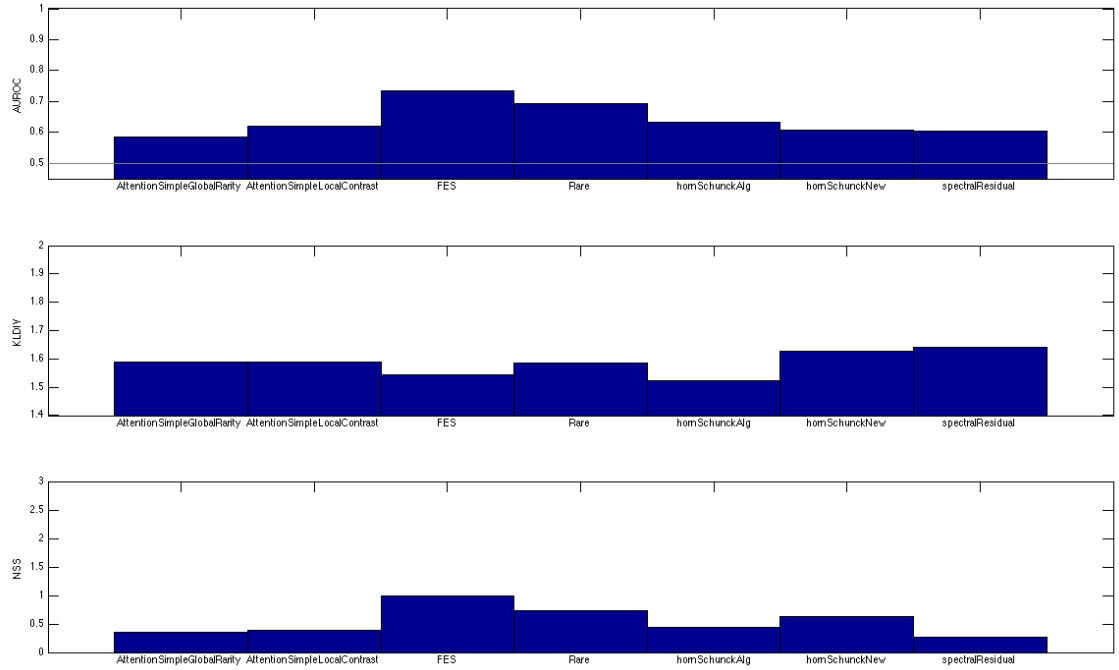
Obr. 4.9: Vizualizácia porovnania pre dataset ASCMN[24]



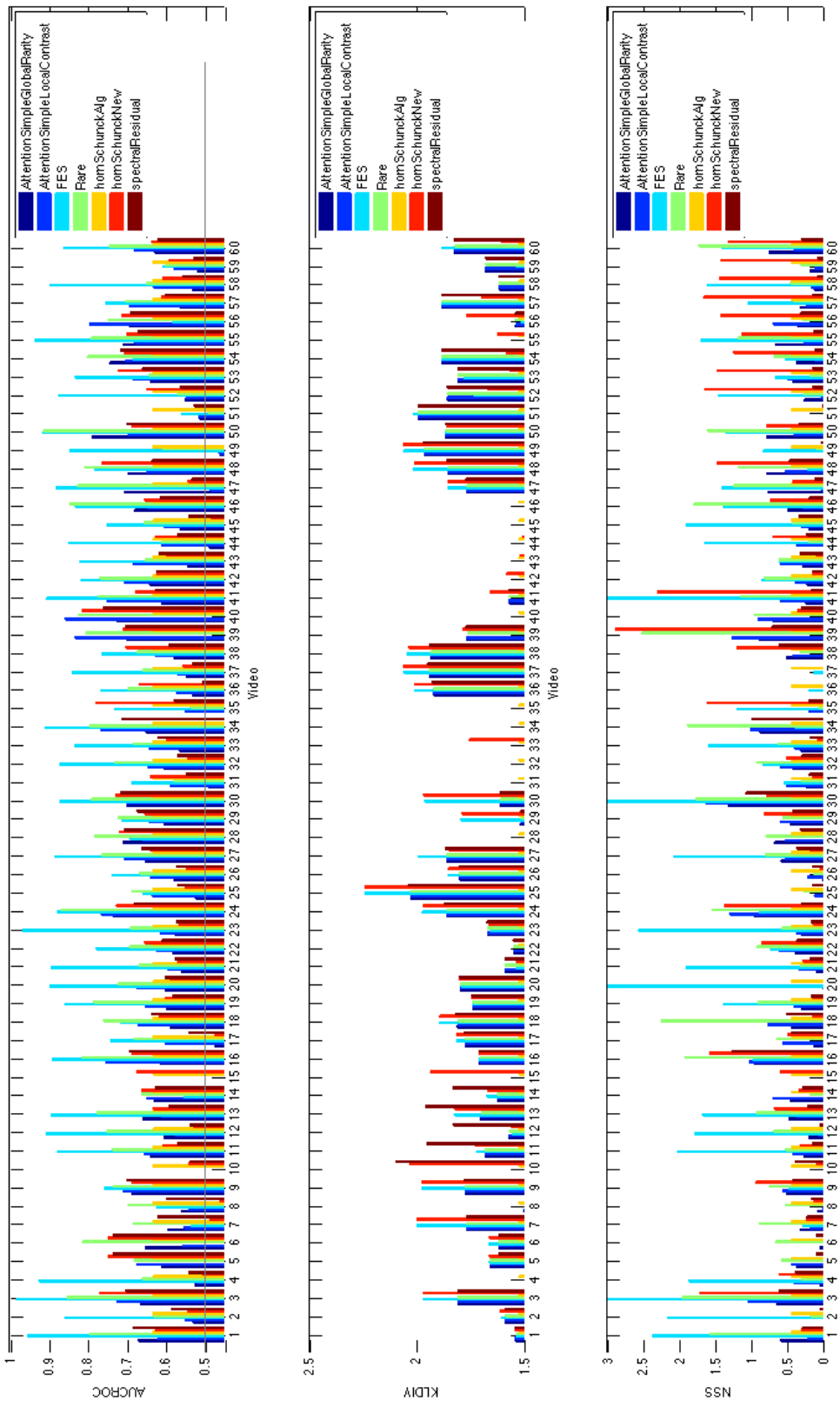
Obr. 4.10: Vizualizácia porovnania pre dataset ASCMN[24]

4.3.2.2 Coutrot 1

V datasete Coutrot 1 boli z testovania vyňaté niektoré videá z dôvodu chybového spracovania v jednom alebo vo viacerých testovaných modeloch.

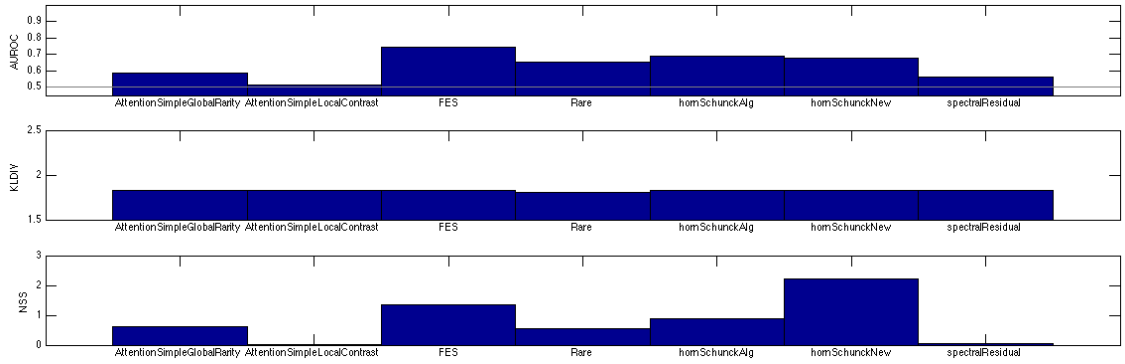


Obr. 4.11: Vizualizácia porovnania pre dataset Coutrot 1[8]



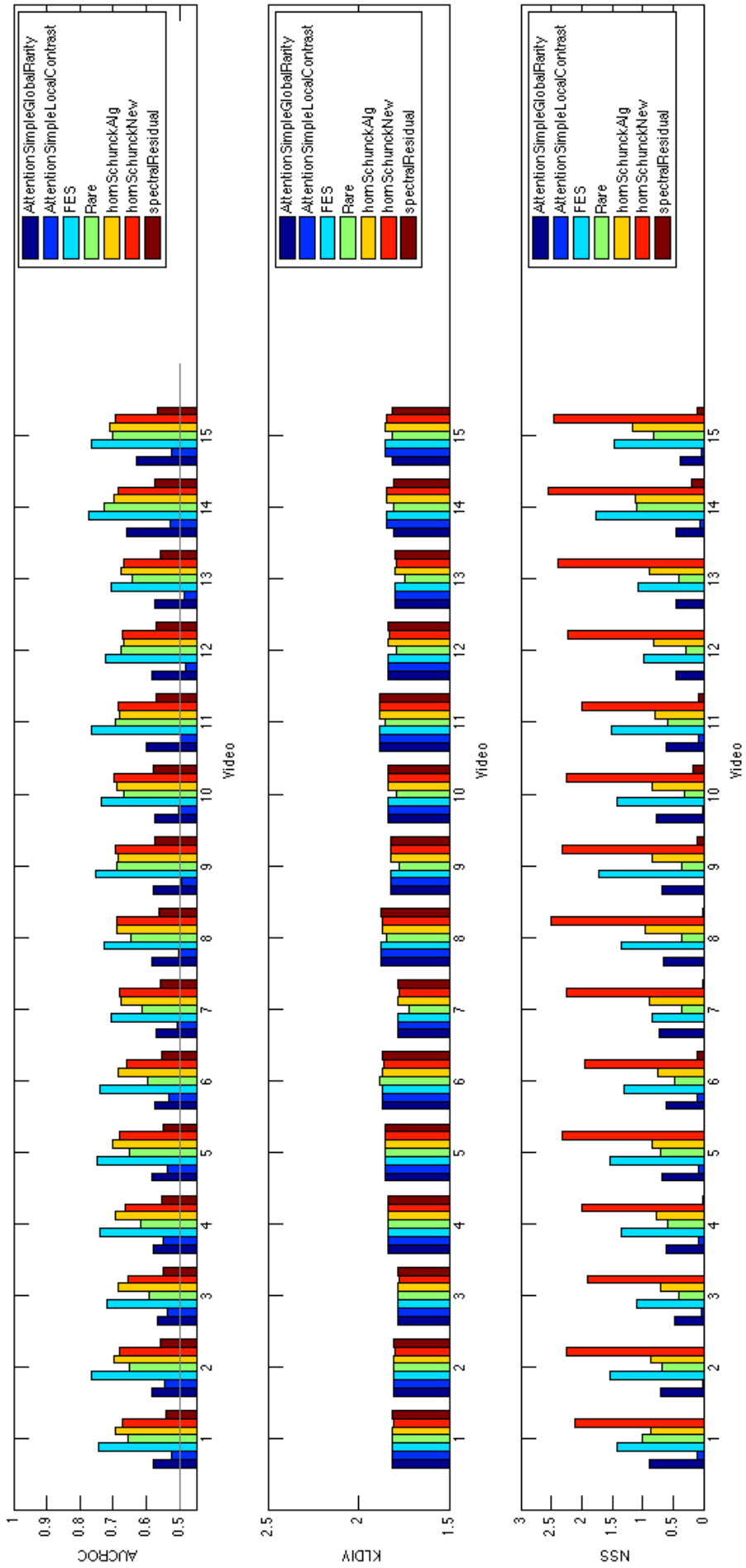
Obr. 4.12: Vizualizácia porovnania pre dataset Coutrot 1[8]

4.3.2.3 Coutrot 2



Obr. 4.13: Vizualizácia porovnania pre dataset Coutrot 2[7]

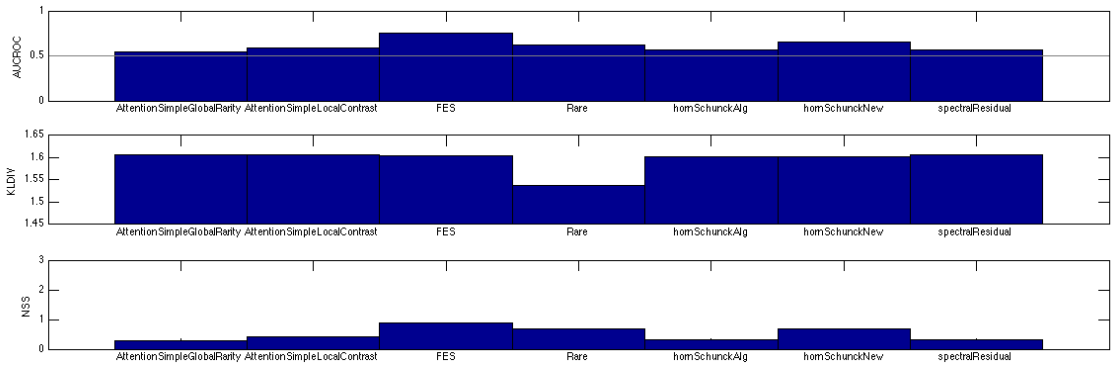
V datasete Coutrot 2 bolo testovanie uskutočnené na všetkých videách a na všetkých vyššie uvedených modeloch.



Obr. 4.14: Vizualizácia porovnania pre dataset Coutrot 2[7]

4.3.2.4 Zhrnutie benchmarku

Celkové hodnotenie benchmarku znázorním porovnanním priemerov všetkých datasetov spolu.



Obr. 4.15: Vizualizácia porovnania pre všetky datasety

4.3.3 Zhrnutie validácie

Všetky metriky potvrdzujú tvrdenie, že narvhovaný model má značnú koreláciu k skutočným dátam nameraným na reálnych používateľoch. Zároveň bola validácia vykonaná na typovo rozdielnych videách, keďže obsahujú pohybujúcu sa kameru aj statickú pozíciu kamery, konverzačné scény aj scény s prírodnými motívami. Súčastne na základe vypracovaného benchmarku môžeme konštatovať, že nová metóda je efektívnejšia vo väčšine prípadov ako dve základné metódy použité na získanie dynamických aj statických príznakov, pričom vo všetkých videách dosahuje lepšie výsledky ako aspoň jeden zo základných algoritmov.

4.4 Diskusia

Možnosť na zlepšenie algoritmu vidieť v celom benchmarku, kde model Rare[23] dosahoval výrazne lepšie výsledky aj napriek používaniu iba statických príznakov. Preto použitie algoritmu Rare[23] by viedlo k zlepšeniu aktuálnych výsledkov.
Ďalšia možnosť ako vylepšiť je v rýchlosti spracovania, ktorá nie je použiteľná na realtime spracovanie videa v štandardnej kvalite videa. Algoritmus je svojou časovou náročnosťou vhodný na spracovanie videí v nízkej obrazovej kvalite. Pri vysokej obrazovej kvalite však algoritmus nevykazoval vyššiu efektivitu (otestované na datasete savam[12], ktorý poskytuje videá vo vysokej kvalite), výpočet ale trval výrazne dlhšie ako v porovnaní s videom s nízkou obrazovou kvalitou.
Ďalšiu možnosť pre zlepšenie identifikuje validácia datasetu Coutrot 2[7], kde pôvodný algotritmus horn-struck dosiahol hodnotenie porovnateľné s navrhovaným modelom. Takéto výsledky sú spôsobené výberom rozostupu frameov, podľa ktorých sa počíta dynamická zložka. Keďže rozpätie bolo zvolené na každú dvojicu framov, pri týchto videách sa často

stávalo, že algoritmus detekoval iba minimálny pohyb. Čo bolo považované za šum a z toho dôvodu bola dymamická zložka potlačená alebo úplne eliminovaná (čo bolo v tomto prípade chybné). Riešením by bolo porovnávanie viacej framov a následná extrakcia pohybu všetkých dvojíc spolu. Týmto postupom by už nebol pohyb považovaný za šum a dynamická zložka by nebola eliminovaná. Ďalšou možnosťou je použiť inú formu rozloženia videa na dynamicky sa meniace keyframy, medzi ktorými sa bude vypočítavať dynamická zložka mapy, zatiaľ čo statická sa bude generovať počas kažkého framu.

5. Záver

Cieľom diplomovej práce bolo preskúmať možnoti extrakcie príznakov významných oblastí pre videá pomocou príznakov, ktoré nemožno extrahovať z čisto obrazovej informácie a navrhnúť metódu používajúcu takéto príznaky. Ďalším cieľom bolo vypracovať validáciu navrhovaného modelu pomocou štandartných metrík používaných pri hodnotení modelov na detekciu významných oblastí.

V práci som navrhol a zvalidoval metódu, ktorá využíva kombináciu štandardných metód pre detekciu významných oblastí vo videu. Naviac bol vypracovaný benchmark, ktorý obsahuje šesť iných modelov/algoritmov, za účelom porovnania a analýzy výsledkov navrhovaného modelu. Sekundárnym prínosom práce je ucelená aplikácia, ktorá má potenciál výrazne zjednodušiť prototypovanie, testovanie a validovanie podobných modelov pre budúcich študentov. Aplikácia je navrhnutá pre výrazné ušetrenie práce, vďaka jednoduchému pridávaniu modelov pozornosti a možnosti ich vizuálneho porovnania. Taktiež je ňou možné validovať výsledky pomocou oficiálne publikovaných troch datasetov a troch štandardne používaných metrík. Validácia prebieha automaticky a dokáže z výsledkov generovať aj výsledky vo forme grafov používaných v tejto práci. Aplikácia je dostupná v elektronickej prílohe a voľne dostupná na internete.

Ďalší možný rozvoj je popísaný a analyzovaný v sekcii 4.4.

Zoznam použitej literatúry

[1] "A threshold selection method from gray-level histograms", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, s. 62–66, Jan. 1979, ISSN: 0018-9472. DOI: 10.1109/TSMC.1979.4310076.

[2] AL KANAWATHI, J. - MOKRI, S.S. - IBRAHIM, N. - HUSSAIN, A. - MUSTAFA, M.M. "Motion detection using horn schunck algorithm and implementation", in *Electrical Engineering and Informatics, 2009. ICEEI '09. International Conference on*, vol. 01 : Aug. 2009. S. 83–87. DOI: 10.1109/ICEEI.2009.5254812.

[3] AN, Kwang-Hwan - LEE, Minho - SHIN, Jang-Kyoo, "Saliency map model based on the edge images of natural scenes", in *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on*, vol. 1 : 2002. S. 1023–1027. DOI: 10.1109/IJCNN.2002.1005616.

[4] B.D. LUCAS, & Kanade. 1981. *An Interative Image Registration Technique with an Application to Stereo Vision*. [online]. : 1981. [cit. 8.4.2013]. Dostupné na internete: https://www.ri.cmu.edu/pub_files/pub3/lucas_bruce_d_1981_1/lucas_bruce_d_1981_1.pdf.

[5] BYLINSKII, Zoya - JUDD, Tilke - BORJI, Ali - ITTI, Laurent - DURAND, Frédo - OLIVA, Aude - TORRALBA, Antonio, *Mit saliency benchmark*.

[6] COUTROT, A. - GUYADER, N. "An audiovisual attention model for natural conversation scenes", in *Image Processing (ICIP), 2014 IEEE International Conference on* : Oct. 2014. S. 1100–1104. DOI: 10.1109/ICIP.2014.7025219.

[7] —, "An efficient audiovisual saliency model to predict eye positions when looking at conversations", in *Signal Processing Conference (EUSIPCO), 2015 23rd European* : Aug. 2015. S. 1531–1535. DOI: 10.1109/EUSIPCO.2015.7362640.

[8] —, "Toward the introduction of auditory information in dynamic visual attention models", in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)* : Jul. 2013. S. 1–4. DOI: 10.1109/WIAMIS.2013.6616164.

[9] COUTROT, Antoine - GUYADER, Nathalie, "How saliency, faces, and sound influence gaze in dynamic social scenesshort title??", *Journal of Vision*, vol. 14, no. 8, p. 5, 2014. DOI: 10.1167/14.8.5. eprint: /data/Journals/JOV/933549/i1534-7362-14-8-5.pdf. Dostupné na internete: +%20http://dx.doi.org/10.1167/14.8.5.

[10] COUTROT, Antoine - GUYADER, Nathalie - IONESCU, Gelu - CAPLIER, Alice, "Video viewing: do auditory salient events capture visual attention?", *annals of telecommunications-annales des télécommunications*, vol. 69, no. 1-2, s. 89–97, 2014.

[11] DUNCAN, K. - SARKAR, S. "Saliency in images and video: a brief survey", *Computer Vision, IET*, vol. 6, no. 6, s. 514–523, Nov. 2012, ISSN: 1751-9632. DOI: 10.1049/iet-cvi.2012.0032.

[12] GITMAN, Y. - EROFEEV, M. - VATOLIN, D. - ANDREY, B. - ALEXEY, F. "Semiautomatic visual-attention modeling and its application to video compression", in *Image Processing (ICIP), 2014 IEEE International Conference on* : Oct. 2014. S. 1105–1109. DOI: 10.1109/ICIP.2014.7025220.

[13] HORN, Berthold K.P. - SCHUNCK, Brian G. "Determining optical flow", Cambridge, MA, USA, Tech. Rep., 1980.

[14] HOU, Xiaodi - ZHANG, Liqing, "Saliency detection: a spectral residual approach", in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on* : Jun. 2007. S. 1–8. DOI: 10.1109/CVPR.2007.383267.

[15] ITTI, L. - KOCH, C. - NIEBUR, E. "A model of saliency-based visual attention for rapid scene analysis", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, s. 1254–1259, Nov. 1998, ISSN: 0162-8828. DOI: 10.1109/34.730558.

[16] JUDD, Tilke - DURAND, Frédo - TORRALBA, Antonio, "A benchmark of computational models of saliency to predict human fixations", in *MIT Technical Report* : 2012. Dostupné na internete: http://hdl.handle.net/1721.1/68590.

[17] KULLBACK, S. - LEIBLER, R. A. "On information and sufficiency", *Ann. Math. Statist.*, vol. 22, no. 1, s. 79–86, 1951.

[18] LI, Jia - TIAN, Yonghong - HUANG, Tiejun - GAO, Wen, "A dataset and evaluation methodology for visual saliency in video", in *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo*, ICME'09, New York, NY, USA : IEEE Press, 2009. S. 442–445, ISBN: 978-1-4244-4290-4. Dostupné na internete: http://dl.acm.org/citation.cfm?id=1698924.1699033.

[19] MANCAS, M. - MANCAS-THILLOU, C. - GOSSELIN, B. - MACQ, B. “A rarity-based visual attention map - application to texture description”, in *2006 International Conference on Image Processing* : Oct. 2006. S. 445–448. DOI: 10.1109/ICIP.2006.312489.

[20] PROF. ROBERT FISHER, Prof. James Crowley. 2005. *CAVIAR*. [online]. : 2005. [cit. 8.4.2013]. Dostupné na internete: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[21] REZAZADEGAN TAVAKOLI, Hamed - RAHTU, Esa - HEIKKILÄ, Janne, “Image analysis: 17th scandinavian conference, scia 2011, ystad, sweden, may 2011. proceedings”, in, Heyden, A. -Kahl, F., Eds. Berlin, Heidelberg : Springer Berlin Heidelberg, 2011. Ch. Fast and Efficient Saliency Detection Using Sparse Sampling and Kernel Density Estimation, s. 666–675, ISBN: 978-3-642-21227-7. DOI: 10.1007/978-3-642-21227-7_62. Dostupné na internete: http://dx.doi.org/10.1007/978-3-642-21227-7_62.

[22] RICHE, N. - DUVINAGE, M. - MANCAS, M. - GOSSELIN, B. - DUTOIT, T. “Saliency and human fixations: state-of-the-art and study of comparison metrics”, in *Computer Vision (ICCV), 2013 IEEE International Conference on* : Dec. 2013. S. 1153–1160. DOI: 10.1109/ICCV.2013.147.

[23] RICHE, N. - MANCAS, M. - GOSSELIN, B. - DUTOIT, T. “Rare: a new bottom-up saliency model”, in *Image Processing (ICIP), 2012 19th IEEE International Conference on* : Sep. 2012. S. 641–644. DOI: 10.1109/ICIP.2012.6466941.

[24] RICHE, Nicolas - MANCAS, Matei - CULIBRK, Dubravko - CRNOJEVIC, Vladimir - GOSSELIN, Bernard - DUTOIT, Thierry, “Computer vision – accv 2012: 11th asian conference on computer vision, daejeon, korea, november 5-9, 2012, revised selected papers, part iii”, in, Lee, K. M. - Matsushita, Y. - Rehg, J. M. -Hu, Z., Eds. Berlin, Heidelberg : Springer Berlin Heidelberg, 2013. Ch. Dynamic Saliency Models and Human Attention: A Comparative Study on Videos, s. 586–598, ISBN: 978-3-642-37431-9. DOI: 10.1007/978-3-642-37431-9_45. Dostupné na internete: http://dx.doi.org/10.1007/978-3-642-37431-9_45.

[25] SHARMA, P. - CHEIKH, F.A. - HARDEBERG, J.Y. “Face saliency in various human visual saliency models”, in *Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on* : Sep. 2009. S. 327–332. DOI: 10.1109/ISPA.2009.5297732.

[26] ŠIKUDOVÁ, E. - ČERNEKOVÁ, Z. - BENEŠOVÁ, W. - HALADOVÁ, Z. - KUČEROVÁ, J. 2014. *Počítačové videnie. Detekcia a rozpoznávanie objektov*, first : Wikina, Livornská 445, 109 00 Praha 10, 2014. .

[27] ZHANG, Lingyun - TONG, Matthew H. - MARKS, Tim K. - SHAN, Honghao - COTTRELL, Garrison W. “Sun: a bayesian framework for saliency using natural statistics”, *Journal of Vision*, vol. 8, no. 7, p. 32, 2008. DOI: 10.1167/8.7.32. eprint: /data/Journals/JOV/933536/jov-8-7-32.pdf. Dostupné na internete: +%20http://dx.doi.org/10.1167/8.7.32.

Prílohy

CD obsahujúce:

- Elektronickú verziu práce
- Zdrojový kód navrhovaného modelu
- Zdrojový kód aplikácie pre testovanie
- Jednoduchý manuál na používanie aplikácie pre testovanie