Projektu Č. VI20172020068

Nástroje a metody zpracování videa a obrazu pro zvýšení efektivity operací bezpečnostních a záchranných složek (VRASSEO)

Detekce chodců z výšky pomocí neuronových sítí

Technická zpráva

Kanich O., Drahanský M., Goldmann T.

Vysoké Učení Technické v Brně
Fakulta Informačních Technologií
Božetěchova 1
612 66 Brno, Česká republika

Prosinec 2019

Obsah

Abstrakt

Zpráva se zabývá popisem detekce osob v obrazu a následné monitorování a vizualizace jejich pohybu. Práce se soustředí na datasety nasnímané z výšky, zaznamenané např. pomocí dronu. K detekci je využívaná model neuronové sítě RetinaNet. Každé detekované osobě je přiřazen příznakový vektor. Pomocí něj se monitoruje samotný pohyb osob. Trajektorie pohybu všech detekovaných osob je vykreslena do složeného panoramatického snímku. Úspěšnost detekce je 58,6 %. Tato metoda je pak porovnávána s metodou využívající termosnímky. Ta pro detekci využívá extrakci blobů a jejich klasifikaci. Pro sledování je využita metoda KLT. Vzhledem k tomu, že oba přístupy využívají jiné algoritmy nicméně na téměř totožný problém je možné je porovnávat, inspirovat se a hledat společné problémy při řešení monitorování osob z pohledu shora.

1 Úvod

Technologický vývoj v oblasti hardware i software zaznamenal v poslední dekádě obrovský pokrok. I díky tomu se rozšířil počet statických i pohyblivých kamer používaných pro různé účely. Mimo jiné například prevenci kriminality umístěním kamerových systémů na místa nejen kritické infrastruktury, ale i běžnějšího použití (stadiony, kulturní domy, parkoviště atp.). Největším pokrokem v oblasti kamerových systému je využití dronů – jako plně pohyblivých a ovladatelných kamer.

Na druhou stranu roste i nutnost využít softwarových řešení pro zpracování těchto záznamů. Hlavním důvodem je prosté množství hodin záznamů vytvořených různými kamerovými systémy. Druhotným důvodem je potom náročnost analýzy mnohahodinového záznamu člověkem. Řešením je tak využít ať už plně automatizovaného nebo alespoň polo automatizovaného zpracování záznamů a soustředit tak lidskou sílu na důležité části záznamu, kde je manuální zpracování nutné nebo dokonce vyžadované. V těchto oblastech se využívají metody strojového učení konkrétně neuronové sítě, schopné na základě velkého množství vstupních dat (určených pro naučení sítě) vytvořit rychlou a zároveň relativně přesnou analýzu obrazu (resp. videa).

Cílem práce je prostudovat možnosti detekce osob v obraze z výšky, z videozáznamu pořízeným dronem. Každý detekovaný člověk bude v průběhu zpracování videa identifikován. Jednotliví lidé budou od sebe rozlišeni, budou monitorováni v průběhu celého záznamu, a nakonec trajektorie jejich pohybů budou vizualizovány v panoramatickém snímku. Bude implementována aplikace pro otestování a demonstraci funkcionality.

V druhé kapitole jsou informace ohledně detekce objektů v obraze a jsou popsány vícevrstvé neuronové sítě, včetně často používaných architektur. Třetí kapitola se věnuje použitému datasetu a průběhu trénovaní použité neuronové sítě, která detekuje osoby na snímcích z výšky. Dále pak uvádí algoritmus pro monitorování chodců a popis vizualizace trajektorie. Experimenty a zhodnocení výsledků metod z předchozích kapitol jsou popsány v kapitole 4. Závěrečná kapitola pět pak shrnuje celou zprávu včetně potencionálních rozšíření. Tato zpráva vychází ze dvou hlavních zdrojů a tím jsou [1] a [2]. Informace v [1] představují námi testovanou variantu, zatímco zdroj [2] je uveden pro porovnání jiného přístupu pro dosažení stejného cíle.

2 Metody detekce osob z výšky

Dříve než přistoupíme k popisu samotných metod je vhodné být obeznámen se základy detekce objektů v obraze pomocí neuronových sítí i klasických algoritmů strojového učení a konvolučních neuronových sítí. Jedná se hlavně o informace ohledně principu fungování vícevrstvých sítí, technice backpropagation, různých aktivačních funkcí, HOG (*histogram of oriented gradients*), DCT (*discrete cosine transform*) SVM (*support vector machines*) – tyto informace je možné dohledat např. v [1] přesněji v kapitolách 2 až 2.5 nebo v [3] v podkapitole 3.1.1. případně

v [4] podkapitola 2.3. V kategorii konvolučních neuronových sítí jsou důležitá témata samotného fungování těchto sítí, používaných vrstev, různých detekčních sítí (RetinaNet, R-CNN, Fast R-CNN, YOLO atp.). Informace o těchto tématech je možnost dohledat v [1] v podkapitolách 2.6 a 2.7 nebo v [3] v podkapitolách 3.1.2 a 3.2.

V této kapitole bude jako první popsána metoda detekce vycházející z běžných snímků (využívající viditelné osvětlení) s využitím neuronových sítí. Bude popsán využitý dataset a průběh trénování. Jakmile bude tento krok hotový, měla by neuronová síť být schopná s jistou přesností detekovat osoby v obraze. Díky tomu je možné určit, na jakých souřadnicích (a s jakou pravděpodobností) se na snímku člověk nachází. Další velkou částí bude potom popis metody využívající termosnímky. Ta funguje na získání pravděpodobných snímků z osobu (za pomocí extrakce skvrn – *blob*). Tyto snímky jsou pak dále filtrovány pomocí SVM, které jako příznakový vektor využívá kombinací HOG a DCT. Klasifikace SVM určí, že na výřezu z obrazu je (nebo není) osoba. I tato část je zakončena popisem datasetu a trénování SVM.

2.1 Metoda použitá pro snímky s viditelným osvětlením

Na detekci osob z výšky bude využita síť RetinaNet. Toto rozhodnutí bylo uděláno po analýze hlavních kritérii pro výběr detekčních sítí. Těmi jsou přesnost a rychlost detekce. Vzhledem k tomu, že se snažíme detektovat osoby z výšky což je docela náročný úkol, a naopak nevyžadujeme detekci v realném čase (*realtime*) byl kladen důraz hlavně na přesnosti. Z tohoto pohledu byl po analýze přesnosti z [5] zvolena právě síť RetinaNet. Zásadní pro využití neuronových sítí je použitý dataset a samotný process trénování. Ty jsou popsány v dalších podkapitolách (a jak bylo předesláno vychází z [1]).

2.1.1 Dataset

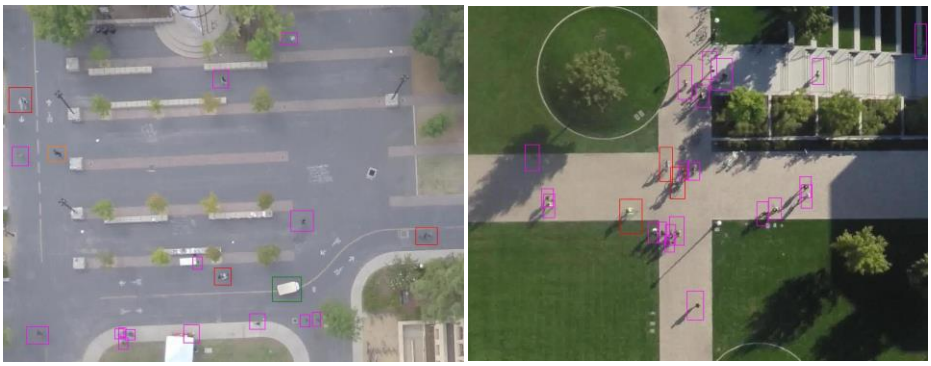
Proto pro úspěšné natrénování modelu bylo potřeba obstarat dataset obsahující záběry z pohledu z výšky a souřadnice ohraničující pozici člověka ve snímku (anotovaný dataset). Ačkoliv klasické uložiště datasetů (Google Open Images, Kaggle, Pascal VOC, COCO) obsahují datasety různých kategorií, ten obsahující specificky osoby z výšky mezi nimi nebyl.

Stanford Drone Dataset [1] (dále už jenom SDD) je dataset záběrů frekventovaných míst univerzitního kampusu na americké univerzitě ve Stanfordu. Dataset vznikl v rámci práce *Learning Social Etiquette: Human Trajectory Understanding in Crowded Scenes* [6], která se zabývá analýzou a předpovídáním pohybů nejenom chodců, ale i cyklistů, aut, autobusů apod. Dataset obsahuje videozáznamy ve formátu mp4 z 8 různých lokací o celkové velikosti 72 GiB. Každému videozáznamu přísluší anotační textový soubor, kde je zaznamenána poloha objektů na každém snímku videa. Druhy zaznamenávaných objektů jsou chodec, cyklista, člověk na skateboardu, auto, autobus a golfový vozík. Každá anotace také obsahuje tři příznaky: *lost*, *occluded* a *generated*. Příznak *lost* je zde z důvodu, pro který byl dataset vytvářen – predikce pohybu

[1] Stanford Drone Dataset je dostupný na http://cvgl.stanford.edu/projects/uav_data

cíle. Značí, zda se objekt nachází mimo zaznamenávaný obraz. Tedy cíl se objeví v zaznamenávané ploše, a když ji opustí, je odhadováno kudy se mohl pohybovat dále. Příznak *occluded* značí, že něco objekt překrývá směrem ke kameře. Například chodec zašel pod strom a není na kameře vidět. A *generated* říká, zda byla anotace automaticky interpolována. Některé anotace byly tvořeny manuálně, zbylé byly vygenerovány (většina).

Kvůli syntaxi anotací bylo nutné videozáznamy rozložit po jednotlivých snímcích a upravit informace z původních anotací do požadovaného formátu. Zároveň pro správné naučení sítě bylo třeba z anotací vyřadit objekty jiných tříd než chodců a také všechny objekty označené jako *lost* a *occluded*. Upravený dataset obsahoval přes 520 000 obrázků ve formátu jpg o celkové velikosti 150 GiB, anotovaných chodců bylo více než 3 800 000. Ukázka datasetu je na obrázku 1.



Obrázek 1: Ukázka z použitého datasetu SDD.

2.1.2 Proces trénování

Váhy byly u každého trénování inicializovány z předtrénovaného modelu na datasetu Image-Net. To by mělo zajistit rychlejší konvergenci než u náhodné inicializace. Jako páteřní síť byla zvolena dopředná konvoluční síť ResNet s 50 vrstvami. U každého trénovacího snímku byla jistá šance, že se před trénováním otočí, jelikož síť by neměla být závislá na tom, jak je chodec natočený ke kameře a jakým směrem jde. Každé trénování probíhalo 60 epoch – 60× byl síti představen celý dataset. Při trénování neuronové sítě je běžná praxe, že se obrázky síti nepředkládají po jednom, ale po balíčcích (*batches*) určité velikosti (*batch size*). Trénování se poté skládá z kroků (*steps*), které jsou definovány tímto vzorcem:

$$steps = \frac{velikost\ datasetu}{batch\ size} \tag{1}$$

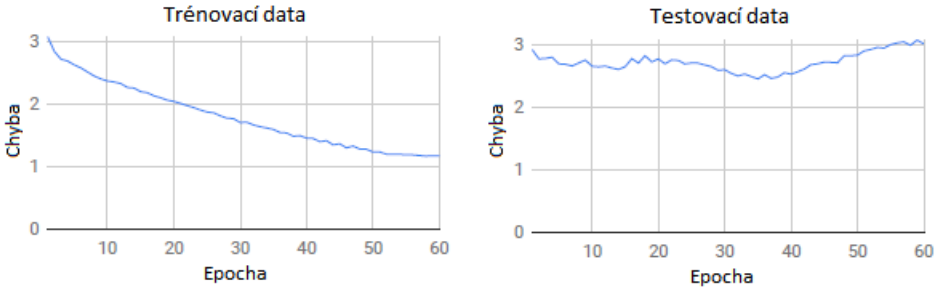
Počet kroků byl zvolen na 2 500, což s množstvím snímků v průměru 10 000 znamená, že síti byly předkládány anotační snímky po zhruba 4 kusech. Dále z každé scény byly vybrány snímky z jednoho videa, které sloužily jako validační

data. Jednalo se zhruba o 10 % dat. Každé trénování, celých 60 epoch, proběhlo za čas kolem 20 hodin.

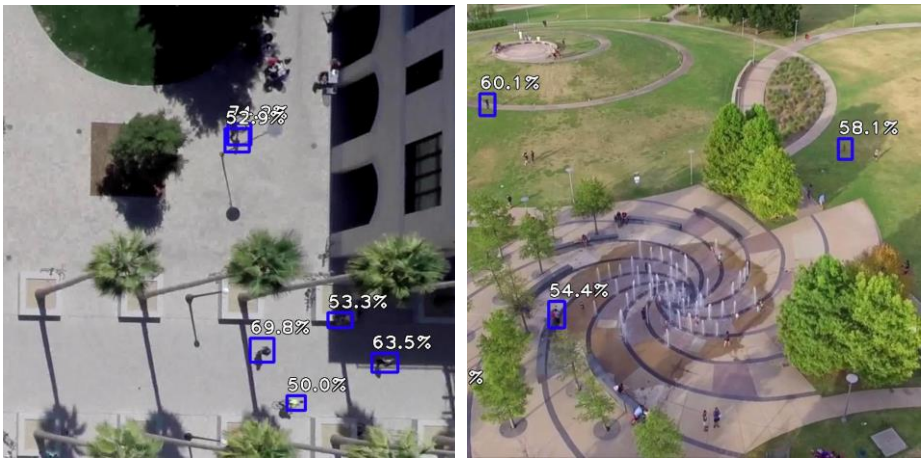
První trénování

Snímky získané z datasetu jsou však pořízeny z videa o 30 fps (frames per second). Dá se předpokládat, že v průběhu jedné sekundy, se příliš věcí v záběru nezmění. Chodci budou posunuti o pár pixelů dále, ale že by každý snímek přinášel nové relevantní informace o tom, jak chodec vypadá, je nepravděpodobné. Použití všech snímků by znamenalo pouze prodloužení procesu trénování, protože v podstatě stejná, či velmi podobná informace by byla zaznamenána na několika snímcích. Proto byl pro trénování použit až každý 50. snímek. To znamenalo, že pro první učení bylo pro trénování anotováno 67 386 chodců a pro testování 9 907.

Postup trénování je vidět na obrázku 2. Nejlépe si trénovaný model vedl kolem 35. epochy, potom chyba trénovacích dat začala růst a model začal jevit známky přetrénování. Z podrobnějšího prozkoumání a otestování výsledků prvního trénování vyplynulo, že spousta cyklistů je mylně rozpoznáno jako chodci a spousta chodců zase jako cyklisti. V původním datasetu tvořili anotovaní chodci zhruba 45 % a cyklisté 40 %. Z výšky je velmi těžké, někdy až nemožné rozeznat co je cyklista a co chodec. Na snímcích je zachycena spousta cyklistů, kteří vypadají jako chodci. Tyto anotace síť mátly, a to byl jeden z důvodů vysoké validační chyby a nepřesnosti modelu. Ukázka validačních snímků je obrázku 3.



Obrázek 2: Porovnání chyb trénovacích a testovacích dat z prvního učení v závislosti na množství odtrénovaných epoch.

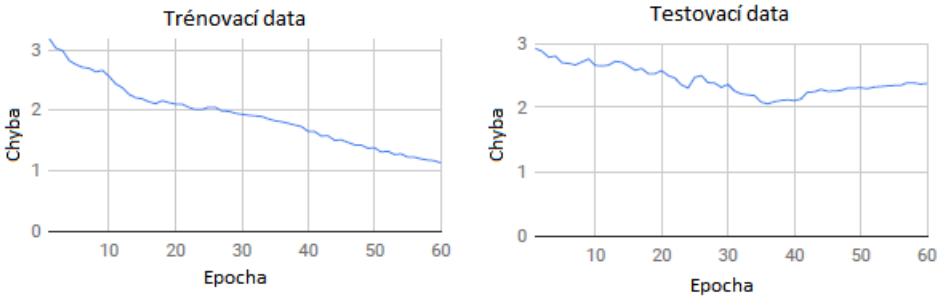


Obrázek 3: Validační snímky z prvního učení z trénovacích (vlevo) a testovacích (vpravo) dat.

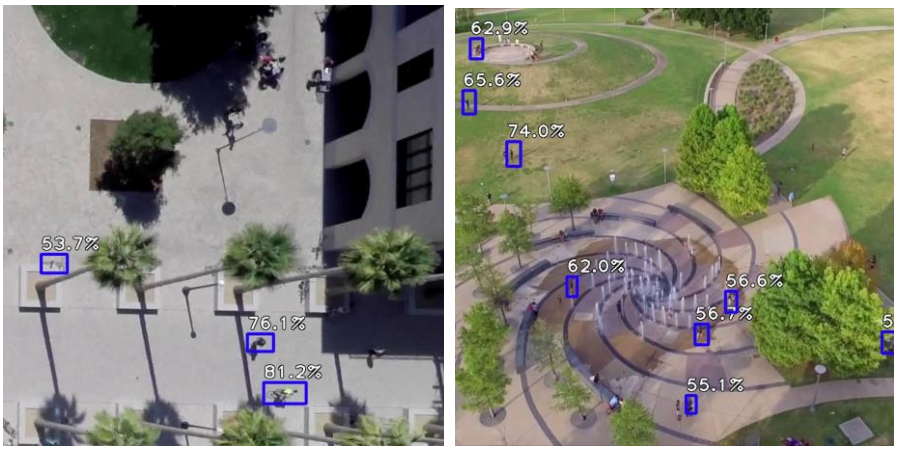
Druhé trénování

Na základě poznatků z prvního trénování byl dataset pro další trénování upraven. Do anotací byli zahrnuti nejen chodci, ale také cyklisté a lidé na skateboardu. To vše pod jednou společnou třídou. Druhá úprava spočívala v úpravě počtu snímků z jednotlivých scén. Videa z některých scén byla výrazně kratší, či obsahovala výrazně odlišný počet anotací. To mohlo mít za následek, že některé typy záběrů byly sítí preferovány. Například z určité výšky, z určitého úhlu či daného barevného spektra ovlivněné světelnými podmínkami ve scéně. Proto pro další trénování nebyl vybrán každý 50. snímek z každé scény, ale snímkování bylo ovlivněno počtem anotací z jednotlivých scén. A to tak, aby každá scéna obsahovala zhruba stejný počet anotací. Pro druhé učení bylo trénovacích anotací 55 241 a testovacích 8 263.

Průběh učení je vidět na obrázku 4. Jako u prvního trénování si model nejlépe vedl kolem 35. epochy. Výsledky byly lepší, ale ne uspokojivé. Dalším zkoumáním bylo zjištěno, že spousta anotací je uvedena úplně mylně. Například příznaky *occluded* nebyly uvedeny ve 100 % případech, a tak v anotacích byla spousta anotovaných objektů například pod stromem či pod střechou domu. Jindy byly zase anotace úplně mimo objekt nebo objekt pokrývaly jen zčásti. Tyto neduhy datasetu byly pravděpodobně způsobeny skutečností, s jakým účelem byl dataset vytvářen – predikce pohybů objektů. V takovém případě není důležité, zda je chodec na kameře skutečně vidět, je skrytý pod stromem nebo zda je anotován pouze zčásti. Bohužel chybných anotací bylo poměrně velké množství a pro lepší a přesnější výsledky bude nutné dataset ručně projít a obrázky s chybnými anotacemi vymazat. Příklady výsledků z druhého trénování jsou na obrázku 5.



Obrázek 4: Porovnání chyb trénovacích a testovacích dat z druhého učení v závislosti na množství odtrénovaných epoch.

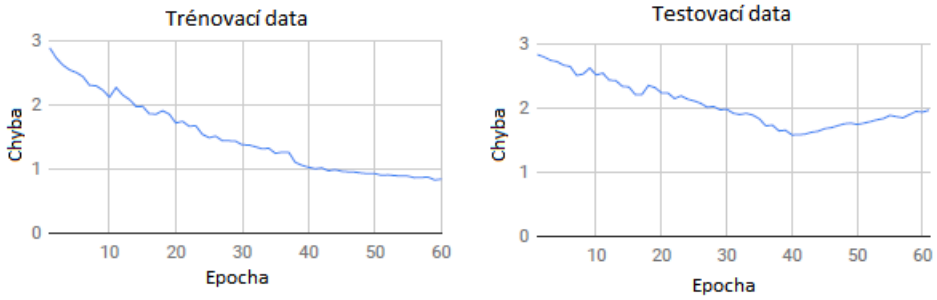


Obrázek 5: Validační snímky z druhého učení z trénovacích (vlevo) a testovacích (vpravo) dat.

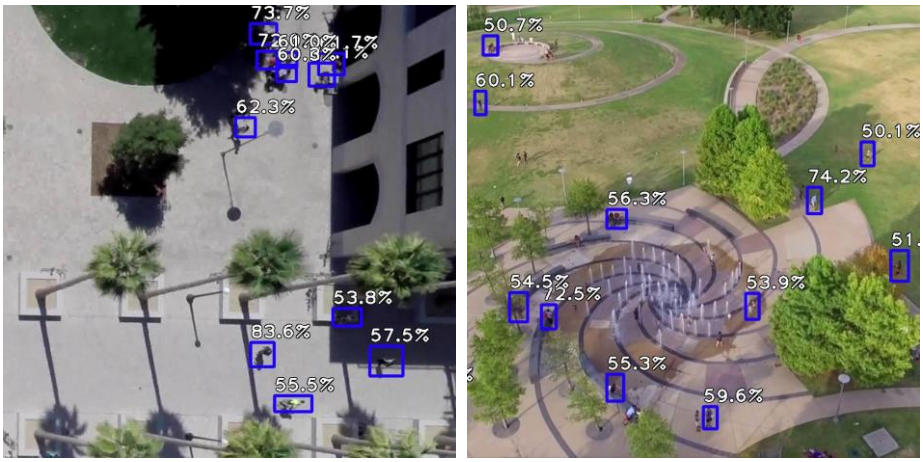
Třetí trénování

Pro třetí trénování byl dataset ručně protřízen. Ty snímky, kde byly anotace nepřesné či úplně chybné, byly vyřazeny. Celkově bylo vymazáno 1 926 snímků z 9 549 vygenerovaných stejným způsobem jako u druhého trénování, aby anotace z žádné scény nepřevažovaly. Anotací pro trénování zůstalo 63 433 a anotací pro testování 14 010.

Model tentokrát dosáhl nejlepších výsledků kolem 40. epochy, ty byly opět lepší než v předchozím trénování (viz obrázek 6). Hodnota chyby (pro testovací data) dosáhla minimální hodnoty 1,58. Tyto výsledky jsme vzhledem k rozmanitosti datasetu považujeme za dostačující. Příklady snímků z třetího učení jsou vidět na obrázku 7.



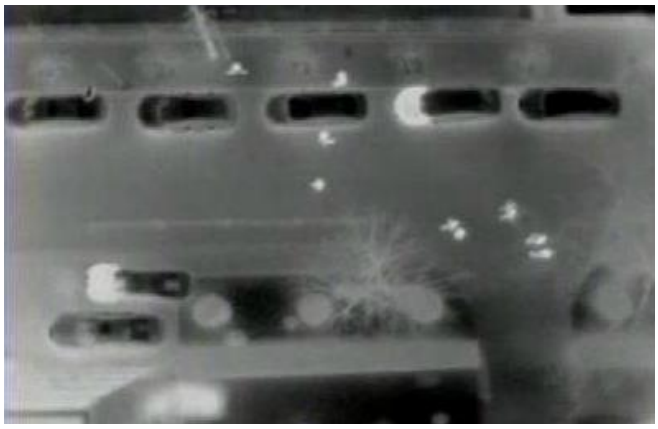
Obrázek 6: Porovnání chyb trénovacích a testovacích dat z třetího učení v závislosti na množství odtrénovaných epoch.



Obrázek 7: Validační snímky z třetího učení z trénovacích (vlevo) a testovacích (vpravo) dat.

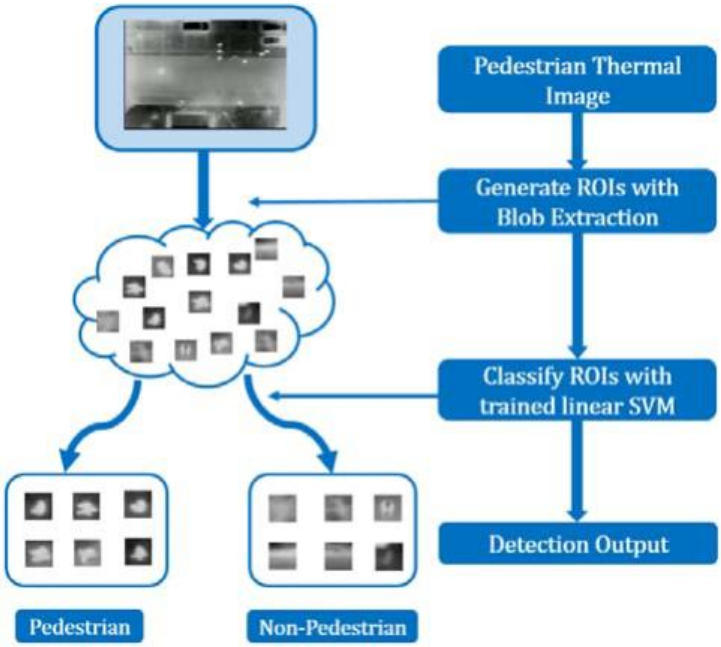
2.2 Metoda použitá pro termosnímky

Tato podkapitola vychází z [2]. Rád bych zdůraznil, že tato metoda nebyla implementována v rámci projektu, je zde jen pro porovnání. V úvodu textu se dočteme, že jeden z hlavních důvodů, proč bylo využito termosnímků (resp. kamery snímající infračervené spektrum) je fakt, že detekce osob z výšky je při použití běžných snímků extrémně náročná činnost. Termosnímky zjednodušují tuto činnost tím, že není třeba řešit stíny, nízký kontrast s pozadím – resp. úmyslnou či neúmyslnou snahu o kamufláž. Nicméně tu jsou jiné problémy a to např. variabilita lidské teplotní stopy, obzvláště ve spojitosti s různými meteorologickými podmínkami a teplotním pozadím scény. Příklad termosnímku pro detekci osob z výšky je vidět na obrázku 8.



Obrázek 8: Ukázka termosnímku ukazující osoby z výšky.

Navržená metoda se skládá ze dvou fází extrakce a klasifikace *blobu* (z angličtiny: kaňka, skvrna – tento termín necháváme dále bez překladu). Graficky je proces detekce znázorněn na obrázku 9. Extrakce blobu si klade za cíl najít ve snímku bloby, které by mohly být osobou, tedy najít ROI (*region of interest* – zájmové oblasti). K tomu využívá lokálních příznaků gradientu a geometrického filtrování. Jak světlé, tak tmavé oblasti mohou být detekovány. V druhé fázi klasifikace je každý blob klasifikován za pomocí SVM. Ten k tomu využívá fúze příznaku získaných za pomocí HOG a DCT (*discrete cosine transform* – diskrétní kosinova transformace). Klasifikátor má určit, zda vybrané ROI (blob) je nebo není osobou a tím zpřesnit proces detekce.

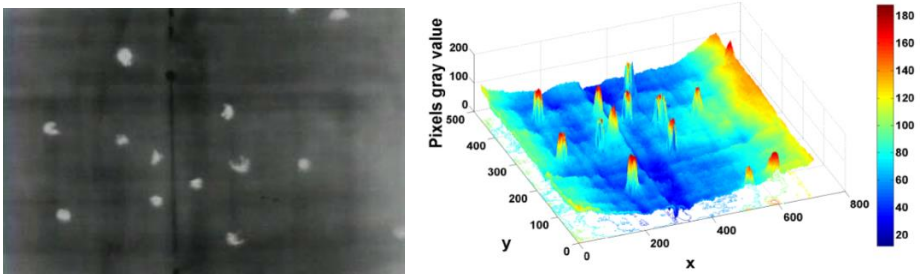


Obrázek 9: Procesu detekce osob z termosnímků.

2.2.1 Extrakce blobu

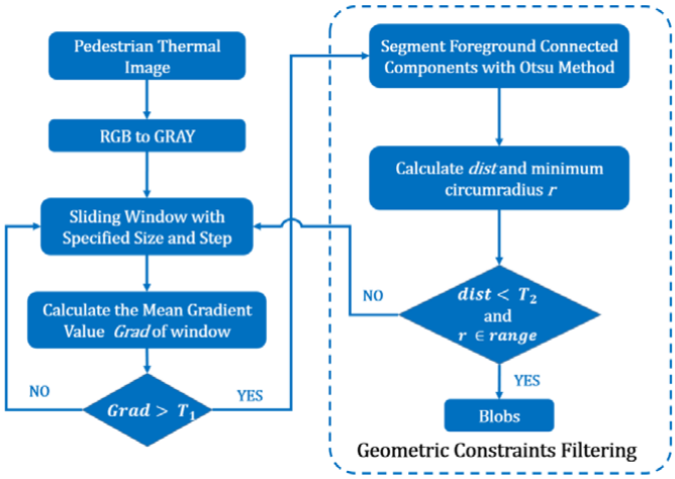
Standardní přístupy k výběru ROI (např. background substraction, intenzita kontrastu, CSM reprezentace, adaptivní prahování, MSER) se často opírají o vlastnosti, které v zadané situaci není možno použít, jsou to např. stacionární kamera, detekce pouze objektů pohybujících se v popředí, fragmentace těla nebo sloučení osoby s pozadím (což by přineslo problémy pro následující krok klasifikace). Detekce osob z výšky je natolik specifická úloha, že byl pro ni vytvořen nový přístup.

Ten využívá toho, že chodci jsou na snímcích z termokamery světlejší (nebo tmavší) než okolí. Tento efekt je hezky znázorněn na obrázku 10. Obrázek 10a ukazuje snímek s termokamery a obrázek 10b potom jeho převedení do 3D kdy výška znázorňuje intenzitu barvy pixelu. Je zřejmé, že na okrajích osob dochází k výrazné změně gradientu. Myšlenkou tedy je oddělit potencionálně zájmové oblasti na základě hran gradientu. V tomto procesu je využita metoda sliding window. Jednotlivé kroky jsou přehledně vidět na obrázku 11, kde T_1 je předdefinovaný práh gradientu, T_2 je práh vzdálenosti, *range* je hodnota minimálního průměru opsané kružnice osoby a *grad* je průměrná hodnota gradientu v lokální oblasti.



a) b)

Obrázek 10: a) Termosnímek osob z výšky, b) převedení snímku a) do 3D.



Obrázek 11: Vývojový diagram algoritmu pro extrakci blobu.

Dále bude podrobněji popsána průměrná hodnota gradientu (*grad*). *Sliding window* je čtverec o straně *k*. Aby byla oblast určena jako zájmová je nutné, aby průměrná hodnota šedé byla ve vnitřních kruzích větší (nebo menší) než hodnoty ve vnějších kruzích (viz obrázek 12). Vzhledem k různým směrům a fázím chůze používáme průměrnou hodnotu šedé ze všech buněk v kružnici (viz obrázek 12). Poté může být změna jasu (tj. hodnota gradientu G_i průměrné hodnoty šedé mezi *i* a (*i*+1) kružnicí vypočtena pomocí této rovnice:

$$G_i = |\bar{C}_i - \bar{C}_{i+1}| \tag{2}$$

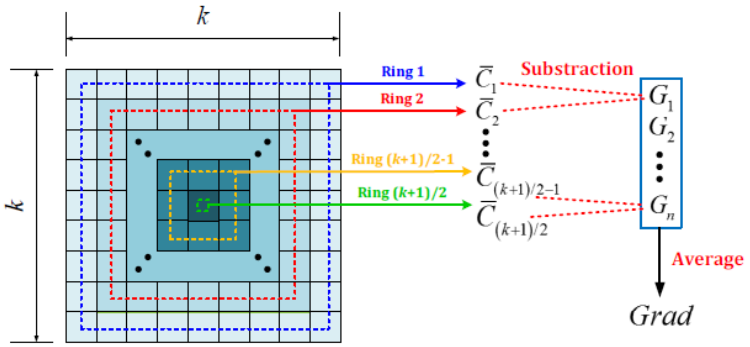
kde $\bar{C}_i$ reprezentuje průměrnou hodnotu šedi ve všech buňkách v i-té kružnici. Pokud máme G_i jako průměrnou hodnotu gradientu pro oblast pak *grad* je vypočten takto:

$$Grad = \frac{\sum_{i+1}^{n} G_i}{n} \tag{3}$$

kde je *n* je počet G_i v dané lokální oblasti. Důležitým parametrem je pak hodnota *k* (velikost *sliding window*). Ta se odvíjí od letové výšky, empirický vzorec pro určení této hodnoty je:

$$k = \frac{110 - h}{2} \tag{4}$$

kde *h* je letový výška v metrech (očekává se, že bude menší než 110 m) a výsledek je zaokrouhlen dolů.



Obrázek 12: Ilustrace výpočtu hodnoty *grad*.

Důležitou částí je pak samotná filtrace na základě geometrických omezení. Předchozí metoda nalezne mnoho ROI, avšak dost z nich je falešně pozitivních. Pro další filtrování je použita metoda segmentace pomocí Otsu prahování [7]. Tato metoda vytvoří černobílý obrázek, v němž je mnohoúhelník reprezentující blob. Filtrování probíhá na základě dvou geometrických omezení:

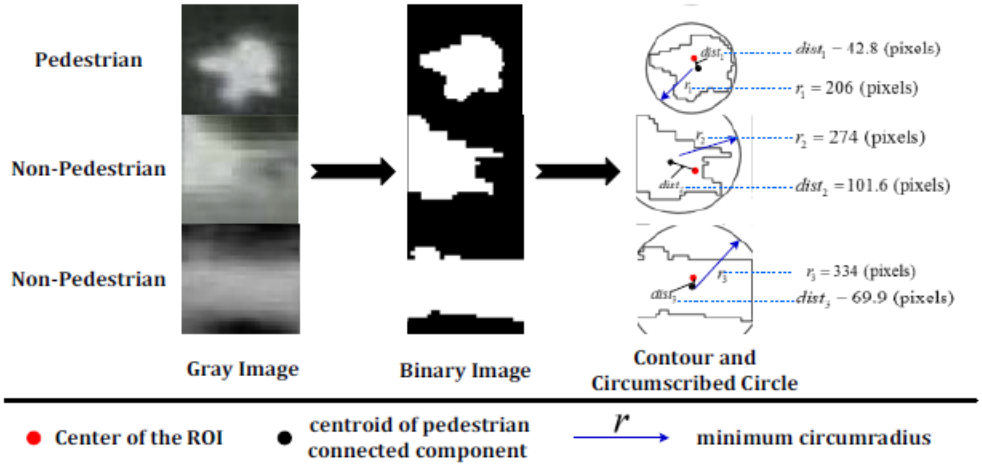
První z nich je **pozice středu**. Přihlédneme-li k obvyklým vlastnostem pohledu shora na termogram osoby měla by být vzdálenost mezi těžištěm

mnohoúhelníku a středem ROI nepříliš velká. Vypočítáme Euklidovu vzdálenost mezi těmito dvěma body a porovnáme to s prahem T_2 dle následujícího vzorce:

$$\begin{cases} Pedestrian\ ROI: pokud\ dist\ = \|P_m - P_0\| \leq T_2 \\ Non-Pedestrian\ ROI: pokud\ dist\ = \|P_m - P_0\| > T_2 \end{cases} \quad (5)$$

Kde P_m je těžiště osoby (mnohoúhelníku) a P_0 střed oblasti. Tato podmínka eliminuje oblasti, které vycházejí z rohu nebo okrajů světlých oblastí na termosnímku a neúplnými fragmenty blobu osob. Dále zajišťuje, že je osoba uprostřed oblasti což je výhodné pro druhou fázi (klasifikaci).

Druhým geometrickým parametrem je **minimální průměr** opsané kružnice. Při pohledu shora je velikosti osob relativně podobná, a tak se dá definovat rozumná velikost blobu. To je provedeno tak, že se udělá opsaná kružnice mnohoúhelníku a její průměr musí být pod stanovenou hranici *range*. Obě metody přehledně prezentuje obrázek 13. Jeden blob je obvykle detekován ve více oknech (*sliding window*) v tomto případě se tak jako výsledná oblast bere průměr pozic středů nalezených oken (oblastí).



Obrázek 13: Ukázka aplikace geometrických omezení.

2.2.2 Klasifikace blobu

I přes snahu ve fázi extrakce odfiltrovat všechny nežádoucí ROI, je v nich pořád mnoho oblasti, které chodce neobsahují (jedná se o typicky kulaté objekty – lampy, motory aut, mlází atp.). Tyto objektu budou dále odfiltrovány pomocí klasifikátoru SVM. K tomu, je potřeba určit charakteristické rysy jednotlivých blobů, na ty jsou využity dvě metody, které data popíšou – HOG a DCT.

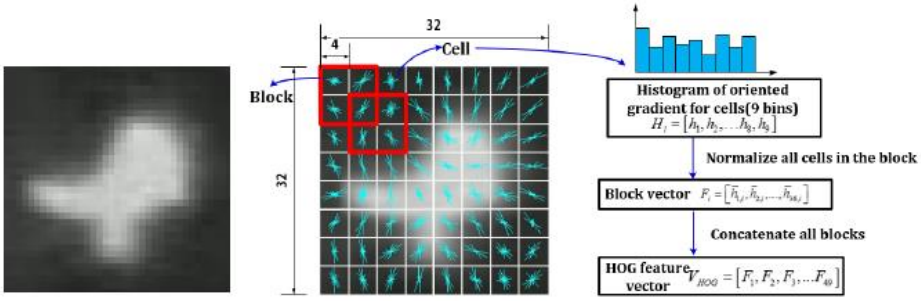
HOG využívá normalizované lokální histogramy v husté mřížce. Prvním krokem je výpočet gradientů a orientací všech pixlů v obrázku použitím 1D masky s jádrem filtru [-1, 0 1]. Druhým krokem je vytvoření histogramů buněk. V tomto kroku je obrázek rozdělen na buňky a histogram H_i je vygenerovaný rozdělením lokálních gradientů do skupin („košů“) podle orientace každý buňky. Třetím

krokem je složení buněk dohromady do větších, prostorově spojených, bloků F_i. Tyto jsou potom normalizovány pro snížení vlivu osvětlení a stínů. Vzhledem k tomu, že každý blok obsahuje skupinu buněk, může se stát, že je buňka součástí několika bloků, a tedy i jejich normalizací (když se bloky překrývají). Poslední krok je získání vektorů rysů V_{HOG} konkatenací histogramů všech bloků.

Zjednodušeně HOG vektor popisuje lokální intenzity gradientů a směry hran. V provedených experimentech se výška dronu pohybovala mezi 40-60 m velikost příslušných objektů pak byl čtverec o hraně 35-25 px. Normalizace velikosti objektu je tedy někde mezi tím, škálováním vzorků na čtverec o hraně 32 px. Využije se 8×8 bloků s 4 pixelovým krokem a 9 „košů" pro histogramy (po 20° orientace). Vznikne tedy 49 bloků a podle toho i výsledný vektor V_{HOG} a příslušné hodnoty F_i:

$$V_{HOG} = [F_1, F_2, \ldots, F_i, \ldots, F_{49}]$$
$$F_i = [\bar{h}_{1,i}, \bar{h}_{2,i}, \ldots, \bar{h}_{j,i}, \ldots, \bar{h}_{36,i}] \quad (6)$$

kde V_{HOG} je vektor rysů pro HOG a F_i je normalizovaný vektor pro i-tý blok. Definice F_i pak vychází z toho, že každý blok obsahuje 4 buňky a každá buňka 9 „košů". $\bar{h}_{j,i}$ je j-tá normalizovaná hodnota pro i-tý blok. Obrázek 14 ukazuje celý proces získání vektoru pro HOG. Obrázek 14a ukazuje původní obrázek, 14b vizualizace gradientu orientací pro všechny buňky (červeně jsou ohraničeny bloky) a 14c ukazuje získání údajů pro vektor rysů.



a) b) c)

Obrázek 14: Ilustrace procesu získání vektoru rysů pomocí HOG metody.

DCT transformuje signál (prostorové informace) do frekvenční reprezentace. K získání vektoru rysů pro DCT je potřeba vypočítat DCT koeficienty. Obecně se výpočet koeficientů probíhá podle tohoto vzorce:

$$G(u,v) = \frac{2}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} C(u)C(v)\, f(x,y)\, \cos\left[\frac{\pi u(2x+1)}{2M}\right] \cos\left[\frac{\pi v(2y+1)}{2N}\right] \quad (7)$$

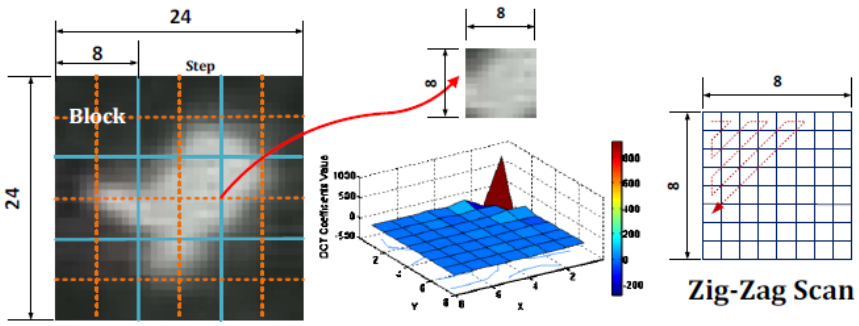
kde $G(u, v)$ je matice koeficientů frekvenčních komponent (u, v), M×N jsou rozměry obrázku, $f(x, y)$ je intenzita barvy pixelu v řádku x a sloupci y vstupního obrázku. $C(u)$ a $C(v)$ jsou definovány takto:

$$C(u), C(v) = \begin{cases} 0{,}5 \ \ pokud \ u = 0 \ a \ v = 0 \\ 1 \ \ v \ ostatních \ případech \end{cases} \tag{8}$$

Vektor rysů získaný pomocí **DCT** by měl zmírnit varianci ve vzhledu chodců. Většina termogramů osob z výšky má energii signálů v malých frekvencích což odpovídá vysokým magnitudám DCT koeficientů v levém horním rohu matice koeficientů. Vektor rysů složíme z 21 koeficientů pomocí *zig zag* (hadovitého) průchodu levého horního rohu matice koeficientů výsledek je pak:

$$V_{DCT} = [v_1, v_2, \ldots, v_{21}] \tag{9}$$

kde v_i jsou jednotlivé koeficienty. Získání vektorů rysů přehledně ukazuje obrázek 15. Vstupní obrázek je rozdělen na bloky 8×8 pixelů (obrázek 15a). Diskrétní kosinová transformace je aplikována na každý takový blok. Z každého bloku tak získáme příslušnou DCT matici koeficientů viz obrázek 15b. Obrázek 15c ukazuje *zig zag* metodu získání 21 koeficientů. 21 koeficientů bylo určeno experimentálně, využití většího množství koeficientů nezaručuje lepší výsledky. A tyto zvolené obsahují výrazně větší amplitudy než ostatní (jak je vidět na obrázku 15b).



a) b) c)

Obrázek 15: Ilustrace procesu získání vektoru rysů pomocí DCT metody.

Posledním popsaným bodem je samotný **SVM** klasifikátor. Ten kombinuje vektory obou metod (HOG i DCT). Bohužel samostatně tyto metody nebyly dostatečně přesné. Proto byl vytvořen vektor V_F, který kombinuje vektory z obou metod. Vzhledem k tomu, že DCT i HOG adresují jiný problém dává takové řešení smysl. K tomu je však nutné oba vektory normalizovat, pro V_{HOG} byla použita metoda L2-norm. DCT je pak normalizován pomocí tohoto vzorce:

$$v_i^* = \frac{v_i}{max(v_1, v_2, \ldots, v_i, \ldots, v_{21})} \tag{10}$$

výsledný vektor V_F je pak vytvořen jako vážená kombinace V_{HOG} a V_{DCT} takto:

$$V_F = [\alpha \, V_{HOG}, \beta \, V_{DCT}] \tag{11}$$

hodnoty α a β jsou hyperparametry, které jsou určeny křížovou validací s cílem minimalizovat počet falešných klasifikací v trénovacích datech. S tímto

kombinovaným vektorem bylo SVM relativně úspěšné v klasifikaci chodců. Viz kapitola 2.2.4. Poznámka ke klasifikaci: využito bylo pouze lineárního jádra bez metod selekce parametrů (jako jsou křížová validace nebo vyhledávání v mřížce). To ze dvou důvodu, prvním z nich je časová náročnost takových algoritmů a druhým je těžké určit typ jádra.

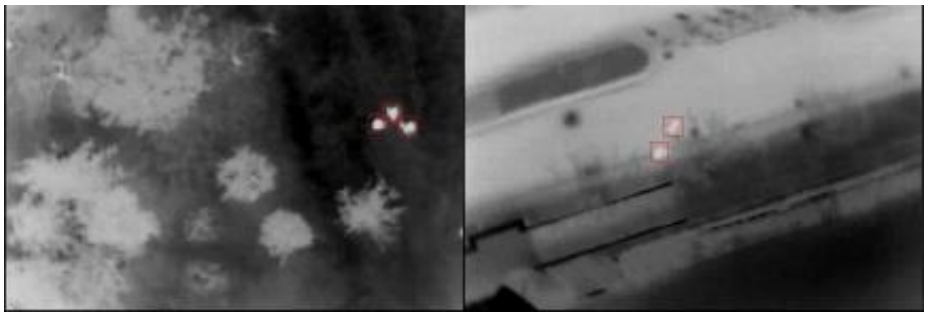
2.2.3 Dataset a trénování

K tomu, aby bylo možné metody popsané v předcházející kapitole 2.2.1. vůbec vyhodnotit, byla vytvořena série měření (resp. snímání). Ty byly zachyceny LWIR (*long-wave infrared*) kamerou s rozlišením 720×480 připevněnou na dron letící ve výšce 40–60 m. Pět různých měření (dále nazývané scény) se liší pozadím, parametry letu, délkou snímání i venkovní teplotou. Parametry snímání ukazuje přehledně tabulka 1. Scéna 1 byla nasnímaná jako prvotní testování celého procesu. Snímána byla jednoduchá pozadí (louka, hřiště atd.), kde je vysoký kontrast mezi chodci a pozadím. Pro takové typy scén nakonec není nutné ani trénovat SVM, stačí jen extrakce blobů. V testu bylo vybráno 176 typických snímků a ty byly prohlášeny za scénu 1.

Tabulka 1: Základní informace o jednotlivých scénách.

Dataset	**Rozlišení**	**Letová výška**	**Pozadí**	**Datum/Čas**	**Teplota**
Scéna 1	720×480	40-70 m	Různá místa	Duben/ Odpoledne, noc	10 °C-15 °C
Scéna 2	720×480	50 m	Cesta	20.1./ 20:22	5 °C
Scéna 3	720×480	60 m	Cesta	19.1./ 19:46	6 °C
Scéna 4	720×480	40 m	Cesta	8.4./ 23:12	14 °C
Scéna 5	720×480	50 m	Cesta	8.7./ 16:12	28 °C

Snímání u ostatních scén probíhalo v komplikovanějších podmínkách, které obsahují lampy, motory aut, klestí a jiné objekty které jsou podobné osobám (z pohledu termogramů samozřejmě). V těchto případech již bylo nutné trénovat klasifikátor. Trénovaní probíhalo tak, že metoda extrakce blobů našla určitý počet objektů. Tyto objekty pak byly manuálně klasifikovány a rozděleny na testovací a trénovací. U scény 2 tak bylo detekováno 2783 blobů, z nich 544 bylo využito jako trénovacích a 2239 jako testovacích. U Scény 3 se jedná o podobný počet 2817 celkově, 512 trénovacích a 2305 testovacích. Scéna 4 obsahovala pouze 1446 blobů z toho 365 trénovacích a 1081 testovacích. Poslední Scéna 5 pak byla složena z 1270 obrázků, 340 trénovacích a 930 testovacích. Obrázek 16 ukazuje výsledky detekce ve Scéně 1.



Obrázek 16: Příklady detekce blobů v ukázkových obrazcích ze Scény 1.

3 Aplikace pro monitorování chodců

Dále je navržen a popsán algoritmus identifikace osob – ten má na starosti jednotlivé chodce od sebe navzájem rozlišit a vizualizovat trajektorie jejich pohybů v průběhu celého videa. Je potřeba také vyřešit otázku zpracování videa, tedy postupného dávkování snímků detektoru. A nakonec implementovat demonstrační aplikaci pro kompletní otestování funkcionality a provedení experimentů.

Z pohledu teorie se obě metody velmi liší, u viditelného osvětlení je využito hlavně barevných histogramů a potom různých vzdáleností (manhattanovská, korelační). Informace o těchto metodách lze najít např. v [4] v kapitole 2. U termogramů je nejdůležitější metoda sledování KLT viz [8] a princip sledování byl hodně ovlivněn [9] a [10].

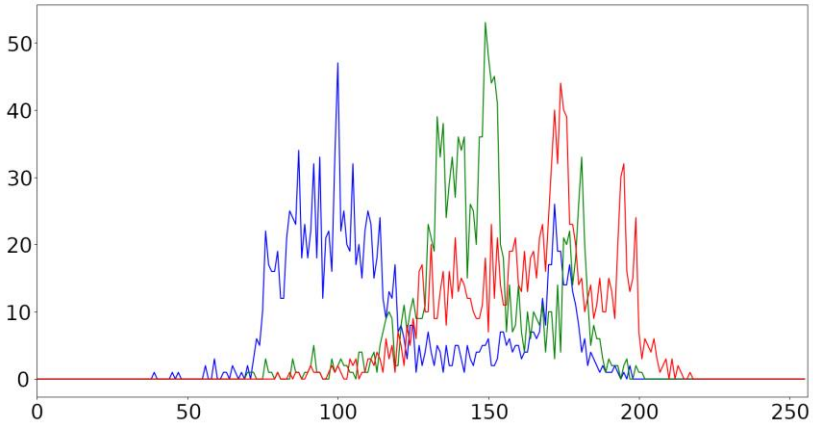
3.1 Metoda použitá pro snímky s viditelným osvětlením

Tato podkapitola vychází z [1]. Vzhledem k tomu, s jakými typy záběrů se pracuje, jsou detekovaní lidé běžně zaznamenáni v matici o výšce a šířce řádově desítek pixelů. To není mnoho, navíc z toho značnou část tvoří pozadí. Lidé také mohou stát přímo kolmo ke kameře. V takových záběrech pak tělo člověka není zachyceno téměř vůbec. Jak je vidět na obrázku 17, síť si také někdy pomáhá detekcí typického stínu člověka, pokud jsou záběry pořízeny ve světelných podmínkách, které to umožňují. Díky těmto aspektům není možné využít nové sofistikované algoritmy pro reidentifikaci lidí, jako například AlignedReID [11].

Obrázek 17: Výřezy detekovaných osob neuronovou sítí.

Díky výše zmíněným skutečnostem je zvolen jednodušší a přímočařejší přístup, a sice extrakce příznaků pomocí **barevných histogramů**. Každé

detekované osobě ve snímku bude spočítán barevný histogram – jak jsou jednotlivé barevné kanály v daném segmentu rozloženy. Příklad takového histogramu je vidět na obrázku 18. Na ose *x* je vyznačeno 256 možných hodnot jednotlivých barevných složek. Na ose *y* počet pixelů s touto barvenou hodnotou. Barevný histogram bude sloužit jako příznakový vektor (*feature vector*). Jedná se o seznam hodnot, který popisuje konkrétní obrázek a může být porovnáván s jinými příznakovými vektory. Detekované osoby, kterým bude spočítán barevný histogram, budou označeny jako viděné a jejich příznakové vektory budou uloženy. V dalším snímku budou opět vypočítány příznakové vektory pro každého detekovaného chodce. Následně budou porovnávány s příznakovými vektory již viděných chodců. Pokud bude nalezena dostatečná shoda, chodec bude považován za reidentifikovaného, jeho příznakový vektor bude nahrazen aktuálnějším a jeho souřadnice v aktuálním snímku budou uloženy.

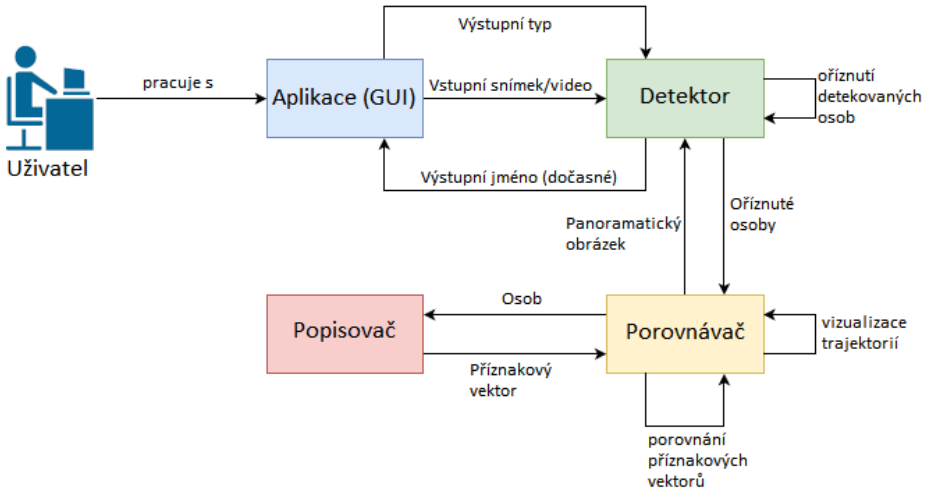


Obrázek 18: RGB histogram jednotlivých barevných složek pro první osobu zleva z obrázku 17.

Předpokládaná funkcionalita je tato. Uživatel přes grafické uživatelské rozhraní specifikuje vstup a spustí detekci. Zároveň definuje, jestli jde jen o detekci osob (zobrazí ohraničující obdélníky a pravděpodobnosti detekce) nebo vizualizovat trajektorie. Aplikační okno dávkuje vstupní soubor po jednotlivých snímcích a postupně je předává detektoru. Ten v každém snímku detekuje osoby, vyřízne je a předá porovnávači (*matcher*). Porovnávač předává výřezy dále popisovači (*describer*), který mu vrací příznakové vektory. Porovnávač příznakové vektory ukládá a porovnává. Po zpracování všech výřezů vizualizuje trajektorie chodců a výslednou mapu předá detektoru. Uživatel poté může pokračovat v práci s aplikací. Například si výstup uložit, zadat nový soubor či spustit detekci znovu. Návrhový diagram celé aplikace je na obrázku 19.

Není implementován žádný *image stitching* algoritmus pro spojování snímků napříč videem. Trajektorie pohybu lidí jsou vykreslovány do prvního snímku videa. Z toho plyne zásadní omezení, a sice že pořízené video musí být ze statického bodu – kamera se v průběhu pořizování záznamu nesmí hýbat. Každému rozpoznanému chodci jsou propojena všechna místa, kde byl v průběhu

videa identifikován. Kolem chodce je vyznačen ohraničující obdélník (*bounding box*) a z jeho středu vede linie, kudy se bude dál ve videu pohybovat.



Obrázek 19: Návrhový diagram aplikace pro monitorování osob.

3.1.1 Extrakce a porovnávání příznakových vektorů

Skutečně použitý histogram se od toho na obrázku 18 liší. Byl použit třídimenzionální histogram a jednotlivé barevné kanály byly rozděleny do 8 binů. Kategorizace obrazových bodů probíhá následujícím způsobem: Kolik pixelů má hodnotu červené složky z daného intervalu, zelené z jiného a modré zase z jiného. Například červené z ⟨0, 31⟩, zároveň zelené z ⟨32, 63⟩ a modré z ⟨64, 95⟩. Reálný počet binů je tak $8^3 = 512$.

Histogram je následně normalizován, aby velikost rozlišení vstupního snímku nezkreslovala výsledky. Každá hodnota v histogramu je vydělena maximální možnou hodnotou, tím jsou všechny hodnoty přeškálovány do intervalu ⟨0, 1⟩. Spočítání podobnosti výřezů obrázků lze realizovat porovnáním jejich příznakových vektorů. Cílem je spočítat jejich vzájemnou vzdálenost. K tomu lze použít různé distanční metriky, například euklidovskou vzdálenost,

$$d(u,v) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2} \tag{12}$$

manhattanskou vzdálenost,

$$d(u,v) = \sum_{i=1}^{n}|u_i - v_i| \tag{13}$$

nebo korelační vzdálenost,

$$d(u,v) = \frac{\sum_{i=1}^{n}(u_i - \hat{u})(v_i - \hat{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \hat{u})\sum_{i=1}^{n}(v_i - \hat{v})}} \quad (14)$$

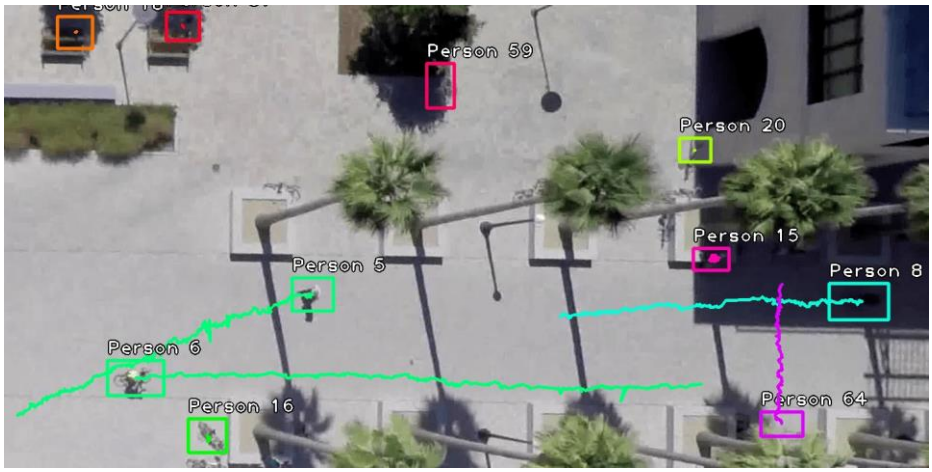
kde u, v jsou příznakové vektory a $\hat{u}$, $\hat{v}$ jsou průměrné hodnoty. Experimentálně bylo zjištěno, že pro tento typ dat bude nejvhodnější korelační vzdálenost, kde byly rozdíly ve vzdálenostech mezi stejnými objekty napříč snímky vůči jiným objektům nejvýraznější.

3.1.2 Testování funkčnosti příznakových vektorů

V rámci testování funkčnosti výše zmíněného přístupu bylo nutné provést několik experimentů, a tím zjistit, zda porovnání pouze příznakových vektorů je dostačující. Hned první experiment (obrázek 20) ukázal, že tento přístup není ideální. Pokud je ve snímku zachycen větší počet lidí, je pravděpodobné, že si jejich příznakové vektory mohou být velmi podobné, a díky tomu jako nejpodobnější může být vyhodnocen úplně jiný.

Obrázek 20: Vizualizace trajektorií chodců první verzí algoritmu. Chodci jsou porovnáváni pouze na základě příznakových vektorů.

Je nemožné, aby se člověk v rámci jednoho snímku videa přesunul o výraznou vzdálenost. Vzhledem k tomu byl algoritmus upraven, aby příznakový vektor osoby byl porovnáván pouze s příznakovými vektory osob v její blízkosti. To zajistilo, že detekované osoby nejsou spojovány napříč snímkem. Z tohoto přístupu vyplývá problém v následujícím případě: Člověk projde pod nějakou překážkou, a tím se kameře schová. Jeho další detekce, ačkoliv o spoustu snímků později, může tak být velmi vzdálená od té předchozí. Stejný případ může nastat, pokud člověk záběr opustí a poté se do něho vrátí v jiném místě. Výsledek z druhé verze je zobrazený na obrázku 21.



Obrázek 21: Vizualizace trajektorií chodců druhou verzí algoritmu. Při porovnávání příznakových vektorů je brána v potaz i jejich reálná vzdálenost.

Výsledek nyní vypadá podstatně lépe. Avšak ve snímku jsou vyznačeny statické objekty, které jistě nejsou lidé. To je dáno chybou neuronové sítě, která v některém snímku vrátila jako osobu objekt, který osoba není. Druhý problém je vidět na osobě 64. Tato osoba se ve videu objevila až později, nebyla na prvním snímku. Proto je zakreslen ohraničující obdélník kolem prázdného místa. Dá se předpokládat, že ve většině záznamů se lidé budou pohybovat, zvlášť bude-li videozáznam delší. Pokud se člověk pohybovat nebude, nemá pro něho vykreslení trajektorie pohybu smysl. Algoritmus byl upraven tak, aby pro statické objekty, které se v rámci celého pořízeného záběru nepohnou, nebylo vykresleno nic. Tato úprava vyřešila problém občasné chybné detekce neuronovou sítí nějakého statického objektu jako osoby. Druhá úprava se pak postarala o problém prázdných obdélníků pro osoby, které se v záběru objevily až později. Při jejich prvním výskytu je obraz s nimi vyříznut a překopírován do panoramatického snímku. Snímek z třetí verze algoritmu je vidět na obrázku 22.

Obrázek 22: Vizualizace trajektorií chodců třetí verzí algoritmu. Nyní jsou ve snímku vyznačeni pouze pohyblivé objekty a směr jejich pohybu.

3.2 Metoda použitá pro termosnímky

Základní myšlenka použité metody je sledovat pozici všech osob detekovaných pomocí předchozího algoritmu najednou a ukládat informace o trase jejich pohybu. Důležitou součástí této metody pro termosnímky je tzv. kalibrace videa. Jejím účelem je vyrovnat se z nepravidelným pohybem dronu ve všech osách (roll, pitch a yaw) [3] a tím i pohybu získaných snímků. Celá tato podkapitola se opírá o [2].

3.2.1 Kalibrace videa

Cílem kalibrace je prostorové zarovnání snímků videa tak, aby využívali stejný souřadnicový systém definovaný vzorovým snímkem (*frame*). Je zřejmé, že chyba v tomto algoritmů zanese chybu i dalšího kroku sledování trajektorie. Obecně metody na zarovnání na základě význačných oblastí ve videu se skládají z těchto kroků: získání význačných (stacionárních) bodů, určení korespondence mezi množinou souřadnic všech odpovídajících si význačných bodů, výpočet zobrazení mezi těmito souřadnicemi a aplikace zobrazení z předchozího bodu na daný snímek.

V tomto případě bylo využito registrace na základě **metody sledování KLT** (Kanade–Lucas–Tomasi) [12]. Prvním krokem této metody je získání KLT příznaků. To se děje na základě výpočtu přemístění příznaku mezi dvěma následujícími snímky a dá se to popsat touto rovnicí:

$$\begin{aligned} J(x) &= I(\boldsymbol{x} - \boldsymbol{d}) + \eta(\boldsymbol{x}) \\ \boldsymbol{x} &= [x, y]^T \end{aligned} \tag{15}$$

$I(x)$, resp. $J(x)$ reprezentuje intenzitu výřezu, který má uprostřed význačný bod, z aktuálního snímku, resp. z dalšího snímku. $\boldsymbol{d}$ je tzv. vektor přemístění a $\eta(x)$ je vliv šumu.

Cílem pro sledování je najít $\boldsymbol{d}$ takové, které minimalizuje součet čtverců rozdílů intenzit mezi $J(x)$ a $I(x\text{-}\boldsymbol{d})$ ve výřezu – to se dá reprezentovat takto:

$$\varepsilon = \iint [I(\boldsymbol{x} - \boldsymbol{d}) - J(\boldsymbol{x})]^2 \, \omega \, dx \tag{16}$$

kde ω je váhovací funkce. Pro krátké přemístění může být funkce intenzity odhadnutá pomocí Taylorovy řady takto:

$$\begin{aligned} I(\boldsymbol{x} - \boldsymbol{d}) &= I(\boldsymbol{x}) - g \cdot \boldsymbol{d} \\ g &= \left[\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right]^T \end{aligned} \tag{17}$$

Pokud rovnici 16 derivujeme podle $\boldsymbol{d}$ a určíme, že výsledek musí být roven nule tak finální rovnice přemístění se dá vyjádřit tímto způsobem:

$$G\boldsymbol{d} = e$$
$$G = \iint gg^T \omega dA \tag{18}$$
$$e = \iint (I - J)\, g\omega dA$$

kde G je matice koeficientů 2×2 a e je 2D vektor. Je důležité poznamenat, že pro zisk spolehlivého řešení rovnice 18 je nutné, aby byly vlastní hodnoty (*eigenvalues*) koeficientu matice G byly velké [13].

KLT příznak je definován jako střed výřezu V, kde je minimální vlastní hodnota G je větší než předdefinovaný práh. Tyto body jsou obvykle v rozích a nebo mají vysokou variaci intenzity gradientů. Po detekci těchto příznaků v prvním snímku videa jsou řídící body pro registraci vybrány manuálně abychom se vyhnuli nevhodným bodům vázaných na pohybující se objekty. Metoda sledování KLT potom sleduje vybrané příznaky ve videu a určuje jejich korespondenci pro výpočet zobrazení. Je použito afinní zobrazení pro zobrazení aktuálního snímku na první snímek – což vyžaduje čtyři zvolené příznaky. Mapování na první snímek zabraňuje kumulaci chyb mezi dvěma následujícími snímky. Tato metoda je vhodná pro eliminaci drobných pohybů dronu. Na obrázku 23 je ukázka kalibrace, zleva vidíme 1., 317. a 392. snímek namapovaný na první snímek. Červené body ukazují KLT řídící body.

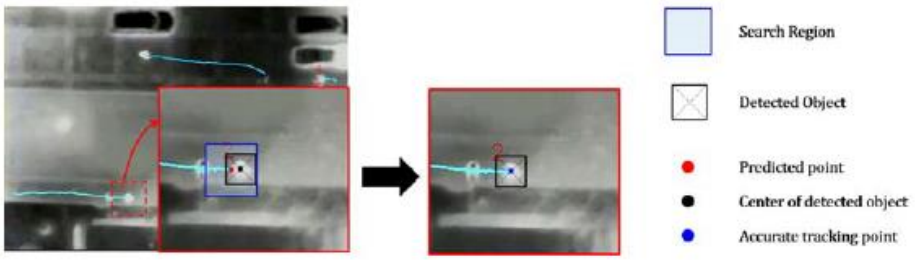
Obrázek 23: Ukázka zarovnání videa.

3.2.2 Metoda pro samotné sledování chodců

Metoda KLT je využita i pro samotné monitorování chodců. KLT příznaky jsou však obvykle na okrajích detekovaných objektů (tam je největší rozdíl intenzit), kvůli toho není tato metoda dostatečně přesná pro lokalizaci chodce. Tento problém je vyřešen tím, že je určen střed osoby na začátku sledování a potom pyramidová implementace Lucas-Kanade metody sledování je použita pro přesné nalezení přemístění. Pyramidová reprezentace je schopna pracovat s relativně velkým přemístěním (světelný tok je vypočten a propagován z vyšších do nižších úrovní). Vlivem nízké kvality infračervených snímků a šumu s pohybu dronu se může tento střed odchýlit v průběhu sledování. To je díky tomu, že předpoklad metod založených na optickém toku je konzistentní úroveň šedi sledovaných objektů. Tuto podmínku však termogramy nemohou splnit. Je tedy využitá další fáze detekce (lokální) ta spraví odchylku pomocí několika kroků:

1. Jakmile je nalezena osoba v aktuálním snímku jsou středové souřadnice vloženy na vstup metody pro sledování.
2. Použitím pyramidové Lucas-Kanade metody pro sledování je odhadnuto přemístění mezi následujícími snímky a tento odhad je využit pro predikci souřadnic v dalším snímku.
3. Kolem těchto předpokládaných souřadnic je provedena druhá detekce zaměřena na přesné určení osoby.

Tento proces je ukázán na obrázku 24. Pokud by se v oblasti kolem předpokládaných souřadnic nalezlo více objektů, bude propočítána minimální vzdálenost mezi středy objektů. Pokud bude tato vzdálenost menší, než polovina předdefinovaná velikost blobu bude daný střed považován za osobu. V situaci, kdy by se naopak nenašel objekt vůbec bude se předpokládat, že se nepohnul (a tedy souřadnice zůstávají stejné jako předtím). Získaný „přesný" střed bude použit pro opakování celé metody a sledování objektu v dalším snímku. Pro zvýšení přesnosti je detekce v celém obrázku provedena každých 15 snímků – tím jsou zároveň zachyceny nové osoby. Celkově se metoda podobá obvyklým metodám sledováním na základě detekce. Nicméně díky využitím natrénovaného klasifikátoru blobů (viz podkapitola 2.2.2) je celý proces o něco rychlejší a přesnější.



Obrázek 24: Představení získání přesného středu pomocí druhé detekce.

4 Experimenty

Tato kapitola popisuje vyhodnocení detektorů a sledování osob. Popisuje jednotlivé scénáře a experimenty. Rozhodnutí, jestli detekce byla správná či nikoliv, se uskutečňuje pomocí metody Intersection over Union (IoU). Je spočítán podíl průniku a sjednocení predikované oblasti a oblasti skutečně anotované. Oblast predikovaná je označena A a oblast skutečně anotovaná B. Výpočet pak vypadá následovně.

$$IoU(A,B) = \frac{A \cap B}{A \cup B} \tag{14}$$

Pokud podíl vyjde vyšší než 0,5 je detekce považována za úspěšnou a označena jako skutečně pozitivní (*true positive*). V opačném případě je označena za falešně pozitivní (*false positive*). Případ, kdy není anotovaný objekt nalezen vůbec, je označen jako falešně negativní (*false negative*). Z těchto ukazatelů lze vypočítat metriky *precision* a *recall* takto:

$$precision = \frac{true\ positives}{true\ positives\ +\ false\ positives} \tag{15}$$

$$recall = \frac{true\ positives}{true\ positives\ +\ false\ negatives} \tag{16}$$

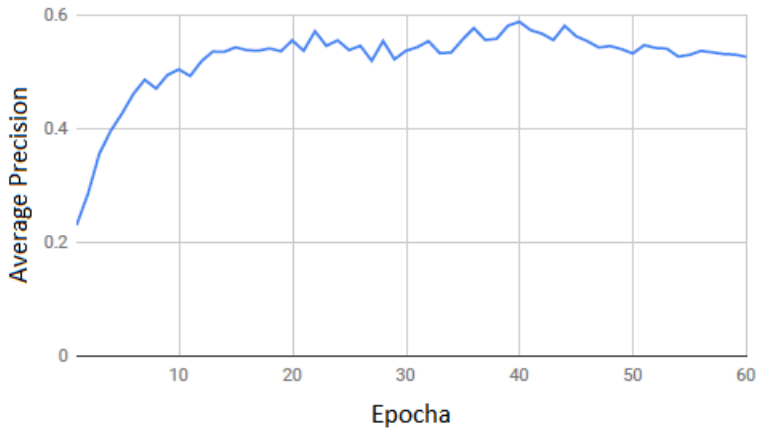
U hodnocení úspěšnosti sledování osob se zvolené metody liší. U viditelného osvětlení se jedná o manuální subjektivní zhodnocení na několika demonstračních videí. U termogramů byla využita metrika MOTA (viz podkapitola 4.2.2). Nicméně ta vyžadovala speciální podmínky, které se daly použít u relativně malého vzorku dat.

4.1 Vyhodnocení pro snímky s viditelným osvětlením

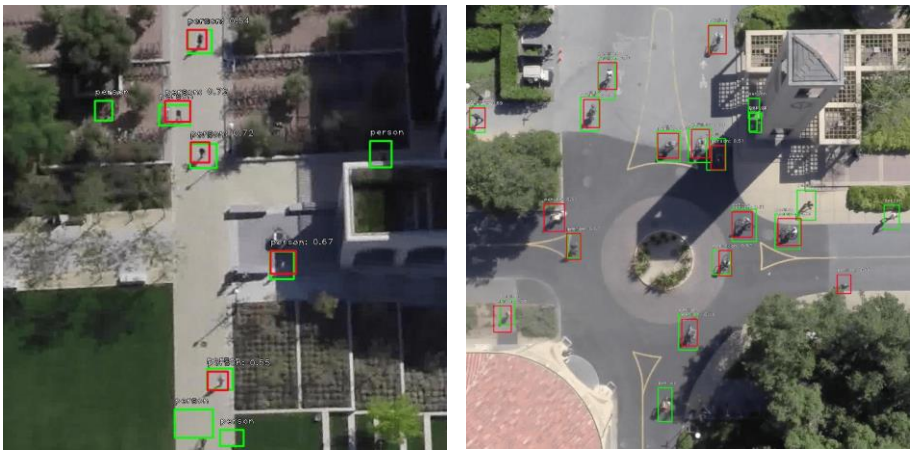
Vyhodnocení přesnosti detektoru bylo provedeno metrikou Average Precision (AP). Jedná se o způsob, který se používá na soutěžích Pascal VOC Challenge. Ten využívá předchozí vzorce pro *precision* a *recall*. Tvar křivky AP je vytvořen pomocí poměru mezi nimi $precision/recall$ [14]. Zhodnocení úspěšnosti reidentifikace osob a správnosti vykreslených trajektorií bylo provedeno subjektivně po otestování na několika demonstračních videích. Bylo vyhodnoceno, s čím má algoritmus problémy a kde naopak funguje. I tato podkapitola vychází z [1].

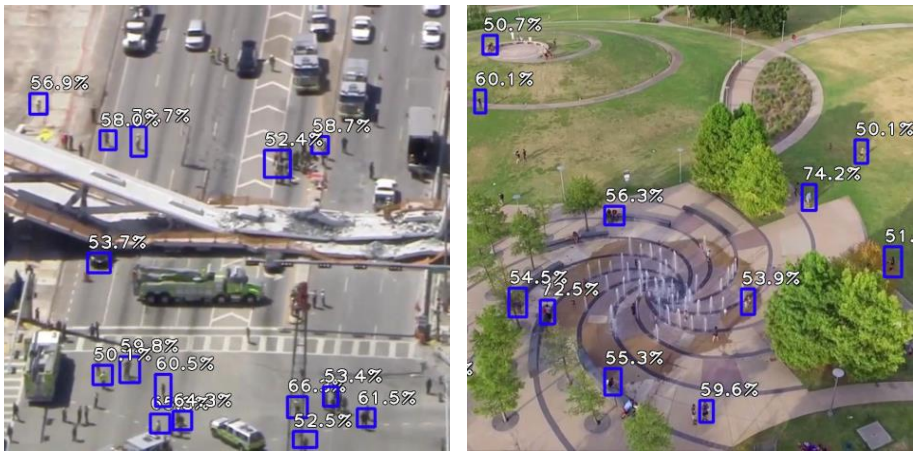
4.1.1 Úspěšnost a rychlost detektoru

Vyhodnocení přesnosti detektoru bylo měřeno na testovací části datasetu SDD (sekce 3.1). Přesnost byla měřena na modelech z finálního třetího trénování, kde se dosáhlo nejnižší chyby na testovacích datech. Nejvyšší přesnosti AP dosáhl model po 40. trénovací epoše a to sice 58,6 %. Graf *average precision* je vidět na obrázku 25. Na obrázku 26 je ukázka vyhodnocení přesnosti na testovacích datech. Zeleně jsou vyznačení anotované osoby, červeně rozpoznané. Obrázek 27 pak ukazuje snímky z YouTube, které nebyly vůbec součástí použitého datasetu SDD. Je vidět, že některé osoby nejsou detekovány, resp. některé objekty mohou rozpoznávání zmást.



Obrázek 25: Vývoj přesnosti v závislosti na množství epoch.


Obrázek 26: Vyhodnocení přesnosti na testovacích snímcích.



Obrázek 27: Ukázka detekce osob na snímcích mimo dataset (převzato z[2]).

[2] https://youtu.be/zJq5_7Au13o a https://youtu.be/7ns6fFhahbw

4.1.2 Úspěšnost reidentifikace lidí

Vyhodnocení úspěšnosti reidentifikace lidí a zakreslení správnosti jejich trajektorií probíhalo na záběrech z validační části použitého datasetu a na pořízených záznamech z dronu z výšky 35 m. Z datasetu byly vybrány takové záběry, kde je možnost reidentifikace aspoň trochu možná, tedy spíše záběry v lepší kvalitě a pořízené z nižší výšky. Dataset jinak obsahuje záběry z velmi vysoké výšky, případně v nízké kvalitě, kde je v podstatě člověk viděn pouze jako tmavá čmouha. Tam pochopitelně není možné od sebe jednotlivé lidi rozlišit.

Jako první je scéna **Coupa** jedná se o cca 10 s video. V pravé horní části obrazu se nachází po celou dobu několik sedících osob. V průběhu kolem nich projde další dvojice osob. V dolní části se pak pohybuje asi čtveřice osob, kdy jedna z nich se v záběru objeví až později. V pravém dolním rohu odchází ze záběru dvojice osob. Výsledek z této scény je na obrázku 28.

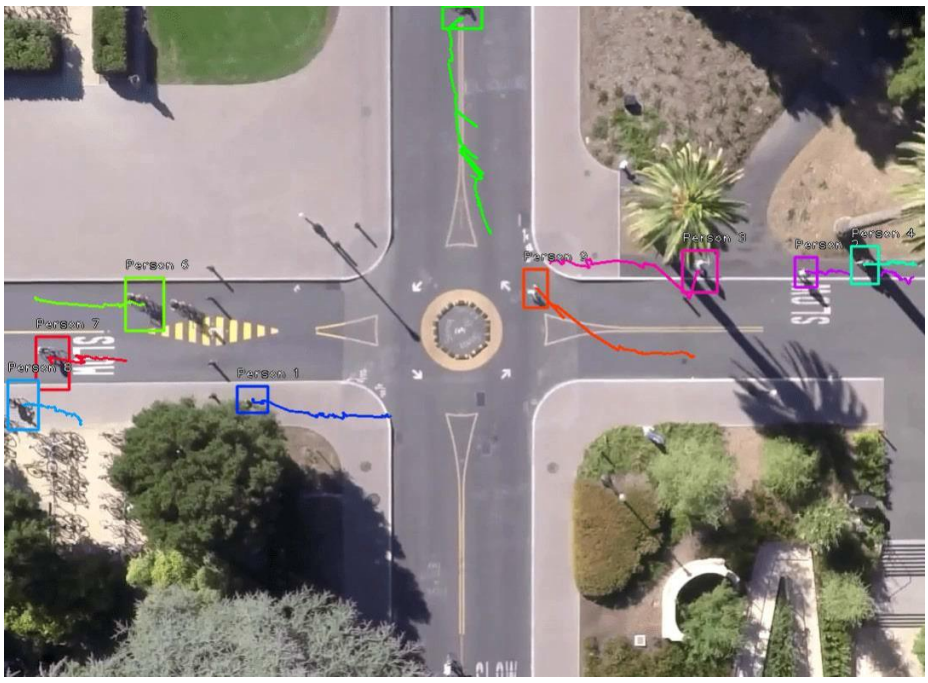
Pravý horní roh způsobil velký zmatek. Dvojice chodců, která procházela kolem skupinky sedících lidí, nebyla dostatečně rozlišena. Procházeli velmi blízko sebe a díky tomu, že není nijak vyřešeno oříznutí pozadí chodce a histogram je počítán pro celou výseč, je výsledek takto špatný. Vektory příznaků si jsou příliš podobné. Dolní část obrazu dopadla o poznání lépe. Všem čtyřem osobám byly trajektorie vykresleny správně. Osoby se potkaly, ale pomocí vektoru příznaků od sebe byly správně rozpoznány. Osoba 7 navíc vyšla ze stínu, algoritmus se přesto zachoval správně. Další problém nastal s dvojicí osob v pravém dolním rohu, která byla v obraze pouze chvilku a procházela přes částečný stín. Osoby nebyly úspěšně reidentifikovány, jejich příznakové vektory se napříč snímky příliš lišily.

Obrázek 28: Výsledná mapa s trajektoriemi osob ze scény Coupa.

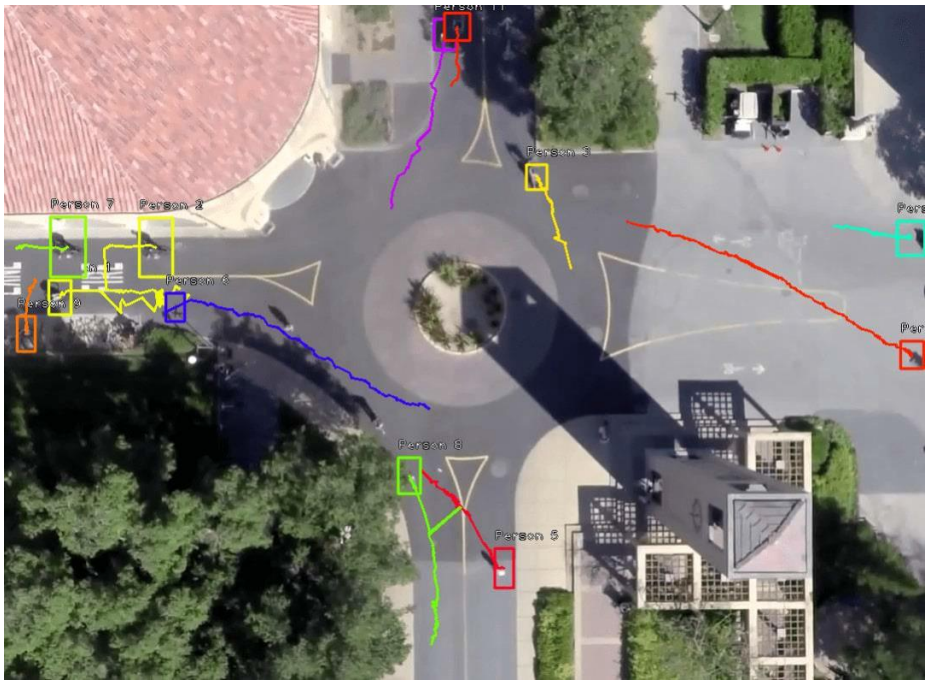
Vstupní 7 s video ze scény **Little**. V obrázku se pohybují chodci a cyklisté. Někteří procházejí skrz stíny, pod lehkým zákrytem stromů nebo do záběru vstupují až v průběhu záznamu. Výsledek zpracování této scény je vidět na obrázku 29. Cyklistu 6 síť nerozpoznala ihned, ale až po několika snímcích, proto

je jeho výřez vidět dvakrát. Výřez více vpravo je pozice cyklisty v prvním snímku. Označený výřez vlevo je moment, kdy byl sítí poprvé rozpoznán. Osoby 7, 8 se v záběru objevily až později a byly správně detekovány a reidentifikovány. Osoba 1 je na počátku schovaná pod stromem, v záběru se ukáže až v průběhu videa a je správně rozpoznána. Osoby 2, 4 projdou stínem, přesto jsou správně reidentifikovány. Osoba 3 projde stínem, částečně pod zákrytem stromu a je opět správně rozpoznána. Cyklista 9 se uprostřed křižovatky ztratí, síť ho přestane na několik snímků detekovat. Po jeho zpětném rozpoznání je už považován za jinou osobu, protože se mezitím dostal příliš daleko.

Vstupní 5 s video ze scény **Death Circle**. Záznam je podobného typu jako ze scény Little. V záběrech se objevují cyklisté a chodci, procházejí skrz stíny, případně pod zákryty stromů. Výsledná mapa trajektorií je na obrázku 30. Ve spodní části videa jdou proti sobě lidé označeni čísly 5 a 8. Osoba 8 se na krátký okamžik detektoru ztratí. Jako nejbližší a nejpodobnější detekce je pak vyhodnocena osoba 5, proto ta zelená linie k červené. V pravé horní části obrazu se vyskytuje 5 osob, jejichž detekce i predikce trajektorií je bezchybná. V levé části jede mimo záznam cyklista 2 a proti němu cyklista 1, později identifikován jako cyklista 6. V momentě, kdy si jsou nejblíže, nastává pro cyklistu 2 zmatek a jeho trajektorie uhýbá směrem k cyklistovi 1. O několik snímků později je cyklista 2 identifikován jako cyklista 7 a dále je jeho pohyb zaznamenán správně. Cyklista 1 je od doby, co není zaměňován s cyklistou 2 identifikován jako cyklista 6 a jeho trajektorie je také vizualizována správně. V této části se nachází ještě osoba 9, pro kterou je detekce korektní.



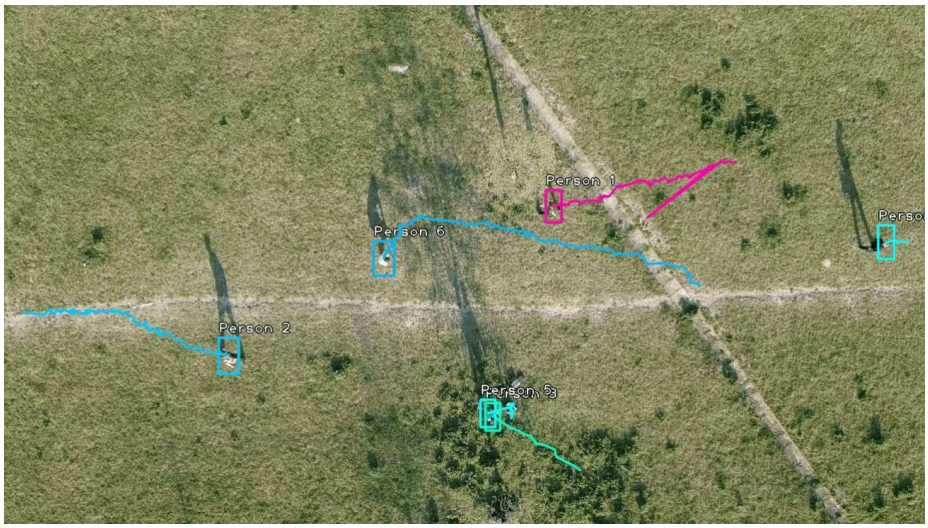
Obrázek 29: Výsledná mapa s trajektoriemi osob ze scény Little.



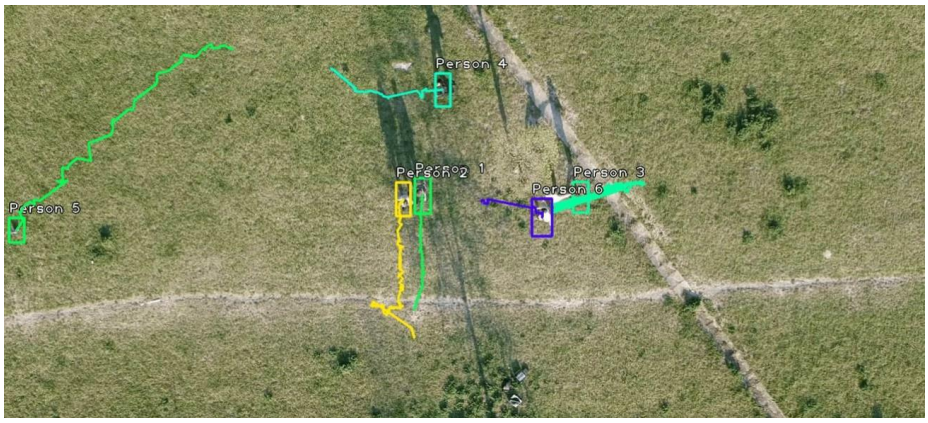
Obrázek 30: Výsledná mapa s trajektoriemi osob ze scény Death Circle.

Pro další experimentování a vyhodnocování úspěšnosti byly pořízeny záběry dronem. Skupinka osob se náhodně pohybovala po natáčeném prostoru. Vzniklo několik záběrů, které byly následně sestříhány a využity pro další otestování algoritmu. Hranice, pro vyhodnocení detekce neuronovou sítí jako úspěšné, byla snížena z výchozí hodnoty 0,5 na 0,4. **První záznam pořízený dronem** je 6 s video. Na začátku je v obraze 5 lidí a postupně všichni odcházejí mimo záběr. Výsledek z tohoto záznamu je zobrazen na obrázku 31. Všechny osoby byly v průběhu záznamu detekovány a jejich trajektorie zakresleny. Osoba 1 byla ke konci mylně spojena s jiným bodem. Pravděpodobně síť osobu na moment ztratila a zároveň rozpoznala jiný bod jako člověka o kus vedle. Osoba 5 a 3 jsou na počátku detekováni jako dvě osoby, jedná se pouze o jednu. Zbytek reidentifikací a vykreslení trajektorií je správné.

Druhý záznam z dronu je 5 s video. Ukazuje 6 osob pohybujících se v záběru. Výsledné trajektorie jsou vidět na obrázku 32. Osoby 1 a 2 jdou vedle sebe, jsou od sebe správně rozlišeni na základě vektoru příznaků a jejich trajektorie jsou korektně vykresleny. Osoby 6 a 3 běží od sebe. 6 je rozpoznána až později, nejdříve je několikrát mylně zaměněna s osobou 3. Výsledkem je několik dlouhých, rovných linií. Osoby 5 a 4 jsou reidentifikovány po celou dobu správně a jejich trajektorie jsou korektní.



Obrázek 31: Výsledná mapa s trajektoriemi osob z první scény z dronu.



Obrázek 32: Výsledná mapa s trajektoriemi osob z druhé scény z dronu.

4.2 Vyhodnocení pro termosnímky

Vyhodnocení detektoru osob na termosnímcích je založeno přímo na *precision* a *recall* v jednotlivých snímaných scénách. Jediné, co je ještě třeba definovat je podmínka pro určení vztahu mezi detekovanou a anotovanou plochou ve snímcích. U vyhodnocení sledování osob a zjištění jejich trajektorií byla využita metoda MOTA (viz dále 4.2.2). V [2], z čehož vychází tato podkapitola, je také analyzována trajektorie a odhad rychlosti.

4.2.1 Vyhodnocení úspěšnosti detektoru

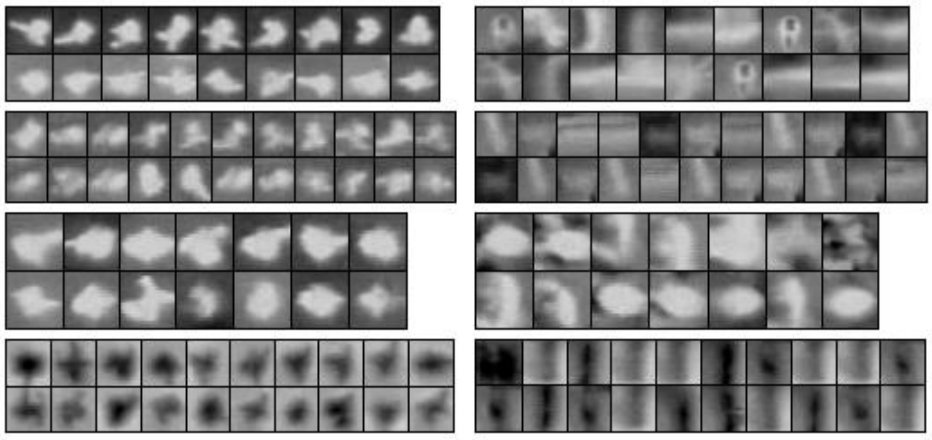
Při hodnocení úspěšnosti detektoru byl blob určen jako TP (*true positive*) pokud splňuje tuto podmínku:

$$\frac{A_{anno} \cap A_{det}}{min\{A_{anno}, A_{det}\}} \geq 0{,}7 \qquad (17)$$

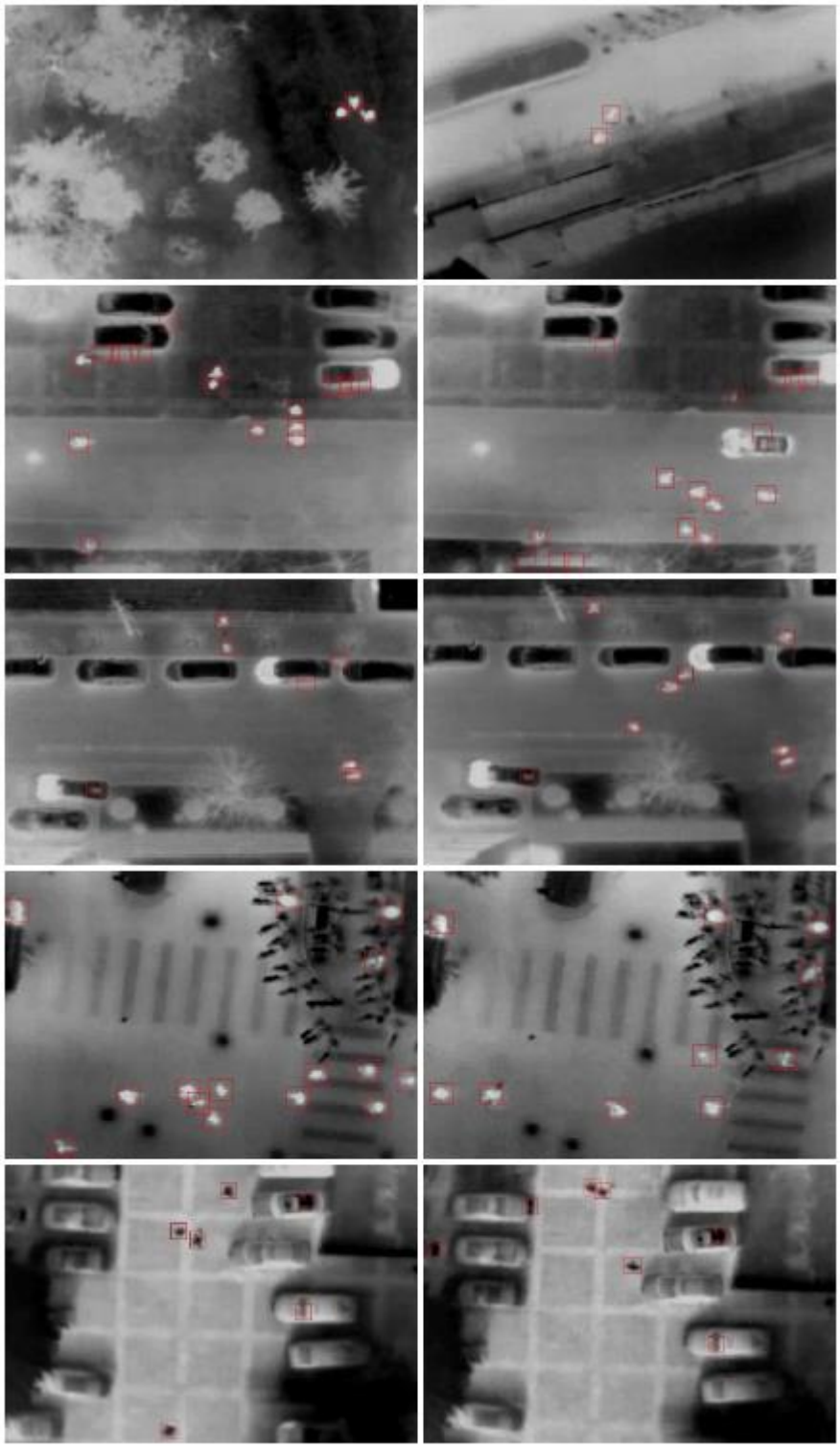
kde A_{anno} je anotovaná plocha a A_{det} je plocha určená detektorem. V tabulce 2 jsou zaznamenány výsledky ve všech scénách (včetně hodnot TP, FP a FN). Definice scén a průběhu učení již byla zmíněna v podkapitole 2.3.3. Nicméně i tak shrňme informace mající vliv na vyhodnocení. **Scéna 1** (testovací) definuje spíše kvalitní a dobře detekovatelné snímky, z tohoto důvodu je zde použit pouze detektor blobů, klasifikace se tak, na rozdíl od ostatních scén, nepoužívá. **Scéna 2** (zimní) obsahuje dobře rozlišitelné snímky hlavně kvůli vysoké intenzitě barvy (bílé) chodců. Podobné snímky obsahuje i **scéna 3** (ve výšce) nicméně u ní je vlivem výšky množství chodců rozmazaných. **Scéna 4** (jarní) se nakonec ukázala jako nejhorší. Díky nižší výšce letu jsou sice bloby větší a přehlednější, problémem je však vliv okolí. Lampy i světla aut se v této scéně zobrazují velmi podobně jako chodci a jejich rozlišení je problematické. Závěrem popis **scény 5** (letní) u ní dochází ke zvýšení okolní teploty natolik, že chodci již nejsou bílé skvrny ale černé. Zmizely však problémy s falešnými detekcemi, a tak jsou výsledky velmi podobné jako u scény 2. Pro lepší představu obsahuje obrázek 33 detekované bloby (osoby i falešné detekce) a obrázek 34 ukázkový snímek z každé scény.

Tabulka 2: Statistické výsledky detektoru chodců.

Dataset	Osoby	TP	FP	FN	Precision	Recall
Scéna 1	1320	1328	78	92	94,03 %	93,03 %
Scéna 2	9214	8216	686	998	92,30 %	89,17 %
Scéna 3	10029	9255	344	774	96,41 %	92,28 %
Scéna 4	5094	4322	396	772	91,61 %	84,84 %
Scéna 5	3175	2923	197	252	93,69 %	92,06 %



Obrázek 33: Detekované bloby (ze scén 2-5 – každý řádek jedna scéna). Vlevo jsou osoby (TP), vpravo falešné detekce (FP).



Obrázek 34: Ukázky snímků z jednotlivých scén (každý řádek jedna scéna).

Celkově vidíme relativně vysokou výkonnost, kdy se *precision* je vždy nad 90 % a *recall* se kolem této hodnoty pohybuje. Důležitá poznámka je, že i zde byly jako TP považování cyklisti, jelikož jejich rozlišení z výšky je velmi problematické.

4.2.2 Vyhodnocení úspěšnosti monitorování chodců

Při hodnocení samotného sledování chodců byla využita metrika MOTA (*multiple object tracking accuracy*). Ta je dle [15] definována takto:

$$MOTA = \frac{\sum_t (m_t + FP_t + mme_t)}{\sum_t g_t} \tag{18}$$

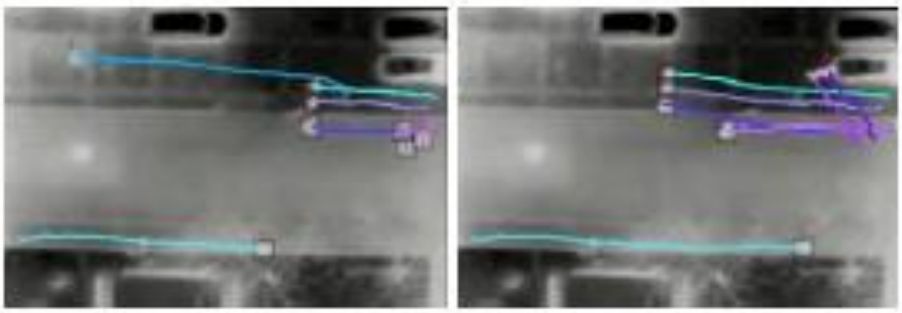
kde m_t je počet vynechaných chodců, FP_t je počet falešně pozitivních a mme_t je počet špatných shod, vše pro čas t. Poslední proměnná g_t ukazuje aktuální celkový počet chodců. Pro použití takové metody je ovšem nutné mít jednotlivé snímky tzv. registrované. Vzhledem k tomu, že video se hodně třese, pouze několik částí se povedlo zaregistrovat v pořádku. Popis těchto částí je v tabulce 3. U každé z nich byly vypočítaný příslušné proměnné, získané výsledky jsou pak přehledně v tabulce 4. Jak je vidět výsledky jsou celkem slušné průměrná hodnota MOTA se pohybuje kolem 0,86. Ukázky několika získaných trajektorií jsou na obrázku 35.

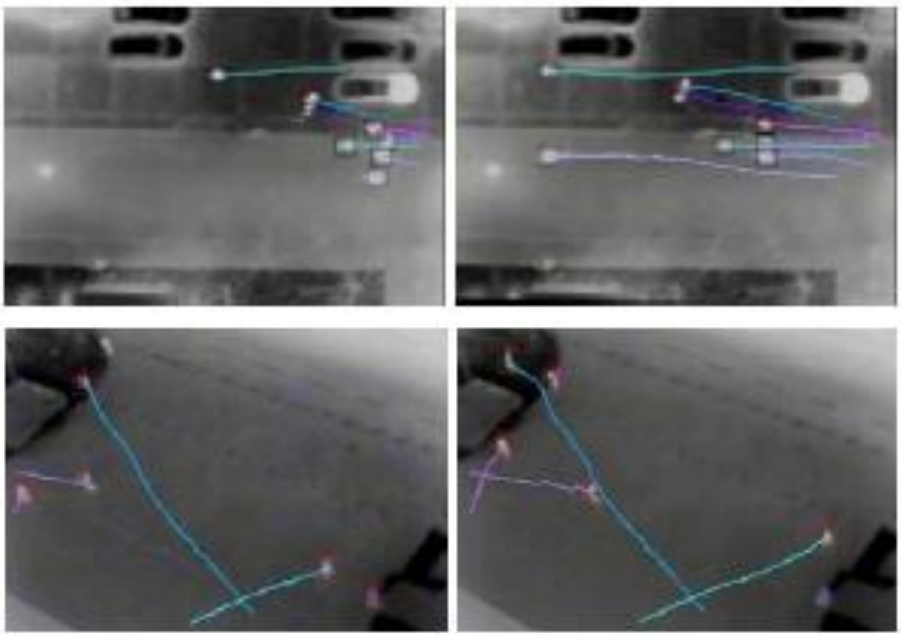
Tabulka 3: Informace ohledně vyhodnocovaných částí videa.

Dataset	**Rozlišení**	**Letová výška**	**Pozadí**	**Datum/Čas**	**Teplota**
Část 1	720×480	50 m	Cesta	20.1./ 20:22	5 °C
Část 2	720×480	50 m	Cesta	20.1./ 20:32	5 °C
Část 3	720×480	50 m	Náměstí	13.4./ 16:12	18 °C

Tabulka 4: Vyhodnocení sledování chodců pomocí metody MOTA.

Dataset	**Snímky**	***g***	***m***	***FP***	***mme***	**MOTA**
Část 1	635	3404	374	193	0	0,8334
Část 2	992	4957	161	597	0	0,8471
Část 3	1186	3699	318	0	0	0,9140





Obrázek 35: Ukázka monitorování chodců (každý řádek odpovídání jedné části).

4.3 Shrnutí výsledků obou přístupů

Tato podkapitola si klade za cíl porovnat a zhodnotit výsledky, kterých bylo dosaženo oběma metodami. Detektor u snímků s viditelným osvětlením dosáhl 59 % přesnosti (*precision*) a byl otestován i na snímcích z úplně jiných scénářů (tudíž i uhlů a komplexity prostředí). Termogramy ve stejné úloze měli přesnost cca 94 %, nicméně tyto výsledky byly dosaženy hlavně díky kvalitní klasifikaci kandidátních blobů (bez této klasifikace by se přesnost výrazně propadla). I termogramy byly testovány na scénářích mimo testované varianty s relativně slušnou přesností.

U sledování osob shodně oba zdroje uvádí, že provést verifikaci je velice obtížné. Na viditelných snímcích byla trajektorie analyzována manuálně a algoritmy postupně vylepšovány. V případě termogramů byla spočítaná metrika MOTA s výsledkem cca 0,86. Ovšem ta mohla být provedena jen na části z nasnímaných dat. Obě metody také došli k závěru, že rozlišit cyklistu, běžce a chodce jde pouze velmi obtížně. U termogramů byla snaha je rozlišit podle rychlosti, ovšem tato metoda nerozliší běžce a cyklisty.

5 Závěr

Zpráva se zabývá detekcí a monitorování lidí v obrazovém materiálu pořízeném z výšky. Detekované osoby jsou od sebe vzájemně rozlišeny a jejich pohyb je monitorován v průběhu celého videozáznamu. Po zpracování celého videa jsou trajektorie jednotlivých osob vizualizovány. Zpráva popisuje námi vyzkoušenou

metodu založenou na snímcích s viditelným osvětlením. Ta je popisována společně s metodou zabývající se podobnou problematikou založenou na termosnímcích.

U snímku s viditelným osvětlením byla k rozpoznání osob využita detekční síť RetinaNet. Předtrénované modely nebyly schopny detekovat lidi z dostatečné výšky. Z toho důvodu byl natrénován vlastní model na datasetu Stanford Drone Dataset. Proběhly celkem 3 trénování, kdy bylo s datasetem manipulováno na základě průběžných výsledků. U termosnímků se pomocí gradientů a geometrických vlastností získali bloby (skvrny). Ty reprezentují potencionální chodce (vyhledání funguje pro tmavé i světlé skvrny). Dále jsou pak klasifikovány za pomocí metod DCT, HOG a SVM.

Dále byl implementován algoritmus identifikace chodců, podle kterého byla každé osobě vykreslena trajektorie pohybu. Snímky s viditelným osvětlením využívají pro toto příznakových vektorů založených na barevných histogramech. Detekované objekty, které se. v průběhu videa nepohybovaly, jsou považovány za chybu detektoru a nejsou vizualizovány. Trajektorie pohybů osob jsou vykresleny do prvního snímku videa. Termosnímky využívají metodu KLT s tím, že jsou schopny obraz zarovnat.

Dosáhnutá přesnost detektoru snímků s viditelným osvětlením je 58,6 % na testovací části datasetu je spíše slabá. Tato nízká hodnota je způsobena jednak složitostí problému, kdy lidé z velké výšky mohou být snadno zaměnitelní za jiné objekty, tak také nepřesnými anotacemi použitého datasetu, a to i přes jeho ruční protřídění. Přesnost detektoru termogramů je 94 %, k tomu hodně pomáhá klasifikační část detektoru a samotný fakt, že se jedná o termosnímky kde je detekce o něco jednodušší než u viditelného osvětlení.

Identifikace chodců u snímků s viditelným osvětlením pomocí porovnávání jejich příznakových vektorů založených na barevných histogramech nefunguje úplně perfektně. Nijak není realizováno oříznutí pozadí kolem chodce. Histogramy tak jsou počítány pro celý segment vrácený detektorem. To znamená, že histogram se může výrazně změnit, pokud chodec změní pozadí. Proto si algoritmus identifikace pomáhá využitím vzájemné fyzické vzdálenosti detekovaných osob. Výsledné trajektorie jsou zaznamenávány do prvního snímku videa. Z toho vyplývá omezující podmínka pro korektní vykreslování trajektorií pohybu – pořízený záznam musí být statický. U termosnímků bylo dosaženo relativně slušných výsledků (0,88 dle metriky MOTA). I zde však byly problémy způsobené např. přílišnou blízkostí osob (což znesnadňuje samotnou detekci) nebo křížením trajektorií. Obě metody uznávali cyklisty jako chodce vzhledem k velmi problematickému rozlišení mezi těmito objekty.

Potenciální budoucí rozšíření může spočívat v implementaci nějaké formy *image-stitching* algoritmu. Pomocí něho by jednotlivé snímky videozáznamu mohly být pospojovány a vznikla by panoramatická mapa, která by zaznamenávala celou zabranou oblast ve videu. Celý program by pak fungoval i pro videozáznamy, kde se kamera pohybuje. Dosáhnutí vyšší přesnosti detektoru by bylo možné za použití lépe anotovaného datasetu. Dále by bylo možné inspirovat se u řešení použitého při termosnímcích a zkusit využít tyto metody.

Literatura

[1] Dušek, V.: *Monitorování chodců pomocí dronu*. Vysoké učení technické v Brně, Fakulta informačních technologií, Brno, 2019.

[2] Ma, Y.; Wu, X.; Yu, G.; aj.: *Pedestrian Detection and Tracking from Low-Resolution Unmanned Aerial Vehicle Thermal Imagery*. Sensor 2016, ročník 16, číslo 446, 2016. DOI 10.3390/s16040446.

[3] Kanich, O.; Dvořák, M.; Drahanský, M.: *Generátor a detektor modelů zbraní*, Vysoké učení technické v Brně, Fakulta informačních technologií, Brno, TZ k projektu VI20172020068 (VRASSEO), 2019.

[4] Drahanský, M., Orság, F., Doležel, M. aj.: *Biometrie*. Computer Press a.s., 2011, p. 294. ISBN 978-80-254-8979-6.

[5] Hui, J.: *Object detection: speed and accuracy comparison*. Březen 2018, [Online; navštíveno 5.5.2019]. URL https://medium.com/@jonathan_hui/object-detection-speed-andaccuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359

[6] Robicquet, A., Sadeghian, A., Alahi, A. aj.: *Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes*. In Computer Vision – ECCV 2016, Cham: Springer International Publishing, 2016, s. 549–565.

[7] Otsu, N.: *A Threshold Selection Method from Gray-Level Histograms*. IEEE Trans. Syst. Man Cybern. 1979, 9, 62–66.

[8] Bouguet, J.: *Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the Algorithm*. [Online navštíveno 10.11.2019] URL http://robots.stanford.edu/cs223b04/algo_tracking.pdf

[9] Xiao, J., Yang, C., Han, F., Cheng, H.: *Vehicle and Person Tracking in Aerial Videos. In Multimodal Technologies for Perception of Humans*. Stiefelhagen, R., Bowers, R., Fiscus, J., Eds.; Springer: Berlin/Heidelberg, Německo, 2008; pp. 203–214.

[10] Bhattacharya, S., Idrees, H., Saleemi, I., aj.: *Moving Object Detection and Tracking in Forward Looking Infra-Red Aerial Imagery*. In Machine Vision Beyond Visible Spectrum; Hammoud, R., Fan, G., McMillan, R.W., Ikeuchi, K., Eds.; Springer: Berlin/Heidelberg, Německo, 2011; pp. 221–252.

[11] Zhang, X.; Luo, H.; Fan, X.; aj.: *AlignedReID: Surpassing Human-Level Performance in Person Re-Identification*. CoRR, ročník abs/1711.08184, 2017.

[12] Shi, J., Tomasi, C.: *Good Features To Track*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle,WA, USA, 21–23 června 1994; pp. 593–600.

[13] Tomasi, C.: *Detection and Tracking of Point Features*. Tech. Rep. 1991, 9, 9795–9802.

[14] Everingham, M., Gool, L., Williams, C. K. aj.: *The Pascal Visual Object Classes (VOC) Challenge*. Int. J. Comput. Vision, Červen 2010: s. 303–338, ISSN 0920-5691.

[15] Bernardin, K., Stiefelhagen, R.: *Evaluating multiple object tracking performance: The CLEAR MOT metrics*. J. Image Video Process. 2008, 2008.