Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Automatické rozpoznání výrazu tváře s libovolným natočením v prostoru

Ing. Jana Trojanová

disertační práce

k získání akademického titulu doktor

v oboru Kybernetika

Školitel: Doc. Ing. Miloš Železný, Ph.D.

Katedra: Kybernetika

Plzeň, 2013

University of West Bohemia

Faculty of Applied Science

Automatic Facial Expression Recognition For Arbitrary Head Pose

Ing. Jana Trojanová

doctoral thesis

submitted for the degree Doctor of Philosophy

in the field of Cybernetics

Advisor: Doc. Ing. Miloš Železný, Ph.D.

Department: Cybernetics

Pilsen, 2013

Prohlášení

Prohlašuji, že jsem tuto disertační práci vypracovala samostatně, s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

V Plzni dne

Jana Trojanová

Poděkování

Tato odborná práce vznikla za odborného vedení mého školitele Doc. Ing. Miloše Železného, Ph.D. Ráda bych zde poděkovala za odborné konzultace Ing. Zdeňku Krňoulovi, Ph.D.

Dále bych chtěla poděkovat své rodině za vytvoření dobrých pracovních podmínek a kolegům z oddělení umělé inteligence katedry kybernetiky za cenné rady a pomoc při vypracovávání této práce.

Obsah

3DMM 3D Morphable Model

AAM Active Appearance Model

AU Action Units

CK Cohne Kanade

FACS Facial Action Coding System

FER Facial Expression Recognition

HCI Human-Computer-Interaction

HMM Hidden Markov Model

LBP Local Binary Pattern

PCA Principal Component Analysis

RBF Radial Basis Function

SIFT Scale-invariant feature transform

SVM Support Vector Machine

Seznam obrázků

Seznam tabulek

Kapitola 1

Úvod

Komunikace člověka a počítače (HCI - Human Computer Interaction) se stala součástí každodenního života. Pro rozhraní se nejčastěji používají zařízení jako je klávesnice, myš anebo dotyková obrazovka. Postupně se rozhraní přesouvá do nové roviny, kde je počítač schopen reagovat na víceslovné povely (např. otevři Internet Explorer) či gesta pomocí ruky. Zmíněná rozhraní poskytnou počítači jednoznačnou informaci co má provést. Ignorují však nevyslovenou informaci o emocionálním stavu člověka, která je důležitou složkou při mezilidské komunikaci. Některé emoce ovlivňují lidské jednání a jiné rozšiřují význam komunikované informace (překvapený výraz může značit nově nabytou informaci).

Informace o emocionálním stavu člověka je významným prvkem nejen pro vytvoření kvalitního HCI systému. Lze ji použít také pro psychologické studie (výraz tváře vzhledem k různé stimulaci), edukativní systémy (elektronický učitel přizpůsobující rychlost výuky dle emočního stavu studenta např. znudění vede k nepozornosti při výuce), a nebo průzkum trhu (jak lidé reagují na reklamy, nový produkt).

Chceme-li definovat systém pro rozpoznání emocí, nabízejí se dva typy úlohy, a to rozpoznání emoce z výrazu tváře (FER - Facial expression recognition) a multimodální rozpoznání emoce člověka (Multimodal Human Emotion Recognition). V případě systému FER jsou jediným vstupem obrazová data zachycující hlavu člověka. V případě druhém do systému vstupuje hned několik měření najednou (obrazová data zachycující výraz, postoj a gesta, akustická složka, psychologické signály jako data z elektroencefalogramu nebo elektrický odpor kůže). Přehled současného stavu multimodálního rozpoznávání emocí lze nalézt v [131]. Tato práce se zabývá systémy FER - vstupem systému jsou obrazová data zachycující tvář, výstupem je rozpoznaný výraz tváře.

Většina existujících FER systémů se zaměřila na rozpoznání základních emocí (radost, smutek, překvapení, strach, znechucení, hněv a neutrálního výraz). Komerčním příkladem jsou systémy 3D Facial Imaging[1] a FaceReader[2]. 3D Facial Imaging je určen pro měření emocionálních reakcí spotřebitelů na internetovou reklamu a média. FaceReader byl vytvořen pro osoby trpící autismem, ty jej mohou využívat jako pomůcku při učení rozpoznávání výrazu tváře (pozn. autisté mají problémy rozpoznat základní emoce). Mimoto je FaceReader využíván psychology k automatické anotaci dat pro studium vztahu emocí a výrazu tváře.

[1] 3D Facial imaging - http://www.nviso.ch
[2] FaceReader - www.noldus.com

Obrázek 1.1: Ukázka popisu svalové aktivity tváře pro libovolný výraz tváře systémem FACS (Facial Action Coding System). Každý elementární výraz tváře má své kódové číslo a jméno, popis FACS je v sekci 2.2.

Nevýhodou systémů pro rozpoznání základních emocí je interpretace jakéhokoliv výrazu do předem definovaných kategorií. Výraz obličeje však vyjadřuje daleko více než jen základní sadu emocí. Lidská tvář je schopná vyjádřit až 7000 různých výrazů tváře [23]. Paul Ekman, zabývající se studiem výrazu tváře již více než půl století, vytvořil seznam elementárních výrazů obličeje (FACS - Facial Action Coding System). Tyto elementární jednotky popisují svalovou aktivitu tváře. Zachycují například zvednuté obočí, našpulené rty a jiné, ukázka na obrázku 1.1. Kombinací elementárních výrazů tváře lze popsat jakýkoliv výraz tváře (více o FACS viz kapitola 2.2).

V posledních letech byly vytvořeny FER systémy schopné rozpoznat svalovou aktivitu tváře. Jejich komerční příkladem je systém Affdex[3]. Ten nejprve rozpozná elementární jednotky a ty poté interpretuje jako různé výrazy tváře a to například souhlas, koncentraci, nesouhlas, zájem, přemýšlení a nejistotu pozorované osoby, ukázka na obrázku 1.2. Affdex je používán k analýze reakce spotřebitelů na reklamy.

Navzdory úspěchu metod pro automatické rozpoznání výrazu tváře v praxi, pracují FER systémy spolehlivě pouze v kontrolovaném prostředí (čelní pohled na řečníka s minimálním natočením hlavy, konstantní neměnné pozadí, rovnoměrné osvětlení tváře bez překryvu rukou či vlasů). Úspěšnost rozpoznávání výrazu tváře je přímo úměrná kvalitě dat. Nejvýznamnějšími faktory jsou natoče-ní hlavy v prostoru a překryv tváře [115]. Tato práce se zaměřuje na rozpoznání výrazu tváře v případech, kdy subjekt pohybuje hlavou v prostoru bez omezení.

[3] Affdex - http://www.affectiva.com

Obrázek 1.2: Systém Affdex ukázka vstupních dat a rozpoznaných výrazů, převzato z webových stránek firmy Affectiva.

1.1 Definice Problému

Jeden z největších problémů při rozpoznávání výrazu tváře je libovolné natočení hlavy v prostoru [115, 100, 63]. S odklonem tváře od čelního pohledu se spolehlivost rozpoznání výrazu tváře výrazně snižuje [115]. K vyřešení tohoto problému je nutné odstranit vliv natočení tváře. Nabízejí se dvě řešení a to trénování samostatných klasifikátorů pro každé natočení [85] a nebo při rozpoznávání výrazu tváře pracovat s trojdimenzionálními (3D) daty. Vzhledem k velkému počtu natočení hlavy v prostoru a nedostupnosti dostatečného množství dat pro trénování klasifikátorů je vhodnější pracovat s 3D daty.

V současné době jsou k dispozici různé typy 3D video senzorů, nejdostupnější pro širokou veřejnost je Kinect[4]. Subjekt však musí být blízko 3D snímače a nesmí provádět rychlé pohyby. Vizuální projev emoce je velmi dynamický děj a trvá řádově stovky milisekund. Při spontánním projevu emoce subjekt často rychle mění orientaci hlavy. Data pořízená pomocí Kinect obsahují hodně šumu a v některých místech dokonce chybí (viz obrázek 1.3). Vývoj 3D technologií jde stále vpřed. Přesto v současné době nelze v reálném čase pořídit kvalitní 3D video data obsahující spontánní projev emoce [100].

Alternativní cestou jak získat 3D data z 2D obrázku je 3D deformovatelný model (3DMM - 3D Morphable Model) navržený v [7]. 3D tvář je rekonstruována na základě 2D snímku a 3DMM. Kvalitní 3DMM je nutné vytvořit z velkého počtu 3D modelů tváří s různými výrazy, aby postihl variaci (deformaci) pro velkou skupinu subjektů a emocí. Navíc je metoda rekonstrukce 3D tváře časově náročná, jeden snímek se zpracovává řádově minuty. Pro HCI systém je tedy 3DMM metoda nevhodná. V posledních letech se objevila řešení jak obejít výpočetně náročnou 3DMM úlohu (popsané jsou v kapitole 4). Metody byly navrženy pro úlohu identifikace subjektu, kde je žádoucí emoci z výrazu odstranit. Nejsou tedy schopné provést rekonstrukci tváře včetně zachování výrazu tváře. Pro rozpoznávání výrazu tváře s libovolným natočením hlavy v prostoru je nutné provést rekonstrukci tváře respektující aktuální výraz tváře.

[4] http://www.kinect.cz/

Obrázek 1.3: Ukázka pořízených dat z hrací konzole Kinect. Při rychlém pohybu hlavy část 3D dat chybí (díry) a navíc data obsahují hodně šumu. Obrázek převzat z [121]

1.2 Cíle disertační práce

Hlavním cílem disertační práce je návrh a realizace systému pro rozpoznávání výrazu tváře schopného pracovat s různým natočením tváře. Jednotlivé dílčí cíle jsou následující:

- Popis psychologického pojetí pojmu emoce se zaměřením na vztah mezi emocionálním stavem člověka a výrazem tváře. Definice faktorů nutných k rozpoznání výrazu tváře pomocí počítače.

- Analýza problematiky FER systémů a jejich současný stav. Určení a implementace algoritmů pro základní systém rozpoznávání výrazu tváře.

- Výzkum a implementace nových přístupů a metod pro 3D rekonstrukci tváře splňující požadavky kladené na FER systém schopný pracovat v reálných podmínkách (tj. s libovolným natočením tváře v prostoru).

- Experimentální ověření navrženého řešení na datech s libovolným natoče-ním tváře.

1.3 Přehled práce

Práce je strukturována následovně. Kapitola 2 obsahuje popis emoce ve vztahu k výrazu tváře a přehled existujících FER systémů. Kapitola 3 uvádí přehled stávajících metod používaných pro rozpoznávání výrazu tváře, a to předzpraco-vání dat, extrakci příznaků a jejich následnou klasifikaci. Kapitola 4 shrnuje současný stav 3D rekonstrukce tváře a shrnuje postup provedeného výzkumu nové metody pro získání 3D modelu tváře z 2D snímku vhodné pro automatickou analýzu výrazu tváře. V kapitole 5 jsou uvedeny výsledky experimentů určující vliv 3D rekonstrukce tváře na rozpoznání výrazu tváře. Získané poznatky a hlavní přínosy této disertační práce shrnuje kapitola 6. V příloze A jsou ukázky stávajících FER systémů, v příloze B je přehled databází a v příloze C jsou alternativní přístupy pro popis výrazu tváře.

Kapitola 2

Vztah mezi emocí a výrazem tváře

Při každodenní komunikaci s lidmi používáme mimo mluveného slova i výraz naší tváře. Výrazem vyjadřujeme náš vnitřní -emocionální- stav. Stejná informace vyslovena s různým výrazem ve tváři může mít naprosto jiný význam. Uvedu několik příkladů. Pokud vám někdo řekne 'zlatíčko' a bude se u toho tvářit rozlobeně, budete se cítit nepříjemně. Odpoví-li žák učiteli na otázku „můžeme pokračovat?" - „no jasně" se znuděným výrazem, může na to učitel reagovat zpomalením výuky či změnou tématu. Takovýchto příkladů se najde v komunikaci velké množství. Chceme-li efektivně komunikovat s počítačem, je potřeba, aby byl schopen rozpoznat náš výraz, potažmo naše aktuální rozpoložení.

Již v sedmnáctém století se objevily psychologické studie zabývající se rozpoznáváním emoce z výrazu obličeje. V sekci 2.1 jsou shrnuty poznatky o propojení výrazu tváře s emočním stavem člověka. Způsob měření výrazu tváře je popsán v sekci 2.2. Technické obory (respektive automatické zpracování obrazu) se rozpoznáváním emocí začaly zabývat až v 70. let minulého století. Většina existujících systémů pro rozpoznání výrazu tváře (dále jen FER systém) je schopna rozpoznat šest základních emocí (radost, smutek, překvapení, strach, znechucení a hněv). Mimo nich bylo vytvořeno několik systémů, zabývajících se mentálním stavem subjektu (souhlas, zájem, přemýšlení, koncentrace), nebo systémy, které namísto emocí rozpoznají aktivitu obličejových svalů (např. zvednuté koutky úst). Přehled a srovnání existujících FER systémů je uveden v sekci 2.3. Požadavky na systém rozpoznávající výraz tváře jsou shrnuty v sekci 2.4.

2.1 Psychologické pojetí emoce a výrazu tváře

Pojem emoce je významově ztotožňován s pojmem cit a označuje se jím prožívání stavů jako jsou radost, smutek, hněv, lítost, strach atd. Psychologické vymezení pojmu emoce je však obtížné a jeho definovávání přímo nemožné. Nejednotný přístup k tématu emoce vedl na stovky různých definic [78]. Namísto definice uznává většina psychologů základní množinu složek (komponent) ovlivňujících vznik a projev emoce[1]:

- **Subjektivní/Pocitová komponenta.** Tato komponenta se většině lidí zdá být hlavní komponentou při vnímání emocí. Lidé obecně souhlasí s názorem, že stav mysli je odlišný porovnáme-li prožívání hněvu a radosti. Subjektivní komponenta je nejvíce kontroverzní

[1] http://www.paulekman.com/

a existuje velké množství nezodpovězených otázek, jako např. „ Je kvalita pocitu určité emoce stejná mezi lidmi? Co je podstatou rozdílnosti povahy mezi emocemi? Jakou důležitost nebo funkci mají tyto pocity?“

- **Komponenta vnějšího vyjádření**. Tato komponenta je pozorována ve spojení s emocí, a je vytvořená díky příčně pruhovaným svalům. Rozděluje se na dva obecné typy: reakce těla (např. poklepávání nohou při nervozitě) a výraz tváře (např. úsměv při prožívání radosti).

- **Fyziologická komponenta** je relativně skrytá. Popisuje změny uvnitř těla způsobené hladkým svalstvem (např. tlukot srdce), žlázami (různé chemické reakce vedoucí k pocení, zčervenání tváří) a nervovým systémem (mozková aktivita).

- **Kognitivní komponenta** slouží k vyhodnocení situace. Skládá se z před-stav a myšlenek, které nám přijdou na mysl během prožívání emoce. Může vést ke vzniku emoce (např. vzpomínka na zemřelou osobu vyvolá smutek) a zároveň se učíme jak reagovat na novou situaci (strach jedince připraví na nebezpečí a vtiskne do paměti situaci jako nebezpečnou). Tato komponenta je těžko změřitelná a stále není jasné, jak zapadá do celkového obrazu emoce.

- **Komponenta chování.** Některé události vyvolávají stejné emoce v rů-zných kulturách (smrt vlastního dítěte vyvolá smutek). Jiné věci, třeba reakce na určitý druh jídla, se můžou v závislosti na kultuře hodně odlišovat.

Psychologové studují nejen jednotlivé komponenty, ale i vztahy mezi nimi, respektive příčiny a funkce jednotlivých emocionálních reakcí. Tato práce se zamě-řuje pouze na komponentu vnějšího vyjádření a to na rozpoznání výrazu tváře.

Nejčastěji je v souvislosti s komponentou vnějšího vyjádření, respektive vzta-hu mezi emocí a výrazem tváře zmiňovaná práce od Charlese Darwina z roku 1872 The Expression of the Emotions in Man and Animals. Darwin argumentoval, že **určité emocionální výrazy tváře jsou vrozené a stejné pro všechny lidi**. Darwinova teorie byla po dlouhá léta zpochybňována a psychologové nebrali výraz tváře jako indikátor emoce. Změna nastala až v polovině dvacátého století, kdy bylo vytvořeno několik studii zabývajících se emocemi a výrazem v obličeji [44], [25]. Studie poukázaly na chybnou metodologii vedoucí k zamítnutí vztahu mezi emocí a výrazem. Dále experimentálně ověřili schopnost pozorovatele označit hraný i spontánní výraz tváře pomocí emočních kategorií (smutek, radost, aj.). Pozorovatelem označené hrané emoce se shodovaly s tím, co měl předvádějící v úmyslu ukázat. Pro spontánní emoce vybral pozorovatel emoci příslušející situaci, která emoci vyvolala. Studie jasně potvrdily existenci vztahu mezi emocí a výrazem tváře, tedy že **výraz tváře reprezentuje emoční stav člověka**.

Další důležitou otázkou ve vztahu emoce a výrazu tváře je jejich univerzálnost pro různé kultury. Studie [26] provedená Paulem Ekmanem pro pět gramotných a dvě negramotné kultury potvrdila existenci obecného emocionálního výrazu tváře. Lidé různých kultur označily shodně výrazy ve tváři, které jim byli ukázány. Ekman tak dokázal, že existuje základní množina emocí (radost, smutek, vztek, strach, znechucení a překvapení), která je stejná pro všechny lidi na světě. Jiná studie [22] se zaměřila na výraz tváře u slepých dětí. Tyto děti se výraz ve tváři nemohly naučit pozorováním okolí, přesto se u nich objevoval stejný výraz ve tváři jako

Tabulka 2.1: Přehled kategorií emocí definovaných v rámci projektu HUMAINE.

Negativní & prudký	**Pozitivní & živý**
Zlost, Mrzutost, Opovržení, Znechucení, Podrážděnost	Veselost, Euforie, Vzrušení, Radost, Blaho, Příjemnost požitku
Negativní & bez kontroly	**Laskavý**
Úzkost, Rozpaky, Strach, Bezradnost, Bezmocnost, Starostlivost	Náklonnost, Rozhodnost, Přátelskost, Zamilovanost
Negativní myšlenky	**Pozitivní myšlenky**
Pochyby, Závist, Frustrace, Vina, Zahanbení	Odvaha, Naděje, Víra, Uspokojení, Důvěra
Negativní & pasivní	**Klidný pozitivní**
Znudění, Zoufalství, Zklamání, Zarmoucení, Smutek	Sebejistota, Spokojenost, Uvolněnost, Osvobození, Vyrovnanost
Agitace (Rozruch)	**Reaktivnost**
Šok, Stres, Napětí	Zájem, Zdvořilost, Překvapení

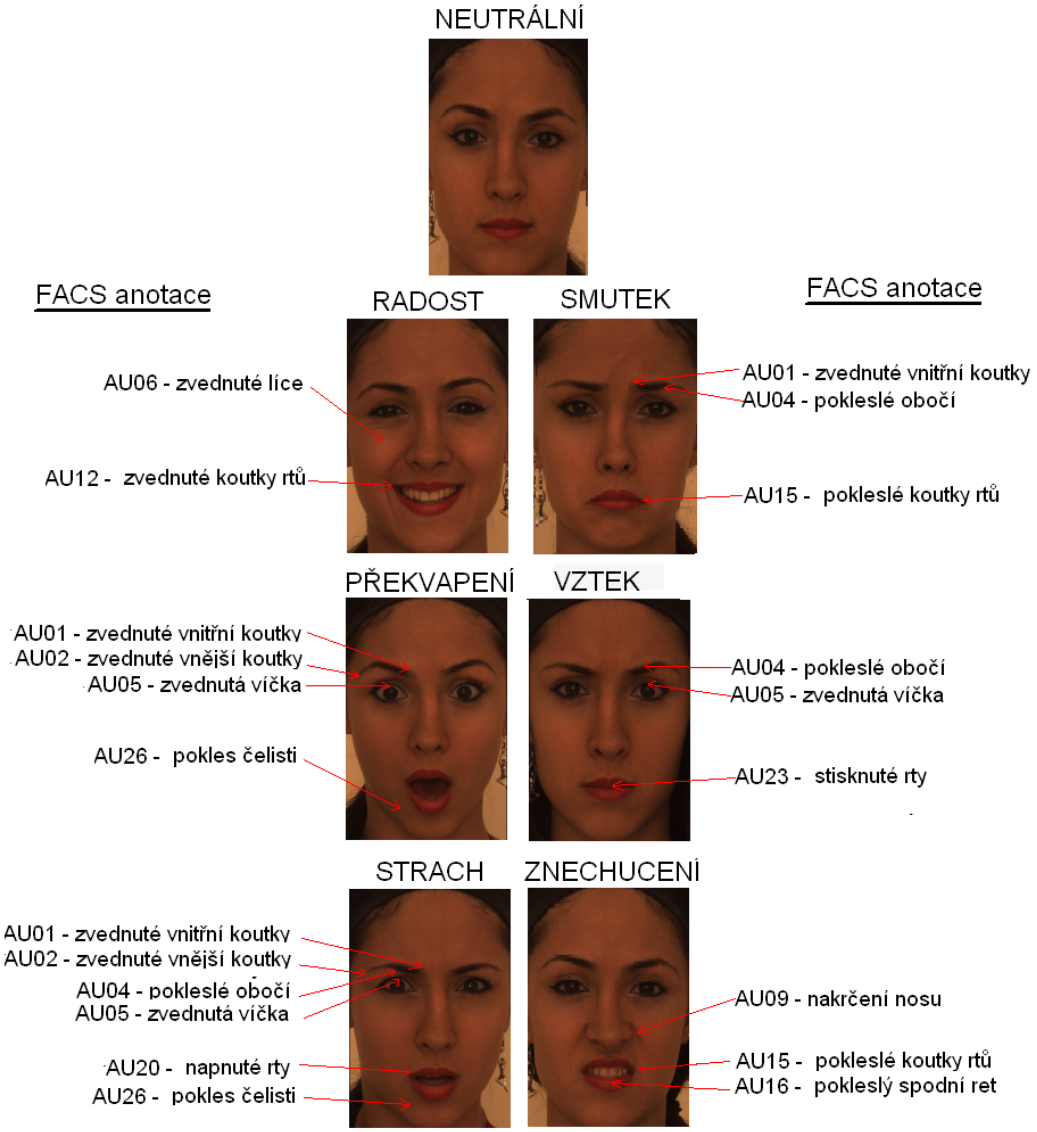
u dětí které vidí. Tyto studie přispěli k podpoření Darwinovi hypotézu, že určité výrazy tváře jsou vrozené a stejné pro všechny lidi.

2.2 Měření výrazu tváře v psychologii

Mezi nejčastěji používanou reprezentaci emoce patří dělení do kategorií (strach, radost, aj., viz obrázek 2.1), navrženou Robertem Woodworthem v knize Experimental Psychology (informace převzata z knihy [44]). Tato reprezentace je velmi intuitivní, neboť ji lidé používají k popisu emoce v každodenním životě. Většina dostupných databází je anotována právě pomocí kategorického dělení. Anotace dat je velmi snadná, neboť při pořizování dat je subjekt požádán, aby předvedl konkrétní emoci. Případně jsou nahrávány reakce na stimul vyvolávající určitou emoci (např. subjektu je promítáno vtipné video). V rámci projektu HUMAINE vytvořil tým Dougles-Cowie, Cox et al. rozšířený seznam 48 kategorií, uvedený v tabulce 2.1. Kompletní seznam kategorií schopný popsat všechny možné výrazy tváře vyskytující se v přirozeném prostředí neexistuje.

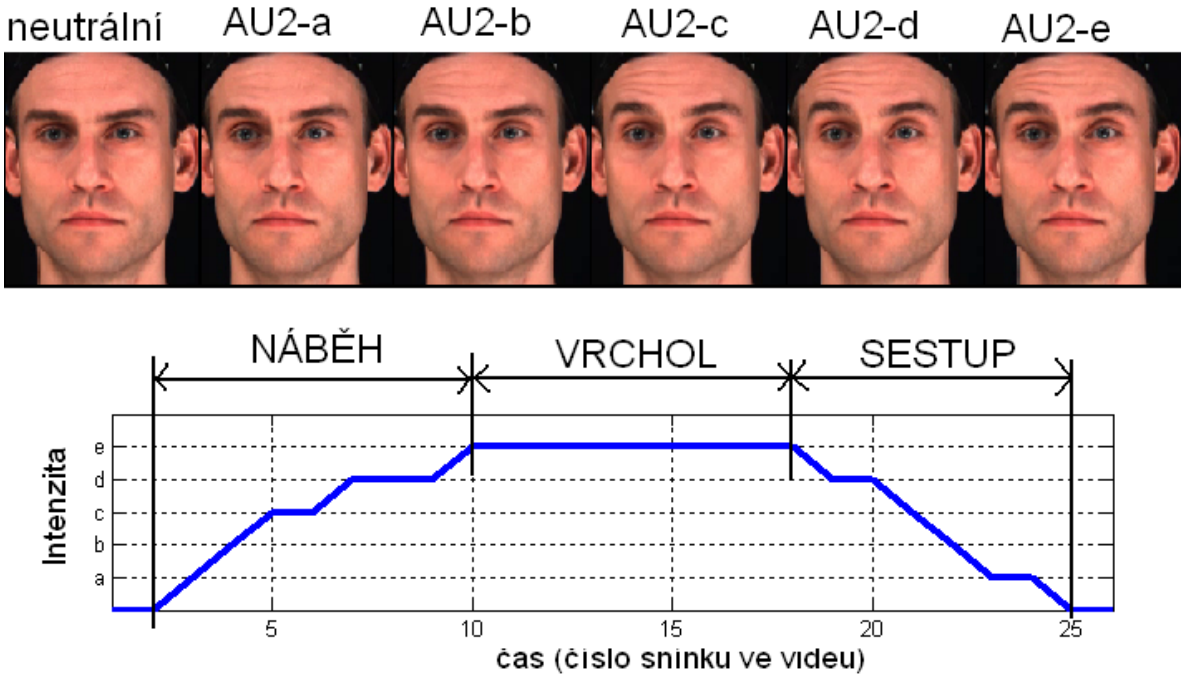
Další velmi používanou reprezentací je systémem Facial Action Coding System (FACS). Oproti kategorickému dělení slouží FACS k měření a popisu libovolného výrazu tváře. Důraz je kladen na měření svalové aktivity a úsudek o významu výrazu (tj. konkrétní emoce) je při popisu ignorován. Paul Ekman a W.V.Friesen vytvořili základní systém v 70-tých letech studiem vztahů mezi svaly v obličeji a výrazem tváře. FACS popisuje změnu výrazu tváře pomocí změny pohybu kůže (např. zvrásnění, vyboulení způsobené nafouknutím tváře), a dočasné změny ve velikosti a pozici klíčových bodů (např. zvednuté koutky úst). FACS neměří neviditelné změny jako je svalové napětí nebo cévní a žlázové změny (pot, slzy) a také neměří

trvalé obličejové charakteristiky (různá znamínka či deformace tváře). Díky zaměření systému FACS na změnu pohybu svalů vznikl popis nezávislý na rozdílech mezi subjekty (rozdíly jako velikost, tvar a umístěni klíčových bodů, nebo trvalé vrásky).



Obrázek 2.1: Kategorické dělení emocí a jejich anotace v systému FACS.

Při anotaci dat pomocí FACS je analyzována změna pohybu svalu. K popisu výrazu jsou definovány tzv. akční jednotky (Action Units, dále AUs), ukázka na obrázku 2.1. Tyto jednotky jsou přímo spojeny se svalem, který je způsobuje. Interpretace jednotlivých AU jednotek není závislá na významu výrazu tváře, k rozhodování o významu výrazu tváře je nutné vytvořit rozhodovací systém. Pro základní emoce existuje slovník Facial Action Coding System Affect



Obrázek 2.2: Anotace AU2-zvednuté vnější kouty obočí v čase. V horní části jsou ukázky pro jednotlivé intenzity AU jednotky, v dolní části je zobrazen typický vývoj AU jednotky v čase.

Interpretation Dictionary (FACSAID), který definuje z jakých AU jednotek se skládá emoce. Pro další výrazy tváře jako např. (ne)souhlas, zmatení, znudění, bolest si uživatel může kombinaci AU jednotek definovat sám. Pro každou AU jednotku je ve FACS manuálu [93] poskytnuta následující informace:

- Svalový základ pro každou AU jednotku
- Detailní popis vzhledu tváře doplněný o fotografie a video ukázky
- Instrukce jak provést AU jednotku na vlastním obličeji
- Kritéria k ocenění intenzity AU jednotky. Intenzita je hodnocena od A (nejmenší intenzita) až do E (největší intenzita)

Celkem existuje 56 jednotek, rozděleny jsou do tří skupin a to horní obličejové pohyby, dolní obličejové pohyby a akce popisující pohyb hlavy a očního okolí. Každá AU jednotka je identifikována číslem a jménem (AU12-zvednuté koutky úst). Jméno poskytuje smysluplnou pomůcku oproti nicneříkajícímu číslu.

Výraz tváře (intenzita AU jednotky) se mění v čase, ukázka na obrázku 2.2. Při anotaci dat je přesnost závislá na tom, co je anotátor schopen vidět když si opakovaně přehrává video data. Každá AU jednotka lze rozdělit do tří komponent a to náběh, vrchol a sestup. Tyto komponenty vyjadřují úroveň intenzity AU jednotky v čase. Vývoj intenzity AU jednotky v čase se používá k rozlišení hraného a spontánního výrazu tváře. Většina databází obsahuje

pouze anotaci určující AU jednotku. Databáze obsahující informaci o intenzitě nejsou veřejně dostupné.

Kategorické dělení emocí a systém FACS nejsou jediné možné způsoby reprezentace výrazu tváře, alternativní systémy jsou popsány v příloze C.

2.3 Přehled FER systémů

Výzkum v oblasti rozpoznávání výrazu tváře se v současnosti dělí na dva hlavní směry a to rozpoznání kategorických emocí a rozpoznání svalové aktivity tváře, respektive rozpoznání AU jednotky systému FACS. Tyto dva směry vznikly na základě nejpoužívanějších reprezentaci při měření výrazu tváře v psychologii. Podrobně jsou popsané v předešlé sekci. Většina existujících FER systémů se zaměřila na kategorické rozpoznání emocí. Nevýhodou těchto systémů je zařazení jakéhokoliv výrazu do předem definovaných kategorii. Proto se vývoj FER systémů v posledních letech zaměřil na rozpoznání svalové aktivity. AU jednotky výrazně snižují dimenzi klasifikace výrazu tváře, neboť s pomocí jen 32 AU jednotek lze popsat tisíce různých výrazů tváře [91].

2.3.1 FER systémy pro rozpoznání kategorických emocí

První studie na rozpoznávání emocí se objevily již v 70. letech [48, 107, 99]. V průběhu 90. let se objevilo několik dalších studií jejichž přehled lze nalézt v [86], [30], [88]. Všechny tyto systémy rozpoznávaly pouze základní emoce. Používané databáze, nedostupné pro širokou veřejnost, obsahovaly malý počet subjektů, emoce byla hraná (subjekt byl vyzván, aby předvedl emoci) a zachycena pouze na jediném statickém snímku. Většina systémů, které vznikly na začátku tisíciletí, se nadále zabývala rozpoznáním základních emocí, změna oproti 90 letům byla v rozpoznání emoce z video dat (např. [62],[39],[102], [130], [114]). Od roku 2004 se objevují systémy schopné rozpoznat jiné než základní emoce. Příkladem je MindReader od Kaliouby [28] pro rozpoznávání mentálního stavu subjektu, Kapoor05 [51] pro rozpoznávání zájmu a nezájmu u dětí při hraní her, a Kapoor07 [50] rozpoznávající frustraci. Systémy schopné rozeznat bolest od emoce [1], [64], a Dmello08 [19] schopný rozpoznat znudění, zaujetí, zmatení a frustraci.

Rozpoznání základních emoci je považováno za vyřešenou úlohu pokud jsou data pořízena v kontrolovaných podmínkách (čelní pohled na subjekt, který předvádí v jeden okamžik jedinou emoci ohraničenou neutrálním výrazem). Aktuální výzvou pro rozpoznávání kategorických emocí jsou problémy související s přechodem na reálná data. Nevyřešené problémy jsou libovolné natočení hlavy a rozpoznání překrývajících se emocí (např. přechod od překvapení k radosti). Částečně vyřešené problémy jsou překryv části tváře a sledování klíčových bodů tváře pro nové subjekty [36, 115].

2.3.2 FER systémy pro rozpoznání svalové aktivity

Na začátku tisíciletí se začaly objevovat studie zabývající se rozpoznáním aktivity obličejových svalů, založených na FACS. Zpočátku systémy rozpoznávaly pouze izolované AU jednotky s maximální intenzitou. ([52], [6], [87], [112], [11], [12], [72], [90], [89], [84], [73], [117],

[124], [85], [66], [56]). Postupně se FER systémy zaměřily na rozpoznávání jednotlivých komponent AU jednotek (náběh, vrchol, sestup). Což vedlo na systémy schopné rozlišit hraný od spontánního výrazu tváře [117], [118], [6], [76], [14]. Existuje relativně málo studií zabývajících se kombinací výrazu tváře s dalšími vizuálními vstupy jako je např. pohyb hlavy [14, 134], nebo gesta [118, 35].

V současné době jsou výzvy pro FER systémy rozpoznávající svalovou aktivitu stejné jako pro kategorické dělení (libovolné natočení hlavy, rozpoznání překrývajících se AU jednotek, překryv tváře a sledování klíčových bodů). Navíc jsou zde problémy s rozpoznáním intenzity AU jednotky a trvání jednotlivých komponent, které jsou stále považovány za částečně vyřešený problém [36, 115].

2.4 Shrnutí

Výzkum v oblasti rozpoznávání výrazu tváře urazil za posledních dvacet let velký kus cesty. Současné systémy jsou plně automatické a jsou schopné rozpoznávat spontánní emoce či mentální stav člověka (znudění, zájem, přemýšlení, nesouhlas, rozpaky, aj.). Velkou zásluhu na tom má dostupnost rozsáhlých databází, nově vytvořené metody pro detekci hlavy, popis tváře a bližší spolupráce s psychology. I přesto stále existují problémy, které nejsou kompletně vyřešeny. Ideální systém pro rozpoznávání výrazu tváře by měl splňovat následující (nevyřešené problémy jsou označené *):

- Systém by měl být **plně automatický**. Což znamená, že není potřeba žádná manuální práce (např. vyznačení pozice hlavy, obočí a úst). Měl by **fungovat v reálném čase** (tzn. reakce systému by neměla mít znatelné zpoždění rozpoznatelné uživatelem).
- Systém by měl být **nezávislý na uživateli**, respektive na pohlaví, etnice (Evropan, Afričan, Asiat) a na věku.
- Systém by měl být schopen rozpoznat **výraz ovlivněný promluvou**. Neměl by se omezovat jen na případy, kdy člověk nemluví.
- Systém by měl **pracovat i s malým rozlišením**. **Neměl by být závislý na změně prostředí** (tzn. neměl by vyžadovat konstantní pozadí).
- * Systém by měl být schopný **rozpoznat přechod od jednoho výrazu k druhému**.
- * Systém by měl být schopný **pracovat při různém osvětlení**. Neměl by vyžadovat konstantně osvětlenou tvář.
- * Systém by měl umožňovat **pracovat s různým natočením hlavy** (ne jen čelní pohled či pohled z profilu).
- *Systém by by si měl **poradit s překryvem obličeje** (ruka přes obličej, brýle, vlasy v obličeji).

Kapitola 3

Stávající metody pro rozpoznání výrazu tváře

Variabilita lidské tváře činí z úlohy rozpoznání výrazu tváře velmi složitý pro-blém. Lidské tváře se mezi sebou odlišují tvarem a vzhledem. Navíc každá tvář sama o sobě vytváří další variace [133], vzhledem k příčině lze tyto variace rozdělit do čtyř skupin

- Změna **orientace tváře v prostoru** mění vstupní data do systému. Data pak zobrazují snímky pro různé natočení tváře. Při velkém natočení dochází až k zákrytu různých částí tváře.
- Změna **osvětlení** ovlivňuje vhled tváře i v případech stejně natočené tváře. Umístění a distribuce světla na tváři mění intensitu snímku, umístění stínu a odlesky. Navíc vytvořené stíny vytvářejí výrazné kontury na tváři.
- Změna **pozice** většiny **klíčových bodů** způsobená změnou výrazu tváře. Body mění svojí pozici díky pohybu čelisti nebo svalům pro pohyb obočí, rtů a tváří. Existuje však několik klíčových bodů, které nejsou ovlivněné výrazem tváře. Tyto body jsou přímo navázané na strukturu lebky jedná se například o vzdálenost mezi očima.
- **Dlouhodobá změna tváře** způsobená v průběhu času. Například díky stárnutí, nalíčení tváře, vousům, náušnicím či změně účesu.

Žádnou ze zmíněných variací nelze považovat za konstantní přes větší množinu snímků. Je tedy nutné aby systém pro rozpoznávání výrazu tváře uměl rozlišit a popsat jednotlivé variace.

Tato kapitola popisuje metody pro stávající systémy na rozpoznávání výrazu tváře (ať kategorické emoce či AU jednotky). FER systém se typicky skládá ze tří částí: předzpracování dat, extrakce příznaků a klasifikace výrazu. **Předzpraco-vání** zajišťuje konzistentní data pro zbývající dvě části. Obsahuje metody pro nalezení tváře případně nalezení pozice klíčových bodů, dále provádí normalizaci tváře na stejnou velikost, také může obsahovat metody pro odstranění vlivu osvětlení, čí natočení tváře. Přehled používaných metod pro předzpracování dat je v sekci 3.1. **Extrakce příznaků** vytváří popis tváře. Cílem je zvýraznění rysů tváře pro různé výrazy nezávislé na osobě. Dle zvolené metody pro extrakci příznaků se FER systémy dají dělit na systémy používající geometrické příznaky, systémy používající vizuální příznaky a hybridní systémy používající kombinaci obou. Přehled používaných metod je v sekci 3.2.

Poslední částí je **klasifikace výrazu** tváře. Vzhledem k tomu jak klasifikátor pracuje s časovou informací dělí se systémy na statické (pracující s informaci v aktuálním snímku) a dynamické (integrující informaci přes sekvenci snímků). Přehled klasifikátorů je v sekci 3.3. Srovnání úspěšnosti existujících systému je uvedeno v sekci 3.4. Shrnutí výhod a nevýhod jednotlivých metod pro analýzu výrazu tváře je v sekci 3.5.

3.1 Předzpracování dat

Předzpracování dat je velmi důležitá část FER systému. Přesná informace o aktuální pozici, velikosti a orientaci tváře je zásadní pro rozpoznání výrazu tváře. Předzpracování dat se může skládat z následujících kroků (ne nutně v tomto pořadí): nalezení tváře ve snímku, odstranění vlivu osvětlení, nalezení klíčových bodů tváře, určení natočení tváře a zarovnání tváře. První krok, nalezení tváře, obsahuje každý FER systém, bez něj nelze provést extrakci příznaků sloužící k popisu tváře. Pro dosažení vyšší korelace mezi extrahovanými příznaky a aktuálním výrazem tváře je vhodné provést i další kroky předzpracování. Díky těmto krokům lze odstranit variace tváře způsobené natočením tváře nebo proměnlivým osvětlením, atd. Každý krok předzpracování by vydal na téma samostatné disertační práce, následující text obsahuje pouze krátký úvod pro jednotlivé kroky předzpracování dat.

3.1.1 Nalezení tváře ve snímku

Nalezení tváře je prvním krokem při rozpoznání výrazu tváře. Existují dva přístupy k nalezení tváře: metody založené na analýze pixelů a metody založené na apriorní informaci o hledaném objektu [74]. V případě analýzy pixelů je pro každý pixel na základě modelu rozhodnuto zda jde o hledaný objekt nebo ne (model často představuje referenční snímek bez hledaného objektu). Výstupem je silueta objektu. V případě druhém se prohledává celý snímek pro různě velké výřezy. Pro každý výřez je rozhodnuto zda je objekt obsažen nebo ne. Výstupem je obdélník označující místo a velikost hledaného objektu. Většina FER systémů používá přístup založený na apriorní informaci o hledaném objektu, nejčastěji detektor tváře od Viola & Jones [119]. Pro zvýšení úspěšnosti nalezení tváře ve snímku lze doplnit detektor o sledovač tváře (face tracking) [135]

3.1.2 Odstranění vlivu osvětlení

Nerovnoměrné osvětlení tváře může způsobit velké problémy při hledání klíčových bodů a ovlivní tak všechny následující kroky předzpracování dat. V případech kdy je tvář nerovnoměrně osvětlena lze použít speciální filtry odstraňující proměn-livé osvětlení tváře. Obrázek 3.1 zobrazuje ukázku odstranění nerovnoměrného osvětlení tváře (způsobující lokální stíny a přesvícení určitých části tváře) v úloze rozpoznávání identity tváře[109]. Dostupné databáze pro rozpoznávání výrazu tváře jsou pořízeny v kontrolovaných podmínkách. Osvětlení je rovnoměrně rozložené a vhodnou volbou popisu tváře (např. Lokální binární vzory [79]) je problém rozdílné intenzity světla vyřešen.

Obrázek 3.1: Odstranění nerovnoměrného osvětlení tváře, převzato z [109]

3.1.3 Nalezení pozice klíčových bodů

Tvar objektu je nepostradatelná informace díky které lze snadno charakterizovat a klasifikovat typ objektu. Pro rozpoznávání výrazu tváře lze tvar jednotlivých částí tváře použít přímo k vytvoření popisu (viz geometrické příznaky 3.2.1). Z tvaru tváře lze odhadnout natočení tváře v prostoru. Je tedy vhodné nejprve získat tvar tváře a porozumět variacím kterých tvář může nabývat dříve než začneme pracovat z texturou.

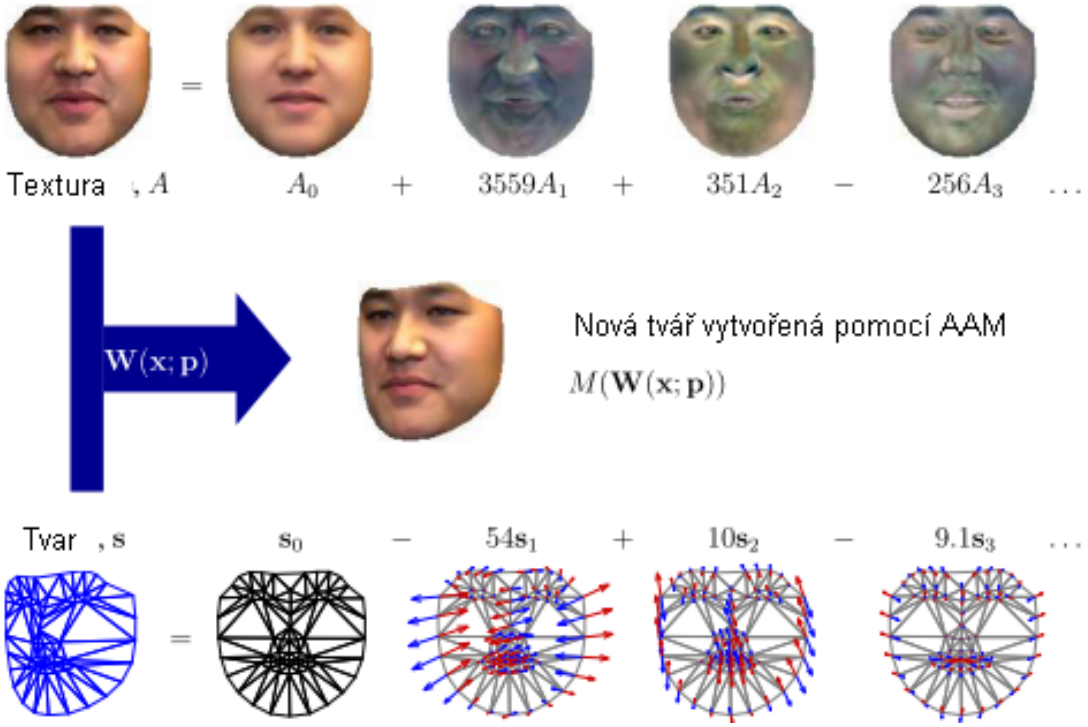
Tvar tváře lze reprezentovat pomocí klíčových bodů. K nalezení pozice klíčových bodů se ve FER systémech nejčastěji používá aktivní vzhledový model (AAM - active appearance model) [66, 56, 113]. Další metody pro nalezení klíčových bodů jsou například Piecewise Bézier Volume Deformation [12, 102], Constrained Local Models [9], nebo Partical Filter [117].

Aktivní vzhledový model (AAM - Active Appearance Model) Na obrázku 3.2 je ukázka aktivního tvarového modelu pro tvář. Model se skládá z informace o tvaru a textuře tváře. Tvar je definován pomocí bodů spojených do 2D sítě trojúhelníků. Textura je jas jednotlivých bodů uvnitř trojúhelníkové sítě.

Při trénování jsou k dispozici manuálně anotované snímky tváře obsahující požadovaný počet klíčových bodů. Nejprve je nutné odstranit z trénovacích dat variaci způsobenou rozdílnou velikostí a natočením. Poté lze spočítat střední tvar modelu a pomocí metody PCA se vypočte transformační matice a vektor parametrů definující povolenou deformaci modelu. Model pro tvar je definován

$$s = s_0 + \sum_{i=1}^{m} p_i s_i \tag{3.1}$$

kde s_0 je střední tvar modelu, koeficienty $p = (p_1, ..p_m)^T$ reprezentují tvarové deformační parametry naučené při trénování a s_i reprezentuje vektory transformační matice získané pomocí PCA. Model pro texturu je definován



Obrázek 3.2: Ukázka vytvoření nové tváře z AAM modelu. Horní řada zobrazuje model textury, dolní řada zobrazuje model tvaru. Uprostřed je ukázka nové tváře získané změnou parametrů tvaru p a parametrů textury λ. Převzato z [70]

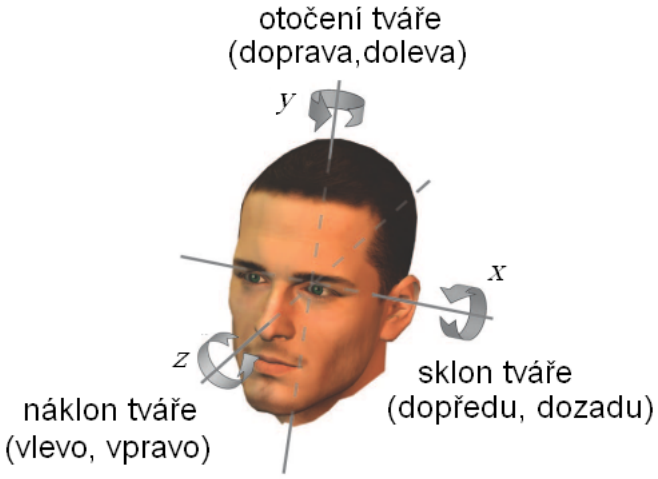
$$A(u) = A_0(u) + \sum_{i=1}^{l} \lambda_i A_i(u) \tag{3.2}$$

kde $u = (x, y)^T$ představuje pozici pixelů uvnitř 2D sítě trojúhelníků, $A_0(u)$ je střední textura modelu, koeficienty $\lambda = (\lambda_1, ..\lambda_l)^T$ reprezentují parametry pro povolenou změnu textury a $A_i(u)$ jsou vektory transformační matice získané pomocí PCA.

Rovnice 3.1 a 3.2 definují strukturu AAM modelu, nepopisují však jak vytvořit nové tváře z AAM modelu. K tomuto účelu se používá afinní transformace $W(u, p)$. AAM model je upraven pro 2D snímek M tak, že pro každý pixel u ve středním tvaru s_0 je spočtena jeho nová pozice pomocí afinní transformace

$$M(W(u; p)) = A(u) \tag{3.3}$$

V případě nasazení modelu na nový snímek I chceme minimalizovat chybu mezi $I(u)$ a $M(W(u; p)) = A(u)$. Chyba může být počítána buď v souřadnicích nového snímku nebo v souřadnicích AAM modelu. Pro zvýšení efektivity algoritmu je lepší chybu počítat v souřadnicích AAM, tzn. pro střední tvar s_0. Pokud je tedy pixel u uvnitř tvaru s_0, potom korespondující pixel ve vstupním obrázku I je na pozici $W(u, p)$. Pokud je jas na pozici u definován rovnicí 3.2, potom je jas v novém snímku definován jako$I(W(u, p))$. Cílem tedy je minimalizace následující chyby



Obrázek 3.3: Tři stupně volnosti orientace tváře v prostoru, převzato z [77]

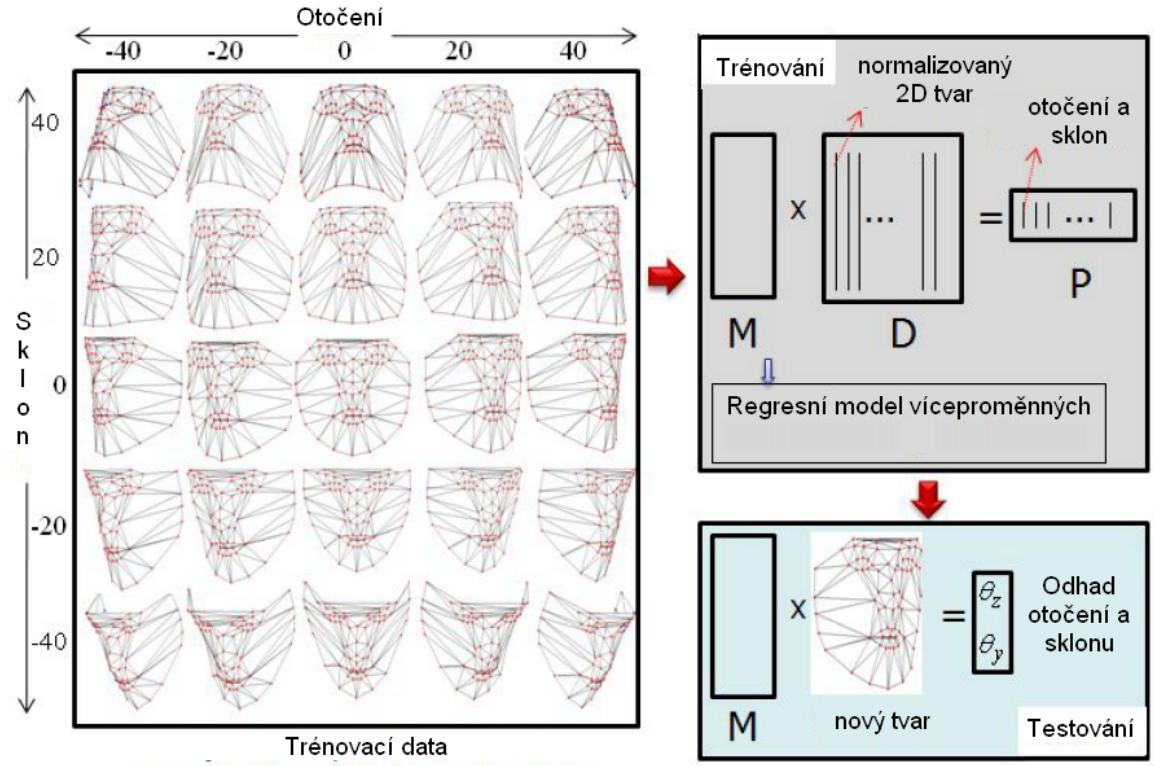
$$\sum_{u \epsilon s_0} \left[A_0(u) + \sum_{i=1}^{l} \lambda_i A_i(u) - I(W(u,p)) \right]^2 \tag{3.4}$$

respektující povolenou deformaci tvaru p a povolenou změnu textury λ. Jinými slovy při hledání AAM modelu je vstupní snímek zpětně transformovaný na střední tvar s_0 tak, aby chyba definovaná v rovnici 3.4 byla minimální. Porovnání úspěšnosti různých algoritmů pro usazení AAM modelu na nový snímek je popsáno v [2].

3.1.4 Určení (Odhad) natočení tváře v prostoru

Určení natočení tváře v prostoru znamená v počítačovém vidění schopnost určit relativní orientaci tváře vůči kameře. Obrázek 3.3 zobrazuje tři stupně volnosti orientace lidské tváře a to náklon, sklon a otočení. Náklon představuje natočení tváře kolem z-ové osy, sklon je natočení kolem x-ové osy a otočení je vzhledem k y-ové ose. Aktuální snímek může být zařazen buď do jedné diskrétní kategorie orientace tváře (při trénovaní se data rozdělí na konečný počet tříd např. tvář je natočená dolů, vpravo, čelně, vlevo, nahoru) a nebo je určen kontinuální odhad natočení tváře vzhledem k osám x, y a z. Pro úlohu rozpoznání výrazu tváře přináší orientace tváře další informaci. Kývnutí hlavou s neutrálním výrazem ve tváři může znamenat souhlas.

Metody pro odhad natočení tváře lze dělit podle různých kategorií. Jednou možností je dělení vzhledem k použitému vstupu na vizuální přístup a geometrický přístup [37]. V případě vizuálního přístupu je nutné na vstupní data použit regresní model, např. neuronové sítě [96]. V případě geometrického přístupu se používají buď metody popisující vztah mezi klíčovými body [32] nebo lze použít lineární regresní model více proměnných [37] popsaný v dalším odstavci. Rozsáhlý přehled metod pro určení orientace tváře v prostoru lze nalézt v [77].



Obrázek 3.4: Odhad natočení tváře v prostoru, převzato z [37].

Lineární regresní model více proměnných pro odhad natočení tváře Máme-li k dispozici přesně nalezené klíčové body, je pro odhad natočení tváře vhodné použít metodu popsanou v [37]. Metoda dosahuje vysoké přesnosti odhadu natočení tváře v prostoru. Pro rozsah sklonu a otočení v rozmezí -40 až 40 stupňů je chyba odhadu pro otočení 5 stupňů a pro sklon 2 stupně.

Metoda využívá 3D model tváře a jeho projekci do 2D prostoru. Lineární regresní model je použitý k naučení vztahu mezi natočením tváře ve 3D prostoru (sklon a otočení) a promítnutým 2D tvarem tváře (náklon je odhadnutý na základě pozice středu očí). Celý postup je zobrazen na obrázku 3.4, rozdělen je do dvou kroků: trénování a testování regresního modelu.

V kroku trénování se regresní model učí vztah mezi natočením tvaru tváře R_{Θ_x} (otočení) a R_{Θ_y} (sklon) ve 3D prostoru a jeho projekcí ve 2D. Tento vztah je definován následovně

$$S_{2dn} = PR_{\Theta_x}R_{\Theta_y}S_{3dn} \tag{3.5}$$

kde $s_{2di} = (x, y)^T$ reprezentuje 2D projekci 3D tvaru $S_{3di} = (x, y, z)^T$, P je projekční matice z 3D prostoru do 2D prostoru definována jako

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{3.6}$$

Trénovací data jsou normalizována tak aby pro každé natočení měla projekce 3D tvaru do 2D prostoru stejnou velikost a střed v počátku.

$$f_{\Theta_x\Theta_y} = (s_{1d2n}(\Theta_x, \Theta_y) - \overline{s_{1d2n}(\Theta_x, \Theta_y)}) / \left\| s_{1d2n}(\Theta_x, \Theta_y) \right\|^2 \tag{3.7}$$

kde s_{1d2n} je vektorová verze tvaru s_{2dn}. Vztah mezi natočením a 2D projekcí je pak definován následovně

$$M_{\Theta_x,\Theta_y} = R_{\Theta_x,\Theta_y}(F_{\Theta_x,\Theta_y} F^T_{\Theta_x,\Theta_y} + \varepsilon I)^{-1} \tag{3.8}$$

kde M_{Θ_x,Θ_y} představuje regresní model dvou proměnných, εje regularizační parametr a I je jednotková matice.

V kroku testování se odhad natočení z 2D tvaru určí jako

$$(\Theta_x, \Theta_y)^T = M_{\Theta_x,\Theta_y} s'_{1d2n} \tag{3.9}$$

kde s'_{1d2n} je normalizovaný 2D tvar dle rovnice 3.7.

3.1.5 Zarovnání tváře

Hlavním cílem zarovnání tváře je normalizace tváře na stejnou velikost a odstranění vlivu natočení tváře v prostoru. Nejrychlejším řešením je použít výstup z Viola&Jones detektoru tváře [119]. Výstupní obdélník je vždy normalizován na stejnou velikost tváře. Nevýhodou je proměnlivá velikosti a pozice detekovaného obdélníku, i pro skoro statickou tvář je standardní odchylka výstupu detektoru okolo 5% [115]. Aby následné rozpoznávání výrazu tváře nebylo ovlivněno touto chybou je vhodnější zarovnat tvář vzhledem k fixní pozici středu očí. Středy očí jsou přímo spojené se strukturou lebky a jejich pozice není ovlivněna výrazem tváře [133].

Většina FER systémů zarovnává tvář vzhledem k fixní pozici středu očí. Nejprve je odstraněn vliv naklonění tváře v x-ové ose tak, aby souřadnice levého i pravého oka byly ve stejné rovině. Následně je změněna velikost tváře vzhledem k fixní vzdálenosti mezi očima. FER systémy se liší metodou použitou k nalezení pozice středu očí.

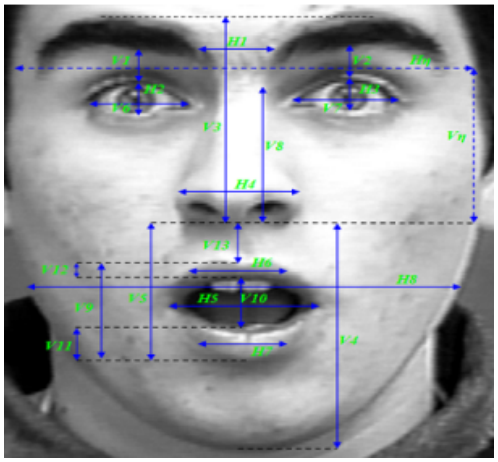
Zarovnáním tváře vzhledem k fixní vzdálenosti středu očí lze odstranit pouze natočení v z-ové ose. Pro odstranění natočení v x-ové a y-ové ose je potřeba mít k dispozici 3D data. Pokud nejsou k dispozici 3D data lze provést 3D rekonstrukci tváře z 2D snímku. Podrobněji se tomuto postupu věnuje následující kapitola 4.

3.2 Extrakce příznaků

Máme-li k dispozici zarovnanou tvář můžeme přistoupit k jejímu popisu. Cílem extrakce příznaků je zvýraznění rysů tváře pro různé výrazy nezávisle na subjektu. Popis tváře se může zaměřit jak na celý obličej, tak i na jednotlivé části. Rozlišit lze dva typy rysů tváře: trvalé a dočasné. Trvalé rysy jsou vždy viditelné, mohou však měnit svůj tvar v důsledku změny výrazu. Příkladem trvalých rysů jsou oči, obočí a ústa. Dočasné rysy jsou viditelné při změně výrazu. Tyto rysy zahrnují vrásky a různé deformace tváře (např. nafouklá pusa).

Metody pro extrakci příznaků mohou být rozděleny vzhledem k typu získa-ných příznaků na geometrické (zachycují trvalé rysy tváře) či vizuální (zachycují trvalé i dočasné rysy tváře).

Příkladem **geometrických příznaků** je změna pozice klíčových bodů vůči neutrálnímu výrazu, trajektorie pohybu nebo rychlost pohybu klíčových bodů. **Vizuální příznaky** reprezentují jednotlivé pixely snímku tváře (intenzita, barva) zpracované filtrem (např. Gabor filtr), nebo rozdílový obraz mezi snímky.



Obrázek 3.5: Ukázka geometrických příznaků. Normalizační vzdálenosti Vn a Hn jsou použité k normalizaci vektoru příznaků aby byla respektována rozdílná velikost tváře mezi subjekty. Popis jednotlivých příznaků Vi a Hi je v tabulce 3.1. Obrázek převzat z [2]

Tabulka 3.1: Popis geometrických příznaků zobrazených na obrázku 3.5.

AU jednotky	Popis příznaků	identifikátor příznaků
AU1,AU2,AU4	pohyb mezi obočím a očima, změna vzdálenosti mezi obočím	V1, V2, V3, H1
AU5,AU7,AU41, AU42,AU43,AU44, AU45,AU46	pohyb očí	V6, V7, H2, H3
AU6	pohyb tváří	H8
AU9	pohyb nosu	V8, H4
AU10	pohyb mezi nosem a ústy	V13
AU12,AU15,AU16 AU18,AU20,AU22,AU23	pohyb úst	V9, V10, V11, V12, H5 H6, H7
AU27	pohyb mezi nosem a ústy	V5
AU17, AU26	pohyb mezi nosem a bradou	V4

3.2.1 Geometrické příznaky

Výraz tváře ovlivňuje relativní pozici klíčových bodů a tvar jednotlivých částí tváře. Získáme-li pozici klíčových bodů v několika po sobě jdoucích snímcích, můžeme měřit geometrické příznaky. Příkladem geometrických příznaků je změna polohy významných bodů (či změna tvaru) vůči neutrálnímu výrazu, trajektorie pohybu nebo rychlost pohybu. Obrázek 3.5 společně s tabulkou 3.1 obsahuje ukázku geometrických příznaků pro rozpoznání svalové aktivity. Efektivita geometrických příznaků je přímo úměrná přesnosti nalezení pozice klíčových bodů. Body jsou proto vybírány tak, aby byly dostatečně kontrastní vůči svému okolí (kontura očí, obočí a rtů). Pro nalezení klíčových bodů máme na výběr z metod popsaných v sekci nalezení klíčových bodů 3.1.3, nejčastěji se používají metody založené na AAM [115]. Srovnání úspěšnosti jednotlivých AAM metod na Cohn-Kanade databázi lze nalézt v [2].

3.2.2 Vizuální příznaky

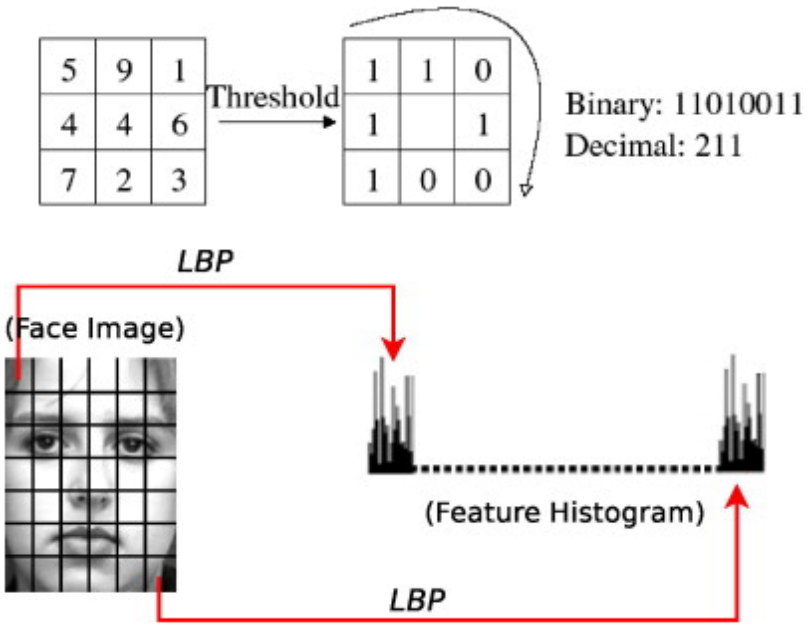
Vizuální příznaky pracují bez hlubší znalosti o vlastnostech tváře (trvalé a dočasné rysy). Při vytváření popisu nevyžadují vzorová data. Příznaky mohou reprezentovat jednotlivé pixely tváře (intenzita, barva), nebo rozdílový obraz mezi snímky. Jejich výhodou je rychlá a snadná dostupnost. Na druhou stranu mění-li se orientace tváře v průběhu extrakce příznaků stávají se nespolehlivými [30].

V poslední době jsou pro reprezentaci jednotlivých pixelů tváře hojně využí-vány příznaky označované jako dense local appearance descriptors (husté lokální vizuální příznaky). Nejprve se použije lokální filtr (nebo maska, operátor) na každý pixel a poté se vytvoří histogram přes jednotlivé bloky snímku. Díky histogramu je snížena dimenze příznaků a také citlivost na zarovnání tváře. Příkladem takových příznaků jsou Lokální Binární vzory (LBP) [47, 132, 104], Lokální Gaborovi binární vzory [75], Local Phase Quantisation [47], Histogram orientovaných gradientů [18], nebo Haarovi filtry [63]. Dalším příkladem vizuálních příznaků reprezentující jednotlivé pixely tváře jsou Gáborovi filtry [63]. Snímek je zpracován sadou filtrů s měnícími se parametry (velikost a orientace filtru). Výstupy vybraných filtrů pak představují vizuální příznaky.

Příkladem vizuálních příznaků pracujících s rozdílovým obrazem mezi více snímky jsou Optický tok [125], dočasné vzory (Temporal Templates) [72], nebo historie změny intenzity (Motion History images) [54].

Lokální binární vzory (LBP - Local Binary Patterns) LBP byli originálně navrhnuty pro popis textury [80], celkem rychle se uplatnily i v úloze zpracování výrazu tváře [47, 132, 104]. Při vytváření popisu se zpracuje lokální okolí pixelu a vytvoří se jeho binární kód (nazývaný také jako lokální binární vzor). Poté se spočte četnost výskytu binárního kódu, která reprezentuje vizuální příznakový vektor.

Nejjednodušším příkladem je LBP operátor aplikovaný na oblast snímku o velikosti 3x3 pixel. Ukázka na obrázku 3.6. Binární kód se získá porovnáním osmi okolních pixelů vůči centrálnímu pixelu. Pokud je hodnota centrálního pixelu vyšší je výsledek porovnání 1, v opačném případě 0. Výsledky porovnání jsou seřazeny za sebe a tvoří tak 8-bitový lokální binární kód. Centrálnímu pixelu je pak přiřazena decimální hodnota binárního kódu/vzoru. Celkem tedy existuje 256 vzorů pro LBP operátor aplikovaný na oblast 3x3 pixel. LBP operátor



Obrázek 3.6: Ukázka vytvoření vektoru příznaků pro Lokální Binární Vzor. LBP operátor (horní část obrázku) a vektoru příznaků (dolní část obrázku). Při tvorbě vektoru příznaků je snímek rozdělen do malých bloků a z každého bloku je spočten LBP histogram. Výsledný vektor příznaků (Feature Histogram) je vytvořen poskládáním jednotlivých LBP histogramů. Převzato z [104]

se značí $LBP_{P,R}$ může se lišit jak počtem vybraných okolních pixelů P tak i velikostí poloměru R od centrálního pixelu [79].

Poté co je celý snímek zpracován LBP operátorem, spočte se četnost výskytu jednotlivých vzorů. Pokud bychom spočetli četnost přes celý snímek ztratila by se informace o prostorovém umístění vzorů. Snímek je tedy nejprve rozdělen na jednotlivé bloky. Pro každý blok se spočte histogram pro lokální binární vzor. Nakonec se jednotlivé histogramy zařadí za sebe a vytvoří tak vizuální příznakový vektor.

Existuje několik vylepšení lokálních binárních vzorů, nejpoužívanější je tzv. uniformní vzor $LBP^{u}_{P,R}$ [79]. Binární vzor je označen jako uniformní pokud obsahuje maximálně dvě změny z 0 na 1 nebo naopak (např. 11110111). Pro LBP operátor aplikovaný na okolí 3x3 pixel se díky tomu sníží počet binárních vzorů z 256 na 59.

3.3 Klasifikace výrazu tváře

Klasifikace je posledním krokem v automatické analýze výrazu tváře. Podle toho jak klasifikátor pracuje s časovou informací dělí se FER systémy na statické a dynamické. V případě **statického přístupu** (pracujícího pouze s informací obsaženou v aktuálním snímku) bylo již odzkoušeno velké množství klasifikátorů. Z prvního oficiálního srovnání FER systémů [115] vyplývá, že současné systémy nejčastěji používají statický přístup a to klasifikátor Support

Vector Machine (SVM) použitý v 10 z 12 FER systémů. Stručný přehled klasifikátorů pro statický přístup je popsán v 3.3.1. Studie zabývající se **dynamickým přístupem** (kde klasifikátor zpracovává informaci přes sekvenci snímků) se začali objevovat až v posledních letech. Objevují se především pro rozpoznání svalové aktivity, respektive segmentace výrazu tváře na jednotlivé fáze: náběh, vrchol a sestup (viz obrázek 2.2). Nejčastěji se používají skryté Markovovi modely (HMM - Hidden Markov Model). Stručný přehled klasifikátorů pro dynamický přístup je popsán v 3.3.2.

Studii porovnávající statický a dynamický přístup pro několik různých klasifikátoru lze nalézt v [12]. V závěrečné diskuzi Cohen odpovídá na otázku volby mezi statickým a dynamickým přístupem následovně. Dynamický přístup je vhodný pro úlohu kdy je subjekt obsažen i v trénovacích datech, a to díky variabilitě trvání jednotlivých fází výrazu tváře (viz obrázek 2.2). Statický přístup je vhodný v případech, kdy je výraz v maximální intenzitě (na vrcholu), v době náběhu a sestupu poskytuje nespolehlivé výsledky. Statický systém se oproti dynamickému snadněji trénuje a implementuje. Nevyžaduje tolik trénovacích dat a nastavování velkého množství parametrů jako v případě dynamického přístupu. Vývoj FER systému by se měl zaměřit na vytvoření hybridního systému spojujícího výhody a nevýhody obou přístupů.

Nehledě na volbu statického či dynamického přístupu čelí klasifikátory problému velké dimenze vstupních dat (dimenze příznaků pro jeden vzorek se pohybuje v řádech několika tisíců). Proto se často využívají metody pro snížení dimenze dat. Nejčastěji se ve FER systémech využívá metoda hlavních komponent (PCA - Principal Component Analysis) [115].

3.3.1 Statický přístup - klasifikace aktuálního snímku

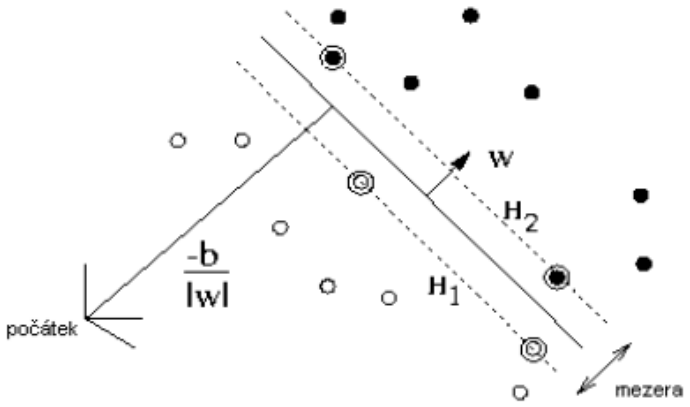
Klasifikace z aktuálního snímku nevyužívá časovou informaci. Klasifikuje se přímo vstupní snímek případně je navíc použit referenční snímek (obsahující neutrální výraz). Pro rozpoznání výrazu tváře bylo vyzkoušeno velké množství metod a to například k-nejbližší soused [56, 102, 72], Support Vector Machines (SVM) [50, 6, 73], lineární diskriminační analýza [61], bayesovské sítě [12, 102], nebo rozhodovací stromy [102]. Statické klasifikátory mohou být dále vylepšeny váženým hlasováním, např pomocí Adaboost algoritmu [102]. Nejpopulárnější metodou je dle první veřejné soutěže FERA 2011 klasifikátor SVM [115].

Support Vector Machine (SVM) Klasifikace pomocí SVM se ukázala být jako extrémně užitečná pro velké množství úloh rozpoznávání, např. rozpoznání řeči nebo rozpoznání výrazu tváře. Cílem klasifikátoru je nalezení optimální rozdělující nadroviny maximalizující vzdálenost (mezeru) mezi dvěma třídami.

Obrázek 3.7 zobrazuje lineárně separabilní data a jejich optimální rozdělující nadrovinu. Vysvětleme si nyní princip SVM pro tento jednoduchý příklad. K dispozici máme trénovací data ve tvaru $\{x_i, y_i\}$, $i = 1..N$, kde $x_i \epsilon R^d$ je vektor souřadnic a $y_i = \{-1, 1\}$ definuje třídu (negativní a pozitivní vzorek), do které daná souřadnice patří. Body ležící na oddělující nadrovině vyhovují rovnici

$$w \cdot x + b = 0 \tag{3.10}$$

kde w je normálový vektor oddělující nadroviny a $|b|/|w|$ je vzdálenost nadroviny od počátku souřadnic. Všechna trénovací data vyhovují následujícím podmín-kám



Obrázek 3.7: Ukázka rozdělující nadroviny ve 2D prostoru pro lineárně separabilní data, převzato z [8]

$$\begin{array}{l} H_1: \ x_i \cdot w + b \geq +1, y_i = +1 \\ H_2: \ x_i \cdot w + b \leq -1, y_i = -1 \end{array} \Rightarrow \vee i: \ y_i(x_i \cdot w + b) - 1 \geq 0 \tag{3.11}$$

změnou parametru w budeme měnit šířku mezery. Body ležící na nadrovině H_1, H_2se nazývají support vectors a jsou v obrázku 3.7 označeny kroužkem. Hledáme tedy support vectors vyhovující podmínce

$$\underset{x_i}{min}|x \cdot w + b| = 1 \tag{3.12}$$

Chceme-li maximalizovat mezeru musíme maximalizovat výraz $2/|w|$ nebo-li minimalizovat výraz $|w|$. K tomuto účelu je použit princip Lagrangeových koeficientů pro nalezení optimální oddělující nadroviny

$$L(w, b, \alpha) = \frac{1}{2}|w|^2 - \sum_{i=1}^{l} \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^{l} \alpha_i \tag{3.13}$$

kde α_i jsou Lagrangeovy koeficienty pro které platí $\alpha_i > 0 \Leftrightarrow x_i$ je support vector. Řešení Lagrangeovy rovnice je problém kvadratického programování. Výsledná nadrovina pro lineární SVM je následující

$$\mathcal{H}: f(x_j) = \sum_i \alpha_i y_i x_i^T x_j + b \tag{3.14}$$

Detailní popis řešení je v [8].

Pokud nejsou data v daném prostoru lineárně separabilní je možné nalézt mapování do vyšší dimenze, kde je již lineární separabilita dat splněna. Výsledný klasifikátor bude nelineární. Řešení tohoto problému je možné díky skalárnímu součinu $x_i \cdot x_j$. Trénovací data jednoduše přetransformujeme z prostoru R^d do $\mathcal{H}$ pomocí zobrazení

$$\Phi : R^d \rightarrow \mathcal{H} \tag{3.15}$$

Trénovací algoritmus nyní závisí pouze na datech ze skalárního součinu v prostoru $\mathcal{H}$. Pokud by existovala kernel funkce K splňující $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, mohli bychom použít pouze K a nebylo by nutné explicitně znát zobrazení Φ. Příkladem takové kernel funkce je RBF (Radial Basis Function) kernel

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2} \tag{3.16}$$

Aby byla splněna podmínka platné kernel funkce, stačí dodržet Mercerovu podmínku. Kernel musí být symetrický a pozitivně definitní. Rozhodovací funkce klasifikátoru pak je

$$\mathcal{H} : f(x) = \sum_i \alpha_i y_i K(s_i, x) + b \tag{3.17}$$

kde s_i reprezentuje support vektor. Detailní odvození pro nelineární SVM lze nalézt v [8].

3.3.2 Dynamický přístup - klasifikace sekvence snímků

Klasifikace sekvence snímků využívá časovou informaci. Tato informace je získána ze vzájemného vztahu příznaků z několika po sobě jdoucích snímků (např. změna polohy významných bodů) nebo se používají příznaky, které časovou informaci obsahují přímo v reprezentaci (např. Temporal Templates). Pro zpracování časové informace se využívají klasifikátory, které se přizpůsobí proměnlivé dynamice výrazu obličeje (stejná emoce pozorována u různých lidí nemusí mít stejnou délku trvání). Nejčastěji se využívají skryté Markovovi modely (HMM) [28, 85, 12, 54], dále byli vyzkoušeny klasifikátory jako stromově rozšířený naivní Bayes (Tree Augmented Naive Bayes) [12, 102, 114], nebo klasifikátor založený na vlastních pravidlech [90, 85].

3.4 Srovnání úspěšnosti FER systémů

K otestování úspěšnosti FER systémů se nejčastěji používá Cohn-Kanade Database (dále jen CK databáze)[49]. CK databáze podrobněji popsaná v příloze B obsahuje video data pro 100 subjektů předvádějících šest základních emocí. Databáze je anotovaná jak pro kategorické dělení tak i pro FACS systém. Data jsou pořízená v kontrolovaných podmínkách, tj. konstantní osvětlení a čelní pohled na subjekt. Předváděný výraz tváře je zachycen od neutrálního výrazu po maximální intenzitu.

Tabulky 3.2 a 3.3 uvádějí přehled dosažených výsledků na CK databázi pro rozpoznání kategorických emocí a svalové aktivity. Tabulky obsahují informace o metodách použitých v jednotlivých krocích rozpoznávání výrazu tváře. Nejúspěšnější systém pro kategorické emoce vytvořili Image Formation and Processing (IFP) společně s Intelligent Systems Lab Amsterdam (ISLA). Úspěšnost systému je 95%. Pro svalovou aktivitu je nejlepší systém od týmu Face group (FG) společně s Affect analysis group (AAG), který dosáhl 97%. Díky dosaženým výsledkům je rozpoznávání výrazu tváře v kontrolovaných podmínkách považo-váno za vyřešený problém.

Tabulka 3.2: Přehled FER systémů pro rozpoznání kategorických emocí na CK databázi. První sloupec obsahuje zkratky týmů, ukázky systémů v příloze A. Další tři sloupce obsahují zvolené metody a poslední sloupec úspěšnost systému.

Tým	Předzpracování	Extrakce Příznaků	Klasifikace	Úsp [%]
IFP+ ISLA [102]	Piecewise Bézier Volume Deformation - vektory definující změnu klíčových bodů mezi snímky		K-nejbližší soused	95
MPL [76]	Detektor tváře	Gáborovi filtry	SVM	94
FG[2]	AAM - normalizovaná vzdálenost mezi klíčovými body		SVM	93
IFP[113]	AAM - vektory definující změnu klíčových bodů mezi snímky		Naivní Bayes	84
AC [52]	Fixní střed očí	změna klíčových bodů vůči neutrálnímu výrazu	SVM	81

Tabulka 3.3: Přehled FER systémů pro rozpoznání svalové aktivity na CK databázi. První sloupec obsahuje zkratky týmů, ukázky systémů v příloze A. Další tři sloupce obsahují zvolené metody a poslední sloupec úspěšnost systému.

Tým	Před- zpracování	Extrakce Příznaků	Klasifikace	Úsp [%]
AAG +FG [65]	AAM - normalizovaný tvar a textura		SVM	97
MPL[124]	Detektor očí a úst	Gáborovi filtry	SVM	95
HCI^2[84]	Partical filtr - normalizovaná pozice klíčových bodů		Vlastní pravidla	90

Donedávna byli všechny veřejně dostupné databáze natáčeny v kontrolovaných podmínkách. Navíc neobsahovaly popis jak provést testování systému (tj. rozdělení dat na trénovací a testovací data, metrika na měření úspěšnosti). Nebylo tedy možné srovnat jednotlivé FER systémy mezi sebou. V roce 2011 proběhla první veřejná soutěž pro srovnání FER systému The First Facial Expression Recognition Challange (dále FERA2011) [115]. Testovány byli jak sys-témy na rozpoznávání emocí tak na rozpoznávání AU jednotek. Cílem bylo ukázat jaká je úspěšnost existujících FER systémů a poukázat na problémy které jsou stále nevyřešené.

FERA2011 se zúčastnilo celkem 12 týmů, z toho 10 týmů otestovalo úspěšnost na rozpoznávání kategorických emocí a 5 týmu otestovalo rozpoznávání AU jednotek. Pro testování byla zvolena databáze GEMEP [4] neboť splňovala dvě důležité kriteria. V době vyhlášení soutěže nebyla databáze přístupná pro širokou veřejnost. Dále databáze obsahuje anotaci AU jednotek pro každý snímek včetně informaci o kategorické emoci pro jednotlivá videa.

Tabulka 3.4: Přehled FER systémů pro rozpoznání emocí na datech z FERA 2011.

Tým	Před-zpracování	Extrakce Příznaků	Klasifikace	Úsp [%]
Riverside[127]	SIFT registrace	LBP + Local Phase Quantisation	SVM	83.8
UIUC-UMC[111]	fixní střed očí	SIFT + Optický tok	SVM	79.8
KIT[31]	fixní střed očí	Histogram z diskretní kosinovi transformace	regrese metodou částečných nejmenších čtverců	77.3
MPL[63]	fixní střed očí	Extrémní změna rychlosti a zrychlení intenzity pixelů	Logistická regrese	76.1
ANU[18]	detektor tváře	Histogram orientovaných gradientů + Local Phase Quantisation	SVM + K-nejbližší sousedd	73.4
UCLIC[71]	detektor tváře	LBP + Histogram z historie pohybu	multi-kernel SVM	70
Montreal[17]	fixní střed očí	Histogram orientovaných gradientů	SVM	70
NUS[105]	detektor tváře	akumulovaná energie snímků	SVM	67.2
FG[10]	normalizace - referenční tvar	poloha a intenzita okolí klíčových bodů	SVM	60
AC[3]	fixní střed očí	Gáborovi filtry	Dynamická Bayesovská síť	44.4

GEMEP databáze obsahuje 7000 videí zachycujících 18 různých emocí ztvár-něných 10 herci. Pro soutěž byla vybrána část databáze nazvaná GEMEP-FERA obsahující 5 kategorických emocí (radost, smutek, vztek, strach a úleva) a 12 AU jednotek nejčastěji se projevující v datech (1,2,4,6,7,10,12,15,17,18,25,26). Podrobný popis GEMEP-FERA a protokolu o testování je uveden v kapitole 5

FER systémy mohli svou úspěšnost srovnat nejen oproti sobě, ale i proti základnímu systému vytvořenému organizátory soutěže. Základní systém byl vytvořen tak, aby bylo pro kohokoliv snadné zopakovat dosažené výsledky. Popis základního systému je uveden v kapitole 3.Tabulka 3.4 shrnuje dosažené výsledky pro rozpoznávání kategorických emocí. Vyšší úspěšnost než základní systém dosáhlo 9 z 10 týmů. Výhercem se stal tým z Riverside [127] s úspěšností 83.8% (průměr přes všechny emoce). Pro rozpoznávání AU jednotek všech 5 týmů překonalo základní systém. Výhercem se stal tým ISIR [103] s úspěšností rozpoznávání 62% (průměr přes všechny AU jednotky).

Tabulka 3.5: Přehled FER systémů pro rozpoznání svalové aktivity na datech z FERA 2011.

Tým	Předzpracování	Extrakce Příznaků	Klasifikace	Úsp [%]
ISIR[103]	fixní střed očí	Locaní Gáborovi Binární vzory	multi-kernel SVM	62
MPL[63]	fixní střed očí	Gáborovi filtry	SVM	58.3
KIT[31]	fixní střed očí	Histogram z diskretní kosinovi transformace	regrese metodou nejmenších čtverců	52.3
FG[10]	normalizace - referenční tvar	poloha a intenzita okolí klíčových bodů	SVM	51
AC[3]	fixní střed očí	Gáborovi filtry	HMM + Dynamická Bayesovská síť	46.1

3.5 Shrnutí

Prvním krokem při rozpoznání výrazu tváře je předzpracování dat. Přesná informace o aktuální pozici, velikosti a orientaci tváře je zásadní pro rozpoznání výrazu tváře. Většina FER systémů se zaměřuje pouze na rozpoznání výrazu tváře v čelním pohledu. Neuvažuje změnu orientace tváře v prostoru či dlouhodobou změnu způsobenou např. změnou účesu.

Druhým krokem při rozpoznávání výrazu je popis tváře. Na výběr jsou buďto příznaky geometrické, vizuální nebo kombinace obou. Volba popisu tváře je závislá na typu úlohy. V případě že chceme rozpoznat jednotlivé fáze výrazu tváře (náběh, vrchol a sestup) jsou vhodnější příznaky geometrické [116]. Pokud rozpoznáváme výraz tváře v maximální intenzitě a okolí je vhodnější použít příznaky vizuální, nebo kombinaci vizuálních a geometrických příznaků [115, 12]. Dostupné metody pro extrakci příznaků vyžadují data s minimální změnou natočení tváře (nejčastěji čelní pohled).

Posledním krokem je klasifikace výrazu tváře. Existujíc FER systémy se zaměřují především na rozpoznávání kategorických emocí pro čelní pohled. Pro statický přístup, kdy je výraz rozpoznáván v každém snímku, bylo vyzkoušeno velké množství klasifikátorů. Nejčastěji se používá SVM [115]. Dynamický přístup klasifikace se využívá především k rozpoznání jednotlivých fází výrazu (náběh, vrchol a sestup). Úspěšnost rozpoznávání výrazu se s natočením mimo čelní pohled snižuje [115, 100, 63, 131].

Závěry z první soutěže v rozpoznávání výrazu tváře FERA2011[115] jsou následující. FER systémy pro rozpoznávání emocí jsou pro vědeckou komunitu stále nejpopulárnější, a to i přes kritiku praktického použití (tj. zařazení jakéhokoliv výrazu do předem definovaných kategorií). Nejčastěji používaný učící mechanismus je SVM, z 12 týmů jej použilo 10. Rozpoznávání emocí pro známý subjekt je považováno za vyřešený problém, jakékoliv jeho vylepšení je už věcí průmyslu. Rozpoznávání AU jednotek se stane hlavním tématem pro budoucí vývoj FER systémů. Organizátoři FERA2011 plánují v příští soutěži rozšířit počet AU jednotek z 12 na 32. Dále poukazují na nutnost ohodnocení intenzity AU jednotky a určení jejích komponent (náběh, vrchol, sestup). Neboť tyto charakteristiky jsou zásadní pro porozumění chování subjektu. Libovolné natočení tváře v prostoru je v současné době považováno za největší problém FER systémů.

Kapitola 4

Natočení tváře na čelní pohled

Stávající systémy pro rozpoznávání výrazu tváře pracují převážně s daty poříze-nými v kontrolovaném prostředí. Osoba je vyzvána k provedení výrazu tváře s omezeným pohybem hlavy v prostoru. Natočení tváře mimo čelní pohled je tedy minimální. Systémy navržené pro tento typ dat se nedokáží vyrovnat s reálnými daty obsahující libovolné natočení tváře. Jejich úspěšnost se výrazně snižuje s natočením tváře mimo čelní pohled[115, 100, 63, 131].

Několik FER systémů se zaměřilo na problematiku libovolného natočení tváře v prostoru. Jedním z řešení je trénovat samostatné klasifikátory pro různá natočení. Studie [41] popisuje experiment pro natočení tváře do pěti různých úhlů. Metoda je závislá na dostupných datech. Není schopná pracovat s natočením tváře, které nebylo k dispozici při trénování. Pro reálnou situaci (libovolné natočení tváře) vyžaduje tento postup velké množství trénovacích dat, které není snadné pořídit. Další možností jak odstranit vliv natočení tváře je normalizace na čelní pohled [97, 46]. K zarovnání tváře na čelní pohled se učí model zachycující závislost mezi 3D a 2D daty. Naučené modely lze pak aplikovat na 2D data a provést normalizaci tváře na čelní pohled. Podrobněji jsou tyto metody popsané v sekci 4.1.

Problém libovolného natočení tváře je dlouhodobě řešen v oblasti identifikace osob. Zde je natočení na čelní pohled provedeno technikou 3D rekonstrukce tváře z 2D snímku. Stávající metody, popsané v sekci 4.2 výraz tváře potlačují. Identifikace osob se tím zjednoduší, pro FER systém nejsou tyto metody vhodné. I přesto ukazují cestu, jak provést 3D rekonstrukci tváře pro libovolné natočení.

Metoda navržená v této práci vychází z technik používaných při identifikaci osob. Nová metoda je schopná rekonstruovat 3D objekt tváře respektující výraz tváře, popsána je v sekci 4.3. Navržená metoda potřebuje na vstupu 2D snímek s lokalizovanými klíčovými body a 3D statický model tváře s vyznačenými body korespondující s 2D snímkem. Výstupem je 3D objekt tváře včetně aktuálního výrazu tváře.

3D rekonstrukce tváře z jediného snímku nabízí alternativní řešení ke klasickým metodám pro pořízení 3D dat (strukturované světlo, stereo viděni, či fotometrie). Nevyžaduje žádné speciální zařízení a při sběru dat je osoba minimálně ovlivněna nahrávacím zařízením (není ji na tvář promítáno světlo jako v případě fotometrie, či kladeno omezení na rychlé pohyby jako v případě strukturovaného světla). Navíc žádná z klasických metod pro pořízení 3D dat nelze použít v reálné situaci obsahující spontánní emoce. Zařízení využívající metodu strukturovaného světla nezvládá rychlé pohyby hlavy. Použití více kamer vyžaduje přesnou kalibraci a navíc díky vysoké výpočetní náročnosti je 3D rekonstrukce provedena offline. Fotometrie

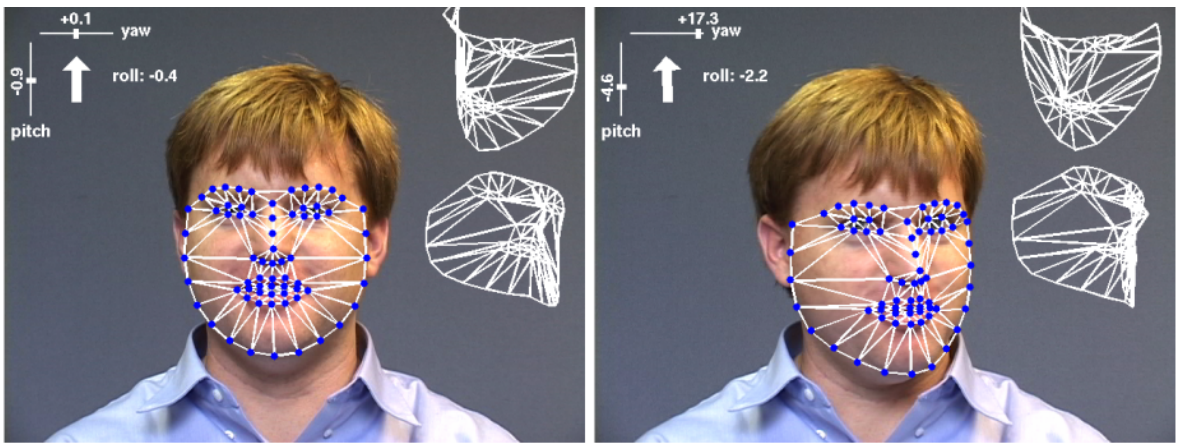
vyžaduje kalibraci pro každý výraz tváře. Subjekt musí během kalibrace předvést stejný výraz, který bude poté natočen. Další výhodou 3D rekonstrukce je, že většina dostupných 2D databází obsahuje dynamická data. Oproti dostupným 3D databázím, kde jsou až na vyjímky (databáze BU-4DFE[128] a D3DFACS[15]) k dispozici data statická. Navíc databáze obsahující dynamická 3D data jsou pouze hrané. 3D rekonstrukce tváře z jediného snímku má tedy pro FER systémy velký význam.

4.1 FER systémy s normalizací tváře

Většina FER systémů je vysoce závislá na natočení hlavy a vyžaduje,aby maximální natočení do všech stran (sklon nahoru a dolu, otočení doprava a doleva) bylo ±10 stupňů. V posledních letech se několik systémů zaměřilo na odstranění natočení tváře pomocí normalizace tváře na čelní pohled. Následující text popisuje současný stav řešení. Normalizován je buď tvar tváře, textura nebo obojí.

Normalizace tvaru tváře

K normalizaci geometrického tvaru tváře se nejčastěji využívají 3D modely tváře, ukázka na obrázku 4.1. Ve studii [126] byl navržen přístup kombinující 3D aktivní vzhledový model (3D-AAM) s technikou structure-from-motion, stejný přístup byl použit i v [67]. Zatímco první studie testovaná na rozpoznávání emocí poukazuje na zlepšení rozpoznávání výrazu pro různá natočení tváře. Druhá studie se zaměřila na ověření přínosu 3D-AAM vůči 2D-AAM. Studie poukazuje na nepřesnost rekonstruovaného 3D tvaru tváře a zhoršení rozpoznávání oproti 2D aktivnímu vzhledovému modelu.



Obrázek 4.1: Výsledek normalizace tvaru tváře pomocí 3D Aktivních vzhledových modelů. Převzato z [126]

Dalším příkladem s 3D modelem je systém [46]. Zde autoři vytváří ze sekvence snímků 3D personal mean shape (3D-PMS) s neutrálním výrazem. 3D-PMS je vytvořen jako průměr pro několik výrazů přes všechny body definující tvar tváře. Při rozpoznávání výrazu je tvar tváře v

každém snímku porovnán s 3D-PMS. Výsledky rozpoznávání výrazu nebyly závislé na otočení tváře do stran v rozsahu 0-45 stupňů. Přehled dalších metod pro normalizaci tvaru založeného na 3D modelu lze nalézt v *[55] a* [97].

Alternativní přístup k normalizaci tvaru tváře je navržený v [97]. Systém trénuje regresní model schopný normalizovat tvar tváře na čelní pohled. Při trénování jsou systému pro různá natočení tváře předkládány vždy dva tvary tváře se stejným výrazem. Jeden tvar je v čelním pohledu a druhý je natočený. Systém je schopný rozpoznat výraz tváře, respektive emoci, pro jakékoliv natočení tváře v úhlu 0 až 45 stupňů pro otočení do stran a 0 až 30 stupňů pro sklon hlavy. A to přesto. že byl trénován jen pro 12 diskrétních natočení tváře. Navíc je systém schopný zobecnění ve smyslu neúplných trénovacích dat. Pokud je některá emoce vynechána při trénování, model je stále schopen provést normalizaci tváře na čelní pohled.

Všechny uvedené FER systémy rozpoznávají výraz tváře nezávisle na jeho natočení. Pracují však pouze s tvarem tváře a ignorují texturu. Tvar tváře popisují pouze klíčové body (např. konturu rtů, očí, obočí), které nezachycují změny jako jsou zvednuté líce nebo nafouklé tváře. FER systémy pracující s texturou tváře dosahují výrazně lepších výsledků při rozpoznávání výrazu než systémy pracující pouze s tvarem tváře [115]. Systém schopný provést normalizaci tváře na čelní pohled včetně textury by měl dosáhnout výrazného zlepšení při rozpoznávání výrazu tváře.

Normalizace textury tváře

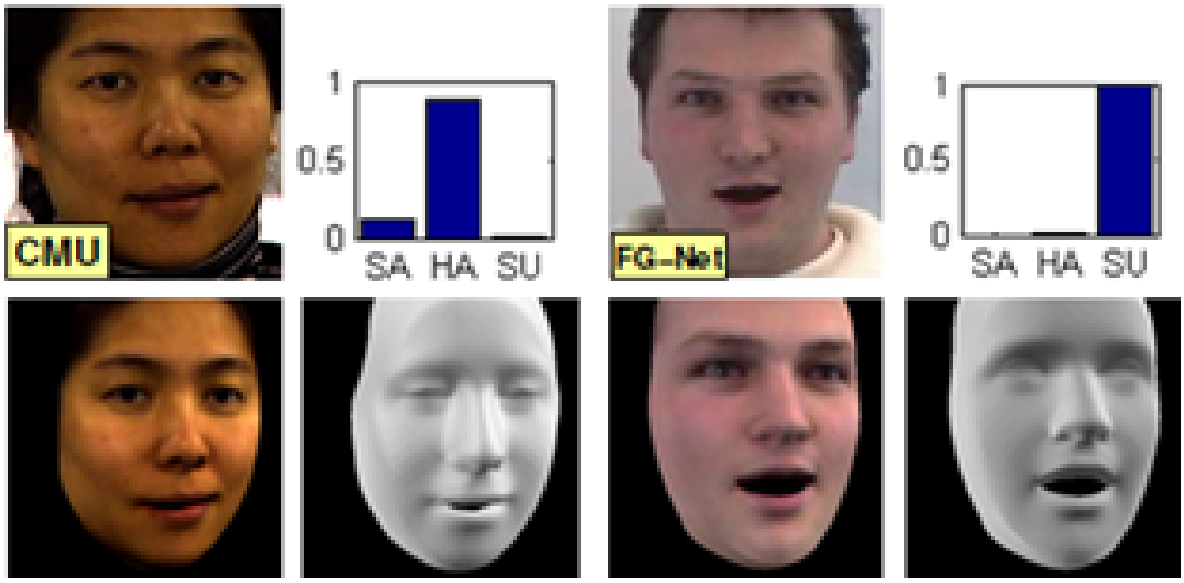
Normalizace textury tváře není ve FER systémech v současné době prozkoumána. Prvotní kroky učinila studie [59] provádějící normalizaci na čelní pohled za pomoci mapování textury na povrch 3D válce. Systém byl testován na vlastní databázi obsahující různé natočení tváře v prostoru. Výsledky pro celou databázi jsou uvedeny pomocí jediné matice záměn. Nelze tedy posoudit vliv normalizace textury na rozpoznávání výrazu tváře pro libovolné natočení. Není zde ani provedeno srovnání se systémem bez normalizace tváře.

Další studie [120] je inspirována 3D rekonstrukcí z oblasti identifikace osob. Hlavním přínosem je návrh metody schopné provést 3D rekonstrukci tváře včetně aktuálního výrazu tváře, ukázka na obrázku 4.2. K normalizaci textury je využit 3D morphable model (3DMM) [7] v kombinaci s Gaussian Mixture Model (GMM). Samotný 3DMM neumí rozhodnout o aktuálním výrazu tváře. Studie navrhuje řešení jak odstranit nejednoznačnost deformace tváře mezi neutrálním výrazem a aktuálním výrazem tváře. Určení výrazu je závislé na GMM modelu. Pravděpodobnostní model GMM se z klíčových bodů tváře naučí deformaci pro jednotlivé výrazy. Při rekonstrukci neznámého snímku jsou odhadovány parametry textury, osvětlení a 3D objektu. Minimalizován je rozdíl mezi aktuálním snímkem a deformovaným 3D modelem následujícím způsobem

$$\min_{\beta,\ell} \|\mathrm{I}_{input} - \mathrm{B}(T(\beta), \mathrm{n})\ell\|$$

kde sférická harmonická báze $\mathbf{B}$ je definována povrchovou normálou $\mathbf{n}$ a intensitou textury $T(\beta)$. Povrchová normála $\mathbf{n}$ je definována aktuálním 3D modelem. Odhadovány jsou tedy koeficienty textury β a sférické harmonické koeficienty ℓ. 3D model je při inicializaci deformován na základě rozpoznaného výrazu z 2D tvaru tváře.

Studie se zaměřila na vyhodnocení kvality provedené 3D rekonstrukce. Nelze tedy vyhodnotit vliv normalizované tváře na rozpoznávání výrazu. Navíc základním principem metody je rozpoznání výrazu z geometrického tvaru tváře a následná deformace 3D modelu. Rekon-



Obrázek 4.2: Výsledek normalizace textury tváře pro různé výrazy: zleva radost, překvapení. Převzato z [120]

strukce 3D objektu a textury je tedy závislá na rozpoznání výrazu z 2D geometrického tvaru tváře. Pokud je toto rozpoznání nepřesné je vytvořený 3D objekt jiný než skutečný výraz.

4.2 3D rekonstrukce tváře pro identifikaci osob

V oblasti identifikace subjektu (face recognition) se 3D rekonstrukce tváře z jediného snímku stala velkým tématem. A to i přesto, že úloha rekonstrukce tváře z jediného snímku je matematicky špatně podmíněná. Může mít více než jedno řešení pro stejná vstupní data. V oblasti počítačového vidění se pro řešení této úlohy využívá přístup založený na 3D statistickém modelu. Nejčastěji je využíván 3D morphable model (3DMM) [7], který je nutné vytvořit z velkého počtu 3D modelů tváří, aby postihl variaci (deformaci) pro velkou skupinu subjektů. S ohledem na stávající přístupy 3D rekonstrukce tváře z jediného snímku lze sepsat několik obecných kroků algoritmu pro 3D rekonstrukci tváře. Odlišnost jednotlivých přístupů je dána způsobem řešení jednotlivých kroků. Seznam těchto kroků je následující

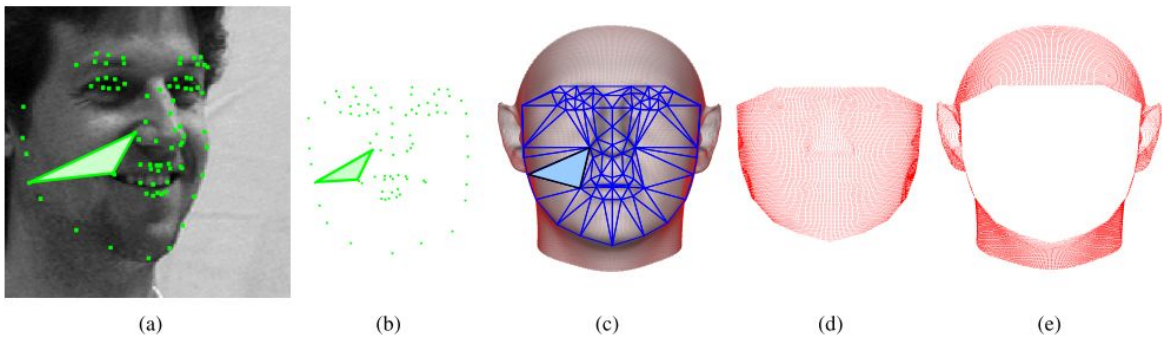
1. **Nalezení tváře a klíčových bodů.** Nejprve musí být nalezena oblast, kde se nachází tvář. Poté mohou být lokalizovány jednotlivé klíčové body (např. oči, nos, ústa).

2. **Odhad hloubky pro aktuální natočení tváře**. Pro přesnou a realistickou rekonstrukci je nutné mít k dispozici jak pozici jednotlivých klíčových bodů tváře tak i jejich hloubku. Při odhadu hloubky je respektována orientace tváře ve snímku, vzhledem k odhadu natočení tváře. Vytvořená hloubková mapa ze vstupního obrázku je v dalším kroku využita k tvorbě celkového modelu.

3. **3D rekonstrukce**. Máme-li k dispozici jednotlivé klíčové body tváře a hloubkovou mapu provede se deformace předpřipraveného 3D modelu tak, aby co nejlépe odpovídal danému

subjektu. Nakonec je na deformovaný 3D model namapována 2D textura a provedeny případné úpravy jako doplnění chybějící textury ze symetrické části obličeje a vyhlazení textury pomocí interpolace chybějících bodů.

V této sekci je stručně popsána vybraná metoda schopná plně automaticky provést 3D rekonstrukci tváře z jediného 2D snímku během několika vteřin. Přehled dalších metod lze nalézt v [108, 106, 58].



Obrázek 4.3: 3D rekonstrukce pro identifikaci subjektu (a) vstupní snímek, (b) normalizovaný 2D tvar tváře na čelní pohled, (c) 3D model včetně trojúhelníkové sítě, (d) 36,112 bodů uvnitř sítě trojúhelníků, (e) 39,805 bodů mimo síť trojúhelníků. Převzato z[92]

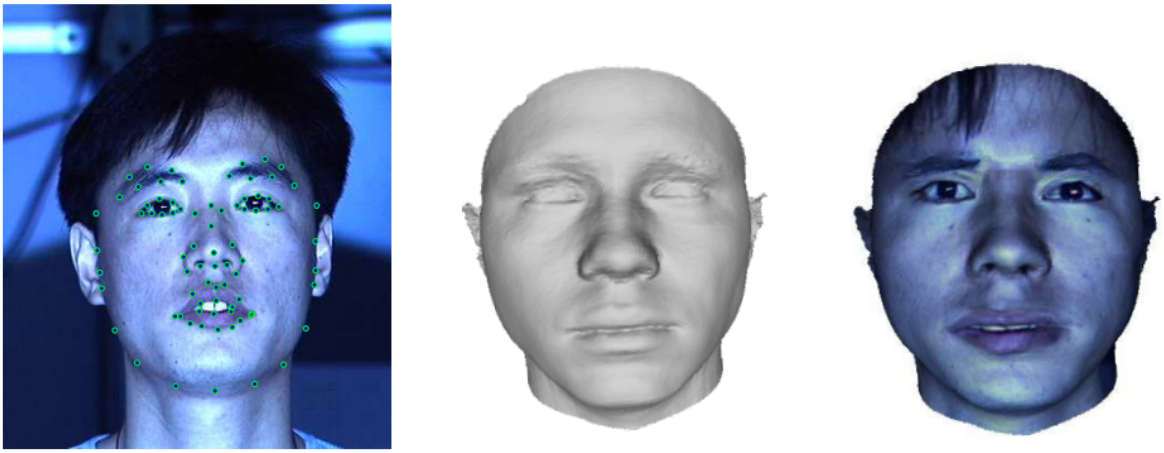
Metoda 3D rekonstrukce popsaná v [92] je typický představitel použití 3D morphable modelu (3DMM). K nalezení tváře je použit Viola&Jones detektor, k nalezení klíčových bodů je použita metoda Aktivních Vzhledových Modelů (AAM). Na tváři je nalezeno 79 klíčových bodů. Pomocí Expectation-Maximization (EM) algoritmu je odhadnuta hloubka těchto bodů a natočení tváře v prostoru. Postup určení hloubky zbývajících bodů je zachycen na obrázku 4.3. 3D model použitý k odhadu hloubky byl vytvořen z dat USF Human-ID database [7].

Výstupem EM algoritmu jsou klíčové body v 3D a odhad natočení tváře. Nejprve jsou klíčové body normalizovány na čelní pohled. Body jsou natočeny na čelní pohled a je provedena ortogonální projekce ze 3D do 2D prostoru. V 3D modelu je jednorázově vyznačeno 79 bodů korespondujících s body 2D tvaru tváře. Normalizované klíčové body (2D tvar tváře natočený na čelní pohled) lze snadno porovnat s vyznačenými body 3D modelu v čelním pohledu. K určení zbývajících bodů 3D modelu je využita trojúhelníková síť a princip barycentrických souřadnic. . Interpolací barycentrických souřadnic jsou získány body uvnitř trojúhelníkové sítě normalizovaného 2D tvaru. K odhadu hloubky jsou z trénovacích dat pomocí PCA naučeny dva lineární deformovatelné modely

$z_{all} = \mathbf{W}_{all} c_{all} + \mu_{all}$ a $z_{in} = \mathbf{W}_{in} c_{in} + \mu_{in}$,

kde $\mathbf{W}_{all} \epsilon \mathcal{R}^{\mathrm{N_{all}} \times \mathrm{m}}$ a $\mathbf{W}_{in} \epsilon \mathcal{R}^{\mathrm{N_{in}} \times \mathrm{m}}$reprezentují vlastní vektory pro všechny body 3D modelu N_{all} , respektive pro body uvnitř trojúhelníkové sítě N_{in}. Body N_{in} jsou podmnožinou bodů N_{all}. Na základě toho autoři zavedli předpoklad, že koeficienty c_{all}a c_{in} jsou v ideálním případě stejné. μ_{all} a μ_{in} jsou střední tvary 3D modelu, respektive 3D modelu uvnitř trojúhelníkové sítě.

Pro neznámý snímek jsou koeficienty c_{in} spočteny z interpolovaných bodů zpětnou projekcí $c_{in} = \mathbf{W}_{in}^{T} z_{in}$. Body mimo trojúhelníkovou síť jsou spočteny z modelu $z_{all} = \mathbf{W}_{all} c_{in} + \mu_{all}$. Nakonec je pro každý bod namapována textura.



Obrázek 4.4: Ukázka 3D rekonstrukce s potlačením výrazu tváře. Vstupní snímek s lehce našpulenými a otevřenými ústy. 3D rekonstrukce včetně textury s ústy v neutrálním výrazu. Převzato z [92]

Cílem identifikace osob je potlačení výrazu tváře. Díky trénování 3D modelu pouze pro neutrální výraz je tato podmínka při rekonstrukci splněna, ukázka na obrázku 4.4. Aby se tato metoda dala použít pro FER systém, musí být deformovatelný model natrénován pro všechny možné výrazy. Vysoká variabilita 3D modelu však vede na více řešení pro jeden vstupní snímek. Možnost, jak tento problém vyřešit, byla navrhnuta ve studii [120] popsané v předchozí sekci. Metoda však vyžaduje rozpoznání výrazu během rekonstrukce 3D objektu tváře.

4.3 3D rekonstrukce tváře pro rozpoznání výrazu

Stávající metody rekonstrukce tváře dokáží pro libovolné natočení buď rekonstruovat 3D objekt tváře s neutrálním výrazem [92], a nebo rekonstruovat 3D objekt tváře včetně výrazu s nutností rozpoznat výraz již během rekonstrukce 3D objektu tváře [120]. **Přínosem nově navržené metody popsané v této sekci je rekonstrukce 3D objektu tváře z jediného 2D snímku včetně aktuálního výrazu pro libovolné natočení tváře bez nutnosti rozpoznání výrazu tváře.** Výstupem je tvář s aktuálním výrazem normalizovaná na čelní pohled, ukázka vstupu a výstupu systému pro několik různých výrazů je na obrázku 4.5.

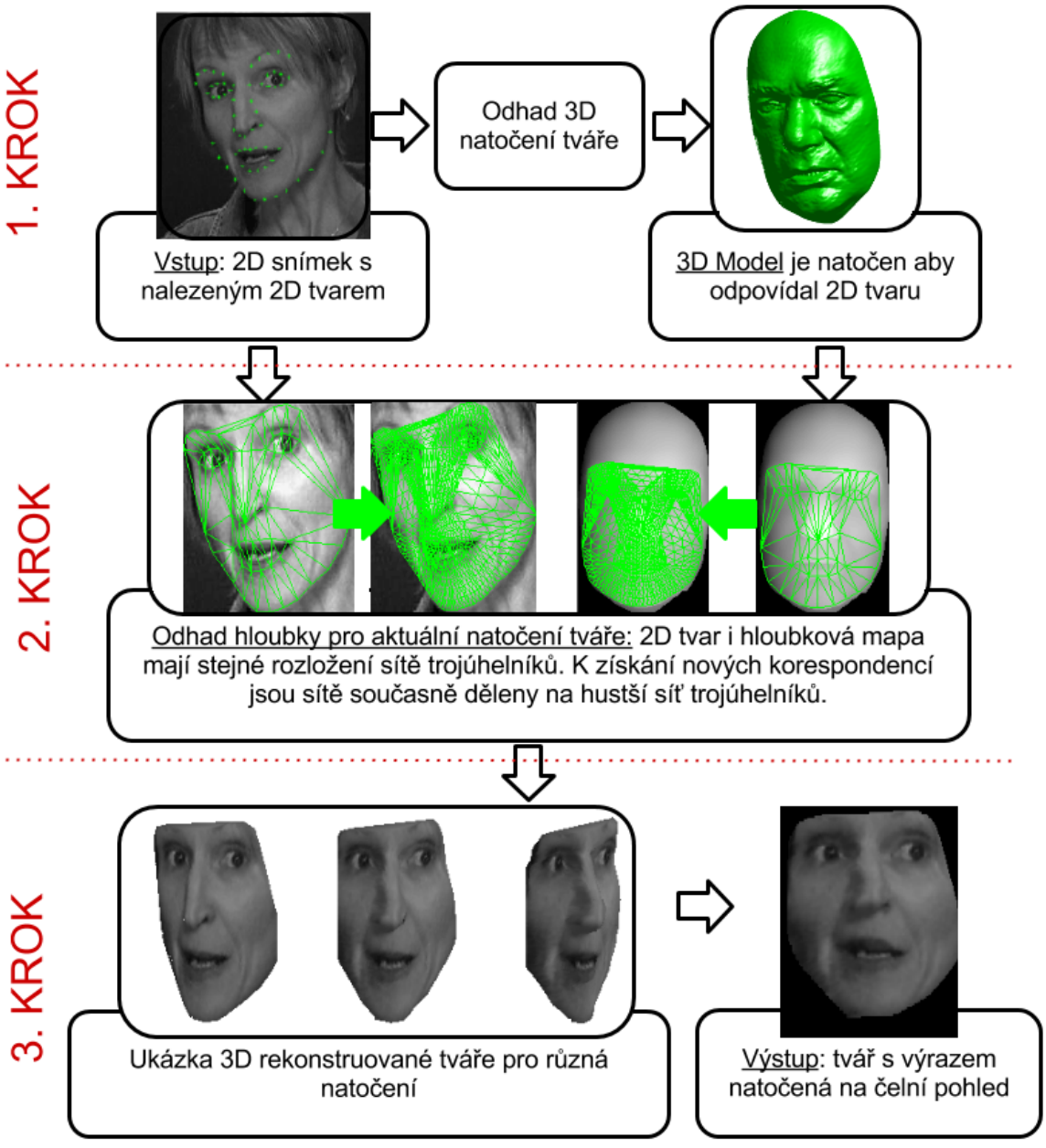
Jednotlivé kroky normalizace tváře na čelní pohled jsou zachyceny na obrázku 4.6. Na vstupu je požadován 2D snímek s klíčovými body reprezentující 2D tvar tváře. V prvním kroku je z 2D tvaru tváře odhadnuto natočení tváře a 3D model je natočen tak, aby odpovídal 2D tvaru. V druhém kroku je odhadnuta hloubka pro každý bod snímku uvnitř 2D tvaru tváře. V posledním kroku je vytvořen 3D objekt tváře, který je natočen na čelní pohled a promítnut do 2D prostoru.

Obrázek 4.5: Ukázka vstupu a výstupu systému pro 3D rekonstrukci tváře. První sloupec je vstup do systému, druhý sloupec je rekonstruovaný 3D objekt a třetí sloupec je tvář normalizovaná na čelní pohled.

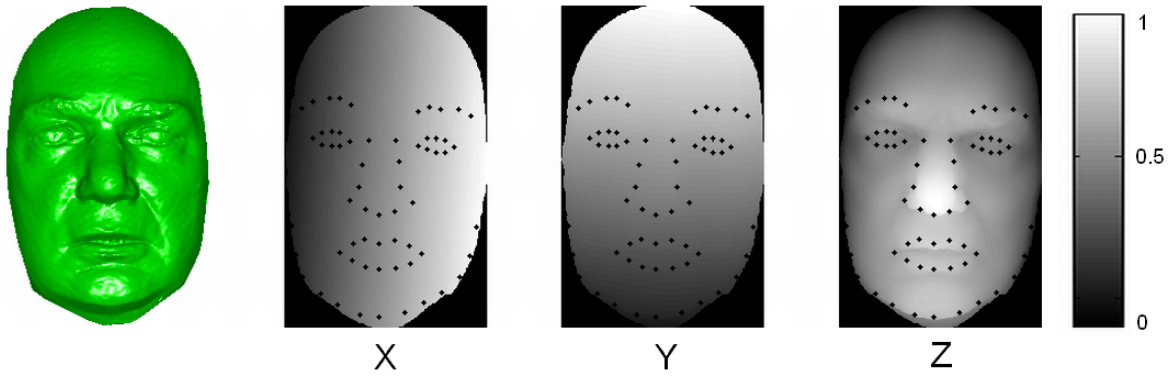
V prvním kroku je z tvaru tváře odhadnuto natočení vzhledem k osám x,y,z. Tvar je reprezentován klíčovými body. Klíčové body umístěné v okolí očí, obočí, nosu, úst a linie tváře jsou nalezeny pomocí Aktivních Vzhledových Modelů (AAM). Tvar tváře reprezentovaný maticí T

$$\mathbf{T} = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \\ y_1 & y_2 & \cdots & y_k \end{bmatrix},$$

obsahuje souřadnice klíčových bodů ve 2D prostoru. 3D referenční model je natočen tak, aby odpovídal natočení tváře ve vstupním snímku. Metoda pro odhad natočení je popsaná v sekci 3.1.4.



Obrázek 4.6: Blokové schéma 3D rekonstrukce tváře z 2D snímků včetně aktuálního výrazu tváře.



Obrázek 4.7: Referenční 3D model v čelním pohledu včetně reprezentace pomocí matic X,Y a Z. Černé hvězdičky reprezentují indexy mapovací matice M (body na 3D modelu jsou zvoleny tak aby odpovídaly rozmístění ve 2D tvaru). Hodnoty v matici X lineárně narůstají zleva doprava. Hodnoty matice Y lineárně narůstají zdola nahoru. Hodnoty v matici Z zobrazují hloubku (neblíže je nos).

Referenční 3D model byl jednorázově vybrán z Face Recognition Grand Challange databáze (FRGC)[94]. Zvolen byl snímek s neutrálním výrazem, ukázka na obrázku 4.7. FRGC databáze byla natočena pomocí zařízení využívající techniku strukturovaného světla. Referenční 3D model je reprezentován maticemi **X**, **Y**, **Z**

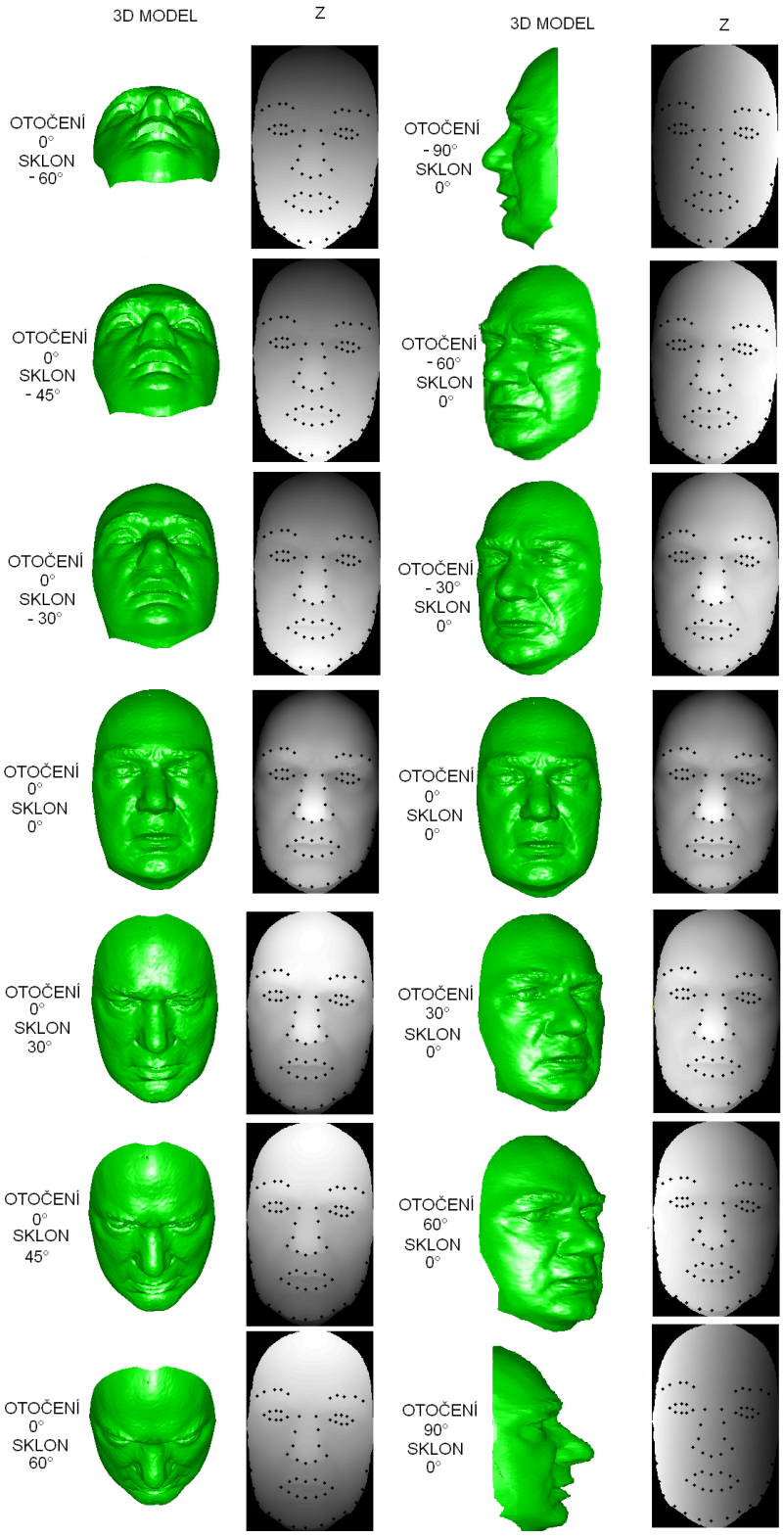
$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ x_{12} & & \\ \vdots & \ddots & \\ x_{m1} & & x_{mn} \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} y_{11} & \dots & y_{1n} \\ y_{12} & & \\ \vdots & \ddots & \\ y_{m1} & & y_{mn} \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} z_{11} & \dots & z_{1n} \\ z_{12} & & \\ \vdots & \ddots & \\ z_{m1} & & z_{mn} \end{bmatrix}$$

Prvky v maticích popisují body tak, že prvky obsahují souřadnice x,y,z 3D modelu v čelním pohledu. Díky metodě strukturovaného světla má 3D model strukturu, kde souřadnice v maticích **X** a **Y** lineárně narůstají a souřadnice v matici **Z** reprezentují jejich hloubku (viz obrázek 4.7).

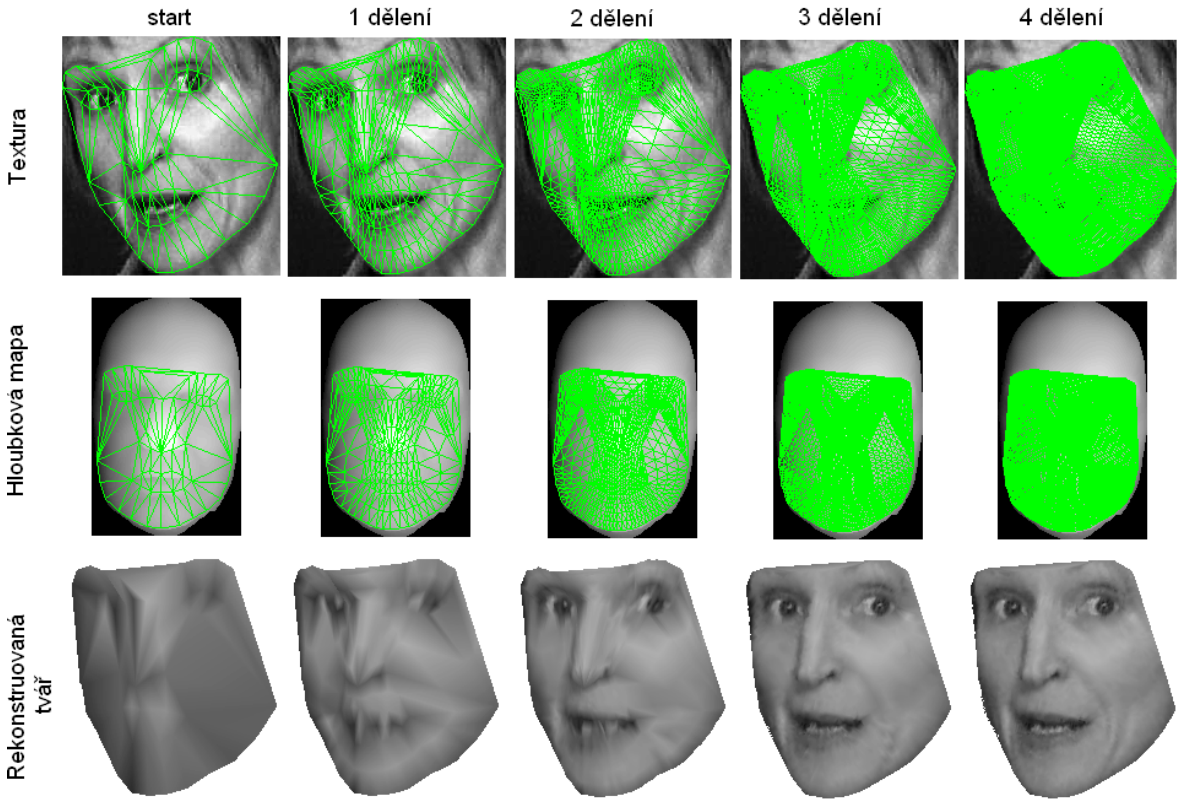
Manipulace s každým bodem 3D modelu, v našem případě natočení, je provedeno přenásobením rotační maticí R

$$\mathbf{R} = \begin{bmatrix} \cos\alpha + u_x^2(1-\cos\alpha) & u_x u_y(1-\cos\alpha) - u_z \sin\alpha & u_x u_z(1-\cos\alpha) + u_y \sin\alpha \\ u_x u_y(1-\cos\alpha) + u_z \sin\alpha & \cos\alpha + u_y^2(1-\cos\alpha) & u_y u_z(1-\cos\alpha) - u_x \sin\alpha \\ u_x u_z(1-\cos\alpha) - u_y \sin\alpha & u_y u_z(1-\cos\alpha) + u_x \sin\alpha & \cos\alpha + u_z^2(1-\cos\alpha) \end{bmatrix}$$

kde $u = \begin{bmatrix} u_x & u_y & u_z \end{bmatrix}$ reprezentuje osu otáčení a α hodnotu úhlu otočení (např. pro otočení kolem osy x se $u = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$).



Obrázek 4.8: Referenční 3D model a matice Z (hloubková mapa) pro různá natočení. V čelním pohledu jsou nejblíže body na nose (bíla barva). Při sklonu hlavy dopředu (v druhém sloupci) je vidět že body na čele jsou blíže než body na bradě. Při otáčení vpravo/vlevo (čtvrtý sloupec) jsou blíže body v pravé/v levé části tváře.



Obrázek 4.9: Vizuální ukázka odhadu hloubky 2D snímku pomocí vektorového dělení.

K zajištění korespondence mezi 2D tvarem a 3D modelem je jednorázově vytvořena mapovací matice M

$$\mathbf{M} = \begin{bmatrix} i_1 & i_2 & \dots & i_k \\ j_1 & j_2 & \dots & j_k \end{bmatrix}$$

indexy i a j určující řádek a sloupec v maticích **X,Y** a **Z**. Indexy jsou zvoleny tak, aby odpovídaly klíčovým bodům tvaru tváře z 2D snímku, ukázka je na obrázku 4.7. Při natáčení 3D modelu se mění hodnota dané souřadnice odpovídající 3D vrcholu 3D modelu. Při odhadu hloubky (2. krok) pracujeme jen s maticí **Z**. Ta reprezentuje hloubkovou mapu pro aktuální natočení tváře. Na obrázku 4.8 je ukázka změny hodnot souřadnic matice **Z** pro různá natočení tváře.

Pro odhad hloubky máme k dispozici 2D tvar s aktuálním natočením tváře, matici **Z** v aktuálním natočení tváře (hloubková mapa) a mapovací matici **M** (definující 2D tvar v hloubkové mapě). K získání hloubky pro každý bod uvnitř 2D tvaru tváře je nutné získat velké množství korespondenčních bodů mezi 2D snímkem a hloubkovou mapou. Tvar tváře **T** a matice **Z** jsou reprezentovány stejnou trojúhelníkovou sítí (obrázek 4.9 první sloupec). Ke zvýšení počtu korespondenčních bodů jsou oba tvary současně děleny pomocí vektorového dělení trojúhelníků. Dělením obou tvarů zároveň zajistíme ke každému novému bodu z 2D

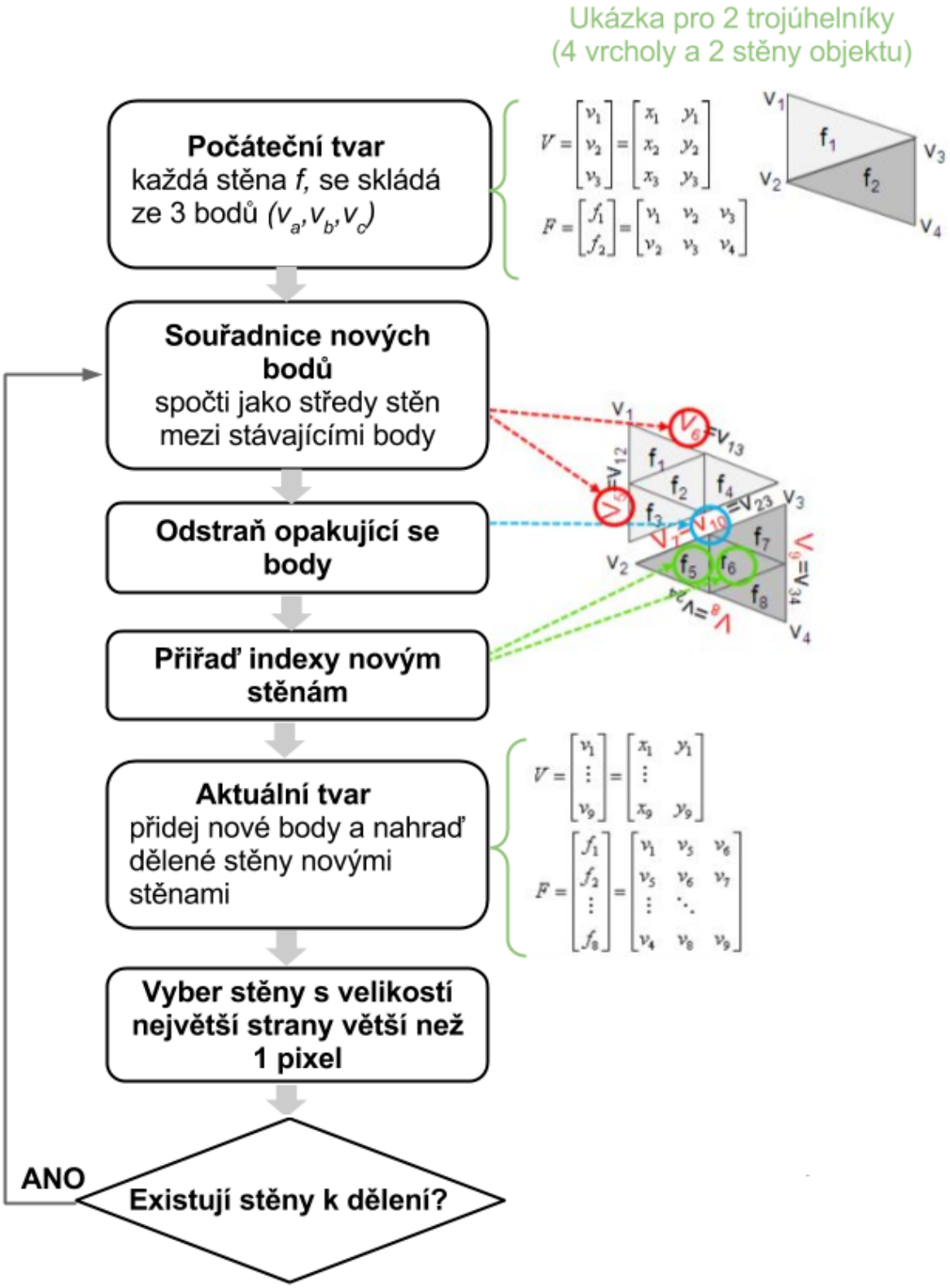
tvaru (souřadnice x a y) přibližnou hloubku z matice Z (souřadnice z).

Na obrázku 4.11 je naznačen postup vektorového dělení trojúhelníků. Vrcholy trojúhelníků v_{Ti} reprezentují 2D tvar ve snímku a v_{Zi} reprezentují 2D tvar v hloubkové mapě, kde i odpovídá aktuálnímu počtu bodů. Vrcholy jsou uloženy v maticích $\mathbf{V_T}, \mathbf{V_Z}$, řádek obsahuje jeden bod. Každý trojúhelník se skládá ze tří bodů. Trojúhelník $f_i = [idx_{v_A}, idx_{v_B}, idx_{v_C}]$ je reprezentován pomocí indexů odkazujících se na řádek v maticích $\mathbf{V_T}, \mathbf{V_Z}$. Trojúhelník reprezentuje stěnu objektu. Vektorovým dělením trojúhelníků získáme pro každou stěnu čtyři nové stěny objektu. Začínáme-li s počtem 102 stěn (62 bodů) získané po jednom dělení 408 stěn (220 bodů) a po pěti děleních 3D model s 80700 stěn (44100 bodů).

Dělení trojúhelníkové sítě se řídí tvarem ve 2D snímku, respektována je velikost strany každé stěny. Děleny jsou pouze stěny, jejichž největší strana je větší než 1 pixel. V každém kroku vektorového dělení jsou nejprve spočteny souřadnice nových bodů. Nový bod je spočten jako střed strany trojúhelníku. Opakující se body jsou odstraněny a poté jsou novým stěnám objektu přiřazeny indexy bodů. Dělení trojúhelníků se provádí dokud existuje stěna se stranou o velikosti větší než 1 pixel. Pro tvář širokou 200 pixelů trvá vektorové dělení trojúhelníků méně než 100ms.

Přímá korespondence mezi body ve vstupním 2D snímku a hloubkové mapě (získaná dělením obou tvarů zároveň) je využita při rekonstrukci tváře. Každé souřadnici x,y je přiřazena hloubka z čímž vznikne 3D objekt tváře. Textura každého bodu odpovídá intenzitě 2D snímku na souřadnici x, y.

Obrázek 4.10: Ukázka interpolace textury v oblasti nosu. Snímek vpravo je předzpracovaný a připravený pro popis a klasifikaci výrazu.



Obrázek 4.11: Blokové schéma vektorového dělení trojúhelníkové sítě

Rekonstruovaný 3D objekt tváře je natočen na čelní pohled a promítnut do 2D prostoru. Pokud dochází k zákrytu části tváře díky natočení, je textura v daném místě interpolována z okolí, ukázka na obrázku 4.10. Pro otočení do stran lze texturu také doplnit ze symetrie tváře. Takto doplněná textura pouze zlepší vizuální vzhled 3D objektu tváře, následné rozpoznávání tváře nezíská novou informaci. Další ukázky 3D rekonstrukce jsou k dispozici na webu[1].

Navržená metoda je experimentálně ověřena na GEMEP-FERA databázi [115]. Pozice klíčových bodů byla vybrána vzhledem k datům v GEMEP-FERA databázi. Tato databáze obsahuje maximální natočení 30 stupňů do obou stran a 30 stupňů pro sklon. Rozmístění klíčových bodů pokrývá natočení až 45 stupňů pro otočení do stran a až 30 stupňů pro sklon hlavy. V případě většího natočení tváře dochází k zákrytu klíčových bodů a je nutné změnit rozmístění bodů definující tvar tváře jak ve 2D snímku tak v 3D modelu.

4.4 Shrnutí

Libovolné natočení tváře představuje v současnosti největší problém při rozpoznávání výrazu tváře. V této kapitole byla popsána řešení z oblasti FER systémů a identifikace osob, jak natočení tváře odstranit.

Stávající řešení pro FER systémy nabízejí řešení, jak odstranit vliv natočení pro geometrický tvar tváře i texturu. Zatímco metody pro geometrický tvar tváře byly experimentálně ověřeny, vliv normalizace textury tváře na rozpoznávání výrazu nebyl zatím prozkoumán. Systémy pro geometrický tvar mají úspěšnost rozpoznávání výrazu nezávislou na natočení tváře. Nejlepší systém je schopný rozpoznat výraz tváře v úhlu 0 až 45 stupňů pro otočení do stran a 0 až 30 stupňů pro sklon hlavy [97]. Na druhou stranu hodnocení úspěšnosti systémů pro normalizaci textury neobsahuje srovnání vůči systému bez normalizace tváře [59], nebo je zhodnocena pouze kvalita rekonstrukce tváře [120].

Stávající řešení pro identifikaci osob jsou schopná rekonstruovat 3D tvář z 2D snímku s libovolným natočení. Metody však pracují pouze s neutrálním výrazem. Pokud tvář obsahuje jiný výraz, je tento výraz během rekonstrukce potlačen.

Nově navržená metoda, popsaná v sekci 4.3 vychází z technik používaných při identifikaci osob. Výstupem je 3D objekt tváře včetně aktuálního výrazu tváře. Cílem metody je normalizace tváře na čelní pohled. Rekonstruovaná tvář je natočena na čelní pohled a body jsou ortogonální projekcí promítnuty z 3D do 2D prostoru. Takto normalizovaná tvář lze použít k rozpoznání výrazu.

V následující kapitole je experimentálně ověřena normalizace tváře na čelní pohled. Testován je vliv metody na úspěšnost rozpoznávání výrazu tváře pro GEMEP-FERA databázi [115]. Vytvořen byl základní systém pro rozpoznávání výrazu tváře z 2D dat. Systému jsou předkládány data normalizovaná buď na fixní pozici středu očí nebo data normalizovaná na čelní pohled s využitím 3D rekonstrukce.

[1] 3D ukázky https://picasaweb.google.com/104238338003370449555/3D02

Kapitola 5

Experimentální ověření normalizace tváře

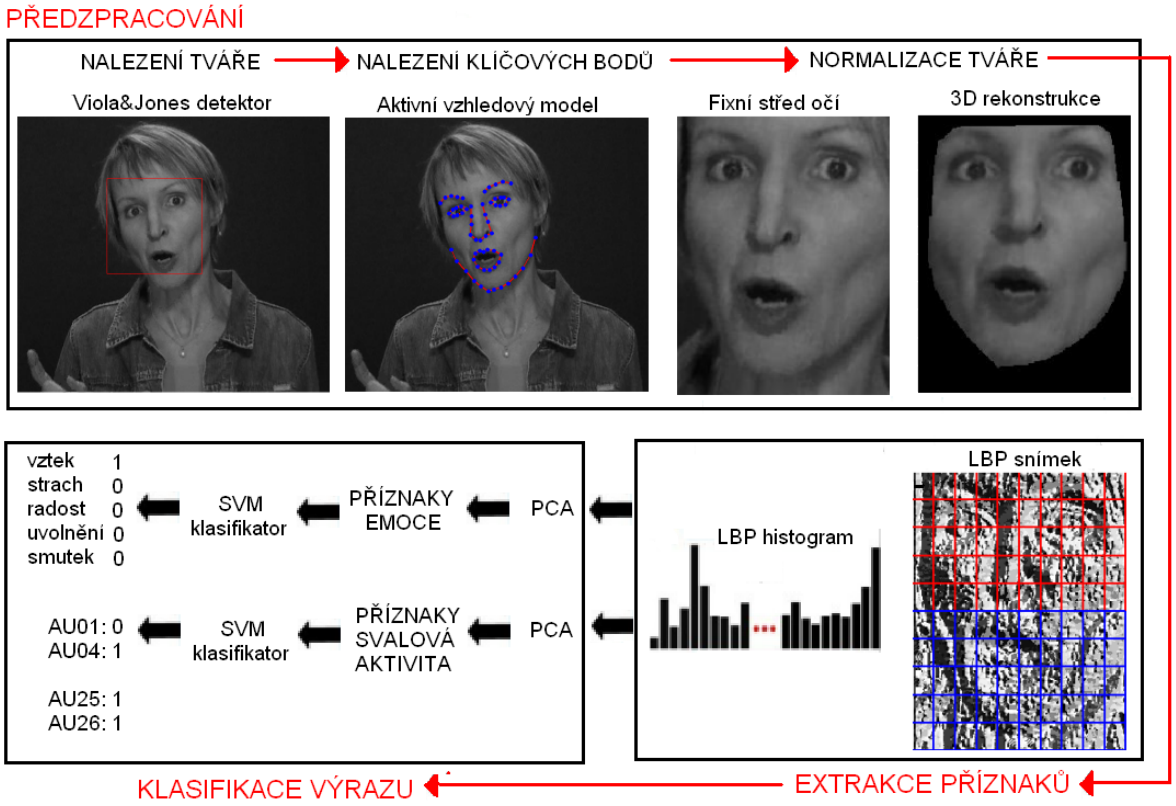
Metoda normalizace tváře na čelní pohled je experimentálně ověřena na datech z první soutěže rozpoznání výrazu tváře FERA 2011 (Facial Expression Recognition and Analysis Challenge [115]). Pro stanovení vlivu 3D rekonstrukce tváře z 2D snímku na rozpoznávání výrazu tváře jsem vytvořila FER systém podobný systému uvedenému na FERA 2011. Popis základního systému je v sekci 5.1.

Ověřen je vliv normalizace tváře na čelní pohled, systém nepracuje s hloubkou získanou z 3D rekonstrukce. Systému je předkládán buď snímek normalizovaný na fixní pozice středu očí (základní systém) nebo snímek normalizovaný na čelní pohled s využitím 3D rekonstrukce (systém s 3D rekonstrukcí). Ukázka výstupu pro obě normalizace je na obrázku 5.1.

Pro experimenty byla použita data a protokol testování ze soutěže FERA 2011 [115]. Porovnání úspěšnosti systému, využívajícího 3D rekonstrukci, bylo provedeno jak pro rozpoznávání kategorických emocí tak pro rozpoznání svalové aktivity. Data pro **rozpoznávání kategorických emocí** jsou rozdělena do pěti tříd: vztek, strach, radost, smutek a úleva. Čtyři z těchto emocí jsou považovány za základní emoce [27]. Pátá emoce, úleva, byla přidána k vytvoření rovnováhy mezi pozitivními a negativními emocemi. Emoce vyjadřující úlevu se zatím v rozpoznávání kategorických emocí neobjevila, představuje tak novou výzvu pro FER systémy. Každé video je zařazeno do jedné z pěti tříd. Podrobný popis rozdělení dat a dosažených výsledků pro rozpoznávání kategorických emocí je v sekci 5.2 Data pro **rozpoznání svalové aktivity** jsou anotována pomoci FACS [24]. Vybráno bylo 12 AU jednotek nejčastěji se zobrazujících v datech. Jejich seznam je v tabulce 5.5. Anotován je každý snímek videa. Pro danou jednotku je hodnota anotace '1' v případě, že je jednotka obsažena, '0' jednotka není obsažena. Popis dat a dosažené výsledky jsou v sekci 5.3. Přehledné shrnutí výsledků mezi základním systémem a systémem využívajícím 3D rekonstrukci tváře je v sekci 5.4.

5.1 Systém pro rozpoznání výrazu tváře

Obrázek 5.1 zobrazuje blokové schéma systému pro rozpoznání výrazu tváře vytvořeného k experimentálnímu ověření normalizace tváře na čelní pohled s využitím 3D rekonstrukce tváře z 2D snímku. Systém vytvořený v této práci je podobný základnímu systému ze soutěže



Obrázek 5.1: Blokové schéma systému pro rozpoznání výrazu tváře. Základní systém používá snímek s normalizací na fixní pozici středu očí. Systém s 3D rekonstrukcí používá snímek s normalizací rekonstruované 3D tváře na čelní pohled. Pro rozpoznávání emocí je při extrakci příznaků použit celý snímek. Pro svalovou aktivitu je snímek rozdělen na horní a dolní část v závislosti na tom, která AU jednotka se rozpoznává, rozdělení viz sekce 5.3. Ukázka zobrazuje emoci vztek jedná se o 41 snímek z videa train_095.avi.

FERA 2011[1], vůči kterému porovnávali úspěšnost svých systémů účastníci soutěže. Systém se liší pouze v kroku předzpracování, respektive v kroku nalezení klíčových bodů a normalizace tváře. Zatímco systém FERA 2011 [115] použil k nalezení klíčových bodů detektor z knihovny OpenCV[2], systém v této práci používá metodu Aktivních Vzhledových Modelů (dále jen AAM) 3.1.3 z Matlab file exchange[3]. Na základě nalezených klíčových bodů je provedena normalizace tváře pro dvě různé metody. Jednou je tvář normalizována vzhledem k fixní pozici středu očí a podruhé je tvář normalizována na čelní pohled s využitím 3D rekonstrukce tváře z 2D snímku. Následná extrakce příznaků a klasifikace výrazu je stejná jako u systému FERA 2011[115]. Dle metody použité k normalizaci tváře se v této práci rozlišuji dva systémy.

[1] fera2011 (http://sspnet.eu/fera2011/)
[2] openCV (http://opencv.org/)
[3] icaam (http://www.mathworks.com/matlabcentral/fileexchange/32704-icaam-inverse-compositional-active-appearance-models)

- **Základní systém** - normalizace tváře na fixní pozice středu očí
- **Systém s 3D rekonstrukcí** - normalizace tváře na čelní pohled s využitím 3D rekonstrukce z 2D snímku.

Blok pro předzpracování snímku (horní řádek v obrázku 5.1) je stejný jak pro úlohu rozpoznávání emocí tak pro rozpoznávání svalové aktivity. Jednotlivé kroky pro předzpracování snímků videa jsou následující. Pro nalezení tváře je použita vlastní implementace Viola&Jones detektoru tváře schopná detekovat tvář natočenou v rozsahu $\pm 20°$ pro náklon a $\pm 45°$ pro sklon a otočení tváře. K nalezení pozice klíčových bodů je využita metoda AAM 3.1.3. Pro každého řečníka byl vytvořen jeden AAM model vytvořený z 62 manuálně označených klíčových bodů. Při hledání pozice klíčových bodů nebyl AAM model vždy úspěšný (ukázka na obrázku 5.2). Nepřesně nalezená pozice klíčových bodů by značně ovlivnila následnou normalizaci tváře. Pro každé video byla tedy spočtena chyba usazení AAM modelu. Chyba byla vypočtena jako střední kvadratická odchylka (MSE - mean square error) rozdílu mezi tvary ve dvou po sobě jdoucích snímcích. Tvary byly nejprve zarovnány na fixní střed očí. Teprve poté byl spočten rozdíl mezi jednotlivými body tvaru a z těchto rozdílů byla spočtena střední kvadratická odchylka. Pokud hodnota MSE překročila experimentálně stanovený práh $T = 0,3\, pixel^2$ byl snímek vyřazen. Celkem bylo vyřazeno 15% snímků jak z trénovacích tak i testovacích dat.

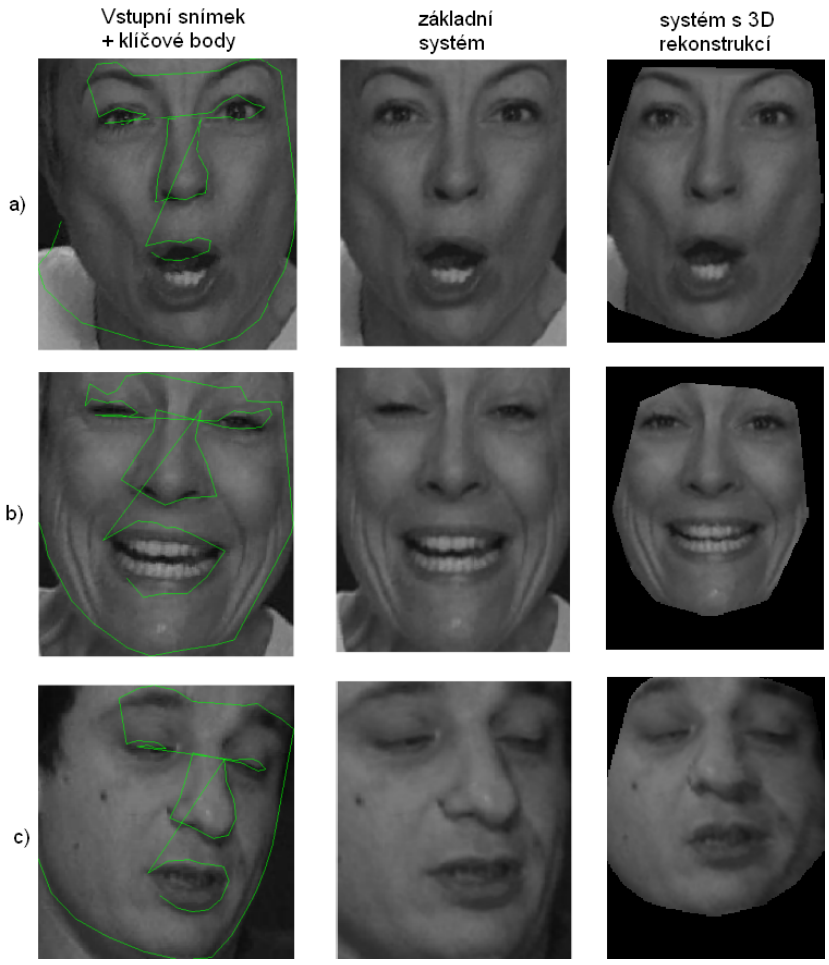
Pro základní systém je tvář normalizována na fixní pozice středu očí. Nejprve je odstraněn náklon tak, aby přímka mezi středy očí byla paralelní s x-ovou osou. Poté je vzdálenost mezi očima nastavena na 100 pixel a nakonec je vyříznuta tvář o velikosti 200x200 pixel. Pro systém s 3D rekonstrukcí se nejprve odhadne natočení tváře v prostoru. Poté se pro každý bod 2D snímku odhadne hloubka. Nakonec je tvář natočena do čelního pohledu a promítnuta zpět do 2D. Tvář o velikosti 200x200 pixel je vyříznutá tak, aby vzdálenost mezi středy očí byla 100 pixel. Podrobný popis 3D rekonstrukce tváře je v sekci4.3.

Pro extrakci příznaků je použita implementace lokálních binárních vzorů [4], metoda je popsána v sekci 3.2.2. Pro klasifikaci výrazu, přesněji selekci příznaků pomocí metody hlavních komponent a klasifikátor Support Vector Machine, je použit pattern recognition toolbox [5]. Bloky extrakce příznaků a klasifikace výrazu (spodní řádek obrázku 5.1) mají rozdílné nastavení pro úlohu rozpoznávání emocí a rozpoznávání svalové aktivity.

Pro rozpoznávání kategorických emocí je výstupní snímek z předpracování rozdělen na 10x10 bloků a vektor příznaků je vytvořen poskládání LBP histogramů pro jednotlivé bloky. Ke snížení dimenze vektoru příznaků je použita metoda hlavních komponent, použito je tolik vlastních vektorů, aby výsledná transformace popsala 90% rozptylu originálních dat. Každý snímek videa je zařazen do jedné z pěti tříd (vztek, strach, radost, smutek a úleva.). SVM klasifikátor je tedy trénován pro úlohu jeden-proti-všem (tj. videa obsahující vztek jsou brána jako pozitivní vzory a všechny ostatní emoce jsou považovány za negativní vzory). Pro každou emoci existuje jeden SVM klasifikátor. Výstup klasifikátorů $y_j^e \epsilon \{-1, 1\}$ rozhoduje o výskytu emoce e pro každý snímek j testovacího videa. Výsledné přiřazení identifikátoru Y pro celé video závisí na nejvyšším počtu snímků klasifikovaných jako emoce e.

$$Y = \underset{e}{argmax} \sum_{j=1}^{n} y_j^e \tag{5.1}$$

[4] LBP - http://ljk.imag.fr/membres/Bill.Triggs/src/
[5] PRT - http://www.newfolderconsulting.com/prt



Obrázek 5.2: Normalizaci tváře, ukázka nevýhod jednotlivých metod. a,b) Díky nepřesně nalezené pozice klíčových bodů je výsledek po normalizaci na čelní pohled pro systém s 3D rekonstrukcí nevhodný k rozpoznávání výrazu. V prvním řádku je kromě tváře zahrnuta i část oblečení, v druhém řádku zase chybí obočí. c) Metoda normalizace na fixní pozici středu očí dokáže odstranit pouze natočení v x-ové ose, neumí se vyrovnat s natočením do stran a se sklonem hlavy.

Pro rozpoznávání svalové aktivity je výstupní snímek z předzpracování také rozdělen na 10x10 bloků. Vektor příznaků je poskládán z LBP histogramů. Pro AU jednotky popisující svalovou aktivitu horní části tváře je použito prvních pět řádků (červené bloky LBP snímku ve obrázku 5.1), pro svalovou aktivitu dolní části tváře je použito spodních pět řádek (modré bloky LBP snímku ve obrázku 5.1). Dimenze vektoru příznaků je snížena pomocí metody hlavních komponent. Vybráno je tolik vlastních vektorů, aby výsledná transformace popsala 95% rozptylu originálních dat. Příznaky jsou poté normalizovány na interval $[-1, 1]$. SVM klasifikátor s RBF kernelem je trénován pro úlohu jeden-proti-všem. Celkem existuje 12 klasifikátorů, seznam AU jednotek je v tabulce 5.5. Na každý snímek videa je aplikováno všech 12 klasifikátorů, výstupem každého klasifikátoru je buď '1' AU jednotka obsažena, nebo '0' AU jednotka není obsažena (v jednom snímku je zachycen výraz skládající se i z několika AU jednotek).

Tabulka 5.1: Data z FERA 2011 rozdělení kategorických emocí na trénovací a testovací videa. Čísla uvádí počet videí, $N_{celkem} = 289$, $N_{trénovací} = 155$, $N_{testovací} = 134$. ZS označuje známý subjekt, obsažený v trénovací i testovací sadě. NS označuje neznámý subjekt, který se objevuje pouze v testovací sadě.

Emoce	Slovní definice	Trén. sada	ZS-Test. sada	NS-Test. sada	$\sum$
Vztek	Extrémní nespokojenost způsobená nepřátelským chováním jiného subjektu	32	13	14	59
Strach	Být vystaven bezprostřednímu nebezpečí ohrožující přežití nebo psychickou pohodu	31	10	16	57
Radost	Cítit se unešeně díky skvělé věci, která se stala nečekaně	30	11	19	60
Úleva	Cítit se klidně po absolvování nepříjemné nebo dokonce nebezpečné situace	31	10	15	56
Smutek	Cítit se na nic díky ztrátě osoby, místa nebo věci	31	10	16	56

5.2 Výsledky rozpoznání kategorických emocí

Rozpoznávač kategorických emocí má rozhodnout o zařazení emoce do jedné z pěti tříd. Každé video má přiřazen identifikátor emoce $e \epsilon E$, kde $E=\{vztek,\ strach,\ radost,\ smutek,\ úleva\}$. Výraz tváře se v průběhu videa mění, není však ohraničen neutrálním výrazem. Všechny snímky videa sdílejí stejný identifikátor emoce. Z toho důvodu jsou všechny snímky videa využité k trénování i testování základního systému.

Tabulka 5.1 obsahuje podrobný přehled rozdělení videí na testovací a trénovací sadu pro jednotlivé emoce. Celkem je k dispozici 289 videí pro 10 různých subjektů. Pro trénování bylo vybráno 155, pro testování 134 videí. Přibližně v 17% dat subjekty vyslovovali samohlásku 'aaa', ve zbývající datech vyslovovali dvě pseudo-lingvistické sekvence fonémů. Trénovací sada obsahuje 7 subjektů se 3 až 5 příklady pro každou emoci předvedenou jedním subjektem. Testovací sada obsahuje 6 subjektů. Z toho polovina subjektů nebyla obsažena v trénovací sadě. Každý subjekt předvedl 3 až 10 příkladů pro jednotlivé emoce.

Cílem rozpoznání kategorických emocí je určení, která z pěti emocí je obsažena ve videu. Kvalita rozpoznání je měřena pomocí F1-measure definované následovně

$$F1 = 2\frac{P \cdot S}{P+S}, \quad \begin{matrix} P = \frac{tp}{tp+fp} \\ S = \frac{tp}{tp+fn} \end{matrix}$$

$$\begin{matrix} tp = & počet\ správně\ zařazených\ pozitivních\ vzorů \\ fp = & počet\ nesprávně\ zařazených\ pozitivních\ vzorů \\ fn = & počet\ nesprávně\ zařazených\ negativních\ vzorů \end{matrix} \tag{5.2}$$

kde P vyjadřuje poměr správně zařazených pozitivních vzorů a S vyjadřuje úspěšnost klasifikátoru při nalezení pozitivních vzorů. F1-measure je nejprve spočtena pro každou emoci a poté je spočten průměr přes všech 5 emocí.

Tabulka 5.2 obsahuje výsledky pro kategorické rozpoznávání emocí dle F1-measure. Tabulka porovnává úspěšnost mezi základním systémem a systémem s 3D rekonstrukci. Výsledky jsou rozděleny na tři části. První část obsahuje srovnání pro testovací sadu obsahující známé subjekty (obsažené i v trénovací sadě), druhá část obsahuje srovnání pro testovací sadu obsahující neznámé subjekty (neobsažené v trénovací sadě) a poslední část srovnává výsledky pro celou testovací sadu.

Tabulka 5.2: Výsledky dle F1-measure [%] (vzorec 5.2) pro rozpoznávání kategorických emocí. Srovnání mezi základním systémem a systémem s 3D rekonstrukci (označeným indexem $_{3D}$). ZS označuje známý subjekt, obsažený v trénovací i testovací sadě. NS označuje neznámý subjekt, který se objevuje pouze v testovací sadě. $\sum$ jsou výsledky pro celou testovací sadu.

Emoce	ZS	ZS_{3D}	NS	NS_{3D}	$\sum$	$\sum_{3D}$
vztek	86	**96**	58	**62**	71	76
strach	84	**100**	12	**61**	50	79
radost	88	**90**	59	**64**	68	72
smutek	78	**95**	**42**	24	67	70
úleva	89	**91**	48	**57**	57	62
průměr	85	**94**	44	**54**	63	73

Systém s 3D rekonstrukcí dosahuje lepších výsledků pro rozpoznávání všech kategorických emocí pro známý subjekt a většiny kategorických emocí pro nezná-mý subjekt. Horšího výsledku dosahuje pouze v případě rozpoznávání kategorické emoce smutek pro neznámý subjekt. Dle analýzy pomocí matice záměn uvedený v tabulce 5.4 je smutek pro neznámý subjekt klasifikován jako úleva nebo vztek. Pro případy, kdy je smutek klasifikován jako úleva je tento výsledek způsoben díky odstranění natočení tváře v prostoru, respektive sklonu tváře. Výraz pro úlevu a smutek je velmi podobný. Pokud klasifikátor získá vektor příznaků s odstraněným natočením tváře dochází tak k záměně těchto dvou výrazů. K záměně mezi emocemi vztek a smutek dochází díky podobnosti emoce smutek s neutrálním výrazem. Trénovací sada neobsahuje anotaci pro neutrální výraz a testovací sada pro emoci vztek často obsahuje snímky s neutrálním výrazem. Tento závěr potvrzuje i matice záměn pro základní systém, kdy je emoce smutek daleko častěji klasifikována jako vztek (celkem v 8 z 15 videí).

Základní systém dosahuje v průměru 85% úspěšnosti v rozpoznávání kategorické emoce pro známý subjekt. Pro neznámý subjekt dosahuje u většiny emocí úspěšnosti rozpoznávání kolem 52%. Nejhoršího výsledku rozpoznávání dosahuje emoce strach pro neznámý subjekt. Dle analýzy pomocí matice záměn 5.4 je strach klasifikován jako radost nebo vztek. Přitom emoce strach pro neznámý subjekt obsahuje 91% dat v čelním pohledu, viz obrázek 5.6. Dle srovnání s ostatními emocemi je to jediná emoce s tolika daty pro čelní pohled. Oproti tomu dosahuje systém s 3D rekonstrukcí o 50% vyšší úspěšnosti při klasifikaci. Tento výsledek je způsoben odstraněním natočení z emocí radost a vztek, díky tomu nedochází tak často k záměně s emocí strachu.

Z tabulek 5.3 a 5.4 obsahující matici záměn pro jednotlivé emoce je vidět, že systém s 3D rekonstrukcí daleko častěji zařadí emoce do správné kategorie. Pro známý subjekt je pro systém s 3D rekonstrukcí správně zařazeno 51 z 54 příkladů emocí, oproti základnímu systému kdy je zařazeno 46 z 54. Pro neznámý subjekt je pro systém s 3D rekonstrukcí správně zařazeno 50 z 80 příkladů emocí, oproti 36 z 80 pro základní systém. Systém s 3D rekonstrukci dosahuje dle F1-measure v průměru zlepšení rozpoznávání výrazu o 10%.

Tabulka 5.3: Matice záměn pro známý subjekt (ZS)

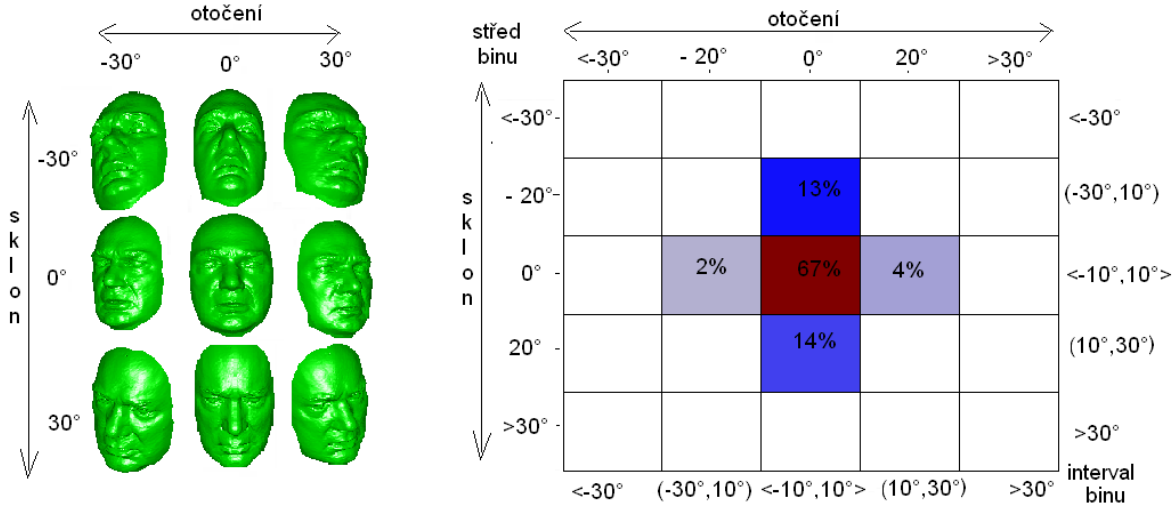
	základní systém	predikovaná emoce				
		vztek	strach	radost	smutek	úleva
skutečná	vztek	**12**	0	1	0	0
	strach	1	**8**	1	0	0
	radost	0	0	**11**	0	0
	smutek	2	1	0	**7**	0
	úleva	0	0	1	1	**8**

	systém s 3D rekon.	predikovaná emoce				
		vztek	strach	radost	smutek	úleva
	vztek	**13**	0	0	0	0
	strach	0	**10**	0	0	0
	radost	0	0	**9**	0	2
	smutek	1	0	0	**9**	0
	úleva	0	0	0	0	**10**

Tabulka 5.4: Matice záměn pro neznámý subjekt (NS)

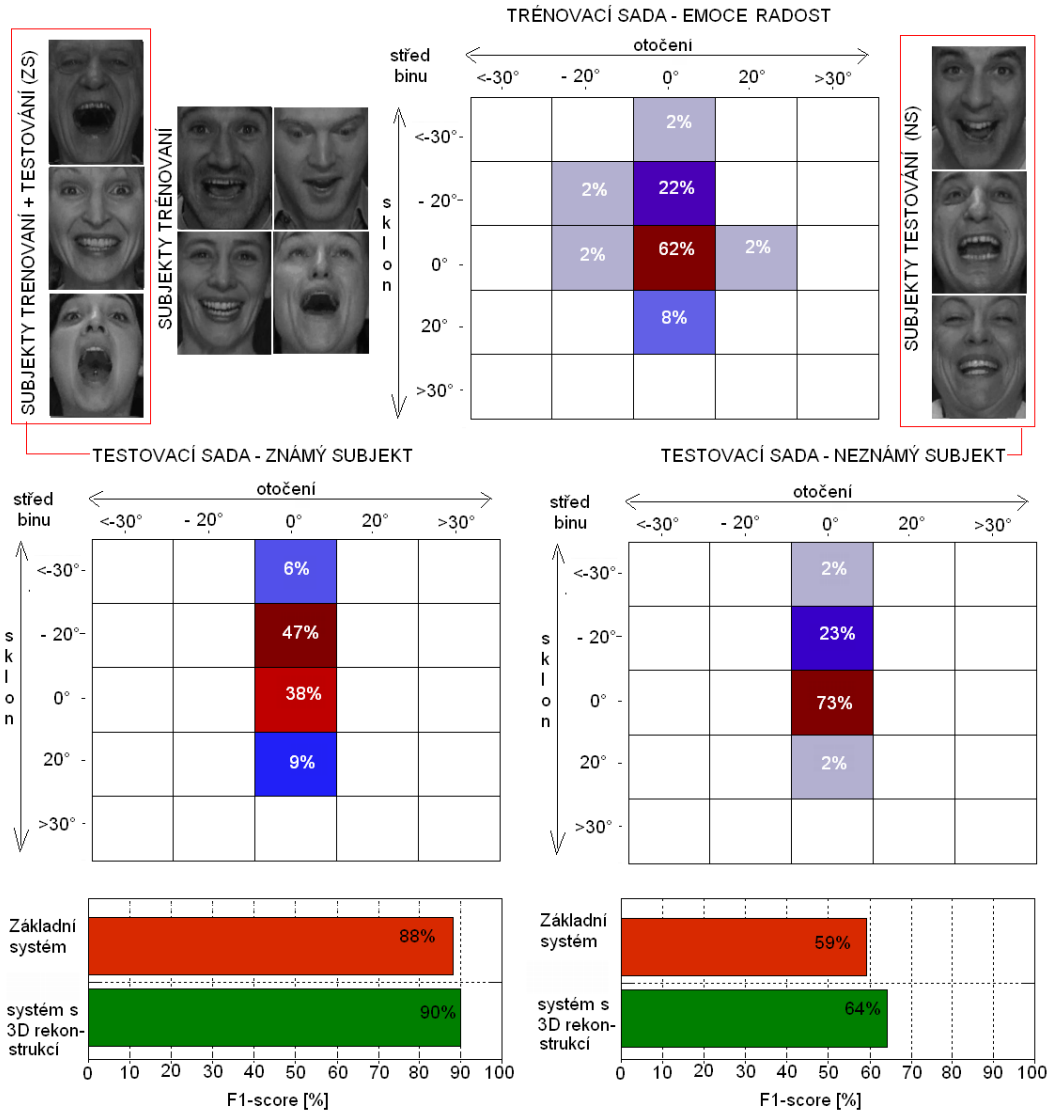
	základní systém	predikovaná emoce				
		vztek	strach	radost	smutek	úleva
skutečná	vztek	**9**	0	5	0	0
	strach	4	**1**	10	1	0
	radost	2	0	**16**	0	1
	smutek	8	0	2	**5**	0
	úleva	2	0	6	3	**5**

	systém s 3D rekon.	predikovaná emoce				
		vztek	strach	radost	smutek	úleva
	vztek	**13**	0	1	0	0
	strach	0	**7**	8	0	1
	radost	1	0	**18**	0	0
	smutek	6	0	0	**2**	7
	úleva	0	0	6	0	**10**

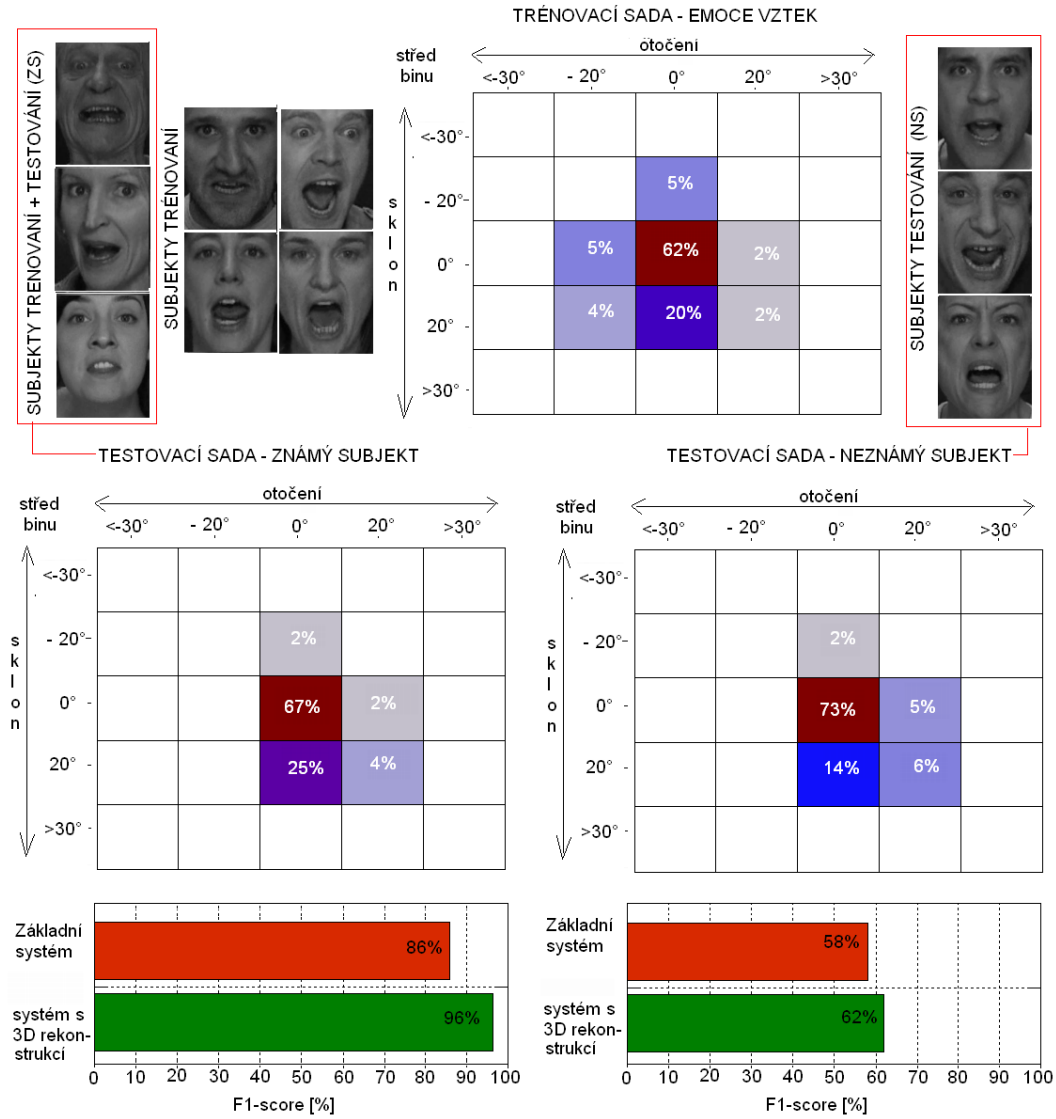


Obrázek 5.3: Rozložení dat vzhledem k natočení tváře v prostoru pomocí 2D histogramu. Vlevo je vizuální ukázka pro různé kombinace sklonu a otočení tváře. Vpravo je rozložení dat pro trénovací sadu rozpoznávání kategorických emocí. Velikost každého binu je 20° pro sklon i otočení tváře. Celkem 67% dat je v čelním pohledu ±10°, zbývajících 33% dat obsahuje natočení tváře větší než 10°. Trénovací sada obsahují pouze 6% dat pro otočení tváře vpravo a vlevo. Daleko více dat, 27%, obsahuje subjekty s pohybem tváře nahoru a dolů.

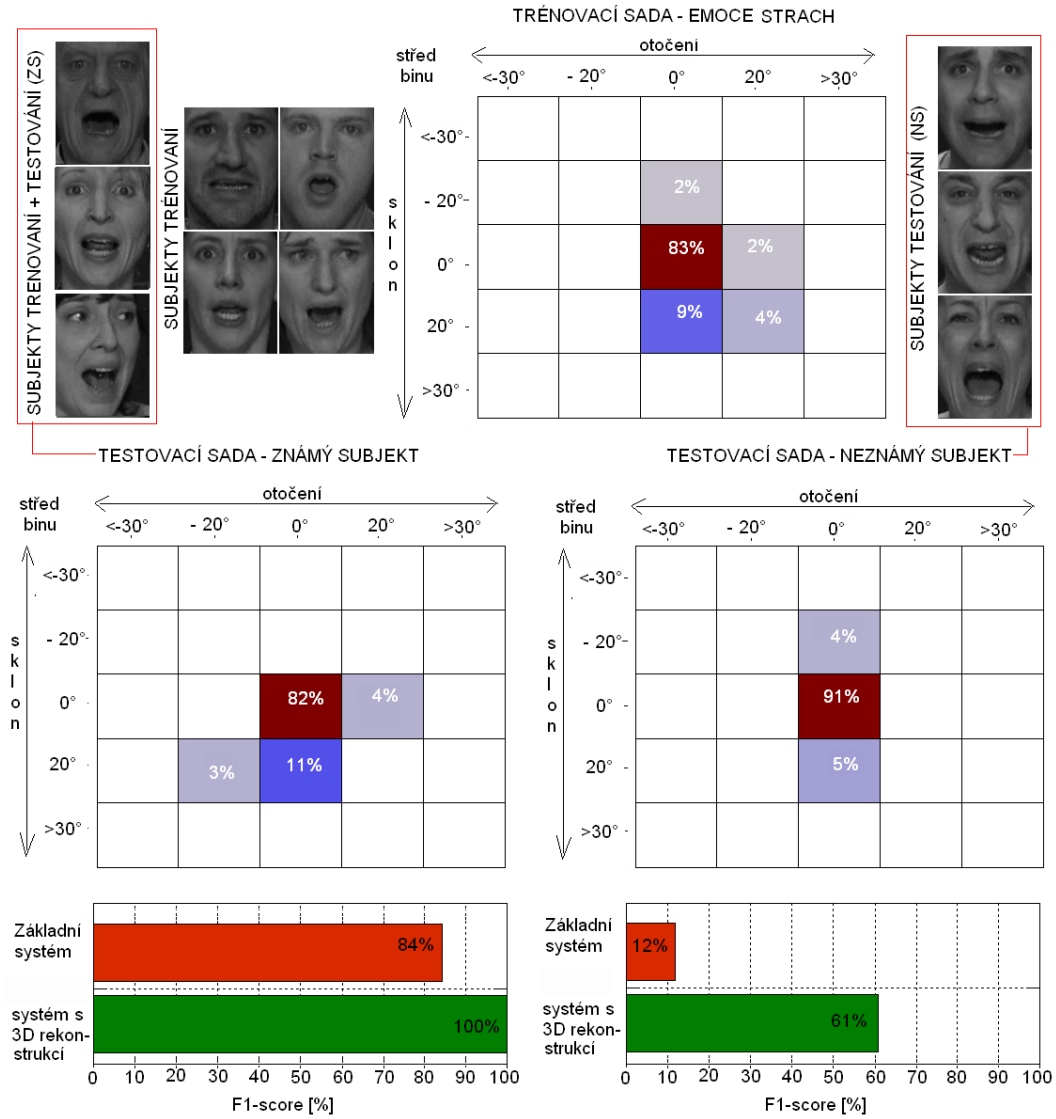
Výsledky pro jednotlivé emoce jsou podrobně zobrazeny na obrázcích 5.5 až 5.8. Pro každou emoci je zobrazeno rozložení dat vzhledem k natočení tváře v prostoru zvlášť pro známý i neznámý subjekt. Ukázka na obrázku 5.3. Rozložení je reprezentováno 2D histogramem pro sklon a otočení tváře v prostoru. Náklon tváře není uvažován, neboť lze snadno odstranit zarovnáním přímky mezi středy očí paralelně s x-ovou osou.



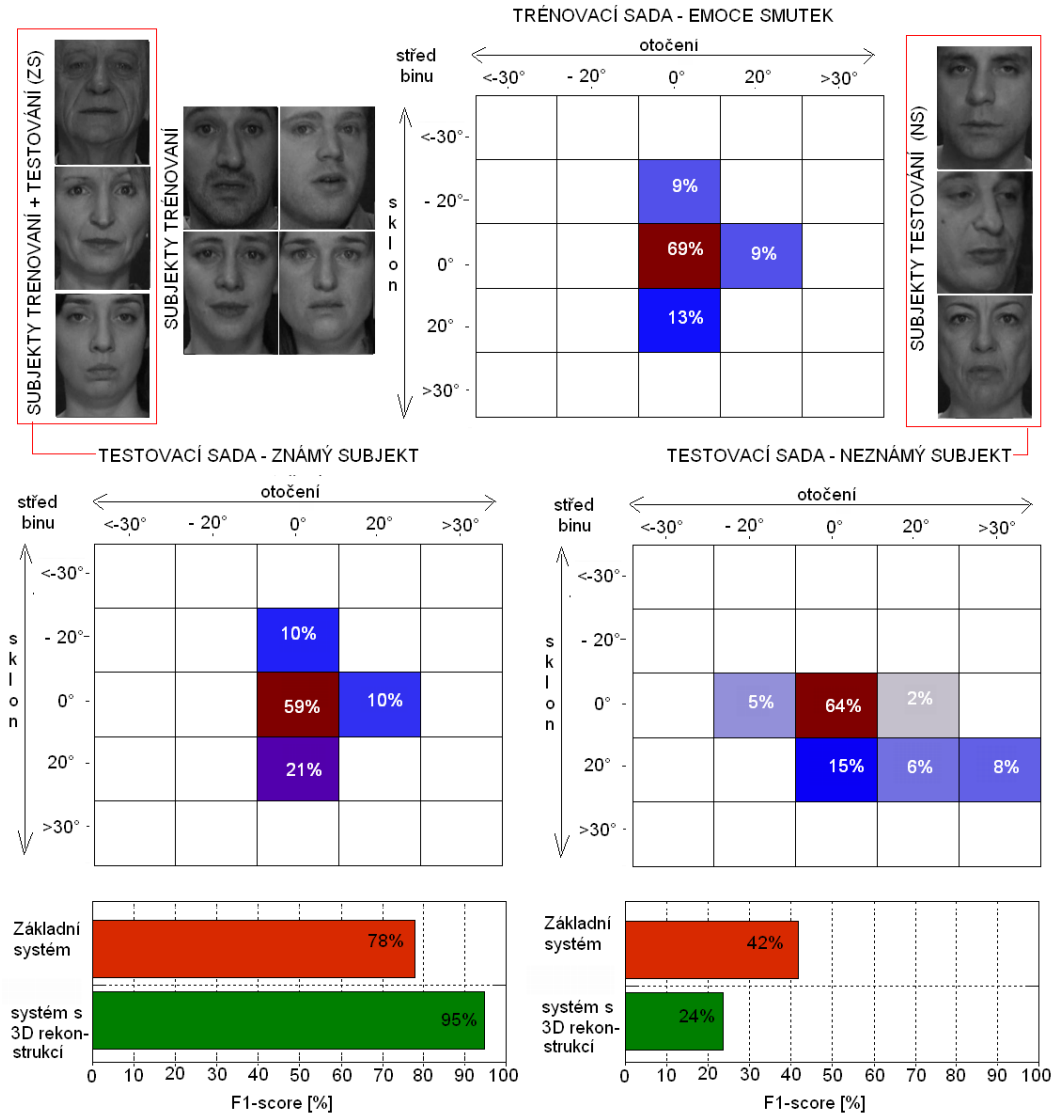
Obrázek 5.4: Projev radosti je vizuálně dobře viditelný (zvednuté koutky úst, otevřená ústa, zvednuté vnitřní i vnější koutky obočí). Subjekty při projevu často zaklání hlavu dozadu. Při rozpoznání výrazu jsou oba systémy velmi úspěšné a dosahují shodně správné detekce ve 27 z 30 případů. Úspěšnost pro neznámý subjekt je snížena díky nadbytečné predikci emoce radost, viz tabulka 5.4. Nejčastěji dochází k záměně s emocí strach a to v 8 případech pro systém s 3D rekonstrukcí a v 11 případech pro základní systém.



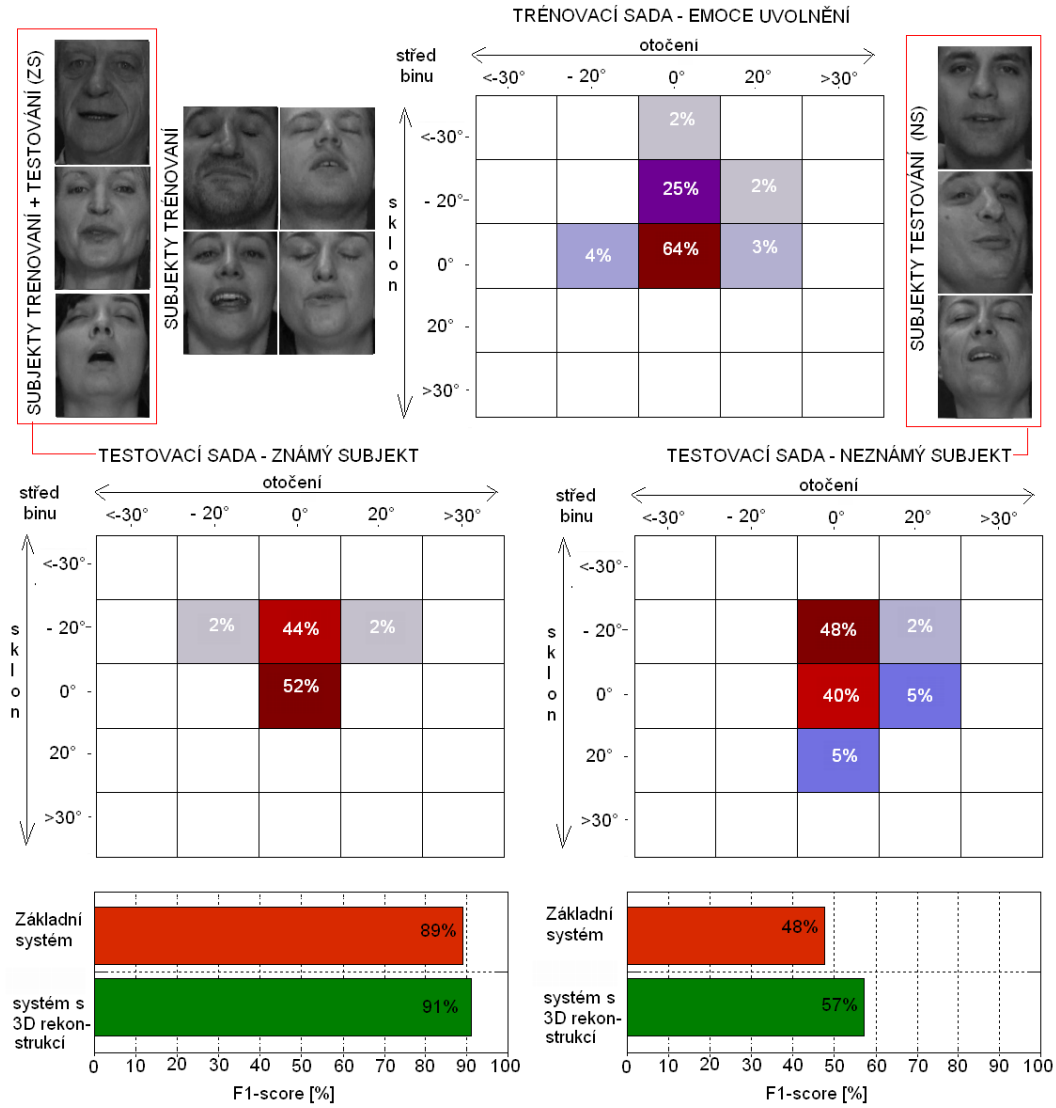
Obrázek 5.5: Emoce vzteku má velmi silný vizuální projev (zvednuté obočí, otevřená ústa, vrásky u kořene nosu), trvá však velmi krátce. Proto velké množství snímků videa pro emoci vztek obsahuje neutrální výraz tváře. Při rozpoznávání výrazu je emoce vzteku ve většině případů správně rozpoznána. Dle matic záměn pro známý 5.3 i neznámý 5.4 subjekt je emoce vzteku pro systém s 3D rekonstrukcí rozpoznána správně ve 26 z 27 případů, pro základní systém je rozpoznána ve 21 z 27 případů. Při trénování klasifikátoru jsou všechny snímky považovány za emoci vztek. Díky velkému počtu snímků s neutrálním výrazem dochází často k záměně s jinými emocemi. Pro systém s 3D rekonstrukcí je emoce vztek nadbytečně klasifikována v 8 případech, pro základní systém dokonce v 19 případech. Úspěšnost obou systémů pro rozpoznání emoce vztek je snížena díky nesprávné klasifikaci emoce vztek, respektive záměně za jinou emoci.



Obrázek 5.6: Emoce strach je vizuálně výrazná (otevřená ústa, doširoka otevřené oči) s dlouhým trváním. Většina snímků obsahuje vyděšený výraz. Subjekt hledí přímo na kameru s minimálním natočením v prostoru. Dle matic záměn 5.3 a 5.4 je výraz strachu správně rozpoznán v 17 z 28 případů pro systém s 3D rekonstrukcí a v 9 z 28 případů pro základní systém. Nejnižší úspěšnost má základní systém pro neznámý subjekt a to jen 12%, emoce je správně rozpoznána pouze v jednom případě. K záměně dochází s emocemi vztek a radost. Pro systém s 3D rekonstrukcí také dochází k záměně emoce strach za radost. Obě emoce jsou velmi podobné a odlišuje je pouze dynamika pohybu hlavy. Zatímco v případě strachu je hlava strnulá, tak v případě radosti se subjekt zaklání dozadu. Pokud nebude zahrnuta dynamika pohybu hlavy, je těžké tyto dvě emoce odlišit.



Obrázek 5.7: Emoce smutek je vizuálně nevýrazná (snížené koutky obočí a úst) s minimálním pohybem hlavy. Navíc má velkou podobnost s emocí úlevy a neutrálním výrazem. Subjekty při projevu emoce smutek často hledí mimo kameru. Systém s 3D rekonstrukcí správně rozpozná smutek v 11 z 25 případů, základní systém ve 12 z 25 případů. Úspěšnost obou systémů je snížena díky záměně s jinými emocemi a to převážně pro neznámý subjekt. Pro systém s 3D rekonstrukcí je namísto smutku predikována emoce úleva nebo vztek 5.4. Výraz tváře pro emoce smutek a úleva je velmi podobný. Odlišnost výrazu je dána pouze natočením tváře v prostoru. Díky tomu že systém s 3D rekonstrukcí odstraní vliv natočení dochází k nesprávnému zařazení emoce do skupiny úleva. K záměně mezi emocí smutek a vztek dochází u obou systémů. Problém je velké množství neutrálních snímků pro emoci vztek, které nejsou anotovány jako neutrální výraz. Pokud nebude zahrnuta dynamika hlavy není možné odlišit emoci smutku od uvolnění.



Obrázek 5.8: Emoce úleva je velmi podobná emoci radost a smutek. Systém s 3D rekonstrukcí správně rozpozná výraz tváře v 20 z 26 případů, základní systém v 13 z 26 případů. Úspěšnost systému pro neznámý subjekt je nejvíce ovlivněna nesprávným rozpoznáním emoce radosti. K záměně dochází díky velké podobnosti obou emocí. Subjekty se totiž při provádění emoce usmívají a zaklání hlavu dozadu. Pro systém s 3D rekonstrukcí dochází také k záměně s emocí smutek, to je způsobeno odstraněním naklonění tváře. Dynamika pohybu hlavy je důležitá k odlišení emocí úleva a smutek.

Tabulka 5.5: Data z FERA 2011, rozdělení trénovacích a testovacích videí pro AU jednotky. ZS označuje známý subjekt, obsažený v trénovací i testovací sadě. NS označuje neznámý subjekt, který se objevuje pouze v testovací sadě.. Počet videí je $N_{celkem} = 158$, $N_{trénovací} = 87$, $N_{testovací} = 71$

AU	Slovní popis	Trén. sada	ZS-Test. sada	NS-Test. sada	$\sum$
1	zvednuté vnitřní koutky obočí	48	9	28	85
2	zvednuté vnější koutky obočí	48	12	21	81
4	pokles obočí	34	10	26	70
6	zvednuté líce	37	8	27	72
7	mhouření očí	43	14	30	87
10	zvednutý horní ret	48	13	21	82
12	zvednuté koutky rtů	56	16	33	105
15	pokleslé koutky rtů	30	6	11	47
17	zvednutá brada	49	14	31	94
18	našpulené rty	28	12	20	60
25	rozevřené rty	67	22	37	126
26	pokles čelisti	46	12	23	81

5.3 Výsledky rozpoznání svalové aktivity

Pro rozpoznání svalové aktivity je pro každý snímek rozhodnuto zda-li obsahuje jednu nebo více AU jednotek. Celkem se rozhoduje o 12 AU jednotkách, jejich seznam je v tabulce 5.5. Díky nezávislosti svalové aktivity mezi horní a dolní částí tváře jsou AU jednotky rozděleny do dvou skupin: AU pro horní část tváře $G_u = \{AU1, AU2, AU4, AU6, AU7\}$ a AU pro dolní část tváře $G_l = \{AU10, AU12, AU15, AU17, AU18, AU25, AU26\}$. Pro každou AU jednotku byl natrénován jeden binární SVM klasifikátor. Z trénovacích dat byly pro každou AU jednotku vybrány pozitivní a negativní vzory. Pozitivní vzor buď obsahuje samostatnou AU jednotku nebo AU jednotku v kombinaci s jinými AU jednotkami ze stejné skupiny (horní, dolní část tváře). Negativní vzor buď obsahuje neutrální výraz nebo zbývající AU jednotky ze stejné skupiny. Při výběru snímků pro trénování bylo každé video nejprve rozděleno do úseků obsahujících odlišnou kombinaci AU jednotek ze stejné skupiny. Úsek typicky trval několik snímků, pro trénování byl vybrán prostřední snímek úseku. Průměrně je pro každou AU jednotku k dispozici 70 pozitivních a 130 negativních vzorů.

Celkem je k dispozici 158 videí rozdělených na 87 videí pro trénovací sadu a 71 videí pro testovací sadu. Na všech videích vyslovují subjekty 2 pseudo-lingvistické fonémy, AU jednotky jsou tedy předváděny během řeči. Trénovací sada obsahuje 7 subjektů. Testovací sada obsahuje 6 subjektů, z toho se 3 subjekty neobjevily v trénovacích datech. I přesto, že testovací sada

obsahuje stejné řečníky jako trénovací sada, neobjevuje se žádné z videí v obou sadách.

Cílem rozpoznání AU jednotek je identifikace, zda-li daná jednotka byla obsažena v aktuálním snímku nebo ne. Pokud je daná jednotka rozpoznána, je hodnota výstupu klasifikátoru '1', pokud není, je hodnota '0'. Kvalita rozpoznávání je měřena pomocí F1-measure 5.2. Nejprve je spočteno skóre pro každou jednotku zvlášť a poté je spočten průměr přes všech 12 AU jednotek.

Tabulka 5.6: Výsledky pro jednotlivé AU jednotky dle F1-measure (vzorec 5.2). Srovnání mezi základním systémem systémem s 3D rekonstrukcí. ZS označuje známý subjekt, obsažený v trénovací i testovací sadě. NS označuje neznámý subjekt, který se objevuje pouze v testovací sadě. $\sum$ jsou výsledky pro celou sadu. Pro rozlišení metody použité k normalizaci tváře je pro 3D rekonstrukci uveden index $_{3D}$.

AU	ZS	ZS_{3D}	NS	NS_{3D}	$\sum$	$\sum_{3D}$
1	21	**37**	29	**73**	23	**62**
2	35	**42**	49	**67**	40	**58**
4	**42**	17	**31**	21	**39**	19
6	**40**	23	35	**54**	39	**47**
7	35	**66**	48	**62**	40	**63**
10	41	**36**	47	**51**	44	**46**
12	58	**72**	72	**78**	63	**76**
15	16	**20**	**20**	14	**18**	16
17	**32**	28	14	**33**	26	**31**
18	12	**32**	11	**25**	12	**27**
25	58	**71**	65	**68**	61	**69**
26	22	**44**	37	**39**	30	**42**
průměr	34	40	38	49	36	46

Tabulka 5.6 obsahuje dosažené výsledky pro rozpoznání jednotlivých AU jednotek dle F1-measure. Tabulka porovnává úspěšnost základního systému vůči systému s 3D rekonstrukcí. Výsledky jsou rozděleny na tři části. První část obsahuje srovnání pro testovací sadu obsahující známé subjekty (obsažené i v trénovací sadě), druhá část obsahuje srovnání pro testovací sadu obsahující neznámé subjekty (neobsažené v trénovací sadě) a poslední část srovnává výsledky pro celou testovací sadu. Systém s 3D rekonstrukcí dosahuje v průměru zlepšení rozpoznávání AU jednotek o 10%. Nejvyšší úspěšnost 76 % je dosažena pro rozpoznání AU-12 (zvednuté koutky rtů), nejhůře je rozpoznaná jednotka AU-15 (pokleslé koutky rtů) s úspěšností pouze 16% .

Podrobnou analýzu, vzhledem k vlivu natočení tváře na jednotlivé AU jednotky, nelze provést, neboť organizátoři FERA 2011 neposkytují anotaci pro testovací sadu. Pro získání úspěšnosti systému obsažené v tabulce 5.6 bylo nutné zaslat výstup systému ve speciálním formátu na FERA 2011[6] a obratem získat F1-measure.

[6] FERA 2011 - http://sspnet.eu/fera2011/

5.4 Shrnutí dosažených výsledků

Systém s 3D rekonstrukcí dosáhl oproti základnímu systému zlepšení o 10% jak pro úlohu rozpoznání kategorických emocí tak pro úlohu rozpoznání svalové aktivity. Tabulky 5.7 a 5.8 obsahují navíc srovnání vůči systému z FERA 2011 [115]. Toto porovnání není přesné, neboť základní systém i systém s 3D rekonstrukcí vyřadil 15% snímků jak z trénovací tak i testovací sady (odstraněno bylo několik snímků z každého videa). Důvodem vyřazení snímků byla nedostatečná kvalita nalezených klíčových bodů, ukázka je na obrázku 5.2. Realizace vlastního základního systému je důležitá z důvodu stanovení vlivu normalizace tváře na čelní pohled s využitím 3D rekonstrukce na rozpoznání výrazu tváře.

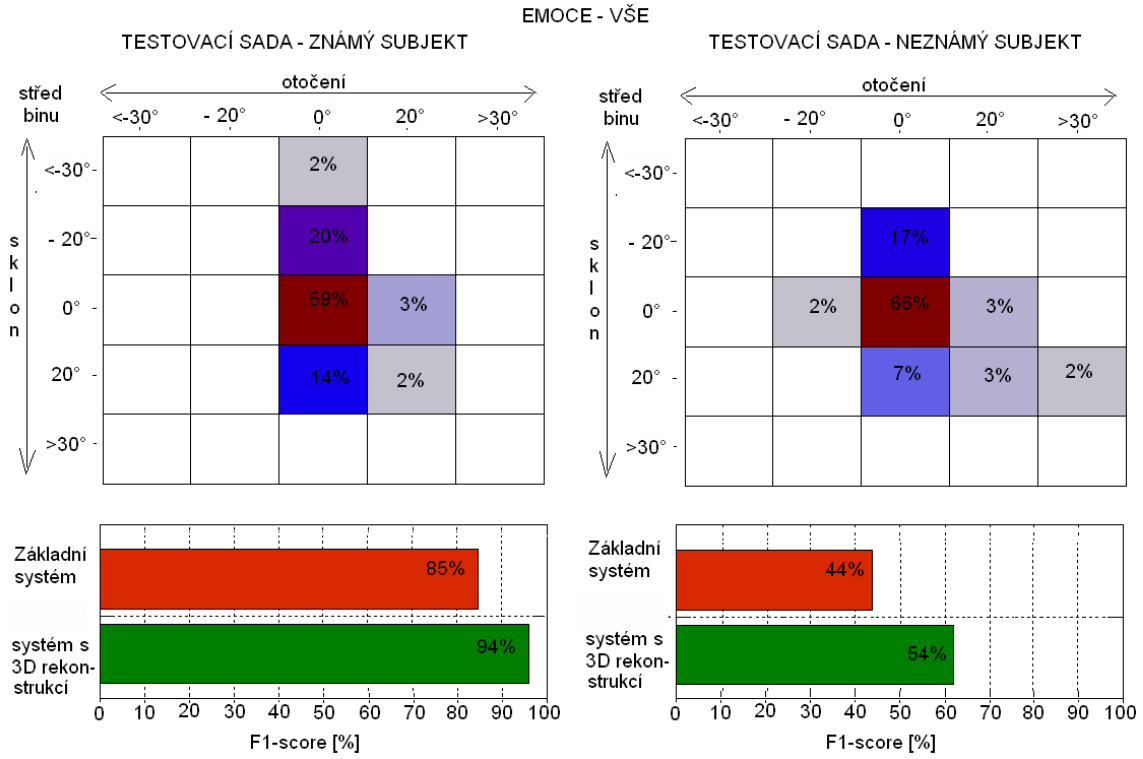
Tabulka 5.7: Srovnání úspěšnosti systému pro rozpoznávání kategorických emocí na datech ze soutěže FERA 2011. NS - neznámý subjekt, ZS - známý subjekt

	ZS	NS	$\sum$
Základní systém z FERA-2011	73	44	56
Základní systém	85	44	63
Systém s 3D rekonstrukcí	**94**	**54**	**73**

V případě rozpoznávání kategorických emocí, tabulka 5.7, je úspěšnost systému hodnocena na základě správně zařazeného testovacího videa do jedné z pěti kategorií (vztek, strach, radost, smutek úleva). V tomto případě vyřazené snímky nemají zásadní vliv na úspěšnost FER systému, protože video i po vyřazení několika snímků obsahuje dostatek dat pro klasifikaci. Navíc díky přesněji nalezenému středu očí dosáhl základní systém oproti systému z FERA 2011 zlepšení o 7%. Toto zlepšení poukazuje na důležitost předzpracovaní dat, respektive díky přesnějšímu nalezení pozice středu očí pomocí AAM modelu byly klasifikátoru předávány data s kvalitnější normalizací tváře (tzn. klasifikátor nemusí řešit variaci způsobenou rozdílnou velikostí tváře či nesprávným odhadem náklonu tváře).

Obrázek 5.9 zobrazuje úspěšnost rozpoznávání kategorických emocí mezi základním systémem a systémem s 3D rekonstrukcí včetně rozložení dat vzhledem k natočení tváře v prostoru. Data pro známý subjekt obsahují 41% dat mimo čelní pohled, z toho většina dat obsahuje sklon hlavy nahoru a dolů. Systém s 3D rekonstrukcí zlepší rozpoznávání pro známý subjekt o 9%. Data pro neznámý subjekt obsahují 34% dat mimo čelní pohled, většina dat obsahuje natočení pro sklon. Oproti známému subjektu obsahují data pro neznámý subjekt dvakrát více dat s otočením tváře na levou či pravou stranu. Systém s 3D rekonstrukcí dosáhl pro neznámý subjekt zlepšení rozpoznávání o 10%. Z dosaženého zlepšení je jasný pozitivní vliv 3D rekonstrukce na rozpoznávání výrazu tváře pro rozpoznávání kategorických emocí.

Při rozpoznávání emocí z výrazu tváře docházelo často k záměně mezi dvojicemi emocí smutek a úleva nebo radost a strach. Pro první dvojici je výraz tváře blízký neutrálnímu výrazu. Emoce smutek má minimální pohyb hlavy, subjekt předvádějící úlevu zaklání hlavu dozadu. Pro dvojici radost a strach má subjekt otevřená ústa. V případě radosti subjekt zaklání hlavu dozadu, zatímco pro emoci strach se hlava hýbe minimálně. Problémem záměny je způsob klasifikace emocí, kdy se s každým snímkem pracuje separátně. To, co tyto emoce rozlišuje, je dynamika pohybu hlavy. Pro zvýšení úspěšnosti rozpoznání mezi jednotlivými



Obrázek 5.9: Výsledky pro rozpoznávání kategorických emocí na datech z FERA 2011. Pro známý i neznámý subjekt dosahuje systém s 3D rekonstrukcí výrazného zlepšení při rozpoznávání výrazu tváře (o 9% pro známý subjekt, o 10% pro neznámý subjekt) .

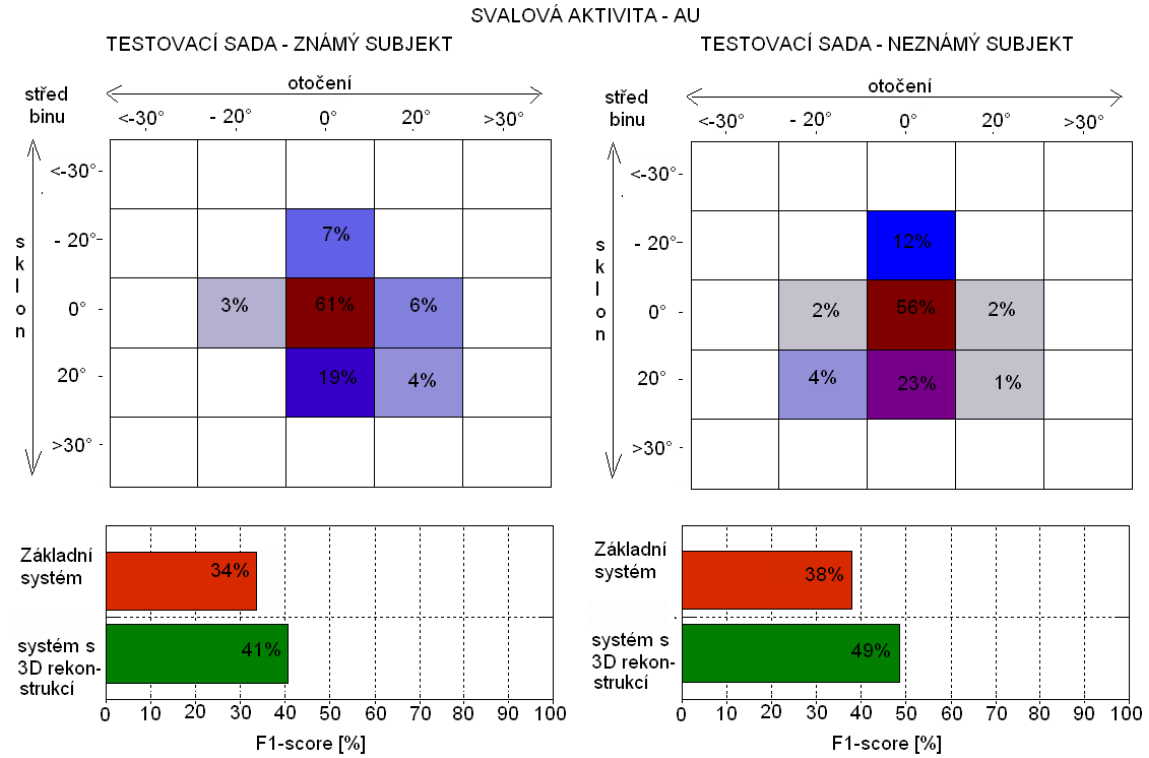
emocemi je nutné vytvořit klasifikátor respektující dynamiku výrazu tváře, respektive pohyb hlavy.

Tabulka 5.8: Srovnání úspěšnosti systému pro rozpoznávání svalové aktivity na datech ze soutěže FERA 2011. NS - neznámý subjekt, ZS - známý subjekt

	ZS	NS	$\sum$
Základní systém z FERA-2011	42	45	45
Základní systém	34	38	36
Systém s 3D rekonstrukcí	40	49	46

Pro rozpoznávání svalové aktivity je úspěšnost systémů hodnocena dle rozpoznání AU jednotek v každém snímku. Vyřazené snímky tedy snižují úspěšnost systému, protože výstup pro všechny AU jednotky je pro tyto snímky automaticky nastaven na '0' (tj. AU jednotka není ve snímku obsažena). Díky tomu je úspěšnost rozpoznání svalové aktivity mezi základním systémem a systémem z FERA 2011 horší o 9%. Organizátoři neposkytují anotaci pro testovací data, namísto toho mají jasně definovaný výstup na základě něhož vygenerují úspěšnost systému dle F1-measure. Srovnání výsledků rozpoznávání svalové aktivity vůči systému FERA

2011 tedy nemá vypovídající hodnotu. Ke stanovení vlivu 3D rekonstrukce bylo nutné realizovat vlastní základní systém pro rozpoznání AU jednotek. Systém s 3D rekonstrukcí dosahuje zlepšení oproti základnímu systému 10%.



Obrázek 5.10: Výsledky pro rozpoznávání svalové aktivity na datech z FERA 2011. Pro známý i neznámý subjekt dosahuje systém s 3D rekonstrukcí výrazného zlepšení při rozpoznávání výrazu tváře (o 7% pro známý subjekt, o 11% pro neznámý subjekt).

Obrázek 5.10 zobrazuje úspěšnost rozpoznávání svalové aktivity základního systému a systému s 3D rekonstrukcí včetně rozložení dat vzhledem k natočení tváře v prostoru. Systém s 3D rekonstrukcí dosáhl zlepšení rozpoznávání AU jednotek pro známý subjekt o 7%, pro neznámý o 11%. Tento výsledek jednoznačně potvrzuje pozitivní vliv 3D rekonstrukce na rozpoznání výrazu tváře.

Úspěšnost rozpoznávání AU jednotek je vyšší pro neznámý subjekt, nehledě na to jaký systém je použitý (viz tabulka 5.8). Stejný závěr vyplývá i z výsledků FERA 2011 [115]. Všechny systémy dosáhly vyšší úspěšnosti pro rozpoznávání na neznámém subjektu. Organizátoři došli k závěru, že data pro známý subjekt nejspíš obsahují výraz tváře s nižší intenzitou a díky tomu je rozpoznávání náročnější. Pro provedení podrobnější analýzy je nutné mít k dispozici anotaci testovací sady, tu však organizátoři neposkytují. Z rozložení dat vzhledem k natočení v obrázku 5.10 je vidět že známý subjekt má o 5% více dat v čelním pohledu než data pro neznámý subjekt. Problém vyšší obtížnosti dat pro známý subjekt tedy není způsoben natočením tváře.

Kapitola 6

Závěr

Výraz tváře je přirozeným prostředkem při mezilidské komunikaci. Automatické rozpoznání výrazu tváře (FER - Facial Expression Recognition) je proto důležité pro vývoj systému komunikace mezi člověkem a počítačem. Počítač využívající informaci o emocionálním stavu člověka může zlepšit interakci s uživatelem (např. elektronický učitel přizpůsobí rychlost výuky s ohledem na výraz tváře žáka). Systém pro rozpoznání výrazu je zajímavý i pro další obory jako psychologie (projev emoce na různé stimuly), medicína (studium autismu) nebo průzkum trhu (reakce spotřebitele na reklamu o novém produktu).

Analýza výrazu tváře a přehled stávajících systému pro rozpoznávání výrazu tváře jsou dílčími cíli této práce. Podrobně jsou popsány v kapitolách 2 a 3. Kapitola 2 se zaměřuje na problematiku výrazu tváře z psychologického pohledu a definuje faktory nutné k rozpoznání výrazu tváře počítačem. Kapitola 3 popisuje metody použité pro jednotlivé kroky při rozpoznávání výrazu tváře (předzpracování, popis tváře a klasifikace výrazu). Dále obsahuje srovnání úspěšnosti stávajících FER systémů.

Díky dostupnosti rozsáhlých databází, nově vytvořeným metodám pro popis tváře a bližší spolupráci s psychology jsou stávající FER systémy plně automatické a schopné rozpoznávat spontánní emoce či mentální stav člověka (znudění, zájem, přemýšlení, nesouhlas, rozpaky, aj.). Problémem stávajících FER systémů je nízká úspěšnost při práci s reálnými daty. Dostupné databáze jsou pořízené v kontrolovaném prostředí. Databáze obsahují data s čelním (či profilovým) pohledem a rovnoměrným osvětlením. Data neobsahují libovolné natočení tváře, překryv tváře (gesta jako mávnutí ruky před obličejem, aj.). Navíc je výraz tváře často hraný. Při konfrontaci s reálnými daty, obsahující libovolné natočení tváře, úspěšnost FER systému významně klesá [115, 100, 63, 131].

Disertační práce se zaměřuje na řešení problému libovolného natočení tváře v prostoru. Nově navržená metoda pro předzpracování vstupních dat je schopná provést 3D rekonstrukci objektu tváře s libovolným výrazem z jediného 2D snímku. Klasifikátoru jsou předkládány snímky v čelním natočením tváře. V této práci byl navržen a experimentálně ověřen FER systém schopný rozpoznat výraz tváře pro sklon hlavy$\pm 30°$stupňů a otočením do stran $\pm 45°$. Podrobný popis nové metody pro normalizaci tváře na čelní pohled včetně aktuálního výrazu je v sekci 4.3.

Pro stanovení vlivu 3D rekonstrukce tváře na rozpoznávání výrazu tváře je systému předkládán buď snímek normalizovaný na fixní pozice středu očí (základní systém) nebo snímek normalizovaný na čelní pohled s využitím 3D rekonstrukce (systém s 3D rekonstrukcí). Popis

obou systému je v sekci 5.1. Ověřen je vliv normalizace tváře na čelní pohled, systém nepracuje s hloubkou získanou z 3D rekonstrukce.

Hlavní cíl práce, vytvoření FER systému schopného pracovat s libovolným natočením tváře, byl splněn. Vytvořený FER systém byl experimentálně ověřen na datech ze soutěže FERA 2011 [115]. Data z FERA 2011 se nejvíce přibližují reálným datům, obsahují libovolné natočení tváře v rozsahu 30° ve všech směrech od čelního pohledu. Porovnání úspěšnosti systémů bylo provedeno jak pro rozpoznávání kategorických emocí tak pro rozpoznání svalové aktivity (např. zvednuté koutky úst). Pro rozpoznávání kategorických emocí dosáhl základní systém 63% a systém s 3D rekonstrukcí 73%. Pro rozpoznávání svalové aktivity dosáhl základní systém 36% a systém s 3D rekonstrukcí 46%. Přehledné shrnutí výsledků mezi základním systémem a systémem využívajícím 3D rekonstrukci tváře je v sekci 5.4.

Systém s 3D rekonstrukcí dosáhl oproti základnímu systému zlepšení o 10% jak pro rozpoznání kategorických emocí tak pro rozpoznání svalové aktivity. Nově navržená metoda pro normalizaci tváře na čelní pohled s využitím 3D rekonstrukce z jediného 2D snímku nabízí alternativní řešení ke klasickým metodám pro pořízení 3D dat (strukturované světlo, stereo viděni, či fotometrie). Nevyžaduje žádné speciální zařízení a při sběru dat je osoba minimálně ovlivněn nahrávacím zařízením. Metoda navržená v této práci vychází z technik používaných při identifikaci osob. Zde existují pokročilé metody schopné rekonstruovat 3D objekt tváře z 2D snímku. Stávající metody, popsané v sekci 4.2 však výraz tváře potlačují. Navržená metoda rekonstruuje 3D objekt tváře včetně aktuálního výrazu tváře.

6.1 Kroky pro další výzkum

V této práci byla ověřena metoda normalizace tváře na čelní pohled používající 3D rekonstrukci tváře z 2D snímku. Systém s 3D rekonstrukcí nevyužívá informaci o hloubce. 3D objekt tváře je ortogonálně promítnut do 2D prostoru. Aby informace o hloubce mohla být využita je nutné realizovat FER systém pracující s 3D daty. Možnými kroky proto jsou. Za prvé získat databázi s 3D daty obsahující různé výrazy tváře. Za druhé ověřit kvalitu 3D rekonstrukce navržené metody. Teprve poté vytvořit FER systém pracující s 3D daty.

Výzkum v oblasti 3D rozpoznávání výrazu tváře využívající informaci o hloubce je v počátečním stádiu. Bude nutné vyřešit ještě mnoho problémů než se 3D rozpoznávání výrazu tváře dostane na úroveň FER systému pracujících s 2D daty. Velkým problémem je časová náročnost registrace 3D dat mezi jednotlivými snímky. Se zvyšujícím se rozlišením a počtem snímků za vteřinu bude nutné tuto registraci optimalizovat. Navíc dostupné databáze obsahují převážně 3D statický sken, který je pro úlohu rozpoznávání výrazu tváře nevhodný. Výraz tváře je dynamický děj a je nutné zachytit pohyby jednotlivých obličejových svalů (více viz kapitola 2).

Další výzkum by měl směřovat k zohlednění dynamiky výrazu tváře. Rychlost změny výrazu tváře a pohyby hlavy se ukázaly být zásadní pro rozlišení kategorických emocí jako je radost a strach nebo smutek a úleva. Návrh klasifikačního schématu spojujícího rozpoznání intenzity výrazu tváře a pohybů hlavy bude potřeba jak pro systémy pracující s 2D daty tak pro systémy pracující s 3D daty.

Resumé

Výraz tváře je přirozeným prostředkem při mezilidské komunikaci. Automatické rozpoznání výrazu tváře je proto důležité pro vývoj systému komunikace mezi člověkem a počítačem. Stávající systémy pro rozpoznávání výrazu tváře pracují převážně s daty pořízenými v kontrolovaném prostředí. Osoba je vyzvána k provedení výrazu tváře s omezeným pohybem hlavy v prostoru. Natočení tváře mimo čelní pohled je tedy minimální. Systémy navržené pro tento typ dat se nedokáží vyrovnat s reálnými daty obsahující libovolné natočení tváře. Jejich úspěšnost se výrazně snižuje s natočením tváře mimo čelní pohled.

Disertační práce se zaměřuje na řešení problému libovolného natočení tváře v prostoru. Nově navržená metoda pro předzpracování vstupních dat je schopná provést 3D rekonstrukci objektu tváře s libovolným výrazem z jediného 2D snímku. Klasifikátoru jsou předkládány snímky v čelním natočením tváře. V této práci byl navržen a experimentálně ověřen systém schopný rozpoznat výraz tváře pro sklon hlavy$\pm$30°stupňů a otočením do stran $\pm$45°.

Vytvořený systém byl experimentálně ověřen na datech ze soutěže FERA 2011 (First Facial Expression Analysis Challange). Data z FERA 2011 se nejvíce přibližují reálným datům, obsahují libovolné natočení tváře v rozsahu 30° ve všech směrech od čelního pohledu. Pro rozpoznávání kategorických emocí dosáhl základní systém 63% a systém s 3D rekonstrukcí 73%. Nově navržená metoda pro normalizaci tváře na čelní pohled s využitím 3D rekonstrukce z jediného 2D snímku nabízí alternativní řešení ke klasickým metodám pro pořízení 3D dat.

Abstract

Facial expressions are a vital part of human communication. Automatic facial expression recognition is essential while designing system for human computer interaction. Existing systems for expression recognition works in controlled environment. Subject is asked to perform an expression while his head is looking straight into camera. Other views than the frontal view are absent. System designed for such data are unable to work with real data containing arbitrary pose.

This work is focused on the problem of facial expression recognition with arbitrary view. A new method able to perform 3D reconstruction of facial expression with arbitrary view from single 2D image has been developed. The method is able to work with face with tilt view $\pm 30^\circ$ and side view $\pm 45^\circ$.

Designed system was experimentally tested on data from The First Facial Expression Recognition Challenge (FERA 2011). This data are very close to real one. Data contains various head pose that vary $\pm 30^\circ$ from frontal view. For the task of emotion recognition the basic system obtain recognition rate of 63%, and system with 3D reconstruction obtained 73%. The new method for face normalization to frontal view using the 3D reconstruction of face from single 2D image offers an alternative way to classical method for acquisition of 3D data.

Literatura

[1] A.B. Ashraf, Simon Lucey, T. Chen, K. Prkachin, P. Solomon, Zara Ambadar, and Jeffrey Cohn. The painful face: Pain expression recognition using active appearance models. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI'07)*, pages 9 – 14, 2007. 2.3.1, A

[2] A. Asthana, J. Saragih, M. Wagner, and R. Goecke. Evaluating aam fitting methods for facial expression recognition. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–8. IEEE, 2009. 3.1.3, 3.5, 3.2.1, 3.2

[3] Tadas Baltrusaitis, Daniel McDuff, Ntombikayise Banda, Marwa Mahmoud, Rana El Kaliouby, Peter Robinson, and Rosalind W. Picard. Real-time inference of mental states from facial expressions and upper body gestures. In *FG'11*, pages 909–914, 2011. 3.4, 3.5

[4] Tanja Banziger, Marcello Mortillaro, and Klaus R Scherer. Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion Washington Dc*, 2011. 3.4

[5] Littlewort G.and Marian Stewart Bartlett, Ian Fasel, Joel Chenu, and Javier R. Movellan. Analysis of machine learning methods for real-time recognition of facial expressions from video. In *Computer Vision and Pattern Recognition*, 2003. A

[6] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In IEEE, editor, *Computer Vision and Pattern Recognition*, volume 2, pages 568 – 573, 2005. 2.3.2, 3.3.1, A

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '99, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 1.1, 4.1, 4.2, 4.2

[8] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, June 1998. 3.7, 3.3.1, 3.3.1

[9] Sien W. Chew, Patrick Lucey, Simon Lucey, Jason M. Saragih, Jeffrey F. Cohn, and Sridha Sridharan. Person-independent facial expression detection using constrained local models. In *FG*, pages 915–920, 2011. 3.1.3

[10] Sien. W. Chew, Patrick Lucey, Simon Lucey, Jason M. Saragih, Jeffrey F. Cohn, and Sridha Sridharan. Person-independent facial expression detection using constrained local models. In *FG'11*, pages 915–920, 2011. 3.4, 3.5

[11] Ira Cohen, Nicu Sebe, and Thomas S. Huang Ashutosh Garg. Facial expression recognition from video sequences. In *International conference on Multimedia and Expo (ICME)*, volume 2, pages 121–124, 2002. 2.3.2, A

[12] Ira Cohen, Nicu Sebe, Fabio G. Cozman, Marcelo C. Cirelo, and Thomas S. Huang. Learning bayesian network classifiers for facial expression recognition with both labeled and unlabeled data. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 2.3.2, 3.1.3, 3.3, 3.3.1, 3.3.2, 3.5, A

[13] Ira Cohen, Nicu Sebe, and Thomas S. Huang Larry Chen, Ashutosh Garg. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91(1-2):160–187, August 2003. A

[14] Jeffrey Cohn, L.I. Reed, Zara Ambadar, Jing Xiao, and Tsuyoshi Moriyama. Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*, volume 1, pages 610 – 616, October 2004. 2.3.2, A

[15] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2296–2303. IEEE, 2011. 4

[16] Roddy Cowie, Ellen Douglas-cowie, Susie Savvidou, Edelle Mcmahon, Martin Sawey, and Marc Schrďż˝der. ďż˝feeltraceďż˝: an instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 19–24, 2000. B, C, C.2

[17] Mohamed Dahmane and Jean Meunier. Emotion recognition using dynamic grid-based hog features. In *FG*, pages 884–888, 2011. 3.4

[18] Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion recognition using phog and lpq features. In *FG*, pages 878–883, 2011. 3.2.2, 3.4

[19] S. D'Mello, T. Jackson, S. Craig, B. Morgan, P. Chipman, H. White, N. Person, B. Kort, R. el Kaliouby, R.W. Picard, and A. Graesser. Autotutor detects and responds to learners affective and cognitive states. In *Workshop on Emotional and Cognitive Issues at the International Conference of Intelligent Tutoring Systems*, 2008. 2.3.1

[20] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60, April 2003. B

[21] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, Noam Amir, and Kostas Karpouzis. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *ACII '07: Proceedings of the*

2nd international conference on Affective Computing and Intelligent Interaction, pages 488–500, Berlin, Heidelberg, 2007. Springer-Verlag. B, B.1, B.2, B.3, B.4, B.5, B.6, B.7

[22] Irenaus Eibl-Eibesfeldt. *Ethology: The biology of behavior.* Oxford, England: Holt, Rinehart, and Winston, 1970. 2.1

[23] P. Ekman. *Handbook of Methods in Nonverbal Behaviour Research*, chapter Methods for measuring Facial Actions, pages 45–90. Cambridge University, 1982. 1

[24] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978. 5

[25] P. Ekman, W. V. Friesen, and S. S. Tomkins. Facial affect scoring technique: A field study. *Semiotica*, 3(1):37–58, 1971. 2.1

[26] Paul Ekman. Universals and cultural differences in facial expressions of emotions. 19:207–283, 1972. 2.1

[27] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, May 1992. 5

[28] R. el Kaliouby and P. Robinson. Mind reading machines: Automated inference of cognitive mental states from video. In *International Conference on Systems, Man and Cybernetics*, 2004. 2.3.1, 3.3.2, A

[29] Rana Ayman el Kaliouby. *Mind-reading machines:automated inference of complex mental states.* PhD thesis, University of Cambridge, 2005. A.1, A

[30] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36, 2003. 2.3.1, 3.2.2

[31] Tobias Gehrig and Hazim Kemal Ekenel. Facial action unit detection using kernel partial least squares. pages 2092–2099. IEEE, 2011. 3.4, 3.5

[32] Gaile G. Gordon. 3d pose estimation of the face from video. In *In Face Recognition: From Theory to Applications*, pages 433–455. Springer-Verlag, 1998. 3.1.4

[33] Ralph Gross. Face databases. In A.Jain S.Li, editor, *Handbook of Face Recognition.* Springer, New York, February 2005. B

[34] Hatice Gunes and Massimo Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 1148–1153, Washington, DC, USA, 2006. IEEE Computer Society. B

[35] Hatice Gunes and Massimo Piccardi. Observer annotation of affective display and evaluation of expressivity: face vs. face-and-body. In *VisHCI '06: Proceedings of the HCSNet workshop on Use of vision in human-computer interaction*, pages 35–42, Darlinghurst, Australia, Australia, 2006. Australian Computer Society, Inc. 2.3.2, B

[36] J.V. Haxby and M.I. Gobbini. Distributed neural systems for face perception. *The Oxford Handbook of Face Perception*, 2011. 2.3.1, 2.3.2

[37] Jingu Heo and Marios Savvides. Generic 3d face pose estimation using facial shapes. In *IJCB*, pages 1–8, 2011. 3.1.4, 3.4, 3.1.4

[38] Florian Hönig. DRIVAWORK - Driving Under Varying Workload. A Multi-Modal Stress Database in the Automotive Context. *not published*, 2007. B

[39] Changbo Hu, Ya Chang, Rogerio Feris, and Matthew Turk. Manifold based analysis of facial expression. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, volume 05, page 81, 2004. 2.3.1, A

[40] Changbo Hu, Ya Chang, Rogerio Feris, and Matthew Turk. Manifold based analysis of facial expression. In *CVPRW '04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 5*, page 81, Washington, DC, USA, 2004. IEEE Computer Society. A.4, A

[41] Yuxiao Hu, Zhihong Zeng, Lijun Yin, Xiaozhou Wei, Jilin Tu, and Thomas S. Huang. A study of non-frontal-view facial expressions recognition. In *ICPR*, pages 1–4. IEEE, 2008. 4

[42] Spiros V. Ioannou, Amaryllis T. Raouzaiou, Vasilis A. Tzouvaras, Theofilos P. Mailis, Kostas C. Karpouzis, and Stefanos D. Kollias. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Netw.*, 18(4):423–435, 2005. C

[43] C. E Izard, C. A. Fantauzzo, J. M. Castle, O. M. Haynes, M.F. Rayias, and P.H. Putnam. The ontogeny and significance of infants' facial expressions in the first 9 months of life. *Developmental Psychology*, 6:997–1013, 1995. C

[44] Carroll Ellis Izard. *Human Emotions*. Springer, 1977. 2.1, 2.2, C, C.1

[45] C.E. Izard. The maximally discriminative facial movement coding system (max). Technical report, University of Delaware, 1979. C

[46] LáSzló A. Jeni, AndráS Lrincz, TamáS Nagy, Zsolt Palotai, Judit Sebk, ZoltáN Szabó, and DáNiel TakáCs. 3d shape estimation in video sequences provides high precision evaluation of facial expressions. *Image Vision Comput.*, 30(10):785–795, October 2012. 4, 4.1

[47] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11)*, pages 314–321, Santa Barbara, CA, USA, March 2011. 3.2.2, 3.2.2

[48] Takeo Kanade. Picture processing system by computer complex and recognition of human faces. In *doctoral dissertation, Kyoto University*. November 1973. 2.3.1

[49] Takeo Kanade, Jeffrey Cohn, and Ying-Li Tian. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, pages 46 – 53, March 2000. 3.4, B.10, B

[50] Ashish Kapoor, Winslow Burleson, and Rosalind W. Picard. Automatic prediction of frustration. *nternational Journal of Human Computer Studies*, 65:724–736, 2007. 2.3.1, 3.3.1, A

[51] Ashish Kapoor and Rosalind W. Picard. Multimodal affect recognition in learning environments. In *ACM Multimedia*, 2005. 2.3.1, A

[52] Ashish Kapoor, Yuan Qi, and Rosalind W. Picard. Fully automatic upper facial action recognition. In *Analysis and Modeling of Faces and Gestures (AMFG)*, 2003. 2.3.2, 3.2, A

[53] Malene Wegener Knudsen, Jean-Claude Martin, Laila Dybkjaer, Marďż˝a Jesďż˝s Machuca Ayuso, Niels Ole Bernsen, Jean Carletta, Ulrich Heid, Sotaro Kita, Joaquim Llisterri, Catherine Pelachaud, Isabella Poggi, Norbert Reithinger, Gijs van Elswijk, and Peter Wittenburg. Survey of multimodal annotation schemes and best practice. Technical report, ISLE Natural Interactivity and Multimodality Working Group, 2002. C

[54] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1940–1954, november 2010. 3.2.2, 3.3.2

[55] Shiro Kumano, Kazuhiro Otsuka, Junji Yamato, Eisaku Maeda, and Yoichi Sato. Pose-invariant facial expression recognition using variable-intensity templates. *International journal of computer vision*, 83(2):178–194, 2009. 4.1

[56] Fernando De la Torre Frade, Joan Campoy, Zara Ambadar, and J. F. Cohn. Temporal segmentation of facial behavior. *International Conference on Computer Vision*, October 2007. 2.3.2, 3.1.3, 3.3.1, A

[57] Milan Legďż˝t, Martin Grďż˝ber, and Pavel Ircing. Wizard of oz data collection for the czech senior companion dialogue system. In *Fourth International Workshop on Human-Computer Conversation*, 2008. B.15, B

[58] Martin D Levine and Yingfeng Yu. State-of-the-art of 3d facial reconstruction methods for face recognition based on a single 2d training image per person. *Pattern Recognition Letters*, 30(10):908–913, 2009. 4.2

[59] Wei-Kai Liao and Isaac Cohen. Belief propagation driven method for facial gestures recognition in presence of occlusions. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 158–158. IEEE, 2006. 4.1, 4.4

[60] Christine L. Lisetti and David E. Rumelhart. Facial expression recognition using a neural network. In *In Proceedings of the 11 th International FLAIRS Conference*, pages 328–332. AAAI Press, 1998. A

[61] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. In *Computer Vision and Pattern Recognition*, 2004. 3.3.1, A

[62] Gwen Littlewort, Ian Fasel, Marian Stewart Bartlett, and Javier R. Movellan. Fully automatic coding of basic expressions from video. In *9th Symposium on Neural Computation*, 2002. 2.3.1, A.2, A

[63] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian R. Fasel, Mark G. Frank, Javier R. Movellan, and Marian Stewart Bartlett. The computer expression recognition toolbox (cert). In *FG*, pages 298–305, 2011. 1.1, 3.2.2, 3.4, 3.5, 3.5, 4, 6

[64] Gwen C. Littlewort, Marian Stewart Bartlett, and Kang Lee. Faces of pain: Automated measurement of spontaneous facial expressions of genuine and posed pain. In *International Conference on Multimodal Interfaces*, 2007. 2.3.1

[65] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010. 3.3

[66] S. Lucey, A.B. Ashraf, and J. Cohn. *Face Recognition Book*, chapter Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face, pages 395–406. Pro Literatur Verlag, 2007. 2.3.2, 3.1.3

[67] Simon Lucey, Iain Matthews, Changbo Hu, Zara Ambadar, Fernando De la Torre Frade, and Jeffrey Cohn. Aam derived face representations for robust facial action recognition. In *Proceedings of the Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FG'06)*, pages 155 – 160, 2006. 4.1, A

[68] J.L. Landabaso M. Pardas, A. Bonafonte. Emotion recognition based on mpeg-4 facial animation parameters. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3624–3627, 2002. C

[69] Ron Rademaker Maja Pantic, Michel Valstar and Ludo Maat. Web-based database for facial expression analysis. In *Proc. IEEE Int'l Conf. on Multimedia and Expo*, pages 317–321, 2005. B, B, B.13, B

[70] Iain Matthews and Simon Baker. Active appearance models revisited. *Int. J. Comput. Vision*, 60(2):135–164, November 2004. 3.2

[71] Hongying Meng, Bernardino Romera-Paredes, and Nadia Bianchi-Berthouze. Emotion recognition by two view svm 2k classifier on dynamic facial expression features. In *FG*, pages 854–859, 2011. 3.4

[72] I. Patras M.F. Valstar and M. Pantic. Facial action unit recognition using temporal templates (pdf file). In *Proc. IEEE Int'l Workshop on Human-Robot Interaction*, pages 253–258, 2004. 2.3.2, 3.2.2, 3.3.1, A

[73] I. Patras M.F. Valstar and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *Computer Vision and Pattern Recognition*, 2005. 2.3.2, 3.3.1, A

[74] T.B. Moeslund. *Visual Analysis of Humans*. SpringerLink : Bücher. Springer-Verlag London Limited, 2011. 3.1.1

[75] S. Moore and R. Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541 – 558, 2011. 3.2.2

[76] Bartlett M.S., Littlewort G.C., and Lainscsek C.and Fasel I.and Frank M.G.and Movellan J.R. Fully automatic facial action recognition in spontaneous behavior. In *7th International Conference on Automatic Face and Gesture Recognition*, pages 223–228, 2006. 2.3.2, 3.2, A

[77] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):607–626, April 2009. 3.3, 3.1.4

[78] Milan Nakonecny. *Lidske emoce*. Academia, 2000. 2.1

[79] Timo Ojala, Matti Pietikae Inen, Senior Member, and Topi Maeenpaw Ae. Multiresolution grayscale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 2002. 3.1.2, 3.2.2

[80] Timo Ojala, Matti Pietikakinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59, 1996. 3.2.2

[81] Alice J. O'Toole, Joshua Harms, Sarah L. Snow, Dawn R. Hurst, Matthew R. Pappas, Janet H. Ayyad, and Herve Abdi. A video database of moving faces and people. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(5):812–816, 2005. B.12, B, B

[82] C Padgett and G Cottrell. Representing face images for emotion classification. In *Advances in Neural Information Processing Systems*, 1997. A

[83] Igor S. Pandzic and Robert Forchheimer. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. MPEG-4 Facial Animation: The Standard, Implementation and Applications, 2002. C.3, C

[84] M. Pantic and I. Patras. Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. In *Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics (*, pages 3358–3363, 2005. 2.3.2, 3.3, A

[85] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man and Cybernetics*, 36:433–449, 2006. 1.1, 2.3.2, 3.3.2, A

[86] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 22:1424–1445, 2000. 2.3.1

[87] M. Pantic and L.J.M. Rothkrantz. 'expert system for automatic analysis of facial expression. *Image and Vision Computing Journal*, 18:881–905, 2000. 2.3.2

[88] M. Pantic and L.J.M. Rothkrantz. Towards an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91:1370–1390, 2003. 2.3.1

[89] M. Pantic and L.J.M. Rothkrantz. Facial action detection from dual-view static face images. In *Proc. IEEE Int'l Conf. on Fuzzy Systems*, pages 39–44, 2004. 2.3.2

[90] Maja Pantic and Ioannis Patras. Temporal modeling of facial actions from face profile image sequences. In *Proc. IEEE Int'l Conf. on Multimedia and Expo*, pages 49–52, 2004. 2.3.2, 3.3.2, A

[91] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas Huang. Human computing and machine understanding of human behavior: a survey. *International Conference on Multimodal Interfaces*, pages 239–248, 2006. 2.3, A.3

[92] Sung Won Park, Jingu Heo, and Marios Savvides. 3d face econstruction from a single 2d face image. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008. 4.3, 4.2, 4.4, 4.3

[93] P.Ekman, W.V. Friesen, and J.C.Hager. Facial action coding system. Technical report, Research Nexus, 2002. 2.2, B

[94] P. Jonathon Phillips, Patrick J. Flynn, Todd Scruggs, Kevin W. Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 947–954, Washington, DC, USA, 2005. IEEE Computer Society. 4.3

[95] Glenn I. Roisman, Jeanne L. Tsai, and Kuan-Hiong Sylvia Chiang. The emotional integration of childhood experience: Physiological, facial expressive, and self-reported emotional response during the adult attachment interview. *Developmental Psychology*, 40:779–789, 2004. B, B

[96] H. A. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '98, pages 38–, Washington, DC, USA, 1998. IEEE Computer Society. 3.1.4

[97] O. Rudovic, I. Patras, and M. Pantic. Facial expression invariant head pose normalization using gaussian process regression. In *Proceedings of IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR-W'10)*, volume 3, pages 28–33, San Francisco, USA, June 2010. 4, 4.1, 4.4

[98] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39:1161–1178, 1980. B

[99] Ashok Samal and Prasana A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: a survey. *Pattern Recogn.*, 25(1):65–77, January 1992. 2.3.1

[100] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012. 3D Facial Behaviour Analysis and Understanding. 1.1, 3.5, 4, 6

[101] Arman Savran, Oya Celiktutan, Aydin Akyol, Jana Trojanová, Hamdi Dibeklioglu, Semih Esenlik, Nesli Bozkurt, Cem Demirkir, Erdem Akagunduz, Kerem Caliskan, Nese Alyuz, Bulent Sankur, Ilkay Ulusoy, Lale Akarun, and Tevfik Metin Sezgin. 3d face recognition performance under adversarial conditions. In *eNTERFACE'07*, pages 87–102, Louvain-la-Neuve, 2007. TELE, Universite catholique de Louvain. B.14, B

[102] N. Sebe, M. Lew, Y. Sun, I. Cohen, T. Gevers, and T.S. Huang. Authentic facial expression analysis. *Image and Vision Computing,*, 2004. 2.3.1, 3.1.3, 3.3.1, 3.3.2, 3.2, A.5, A, B.9, B, B

[103] Thibaud Senechal, Vincent Rapp, Hanan Salam, Renaud Sd'ż˝guier, Kevin Bailly, and Lionel Prevost. Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 860–865, 2011. 3.4, 3.5

[104] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vision Comput.*, 27(6):803–816, May 2009. 3.2.2, 3.2.2, 3.6

[105] Ruchir Srivastava, Sujoy Roy, Shuicheng Yan, and Terence Sim. Accumulated motion images for facial expression recognition in videos. In *FG*, pages 903–908, 2011. 3.4

[106] Ching Y Suen, Arash Zaryabi Langaroudi, Chunghua Feng, and Yuxing Mao. A survey of techniques for face reconstruction. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pages 3554–3560. IEEE, 2007. 4.2

[107] M. Suwa, N. Sugie, and K. Fujimora. A preliminary note on pattern recognition of human emotional expression. In *4th International Joint Conference on Pattern Recognition*, pages 408–410, 1978. 2.3.1

[108] Xiaoyang Tan, Songcan Chen, Zhi-Hua Zhou, and Fuyan Zhang. Face recognition from a single image per person: A survey. *Pattern Recognition*, 39(9):1725–1745, 2006. 4.2

[109] Xiaoyang Tan and Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Trans. Img. Proc.*, 19(6):1635–1650, June 2010. 3.1.2, 3.1

[110] H. Tao and T.S. Huang. Connected vibrations: A modal analysis approach to non-rigid motion tracking. In *CVPR98*, pages 735–740, 1998. A

[111] Usman Tariq, Kai-Hsiang Lin, Zhen Li, Xi Zhou, Zhaowen Wang, Vuong Le, Thomas S. Huang, Xutao Lv, and Tony X. Han. Emotion recognition from an ensemble of features. In *FG*, pages 872–877, 2011. 3.4

[112] Y. L Tian, T.Kanade, and J. F. Cohn. *Transactions on Pattern Analysis and Machine Intelligence*, chapter Recognizing action units for facial expression analysis, pages 97–116. IEEE Computer Society, 2001. 2.3.2, A

[113] R. Valenti, N. Sebe, and T. Gevers. Facial features matching using a virtual structuring element. In *THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING*, volume 6820, page 6820 08. International Society for Optical Engineering; 1999, 2008. 3.1.3, 3.2, A

[114] Roberto Valenti, Nicu Sebe, and Theo Gevers. Facial expression recognition: A fully integrated approach. In *ICIAPW '07: Proceedings of the 14th International Conference of Image Analysis and Processing - Workshops*, pages 125–130, Washington, DC, USA, 2007. IEEE Computer Society. 2.3.1, 3.3.2, A

[115] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analyis of the first facial expression recognition challenge. *IEEE Transactions of Systems, Man and Cybernetics – Part B*, 42(4):966–979, 2012. 1, 1.1, 2.3.1, 2.3.2, 3.1.5, 3.2.1, 3.3, 3.3.1, 3.4, 3.5, 4, 4.1, 4.3, 4.4, 5, 5.1, 5.4, 5.4, 6

[116] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man and Cybernetics*, 42:28–43, 2012. 3.5

[117] M.F. Valstar, M. Pantic, Z. Ambadar, and J.F. Cohn. Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. In *Proceedings of ACM Intd'ž˝l Conf. Multimodal Interfaces*, pages 162–170, 2006. 2.3.2, 3.1.3, A

[118] Michel F. Valstar, Hatice Gunes, and Maja Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pages 38–45, New York, NY, USA, 2007. ACM. 2.3.2

[119] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. *CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION*, pages 511–518, 2001. 3.1.1, 3.1.5

[120] Shu-Fan Wang and Shang-Hong Lai. Reconstructing 3d face model with associated expression deformation from a single face image via constructing a low-dimensional expression deformation manifold. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):2115–2121, 2011. 4.1, 4.2, 4.2, 4.3, 4.4

[121] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Realtime performance-based facial animation. In *ACM SIGGRAPH 2011 papers*, SIGGRAPH '11, pages 77:1–77:10, New York, NY, USA, 2011. ACM. 1.3

[122] Zhen Wen and Thomas S. Huang. Capturing subtle facial motions in 3d face tracking. *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 2:1343, 2003. A

[123] Jacob Whitehill, Marian Bartlett, and Javier Movellan. Measuring the perceived difficulty of a lecture using automatic facial expression recognition. In *ITS '08: Proceedings of the 9th international conference on Intelligent Tutoring Systems*, pages 668–670, Berlin, Heidelberg, 2008. Springer-Verlag. A

[124] Jacob Whitehill and Christian W. Omlin. Local versus global segmentation for facial expression recognition. In *International Conference on Face & Gesture Recognition*, 2006. 2.3.2, 3.3, A

[125] Yu-Kuen Wu and Shang-Hong Lai. Facial expression recognition based on supervised lle analysis of optical flow and ratio image. In *ICS2006*, 2006. 3.2.2

[126] Jing Xiao, Simon Baker, Iain Matthews, and Takeo Kanade. Real-time combined 2d+ 3d active appearance models. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:II–535, 2004. 4.1, 4.1

[127] Songfan Yang and Bir Bhanu. Facial expression recognition using emotion avatar image. In *Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011*, pages 866–871. IEEE, 2011. 3.4

[128] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008. 4

[129] Aurélie Zara, Valérie Maffiolo, Jean Claude Martin, and Laurence Devillers. Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. *Lecture Notes in Computer Science*, pages 464–475, 2007. B, B

[130] Z. Zeng, G.I. Fu, Y.and Roisman, Z. Wen, and T.S. Hu, Y.and Huang. One-class classification for spontaneous facial expression analysis. In *Conf. on Automatic Face & Gesture Recognition*, 2006. 2.3.1, A, B.8

[131] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pages 126–133, New York, NY, USA, 2007. ACM. 1, 3.5, 4, 6, B

[132] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):915–928, June 2007. 3.2.2, 3.2.2

[133] W. Zhao and R. Chellappa. *Face Processing: Advanced Modeling and Methods: Advanced Modeling and Methods*. Elsevier Science, 2011. 3, 3.1.5

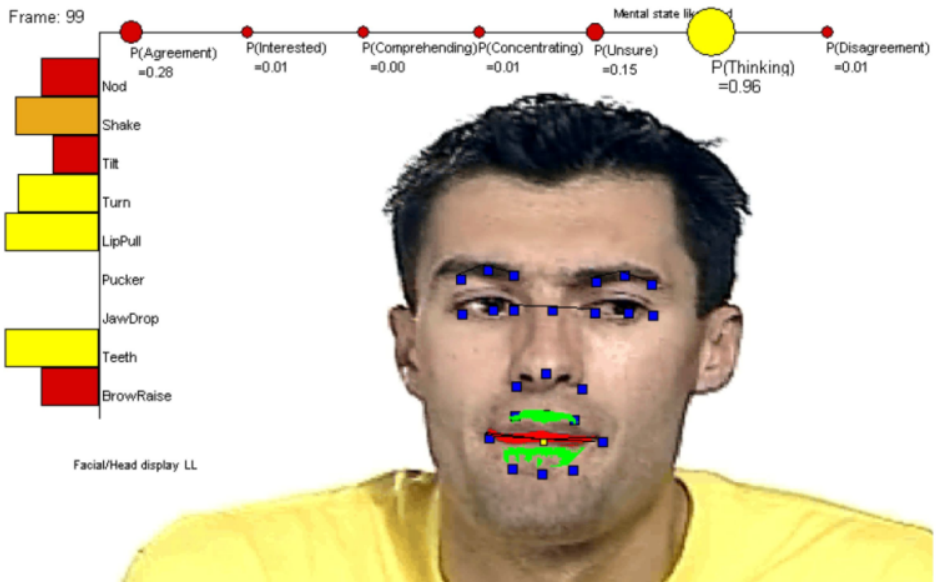
[134] Zhi-Hua Zhou, Ke-Jia Chen, and Hong-Bin Dai. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Trans. Inf. Syst.*, 24(2):219–244, April 2006. 2.3.2

[135] Karel Zimmermann, Jiří Matas, and Thomáš Svoboda. Tracking by an optimal sequence of linear predictors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):677–692, April 2009. 3.1.1

Příloha A

Ukázky FER systémů

Přehled FER systémů je vztažen k výzkumným týmům pracujícím na jeho vývoji. Díky tomuto dělení se snadno sleduje vývoj jednotlivých systémů, jejich výhody oproti ostatním systémům.



Obrázek A.1: MindReader from Massachusetts Institute of Technology rozpoznaný mentální stav je přemýšlení. Převzato z [29]

Tým Affective Computing(AC) [1]z Massachusetts Institute of Technology (MIT) pracuje na vývoji několika systémů pro rozpoznávání emocí. Jedním z nich je MindReader (prezentovaný v Kaliouby04 [28] a Kaliouby05 [29]) pro rozpoznávání kognitivního mentalního stavu z výrazu. Jedná se o první systém, který rozpoznává jiné než základní emoce. MindReader je schopen rozpoznat souhlas, koncentraci, nesouhlas, zájem, přemýšlení a nejistotu. Při rozpoznávání výrazu se kromě popisu obličeje využívá i informace o pohybu hlavy. Dalším je systém

[1] affect.media.mit.edu/index.php

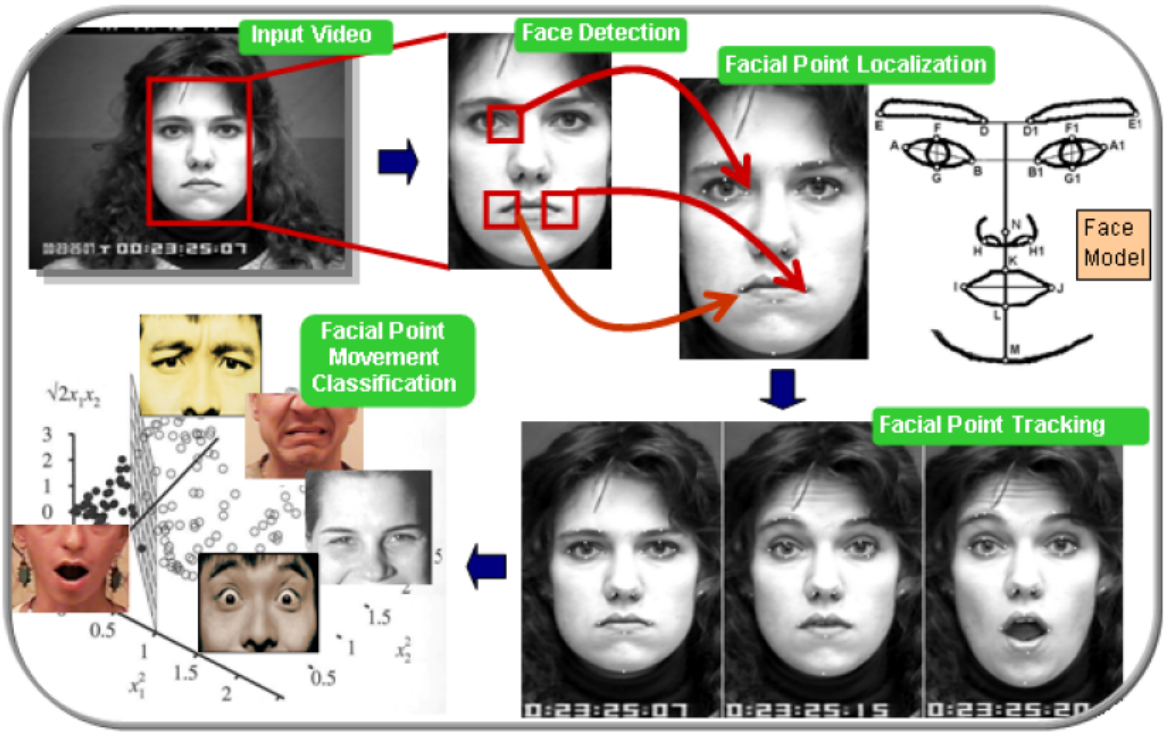
pro rozpoznávání AU jednotek v horní polovině obličeje Kapoor03 [52]. Systém pro rozpoznání zájmu a nezájmu u dětí při skládání puzzle Kapoor05 [51]. Systém pro rozpoznání frustrace Kapoor07 [50]. Zajímavostí těchto systémů je použití termovizní kamery pro lokalizaci očí. Všechny systémy vyvíjené v AC laboratoři se v současné době převádějí do funkčních prototypů. FaceReader: Affective-Cognitive State Inference from Facial Video, RoCo: A Robotic Desktop Computer, SmileSeeker: Customer and Employee Affect Tagging System, ShyBot, pro podrobný popis systémů lze nalézt na webové stránce [2].

Obrázek A.2: Emotion Mirror from University of California, San Diego. Sloupcový graf nalevo zobrazuje detekovanou emoci, červená barva zobrazuje zlost, žlutá barva reprezentuje smutek, šedá reprezentuje neutrální výraz. Vpravo je zobrazen nejbližší rozpoznaný výraz. Převzato z [62].

Tým Machine perception laboratory (MPL)[3] z University of California, San Diego (UCSD), vytváří systémy pro rozpoznání základních emocí (Fully Automatic Recognition of Basic Emotion [62],[5], [61]), rozpoznávání AU jednotek (Fully Automated FACS Coding [6],[76],[124]), a rozpoznávání mentálního stavu z výrazu v obličeji (AutoTutor [123]). Zajímavou studií je práce [124] srovnávající lokální a globální segmentaci obličeje. Předchozí práce v této oblasti [60, 82] došly k závěru, že lokální segmentace dosahuje lepších výsledků než globální. Whitehill a Omlin došli při testování k opačnému závěru. Pro rozpoznávání AU jednotek v horní části obličeje (oči a obočí) dosahuje globální segmentace daleko lepších výsledků než lokální segmentace. Tento výsledek připisují problému korelace AU jednotek v Cohn-Canade B databázi a navrhují natočení databáze, kde by se AU jednotky vyskytovaly separátně.

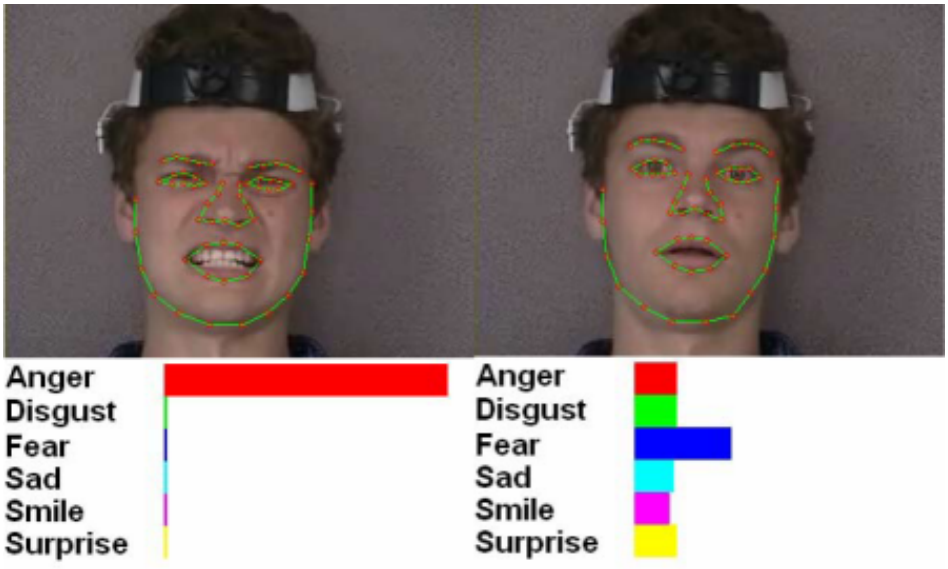
Tým Human-Centered Intelligent Human-Computer Interaction (HCI^2) z Imperial College London (ICL) vyvíjí systém pro rozpoznávání AU jednotek ([72, 73, 84, 73]). Jako jediní se zaměřují na rozpoznávání AU jednotek z profilu obličeje ([90, 85]). Dále se zabývají rozdílem mezi hranými a spontánními emocemi. Ve Valstar06 [117] definovali nové příznaky: maximální intenzita, délka trvání, symetrie, a pořadí výskytu, pomocí nichž rozhodovali, zda se jedná o emoci hranou či spontánní. Také poukázali na rozdílnou délku začátku a konce emoce při

[2] affect
[3] mplab.ucsd.edu/wordpress/?page_id=70



Obrázek A.3: System from Imperial College London Struktura systému pro rozpoznávání AU jednotek. Převzato z [91]

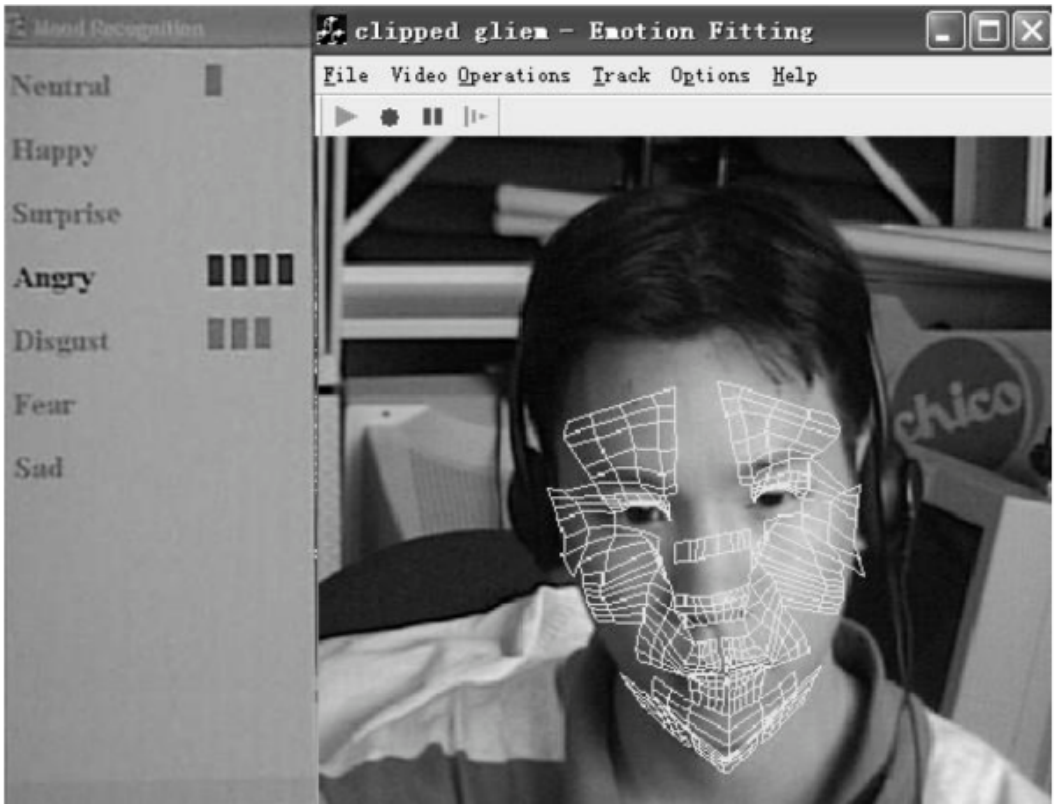
spontánním výrazu. Protože spontánní emoce je často krátká, nové příznaky zvýšily přesnost rozpoznávání emocí z výrazu obličeje.



Obrázek A.4: System from Carnegie Mellon University (CMU) a University of pitsburg (UP) Ukázka výstupu ze systému pro rozpoznávání emocí z výrazu v obličeji. Převzato z [40]

Tým Face group (FG)[4] z Carnegie Mellon University (CMU) a tým Affect analysis group (AAG)[5] z University of Pittsburgh (UP) vyvíjejí ve spolupráci dva typy systémů. Jeden rozpoznávvá emoce ([39, 40]) a druhý rozpoznává AU jednotky ([112, 14, 67, 56]). V článku [14] jako první upozornili na rozpoznávání spontánních emocí. Také se zabývajé rozpoznáním rozdílu mezi výrazem tváře při bolesti a při emoci [1].

[4] www.ri.cmu.edu/research_lab_group_detail.html?type=projects\&lab_id=51\&menu_id=263
[5] www.pitt.edu/\~emotion/research.html



Obrázek A.5: Emotion fitting from University of Illinois at Urbana-Champaign a University of Amsterdam (UA). Na pravé straně je vidět drátový 3D model sledovaného obličeje. Nalevo je zobrazena detekovaná emoce, zlost, (sloupcový graf zobrazuje relativní pravděpodobnost vzhledem k ostatním emocím). Převzato z [102]

Tým Image Formation and Processing (IFP)[6] z University of Illinois at Urbana-Champaign (UIUC) a tým Intelligent Systems Lab Amsterdam (ISLA)[7] z University of Amsterdam (UA) spolupracují společně na vývoji systému pro rozpoznávání emocí ([11, 12, 13, 122, 102, 130, 114, 113]). Pro sledování obličeje využívají Piecewise B´ezier Volume Deformation navrhnutý Huangem v [110]. Jedná se o tracker založený na 3D modelu. Získané příznaky reprezentují směr a intenzitu pohybu ve 3D. Tracker je schopný vyrovnat se s natočením až 40 stupňů. V práci Sebe04 [102] je k nalezení zatím největší srovnání úspěšnosti klasifikátorů (celkem 24 klasifikátorů (nejvýšší úspěšnosti dosáhl K-nejbližší soused v kombinaci s boosting algoritmem). Posledním výstupem spolupráce je projekt Multimodal Computer-aided Learning Systém [8], schopný vzhledem k rozpoznanému výrazu upravit rychlost výuky.

[6] www.beckman.illinois.edu/hcii/ifp.aspx
[7] www.science.uva.nl/research/isla/
[8] itr.beckman.uiuc.edu/

Příloha B

Přehled databází

Databáze je základem při tvorbě jakéhokoliv systému snažícího se napodobovat lidskou komunikaci. Ať už je systém založen na trénování z dat, nebo používá expertní pravidla, musí být otestován na dostatečném množství dat. Pro návrh automatického FER systémů je potřeba mít data zachycující obličej člověka. Data mohou být ve formě sekvence obrázků nebo videa. V této příloze je uveden přehled existujících databází zvlášť jsou uvedené databáze, jejichž natáčení jsem se účastnila v průběhu doktorského studia.

Existující databáze

Natočené databáze pro FER systémy se dají rozdělit do dvou hlavních skupin, a to hrané a spontánní. Databáze obsahující hraná data je natočena v kontrolovaném prostředí, kde je subjekt vyzván, aby předvedl emoci. Spontánní databáze může být natočena jak v přirozeném prostředí, tak i v kontrolovaném prostředí. Příkladem přirozeného prostředí je např. natáčení při venkovních aktivitách, rozhovory se subjektem o jeho názorech, nebo natáčení v průběhu sledování filmů. Kontrolované podmínky jsou různé scénáře, jako např. subjekt je vyzván aby rukou prozkoumal obsah černé skříňky či v průběhu simulace řízení auta jsou subjektu zadávány matematické úkoly. Spontánní databáze je často pořízená bez vědomí subjektu o záměru využití dat (subjekt si není vědom, že se zajímáme o projev jeho emocí). Všechny zde uvedené databáze mají formu video záznamu (v 2D prostoru). Data buď obsahují emoce, nebo aktivitu obličejových svalů. Existuje také mnoho databází se statickými obrázky, kde je emoce či AU jednotka vyjádřená v maximální intenzitě. Přehled těchto databází lze najít v [33], [69]. Přehled audio-vizuálních databází je v [131].

HUMAINE databáze

HUMAINE je EU projekt, jehož cílem je koordinace práce výzkumných skupin pracujících v oblasti rozpoznávání emocí [1]. HUMAINE databáze je jedním z mnoha výstupů tohoto projektu [21]. Hlavním důvodem pro vytvoření této databáze bylo poskytnutí zdroje informací (o formátu emocí v každodenním životě) pro různé typy výzkumných center. Shromážděná data ukazují emoce v různých kontextech. Data poskytují vzory základních principů a nejsou

[1] www.emotion-research.net

primárně zamýšlena pro strojové učení. Formát dat obsahuje jak přirozené situace, tak i situace uměle navozené.

V rámci projektu vznikl i formát pro anotaci dat HUMAINE Emotion Annotation and Representation Language (EARL). EARL je podrobněji rozebrán v kapitole 2.2. Z databází pořízených v rámci HUMAINE projektu bylo vybráno 48 vzorků (klipů) v délce od 3 sekund do 2 minut. Pouze tyto klipy byly anotovány v EARL formátu. Databáze celkově obsahuje 11 různých databází. V této kapitole jsou zmíněny pouze ty, které jsou relevantní pro FER systémy. Následuje jejich popis.

Obrázek B.1: The Belfast Naturalistic Emotional Database. Subjekt vypraví svému starému příteli o svém budoucím zeti. Převzato z [21].

The Belfast Naturalistic Emotional Database
Data obsahují materiály pořízené z televizních diskuzních pořadů. Ukázka dat je na obrázku B.1. Pro každý subjekt je k dispozici nejméně jeden záznam obsahující emoci a jeden obsahující neutrální výraz. Anotace databáze je provedena pomocí Feeltrace [16]. Každý záznam byl anotován třemi nezávislými posluchači. Databáze je distribuovaná na CD a dostupná je po podepsání licenčního ujednání. Obrazová data jsou v MPEG formátu, zvuková ve WAV. Celkově obsahuje databáze 298 klipů pro 125 subjektů, z toho je 31 mužů a 94 žen. Délka klipů je v rozmezí od 10 do 60 vteřin. Vizuální data obsahují pohled ze předu. Kamera snímá obličej, hlavu a ramena [20].

Obrázek B.2: The Belfast Activity/Spaghetti Data. Vlevo subjekt zachycený při sledování pádu člověka z kola. V pravo subjekt, který spustil bzučák při pohybu ruky vně černé bedýnky naplněné špagetami. Převzato z [21].

The Belfast Activity/Spaghetti Data
Data jsou pořízena ve dvou různých situacích. V první situaci jsou subjekty zachyceny při venkovních aktivitách. Data obsahují rychle se pohybující subjekty, zachycené emoce se pře-

krývají. Tato databáze obsahuje několik hodin záznamů, data jsou v hrubé nezpracované formě. V druhé situaci jsou subjekty zachyceny v kontrolovaném prostředí (sedí za stolem). Jejich úkolem je prohledat černou skříňku a napsat svoje pocity na bílou tabuli. Ve skříňce je umístěno několik nepříjemných věcí (včetně špaget a bzučáku). Celkově bylo natočeno 118 subjektů, z toho 50 žen a 68 mužů. Pro každého bylo natočeno 6 různých scénářů v celkové délce 2,5 minuty. Ukázky obou databází jsou na obrázku B.2.

Obrázek B.3: The Castaway Reality Television Dataset. Vlevo subjekt po úspěšném splnění úkolu. Vpravo subjekt po neúspěšném splnění úkolu. Převzato z [21].

The Castaway Reality Television Dataset
Data obsahují materiály z televizní soutěže, kde se 10 lidí účastní různých soutěží na opuštěném ostrově. Záznamy zachycují subjekty při různých aktivitách (např. dotýkání hadů) včetně rozhovorů s účastníky po vykonané aktivitě. Data jsou opravdovou výzvou pro FER systémy. Subjekty jsou natáčeny v průběhu aktivity a data občas neobsahují čelní pohled. Databáze obsahuje velký rozsah emocí.

Obrázek B.4: The DRIVAWORK Dataset. Subjekt při řízení auta. Vlevo je subjekt v uvolněném stavu (radost). Vpravo je subjekt vykonávající úlohu (soustředěnost). Převzato z [21].

Driving under Varying Work-load (DRIVAWORK)
DRIVAWORK byl natočen v Erlangenu [38]. Databáze obsahuje audio, video a fyziologické (EKG, EMG, teplotu prstů) signály během simulovaného řízení auta. Subjekt je v průběhu řízení požádán o vykonání úkolu, který má vyšší náročnost na přemýšlení a vede ke zvýšení stresu. Úkoly mají formu aritmetických úloh. Celkem bylo natočeno 24 subjektů v průběhu 15 hodin.

The Emotaboo Dataset
Korpus Emotaboo byl natočen na univerzitě LIMSI-CNRS ve Francii [129]. Data obsahují interakci mezi dvěma subjekty v průběhu hry TABOO. Princip hry je následující: jeden hráč se snaží druhému vysvětlit slovo pomocí gest a pohybů těla. Zachycené emoce obsahují pozitivní

Obrázek B.5: The Emotaboo Dataset. Vlevo je subjekt ukazující tajné slovo. Vpravo je subjekt hádající slovo. Převzato z [21].

pocity a rozpaky (v případě neznalosti slova). Databáze obsahuje data ze čtyř různých pohledů. Celkem bylo natočeno 8 hodin s 20 hráči.

Obrázek B.6: The Green Persuasive Dataset. Vlevo je subjekt přesvědčující druhého, že v prostředí spřízněném s přírodou nebudou auta potřebná. Vpravo je přesvědčovaný subjekt, který s názorem nesouhlasí. Převzato z [21].

The Green Persuasive Dataset

Data obsahují interakci dvou subjektů, kde se jeden snaží přesvědčit druhého, aby žil více zodpovědně vzhledem k životnímu prostředí (využívajíce emocionální podtext). Databáze zachycuje komplexní emoce spojené s různými kognitivními stavy a mezilidskými signály. Celkově bylo natočeno 8 rozhovorů, každý o délce přibližně 30 minut.

Obrázek B.7: The Sensitive Artificial Listener (SAL). Vlevo subjekt konverzuje se sklíčeným Obadiahem. Vpravo subjekt konverzuje se šťastným virtuálním subjektem. Převzato z [21].

The Sensitive Artificial Listener (SAL)

SAL je virtuální osoba schopná měnit svojí identitu (smutny Poppy, rozlobený Spike, sklíčený

Obadiah, citlivý Prudence). Každá osobnost virtuálního subjektu má zásobu frází, kterými se snaží navodit stejnou náladu u člověka, s nímž si povídá. Zpovídaný subjekt si může sám vybrat, se kterou osobností si bude povídat. Databáze obsahuje záznamy pro 4 subjekty po 20 minut pro každého. Obrazová data subjektu obsahují čelní pohled, v záznamu nejsou obsažena žádná gesta.

Obrázek B.8: The Adult Attachment Interview Database. Ve velkém okně je zobrazen zpovídaný subjekt, v malém okně dotazující. Senzory na uších měří vodivost kůže. Převzato z [130]

The Adult Attachment Interview database (AAI) [95]

Databáze byla původně určena pro psychologický výzkum emocí. Data obsahují spontánní emoce, natočené v průběhu rozhovoru, kde subjekt popisuje svoje zážitky z dětství. Protokol AAI vyžaduje, aby subjekt popsal, jaký měl vztah k rodičům v dětství, vzpomněl si na období, kdy se cítil odloučený, popsal příklady, kdy si uvědomoval odmítnutí, popsal vzpomínky na setkání se ztrátou, a popsal současný vztah k rodičům. Subjekt a tazatel byli natáčeni skrytými kamerami, ukázka je na obrázku B.8. Celkem bylo natočeno 60 subjektů (30 evropanů, 30 asiatů) ve věku 18 až 30 let. Data obsahují kromě audio-vizuálního signálu fyziologické signály (elektrická vodivost kůže, puls). Anotace dat byla provedena dvěma FACS anotátory, kteří se shodovali v 85%.

Obrázek B.9: The Authentic Expression Database. Vlevo neutrální výraz, vpravo výraz radosti. Převzato z [102]

The Authentic Expression Database [102]

Cílem této databáze bylo získání autentického výrazu vyjadřujícího emoční stav člověka. Pro tento účel byl postaven kiosek se skrytou kamerou, ve kterém běžely upoutávky na filmy v kinech. Tento přístup byl zvolen po několika konzultacích na katedře psychologie. Test zohledňoval celkem tři podmínky: subjekt by neměl vědět, že je testován jeho emoční stav, každý subjekt musí být hned po testu vyzpovídán, aby sám popsal svoje pocity, a ve stejné místnosti by se subjektem neměl být nikdo jiný. Celkem použilo kiosek 60 lidí, z nichž pouze 28 souhlasilo s použitím dat pro vědecké účely. Každý subjekt se vyjádřil k výrazům, které zhlédl na videozáznamu. Data obsahují spontánní emoce (radost, překvapení, znechucení). Nevýhodou tohoto přístupu je problém získat větší emoční rozsah (strach a smutek nelze při takovémto experimentu získat).

Obrázek B.10: The CMU-Pittsburgh AU-Coded Facial Expression Database. Vlevo čelní pohled, vpravo pohled z úhlu 30°. Převzato z [49]

The CMU-Pittsburgh AU-Coded Facial Expression Database [49]

Často nazývána jako 'The Cohn-Kanade Database' je databáze natočena v kontrolovaných podmínkách s hranými (pózovanými) emocemi. Natočeno bylo celkem 210 lidí ve věku od 18 do 50 let. Větší část jsou ženy (69%). Celkem 81% subjektů jsou Evropané, 13% Afričané a 6% jiná etnika (Číňani, Indové, aj.). Subjekt je snímán dvěmi kamerami. Jedna je umístěna z čelního pohledu a druhá 30 stupňů napravo od subjektu, příklad je na obrázku B.10. Třetina databáze je nahrána bez přídavného osvětlení, zbylé dvě třetiny byly natáčeny s konstantním osvětlením (2 vysoce-intenzitní lampy). U 60 subjektů byla natočena rotace hlavy s výrazem obličeje. Každý subjekt předvedl celkem 23 různých výrazů. Zdigitalizováno bylo jen 182 subjektů s celkovým počtem 1917 sekvencí. Rozlišení obrazu je 640x480 pixelů. Celkem je pro veřejnost k dispozici 480 anotovaných sekvencí. Anotaci prováděli dva certifikovaní FACS anotátoři. Pozadí za subjektem bylo po celou dobu neměnné.

The Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior (FABO) [34, 35]

Bimodální databáze FABO je první svého druhu. Zachycuje jak výraz v obličeji, tak i pohyb horní poloviny těla. Při natáčení byly použity dvě kamery, ukázka dat je na obrázku B.11. Data jsou hraná, subjekt byl vyzván, aby provedl požadovanou akci. Celkem bylo natočeno 1900 video sekvencí od 23 subjektů ve věku 18 až 50 let. Z toho 10 bylo Evropanů, 9 Asiatů, 3 z

Obrázek B.11: The FABO database. Vlevo záznam z kamery s přiblížením na obličej. Vpravo záznam z kamery s celkovým pohledem. Převzato z

Jižní Ameriky, a 1 Australan. Zachycené emoce jsou základní, plus nejistota, znudění, úzkost a neutrální výraz. Anotace dat byla provedena pomocí dvou metod. Zaprvé obličej byl zařazen do jednotlivých kategorií, s ohledem na Ekmanovu definici [93]. A zadruhé celý video záznam byl kódován do dvoudimenzionálního prostoru s ohledem na práci Russela [98].

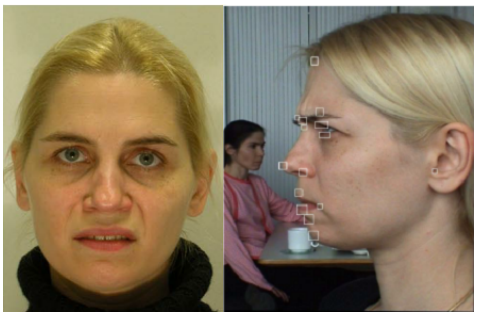
Obrázek B.12: A Video Database of Moving Faces and People from University of Texas. Vlevo výraz znechucení, vpravo radost. Převzato z [81]

A Video Database of Moving Faces and People from University of Texas [81]

Databáze byla natočena v kontrolovaných podmínkách, data zachycují spontánní emoce pro radost, smutek, znechucení, vztek, rozpačitost, smích, překvapení, znudění nebo nedůvěru (ukázka je na obrázku B.12). Během natáčení, subjekt sledoval deseti minutové video obsahující scénky z různých filmů. Celkem je k dispozici pro 284 subjektů 852 video sekvencí. Pro každý subjekt jsou průměrně k dispozici 3 různé výrazy obličeje (v délce 5 vteřin) a neutrální výraz. Anotace dat byla provedena na základě pozorovatele, který daný klip označil jednou z emocí. Data obsahují různé natočení hlavy a některá videa obsahují i kombinaci více emocí dohromady (např. přemýšlení, které přejde v překvapení a je zakončeno smíchem).

The MMI Database (Maja & Michal Initiative) [69]

Databáze byla natočena v kontrolovaných podmínkách, obsahuje jak hrané, tak spontánní emoce. Data obsahují celkem 4000 videosekvencí a 600 statických snímků zachycujících AU jednotky, kombinace AU jednotek, nebo základní emoce. Natočeno bylo celkem 52 subjektů ve věku 19 až 62 let. Data obsahují více národností: 81% jsou běloši, 14% jsou Asiati a 5% jsou Afričani. Přibližně polovina dat je natočena s ženami. Subjekt je natáčen dvěmi kamerami při

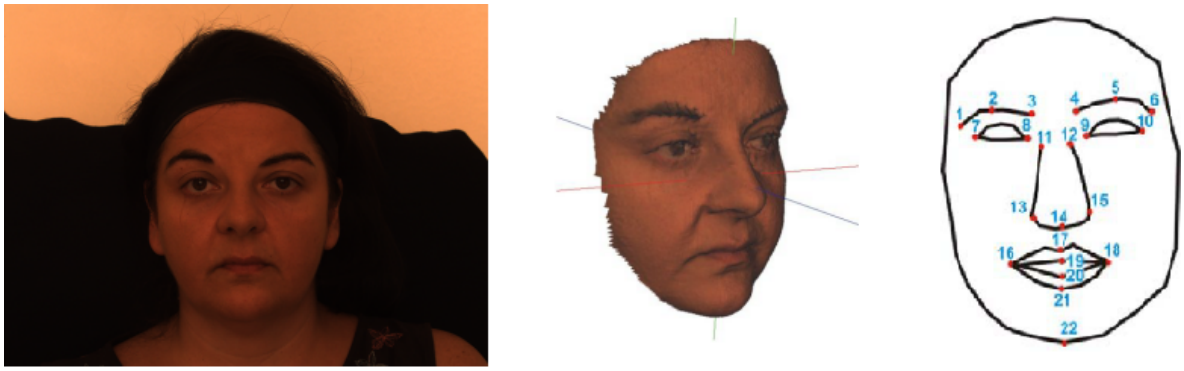


Obrázek B.13: The MMI database. Vlevo čelní pohled s výrazem znechucení, vpravo pohled z profilu. Převzato z [69]

konstatním osvětlení, jedna kamera je zpředu a jedna z profilu (ukázka je na obrázku B.13). Anotaci dat prováděli dva FACS anotátoři. Databáze je stále průběžně doplňována, dostupná je na webových stránkách [2].

Realizované databáze

V průběhu studia jsem se účastnila nahrávání tří databází. Databáze obsahují velmi hodnotná data pro testování systému pro rozpoznání emocí.



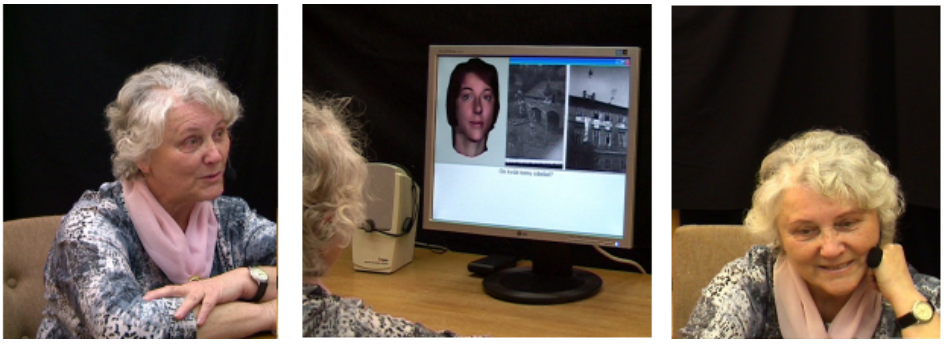
Obrázek B.14: Ukázka dat z databáze Enterface07. Statický 2D snímek, 3D vrml model a ukázka anotace dat (manuální označení klíčových bodů). Převzato z [101]

Enterface07 [101]

Databáze byla natočena v rámci workshopu eNTERFACE[3]. Databáze obsahuje statické 3D modely určeny k rozpoznání lidí a výrazu. Data obsahují různé natočení hlavy, velké množství AU jednotek a překryvy obličeje rukou. Ukázka dat je na obrázku B.14. Celkem bylo natočeno

[2] www.mmifacedb.com/
[3] http://www.cmpe.boun.edu.tr/enterface07/

81 lidí (a databáze je stále rozšiřována) ve věku 25 až 35 let. Pro každého je k dispozici v průměru 40 modelů. Subjekty jsou Evropané, jedna třetina jsou ženy. Všechny dostupná data byla manuálně anotována, pro každý snímek je k dispozici 24 klíčových bodů B.14.



Obrázek B.15: Ukázka dat z databáze Companions. Převzato z [57]

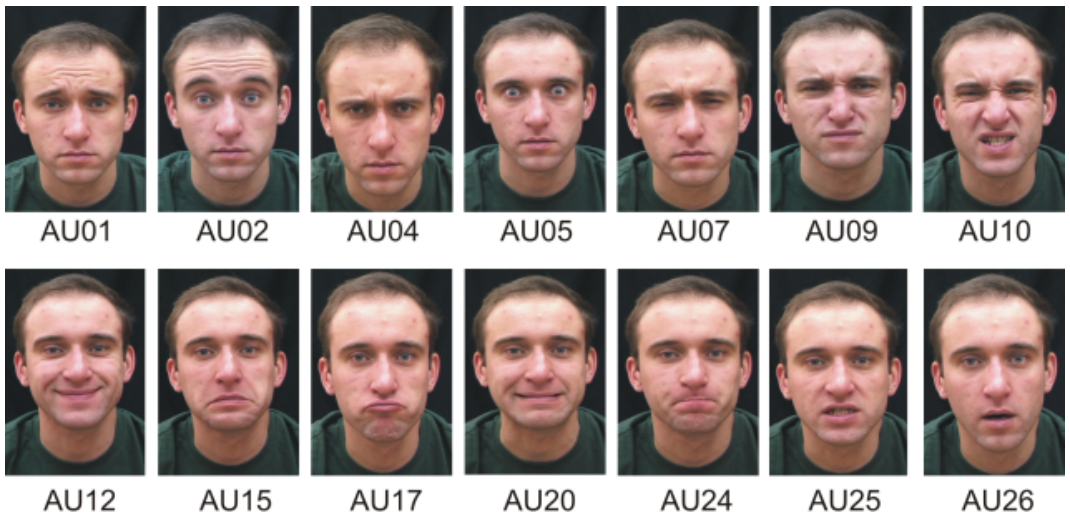
Data Collecttion for the Czech Senior Companion Dialogue System [57]

Databáze byla pořízena v rámci projektu Companions[4], jejím cílem je vytvoření společníka pro staré lidi. Data obsahují záznam z konverzace člověka s virtuální subjektem. Předmětem konverzace jsou fotky. Scéna byla snímána třemi kamerami, ukázka na obrázku B.15. Data obsahují velké množství spontánních emocí a gest. Celkem bylo natočeno 65 osob ve věku od 54 do 86 let. Průměrná délka jednoho rozhovoru je 55 minut. Databáze neobsahuje vizuální anotaci emocí.

The UWB-07-EFER Database

Databáze UWB-07-EFER (University of West Bohemia - 2007 - Emotions and Facial Expressions Recognition) byla vytvořena v rámci projektu MUSSLAP (Multimodal Human Speech and Sign Language Processing for Human-Machine Communication) [5] s cílem rozpoznání emoce z výrazu obličeje při promluvě. Databáze obsahuje hraná data pro 20 subjektů, skládá se ze tří částí: 14 AU jednotek (ukázka na obrázku B.16), šesti základních emocí a 13 krátkých promluv. Průměrně bylo pro každého řečníka nahráno 30 videozáznamů, název souboru určuje nahranou akci. Největší problém měli subjekty s provedením jednotky AU13 (pokles koutků rtů), s výrazem smutku a promluvou se smutným výrazem. Pro výraz smutku je vhodnější použít spontánní data z databáze Companions (B).

[4] www.companions-project.org/
[5] musslap.zcu.cz/cs/o-projektu/



Obrázek B.16: The UWB-07-EFER Database. Ukázka akčních jednotek.

Shrnutí

Pro tvorbu FER systémů je nezbytným předpokladem anotovaná databáze, obsahující velké množství spontánních emocí. Problém ale je, že zatím neexistuje dostatek dat splňujících tuto podmínku, neboť není snadné pořídit záznam se spontánními emocemi a navíc je jejich manuální anotace zdlouhavá a nákladná. Z těchto důvodu většina systémů pracuje převážně s databázemi obsahující hraná data. A pokud už pracují s daty spontánními, jedná se většinou o menší úzce zaměřené databáze v kontrolovaném prostředí s čelním (či profilovým) pohledem, bez překryvu (gesta jako mávnutí ruky před obličejem, aj.) a s vysokou kvalitou obrazu [95, 102, 81, 69]. Pro anotaci dat je na výběr hned několik metod (FACS, MPEG-4, MAX, Feeltrace). Popis výrazu tváře by měl splňovat podmínku oddělení hypotézy(rozpoznaná emoce např. radost) od popisu výrazu (zvednuté obočí a koutky úst). Z uvedených systémů tuto podmínku splňuje systém FACS, MPEG-4 a Feeltrace. Nevýhodou systému FACS je statická interpretace, pro vytvoření dynamické reprezentace je zapotřebí anotovat každý snímek ze sekvence intenzitou výrazu. Nevýhodou systému MPEG-4 je neschopnost popsat všechny viditelné změny v obličeji a složitá interpretace FAP jednotek. Nevýhodou systému Feeltrace je nemožnost rozlišení některých emocí (stejná pozice ve 2D prostoru může být vysvětlena více než jednou emocí). Na druhou stranu umožňuje Feeltrace zobrazení vývoje výrazu v čase. Pro anotaci dat by bylo vhodné mít k dispozici jak popis FACS, tak i časový vývoj emoce z Feeltrace.

Přehled existujících databázi je v tabulce B.1. Realizované databáze pro spontánní emoce mají různé typy scénářů, každý z nich je vhodný pro získání určitých typů emocí. Nejsnažší je získání radosti a překvapení. Na druhou stranu strach, hněv, znechucení, smutek a mnoho dalších emocí je velmi obtížné získat. Seznam možných scénářů vhodných pro získání obtížnějších emocí je následující: Scénář databáze Adult attachment interview [95] je vhodný pro získání smutku a hněvu (B). Scénář databáze EmoTABOO [129] je vhodná pro získání rozpaků (B). Databáze pořízené při venkovních aktivitách (B) jsou vhodné pro získání strachu a deprese. Případy, kdy má subjekt vykonat určitou aktivitu, lze snadno získat znechucení a stres. Při

hraní her nebo sledování filmů lze získat frustraci, zájem a znudění.
Chceme-li vytvořit kvalitní databázi, měla by splňovat následující podmínky :

- Subjekt nesmí vědět, že se snažíme natáčet jeho emoce. Natáčení by mělo být provedeno skrytou kamerou. Pokud nejsou dodrženy tyto podmínky, může být subjekt ovlivněn a data jsou analýzu výrazu nevhodná.
- Databáze by měla obsahovat hraná i spontánní data. Oba typy dat jsou potřeba, chcemeli být schopní rozpoznat rozdíl mezi hranou a spontánní emocí. V případě hraných dat by měl jednotlivé emoce nebo akční jednotky předvádět herec nebo člověk se zkušeností s FACS anotací.
- Při výběru subjektů by se mělo hledět na dostatečnou variabilitu (pohlaví, národnost, věk, různý střih vlasů, brýle, bradka, knír).
- V případě hrané databáze by data měla obsahovat jak samostatné AU, tak jejich kombinace.
- Data by měla obsahovat jak přechod mezi emocemi (např. z překvapení do radosti či znechucení), tak i přechod mezi emocí a neutrálním výrazem.
- V průběhu natáčení by subjekt neměl být omezován v pohybu (různé natočení hlavy, volný pohyb rukou). Scéna, kde se emoce natáčí, by měla být komplexní (nemít konstantní jednolité pozadí, možnost dalších lidí ve scéně, proměnlivé osvětlení).
- Při volbě formátu záznamu by se mělo hledět na kvalitu versus velikost dat, aby bylo snadné data sdílet s ostatními výzkumníky.
- Následná anotace dat by měla být kvalitní. Nejlepší dostupnou reprezentací je FACS a Feeltrace. Spojení obou dohromady vytvoří kvalitní popis dat. Také by se mělo hledět na spolehlivost poskytnuté anotace. Anotaci by měli provádět minimálně dva lidé, aby se mohl porovnat jejich úsudek.
- Formát anotace by měl využívat jazyk EmotionML. Jednotnost formátu ulehčí práci jak při trénování, tak při testování FER systémů.

Nabízí se otázka proč nespojit všechny databáze a nevytvořit tak jednu velkou. Nespornou výhodou tohoto přístupu by byla možnost opakované anotace dat více pracovišti, čímž by se zvýšila věrohodnost jednotlivých anotací a díky jednotnému formátu by se ulehčila práce při trénování a testování FER systémů. Na druhou stranu vytvoření takové databáze má hned několik problémů. První problém jsou licenční ujednání, které je nutné podepsat pro každou databázi zvlášť. Druhým problémem je velký objem dat a tedy nutnost vytvořit centrální úložiště se snadným přístupem k jednotlivým videosekvencím. Třetím problémem je nejednotný formát anotace dat a tím nemyslím rozdíl mezi FACS, FeelTrace nebo kategorickým dělením. Nejednotnou anotací myslím různé druhy formátů souborů obsahujících anotaci dat. Čtvrtým a nejzásadnějším problémem je, že neexistuje osoba, či lépe organizace, která by se takového úkolu chtěla ujmout.

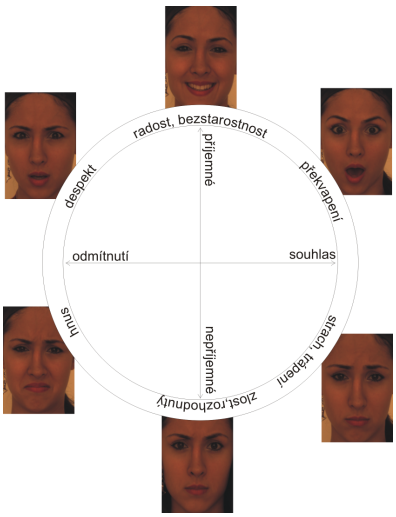
Tabulka B.1: Přehled existujících databází obsahujících video data (nebo sekvence snímků). V horní polovině spontánní data, v dolní polovině data hraná. Databáze označené * vznikly v rámci HUMAINE projektu. (#Sub./#Nah.) - Počet subjektů/počet video záznamů. Typ anotace: Feeltrace, FACS, kategorie - slovní hodnocení. typ: E - emoce, AU - akční jednotky. znak (?) znamená že informace není dostupná.

Jméno	#Sub./ #Nah.	Anotace	typ
AAI B	30/30	kategorie	E
Authentic Exp. B	28/?	kategorie	E
*Bel. naturalistic B	125/209	FeelTrace	E
*Bel. Activity B	?/?	FeelTrace	E
*Castaway R-TV B	10/10	FeelTrace	E
*DRIVAWORK B	24/24	FeelTrace	E
*EmoTABOO B	20/40	FeelTrace	E
*Green Persuasive B	16/32	FeelTrace	E
Moving F&P B	284/852	kategorie	E
*SAL B	4/4	FeelTrace	E
*Spaghetti B	118/708	FeelTrace	E
Companions B	65/65	žádná	E
MMI B	61/1250	FACS	E
Cohn-Kanade B	182/480	FACS	E
EFER B	20/384	kategorie	E
FABO B	23/1900	kategorie	E

Příloha C

Alternativní reprezentace výrazu tváře

Alternativou ke kategorickému dělení emocí je reprezentace z roku 1890 navržena Karlem Spencerem. Emoce je vyjádřena jako míra uvědomění, respektive míra charakteru osobnosti a jeho chování, ukázka na obrázku C.1. Koncept byl nadále rozvíjen (Wundt, Duffy, Schlosberg viz [44]) a v současné verzi jsou k dispozici tři dimenze. Nejčastěji jsou používány první dvě. První dimenze vyjadřuje hodnocení pocitu osoby, a to pozitivní versus negativní (nebo příjemný vs. nepříjemný). Druhá dimenze vyjadřuje, zda se člověk zachová v dané situaci aktivně či pasivně. Třetí dimenze vyjadřuje stupeň kontroly v dané situaci (vysoká versus nízká). V roce 2000 byl na základě této reprezentace vytvořen nástroj Feeltrace, viz obrázek C.2.

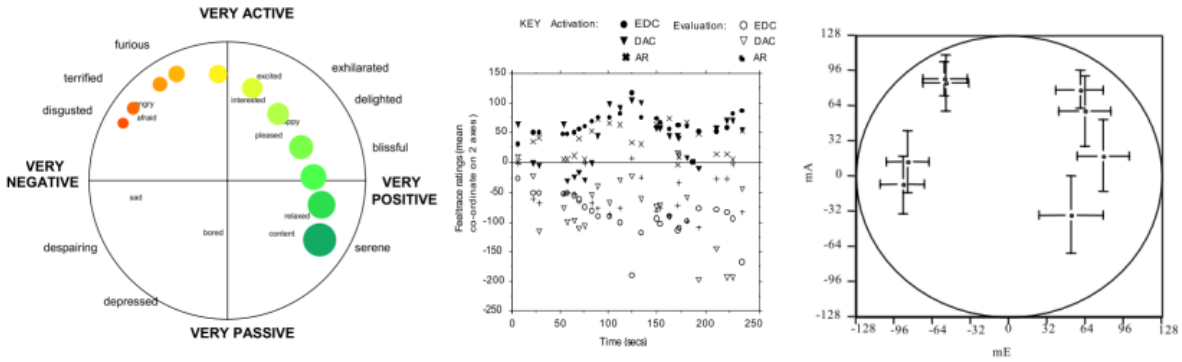


Obrázek C.1: Propojení kategorického dělení emocí s interpretací ve 2D prostoru. Snímek tváře zobrazuje kategorickou emoci, pozice snímku na ose reprezentuje vlastnosti ve 2D prostoru. Upraveno z originálu [44].

Alternativnímy systémy k systému FACS jsou Maximally Discriminative Facial Movement Coding System (MAX) a Moving Pictures Expert Group Facial Animation parameters (MPEG-4 FAP). Kromě zmíněných systémů existují další méně známé systémy, jejich přehled můžete nalézt v [53].

FeelTrace

FeelTrace byl původně navržen pro anotaci emočního obsahu v akustickém signálu [16]. Nicméně je vhodný i pro anotaci výrazu obličeje. Jeho nespornou výhodou je sledování emoce v čase. Na ose x je vyznačená evaluační složka pozitivní versus negativní, na ose y je vyznačena složka aktivační, aktivní versus pasivní. Pozorovatel má k dispozici myš k vyznačení aktuálního stavu ve 2D prostoru. Pro usnadnění práce jsou v prostoru vyznačeny kategorie popisující emoci. Nevýhodou reprezentace ve 2D prostoru je problém rozlišení některých emocí (např. strach od zlosti, které se ve 2D mohou překrývat). Anotace dat touto metodou vyžaduje velké množství zkušeností, a vzhledem k subjektovnímu hodnocení je vhodné aby stejná data anotovalo hned několik osob a výsledek se zprůměroval.



Obrázek C.2: Nástroj FeelTrace. Vlevo: Ukázka vizuální podoby Feeltrace. Uprostřed: zobrazení časové řady, vyplněné značky jsou z y-ové osy, prázdné jsou z x-ové, celkem hodnotily 3 osoby EDC, DAC a AR. Vpravo: výsledek evaluace všech klipů obsahujících emoce, výška a šířka elipsy zobrazují odchylku od středu emoce. Převzato z [16]

The MAX

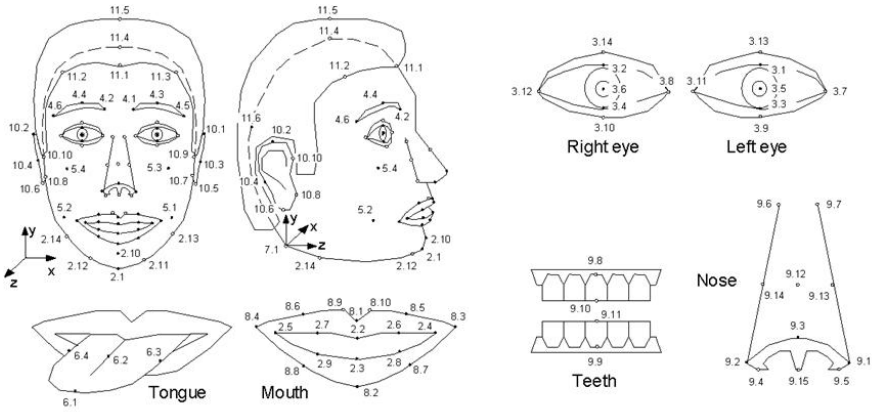
Systém MAX byl vytvořen C.E. Izardem v roce 1979 [45] a aktualizován v roce 1995 [43]. MAX byl vytvořen pro psychologický výzkum výrazu u nemluvňat a malých dětí. Přesto může být s modifikací použit na osoby jiných věkových skupin. MAX se skládá z 27 AC jednotek (apperance changes - změna vzhledu). Analýza výrazu pomocí MAX je provedena ve dvou krocích. V prvním kroku je obličej rozdělen do tří oblastí: 1) obočí, čelo, kořen nosu, 2) oči, nos, tváře, 3) rty a ústa. Jednotky jsou rozděleny následovně. Šest jednotek se vtahuje ke změně svalu v oblasti obočí. Osm jednotek je věnováno změně v oblasti oči-nos-tváře. Třináct jednotek se zabývá oblastí rty-ústa. Tři přidané jednotky jsou rezervovány pro nehodnotitelné, nesrozumitelné a klidové pozice. Regiony obličeje jsou analyzovány odděleně a každý region je

kódován nezávisle na změně ve zbylých dvou regionech. Díky tomu je zachována objektivita.

V druhém kroku jsou jednotky AC v každé obličejové oblasti klasifikovány, buďto jako jedna z osmi odlišných emočních stavů (zájem, radost, překvapení, smutek, zlost, znechucení, opovržení a strach), nebo jako komplexní výraz složený z několika současných dojmů. Stejně jako jednotky AU v systému FACS jsou jednotky AC zakořeněny v anatomii svalu obličeje. Na rozdíl od AU, soubor AC jednotek nezahrnuje všechny vizuálně odlišné obličejové pohyby (např. zvednutí vitřních a vnějších koutků obočí).

The MPEG-4 Facial Animation parameters

MPEG-4 FAP byl původně navržen pro syntézu obličejových výrazů. Objevilo se i několik studií [68, 42], využívajících popis k rozpoznávání. MPEG-4 definuje na neutrálním obličeji 84 charakteristických bodů. Tyto body jsou rozděleny do skupin tváře, oči a ústa. Na obrázku C.3 můžete vidět rozmístění těchto bodů.



Obrázek C.3: The MPEG-4 Facial Animation parameters. Pozice charakteristických bodů (Facial Point). Převzato z [83].

Charakteristické body vytváří základ pro definici 68 facial animation parameters (FAP). FAP jednotky reprezentují kompletní sadu základních obličejových změn, jako je pohyb hlavy, jazyku, očí a úst. Jednotky jsou přímo spjaty s pohybem obličejového svalu. První dvě jednotky slouží k hrubému vytvoření 6 základních emocí. Pro změny jednotlivých částí obličeje se využívají zbylé jednotky, pro více informací doporučuji [83]. Nevýhodou sady FAP jednotek je nekompletní popis všech viditelných změn ve výrazu obličeje.

Seznam publikovaných prací

1. J. Trojanová, P. Císař, and M. Železný. Combined visual parameterization for automatic lip-reading. *In Proceedings of the Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI'07*, Brno, 2007.

2. P. Císař, M. Železný, J. Zelinka, and J. Trojanová. Development and testing of new combined visual speech parameterization. *Proceedings of the workshop on Audio-visual speech processing*, pages 97–100, 2007.

3. A. Savran, O. Celiktutan, A. Akyol, J. Trojanová, H. Dibeklioglu, S. Esenlik, N. Bozkurt, C. Demirkir, E. Akagunduz, K. Caliskan, N. Alyuz, B. Sankur, I. Ulusoy, L. Akarun, and T. M. Sezgin. 3d face recognition performance under adversarial conditions. *In eNTERFACE'07*, pages 87–102, Universite catholique de Louvain., 2007.

4. J. Trojanová and Železný Miloš. Facial expression recognition based on dynamic textures. *In Proceedings of Measuring Behavior 2008*, pages 351–351, Maastricht, 2008.

5. J. Trojanová, M. Hrúz, P. Campr, and M. Železný. Design and recording of czech audio-visual database with impaired conditions for continuous speech recognition. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

6. P. Campr, M. Hrúz, and J. Trojanová. Collection and preprocessing of czech sign language corpus for sign language recognition. In ELRA, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

7. J. Trojanová, J. Vass, K. Macek, and Stluka P. Rojíček, J. and. Fault diagnosis of air handling units. *In Proceeding of the 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes*, pages 366–371, 2009.

8. B. Adessi, R. Ajaj, C. Arrat, I. Chabanaud, M. Courgeon, N. Deflache, N. Erdogmus, F. Gotusso, G. Himmetoglu, Ch. Jacquemin, L. Kessous, J. Martin, M. Osowski, T. Pachoud, and J. Trojanová. Capture and machine learning of physiological signals. *In eNTERFACE'08*, 2009.

9. D. Hurych, T. Svoboda, and US Y. Trojanová, J. and. Active shape model and linear predictors for face association refinement. *In Proceeding of the Computer Vision Workshops, 2009 IEEE 12th International Conference on Computer Vision*, pages 1193–1200, 2009.

10. J. Kukal, K. Macek, and Trojanová J. Rojíček, J. and. From symptoms to faults: Temporal reasoning methods. *In Proceeding of the International Conference on Adaptive and Intelligent Systems*, pages 155–159, 2009.

11. J. Vass, Fišera R. Trojanová, J. and, and J. Rojíček. Embedded controllers for increasing hvac energy efficiency by automated fault diagnostics. *In Proceeding of the First Workshop on Green and Smart Embedded System Technology: Infrastructures, Methods and Tools (GREEMBED 2010)*, 2010.

12. G. Swaminathan, S. Bedros, Y. U. S., and J. Trojanová. Multiple view face tracking, *Patent no. US20100316298 A1.* 2010 (Patent Application)

13. M. Hrúz, J. Trojanová, and M. Železný. Local binary pattern based features for sign language recognition. *Pattern Recognition and Image Analysis*, pages 398–401, 2011.

14. F. Juefei-Xu, M. Cha, M. Savvides, S. Bedros, and J. Trojanová. Robust periocular biometric recognition using multi-level fusion of various local feature extraction techniques. *In Proceeding of IEEE 17th International Conference on Digital Signal Processing*, pages 637–642, 2011.

15. J. Vass, J. Rojíček, and J. Trojanová. Thermal comfort monitoring in commercial buildings. *In Proceeding of the International Symposium on Sustainable Energy in Buildings and Urban Areas*, 2012.

16. J. Vass, J. Rojíček, and J. Trojanová. Setpoint optimization for air handling units. *Patent no. US20120232702 A1.* 2012 (Patent application)

17. V. Líbal, J. Trojanová, and L. Eswara. Detecting retail shrinkage using behavioral analytics, *Patent no. US20120169879 A1.* 2012 (Patent Application).

18. J. Trojanova, and S. J. Bedros. Quality driven image processing for ocular recognition system. *Patent no. US20120321142 A1*, 2012 (Patent Application).

19. K. Macek, J. Rojíček, and J. Trojanová. Fuzzy logic approach in temporal fault reasoning and application in air handling units, *Patent no. US8326790 B2.* Granted 2012

20. J. Trojanová, S. J. Bedros, G. Swaminathan, Y. U. S. Object alignment from a 2-dimensional image. *Patent no. US20130063417 A1.* 2013 (Patent Application)

21. S. J. Bedros, and Jana Trojanova. System and method for ocular recognition. *Patent no. US8385685 B2*, Granted 2013.