VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

IDENTIFIKACE CHODCŮ

PEDESTRIAN IDENTIFICATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **JAN JURČA**
AUTHOR

VEDOUCÍ PRÁCE **MICHAL HRADIŠ, Ing., Ph.D.**
SUPERVISOR

BRNO 2018

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav počítačové grafiky a multimédií (UPGM) Akademický rok 2018/2019

Zadání bakalářské práce

Student: **Jurča Jan**
Program: Informační technologie
Název: **Identifikace chodců**
Pedestrian Identification
Kategorie: Zpracování obrazu

Zadání:

1. Prostudujte základy neuronových sítí, konvolučních neuronových sítí a rekurentních sítí.
2. Vytvořte si přehled o současných metodách pro identifikaci chodců podle tváře, postavy a chůze.
3. Vyberte konkrétní metody a aplikujte je na úlohu rozpoznání chodců ve video sekvencích s fúzí jednotlivých zdrojů informace.
4. Obstarejte si datovou sadu vhodnou pro experimenty.
5. Implementujte navrženou metodu a proveďte experimenty nad datovou sadou.
6. Porovnejte dosažené výsledky a diskutujte možnosti budoucího vývoje.
7. Vytvořte stručné video prezentující vaši práci, její cíle a výsledky.

Literatura:

- Taigman et al.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. CVPR 2014.
- Parkhi et al.: Deep face recognition. Proceedings of the British Machine Vision 1.3 (2015): 6.
- Yang, Jiaolong, et al.: Neural aggregation network for video face recognition.arXiv preprint arXiv:1603.05474 (2016).
- Cao et al.: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, CVPR 2017.
- Zheng et al.: Pedestrian Alignment Network for Large-scale Person Re-identification. arXiv preprint arXiv:1707.00408, 2017.

Pro udělení zápočtu za první semestr je požadováno:

- Body 1 až 3.

Podrobné závazné pokyny pro vypracování práce viz http://www.fit.vutbr.cz/info/szz/

Vedoucí práce: **Hradiš Michal, Ing., Ph.D.**
Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2018
Datum odevzdání: 15. května 2019
Datum schválení: 1. listopadu 2018

Abstrakt

Tato práce se zabývá identifikací osob z videa na základě rozpoznání postavy, obličeje a chůze. Pro rozpoznání postavy a chůze jsou využity předtrénované sítě. Zatímco k rozpoznání chůze je v rámci práce implementováno a srovnáno několik architektur sítí. Finální rozpoznání chodce probíhá na základě multimodální fůze realizované neuronovou sítí. Pro účely práce byl vytvořen vlastní dataset, zároveň se sadou nástrojů umožňující jeho téměř automatickou tvorbu.

Abstract

This thesis deals with pedestrian identification from video sequence based on person, face and gait recognition. For person and face recognition are used pretrained networks. While for gait recognition is implemented and compared many different networks. Final pedestrian recognition is based on multimodal fusion realized by neural network. For the purpose of the work was created dataset, along with a set of tools that allow its almost automatic creation.

Klíčová slova

Rozpoznání tváře, rozpoznání postavy, rozpoznání chůze, multimodální fůze, hluboké neuronové sítě, dataset, konvoluční neuronové sítě, rekurentní neuronové sítě, siamské sítě, extrakce příznakového vektoru, BUT Pedestrians

Keywords

Face recognition, person recognition, gait recognition, multimodal fusion, deep neural networks, dataset, convolutional neural networks, recurrent neural networks, siamese network, feature vector extraction, BUT Pedestrians

Citace

JURČA, Jan. *Identifikace chodců*. Brno, 2018. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Michal Hradiš, Ing., Ph.D.

Identifikace chodců

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Ing. Michala Hradiše, Ph.D.. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Jan Jurča
14. května 2019

Poděkování

Rád bych chtěl velmi poděkovat mému vedoucímu práce Ing. Michalovi Hradišovi, Ph.D., za jeho ochotu, pomoc, cenné rady a investovaný čas, který mi poskytl při odborném vedení této bakalářské práce.

Obsah

Kapitola 1

Úvod

S rozvojem technologií v oblasti strojového učení a počítačového vidění, rostou možnosti v oblasti kamerových systémů. Jednou z úloh, kterou tyto systémy řeší je rozpoznání a identifikace snímaných lidí. Rozvoj kvalitních kamer zároveň umožňuje získat kvalitní záznam i na vzdálenosti desítek metrů. Pomocí takových záběrů lze provést rozpoznání člověka podle obličeje, postavy či chůze. Právě rozpoznáváním osob na základě fúze těchto modalit se zabývá tato práce.

Existuje mnoho projektů zaměřujících se na rozpoznání člověka podle obličeje nebo postavy. Většina z nich[18, 27, 11] pracuje na principu extrakce příznakových vektorů z jednotlivých snímků, které se následně agregují z celé video sekvence, za účelem získaní kvalitnějšího příznakového vektoru. Kromě rozpoznání podle obličeje a postavy, k čemuž jsem využil předtrénovaných sítí[2, 27], jsem se rozhodl realizovat rozpoznání na základě chůze. Konkrétně na základě sekvence 2D souřadnic 18 detekovaných kloubů na lidském těle. Rozpoznání chůze tedy sestává z detekce kloubů[4, 23], normalizace jejich souřadnic a následné extrakce příznaků. Na rozdíl od rozpoznání podle obličeje či postavy, kde stačí pro rozpoznání jediný snímek, je rozpoznání chůze náročnější z důvodu zpracování sekvence jako celku.

Finální identifikace je složena z rozpoznaných tří modalit, které jsou zpracovány neuronovou sítí realizující jejich multimodální fúzi, kde výstupem je příznakový vektor obsahující informace všech modalit. Je zde použit mechanizmus *pozdní fúze (late fusion)*[14], který pracuje až s konkrétními výstupy sítí jednotlivých modalit.

Přes existenci hned několika volných datových sad[26, 7, 16] určených k trénování identifikačních sítí, jsem nenalezl žádnou, která by splňovala všechny požadavky. Především požadavky na viditelnost tváře a minimální existence překážek v obraze, které komplikují detekci kloubů nutných pro rozpoznání podle chůze. Z toho důvodu byla v rámci práce vytvořena datová sada s celkem 659 identitami, které jsou nasnímány až ze čtyř kamer na třech různých natáčecích lokacích. Zároveň byla pro co nejrychlejší tvorbu datasetu vytvořena sada nástrojů, umožnující jeho téměř automatickou tvorbu.

Na vytvořené datové sadě jsou provedeny experimenty, které se zabývají srovnáním různých architektur sítí rozpoznávajících chůzi, vlivem délky sekvence pro rozpoznání chůze a vlivem obličejové modality na výsledky multimodální fúze.

Kapitola 2

Metody rozpoznání člověka a fúze modalit

Při realizaci rozpoznávání lidí pomocí strojového učení si obecně můžeme řešit tyto otázky.

"Kdo je osoba na nahrávce?" Jedná se o první otázku, která se v situaci neznámé snímané osoby nabízí. Řešením by mohl být systém, který by byl natrénován na všech lidech, které má být systém schopný rozpoznat, například k autorizaci vstupu do budovy. Výstupem této klasifikace by byla konkrétní identita daného člověka. Problémem takto vybudovaného systému by byla obtížná rozšiřitelnost o nové identity, protože při nutnosti identifikovat novou osobu by se musel celý systém znovu natrénovat. Proto je lepší položit jinou otázku.

"Je osoba z první nahrávky ta stejná jako z nahrávky druhé?" Je mnohem obecnější dotaz, který je ovšem schopen zodpovědět předešlou otázku. Řešením by mohl být systém, jehož vstupem by byly dvě nahrávky a výstupem by byla informace o míře podobnosti osob na vstupu. Tímto způsobem realizovaný reidentifkační systém by už netrpěl problémem špatné škálovatelnosti identit, ale potýká se z problémem nutnosti ukládání nahrávek, které slouží k následnému porovnání. Proto otázka, kterou řeší většina dnešních identifikačních systémů je následující.

"Jak efektivně vyjádřit identitu dané osoby?" Problém, který se řeší velmi intenzivně a často je právě extrakce příznaků definujících danou osobu. Jednim z prvních projektů[10], který používal tento přístup, realizoval extrakci příznaků nikoliv ze záznamů osob, ale z ručně psaných podpisů. Autoři článku definují co musí příznaky splňovat a jaké operace s nimi lze provádět. Extrakce příznaků je provedena takovým způsobem, aby výsledné příznaky, reprezentované ve formě příznakových vektorů (dále označováno jako deskriptory), splňovali několik kritérií. Hlavní vlastností deskriptorů je to, že podobnost dvou zkoumaných subjektů (v tomto případě podpisů) lze zjistit výpočtem vzdáleností jejich deskriptorů.

Výpočtem vzdálenosti deskriptorů extrahovaných ze snímků osob tedy získáme míru jejich podobnosti. Díky tomu je možné přízpůsobovat citlivost rozpoznávání pouhou změnou prahové hodnoty vzdálenosti.Prahová hodnota se nastavuje podle situace. V situacích, kdy není problémem vyšší počet tzv.: false positive rozpoznání, tedy rozpoznání, kdy jsou za stejné identity prohlášeny i rozdílné, je práh nastaven na benevolentnější hodnotu. Zatímco

při situacích, kdy má být rozpoznání co nejpřesnější a zároveň není problém, že některé výskyty pravděpodobně nebudou identifikovány, je nastavena prahová hodnota striktněji.

Díky tomuto přístupu není třeba uchovávat množství nahrávek a snímků identifikované osoby, ale pouze vyextrahované deskriptory. Tato extrakce deskriptorů byla použita v roce 1994[10], kdy byla nátrénována sít extrahující příznaky z ručně psaných podpisů.

2.1 Sítě pro extrakci příznaků

Příkladem využití sítí pro extrakci příznaků, může být případ kdy je cílem realizovat identifikaci lidí podle tváře. Za tímto účelem je možné natrénovat síť, která bude mít v trénovací sadě fotografie lidí zvolené tak, aby byl každý člověk ideálně zabrán z několika úhlů s různě zachycenou mimikou a v jiném osvětlení tak, aby se síť při trénování nenanučila přisuzovat váhu právě takovým znakům. Při identifikaci člověka nezáleží zda se člověk zrovna usmívá a nebo je zamračený, ale záleží na identitě, respektive vlastnostech tváře, podle kterých lze identitu odhadnout.

Sítě extrahující příznaky lze trénovat několika různými způsoby. Jedním z možných přístupů trénování vychází z trénovaní klasifikátoru[21]. Další dnes používanější technika je trénovaní pomocí siamských sítí[10, 18]. V obou případech je stejným cílem tvorba sítě, která je schopna popsat vstup takovým způsobem, aby jejich výstup charakterizoval vstupní data co nejpřesněji podle vlastností, které se snažíme modelovat.

Klasifikátor

Jednou z možností trénovaní sítí extrahujících příznaky je trénování pomocí klasifikátoru[21]. Jedná se o princip při nemž je trénována klasifikační síť jejíž poslední vrstvou je právě klasifikační plněpropojená vrstva. Tímto způsobem natérnovanou síť je možné použít k extrakci příznaků odstraněním poslední klasifikační vrstvy. Za vyextrahované příznaky jsou tedy prohlášeny aktivace předposlední vrstvy.

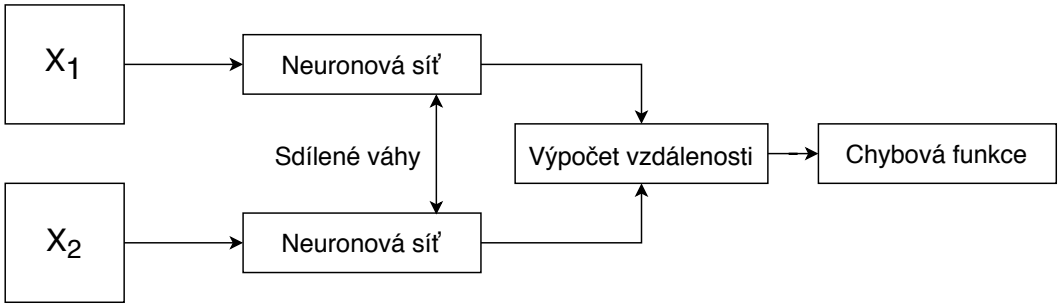
Siamské sítě

Princip siamských sítí spočívá v použití dvou identických sítí, které vzájemně mezi sebou sdílejí váhy. Celková architektura je tedy postavena tak, že na vstupu přijímá dva vstupy a na výstupu generuje pro každý vstup zvlášť výstup, deskriptor, daného vstupu. Stěžejní myšlenkou je fakt, že vzájemná vzdálenost tímto způsobem vygenerovaných deskriptorů vyjadřuje vzájemnou podobnost vstupů na základě vlastností, které se síť naučila rozpoznávat.[10]. Architektura siamské sítě je znázorněna na obrázku 2.1.

Trénovaní siamských sítí probíhá prostřednictvím chybové funkce, *contrastive loss*.

Contrastive loss[10, 6]. Chybová funkce pracující se vzdáleností mezi dvěma deskriptory. Pokud je vstupní dvojice označena jako shodná pak se výsledná vzdálenost mezi deskriptory minimalizuje. Naopak při rozdílných vstupech se vzdálenost mezi deskriptory maximalizuje.

Nechť $\vec{X_1}$ a $\vec{X_2}$ jsou vstupy siamské sítě a Y je binární ohodnocení daného páru. $Y = 0$ v případě shodných vstupů a $Y = 1$ v případě vstupů rozdílných. Nechť D_W je trénovatelná



Obrázek 2.1: Ilustrace architektury siamské sítě, která je složena ze dvou sítí sdílejících mezi sebou váhy. Vstupem sítě jsou vektory x_1 a x_2. Převzato z [6].

funkce vzdálenosti mezi výstupy siamské sítě G_W. Pak platí vztah

$$D_W(\vec{X_1}, \vec{X_2}) = ||G_W(\vec{X_1}) - G_W(\vec{X_2})||_2, \tag{2.1}$$

který je využit v chybové funkci

$$L(W, Y, \vec{X_1}, \vec{X_2}) = (1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{\max(0, m - D_W)\}^2, \tag{2.2}$$

kde zápis $D_W(\vec{X_1}, \vec{X_2})$ je zkrácen na D_W.

Tripletové sítě

Mechanizmus použití dvou sítí byl později rozšířen na trénování prostřednictvím tří shodných sítí [8].

Autoři článku[8] popisují tento typ sítí jako specifický typ siamské sítě. Tripletová síť sestává ze tří siamských sítí, které mezi sebou sdílí váhy. Jak je ilustrováno na obrázku 2.2, vstupem této sítě jsou tři vektory označované jako *anchor* (x), *positive* (x^+) a *negative* (x^-). Výstupem jsou vzdálenosti extrahovaných deskriptorů x a x^+ a deskriptorů x a x^-. Tento vztah lze zapsat jako

$$TripletNet(x, x^-, x^+) = \begin{bmatrix} \|Net(x) - Net(x^-)\|_2 \\ \|Net(x) - Net(x^+)\|_2 \end{bmatrix}. \tag{2.3}$$

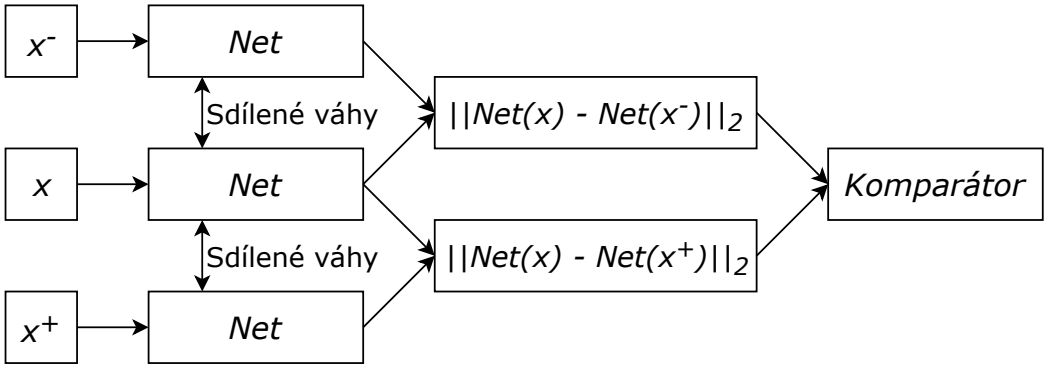
Triplet loss[**8, 18**]. Při trénovaní je vstupem trojice vektorů x, x^+, x^-, kde x a x^+ reprezentují subjekt stejné třídy a x^- reprezentuje subjekt třídy jiné. Cílem je tedy minimalizovat vzdálenosti mezi dekriptorem pro x a deskriptorem pro x^+ a zároveň maximalizovat vzdálenost mezi deskriptorem x a deskriptorem x^-. Matematický zápis toho vztahu je

$$\|f(x_i) - f(x_i^+)\|_2^2 + \alpha \quad < \quad \|f(x_i) - f(x_i^-)\|_2^2 \tag{2.4}$$

$$\forall(f(x_i), f(x_i^+), f(x_i^-)) \quad \in \quad \mathcal{T}, \tag{2.5}$$

kde α je minimální vzdálenost mezi negativními a pozitiními páry a $\mathcal{T}$ je množina všech možných tripletů s kardinalitou N. Finální chybová funkce, jejíž hodnota se v průběhu trénovaní minimalizuje, je

$$Loss = \sum_i^N [\|f(x) - f(x^+)\|_2^2 - \|f(x) - f(x^-)\|_2^2 + \alpha]. \tag{2.6}$$



Obrázek 2.2: Architektura tripletové síte složené ze tří siamských sítí. Převzato z [8]

Hard mining[18]. Je technika úzce související trénovaním pomocí *triplet loss*. Slouží ke zefektivnění a zrychlení trénovacího procesu tím způsobem, že trénovací data jsou záměrně volena složitá. Složitá data jsou volena ze špatně klasifikovaných dat z předešlých tréninkových iterací. Technika se dělí na *hard negative* a *hard positive* mining. *Hard negative mining* se zaměřuje na data v nichž je špatně rozpoznán vztah mezi *anchor* a *negative*. Zatímco *hard positive mining* se zaměřuje na špatně rozpoznaný vztah mezi *anchor* a *positive*.

2.2 Fúze několika modalit

Jedním z cílů této práce je provést fúzi výstupů z několika sítí, které identifikují člověka na základě jedné modality. Konkrétně se zaměřuji na tři modality a to obličej, postavu a chůzi. Aby bylo možné provést identifikaci na základě všech tří modalit je možné postupovat několika způsoby.

Jednoduchou metodou fúze je vyhodnocení každé modality zvlášť a následný výběr nejčastějšího rozhodnutí. Zjednodušeně by se dalo mluvit o technice: **"Rozhodne většina."**. Pro každou modalitu je určena prahová hodnota vzájemné podobnosti (často vyjádřená vzdáleností) mezi deskriptory, tedy hodnotu která určuje zda-li se jedná o stejnou či rozdílnou identitu. Každá modalita je tímto posouzena a konečný výsledek je určen většinovým rozhodnutím.

Přestože se jedná o metodu jednoduchou, nejedná se o metodu ideální. Především z důvodu absence mechanizmu, který by posuzoval všechny dodané informace jako jeden navzájem se doplňující celek. Další nedostatek vychází z použití tohoto mechanizmu při identifikaci podle existující databáze, ve které by bylo nutné uchovávat deskriptor pro každou modalitu zvlášť. S tím souvisí i problém určení prahových hodnot pro každou modalitu zvlášť, což může být problém, při změně sledovaného prostředí, protože podle toho v jakém prostředí je člověk zachycen se mění i kvalita zachycení určitých modalit a tedy i jejich váha.

fúze jediné modality. Specifickým typem fúze je fúze jediné modality tzv.: agregace. Agregace deskriptorů je často prováděna při extrakci příznaků z videa, kdy je pro každý snímek vypočítán deskriptor osoby a následně se ze všech zpracovaných snímků se vytváří

jeden zástupný deskriptor pro dané video. Takovou agregaci lze provést několika způsoby, od jednoduchého průměru všech deskriptorů, po využití natrénované agregační sítě, popsané v článku o neuronové agregační síti[25].

Autoři tohoto článku[25] realizují agregaci deskriptorů na základě adaptivního přidělení váhy každému jednotlivému deskriptoru podle kvality. I když v ideálním případě by informace o kvalitě snímku neměla být obsažena ve výsledném deskriptoru osoby, z praxe se ukazuje, že opak je pravdou.

Multimodální fúze. Problematika fúze deskriptorů několika modalit je velmi podobná jako problematika agregace nad jednou modalitou. Přesto je zde několik rozdílů oproti agregaci, které znemožňují použití některých agregačních metod. Tyto rozdíly vycházejí především z vlastností samotných deskriptorů, které se liší svojí sémantikou a dimenzí. Často nastává i případ kdy informace určité modality chybí úplně. Z těchto vlastností vyplývá, že za učelem fúze není vhodné vektory sčítat či průměrovat a to především z důvodu rozdílné sémantiky dat.

2.3 Metody multimodální fúze

Různým metodám fúze neuronových sítí se věnuje článek *Learn to Combine Modalities in Multimodal Deep Learning*[14]. Z tohoto článku vyberu metody a principy, kterými se budu zabývat v rámci této práce.

Přístupy multimodální fúze lze rozdělit na brzkou fúzi (*early fusion*) a pozdní fúzi (*late fusion*).

Early fusion[22]. Hlavním principem daného přístupu je použití jednoho modelu, který zpracovává všechny modality zároveň. Díky tomu je právě tento jeden model schopen lépe modelovat nízkoúrovňové korelace mezi vstupními daty. Zároveň je ale téměř vždy nutné provádět předzpracování dat jednotlivých vstupních modalit tak, aby tvořily zarovnanou a ideálně sémanticky podobnou datovou strukturu. Velmi podrobně se různým realizacím *early fusion* věnuje článek *CentralNet: a Multilayer Approach for Multimodal Fusion*[22].

Late fusion[14]. Jedná se o obecný přístup, kdy provádíme fúzi výstupů několika oddělených modelů. Každý model provádí nezávislé rozpoznání pro danou modalitu a fúze probíhá na úrovni výsledků z těchto oddělených modelů. Realizace fúze může probíhat několika způsoby(např.: průměrování při agregaci, rozhodnutí dle většiny, natrénovaná fúzní neuronová síť), přičemž výběr vhodné metody je vždy závislý na sémantice dat a počtu jednotlivých modalit. Hlavní výhodou využití přístupu *late fusion* je právě nezávislost unimodálních modelů a tedy možnost využití modelů, které již existují a k fúzi s jinými modely ani nebyly určeny. Matematicky lze přístup *late fusion* chápat jako

$$p = F(h_1(v_1), \ldots, h_m(v_m)), \tag{2.7}$$

kde p je výsledek fúze, F představuje fúzní metodu, $h_{1\ldots m}$ jsou jednotlivé unimodální modely a $v_{1\ldots m}$ jsou vstupy unimodálních modelů (např.: snímek obličeje, postavy nebo záznam pohybů chodce).

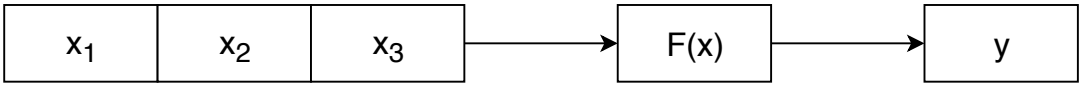
Fúzní neuronová síť[14]**.** Jak bylo zmíněno, jednou z možných realizací *late fusion* je neuronová síť. Vstupem takové neuronové sítě jsou deskriptory jednotlivých modalit a výstupem je jeden deskriptor reprezentující informace všech vstupních modalit. Hlavním účelem fúzní

neuronové sítě je transformace vektorových prostorů jednotlivých modalit do vektorového prostoru společného pro všechny vstupní modality.

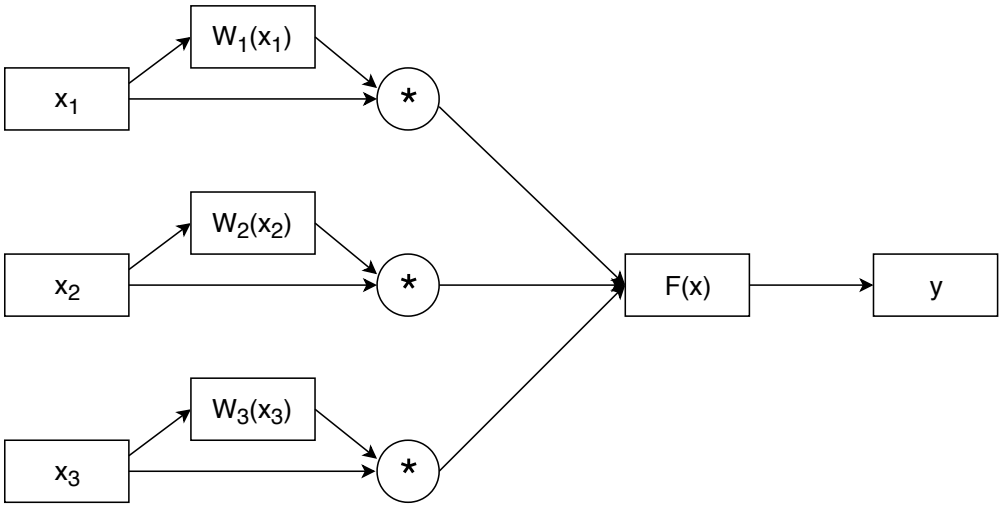
Architektury neuronových sítí, realizujících multimodální fúzi, bývají často minimalistické a rychle natrénovatelné. Velmi se podobají agregačním sítím, jako již zmíněné *Neural aggregation network*[25], právě z důvodu řešení podobného problému. Nejjednoduší architektura fúzní sítě je vyobrazena na obrázku 2.3 a jejím vstupem jsou deskriptory všech modalit zřetězené za sebou bez jakékoliv sémantického rozdělení.

Komplexnější architektura je vyobrazena na obrázku 2.4, kde je vidět plně propojená vrstva zvlášť pro každou modalitu, která slouží k adaptivnímu výpočtu váhy, takto vypočítanou váhou se vynásobí deskriptor dané modality. Následně se ováhované deskriptory zřetězí a poslouží jako vstupy plně propojené vrstvy, která zrealizuje jejich finální fúzi.

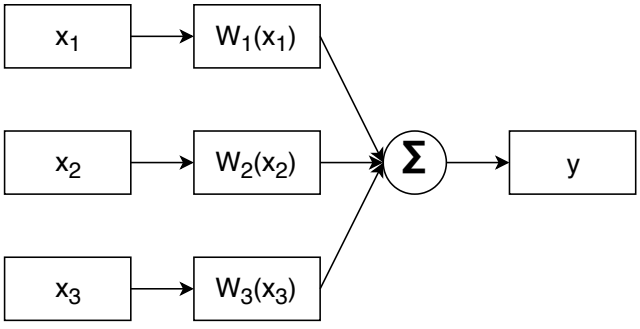
Poslední architektura, se kterou byly v rámci práce provedeny experimenty je vyobrazena na obrázku 2.5. V případě této architektury má každá modalita vlastní vrstvu jejíž úkol je transformace vektorového prostoru příslušné modality do vektorového prostoru jednotné dimenzionality pro všechny modality. Pro výslednou fúzi se provede součet všech transformovaných deskriptorů.



Obrázek 2.3: Minimalistická fúzní síť, jejímž vstupem jsou zřetězené deskriptory všech modalit. $F()$ zde představuje plně propojenou vrstvu. V experimentech označováno jako Simple



Obrázek 2.4: fúzní síť, která adaptivně určí váhový vektor pro každou modalitu a následně provede fúzi takto ováhovaných vektorů. $F()$ a $W_i()$ zde představují plně propojené vrstvy. Inspirováno z [25]. V experimentech označováno jako Concat Fusion



Obrázek 2.5: fúzní síť, která provádí trasformaci deskriptorů jednotlivých modalit do společného vektorového prostoru. $F()$ a $W_i()$ zde představují plně propojené vrstvy. V experimentech označováno jako Sum Fusion

Kapitola 3

Datová sada a tvorba datasetu

Protože se v rámci práce zabývám identifikací člověka na základě fúze několika modalit, je nutné, aby v datové sadě byly jednotlivé modality zřetelně viditelné, pro co největší množství identit. Tato podmíka není problém v případě identifikace člověka na základě postavy, ale podstatně větší problém nastává při extrakci tváře a pozic jednotlivých kloubů v obraze. Požadavky na získaná data jsou tedy tři. Konkrétně záznam na lidské postavy z několika kamer zároveň, ideálně se zřetelným zaběrem obličeje a minimálním počtem rušivých překážek v obraze.

3.1 Existující datové sady

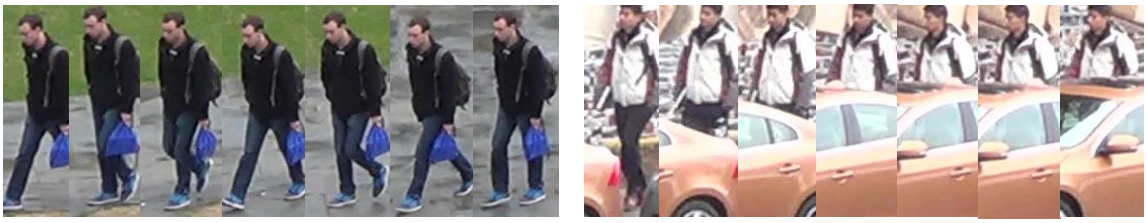
Existuje několik volně dostupných reidentifikačních datových sad zaměřujících se na rozpoznání dle postavy. Jako demonstraci dat, jenž už existují a zároveň ke srovnání jsem vybral čtyři datové sady, které mají velké množství identit a také se blíží tomu, jakou datovou sadu potřebuji. Konkrétně se jedná o datové sady MARS[26], iLIDS-VID [15, 24, 20, 19], PRID2011[7] a DukeMTMC-VideoReID[16]

Z vybraných stop na obrázcích 3.1, 3.2, 3.3 a 3.4 je patrné, že zmíněné datasety nebsahují dostatečná data, která jsou k řešení zadané úlohy potřeba. Především z důvodu nízké kvality snímků s čímž souvisí problém rozpoznání podle obličeje člověka, což téměř znemožňuje použití těchto datasetů k natrénování fúzního modelu.

Obrázek 3.1: Ukázka části sekvence z datasetu MARS[26]

Obrázek 3.2: Ukázka části sekvence z datasetu iLIDS-VID [15, 24, 20, 19]

Obrázek 3.3: Ukázka části sekvence z datasetu PRID2011[7]



Obrázek 3.4: Ukázka části sekvence z datasetu DukeMTMC-VideoReID[16]

3.2 Tvorba datasetu

Z výše zmíněných důvodů jsem se rozhodl pro tvorbu vlastního datasetu, který se bude snažit eliminovat nedostatky ostatních datasetů. Tvorba datasetu zahrnuje terénní natáčení dat z několika kamer současně, synchronizaci kamer, perspektivní transformace souřadnic mezi jednotlivými kamerami, detekci osob a samotné vytvoření strukturovaných dat využitelných k trénování a testování sítí.

Obrázek 3.5: Snímek z natáčení v Brně na Nových sadech

V průběhu řešení práce jsem realizoval tři natáčení, přičemž každé natáčení bylo na jiné lokaci. Vždy byly použity čtyři kamery, tři schopné natáčet ve 4K a jedna FullHD.

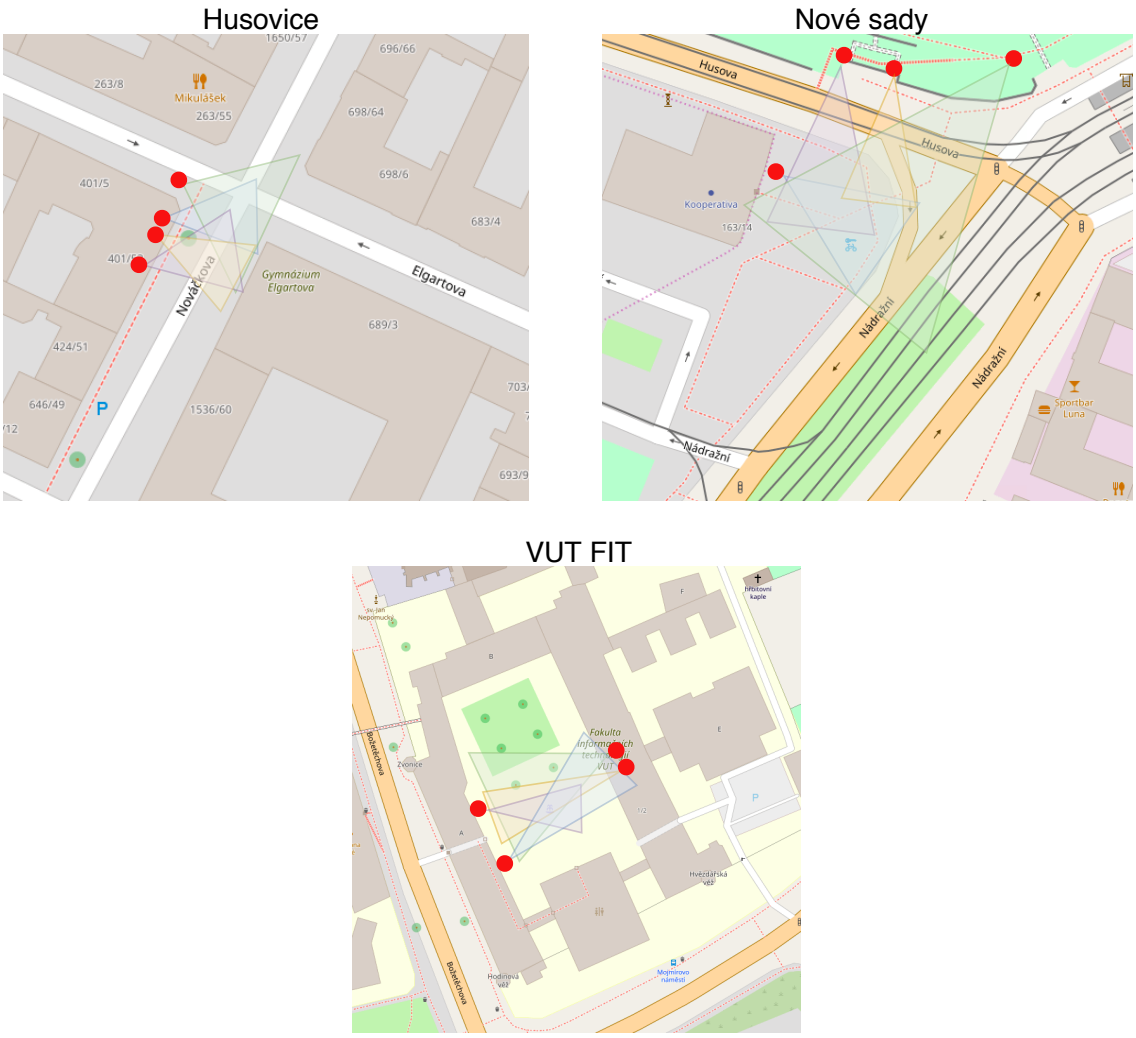
Výběr nahrávací lokace a průběh natáčení

První fází tvorby datasetu je výběr natáčecích lokací. Aby se na určitém místě dalo efektivně natáčet, mělo by splňovat hned několik kritérií:

1. vysoká koncentrace lidí
2. možnost natáčení z vyvýšeného místa
3. lidé by na místě neměli postávat, ale spíše procházet
4. možnost natáčení z různých úhlů pohledu
5. snímaná oblast by měla být co nejrovnější

Většinu kritérií splňuje i prostor náměstí mezi budovami A,B,C a D na VUT FIT, proto první zkušební natáčení proběhlo právě na fakultě v čase účelně vybraném tak, aby bylo v natáčené lokaci co nejvíce lidí. S pomocí nahraných dat jsem vytvořil sadu nástrojů pro téměř automatickou extrakci dat, kterými jsem nahraná data zpracoval. Díky úspěšnému prvnímu natáčení, jsem realizoval ještě dvě další natáčení v centru města. Mapy a rozmístění kamer je zobrazeno na obrázku 3.6.

Rozmístění kamer. Na vybrané natáčecí lokaci, je potřeba v daném místě rozmístit kamery. Kamery by měly být rozmístěny takovým způsobem, aby většinově zabíraly stejný

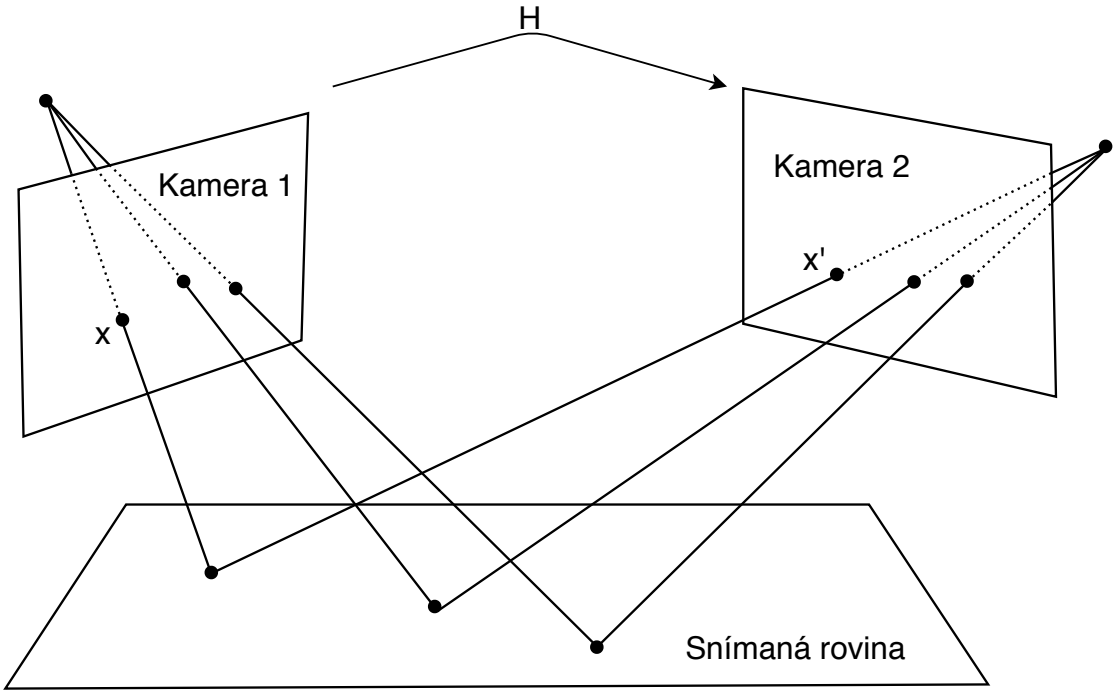


Obrázek 3.6: Ilustrace rozmístění kamer a pohledů na jednotlivých natáčecích lokacích.

prostor. Jedna z kamer plní funkci přehledové kamery a zabírá celkový pohled na snímanou scénu. Ostatní kamery jsou nastaveny na snímání detalnějších pohledů snímané lokace.

Průběh natáčení. Na začátku záznamu je vhodné zkalibrovat kameru. Kalibrace lze provést natočením kalibračního vzoru umístěného na rovné kalibrační tabulce. Díky kalibraci lze eliminovat nežádoucí zkreslení objektivu, především tzv.: soudkovitost objektivu. Přestože je kalibrace vhodná, ukázalo se že není nezbytná, a proto při zpracování nebyla využita.

Protože je velmi obtížné spustit nahrávání na všech kamerách současně,je nutné je synchronizovat. Tato synchronizace byla provedena jednoduše, tlesknutím ve výhledu všech kamer současně.



Obrázek 3.7: Princip homografie mezi dvěma kamerama, které snímají jednu rovinu. Pomocí matice homografie H lze provést přepočet souřadnic bodu x v první kameře a získat souřadnice odpovídaícího bodu x' z kamery druhé. Převzato z [1].

Prvotní zpracování záznamů - synchronizace a perspektivní transformace

Po natočení je nutné záznamy upravit k dalšímu zpracování. První úprava je časová synchronizace záznamů na základě výše zmíněného tlesknutí. Tato synchronizace byla provedena ořezáním videa o specifickou délku, tak aby tlesknutí bylo synchronizované mezi všemi záznami. Po ořezu začátku videa byly, z důvodu nutnosti stejné délky záznamu, oříznuty videa i na konci.

Aby bylo možné provádět přepočty souřadnic podkladové roviny mezi přehledovou kamerou a detailními kamerami, byly vypočítány matice homografií mezi kamerami.

Homografie[1]. Matice homografie je transformační matice o velikosti 3×3. Jak znázorňuje rovnice 3.1, máme-li bod x na podkladové rovine v jedné kameře a provedeme maticové násobení daného bodu s maticí homografie, výsledkem bude bod x' v obraze kamery druhé, který bude odpovídat stejnému místu podkladové roviny. Matematický zápis je

$$\begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ 1 \end{bmatrix}. \tag{3.1}$$

Daný přepočet pomocí matice funguje pouze jednosměrně, tedy jedna matice homografie H určuje převod souřadnic z kamery č. 1 do kamery č. 2 nikoliv obráceně. Pro převod souřadnic z kamery č. 2 do kamery č. 1 je nutné určit inverzní matici H^{-1} s níž lze provést zpětný přepočet. Názorný princip homografie je zobrazen na obrázku 3.7

Obrázek 3.8: Detekované pózy pomocí openpose [4, 23] sítě

Detekce lidí a spojování video stop z několika kamer

Další nezbytnou částí tvorby datasetu je detekce lidí v záznamu. V rámci práce využívám i pózy lidí, tyto pózy se skládají až z 18 detekovaných bodů na lidském těle. Pro detekci jsem se rozhodl využít fungující síť vycházející z projektu Realtime Multi-Person Pose Estimation[4, 23], konkrétně implementaci z projektu *tf-pose-estimation*[12]. Výsledek detekce je znázorněn na obrázku 3.8.

Pro co nejjednodušší a nejrychlejší tvorbu datasetu jsem vytvořil nástroj na automatické zpracování a párování detekovaných stop ze čtyř kamer současně. V rámci zpracování nahrávek do formy datasetu probíhá výpočet spojitých stop jednotlivých osob, filtrace velmi krátkých stop, spojení stop mezi kamerami a tvorba výřezů.

Tvorba spojitých stop jednotlivých osob v obraze. Protože výstupem samotné detekce je pouze pozice detekovaných póz v daném snímku a nikoliv informace o tom, která póza ze snímku x odpovídá póze ze snímku $x+1$, je nutné provést spojení. Pro toto spojení je použita detekovaná pozice krku postavy. U této pozice se stanoví okolí, v němž se v dalším snímku hledá nejbližší detekovaný bod. Takový bod je spárován s bodem předešlým a celý proces se opakuje i pro další snímek. Výstupem tohoto zpracování je seznam stop, pro každou kameru zvlášť.

Filtrace spojených stop. Po spojení stop se odstraní příliš krátké stopy, které většinou vznikají chybnou detekcí kloubů. Tato špatná detekce bývá způsobená překrytím postavy jinou postavou nebo překážkou v obraze.

Spojení stop v rámci několika kamer s ohledem na přehledovou kameru. Právě při spojování stop mezi kamerami jsou použity přepočty souřadnic pomocí homografií. Proces spojování stop probíhá pro každou kameru snímající detail zvlášť. Na detailní kameře se určí nejnižší detekovaný bod u kterého se předpokládá že je na zemi. Souřadnice to-

hoto bodu se pomocí homografie přenesou na odpovídající souřadnice v přehledové kameře. V blízkém okolí přeneseného bodu se hledá nejbližší postava z přehledové kamery. Tento postup se opakuje pro každý snímek. Přičemž se při zpracování každého snímku zaznamená identifikátor nejbližší stopy v přehledové kameře a po zpracování celé stopy se spárují ty stopy s celkovou nejnižší vzdáleností. Vizuální podoba procesu spojování je naznačena na obrázku 3.9.

Finální tvorba výřezů a jejich hierarchické uložení. Poslední fází celého procesu je samotné vytvoření výřezů postav. Z obrazu se postavy vyřezávájí vzhledem k detekované póze a hierarchicky se ukládají na základě zjištěného id osoby a označení kamery. Zároveň je ke každému snímku uložen soubor s normalizovanými souřadnicemi detekovaných kloubů, ve formátu python pickle [1]. Souřadnice detekovaných kloubů normalizovány tak aby souřadnice krku byla vždy v bodu $[0, 0]$ a vzdálenost mezi krkem a pánví byla 1.

Souborová hierarchie datasetu je tedy:
`/dataset/[person-id]/[cam-id]/[track-id]/[frame-id].[jpg|pose]`

Analýza získaných dat a srovnání vůči stávajícím datasetům

Po zpracování nahrávek ze všech natáčecích lokací se podařilo získat celkem 659 jedinečných identit. U všech identit jsou záběry minimálně ze dvou kamer, ale velká část identit má zaběry až ze všech čtyř kamer. Rozložení kvantity záběrů z různých úhlů na identitu je zobrazeno v grafu 3.10

Při zpracování získaných stop nebyly použity stopy kratší než 3 vteřiny. Tato hodnota byla stanovena experimentální cestou v závislosti na pozorování délky průchodů osob přes snímanou plochu. Histogram délky stop všech nátáčecích lokací je zobrazen v grafu 3.11

Natáčelo se celkem na třech různých lokacích. Přehled množství získaných dat z jednotlivých lokací lze najít v tabulce 3.1.

Lokace	Počet identit	Průměrná délka stopy (snímky)	Doba natáčení
FIT	255	12,3s (308 snímků)	20m 17s
Nové sady	326	17,4s (436 snímků)	58m 16s
Husovice	78	6.8s (172 snímků)	13m 14s

Tabulka 3.1: Přehled základních parametrů získaných dat na jednotlivých lokacích

Získaná data lze srovnat na základě kvantitativních parametrů uvedených v tabulce 3.2. Podle počtu identit je vytvořený dataset spíše podprůměrný oproti velkým datasetům jako jsou MARS[26], nebo PRID11[7]. Naopak v počtu kamer (snímaných úhlů) se vytvořený dataset ostatním datasetům plně vyrovná.

Jedna z hlavních motivací pro tvorbu vlastního datasetu bylo získat data na nichž budou rozeznatelné všechny tři zkoumané modality. Snímek postavy je dostupný pro každou identitu v datasetu a stejně tak je pro každou identitu a snímek dostupná detekovaná póza postavy, kterou bylo v ostatních datasetech problém plně detekovat. Problém detekce pózy byl způsoben buď nízkou kvalitou snímků u datasetů iLIDS-VID[15, 24, 20, 19],PRID11[7]

[1] Dokumentace k formátu python pickle je dostupný na adrese https://docs.python.org/3/library/pickle.html

a DukeMTMC-VideoReID[16], nebo vysokým množstvím různých překážek v obraze v případě datasetu MARS[26], čímž vytvořený dataset netrpí.

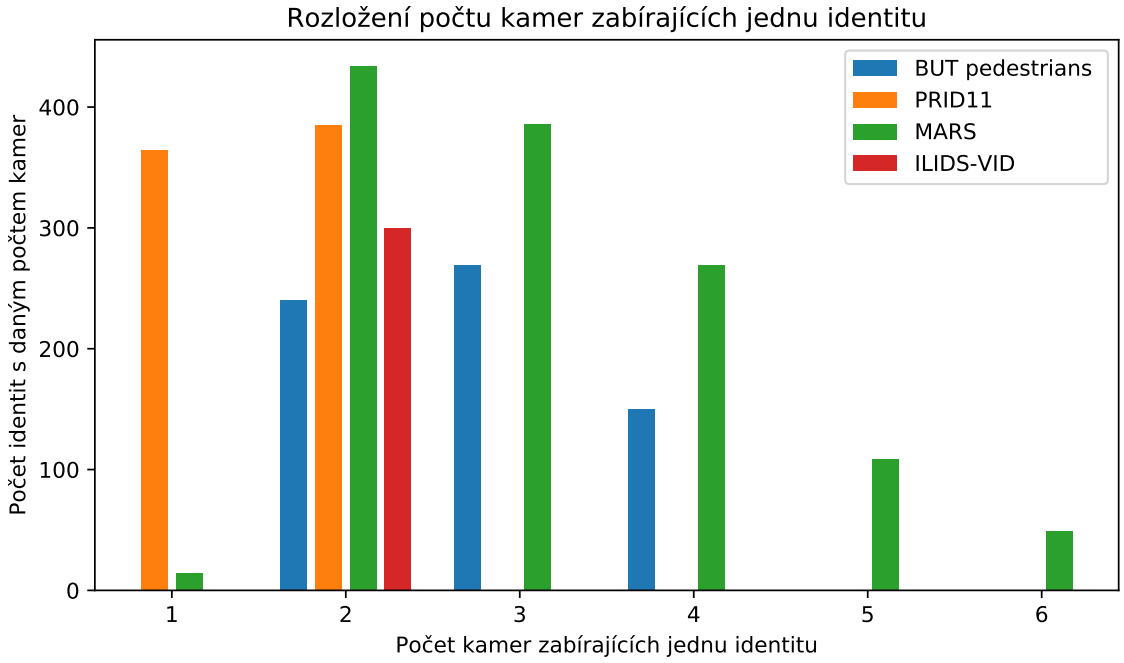
Méně naplněným cílem je snímek obličeje. Při použití detektoru z knihovny dlib[13], byl obličej detekován pouze u 354 identit z celkových 659 (54 % identit). Ve srovnání s datasetem MARS[26], kde byl obličej detekován u 703 identit z celkových 1259 (56 % identit).

Dataset	Počet identit	Počet kamer
BUT pedestrians	659	2 až 4
MARS[26]	1261	1 až 6
iLIDS-VID[15, 24, 20, 19]	300	2
PRID11[7]	1404	1 až 2
DukeMTMC-VideoReID[16]	78	2 až 6

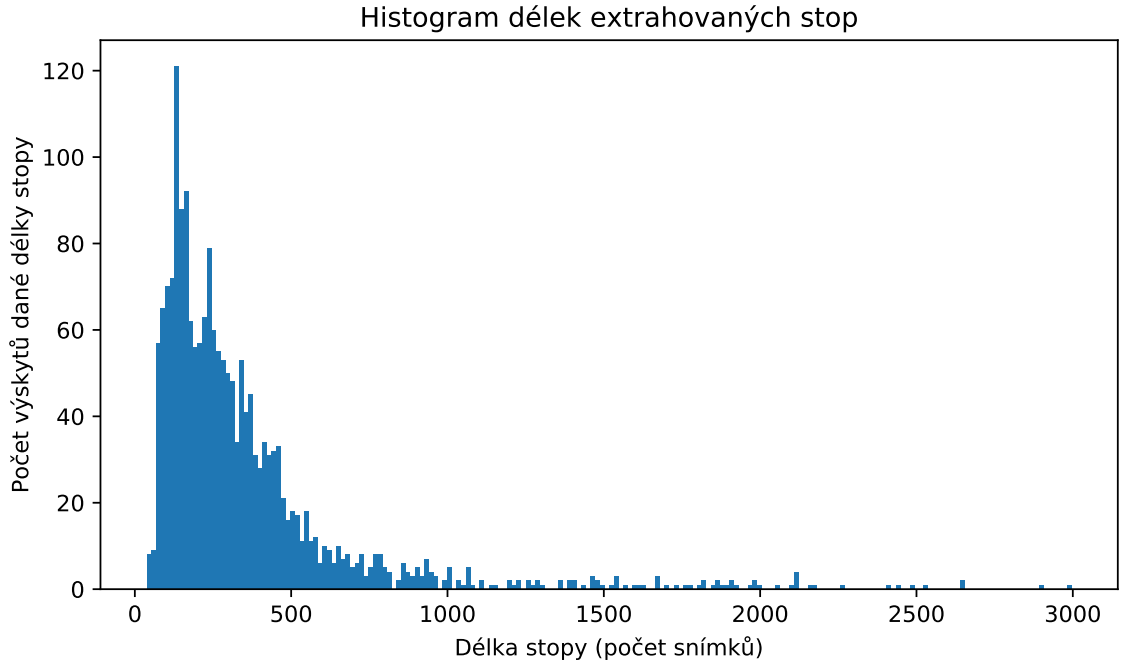
Tabulka 3.2: Porovnání parametrů vytvořeneného datasetu vůči existujícím datasetům.

Obrázek 3.12: Ukázka dat získaných dat.

Obrázek 3.9: Vizualizace párování identit mezi kamerami. Na dolní detalní kameře jsou lidé s přiděleným *id* a detekovaným statickým bodem u země. Tento bod je homografií přenesen na horní přehledovou kameru a znázorněn stejně barevným kroužkem. V okolí přeneseného bodu se vyhledává nejbližší člověk. Párování probíhá na základě takto spárovaných bodů.



Obrázek 3.10: Rozložení počtů kamer (snímaných úhlů) na osobu.



Obrázek 3.11: Histogram délek stop.

Kapitola 4

Návrh řešení identifikačních sítí

Rozhodl jsem se realizovat rozpoznání člověka na základě fúze tří modalit. Konkrétně na základě postavy, obličeje a chůze. K rozpoznání dle postavy a obličeje jsem využil předtrénovaných sítí [2, 27]. Zatímco sítě pro rozpozání dle chůze a síť zajišťující fúzi jsem natrénoval.

Postava. Do rozpoznání podle postavy tedy spadá i aktuální oblečení, včetně batohu a jiných věcí. Z tohoto důvodu nemusí být metoda rozpoznání podle postavy spolehlivá při použití záběrů z jiných časových období, kdy je zkoumaná osoba jinak oblečená. Naopak při rozponávání v rámci krátkého časového úseku se jedná o metodu poměrně spolehlivou.

Obličej. Jedná o často používanou a poměrně spolehlivou metodu v podmínkách dobrého záběru obličeje, v podmíkách rozpoznávání chodců ze vzdálených kamer už je využitelná méně, nikoliv však vůbec, protože na rozdíl od rozpoznání podle záběru celé postavy, je tato metoda téměř nezávislá na aktuálním oděvu člověka, a proto může pomoci při rozponání záběrů z jiných časových období.

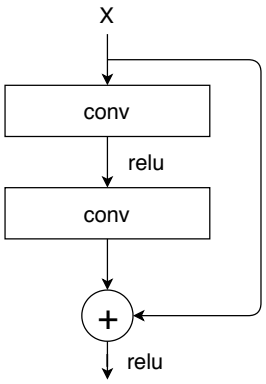
Chůze. Poslední zvolenou modalitou je rozpoznání člověka na základě stylu chůze. Narozdíl od předešlých dvou metod, které se zaměřují na vizuální podobu člověka, se tato metoda zaměřuje výhradně na specifika chůze daného jedince. Přestože existuje množství projektů[3, 5, 17] zabývajících se danou problematikou , nasazení v praxi je velice problematické, protože většina řešení funguje pouze za specifických podminek.

4.1 Rozpoznání postavy

Pro řešení této modality jsem se rozhodl využít předtrénované sítě z projektu *deep-person-reid*[27], který se zabývá právě problematikou reidentifikace osob. Autor projektu poskytuje hned několik předtrénovaných sítí různých architektur. Projekt je implementován v jazyce Python[1] s využitím frameworku Pytorch[2] zaměřeného na strojové učení.

Důvodů k využití právě projektu deep-person-reid, je hned několik. Hlavním důvodem je autory deklarovaná úspěšnost jejich natrénovaných modelů, která odpovídá *state of the art*

[1] Dokumentace aktuální verze jazyka Python je dostupná na stránkách https://docs.python.org/3/
[2] Dokumentace frameworku Pytorch je dostupná na adrese https://pytorch.org/docs/stable/index.html



Obrázek 4.1: Struktura jednoho reziduálního bloku. Převzato z [11]

úspěšnosti dnešní doby. Dalším podružnějším důvodem je právě použitý framework Pytorch, který je použit i v dalších oblastech této práce.

Z modelů, které tento projekt poskytuje, jsem vybral model implementující reziduální síť[11], která podává velmi dobré výsledky právě při rozpoznávání obrazových dat.

Jak je popsáno v článku *Deep Residual Learning for Image Recognition*[11]. Reziduální síť je neuronová síť využívající mechanizmu zkratek, kdy aktivace určitých vrstev v síti se propagují o několik (většinou o dvě) vrstev dopředu právě zmíněnými zkratkami, kde se sečtou s aktivacemi dané vrstvy. Struktura jednoho reziduálního bloku je vyobrazena na obrázku 4.1.

Podle autorů tohoto článku[11] se hluboké neuronové sítě potýkají s problémem tzv.: degradace síťe. Problém degradace sítě je popisován jako stav, kdy přidání další vrstvy do struktury neuronové sítě zhorší úspěšnost neuronové sítě na trénovacích datech. Což je přesný opak očekávaného chování, vzhledem ke skutečnosti, že přidáním vrstvy se síti zvýšil počet trénovatelných parametrů. Přičemž právě tento problém je řešen použitím reziduálních bloků v architektuře sítě.

Z toho důvodu mají reziduální sítě takovou kapacitu a dosahují v dnešní době dobrých výsledků. Konkrétní vybraná síť má 23.5 milionu parametrů a při vyhodnocení EER[3] přesnosti na vytvořeném datasetu *BUT pedestrians* dosahuje 87.2 % korektních rozpoznání. Dimenze výstupního deskriptoru je 2044.

4.2 Rozpoznání obličeje

Druhou vybranou modalitou je záběr obličeje. Stejně jako u rozpoznání postavy, jsem se i tu rozhodl využít předtrénovanou síť. Konkrétně jsem vybral síť z knihovny dlib[13]. Knihovna poskytuje detektor tváře i síť pro extrakci deskriptorů.

Detektor autoři natrénovali na 7198 snímcích. Lokalizuje 5 stěžejních bodů na tváři, konkrétně boční strany očí a spodní část nosu. Díky lokalizaci těchto klíčových bodů je možné

[3] Jedná se techniku vyhodnocení při níž je prahová hodnota nastavena tak aby byl stejný počet *false positive* a *false negative* vyhodnocení.

Obrázek 4.2: Průběh extrakce deskriptoru tváře pomocí knihovny dlib[13]

provést zarovnání tváře z důvodu snadnější extrakce deskriptoru. Vizualizace funkce detektoru a zarovnání je vyobrazena na obrázku 4.2

Rozpoznávací síť[2] je reziduální síť s 29 konvolučními vrstvami. V podstatě se jedná, podobně jako při rozpoznání postavy, o verzi sítě[11], která je využita i při rozpoznání postavy. Dimenze výstupního deskriptoru je 128.

Autoři tuto síť natrénovali na datasetu, který obsahoval téměř 3 milliony snímků tváře s celkem 7485 identitami a na datasetu LFW[9] a dosáhli přesnosti 99.3%.

4.3 Rozpoznání stylu chůze

Poslední vybranou modalitou je styl chůze. Jedná se o v praxi nepříliš využívanou modalitu. Malé využití této metody je způsobeno především již na první pohled nízkou mírou informace, která je pouze ve stylu chůze obsažena. Ale jsou tu i jiné problémy spojené z danou problematikou a to především potřeba delší, několika vteřinové, nepřerušené video stopy člověka, kterého identifikujeme. Dalším problémem je výrazná změna stylu chůze v situacích, kdy pozorovaná osoba má, či naopak nemá nasazený batoh, bundu nebo jiná břemena.

Přes zmíněné problémy má tato modalita určitě potenciál. Tento potenciál vychází z faktu, že i člověk je schopen pouze na základě vzdálené siluety člověka poznat právě podle specifické chůze. Z tohoto faktu vychází několik projektů a různých metod realizujících rozpoznání člověka podle chůze. Vybrané metody a projekty, kterými jsem se při řešení práce inspiroval popíši.

Fisher Motion Descriptor for Multiview Gait Recognition[5]. Projekt od autorů z univerzity ve španělské Córdobě popisuje jeden z možných přístupů rozpoznávání člověka dle chůze. Autoři se konkrétně zaměřují na extrakci příznaků ve formě trajektorií pohybujících se bodů mezi jednotlivými snímky. Extrahované příznaky z několika snímků dále transformují do vektoru, který autoři nazývají *pyramidal fisher vector*. Pomocí takto získaných vektorů je provedeno samotné rozpoznání. Experimenty prováděly na několika datasetech z nichž nějvětší (ale přesto poměrně malý) dataset je *CASIA Gait Dataset B*[17]. Tento dataset obsahuje 124 identit zachycených z 11 úhlů pohledu a je celý vytvořen v laboratorních podmínkách. Podoba dat v datasetu je zobrazena na obrázku 4.3.

Gait Recognition from Motion Capture Data[3]. Rozpoznáním chůze se zabývají i lidé z Masarykovi univerzity a to konkrétně Michal Balážia. Na toto téma sepsal dizertační práci,

Obrázek 4.3: Ukázka snímků z datasetu casia[17]. Záběr stejného člověka bez bundy, s bundou a s batohem.

v níž se zabývá identifikací podle záběrů chůze technologií motion capture. Díky využití motion capture technologie má k dispozici třídimenzionální reprezentaci pózy v podobě pozic specifických částí lidského těla jakou jsou například ruce, ramena nebo nohy. Na základě sekvencí takto detekovaných postav je realizováno rozpoznání.

Rozpoznání ze záběru jedné kamery. V rámci své práce jsem se rozhodl realizovat rozpoznání osoby na základě sekvence detekovaných póz z jedné kamery. Využiji k tomu sestavený dataset popsaný v kapitole 3 a detektor póz z projektu *Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*[4].

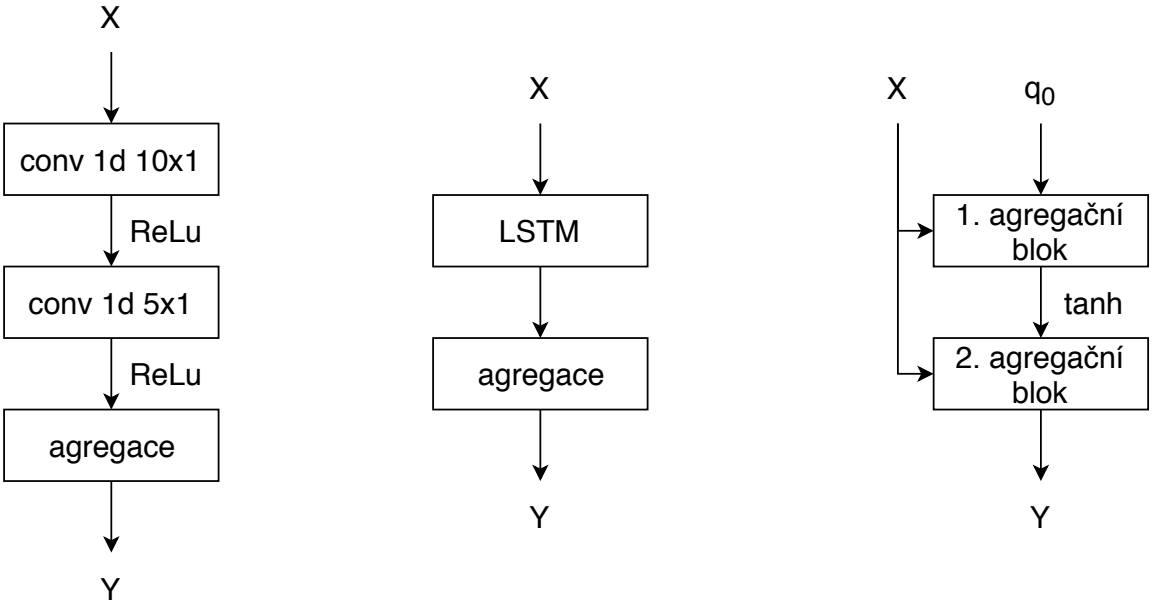
Autoři tohoto projektu vytvořili systém schopný detekovat lidské klouby v obraze. Pomocí techniky, kterou nazvali *part affinity fields* (PAF)(pole afinních částí), tyto detekované klouby spojují tak aby byly vždy od jedné osoby a ve finále vytvořily celistvou pózu. Diky tomu je detekce kloubů a následné odhadování póz, nezávislé na počtu osob v obraze. Kónkrétní informace o této technice lze nalézt v jejich článku [4].

Už při tvorbě datasetu jsem využil tento projekt právě k detekci osob, jejich póz a určení správného výřezu. Díky tomu jsem měl v datasetu k dispozici detekovanou pózu pro každou jednotlivou osobu na každém snímku. Ukázka části sekvence extrahovaných póz je znázorněna na obrázku 4.5.

K realizaci této úlohy jsem se rozhodl natrénovat neuronovou síť na extrakci příznaků, která jako vstupy přijímá seřazené sekvence detekovaných póz a nezávisle na délce sekvence generuje deskriptor osoby. Následné rozponání osoby probíhá na základě výpočtu vzdálenosti mezi vyextrahovanými deskriptory stejně jako je popsáno v sekci 2.1.

Výběr architektury. Všechny architektury jsou vybrány tak, aby byly nezávislé na délce vstupní sekvence a zároveň byl jejich výstupem jediný vektor reprezentující celou sekvenci.Jedním z faktorů ovlivňujících volbu architektury sítě je velikost datasetu. Je-li k dispozici velké množství dat, je možné a vhodné použít větší síť, která by díky tomu měla mít lepší výsledky na trénovacích i na validačních datech. V případě malého množství dat je využití velkých sítí naopak nevhodné z důvodu rychlého přetrénovaní (tzv.: *overfittingu*) sítě na konkrétní trénovací data. Což způsobí degradaci úspěšnosti sítě na validačních datech.

V rámci experimentování jsem vyzkoušel množství různých architektur a ukázalo se, že nejlepší výsledky podávají velmi minimalistické sítě narozdíl od větších sítí, které trpěly rychlý přetrénováním. Proto se při experimentování zabývám pouze těmito minimalistickými sítěmi.

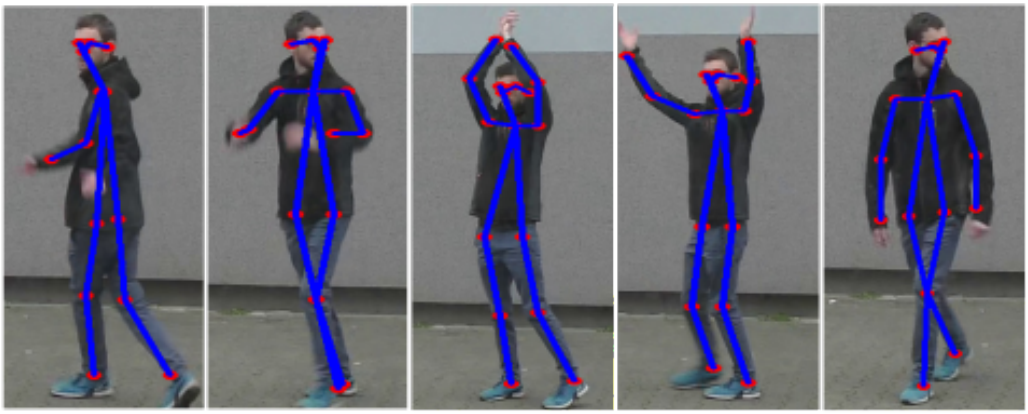


Obrázek 4.4: Sítě použité pro rozpoznání chůze. Vlevo konvoluční architektura. Uprostřed rekurentní a vpravo agregační síť inspirovaná z projektu [25], která je v určitých konfiguracích využita k agregaci výstupů konvoluční a rekurentní sítě.

Implemetoval jsem dva typy sítí. První kovoluční síť složenou pouze ze dvou konvolučních vrstev a druhou rekurentní LSTM síť. Výstupem obou těchto sítí je sekvence vektorů, u nichž je potřeba provést agregaci. Tuto agregaci jsem implementoval dvojím způsobem a to buď pouhým výpočtem průměrného vektoru, nebo pomocí neuronové agregační sítě[25], která je podrobněji popsána v sekci 2.2. Implementované architektury jsou znázorněné na obrázku 4.4.

Síťe jsem trénoval metodou tripletových siamských sítí, popsané v sekci 2.1. Verifikaci modelu jsem provedl metodou ROC křivky[4], u níž posuzuji kvalitu modelu na základě plochy pod křivkou (AUC - Area Under Curve) a pomocí EER (equal error rate) přesnosti.

[4] Jedná se o křivku znázorňující poměr správně detekovaných (true positive) a nesprávně detekovaných (false positive) entit s různým nastavením prahové hodnoty.



Obrázek 4.5: Ukázka extrahovaných póz s jejichž sekvencemi probíhá rozpoznání.

Model jsem, stejně jako ostatní součásti práce, implementoval v jazyce Python[5] s využitím frameworku Pytorch[6], což je v dnešní době společně s Tensorflow[7] nejvíce využívaný framework zaměřený na strojové učení pomocí neuronových sítí.

Příprava dat. Trénovací a validační data jsou použita z vytvořeného datasetu. Jako validační data bylo náhodně zvoleno 100 identit (z celkových 659), na nichž neprobíhalo trénování, ale pouze samotná validace. Vzhledem k faktu, že se zpracovávají celé sekvence a k trénování byly vyžity sekvence fixní délky. Experimentoval jsem i s různým nastavením délky sekvence.

Na základě délky zvolené sekvence, byly data rozděleny a následně zpárovány do formy tripletů, které se přímo využívaly při trénování. Pro účely validace modelu byla vytvořena sada pozitivních a negativních dvojic v poměru 1 : 1.

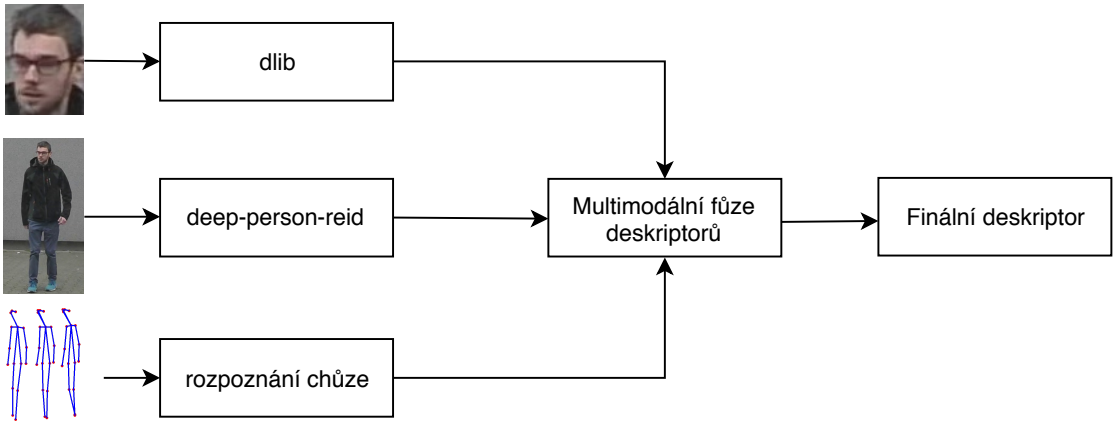
4.4 Multimodální fúze deskriptorů

Finální části práce je využití všech dostupných informací o zkoumané osobě k vytvoření co nejpřesnější reprezentace dané osoby. K tomu využiji metod multimodální fúze, které jsou podrobně popsány v sekci 2.3. Vzhledem k povaze mé práce, především z důvodu využívání již předtrénovaných sítí k rozpoznání obličeje a postavy jsem se rozhodl pro metodu *late fusion* realizovanou pomocí neuronové sítě.

Vstupem této sítě jsou agregované deskriptory jednotlivých zmíněných modalit. Přičemž agregace deskriptorů obličeje a postavy je provedena pomocí aritmetického průměru všech vyextrahovaných vektorů nad danou video stopou. Zatímco agregaci chůze není potřeba provádět z důvodu, že celá sekvence je zpracována jako jeden celek.

Trénování. Trénování sítě je, stejně jako u trénovaní stylu chůze, provedeno s využitím tripletových siamských sítí. V případě fúze zvolených modalit je tedy jeden triplet tvořen devíti vektory, protože dochází k fúzi tří modalit a tudíž každý prvek tripletu by měl

[5]Dokumentace aktuální verze jazyka Python je dostupná na stránkách https://docs.python.org/3/
[6]Dokumentace frameworku Pytorch je dostupná na adrese https://pytorch.org/docs/stable/index.html
[7]https://www.tensorflow.org/



Obrázek 4.6: Znázornění mechanizmu multimodální fúze metodou *late fusion*.

ideálně obsahovat všechny modality. Ideálně zmiňuji proto, že ne vždy jsou ke každé video stopě dostupné všechny modality. Často nastává situace, kdy byl člověk natáčen z takového úhlu, že není možné detekovat obličej a je tedy pro danou stopu nedefinovaný. V reálné implementaci je tato situace vyřešena použitím nulového vektoru.

Výstupem sítě je deskriptor popisující zkoumanou osobu na základě všech modalit. Stejně jako deskriptor extrahovaní sítí pro extrakci příznaků i tento deskriptor splňuje vlastnosti díky kterým lze provádět identifikaci. Především se jedná o vlastnost, že deskriptory popisující jednu konkrétní osobu by měli být ve výsledném vektorovém prostoru umístěny blízko sebe. Díky této vlastnosti lze vyhodnocovat identity lidí na základě vzdáleností jejich deskriptorů.

Mechanizmus fúze zmíněných tří modalit je vyobrazen na obrázku 4.6.

Příprava dat. Trénování sítě probíhalo na datasetu, který byl vytvořen v rámci řešení této práce. Podrobný postup tvorby datasetu je popsán v kapitole 3. Dataset byl rozdělen na trénovací a validační sadu stejně jako v případě trénovaní rozpoznání chůze.

Přestože základní princip trénovacích a validačních dat použitých pro trénovaní sítě na rozpoznávání dle stylu chůze a sítě realizující multiomdální fúzi je stejný. Samotná struktura dat se liší. A to především skutečností, že chůze je sekvence dat, zatímco fúze pracuje s předem stanoveným počtem vektorů. Trénovací data byla vytvořena takovým způsobem, že pro každou identitu se proved kartézský součin dostupných stop, jejichž deskriptory byly použity jako *anchor* a *positive* a ke každé vytvořené dvojici se náhodně vybrala stopa jiné identity, která ve výsledném tripletu plnila roly vektoru *negative*.

Kapitola 5

Výsledky experimentů

Pro všechny experimenty popsané v této kapitole jsem využil vytvořený dataset popsaný v kapitole 3. Architektury a mechanizmy používané při trénování a vyhodnocení jsou popsány v předešlých kapitolách.

5.1 Rozpoznání postavy a obličeje

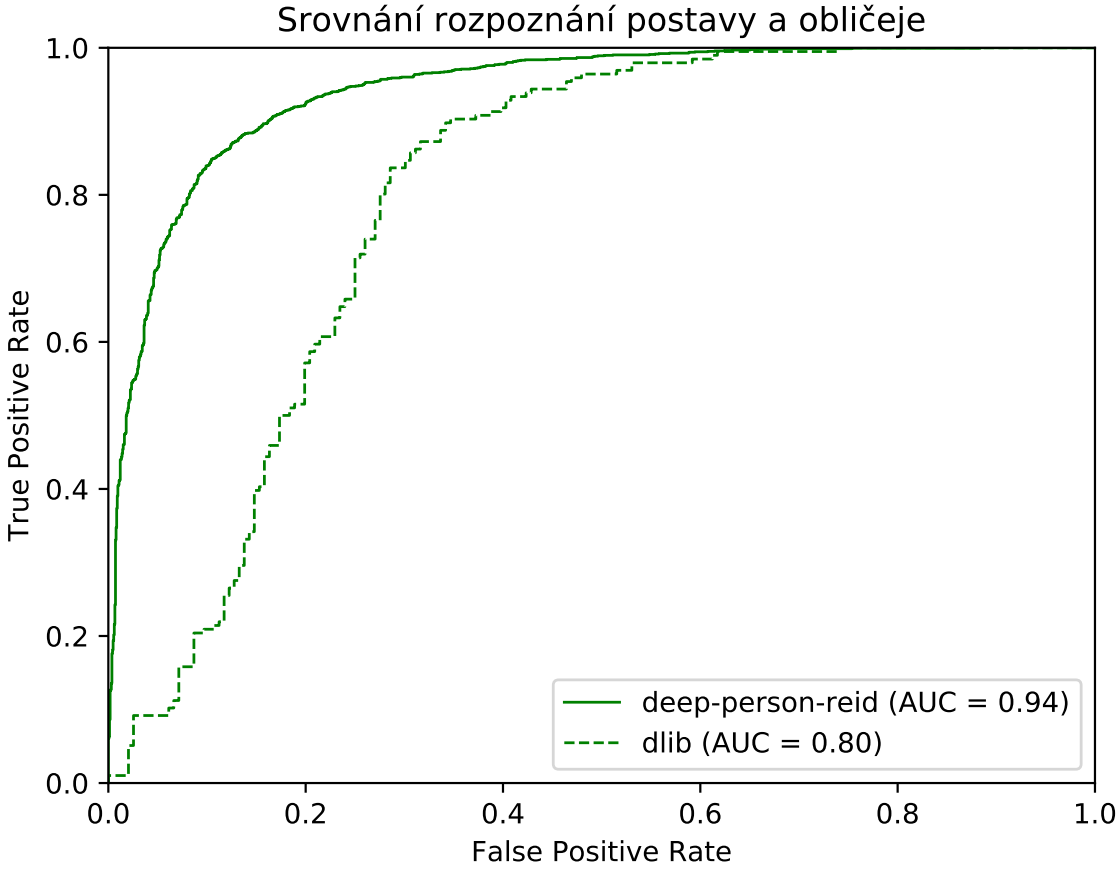
Jak již bylo zmíněno, pro rozpoznání podle tváře a postavy jsem se rozhodl využít předtrénovaných sítí. Zároveň předmětem práce byla tvorba vlastního datasetu a proto jsem se rozhodl provést experiment kterým otestuji úspěšnost převzatých sítí. Tento experiment jsem provedl nad celým datasetem, nikoliv pouze nad testovací sadou.

Rozpoznání tváře. K rozpoznání dle tváře jsem využil síť z knihovny dlib[13]. První fází rozpoznání je samotná detekce tváře, zarovnání a následné extrakce příznaků z každého snímku v datasetu. Po extrakci příznaků je nutné provést agregaci extrahovaných vektorů, aby každé video stopě odpovídal právě jeden deskriptor popisující danou osobu. Agregaci jsem provedl metodou zprůměrování všech deskriptorů extrahovaných z dané video stopy.

Z extrahovaných a agregovaných vektorů jsem sestavil množinu pozitivních a negativních párů, jejichž podobnost byla posuzována podle eukleidovské vzdálenosti. Počet negativních a pozitivních párů byl poměru $1 : 1$. Celková kvalita modelu byla posuzována na základě plochy pod ROC křivkou. Výsledná ROC křivka je vykreslena na obrázku. 5.1

Rozpoznání postavy. Stejně jako pro rozpoznání tváře, jsem i pro rozpoznání postavy využil předtrénované sítě. Narozdíl od zpracování tváře při rozpoznání postavy byly použity přímo výřezy osob ve tvaru, v jakém jsou v datové sadě bez žádného zpracování. Stejně jako u rozpoznání tváře byla agregace deskriptorů provedena metodou průměru a vyhodnocení proběhlo nad pozitivními a negativními páry, které byly v poměru $1 : 1$ pomocí ROC křivky. Výsledná ROC křivka je zobrazena na obrázku 5.1.

Výsledek. Provedený experiment ukázal, že síť rozpoznávající postavu funguje na daných datech dobře což ukazuje, že příprava datasetu, tvorba výřezů a párování identit bylo provedeno správně. Zároveň je podle výsledků zřejmé, že rozpoznání podle obličeje na daných datech zdaleka nefunguje tak dobře. Po prozkoumání dat se ukázalo, že špatně rozpoznané



Obrázek 5.1: Výsledné ROC sítě pro rozpoznání tváře dlib[13] a postavy.

identity byly často v nízkém rozlišení daleko v záběru, nebo téměr z profilu, což pravděpodobně způsobilo tuto nízkou úspěšnost.

5.2 Rozpoznání stylu chůze

Při experimentovaní jsem se zabýval srovnáním různých architektur založených buď na konvolučních, nebo rekurentních vrstvách, které jsou podrobněji popsány v sekci 4.3. Experimenty jsou převážně zaměřeny na vliv délky rozpoznávané stopy na přesnost rozpoznání a srovnání různých implementovaných architektur. Přičemž síte jsem vyhodnocoval pomocí ROC křivky, kde sledovaným parametrem byla především plocha pod křivkou.

Délka sekvence. Důležitým parametrem při identifikaci podle sekvencí je samotná délka sekvence. Respektive jestli se úspešnost rozpoznání s délkou sekvence mění. V ideálním případě by delší sekvence měla znamenat přesnější rozpoznání. Experimenoval jsem se sekvencemi o délce 2 až 6 sekund. Celkem jsem provedl experimenty s pěti architekturami sítí a na základě výsledků uvedených v tabulce 5.1 a obrázku 5.2 je zřejmé, že délka sekvence úspěšnost ovlivňuje, ale pouze v řádu jednotek procent. Dobře zřetelný je tento trend při použití architektury, která kombinuje konvoluční a rekurentní vrsty (CNN + RNN + MEAN).

AUC	Délka sekvence (počet snímků)				
Síť	**50**	**75**	**100**	**125**	**150**
CNN + MEAN	0.775	0.783	0.791	0.801	0.787
CNN + NAN	0.731	0.772	0.776	0.776	0.735
RNN + MEAN	0.778	0.786	0.79	0.802	0.779
RNN + NAN	0.776	0.787	0.807	0.802	0.804
CNN + RNN + MEAN	0.783	0.78	0.799	0.811	0.812

Tabulka 5.1: Tabulka výsledků natrénovaných sítí rozdělených podle délky sekvence.

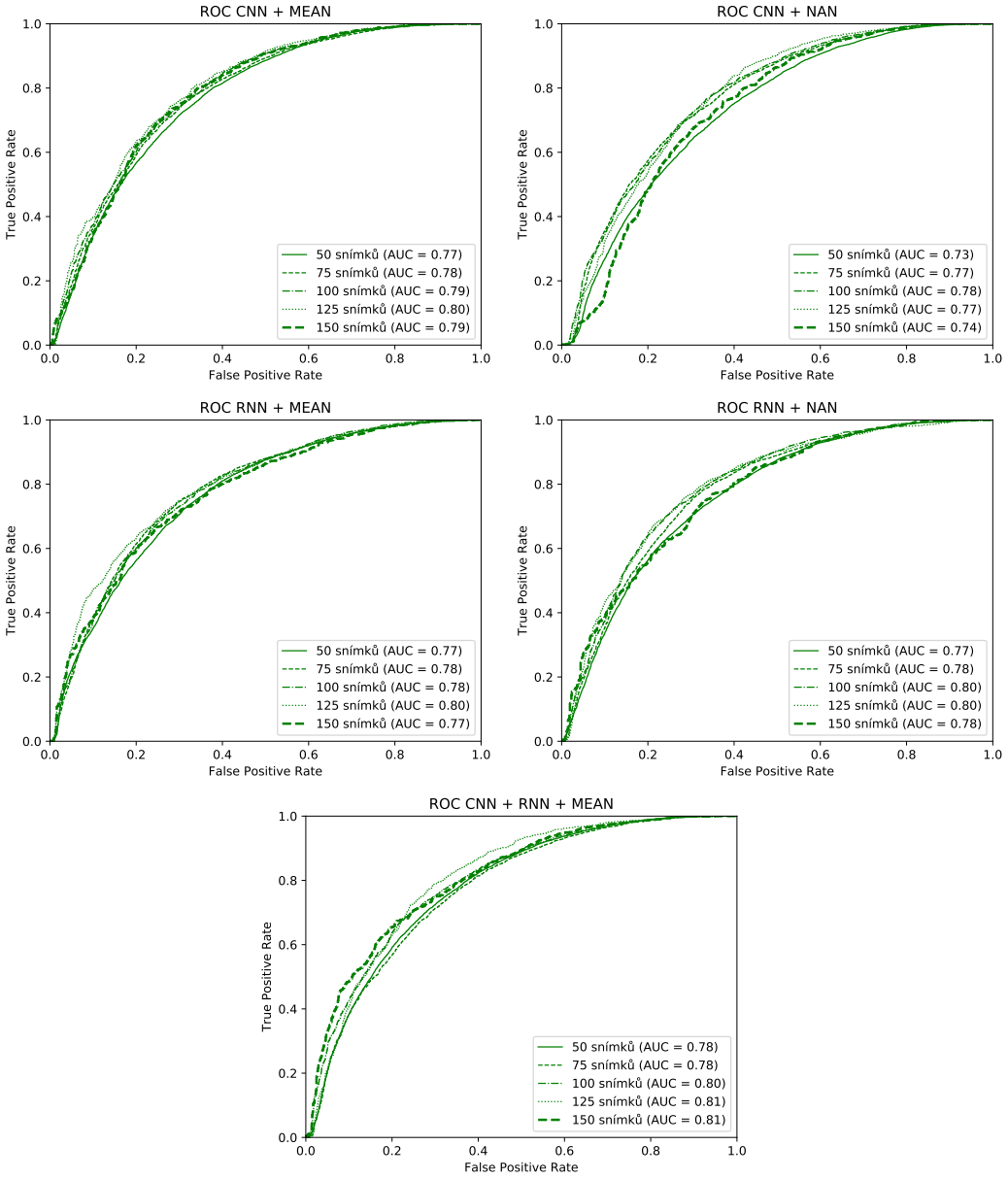
Srovnání architektur. Experimentoval jsem s 5 různými architekturami sítí. Od nejminimalističtejší konvoluční a rekurentní sítě s agregací pomocí průměru, přes agregaci pomocí *Neural agregation network*[25]. Poslední architektura je složena ze dvou konvolučních vrstev a jedné rekurentní vrstvy agregované průměrem.

Výsledky sítí jsou uvedeny v tabulce 5.1. Na základě vásledků jsou vidět rozdíli mezi architekturami. Nejlepší architektura je na základě výsledků kombinace konvoluční a rekurentní sítě. Přestože výsledky této poměrně robustnější sítě jsou lepší, stále se liší pouze o jednotky procent. Srovnání ROC křivek jednotlivých sítí jsou na obrázku 5.3.

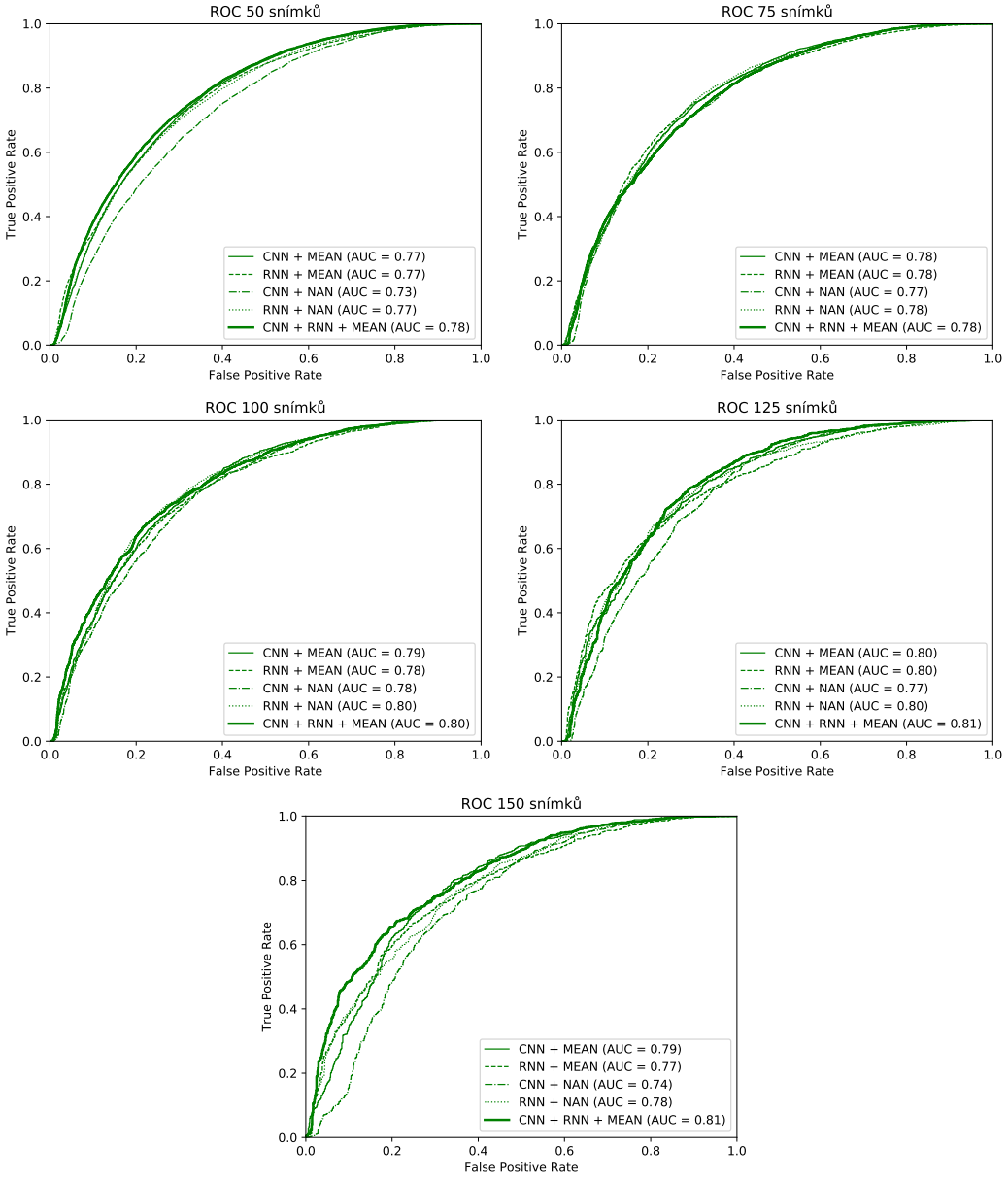
Ověření správné funkce. Velkou otázkou při trénování a testování bylo také to jestli se síť opravdu rozhoduje na základě podoby a vztahu vstupních póz, anebo podle nějakého jiného vzorce, který data obsahují, ale s chůzí vůbec nesouvisí.

Tento problém jsem se rozhodl otestovat natrénováním sítě na datech, které obsahovali statické pózy. Tedy jednu konstantní pózu zreplikovanou do celé sekvence.

Výsledky tohoto experimentu, dle očekávání selhaly. Síť se nebyla schopna naučit rozpoznávat pouze na základě statické pózy a právě tento výsledek slouží k částečnému potvrzení toho, že síť opravdu zpracovává data správně.



Obrázek 5.2: Výsledné ROC křivky v grafech uspořádaných podle typu sítě



Obrázek 5.3: Výsledné ROC křivky v grafech uspořádaných podle délky sekvence

	Postava + obličej + chůze		Postava + chůze	
Síť	**AUC**	**EER ACC**	**AUC**	**EER ACC**
Simple	0.975	91.8 %	0.977	92 %
Sum fusion	0.978	93.3 %	0.983	93.6 %
Concat fusion	0.979	93.3 %	0.979	93.3 %

Tabulka 5.2: Výsledky fúzních sítí. Pojmenovaní sítí vysvětleno v 2.3, 2.4 a 2.5

Modalita	**AUC**	**EER ACC**
Postava	0.941	87.2 %
Tvář	0.8	74 %
Chůze	0.812	72.4 %

Tabulka 5.3: Výsledky jednotlivých modalit k porovnání s výsledky fúze.

5.3 Multimodální fúze

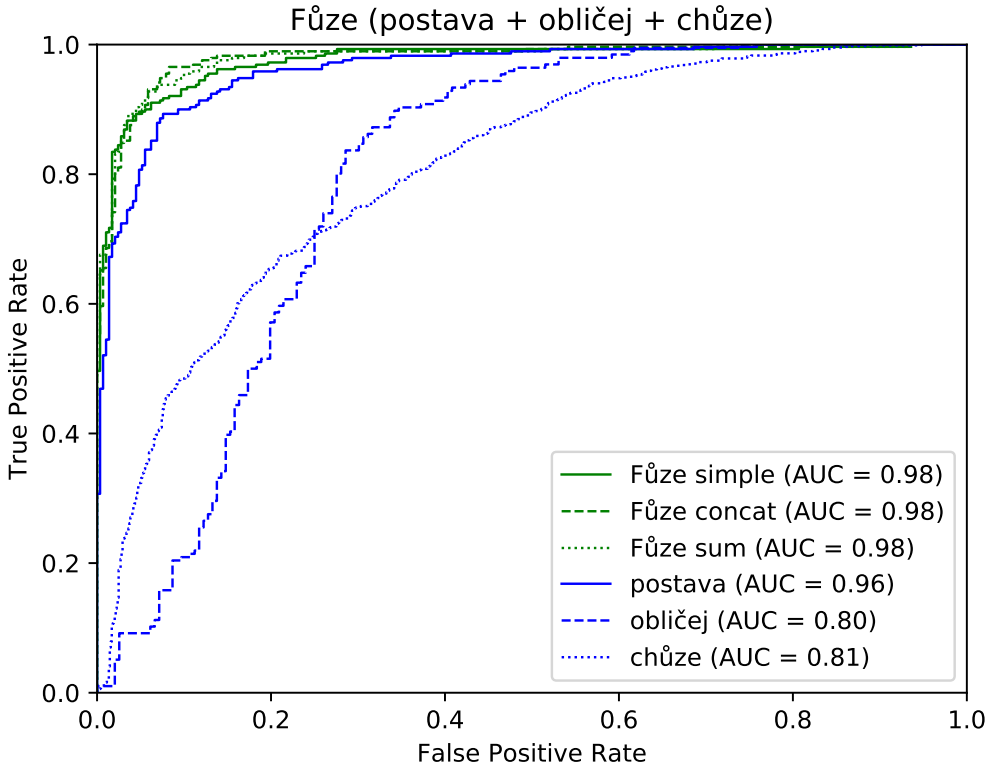
Poslední částí práce byla realizace multimodální fúze zmíněných modalit. Při experimentování jsem se zaměřil na techniku *late fusion*, kterou jsem podrobně popsal v sekci 4.4. Provedl jsem experimenty na třech odlišných architekturách, které zde byly popsány a jejich architektury zobrazeny na obrázcích 2.3, 2.4 a 2.5 v kapitole 2. Stejně jako u trénování předešlých sítí je zde použita metoda siamských sítí s triplet loss chybovou funkcí.

Jako sítě jednotlivých modalit byly použity dlib[2], deep-person-reid[27] a natrénovaná síť rozpoznávající chůzi (CNN + RNN + MEAN).

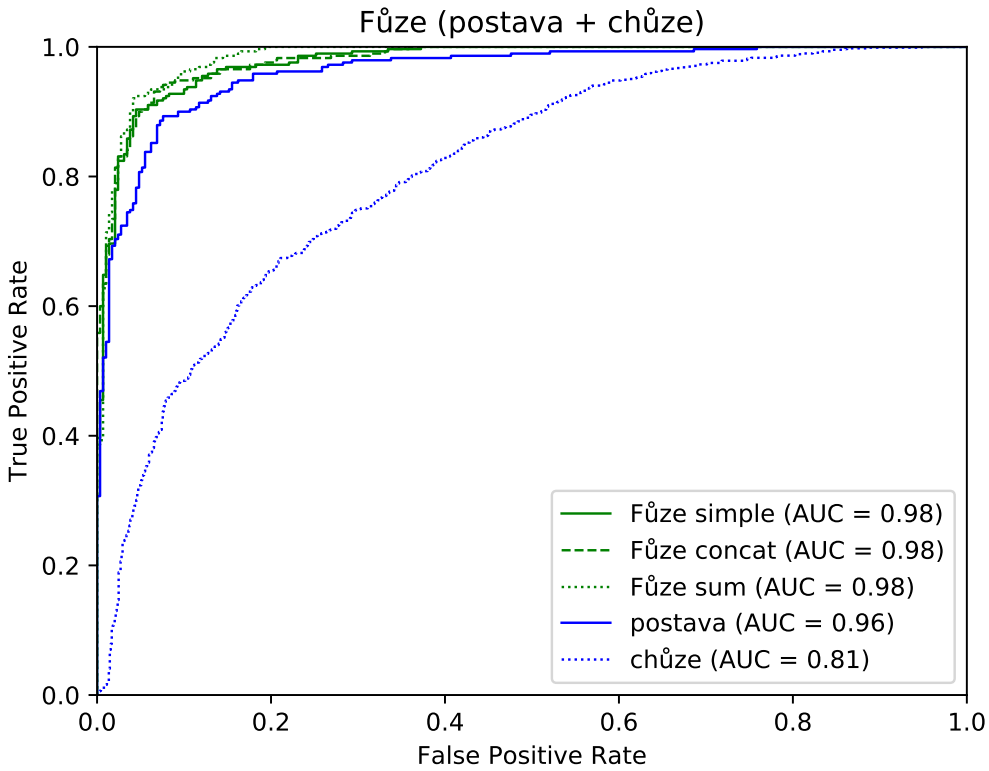
Získané výsledky všech architektur na nichž byly provedeny experimenty, jsou vypsány v tabulce 5.2. Výsledné ROC křivky jsou vykresleny na obrázku 5.4 zároveň s ROC křivkami jednotlivých modalit.

Vliv použití obličeje na výsledek fúze. Z výsledků je vidět pouze minimální vliv použité architektury na úspěšnost fúze. Proto jsem se zaměřil na vliv obličejové modality na celkovou úspěšnost. Konkrétně na vliv obličeje jsem se zaměřil z důvodu relativně vysokého množství identit u nichž obličej není definován a zároveň kvůli ne zcela přesvědčivým výsledkům rozpoznání obličeje, které jsou podrobněji rozebrány v sekci 5.1. Kromě experimentů za použití všech dostupných modalit jsem tedy natrénoval fúzi využívající pouze rozpoznanou postavu a chůzi. Výsledky jsou společně uvedeny v tabulce 5.2 a vykresleny v grafu 5.5.

Výsledky ukázaly mizivý vliv obličejové modality, protože výsledky jsou téměř shodné při použití všech tří modalit tak i při použití pouze dvou modalit. Pro srovnání jsou v tabulce 5.3 uvedeny výsledky jednotlivých modalit.



Obrázek 5.4: ROC křivky sítí realizujících fúzi všech tří modalit.



Obrázek 5.5: ROC křivky sítí realizujících fúzi pouze postavy a chůze.

Kapitola 6

Závěr

Cílem této práce bylo realizovat identifikaci osob za využití několika modalit a jejich fúze. Konkrétně byly vybrány tři modality obličej, postava a chůze.

Rozpoznání podle daných tří modalit vytvořilo velmi specifické požadavky na datovou sadu, které nesplňoval žadný z dostupných datasetů. Proto byl v rámci práce vytvořen dataset, který celkem obsahuje 659 jedinečných identit. Natáčení probíhalo na třech lokacích a při každém natáčení byly použity čtyři kamery, díky čemuž jsou všechny identity natočeny až ze čtyř úhlů. Z důvodu co nejrychlejšího zpracování záběrů byly vytvořeny nástroje umožňující téměř automatickou tvorbu datasetu.

K rozpoznání obličeje byla využita předtrénovaná síť z knihovny dlib[13] a rozpoznání postavy zajišťovala předtrénovaná sít z projektu deep-person-reid[27]. Naopak v rámci práce byly natrénovány sítě rozpoznávající chůzi pomocí sekvence 2D detekovaných pozic 18 kloubů na lidském tělě. Dále byla natrénována síť zajišťující multimodální fúzi extrahovaných deksriptorů jednotlivých modalit.

Experimenty s rozpoznáním chůze ukázaly, že z implementovaných architektur má největší potenciál architektura kombinující konvoluční a rekurentní vrstvy, která v experimentech dosáhla EER přesnosti 72.4 %. Zároveň z experimentů vyplynulo, že délka rozpoznávané sekvence má vliv na přesnost rozpoznání, i když na délkách sekvencí s nimiž bylo experimentováno se jedná pouze o jednotky procent.

Experimenty s multimodální fúzí byly provedeny na třech různých architekturách sítí, přičemž výsledky všech tří sítí byly zcela srovnatelné. EER přesnosti jednotlivých modalit jsou 87.2 % pro postavu, 74 % obličej a 72.4 % chůze. Úspěšnost fúze byla 93.3 % což ukazuje několika procentní zlepšení. Zároveň byl proveden experiment s fúzí pouze dvou modalit a to pouze postavy a chůze, jehož výsledek byl téměř stejný jako při fúzi všech tří modalit. Toto bylo způsobeno malým množstvím detekovaných obličejů ve vytvořených datech a tedy nízkou váhou této modality.

Z pohledu budoucího vývoje by bylo ideální několikanásobně rozšířit vytvořený dataset a zaměřit se na identifkaci na základě snímku postavy a chůze za využití mechanizmu *early fusion*, který provádí fúzi na podstatně nižší úrovni než použitý mechanizmus *late fusion*.

Literatura

[1] Basic concepts of the homography explained with code. [online], [vid. 6.4.2019]. URL https://docs.opencv.org/3.4.1/d9/dab/tutorial_homography.html

[2] High Quality Face Recognition with Deep Metric Learning. [online], [vid. 21.4.2019]. URL http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html

[3] Balážia, M.: *Gait Recognition from Motion Capture Data*. Dizertační práce, Masarykova univerzita, Fakulta informatiky, 2017.

[4] Cao, Z.; Simon, T.; Wei, S.-E.; aj.: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] Castro, F. M.; Marín-Jiménez, M. J.; Guil, N.; aj.: Fisher Motion Descriptor for Multiview Gait Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Leden 2016.

[6] Chopra, S.; Hadsell, R.; LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, ročník 1, June 2005, ISSN 1063-6919, s. 539–546 vol. 1, doi:10.1109/CVPR.2005.202.

[7] Hirzer, M.; Beleznai, C.; Roth, P. M.; aj.: Person Re-Identification by Descriptive and Discriminative Classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*, 2011.

[8] Hoffer, E.; Ailon, N.: DEEP METRIC LEARNING USING TRIPLET NETWORK. Technická zpráva, 2014.

[9] Huang, G. B.; Ramesh, M.; Berg, T.; aj.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technická Zpráva 07-49, University of Massachusetts, Amherst, Říjen 2007.

[10] Jane Bromley, Y. L. E. S., Isabelle Guyon; Shah, R.: Signature Verification using a "Siamese"Time Delay Neural Network. 1994.

[11] Kaiming He, S. R., Xiangyu Zhang; Sun, J.: Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016.

[12] Kim, I.: tf-pose-estimation. https://github.com/ildoonet/tf-pose-estimation/blob/master/etcs/reference.md, 2017, gitHub, [vid. 4.4.2019].

[13] King, D. E.: Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.*, ročník 10, 2009: s. 1755–1758, ISSN 1532-4435.

[14] Liu, K.; Li, Y.; Xu, N.; aj.: Learn to Combine Modalities in Multimodal Deep Learning. 2018.

[15] M. Li, X. Z.; Gong, S.: Unsupervised Person Re-Identification by Deep Learning Tracklet Association. 2018.

[16] Ristani, E.; Solera, F.; Zou, R.; aj.: Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[17] S. Zheng, K. H. R. H., J. Zhang; Tan, T.: Robust View Transformation Model for Gait Recognition, International Conference on Image Processing(ICIP), Brussels, Belgium. Leden 2011.

[18] Schroff, F.; Kalenichenko, D.; Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, ISSN 1063-6919, s. 815–823, doi:10.1109/CVPR.2015.7298682.

[19] T. Wang, X. Z., S. Gong; Wang, S.: Person Re-Identification by Video Ranking. In *In Proc. European Conference on Computer Vision, Zurich, Switzerland*, 2014.

[20] T. Wang, X. Z., S. Gong; Wang, S.: Person Re-Identification by Discriminative Selection in Video Ranking. 2016.

[21] Taigman, Y.; Yang, M.; Ranzato, M.; aj.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, ISSN 1063-6919, s. 1701–1708, doi:10.1109/CVPR.2014.220.

[22] Vielzeuf, V.; Lechervy, A.; Pateux, S.; aj.: CentralNet: a Multilayer Approach for Multimodal Fusion. 2018.

[23] Wei, S.-E.; Ramakrishna, V.; Kanade, T.; aj.: Convolutional pose machines. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] X. Ma, S. G. X. X. J. H. K.-M. L., X. Zhu; Zhong, Y.: Person Re-Identification by Unsupervised Video Matching. Květen 2017.

[25] Yang, J.; Ren, P.; Zhang, D.; aj.: Neural Aggregation Network for Video Face Recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2016.

[26] Zheng, L.; Bie, Z.; Sun, Y.; aj.: MARS: A Video Benchmark for Large-Scale Person Re-identification. In *European Conference on Computer Vision*, Springer, 2016.

[27] Zhou, K.: deep-person-reid. https://github.com/KaiyangZhou/deep-person-reid, 2018, gitHub, [vid. 4.4.2019].

Příloha A

Obsah přiloženého paměťového média

dataset	složka se skripty pro tvorbu datasetu
doc	složka se zdrojovými kódy dokumentace
face	složka se skripy vytvořenými pro extrakci příznaků z tváře
fusion	složka s připravenými skripty pro trénování fůze
gait	složka s připravenými skripty pro trénování rozpoznání chůze
misc	sada podpůrných skriptů převážně pro práci s daty
person	složka se skripy vytvořenými pro extrakci příznaků z postavy
nets	složka s natrénovanými sítěmi
xjurca08-video.mp4	video, které bylo součástí zadání práce
LICENSE	GNU GPLv3 licence
README	podrobnější popis souborů
doc.pdf	text bakalářské práce