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

ODHAD OSOBNOSTNÍCH VLASTNOSTÍ Z VIDEA

APPARENT PERSONALITY ANALYSIS FROM VIDEO

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **PATRIK ČIGÁŠ**
AUTHOR

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

BRNO 2017

Abstrakt

Táto bakalárska práca sa zaoberá experimentami so systémami na odhad dojmových osobnostných vlastností z videa a porovnáva ich úspešnosť. Systémy v experimentoch sú vytvorené pomocou lineárnych regresorov a konvolučných neurónových sietí. Experimenty porovnávajú úspešnosť lineárnych regresorov spracovávajúcich obrazovú a zvukovú modalitu. Na vytvorených spektrogramoch zo zvukových modalít videa práca vyhodnocuje výsledky konvolučných sietí s rôznym počtom konvolučných a plne prepojených vrstiev a následne porovnáva úspešnosť riešenia pomocou regresie a pomocou klasifikácie. Pre obrazovú modalitu práca porovnáva množstvo informácií v pohybe pohľadu človeka a v pohybe orientačných bodov tváre. Najlepšie výsledky v experimentoch dosahuje systém na spracovávanie orientačných bodov tváre.

Abstract

This bachelor thesis deals with experiments with systems for apparent personality analysis from video, and compares accuracy of these systems. Systems from the experiments are created by linear regression and convolutional neural networks. Experiments compare accuracy of linear regressors processing visual and audial modality of video. On spectograms made from audial modality of video, thesis evaluates results of convolutional neural networks with varying number of convolutional and fully connected layers nad subsequently compares accuracy of regression solution and classification solution of the problem. For visual modality of video the thesis compares information values of gaze movement and face landmarks movement. System processing face landmarks movement reaches the best results in the experiments.

Kľúčové slová

Analýza dojmových osobnostných vlastností, prvý dojem, spracovanie videa, konvolučné neurónové siete, veľká päťka

Keywords

Apparent personality analysis, first impression, video processing, convolutional neural networks, big five

Citácia

ČIGÁŠ, Patrik. *Odhad osobnostních vlastností z videa.* Brno, 2017. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Michal Hradiš, Ph.D.

Odhad osobnostních vlastností z videa

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pana Ing. Michala Hradiša, Ph.D. Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

.....
Patrik Čigáš
17. mája 2017

Poďakovanie

Týmto by som sa chcel poďakovať pánovi Ing. Michalovi Hradišovi, Ph.D. za všetok venovaný čas a cenné rady, ktoré ma naviedli na riešenie tejto práce. Ďalej by som sa chcel poďakovať virtuálnej organizácií Metacentrum za poskytnuté výpočetné prostriedky, poďakovanie je uvedené nižšie v požadovanom formáte.

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

Obsah

Kapitola 1

Úvod

Prvý dojem je u ľudí zautomatizovaná činnosť, ktorá sa deje už pri prvom pohľade na cudzieho človeka. Pri tomto pohľade si človek spraví vnútornú predstavu o osobnostných vlastnostiach osoby na základe výzoru, postavy, nálady osoby, alebo na základe vyjadrovania a intonácie reči osoby. Prvý dojem ale nemusí vzniknúť len zo stretnutia naživo, človek si ho môže vytvoriť aj z pohľadu na fotografiu, video alebo vypočutia zvukovej nahrávky. Ak si človek dokáže prezrieť video a odhadnúť z neho osobnostné vlastnosti nahraného človeka, dokáže podobnú analýzu vykonať aj počítač?

Ak by bol počítač schopný si vytvárať prvý dojem z ľudí tak ako to dokážu aj ľudia, mohlo by to viesť k zrýchleniu a zautomatizovaniu niektorých činností v spoločnosti. Jedna z činností, ktorá by sa dala zrýchliť by boli náborové konania a pracovné pohovory, pri ktorých by si uchádzač poslal krátke video a počítač by ho dokázal analyzovať a určiť vhodnosť uchádzača na danú pozíciu.

Pretože osobnosť človeka sa skladá z veľkého množstva osobnostných vlastností, musíme si zvoliť vhodný model na efektívny popis osobnosti človeka, aby sme mohli v počítači túto osobnosť interpretovať. Vhodný model na popis osobnostných vlastností je takzvaná **Veľká päťka** [12]. Veľká päťka modeluje osobnosť za pomoci piatich vlastností: extraverzia, prívetivosť, svedomitosť, neuroticizmus a otvorenosť k novým veciam. Každá z týchto vlastností je modelovaná jednou hodnotou, ktorá značí príslušnosť k danej vlastnosti, alebo vzdialenosť od nej. Aj napriek vhodnému modelu, jednotná analýza osobnostných vlastností z videa nemusí byť efektívna najmä pre rôznorodosť etnicít a individuálne odchýlky v reči ľudí.

V mojej práci som modeloval a porovnával systémy pre odhad dojmových osobnostných vlastností pomocou konvolučných neurónových sietí. Pri modelovaní jednotlivých systémov som používal vizuálnu a audiálnu modalitu videí z dátovej sady *First Impression* vytvorenej pre súťaž *Apparent Personality Analysis and First Impressions Challenge* [13]. Pri modelovaní som porovnával ako fungujú rôzne hĺbky sietí, riešenie pomocou regresie a klasifikácie, alebo aké príznaky fungujú najlepšie pre vizuálnu modalitu.

Pri mojom experimentovaní s dátovou sadou som zistil mnoho zaujímavých skutočností. Napríklad že pri malej dátovej sade je vhodné modelovať systémy s menším počtom parametrov, alebo že klasifikačné riešenie môže fungovať lepšie ako regresné riešenie. Vyhodnotenia jednotlivých experimentov som porovnával s výsledkami tímov v súťaži *Apparent Personality Analysis and First Impressions Challenge* [13], a ak by boli tieto výsledky porovnateľné, tak by môj najlepší systém obsadil v súťaži 8. miesto.

Kapitola 2

Konvolučné neurónové siete

V dnešnej dobre najrozšírenejšie metódy strojového učenia na spracovanie videa sú konvolučné neurónové siete. Konvolučné neurónové siete sa rozšírili do všemožných odvetví spracovávania a generovania zvuku, obrázkov, ale aj iných dát. Ich popularita stále rastie a to najmä preto, že dosahujú veľmi dobré výsledky vo všetkom čo robia, vo veľa prípadoch dosahujú lepšie výsledky ako ľudský náprotivok. Konvolučné siete pri spracovávaní údajov postupujú po krokoch, po vrstvách. Táto kapitola zhŕňa všetky vrstvy použité v tejto práci a prídavné funkcie, ktoré boli pri konvolučných neurónových sieťach použité.

2.1 Konvolučná vrstva

Konvolučná vrstva je základným prvkom konvolučných neurónových sietí. Svoj názov dostala podľa matematickej operácie, ktorú vykonáva, konvolúcie. Konvolúcia je operácia dvoch funkcií, ktorá je definovaná ako váhovaný súčet cez podčast prvej funkcie (vstup) určenej druhou funkciou (jadro). V prípade konvolučných neurónových sietí sa používa diskrétna konvolúcia 2.1, v ktorej f a g sú funkcie, M je veľkosť konvolučného jadra.

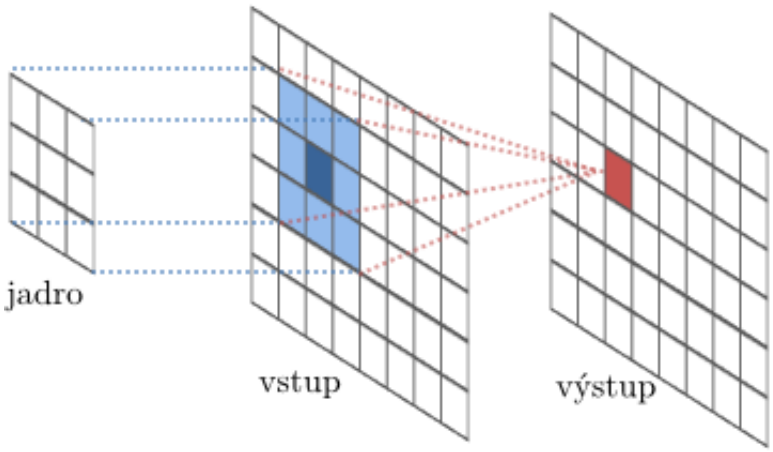
$$(f \star g)[x] = \sum_{m=-M}^{M} f[x-m]g[m] \tag{2.1}$$

Názorný príklad konvolúcie je vidieť na obrázku 2.1[1]. Pri konvolučných neurónových sieťach sú konvolučné vrstvy definované počtom konvolučných jadier, veľkosťou konvolučných jadier a veľkosťou kroku jadra medzi konvolúciami. Počet konvolučných jadier určuje počet kanálov výstupných dát. Veľkosť kroku jadra medzi konvolúciami určuje veľkosť výstupných dát, ak je nastavená na 1, výstupné dáta majú rovnakú veľkosť, ale čím je krok väčší, toľko krát sú výstupné dáta zmenšené.

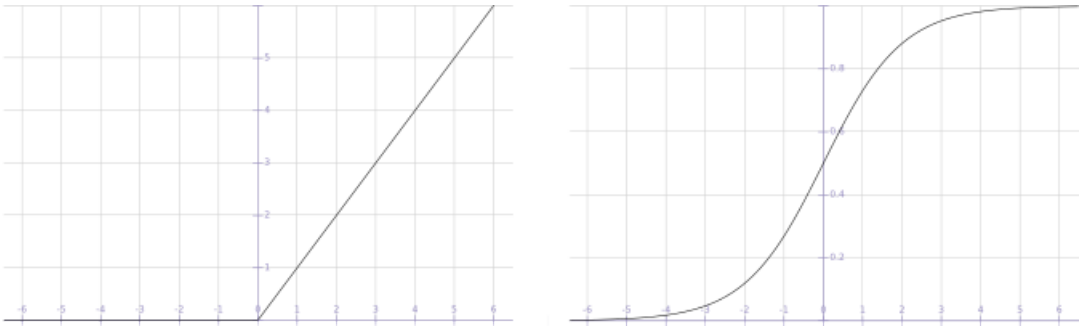
2.2 Aktivačné funkcie

V konvolučných neurónových sieťach často za konvolučnými vrstvami nájdeme nelineárne aktivačné funkcie. Aktivačné funkcie vďaka svojej nelinearite zabraňujú zdegradovaniu siete na lineárne riešiteľný problém, ale niektoré aktivačné funkcie slúžia aj na normalizáciu výstupu. Pri konvolučných neurónových sieťach sa najčastejšie používajú aktivačné funkcie

[1] Zdroj: http://colah.github.io/posts/2014-07-Understanding-Convolutions/img/RiverTrain-ImageConvDiagram.png



Obr. 2.1: Ukážka 2D konvolúcie

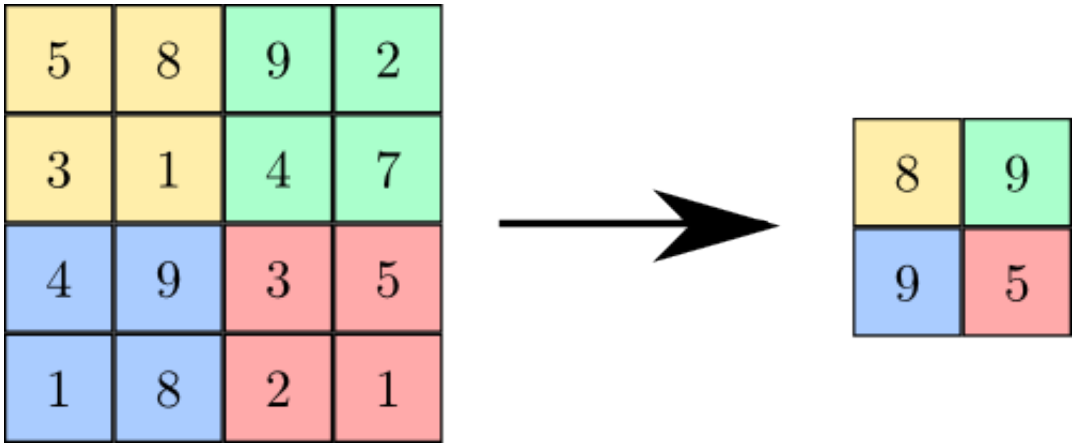


Obr. 2.2: Aktivačné funkcie ReLU (vľavo) a sigmoida (vpravo)

ReLU a sigmoida na odlinearizovanie a softmax na normalizáciu výstupov. ReLU je funkcia, ktorá všetky záporné hodnoty zobrazuje do nuly, tým sa dajú deaktivovať niektoré cesty v neurónovej sieti, čo umožní sieti zamerať sa iba na určité časti vstupov. Sigmoida je funkcia, ktorá zobrazuje vstupy do intervalu $\langle 0, 1 \rangle$, táto funkcia sa využíva menej, pretože v okolí nuly sa funkcia správa lineárne a vo veľkých vzdialenostiach od nuly sa môžu hodnoty saturovať. Porovnanie ReLU a sigmoidy je možné vidieť na obrázku 2.2. Aktivačná funkcia softmax sa používa pri klasifikačných sieťach a to výhradne na konci siete, pretože táto funkcia normalizuje vstupy do kvázi pravdepodobností, čo pri klasifikácií umožňuje výber výslednej triedy.

2.3 Podvzorkovacia vrstva

Podvzorkovacia vrstva je ďalšia často používaná vrstva v konvolučných neurónových sieťach. Podvzorkovanie slúži najmä na zrýchlenie výpočtov siete a zníženie počtu parametrov siete. V konvolučných neurónových sieťach sa najčastejšie používajú podvzorkovania podľa maxima a priemeru. Podvzorkovanie je určené veľkosťou podvzorkovacej matice a kroku. Pri podvzorkovaní sa z časti určenej podvzorkovacou maticou buď vyberie maximum, alebo vypočíta priemer. Tieto hodnoty sú výsledkom jednej podvzorkovacej časti, následne sa podvzorkovacia matica posunie o veľkosť kroku a proces sa opakuje, až kým matica nedôjde do konca vstupných dát. Tento proces je možné vidieť na obrázku 2.3. Od veľkosti



Obr. 2.3: Princíp fungovania podvzorkovania podľa maxima s podvzorkovacou maticou 2x2 a krokom 2

kroku závisí veľkosť výstupných dát, rovnako ako pri konvolučnej vrstve, ak je hodnota kroku 2, výsledné dáta budú 2 krát menšie ako vstupné.

2.4 Plne prepojená vrstva

Častým zakončením konvolučnej neurónovej siete sú plne prepojené vrstvy. Tieto vrstvy sú základným stavebným blokom aj čisto neurónových sietí, pretože modelujú správanie neurónov zo živočíšnej ríše. Podstatou plne prepojenej vrstvy je to, že všetky vstupy sa podieľajú na hodnote každého výstupu. V konvolučných neurónových sieťach sa plne prepojené vrstvy používajú na spracovávanie aktivačných príznakov z konvolučných vrstiev a odhadovanie nelineárnej funkcie popisujúcej vstupné dáta. Pri konvolučných neurónových sieťach sú plne prepojené vrstvy určené iba počtom výstupov. Názornú ukážku plne prepojenej vrstvy je možné vidieť na obrázku 2.4[2]

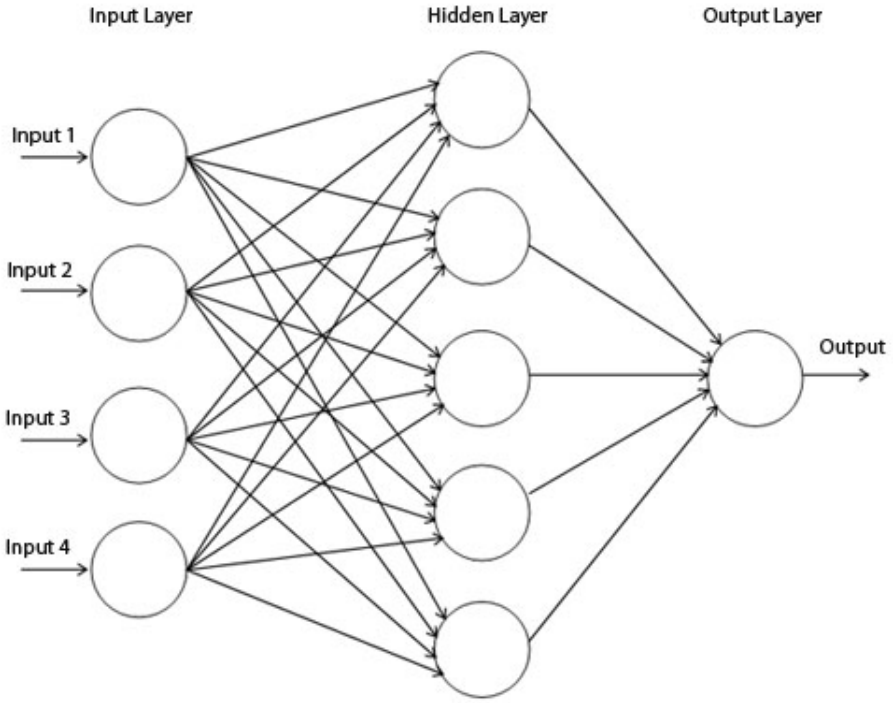
2.5 Dávková normalizácia

Dávková normalizácia [8] je ďalšia vrstva, ktorá sa používa v konvolučných neurónových sieťach. Jej úlohou je normalizovať každý príznak tak, aby cez dávku (mini-batch) mal strednú hodnotu blízku 0 a priemernú odchýlku rovnú 1. Dávková normalizácia je týmto schopná stabilizovať rozloženie aktivačných hodnôt, čo zabraňuje saturácií aktivácií príznakov a zrýchľuje trénovanie. Pri trénovaní sa normalizácia uskutočňuje nad hodnotami z jednej dávky, ale pri testovaní siete sa berú štatistické hodnoty získané počas trénovania.

2.6 Chybové funkcie

Pri trénovaní konvolučných neurónových sietí je nutné po vyhodnotení trénovacích dát určiť chybu od požadovaného vyhodnotenia. Na zistenie chyby sa používajú chybové funkcie. Základnou úlohou chybových funkcií je určiť chybu tak, aby z nej bolo možné vypočítať gradient, ktorý sa používa na úpravu parametrov siete. To znamená, že chybová funkcia je

[2]Zdroj: https://i.stack.imgur.com/gIHAN.jpg



Obr. 2.4: Ukážka plne prepojenej vrstvy s jednou skrytou vrstvou

nevyhnutnou súčasťou trénovania. Rozličné typy problémov vyžadujú rôzne chybové funkcie. Napríklad pre regresné problémy sa používa euklidovská vzdialenosť 2.2 ako chybová funkcia, ktorá určuje odchýlku predpokladanej hodnoty od hodnoty očakávanej. Pri trénovaní sa snažíme čo najviac zmenšiť chybovú funkciu.

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{2.2}$$

Ďalšou chybovou funkciou, ktorá sa používa najmä pri klasifikačných problémoch je krížová entropia. Táto chybová funkcia vyhodnocuje ako veľmi sa líši odhadované pravdepodobnostné rozloženie tried od skutočného pravdepodobnostného rozloženia. Pri trénovaní s touto chybovou funkciou sa snažíme čo najviac zväčšiť pravdepodobnosť správnej triedy spolu so znižovaním pravdepodobností nesprávnej klasifikácie.

Kapitola 3

Odhad osobnostných vlastností

Aj keď v dnešnej dobe je strojové učenie široko rozvinuté v oblastiach zameraných na človeka, či už ide o odhad pózy [2], emócií z textu [3, 18] alebo videa [10, 4], na odhad osobnostných vlastností sa používalo veľmi málo. Najčastejšie použitia na zistenie osobnosti sú z textových príspevkov na sociálnych sieťach [17]. Najviac sa analýza osobnostných vlastností z videa začala využívať po vypísaní sútaže organizáciou *Chalearn*[1], v ktorej si stroje mali vytvoriť prvý dojem o človeku z videa.

3.1 *Looking at People Workshop*

Téma strojového učenia na odhad osobnostných vlastností sa rozšírila v roku 2016, keď organizácia *ChaLearn* usporiadala súťaž *Looking at People Workshop on Apparent Personality Analysis and First Impressions Challenge*[2], ktorej cieľom bolo odhadnúť dojmové osobnostné vlastnosti ľudí z krátkeho videa.

Pre túto súťaž bola vytvorená dátová sada *First Impression* [13] pozostávajúca z 10000 krátkych videí. Tieto videá boli rozdelené do troch tried: 6000 trénovacích videí so zverejnenými ohodnoteniami, 2000 testovacích videí a 2000 evaluačných videí bez zverejneného ohodnotenia. Všetky videá boli stiahnuté z YouTube[3] a rozdelené do 15 sekundových sekvencií. Jednotlivé videá boli pracovníkmi *Amazon Mechanical Turk* po dvojiciach porovnávané v piatich kategóriách podľa osobnostných vlastností modelu „Veľkej Päťky“: extraverzia, prívetivosť, svedomitosť, neuroticizmus a otvorenosť novým zážitkom. Následne boli v každej kategórií videá zoradené a ich poradie prevedené na hodnoty v intervale $\langle 0, 1\rangle$ podľa normálneho rozloženia. Tento spôsob ohodnocovania je vhodný, pretože vierohodne zachytáva prvé dojmy z ľudí. Pri vyhodnocovaní úspešnosti systémov zúčastnených tímov organizátori následujúci vzorec

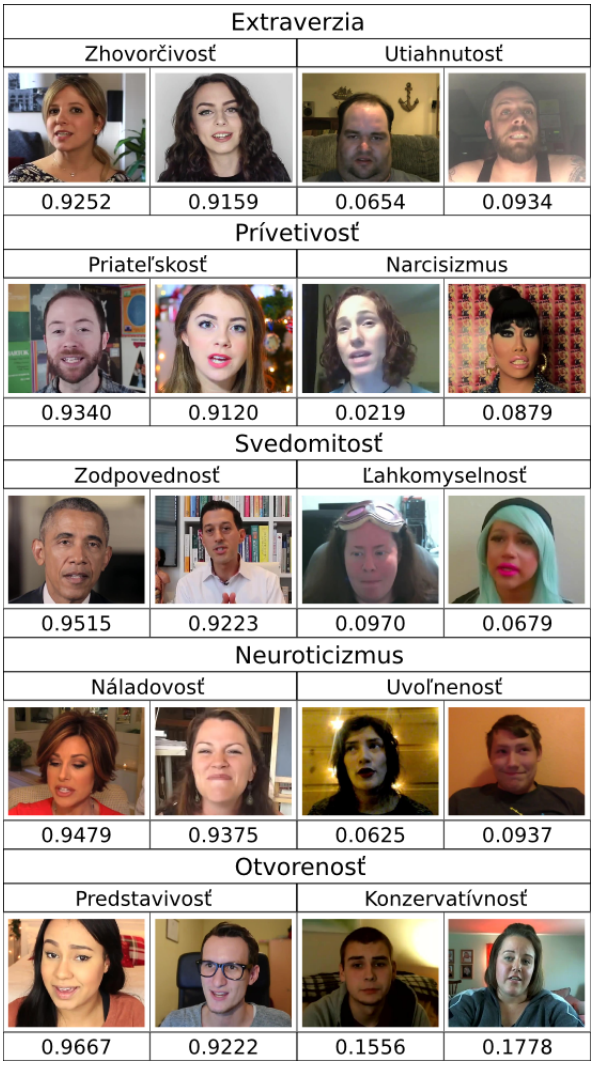
$$A \;=\; 1 \;-\; \frac{1}{N_t} \sum_{i=1}^{N_t} |t_i - p_i| \tag{3.1}$$

V tomto vzorci N_t značí celkový počet vyhodnocovacích dát, t_i je skutočné ohodnotenie videa a p_i je odhadované ohodnotenie videa. Keďže sa definičný obor ohodnotení pohybuje v intervale $\langle 0, 1\rangle$, tak je tento vzorec prepočet priemernej odchýlky ohodnotení na presnosť odhadu.

[1] http://chalearnlap.cvc.uab.es/

[2] http://gesture.chalearn.org/2016-looking-at-people-eccv-workshop-challenge

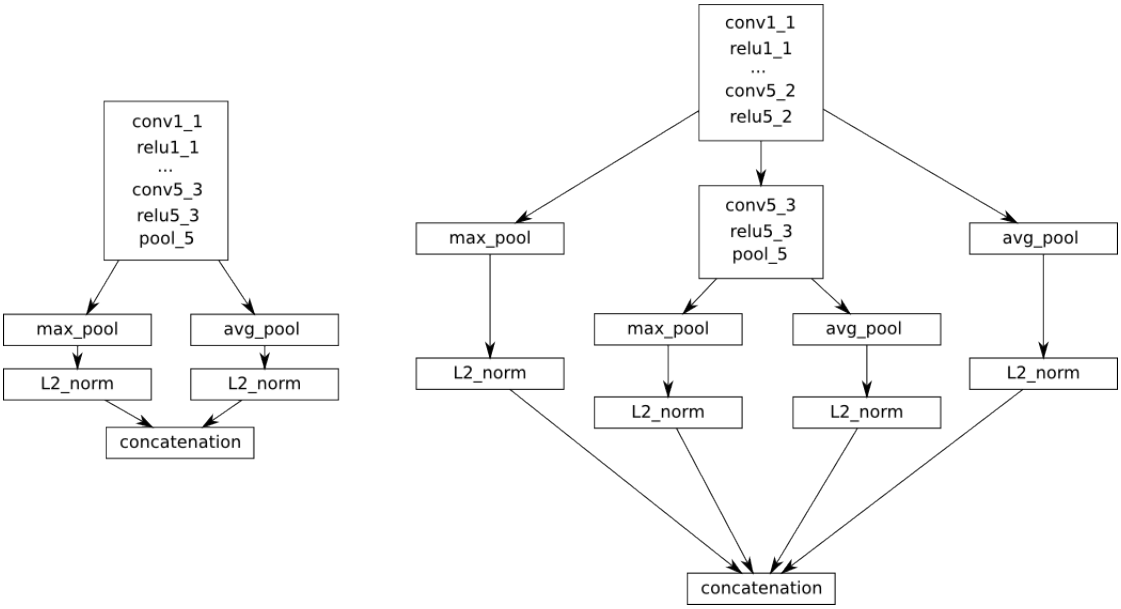
[3] https://www.youtube.com/



Obr. 3.1: Ukážka dátovej sady *First Impression* s hraničnými hodnotami jednotlivých vlastností

Aj napriek vhodnému vytvoreniu dátovej sady je práca s ňou náročná kvôli rôznorodosti kvality videa a audia, diverzite národností a etnicít, ale aj možnou nepresnosťou v anotáciách videí, ktorá môže byť zapríčinená rasovými alebo inými predsudkami voči osobám na videu. Ukážku dátovej sady je možné vidieť na obrázku 3.1.

V súťaži je na vyobrazenie dojmových osobnostných vlastností použitý model veľkej päťky, ktorý pozostáva z piatich vlastností: extraverzia, prívetivosť, svedomitosť, neuroticizmus a otvorenosť novým zážitkom. V tomto modele vysoká hodnota extraverzie značí, že je človek výrečný a zhovorčivý. Naopak nízka hodnota extraverzie naznačuje, že človek pôsobí utiahnuto a ticho. Ak človek pôsobí priateľsky a myslí na druhých, tak má vysokú hodnotu prívetivosti, ale na druhej strane ak pôsobí namyslene a narcisticky, tak má prívetivosť nízku. Hodnoty pre svedomitosť sa zvyšujú, ak človek vyzerá zodpovedne a organizovane, a znižujú, ak sa javí ľahkomyseľne a roztržito. Náladový a nervózny človek sa v modely prejaví vysokou hodnotou pri neuroticizme, ale uvoľnený a pokojný bude mať túto hodnotu nízku. Najzložitejšou osobnostnou vlastnosťou je otvorenosť novým zážitkom,



Obr. 3.2: Architektúra konvloučných neurónovej siete DAN (vľavo) a DAN+ (vpravo) [21]

pretože vysokú hodnotu tejto vlastnosti majú ľudia, ktorý sú umelecky založený, ale aj ľudia, ktorý obľubujú intelektuálne aktivity. Nízku hodnotu otvorenosti majú konzervatívny ľudia, ktorý neradi menia seba alebo svoje okolie.

3.2 Spracovávanie obrazu

Spracovanie videa má dve základné modality. Prvou z nich je obraz a druhou zvuk. Obraz sa dá spracovávať rôznymi spôsobmi, ale takmer všetky popredné tímy zo súťaže si na spracovanie obrazu zvolili konvolučné neurónové siete.

Tím **NJU-LAMBDA** [21], ktorý dosiahol na súťaži najlepšie výsledky, si najprv z videa extrahuje približne 100 snímkov, ktoré sú následne zmenšené na rozlíšenie 224x224, bez ďalších úprav. Tieto snímky spracovával architektúrou siete, ktorú nazval *Descriptor aggregation network* (DAN)[19] a jej rozšírenou verziou DAN+. Tieto architektúry sa od klasických konvolučných neurónových sietí líšia tým, že namiesto koncových plne prepojených vrstiev si zo vstupu získavajú deskriptor. Deskriptor sa zo vstupu získava spojením dvoch pri architektúre DAN, alebo štyroch pri architektúre DAN+ nezávislých podvzorkovacích vrstiev. Každá podvzorkovacia vrstva produkuje 512-prvkový vektor. Vektory z každej podvzorkovacej vrstvy sa nakoniec spoja a tým vznikne 1024 prvkový deskriptor pre architektúru DAN alebo 2048 prvkový deskriptor pre architektúru DAN+. Výsledné deskriptory sú privedené na jednu plne prepojenú vrstvu zakončenú sigmoidou. Použitie takýchto architektúr má za následok rýchlejšie trénovanie a menší počet parametrov siete pri zachovaní až zlepšení výsledkov siete. Na trénovanie architektúry je použitá pred-trénovanú sieť *VGG Face Descriptor* [11]. Pri vyhodnocovaní videa sú získané výsledky všetkých snímkov a následne je z nich vypočítaná priemerná hodnotu pre každú vlastnosť.

Tím **evolgen** [16] si na spracovanie obrazu zvolil rekurentnú LSTM sieť. Riešenie tohto tímu sa sústredilo na predspracovanie obrazu a to tak, že z každého snímku videa si skonštruovali 3D model tváre, tieto modely sa zoradia podľa času a rozčlenia do šiestich po sebe idúcich sekvencií. Jednotlivé sekvencie sú privedené na vstup 3D konvolučnej sieti, ktorá

skúma časové vzťahy. Výsledky z 3D konvolučnej siete sú spolu so zvukom pripojené na vstup vyššie spomínanej rekurentnej siete. Výstupom rekurentnej siete je šesť odhadov pre každú vlastnosť, ktoré sú nakoniec spriemerované aby vznikla jedna hodnota pre každú osobnostnú vlastnosť.

Tretí tím, **DCC**, [5] použil riešenie pomocou reziduálnej siete. Ich architektúra má na vstupe snímok z videa a ten sa spracováva jednou konvolučnou vrstvou a ďalej ôsmimi reziduálnymi blokmi. Každý reziduálny blok pozostáva z dvoch konvolučných vrstiev, dávkovej normalizácie a ReLU aktivačnej funkcie. Posledný reziduálny blok je zakončený podvzorkovaním podľa priemeru. Výsledok je po podvzorkovaní spolu so zvukom napojený na jednu plne prepojenú vrstvu, ktorá priamo počíta hodnoty vlastností.

Ostatné tímy používali metódy podobné vyššie spomenutým. Napríklad tím **BU-NKU** [6] používal pred-trénované VGG siete [11, 15] na získanie príznakov tváre a prostredia, ktoré následne spojili a vyhodnotili pomocou neurónovej siete. Rovnaký prístup zvolil aj tím **pandora** [14], ktorý ale na získanie príznakov používa ešte jednu konvolučnú neurónovú sieť, ktorá by mala extrahovať príznaky, ktoré nepokrývajú prvé dve siete.

3.3 Spracovávanie audia

Dôležitou súčasťou automatického odhadu dojmových osobnostných vlastností je taktiež spracovanie zvuku, pretože prvý dojem nevzniká len z výzoru človeka, ale aj z jeho reči. Na spracovanie audia boli v súťaži použité rôznorodejšie metódy ako pri spracovávaní obrazovej modality.

Tím **NJU-LAMBDA** [21] použil na extrakciu príznakov zo zvukového signálu logaritmované hodnoty energií filtrových bánk (logfbank). Tieto príznaky boli následne privedené na vstup plne prepojenej vrstvy zakončenej aktivačnou funkciou sigmoida. **NJU-LAMBDA** testovali aj použitie Mel-frekvenčných cepstrálnych koeficientov (MFCC) pre ich jednoduchú neurónovú sieť, ale tie však nedosiahli až tak dobré výsledky.

Tím **evolgen** [16] pri spracovávaní zvuku používa metódu, pri ktorej sa najprv rozdelí nahrávka do šiestich neprekrývajúcich sa častí a nad každou časťou je vykonaná spektrálna analýza. Výsledkom spektrálne analýzy je pre každú časť 68-prvkový vektor, ktorý obsahuje príznaky ako energiu, MFCC, entropiu energie a rôzne spektrálne vlastnosti. Tento vektor je spracovaný plne prepojenou vrstvou a zmenšený na 32-prvkový vektor, ktorý je spoločne so spracovanými 3D modelmi tváre vstupom do LSTM rekurentnej konvolučnej neurónovej siete.

Na získanie príznakov zo zvuku pomocou si tím **pandora** [14] zvolil nízko-úrovňové deskriptory (LLD), ktoré produkujú 1428 príznakov. Tieto príznaky sú spracovávané pomocou regresných rozhodovacích stromov. Tím skúšal trénovať regresné rozhodovacie stromy na príznakoch z celých 15 sekúnd zvuku, ale neprodukovalo to prijateľné výsledky, preto si audio rozdelili na približne 2,5 sekundové úseky, ktoré boli použité na trénovanie.

Riešenie tímu **DCC** [5] na spracovanie zvuku je rovnaké, ako ich riešenie na spracovanie obrazu, a to pomocou 17-vrstvovej reziduálnej siete. Podľa informácií od autorov súťaže [13] ostatné tímy na odhad osobnostných vlastností použili buď metódy ako regresia cez podporné vektory (SVR) a regresia pomocou náhodného lesa, alebo zvukovú modalitu videa vôbec nespracovávali.

3.4 Fúzia

Po získaní prvotných odhadov dojmových osobnostných vlastností zo systémov na spracovanie obrazovej a zvukovej modality videa, je vhodné, tieto odhady zjednotiť, aby celkovým výsledkom bola iba jedna hodnota pre každú vlastnosť modelu veľkej päťky. Najjednoduchší spôsob zjednotenia výsledkov je pre každú vlastnosť vypočítať aritmetický priemer zo všetkých výsledkov. Tento spôsob si zvolil aj tím **NJU-LAMBA** [21], ktorý aj napriek jednoduchosti tohto riešenia dosiahol najlepšie výsledky v súťaži.

Riešenia tímov **DCC** [5] a **BU-NKU** [6] sú mierne zložitejšie, pretože nespájajú výsledné odhady, ale získané aktivačné príznaky zo zvuku a obrazu posielajú do dopredných neurónových sietí, ktoré priamo produkujú požadovaný výsledok.

Tím **evolgen** [16] tiež používa spájanie na úrovni aktivačných príznakov, ktoré sú po spojení poslané do LSTM rekurentnej konvolučnej neurónovej siete. Keďže je vstupné video rozdelené do šiestich častí, rekurentná sieť produkuje šesť výsledkov. Z týchto výsledkov je nakoniec vypočítaný aritmetický priemer, ktorého výsledok je jedna hodnota pre každú osobnostnú vlastnosť.

Tím **pandora** [14] pri práci s ich modelom najviac experimentoval so spájaním výsledkov, pretože ich model obsahoval až štyri podsystémy na získavanie predikcií. Všetky podsystémy si vyhodnocovali predikcie iba na základe jedného rámca videa (1 sekundy). Keďže všetky videá v dátovej sade nemajú presne stanovenú dĺžku a ich metóda vyžaduje vektor pevnej dĺžky, tak je nutné upraviť predikcie, aby mali vždy pevnú dĺžku. Tím si zvolil metódu, v ktorej pre videá, ktoré majú väčší počet rámcov ako je požadované, odstráni predikcie z náhodne zvolených rámcov. Na druhej strane, pri videách, ktoré nemajú požadovaný počet rámcov, sú predikcie náhodne zvolených rámcov zduplikované. Po získaní požadovaného počtu predikcií sú všetky predikcie zjednotené a pre každý podsystém je získaná jedná predikcia. Vo finálnej fáze modelu tím skúšal spájanie pomocou aritmetického priemeru, čo viedlo k lepším výsledkom ako dosahovali samostatne fungujúce systémy. K ešte lepším výsledkom viedlo, keď namiesto aritmetického priemeru použili váhovaný priemer, pre ktorý si optimalizovali váhy. Najlepšie výsledky mal ale spôsob, ktorý riešil spájanie predikcií jednotlivých modelov ako ďalší regresný problém.

Kapitola 4

Moje riešenie

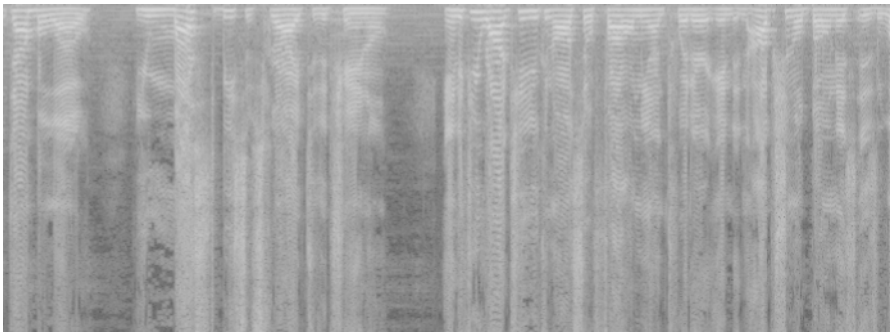
Pri vytváraní vlastného riešenia som sa mierne inšpiroval tímami zo súťaže, ale chcel som vyskúšať spôsoby predspracovania, ako obrazu, tak aj zvuku, ktoré tímy v súťaži nepoužili. Takýto postup som si zvolil, aby som vyskúšal ako jednotlivé spôsoby predspracovania ovplyvnia výsledky a porovnal ich medzi sebou.

4.1 Lineárne regresory

Ako prvú metódu som použil lineárne regresory, pretože sú jednoduché a v niektorých prípadoch produkujú výsledky porovnateľné s neurónovými sieťami. Keďže lineárny regresor dokáže vyprodukovať iba jednu hodnotu, musel som na spracovanie jedného vstupu natrénovať až päť regresorov, jeden pre každú odhadovanú vlastnosť. Pre trénovanie regresorov som si rozdelil dátovú sadu na trénovaciu a testovaciu časť. Trénovacia časť pozostávala zo 4000 videí a testovacia zo zvyšných 2000 videí.

Vo svojej práci som si zvolil, že natrénujem dve sady regresorov, jedny na príznaky z obrazu a druhé na príznaky zo zvuku. Pri spracovaní obrazu som si na získanie príznakov zvolil konvolučnú neurónovú sieť. Najprv som si ale predspracoval všetky videá tak, že som si z nich extrahoval desať snímkov s frekvenciou jeden snímok za sekundu. Tieto snímky som následne orezal tak, aby vznikol štvorcový snímok, ktorý som následne zmenšil na rozlíšenie 227x227. Spracované snímky som posielal do konvolučnej neurónovej siete *Places205-AlexNet* [22], ktorá slúži na klasifikáciu prostredia na obrázku. Túto sieť som si zvolil preto, lebo som si myslel, že prostredie, v ktorom sa ľudia nachádzajú môže ovplyvniť prvý dojem. Po spracovaní snímkov konvolučnou neurónovou sieťou som si zo siete získal aktivačné príznaky, a to z úplne poslednej plne prepojenej vrstvy, v ktorej sú zastúpené hodnoty rozhodnutí pre jednotlivé umiestnenia. Po získaní príznakov zo všetkých videí som pre každú odhadovanú osobnostnú vlastnosť natrénoval jeden lineárny regresor.

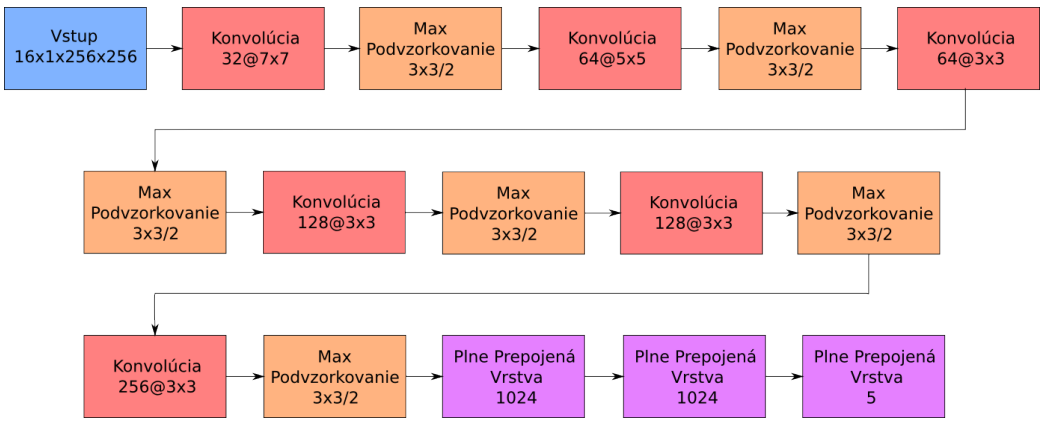
Pri spracovávaní zvukovej modulácie videa som najprv z videa extrahoval audio. Z extrahovaného audia som získal príznaky a ako príznaky som si zvolil logaritmované hodnoty energií filtrových bánk. Tieto príznaky som získaval z rámcov dĺžky 25 milisekúnd a za sebou idúce rámce mali prekryv 10 milisekúnd. Na spracovanie rámca bolo použitých 26 filtrov. Príznaky som získaval vždy z celých nahrávok, čo viedlo na veľký počet parametrov, až 40000 pre každú nahrávku. Aj napriek veľkému počtu vstupných parametrov som zo získaných príznakov namodeloval lineárne regresory pre jednotlivé vlastnosti.



Obr. 4.1: Ukážka vytvoreného spektrogramu zo zvukovej modality videí z dátovej sady *First Impression*

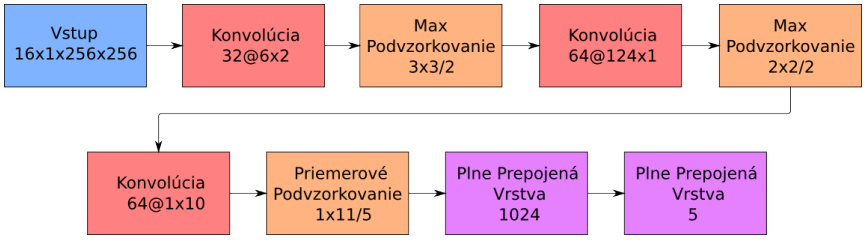
4.2 Spracovanie zvuku konvolučnými neurónovými sieťami

Po natrénovaní lineárnych regresorov som sa rozhodol experimentovať s konvolučnými neurónovými sieťami a to najmä so sieťami na spracovávanie zvuku. Pri experimentovaní som sa inšpiroval autorom Harutyunyan-om [7], ktorý v súťaži na identifikáciu jazyka z nahrávky použil na reprezentáciu audia spektrogramy. Chcel som tento prístup vyskúšať a preto som si z nahrávok, ktoré som extrahoval z videa pri modelovaní lineárnych regresorov, vytvoril sivo tónované spektrogramy pomocou skriptu, ktorý autor Harutyunyan vytvoril pri jeho práci[1]. Získané spektrogramy som použil ako trénovacie dáta pre mnou navrhnutú hlbokú konvolučnú neurónovú sieť 4.2. Ako vstup do siete som nepoužíval spektrogram z celej sekvencie, ale vyberal som v čase náhodný výsek 256x256. Používaním náhodných výsekov v čase dostávam viac trénovacích dát, čo je prospešné pri tak malej dátovej sade ako je *First Impression*. Použitá sieť sa skladá zo šiestich konvolučných vrstiev a troch plne prepojených vrstiev. Každá konvolučná vrstva je bezprostredne nasledovaná aktivačnou funkciou ReLU a podvzorkovaním podľa maxima. Plne prepojené vrstvy sú taktiež nasledované aktivačnou funkciu ReLU, iba posledná plne prepojená vrstva je nasledovaná aktivačnou funkciou sigmoida, ktorej výstup je vektor piatich vlastností v tomto poradí: extraverzia, prívetivosť, svedomitosť, neuroticizmus a otvorenosť novým zážitkom. Pri experimentovaní som skúšal



Obr. 4.2: Architektúra hlbokej konvolučnej neurónovej siete na spracovanie spektrogramov

[1] https://github.com/YerevaNN/Spoken-language-identification/blob/master/create_spectrograms.py



Obr. 4.3: Architektúra konvolučnej neurónovej siete s priemerovaním na spracovanie spektrogramov

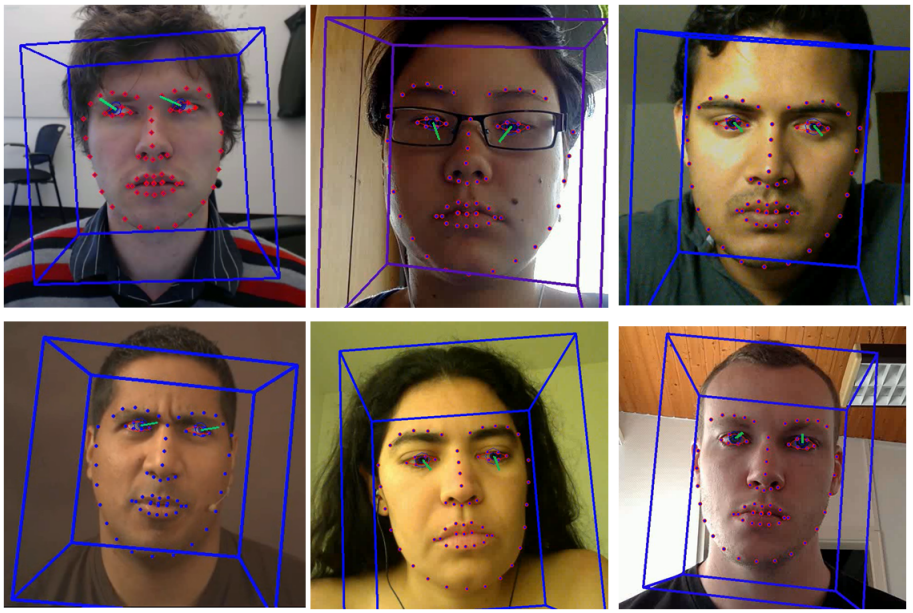
aj variantu malej siete. Navrhol som sieť, ktorá mala dve konvolučné vrstvy so 64-mi konvolučnými jadrami rozmeru 5x5 a jednu plne prepojenú vrstvu, ktorej výstup ešte prešiel cez aktivačnú funkciu sigmoida.

Keďže sa pri spracovávaní spektrogramov jedná o sekvenciu, navrhol som ďalšiu sieť, ktorá nepoužívala štvorcové konvolučné jadrá filtrov, aby sa sieť správala rozdielne k hodnotám energií a ich časovému priebehu. Túto sieť som oproti sieti 4.2 zmenšil tak, že má iba tri konvolučné vrstvy a dve plne prepojené vrstvy. Za prvými dvoma konvolučnými vrstvami sa nachádza dávková normalizácia nasledovaná aktivačnou funkciou ReLU. Za poslednou konvolučnou vrstvou sa nachádza podvzorkovanie podľa priemeru. Podvzorkovanie na konci konvolučných vrstiev slúži na zabezpečenie invariancie siete voči časovému posunutiu príznakov.

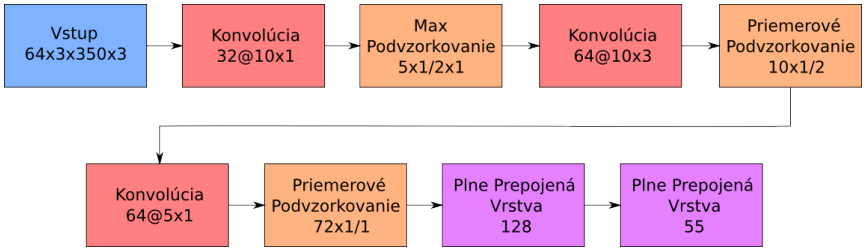
Predchádzajúce siete priamo odhadovali dojmové osobnostné vlastnosti, používali regresiu. Chcel som vyskúšať, ako by sa zmenila úspešnosť výsledkov, ak by som sa na problém pozeral ako na klasifikáciu. Na to, aby som mohol problém riešiť ako klasifikačnú úlohu, som si najprv rozdelil interval $\langle 0, 1\rangle$ do jedenástich tried. Pri rozdeľovaní hodnôt do tried som dodržiaval normálne rozdelenie. Finálne rozdelenie je možné vidieť v rovnici 4.1.

$$trieda(x) = \begin{cases} 0 & \text{ak } x \in \langle 0; 0,2) \\ 1 & \text{ak } x \in \langle 0,2; 0,3) \\ 2 & \text{ak } x \in \langle 0,3; 0,4) \\ 3 & \text{ak } x \in \langle 0,4; 0,45) \\ 4 & \text{ak } x \in \langle 0,45; 0,5) \\ 5 & \text{ak } x \in \langle 0,5; 0,55) \\ 6 & \text{ak } x \in \langle 0,55; 0,6) \\ 7 & \text{ak } x \in \langle 0,6; 0,65) \\ 8 & \text{ak } x \in \langle 0,65; 0,7) \\ 9 & \text{ak } x \in \langle 0,7; 0,8) \\ 10 & \text{ak } x \in \langle 0,8; 1\rangle \end{cases} \tag{4.1}$$

Pri tomto rozdelení majú krajné triedy približne po 700 prvkov a stredné triedy približne 3600 prvkov. Po rozdelení hodnôt do tried som musel upraviť architektúry sietí tak, aby riešili klasifikačný problém a nie regresný. To som dosiahol výmenou úplne poslednej vrstvy s aktivačnou funkciou. Pri regresných problémoch som ako poslednú vrstvu používal aktivačnú funkciu sigmoida, ale pre klasifikačný problém som ju nahradil aktivačnou funkciou *softmax*. Aktivačná funkcia *softmax* prevedie všetky svoje vstupy do kvázi pravdepodobností, ktoré sa používajú na zvolenie výslednej triedy. Pre vyhodnocovacie účely som si musel určiť výsledné hodnoty, ktoré budú produkované systémom v prípade klasifikácie vlastnosti do danej triedy. Jednotlivé výsledné hodnoty pre triedy som počítal ako aritmetický priemer



Obr. 4.4: Ukážka získavania príznakov pohľadu pomocou nástroja *OpenFace*



Obr. 4.5: Architektúra konvolučnej neurónovej siete na spracovávanie pohľadu

hodnôt prvkov, ktoré patria do danej triedy. Pri získavaní koncových hodnôt z klasifikácie som vyskúšal dve metódy. Prvá metóda určovala výslednú hodnotu iba ako vyššie spomenutý aritmetický priemer hodnôt najpravdepodobnejšej triedy. Druhá metóda používala na výpočet odhadovanej hodnoty osobnostnej vlastnosti následujúci vzorec $\sum_{i=0}^{10} \bar{x_i} * p_i$, kde i je trieda, $\bar{x_i}$ je aritmetický priemer hodnôt v triede a p_i je pravdepodobnosť príslušnosti k danej triede získaná z konvolučnej neurónovej siete. Tento vzorec sčíta prenásobia pravdepodobnosti danéj triedy s aritmetickým priemerom jej hodnôt.

4.3 Spracovanie obrazu konvolučnými neurónovými sieťami

Po dokončení experimentov na odhad dojmových osobnostných vlastností zo zvukovej nahrávky, som začal experimentovať s obrazovou modalitou videa. Rozhodol som sa používať predspracované dáta. Najprv som si zvolil, že použijem na odhad dojmových osobnostných vlastností iba pohľad človeka 4.4[2] a jeho zmeny na videu. Príznaky získané z tohto nástroja zahrňujú: natočenie tváre, smer pohľadu ľavého oka a smer pohľadu pravého oka. Získané príznaky som bez použitia normalizácie posielal do mnou navrhnutej siete 4.5. Táto sieť sa skladá z troch konvolučných vrstiev, ktoré rozdielne spracovávajú hodnoty v jednom

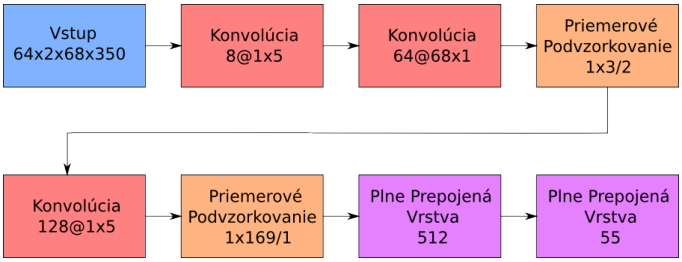
[2]Zdroj: https://github.com/TadasBaltrusaitis/OpenFace/blob/master/imgs/gaze_ex.png

Obr. 4.6: Ukážka zisťovania orientačných bodov tváre pomocou nástroja *OpenFace*

čase a časový priebeh daných hodnôt. Za prvou konvolučnou vrstvou sa nachádza dávková normalizácia a za každou konvolučnou vrstvou nasleduje aktivačná funkcia ReLU. Keďže táto sieť spracováva sekvenciu v čase, za poslednú konvolučnú vrstvu som umiestnil podvzorkovanie podľa priemeru, po ktorom bude výstup jedného konvolučného jadra jedna hodnota, čo zabezpečuje invarianciu príznakov voči časovému posunutiu. Na spracovanie aktivačných príznakov používam dve plne prepojené vrstvy, pričom druhá plne prepojená vrstva má päťdesiatpäť výstupov. Posledná vrstva má taký počet výstupov preto, lebo rieši problém pomocou klasifikácie, podobne ako v podkapitole 4.2.

Pre porovnanie so sieťou z obrázka 4.5 som navrhol ešte jeden spôsob predspracovania obrazovej modality. Tento spôsob zahŕňal zisťovanie polohy orientačných bodov tváre 4.6[3] pre každý snímok z videa. Z každého snímku som získal súradnice pre 68 orientačných bodov tváre. Všetky súradnice som napokon zarovnal na stred pomocou odčítania priemernej hodnoty súradnice zo všetkých snímkov daného videa. Po zarovnaní súradníc orientačných bodov tváre som nad nimi skúšal dva typy normalizácie. Pri prvom type normalizácie som najprv našiel maximálnu absolútnu odchýlku súradnice orientačného bodu zo všetkých snímkov videa. Následne som všetky hodnoty súradníc orientačného bodu vydelil nájdeným maximom. Tento prístup síce nelineárne deformuje priestor obrázka, ale zachováva informácie o pohybe jednotlivých orientačných bodov tváre. Druhý typ normalizácie delil zarovnané hodnoty orientačných bodov tváre podľa veľkosti detekčného okna tváre v snímku, ktoré bolo nájdene ako medziprodukt pri získavaní orientačných bodov tváre. Tento spôsob normalizácie zachováva ako informácie o pohybe jednotlivých orientačných bodov, tak aj proporčné veľkosti jednotlivých pohybov.

[3]Zdroj: https://github.com/TadasBaltrusaitis/OpenFace/blob/master/imgs/multi_face_img.png



Obr. 4.7: Architektúra konvolučnej neurónovej siete na spracovávanie orientačných bodov tváre

Kapitola 5

Experimenty a výsledky

Pre možnosť experimentovania na navrhnutých metódach je nutné si zvoliť nástroje, ktoré sú schopné namodelovať navrhnuté systémy a natrénovať ich. V tejto kapitole je popísané, aké nástroje som si zvolil pre svoje experimenty. Po popise nástrojov je popísané, ako som jednotlivé experimenty vykonával. Pre vyhodnocovanie mojích systémov som použil vzorec 3.1, ktorý bol oficiálnym vyhodnocovacím vzorcom súťaže *Apparent Personality Analysis and First Impressions Challenge* [13]. Tento vzorec počíta úspešnosť ako priemernú odchýlku odhadovaných vlastností od anotácie vide odčítanú od 1.

5.1 Nástroje a trénovanie

Pri predspracovávaní videa som používal nástroj *ffmpeg*[1], ktorý dokáže extrahovať snímky z videa, ale dokáže taktiež extrahovať zvukovú stopu videa. Na získavanie príznakov zo zvuku som používal nástroj *python_speech_features*[2] na spracovávanie zvukových nahrávok pre programovací jazyk *Python*[3]. Na extrakciu príznakov pohľadu človeka a orientačných bodov tváre som si zvolil nástroj *OpenFace*[4] [20, 1], ktorý získava tieto príznaky s veľkou úspešnosťou. Existuje mnoho knižníc, ktoré modelujú metódy strojového učenia, Na prácu s lineárnymi klasifikátormi a regresormi som si zvolil *liblinear*[5]. Túto knižnicu som si zvolil najmä pre jej rýchlosť učenia, pár sekúnd pre milióny vstupných parametrov, v mojom prípade približne štyri sekundy pre vstupný súbor s dvanástimi miliónmi parametrov. Na trénovanie konvolučných neurónových sietí existuje taktiež mnoho rôznych nástrojov, ale ja som si pre svoju prácu zvolil nástroj *Caffe* [9], ktorý vznikol na univerzite v Berkeley a v súčasnosti je vyvíjaný vývojovou skupinou *BAIR*[6]. Pri svojej práci som využíval najmä rozhranie tohto nástroja na programovací jazyk *Python*. Tento nástroj som si zvolil pre jeho rýchlosť, rozsiahlu komunitu a kvalitu rozhrania pre programovací jazyk *Python*. Všetky trénovacie procesy konvolučných neurónových sietí boli vykonávané na strojoch *MetaCentra*[7].

Na trénovanie konvolučných neurónových sietí som zo začiatku používal optimalizačnú metódu *stochastický gradientný zostup* (SGD) s rýchlosťou učenia 0,0005 a rovnako veľkým úpadkom váh (weight decay). Neskôr som začal používať optimalizačnú metódu *AdaDelta*

[1] https://ffmpeg.org/
[2] https://github.com/jameslyons/python_speech_features
[3] https://www.python.org/
[4] https://github.com/TadasBaltrusaitis/OpenFace
[5] https://www.csie.ntu.edu.tw/ cjlin/liblinear/
[6] http://bair.berkeley.edu/
[7] https://metavo.metacentrum.cz/

	Extraverzia	Prívetivosť	Svedomitosť	Neuroticizmus	Otvorenosť	Priemerná úspešnosť
LR pre zvuk	**0,8825**	0,8871	0,8751	**0,8852**	**0,8886**	**0,8837**
LR pre obraz	0,8789	**0,8924**	**0,8797**	0,8786	0,8858	0,8831
Konštantný prediktor	0,8738	0,8888	0,8715	0,8736	0,8826	0,8781

Tabuľka 5.1: Výsledky lineárnych regresorov na odhad dojmových osobnostných vlastností

	Extraverzia	Prívetivosť	Svedomitosť	Neuroticizmus	Otvorenosť	Priemerná úspešnosť
6-vrstvová sieť	0,8751	0,8862	0,8683	0,8692	0,8751	0,8748
3-vrstvová sieť	**0,8840**	**0,8970**	0,8734	0,8802	0,8840	**0,8837**
Najlepší LR	0,8825	0,8871	**0,8751**	**0,8852**	**0,8886**	**0,8837**
Konštantný prediktor	0,8738	0,8888	0,8715	0,8736	0,8826	0,8781

Tabuľka 5.2: Výsledky regresných sietí na odhad dojmových osobnostných vlastností a ich porovnanie s LR

s hodnotou delta nastavenou na 1e-6. Túto metódu som si zvolil pre jej adaptívnosť, ktorá umožňuje podobnú rýchlosť trénovania všetkých vrstiev. Ďalšou výhodou tejto optimalizačnej metódy je jej stabilita, to znamená, že jej výsledok častejšie konverguje. Ako chybovú funkciu som používal Euklidovskú vzdialenosť pri regresných riešeniach a *SoftmaxWithLoss*, ktorá používa krížovú entropiu, pri klasifikačných riešeniach.

Aby som bol schopný trénovať a následne testovať moje systémy musel som si rozdeliť poskytnutú dátovú sadu na trénovaciu a testovaciu časť. Keďže dátová sada obsahovala 6000 ohodnotených videí, rozdelil som ju v pomere 4000 videí pre trénovaciu časť a zvyšných 2000 pre testovaciu časť. Pri rozdeľovaní som vzal do úvahy identity jednotlivých osôb a videá od jednej osoby sa vyskytujú vždy buď iba v trénovacej časti, alebo iba v testovacej časti. Takéto rozdelenie zabezpečuje validitu získaných výsledkov a pri testovaní overuje generalizáciu testovaného systému.

5.2 Lineárne regresory

Ako prvú metódu som si zvolil trénovanie lineárnych regresorov (LR), aby som vyskúšal najjednoduchšiu metódu na riešenie regresného problému. Zároveň s vyskúšaním metódy som si určil základnú hranicu, ktorú som sa snažil v následujúcich pokusoch prekonať. Ako je možné vidieť v tabuľke 5.1, už aj najjednoduchšie riešenie funguje presnejšie ako konštantný prediktor, ktorého odhad je vždy priemerná hodnota ohodnotení danej vlastnosti. Tieto výsledky dokazujú, že mnou zvolené príznaky pre zvuk aj obraz obsahujú informácie o dojmových osobnostných vlastnostiach ľudí. Ako je možné vidieť lineárny regresor spracovávajúci zvuk dosahuje lepšie výsledky, z toho vyplýva, že logaritmované hodnoty energií filtrových bánk obsahujú viac informácie ako prostredie, v ktorom sa človek nachádza

5.3 Regresné konvolučné neurónové siete

Po experimentoch s lineárnymi regresormi som začal experimentovať s konvolučnými neurónovými sieťami a skúšal som ako veľkosť siete ovplyvňuje výsledky. Z tabuľky 5.2 je vidno, že príliš veľká sieť, 6-vrstvová, sa pri tak malej dátovej sade poriadne nenatrénuje, preto je vidieť, že nakoniec nedosiahla ani úroveň konštantného prediktoru. Menšia sieť funguje lepšie ako tá veľká, ale vo výsledku má porovnateľné výsledky s lineárnym regresorom.

	Extraverzia	Prívetivosť	Svedomitosť	Neuroticizmus	Otvorenosť	Priemerná úspešnosť
4-vrstvová klasifikácia, maximum	0,8630	0,8823	0,8294	0,8653	0,8720	0,8624
3-vrstvová klasifikácia, maximum	0,8783	0,8782	0,8678	0,8730	0,8806	0,8756
4-vrstvová klasifikácia, roznásobenie	0,8836	0,8985	0,8719	0,8824	0,8849	0,8843
3-vrstvová klasifikácia, roznásobenie	**0,8904**	**0,8994**	**0,8800**	**0,8852**	**0,8890**	**0,8888**
Regresné vyhodnocovanie	0,8840	0,8970	0,8734	0,8802	0,8840	0,8837
Konštantný prediktor	0,8738	0,8888	0,8715	0,8736	0,8826	0,8781

Tabuľka 5.3: Porovnanie výsledkov riešenia odhadu dojmových osobnostných vlastností pomocou klasifikačného riešenia

5.4 Klasifikačné konvolučné neurónové siete

Následne som porovnával ako funguje riešenie odhadu dojmových osobnostných vlastností ako klasifikačný problém oproti riešeniu problému ako regresiu. Pretože natrénované konvolučné neurónové siete mali presnosť klasifikácie 15-18% pri klasifikácií do 11 tried, rozhodol som sa, že pre rozhodnutie hodnoty vlastnosti použijem strednú hodnotu klasifikovanej triedy. Ako je možné vidieť v tabuľke 5.3, tento spôsob nedosahuje ani úspešnosť konštantného prediktoru. Z tohto dôvodu som sa rozhodol generovať hodnotu súčtom roznásobenia pravdepodobnosti tried s ich strednými hodnotami. Z tabuľky 5.3 vyplýva, že tento spôsob odhaduje vlastnosti s výrazne väčšou presnosťou ako výber triedy s najväčšou pravdepodobnosťou a zároveň dosiahol lepšie výsledky ako konštantný prediktor a riešenie problému pomocou regresie.

5.5 Konvolučné neurónové siete na spracovanie obrazovej modality videa

Pri spracovávaní obrazu som najprv trénoval sieť, ktorá odhadovala vlastnosti z pohľadu a jeho pohybu. Keďže sa dá pohľad charakterizovať troma hodnotami v každom snímku, trénovaná sieť nesmie mať veľa vrstiev, pretože by mohlo dôjsť k pretrénovaniu. Ja som trénoval 3-vrstvovú konvolučnú sieť. Výsledky tejto siete som porovnával so sieťou, ktorá odhadovala vlastnosti z orientačných bodov tváre a ich pohybu. Pri orientačných bodoch tváre som taktiež vyskúšal dva typy normalizácie, ktoré som medzi sebou porovnával. Prvý typ normalizácie orientačného bodu je podľa maximálneho pohybu tohto bodu a druhý typ je podľa veľkosti detekovaného okna tváre. Z tabuľky 5.4 je vidno, že najlepšie výsledky zo sietí využívajúcich obrazovú modalitu videa funguje sieť, ktorá spracováva orientačné body tváre normalizované podľa ich maximálneho pohybu. Taktiež je vidno, že pri normalizácií podľa veľkosti detekovaného okna sieť dosahuje výsledky veľmi blízke konštantnému prediktoru. Toto môže byť spôsobené meniacou sa veľkosťou detekovaného okna medzi jednotlivými snímkami, a tým sa môžu menšie pohyby orientačných bodov stratiť. Výsledky poukazujú, že sieť spracovávajúca pohľad funguje lep3ie ako konštantný prediktor aj napriek malému množstvu vstupných parametrov, čo dokazuje, že aj v troch hodnotách pohľadu sa vyskytuje dostatočné množstvo informácií na odhad dojmových osobnostných vlastností človeka.

5.6 Vyhodnotenie a možné zlepšenie

Ako je možné vidieť z predchádzajúcich výsledkov moje najlepšie systémy pracujú s úspešnosťou 88,88% pri spracovávaní zvuku a 88,93% pri spracovávaní obrazu. Ak tieto výsledky

	Extraverzia	Prívetivosť	Svedomitosť	Neuroticizmus	Otvorenosť	Priemerná úspešnosť
Pohľad	0,8812	0,8932	0,8796	0,8835	0,8892	0,8853
Orientačné body, normalizácia pohybu	0,8894	0,8948	0,8834	0,8865	0,8924	0,8893
Orientačné body, normalizácia veľkosťou okna	0,8785	0,8906	0,8733	0,8777	0,8858	0,8812
Najlepší model zvuku	**0,8904**	0,8994	0,8800	0,8852	0,8890	0,8888
Konštantný prediktor	0,8738	0,8888	0,8715	0,8736	0,8826	0,8781

Tabuľka 5.4: Výsledky modelov spracujúcich obrazovú modalitu videa na odhad dojmových osobnostných vlastností

Umiestnenie	Tím	Úspešnosť
1.	NJU-LAMBDA	0,9130
2.	evolgen	0,9121
3.	DCC	0,9109
4.	ucas	0,9098
5.	BU-NKU	0,9094
6.	pandora	0,9063
7.	Pilab	0,8936
(8.)	Môj systém	0,8893
8.	Kaizoku	0,8826
9.	ITU_SiMiT	0,8815
10.	sp	0,8759

Tabuľka 5.5: Výsledky súťaže s pridaním môjho hodnotenia, ktoré ale nie je plne porovnateľné s výsledkami ostatných tímov

porovnám s výsledkami ostatných tímov, obsadili by moje systémy 8. miesto. Všetky moje výsledky boli vyhodnocované na rozdielnych videách, ako boli trénované, ale aj napriek tomu moje výsledky nie sú porovnateľné s výsledkami zo súťaže. Celkovo moje systémy fungujú s podobnou úspešnosťou ako systémy ostatných tímov, ale možno by sa dali zlepšiť použitím celých snímkov videa, pretože obsahujú viac informácií ako mnou použité príznaky, takýmto spôsobom dosiahol tím **NJU-LAMBDA** [21] najlepšie výsledky v súťaži. Ďalej by sa moje výsledky dali vylepšiť spojením systémov na spracovanie zvuku a obrazu, pretože by sa vzájomne regulovali.

Kapitola 6

Záver

Cieľom tejto bakalárskej práce bolo navrhnúť systémy na odhad dojmových osobnostných vlastností z videa. Rovnaký cieľ mala aj súťaž *Apparent Personality Analysis and First Impressions Challenge* [13], z ktorej bola použitá dátová sada na trénovanie všetkých modelov.

V prvej časti tejto práce som experimentoval s lineárnymi regresormi. Pri práci s lineárnymi regresormi som spracovával obrazovú aj zvukovú modalitu videí z dátovej sady. Pri obrazovej modalite som si z desiatich snímkov videa extrahoval príznaky pomocou konvolučnej neurónovej siete. Pri zvukovej modalite som si zo zvukovej nahrávky extrahoval logaritmované hodnoty energií filtrových bánk, ktoré som používal ako príznaky. Nad získanými príznakmi som namodeloval lineárne regresory, ktoré aj napriek tomu, že sú najjednoduchším možným riešením problému, dosahujú lepšie výsledky ako konštantný prediktor. Lineárny regresor, ktorý spracovával zvuk dosahoval lepšie hodnoty ako ten, ktorý spracovával obraz.

V ďalšej časti som experimentoval s konvolučnými neurónovými sieťami, ktoré spracovávali zvukové údaje vo forme spektrogramov. Prvé experimentovanie pozostávalo zo skúšania rôznych počtov konvolučných a plne prepojených vrstiev. Pri týchto experimentoch najlepšie výsledky dosahovala sieť, ktorá mala dve konvolučné vrstvy a jednu plne prepojenú vrstvu. Menšie siete dosahovali lepšie výsledky, pretože sa trénovali na malej dátovej sade. Následovné experimentovanie bolo tvorené porovnávaním dvoch spôsobov riešenia problému: regresného spôsobu a klasifikačného spôsobu. Pre riešenie problému klasifikáciou som si definičný obor vlastností rozdelil do jedenástich tried. Klasifikačné riešenia dosahovalo lepšie výsledky v prípade, že sa na konečnom výsledku podieľali pravdepodobnosti všetkých tried. Ak sa na výsledku podieľala iba trieda s najvyššou pravdepodobnosťou, systémy nedosahovali ani úspešnosť konštantného prediktoru.

Pri spracovávaní obrazovej modality som porovnával výsledky príznakov získaných z pohľadu človeka na videu a príznakov získaných z orientačných bodov tváre človeka. Všetky porovnávané siete riešili problém ako klasifikáciu, pretože v predchádzajúcich experimentoch dosahovala lepšie výsledky. Pri experimentovaní so systémami na spracovanie orientačných bodov tváre som skúšal dva typy normalizácie, prvý typ podľa pohybu jednotlivých orientačných bodov a druhý podľa veľkosti detekovaného okna tváre. Z experimentov vyplynulo, že prvý typ normalizácie nad touto dátovou sadu funguje lepšie. Systém na spracovávanie orientačných bodov tváre produkoval lepšie ohodnotenia ako systém na spracovávanie pohľadu, to môže byť spôsobené tým, že orientačné body uchovávajú viac informácie ako samotný pohľad.

Systém s najlepšími výsledkami dokáže odhadovať osobnostné vlastnosti s priemernou odchýlkou 0,1107, čo je podobná odchýlka, akú dosahujú tímy na súťaži *Apparent Persona-*

lity Analysis and First Impressions Challenge [13]. Natrénované systémy pracujú s dobrou úspešnosťou, ale pretože používaná dátová sada je malá, úspešnosť by sa mohla dať zlepšiť použitím vstupných dát, ktoré sa dajú upravovať a dajú sa použiť na generovanie ďalších dát. Takéto generovanie môže byť spôsobené pridaním šumu do audia, náhodným otáčaním obrázkov, alebo používaním náhodných výrezov snímkov. Zaujímavé výsledky by mohli produkovať aj iné metódy spracovania príznakov, ako napríklad regresiou cez podporné vektory, alebo regresnými rozhodovacími stromami.

Literatúra

[1] Baltrusaitis, T.; Robinson, P.; Morency, L.-P.: Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild. In *Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops*, ICCVW '13, IEEE Computer Society, 2013, ISBN 978-1-4799-3022-7, s. 354–361, doi:10.1109/ICCVW.2013.54.

[2] Cao, Z.; Simon, T.; Wei, S.; aj.: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *CoRR*, ročník abs/1611.08050, 2016.

[3] Chaffar, S.; Inkpen, D.: *Using a Heterogeneous Dataset for Emotion Analysis in Text*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, ISBN 978-3-642-21043-3, s. 62–67.

[4] Ebrahimi Kahou, S.; Michalski, V.; Konda, K.; aj.: Recurrent Neural Networks for Emotion Recognition in Video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, New York, NY, USA: ACM, 2015, ISBN 978-1-4503-3912-4, s. 467–474.

[5] Güçlütürk, Y.; Güçlü, U.; van Gerven, M. A. J.; aj.: *Deep Impression: Audiovisual Deep Residual Networks for Multimodal Apparent Personality Trait Recognition*. Cham: Springer International Publishing, 2016, ISBN 978-3-319-49409-8, s. 349–358.

[6] Gürpınar, F.; Kaya, H.; Salah, A. A.: *Combining Deep Facial and Ambient Features for First Impression Estimation*. Cham: Springer International Publishing, 2016, ISBN 978-3-319-49409-8, s. 372–385.

[7] Harutyunyan, H.: Spoken language identification with deep convolutional networks. online.
URL `https://yerevann.github.io/2015/10/11/spoken-language-identification-with-deep-convolutional-networks/`

[8] Ioffe, S.; Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR*, ročník abs/1502.03167, 2015.

[9] Jia, Y.; Shelhamer, E.; Donahue, J.; aj.: Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[10] Li, W.; Abtahi, F.; Zhu, Z.: A Deep Feature Based Multi-kernel Learning Approach for Video Emotion Recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, New York, NY, USA: ACM, 2015, ISBN 978-1-4503-3912-4, s. 483–490.

[11] Parkhi, O. M.; Vedaldi, A.; Zisserman, A.: Deep Face Recognition. In *British Machine Vision Conference*, 2015.

[12] Pervin, L.; John, O.: *Handbook of Personality: Theory and Research*. Guilford Press, 1999, ISBN 9781572304833.

[13] Ponce-López, V.; Chen, B.; Oliu, M.; aj.: *ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results*. Cham: Springer International Publishing, 2016, ISBN 978-3-319-49409-8, s. 400–418.

[14] Rai, N.: Bi-modal regression for Apparent Personality trait Recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, s. 55–60.

[15] Simonyan, K.; Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, ročník abs/1409.1556, 2014.

[16] Subramaniam, A.; Patel, V.; Mishra, A.; aj.: *Bi-modal First Impressions Recognition Using Temporally Ordered Deep Audio and Stochastic Visual Features*. Cham: Springer International Publishing, 2016, ISBN 978-3-319-49409-8, s. 337–348.

[17] Sumner, C.; Byers, A.; Boochever, R.; aj.: Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. In *2012 11th International Conference on Machine Learning and Applications*, ročník 2, 2012, s. 386–393, doi:10.1109/ICMLA.2012.218.

[18] Tawari, A.; Trivedi, M. M.: Speech Emotion Analysis: Exploring the Role of Context. *IEEE Transactions on Multimedia*, ročník 12, č. 6, Oct 2010: s. 502–509, ISSN 1520-9210.

[19] Wei, X.; Luo, J.; Wu, J.: Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval. *CoRR*, ročník abs/1604.04994, 2016.

[20] Wood, E.; Baltrusaitis, T.; Zhang, X.; aj.: Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. *CoRR*, ročník abs/1505.05916, 2015.

[21] Zhang, C.-L.; Zhang, H.; Wei, X.-S.; aj.: *Deep Bimodal Regression for Apparent Personality Analysis*. Cham: Springer International Publishing, 2016, ISBN 978-3-319-49409-8, s. 311–324.

[22] Zhou, B.; Lapedriza, A.; Xiao, J.; aj.: Learning Deep Features for Scene Recognition using Places Database. In *Advances in Neural Information Processing Systems 27*, editace Z. Ghahramani; M. Welling; C. Cortes; N. D. Lawrence; K. Q. Weinberger, Curran Associates, Inc., 2014, s. 487–495.

Prílohy

Príloha A

Obsah priloženého CD

Priložené cd obsahuje tieto súbory:

- Dokumentacia → Zdrojové texty k dokumentácií a preložená dokumentácia
- Video → Anglická a slovenská verzia prezentačného vieda
- ZdrojoveKody → Zdrojové kódy všetkých experimentov
  - Bash → Skripty na extrakciu zvuku a snímkov
  - LinearneRegresory → Natrénované modely lineárnych regresorov a skripty k nim
  - CNN → Natrénované modely konvolučných neurónových sietí a skripty k nim