Jihočeská univerzita v Českých Budějovicích

Přírodovědecká fakulta

Forenzní analýza grafických souborů a video souborů

Bakalářská práce

František Dolejš

Vedoucí práce: Ing. Jaroslav Kothánek, Ph. D.

České Budějovice 2014

Bibliografické údaje

Dolejš F., 2014: Forenzní analýza grafických souborů a video souborů.

[Forensic analysis of graphic and video files. Bc. Thesis, in Czech] – 51 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic.

Anotace

Tato bakalářská práce se zabývá problematikou forenzní analýzy grafických souborů. Cílem je vybrat vhodné postupy, které pomohou detekovat pornografii. Na jejich základě má být vytvořen nástroj, který z množiny obrázků vybere potenciálně závadné.

Abstract

This thesis deals with the forensic analysis of graphic files. The goal is to select appropriate procedures to help detect pornography. Create a tool on their basis that selects potentially objectionable images from a set of images.

Klíčová slova

Detekce kůže, detekce obličeje, YCbCr, Haarova vlnka, AdaBoost, OpenCV, EmguCV, sexuální zneužívání dětí

Keywords

Skin detection, face detection, YCbCr, Haar-like features, AdaBoost, OpenCV, EmguCV, child sexual abuse

Prohlašuji, že svoji bakalářskou práci jsem vypracoval/a samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své bakalářské práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích dne 4. prosince 2014

……..

Poděkování

Rád bych poděkoval svému vedoucímu práce Ing. Jaroslavu Kothánkovi, Ph.D. za ochotu, věnovaný čas a poskytování rad. Dále za konzultace doc. RNDr. Ivě Dostálkové, Ph.D. a Mgr. Miloši Prokýškovi, Ph. D..

OBSAH

Úvod

S digitální technikou se dnes setkáváme denně. Používáme ji k práci, usnadnění života, ale i pro zábavu. Samozřejmě, ať už se jedná o výpočetní techniku, různé vstupní či výstupní periferie, úložná zařízení, komunikační technika atd. Všude tam se setkáme s digitální, popřípadě binární informací. V kriminalistice se využívá spíše pojem digitální informace, protože to je obecnější termín [1]. Binární forma je podmnožinou digitální formy.

Rychlé rozvoje počítačových zařízení a levné pořizovací náklady umožnily velké rozšíření do běžného užívání. Forenzní analytici toto musí sledovat a přizpůsobit tomu tak své postupy. Dostupnost výpočetní techniky a zařízení na ukládání dat umožňují pachatelům jejich využití při kriminální činnosti. Dnes jsou běžně k dostání datová úložiště v řádech TB. Zkoumání a bližší analýza obsahu tedy může být časově velice náročná. A procházení jednotlivých adresářů člověkem je tedy zdlouhavá činnost.

Další oblastí je rozšíření internetu, díky němu se objevil nový prostor pro páchání kriminální činnosti. Obrazové soubory se závadným obsahem je možné nekontrolovatelně šířit. Například ukládat na cloudová úložiště, šířit jejich odkazy nebo je posílat emailem atp. Lze tak pohodlně pracovat s velkým množstvím souborů, a ty distribuovat mezi početné množství osob.

Na to musely velké společnosti (Microsoft, Google, Facebook a Twitter) zareagovat a vytvořit nástroje pro detekci závadného obsahu.

Snahou je získání kontroly nad šířením tohoto obsahu a viníky postihovat. Například online nástroji: různá blokující rozšíření do prohlížeče a na vyšší úrovni, kontrola obsahu Gmail účtů, kontrola obsahu příspěvků na Facebooku nebo nástroje využívané policejními složkami a soudními znalci při zajištění výpočetní techniky podezřelého. Z této situace vychází požadavek na aplikaci.

První kapitola seznamuje se zásadami při zajišťování stop a digitální forenzní analýzou. Obecně popisuje termíny forenzní vědy. Druhá kapitola vysvětluje termín dětská pornografie a vyjmenovává právní řády, které ho obsahují. Třetí kapitola se zabývá reakcí velkých společností a organizací na problém šíření dětské pornografie. Čtvrtá kapitola uvádí do problematiky zpracování obrazu. Dále popisuje algoritmy (pro detekci částí lidského těla) využité v aplikaci. Kapitola pátá seznamuje s metodami detekce dítěte v obrazu. Dále v šesté kapitole je návrh a v kapitole sedmé stručný popis implementace. Osmá kapitola obsahuje výsledky testování a statistické vyhodnocení aplikace.

MOTIVACE A CÍLE PRÁCE

Tato bakalářská práce vzniká z několika důvodů. Na Ústavu aplikované informatiky se žádná bakalářská práce zatím touto problematikou nezabývá. Toto téma je důležité. Problematika výroby a šíření dětské pornografie je všeobecný problém. Děti je potřeba chránit. Následky psychického či fyzického zneužívání si nesou celý život.

Seznámit se s problematikou zásad prvotních úkonů při zajišťování digitálních stop pro účely forenzního zkoumání a řešení dětské pornografie.

Seznámit se s problematikou datových formátů grafických souborů a barevných modelů.

Vybrat metody pro analýzu obrazu a na jejich základě vytvořit aplikaci.

Ověření funkčnosti a přesnosti detekce aplikace.

1 DIGITÁLNÍ FORENZNÍ ANALÝZA A PROBLEMATIKA ZÁSAD ÚKONŮ PŘI ZAJIŠŤOVÁNÍ STOP

1.1 DIGITÁLNÍ FORENZNÍ ANALÝZA

Digitální forenzní analýza (DFA) patří mezi nejmladší forenzní vědy. Je známa pod různými názvy pár desetiletí. Za tu dobu prodělala výrazný vývoj.

Mezi forenzní vědy patří: daktyloskopie, forenzní antropologie, forenzní balistika, forenzní chemie, soudní lékařství, forenzní psychologie, forenzní genetika, písmoznalectví a další vědy podle povahy problému.

DFA patřící do široké skupiny forenzních věd, zkoumá jakákoli digitální data. Každá forenzní věda má svůj neforenzní obor jako svoji „matku“. Může se zdát, že pro DFA je takovýmto oborem informatika. Avšak DFA má velice široký okruh využití. Dnes má mnohem širší využití než v minulosti. Zabývá se zkoumáním velkého množství různých druhů trestního nebo jiného protiprávního jednání. Vyskytuje se všude, kde se můžeme setkat s digitálními informacemi. Například u výpočetní techniky, komunikační techniky, digitální zařízení – fotoaparáty, videokamery atd. Digitální informace nás denně provázejí na každém kroku.

1.2 DEFINICE DIGITÁLNÍ STOPY

Digitální zařízení zpracovávají data. Při této činnosti po sobě zanechávají určité záznamy. Z kriminalistického hlediska to jsou stopy.

„Digitální stopa je jakákoliv informace s vypovídající hodnotou, uložená nebo přenášená v digitální podobě.“ [1] Takto definuje digitální stopu Viktor Porada a Roman Rak ve svém článku. Tato definice pokrývá širokou škálu zařízení a to je správně. Zahrnuje veškerou počítačovou a telekomunikační techniku, přenos a pořizování videa a fotografie, dat elektronických zabezpečovacích systémů.

Je potřeba je chápat v širokém kontextu, neboť nové technologie se stále vyvíjejí ajsou tedy zařazeny do stejného okruhu.

1.3 ZÁSADY ÚKONŮ PŘI ZAJIŠŤOVÁNÍ DIGITÁLNÍCH STOP

Jak popisuje Ing. Marián Svetlík ve svém článku [2], ne však každé zkoumání digitálních dat má forenzní charakter. Aby závěry analýzy byly využitelné jako důkaz pro soudní účely, musí splňovat základní požadavky. Toto ovšem není nikde právně nebo jinak zakotveno a vychází se jen z osvědčených postupů a doporučení.

- Legalita – všechny informace, stopy, vzorky atd. musejí být získány legálním způsobem.
- Integrita – veškeré postupy a způsoby práce s informacemi musejí být prováděny tak, aby nemohlo dojít k úmyslné či neúmyslné manipulaci s nimi.
- Opakovatelnost – tj. použít takové postupy a způsoby práce, které mohou být zopakovatelné a došlo při nich ke stejným výsledkům.

2 DĚTSKÁ PORNOGRAFIE A PRÁVNÍ ŘÁDY

Slovo pornografie vychází z řeckého slova porné, což v překladu znamená děvka, nevěstka, prostitutka. Pornografie je neumělecké znázornění lidského těla či sexuálního chování, které nemá jiný účel, než podněcovat sexuální pud [3].

Pojem dítě je v různých právních řádech vysvětlován odlišným způsobem. Podle čl. 1 písm. a) Rámcového rozhodnutí Rady 2004/68/SVV ze dne 22. prosince 2003 o boji proti sexuálnímu vykořisťování dětí a dětské pornografii je hranicí pro pojem dítěte 18 let.

Je možné se ovšem setkat u severních států jako je Švédsko nebo Norsko, že hranice není pevně určena. Pojem je stanoven tak, že dětská pornografie je znázorňování sexuálních úkonů na osobách, jejichž tělesný vývoj ještě není nebo se nezdá být ukončen. V Norsku je hranice stanovena na 16 let.

Dle ustanovení § 126 zákona č. 40/2009 Sb., trestního zákoníku, ve znění pozdějších předpisů, je dítětem osoba mladší 18 let.

Poprvé se dětská pornografie dostává do českého trestního práva přijetím zákona č. 557 /1991 Sb., a dále se rozšiřuje roku 1992. Významného doplnění se dočkal roku 2007 přijetím zákona č. 271/2007 Sb., kdy se začalo postihovat přechovávání dětské pornografie a zneužití dítěte k výrobě pornografie.

3 REAKCE FIREM NA ŠÍŘENÍ

Velké společnosti v oblasti online služeb se stále více zajímají o bezpečnost na internetu.

V praxi například společnosti Google s Microsoftem představily opatření pro blokování vyhledávání dětské pornografie. Kromě toho Google nabídl technickou podporu britské organizaci Internet Watch Foundation a americkému národnímu středisku pro pátrání po pohřešovaných a zneužívaných dětech.

3.1 GOOGLE

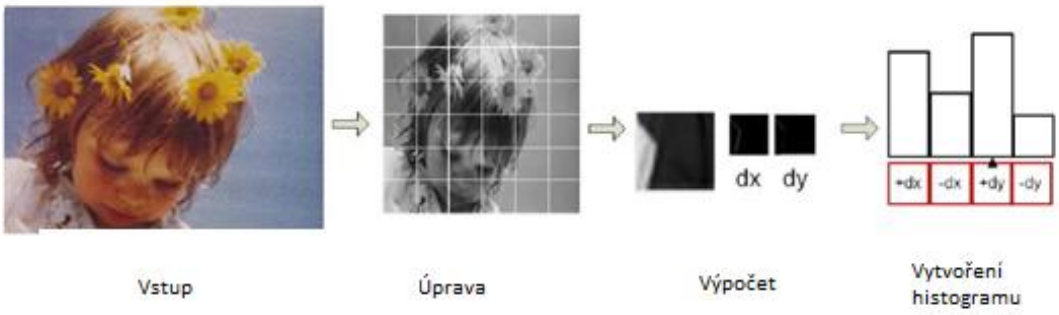
Společnost Google mimo jiné vyvinula program, který prohledává Gmail účty všech uživatelů. Program hledá obrázky dětské pornografie. Podezřelé účty předává dále k prověření. Mimo Google vyhodnocují účty svých uživatelů také například společnosti Microsoft, Facebook a Twitter.

3.2 MICROSOFT PHOTODNA

V roce 2009 Microsoft Digital Crimes Unit (DCU) a National Center for Missing and Exploited Childern (NCMEC) vytvořili produkt PhotoDNA. Jedná se o nástroj pro boj s šířením dětské pornografie.

PhotoDNA umožňuje vytvoření unikátní digitální značky (hash) obrázku. Tento hash lze poté porovnat s již vytvořenými značkami. Je tak možné ve skupině fotografií pomocí vytvořené značky identifikovat stejný obrázek.

Microsoft daroval NCMEC a dalším online službám bezplatnou licenci tohoto nástroje. NCMEC vytvořil PhotoDNA značky nejhorších známých online obrázků dětské pornografie. Databáze obsahuje jen značky, nikoli obrázky samotné. Databáze je sdílena mezi poskytovateli online služeb např. Microsoft (Bing, SkyDrive, Hotmail), Twitter a Facebook.



Obrázek 1 – PhotoDNA [4]

Obrázek se převede do černobílé a upraví se jeho velikost. Dále se rozdělí na části a ty se následně analyzují.

Metoda může připomínat klasický hash. PhotoDNA můžeme nazvat robustnějším hashem. Technologie je založená na podstatě obsahu obrazu, ne souboru. Velkou výhodou je tedy, že obrázku může být změněna velikost, barva, a může být uložen v jiném souborovém formátu nebo jinak podobně pozměněn, přesto si s tím PhotoDNA dokáže poradit.

Orgány činné v trestním řízení mohou získat přímo zdrojové kódy PhotoDNA nebo vybrané nástroje, které tuto technologií využívají, např. NetClean Analyze.

4 METODY ZPRACOVÁNÍ OBRAZU

4.1 POČÍTAČOVÉ VIDĚNÍ (COMPUTER VISION)

Digitální data mají mnoho podob: text, obraz, hudba, video… Tato data se zpracovávají, vyhledává se v nich atd. U textové podoby informace je to jednoduché, vyhledávají se jednotlivá slova. Ovšem u obrazu je to složitější, musí být indexován a popsán. Při zpracování se pracuje s těmito textovými informacemi. U velkých objemů dat je takový popis časově náročný a nákladný, protože tuto práci musí dělat člověk. Dlouhodobě je snaha toto pomalé zpracování nahradit počítačem. A to dalo vzniku oboru počítačové vidění.

Je to relativně nový obor, který se nadále rozvijí. Jde o oblast výpočetní techniky a vývoje softwaru zabývající se zařízeními, která dokážou ze snímaného obrazu získat informaci. Toto odvětví nalezne využití v mnoha oblastech:

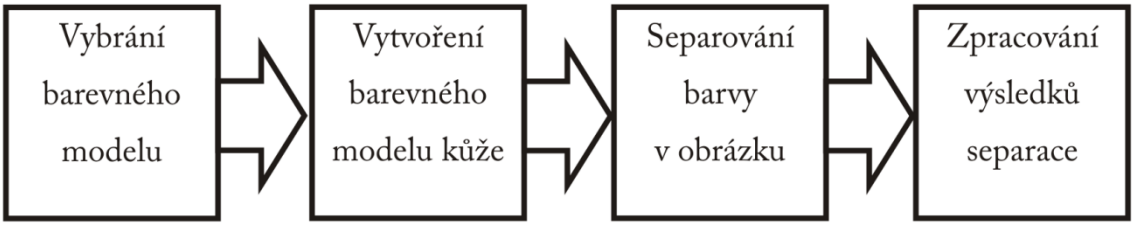
- Ovládání procesů – průmysloví roboti, autonomní vozidla
- Detekce jevů – u kamerových systémů jako čítače a detekce pohybu osob
- Organizace informací – databáze obrázků a videí
- Kontrolní úlohy – kontrola povrchů a výrobků ve výrobě
- Modelování objektů nebo prostředí – analýza lékařských snímků
- Interpretace scény
- Interakce

4.2 METODA ROZPOZNÁVÁNÍ LIDSKÉ KŮŽE

K detekci lidského těla lze přistupovat z různých úhlů. Do jisté míry se jedná o specifický objekt. Pokud je k dispozici barevný obrázek, lze využít detekci těla pomocí jeho barvy [5].

Tato metoda vychází z předpokladu, že lidská kůže má svoji charakteristickou barvu. Zabírá tedy určitou část barevného modelu – pro kůži lze vytvořit vlastní

barevný model. Mezi metodami založenými na invariantních rysech je tato velmi populární. Je to i díky rychlosti zpracování, robustnosti vůči poloze těla a robustnosti vůči kvalitě obrázku.



Obrázek 2 - Schéma procesu.

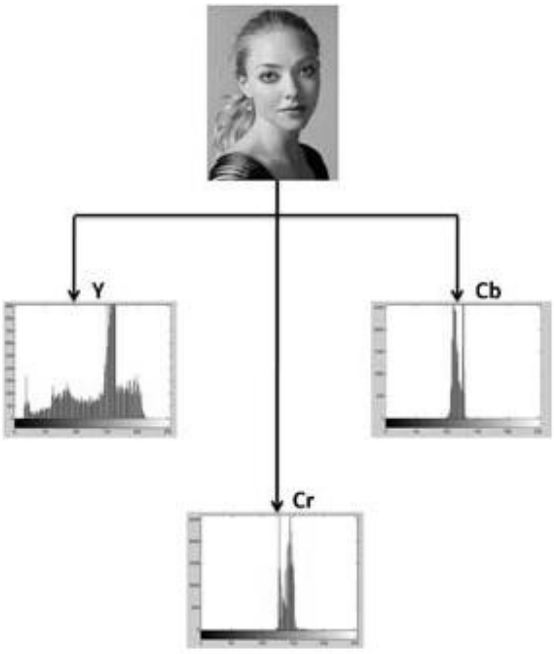
Pro další zpracování lze vybrat jakýkoli barevný model, ale některé jsou pro tuto metodu vhodnější [6]. Obecně se pro detekci lidské kůže hodí barevné modely, které mají složku jasu oddělenou od barevné. N. Sarris a další [5] použil jen Cb a Cr k detekci obličeje v barevných obrázcích. Složka Y je vynechána, jak je vidět na obrázku 3 a 4. Hodnota Y se velmi liší. Oproti tomu složky Cb a Cr mají v histogramech určitě rozmezí. V. Neagoe [7] navrhl systém detekce obličeje využívající tento barevný model.

Byl vybrán model YCbCr, kde Y je složka jasu a pro detekci kůže se nevyužije. Cb a Cr jsou složky barevné informace. Tento model je dále využit v aplikaci.

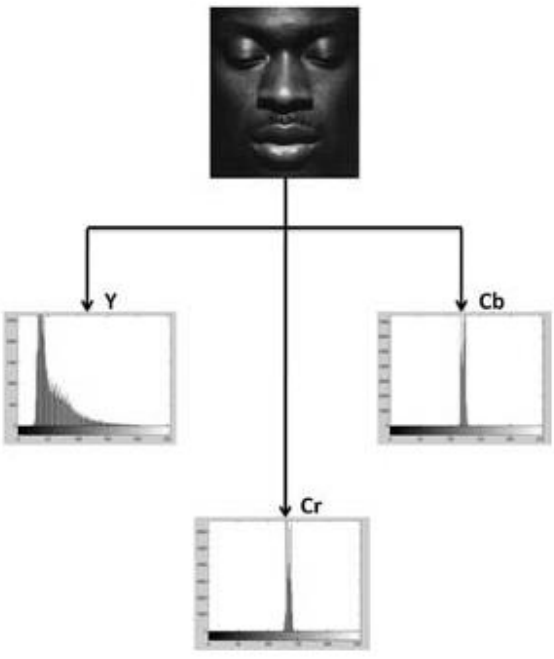
Pro ukládání obrázků se standardně využívá barevný model RGB. Do YCbCr se barvy musejí přepočítat.

$$Y = 0{,}299 * R + 0{,}587 * G + 0{,}114 + B$$
$$Cb = 128 - 0{,}168736 * R - 0{,}331264 * G + 0{,}5 * B$$
$$Cr = 128 + 0{,}5 * R - 0{,}418688 * G - 0{,}081312 * B$$

Nyní, když máme vybrán barevný model, ve kterém budeme zpracovávat obrázek, je potřeba určit barevný model kůže. Tělo má specifický odstín barvy.



Obrázek 3 - Histogram bělošky [6]



Obrázek 4 - Histogram černocha [6]

Z histogramů je patrné, že složka Y je pro náš účel nepoužitelná. A můžeme z nich vyčíst nejlepší možný rozsah hodnot pro různé lidi s odlišnou barvou kůže. Tato metoda je účinná pro detekci různých lidských ras. Podle [6] vychází rozsah hodnot:

$$80 \leq Cb \leq 120$$
$$133 \leq Cr \leq 173$$

Když máme rozsah hodnot odpovídající kůži, můžeme přistoupit ke zpracování obrázku. Vezmeme obrázek s barevným prostorem RGB a pomocí výše uvedených rovnic jej převedeme do YCbCr. Pixel po pixelu RGB hodnoty přepočítáváme do YCbCr. Když máme obrázek v YCbCr, procházíme obrázek znovu a nyní klasifikujeme hodnoty na pozitivní detekci - pixel obsahuje kůži. Jeho hodnota je ve výše uvedeném rozsahu. Anebo naopak jde o negativní detekci - pixel neobsahuje kůži. Tedy jeho hodnota není z výše uvedeného rozsahu.

Jako výsledek máme vyseparovanou oblast. Výsledky je možné zobrazit různými způsoby. Například černé pozadí (oblast bez kůže) a na vyseparovanou oblast zobrazit originální obrázek. V příloze je příklad separace obrázku (Příloha 2).

4.3 DETEKTOR OBJEKTŮ VIOLA-JONES

Studiem literatury a zdrojů dostupných na internetu byla vybrána multiplatformní knihovna OpenCV. Jedná se o svobodnou a otevřenou knihovnu pro manipulaci s obrazem. Je zaměřena především na počítačové vidění a zpracování obrazu v reálném čase. V této práci se zaměříme na její odnož EmguCV. Jedná se o OpenCV pro platformu .NET.

4.3.1 KNIHOVNA OPENCV

Jedná se o otevřenou multiplatformní knihovnu, která je zaměřená především na počítačové vidění a zpracování obrazu v reálném čase. Z počátku byla vyvíjena společností Intel. Knihovnu lze využívat z prostředí C/C++, dále pak s generátorem rozhraní SWIG v Pythonu a Octave. Pomocí wrapperu EmguCV také v prostředí .NET. Že se jedná o skutečně mutiplatformní knihovnu dokazuje fakt, že je možné knihovnu spouštět na Windows, Linux, Mac OS X, iPhone, iPad a Android. V příloze (Příloha 1) je architektura EmguCV [8].

Obrázek 5 - Logo knihovny OpenCV [9]

Knihovna počítačového vidění obsahuje několik stovek algoritmů, především z oblasti filtrace a transformace obrazových informací a strojového učení. Mezi vývojáři je hojně využívaná a stále prochází aktivním vývojem. Podobně zaměřených knihoven je více, ale málokterá je tak komplexní.

4.3.2 VYUŽITÍ EMGUCV V TÉTO PRÁCI

Další metoda, která je začleněna do aplikace je implementována již v EmguCV. Využívá tedy této funkce v knihovně. Jedná se o detektor Viola-Jones [10]. Níže je popsán princip.

Detektor Viola-Jones byl roku 2001 představen Paulem Violou a Michaelem J. Jonesem. Později jej pak vylepšili Rainer Lienhart a Jochen Maydt [11] použitím diagonálních vlnek. Jde o detektor objektů. Technologie je založená na metodách pro detekci jednotlivých objektů a jejich klasifikaci do tříd. Jako vstup požaduje šedotónový obraz. V této metodě detekce vzorů se setkáme se třemi základními pojmy: integrální obraz, Haarova vlnka a klasifikační algoritmus AdaBoost. Metoda je známa svojí rychlostí, spolehlivostí, nezávislostí na jasu obrazu a velikosti objektu.

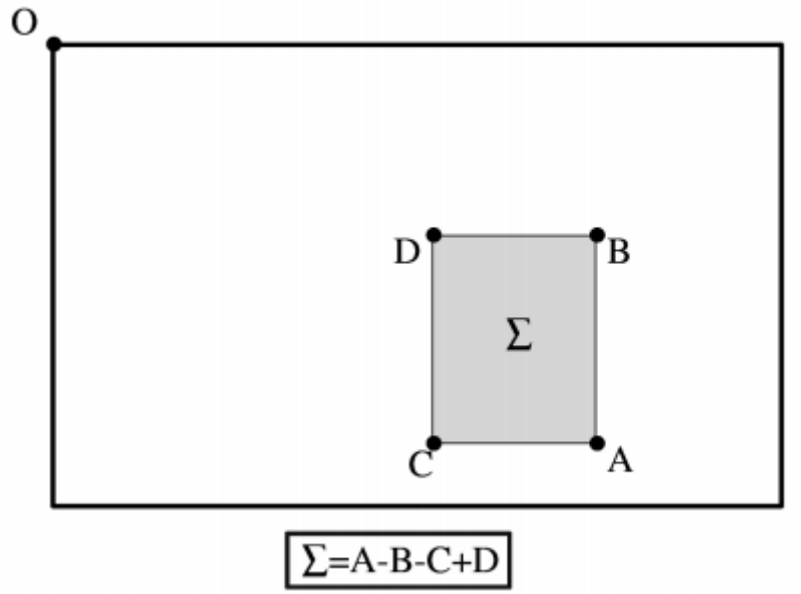
Vlastní detekce objektů se provádí tak, že vstupní obraz je procházen podoknem, které mění svoji pozici a velikost. Podokno vybere příznaky pro klasifikátor a ten rozhodne, zda se shodují s pozitivními vzory.

Výpočetní náročnost je snížena kaskádovým zapojením jednotlivých klasifikátorů.

Většinu podoken vyřadí již první stupeň kaskády a tím dochází k výraznému uspoření času.

4.3.3 INTEGRÁLNÍ OBRAZ

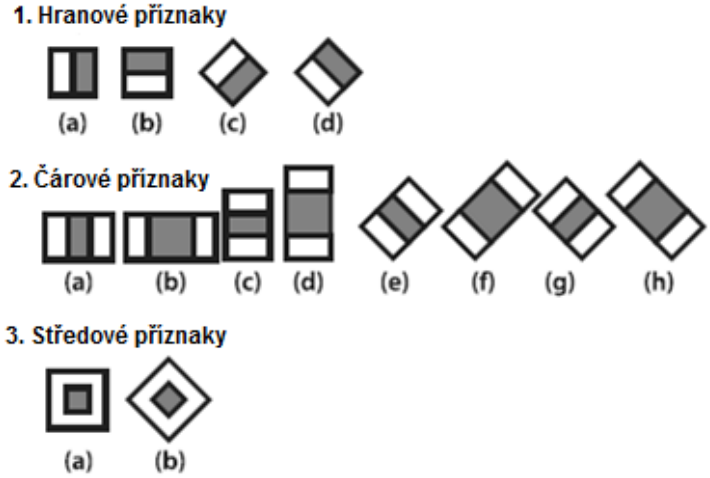
Jedná se o způsob reprezentace obrazu a slouží k zefektivnění výpočtu hodnot jednotlivých příznaků. Každý bod v obrazu představuje součet hodnot předchozích pixelů doleva a nahoru. Z toho vyplývá, že pravý spodní roh představuje součet všech pixelů obrazu.



Obrázek 6 - Výpočet hodnoty ∑ plochy Sigma vymezenou vrcholy A, B, C a D v integrálním obraze.

4.3.4 HAAROVA VLNKA

Jako vstup do procesu učení klasifikačního algoritmu je množina příznaků. Pro detektor Viola-Jones je potřeba větší množství jednoduchých příznaků. Tomu odpovídají příznaky na principu podobnému Haarově vlnky (Haar-like features).



Obrázek 7 - Haar-like features

Zde jsou uvedeny základní používané příznaky mezi něž patří hranové, čárové a středové. Každý příznak je rozdělen na minimálně dvě části a jeho hodnota se vypočítá jako suma pixelů ve světlé části a od nich se odečte suma pixelů tmavé části. Příznaky jsou použity na vstupní obraz, kde mění svoji velikost od 1x1 až po velikost obrazu. Tato sada pak slouží jako vstup pro AdaBoost, který z nich vybere.

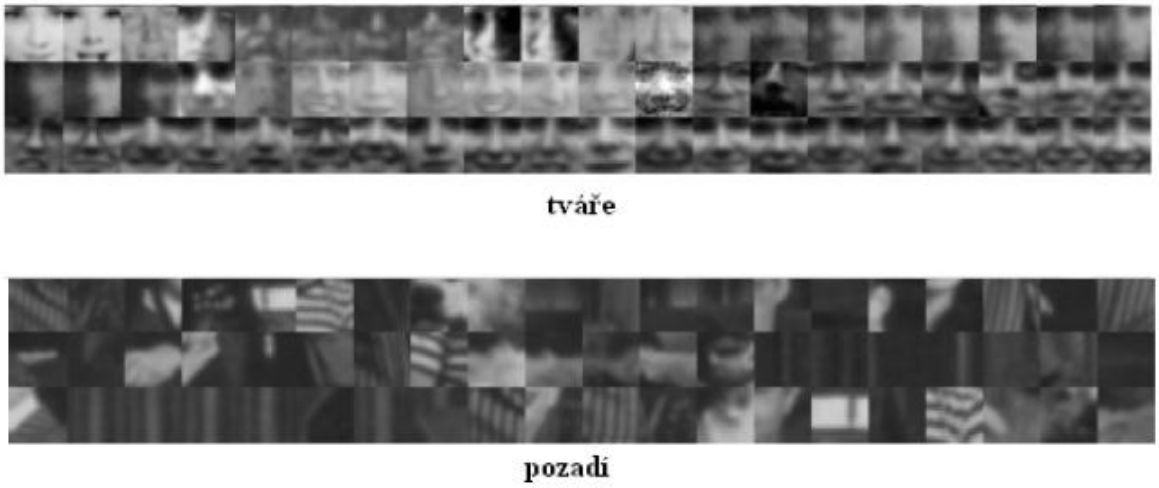
4.3.5 KLASIFIKAČNÍ ALGORITMUS ADABOOST

Vychází z metody strojového učení boosting, která má za cíl zlepšení klasifikační přesnosti libovolného algoritmu strojového učení.

Základem je vytvořit více jednoduchých klasifikátorů (weak learners) s přesností něco málo přes 50%. Tím je vytvořen soubor klasifikátorů, který je označován jako strong learner.

Je potřeba vytvořit dvě sady – trénovací množiny vzorů. Jedna bude obsahovat pozitivní vzory a druhá s negativními. Například při vytváření detektoru obličeje, vypadá taková sada obrázků jako na obrázku číslo 10. První skupina obsahuje obličeje a druhá nikoli.

Na základě trénovací množiny (pozitivní obrázky) se AdaBoost snaží vytvořit soustavu slabých klasifikátorů (kterým jsou nastaveny co nejvhodnější váhy). Celý algoritmus AdaBoost viz. příloha číslo 3.



Obrázek 8 - Sada pozitivních a negativních vzorů

V této práci je využito již předtrénovaného detektoru Viola-Jones pro detekci tváří. Při spojení s detekcí kůže se navzájem tyto metody doplňují a zlepšují výslednou detekci. V dalších kapitolách bude popsána aplikace a její návrh.

4.4 PŘEHLED DALŠÍCH METOD A VYLEPŠENÍ

- Vylepšení rychlosti metody detekce kůže (A new fast skin color detection Technique [12]) – Lze použít v systémech zpracování v reálnem čase. Namísto testování každého pixelu, přeskakujeme sadu pixelů. Důvodem je, že u pixelu kůže budou s vysokou pravděpodobností sousední pixely také s kůží.

- Metoda extrakce a detekce kůže z PDF dokumentů (A new system for extracting and detecting skin color regions from PDF documents [13]) – Návrh nástroje pro extrahování a detekci kůže v PDF dokumentech.

- Metoda adaptivní detekce kůže (Adaptive skin segmentation in color images [14]) – Založena na Bayesovské teorii rozhodování. Navržený

postup je nový ve využití charakteristiky lidské kůže k výběru vhodné prahové hodnoty pro pokožku.

- Metoda detekce kůže podpořena SVM (Adult image detection method base-on skin color model and support vector machine [15]) – Nejprve se detekuje kůže a následně se vyhodnocuje pomocí SVM (Support vector machine). Jedná se o metodu strojového učení. V úloze klasifikace SVM hledá nadrovinu, která v prostoru příznaků optimálně rozděluje trénovací data.

- Metoda detekce kůže s analýzou nalezených oblastí (An algorithm for nudity detection [16]) – Detekce nahoty založena na barevných modelech RGB, normalizovaném RGB a HSV. Nalezené plochy kůže jsou dále analyzovány, například jejich velikost a relativní vzdálenosti od sebe. Na základě této analýzy a procentuálním zastoupení kůže v obrázku je obrázek klasifikován.

- Metoda detekce kůže s vylepšenými vlastnostmi (Appearance-based nude image detection [17]) – Zpracovávaný obrázek prochází víceúrovňovou klasifikací. Nejprve se detekuje obličej, poté kůže (se kterou se dále pracuje – adaptivní segmentace kůže) a následuje SVM klasifikace.

- Metoda detekce kůže s algoritmem RSOR (Approach of RSOR algorithm using HSV color model for nude detection in digital images [18]) – Metoda detekce kůže vylepšená o RSOR algoritmus. Na největší detekovanou oblast kůže se použije RSOR pro výpočet procentuálního zastoupení, se kterým se dále pracuje.

- Metoda vyhledání částí těla a analýza jejich geometrie (Finding naked people [19]) – Metoda dává nový pohled na vyhledávání pornografie. Nejprve se hledají části s kůží, které se následně spojují do skupin. Dále se analyzuje jejich geometrie a vztahy v porovnání s reálným tělem.

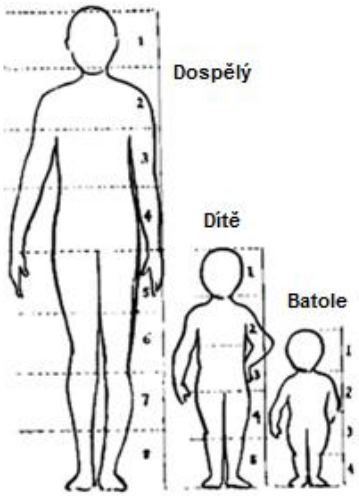
- Metoda detekce kůže založená na adaptivním a rozšiřitelném barevném modelu kůže (Naked image detection based on adaptive and extensible skin color model [20]) – Využití metody je hlavně u obrázků se speciálními světelnými podmínkami, kde kůže může splývat s barvou pozadí.

5 ROZPOZNÁNÍ DÍTĚTE NA OBRÁZKU

Tato kapitola představuje konkrétní metody zpracování a čtení obrazu se zaměřením na určení věku osob. Kapitola obsahuje popis dvou metod. První se zabývá klasifikací osob na základě poměru velikostí částí těla osoby a druhá se zabývá rozdělením osob do tří kategorií (dítě, dospělý, senior) na základě rysů v obličeji.

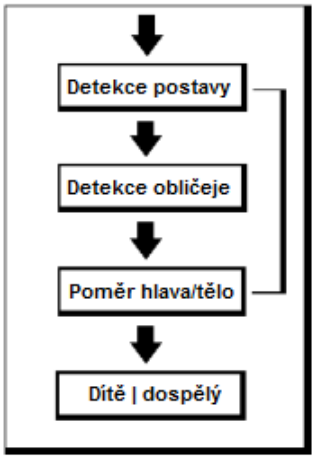
5.1 METODA 1 – POMĚR VELIKOSTI ČÁSTÍ TĚLA OSOBY

Metoda vychází z faktu, že v mládí je poměr velikosti hlavy k tělu jiný než v dospělosti. U dítěte zabírá hlava určité procento výšky postavy. U dospělého člověka je toto procento nižší.



Obrázek 9 – Výška člověka v průběhu života [21]

Blokový diagram na obrázku číslo 10 znázorňuje algoritmus klasifikace osob. Je rozdělen na čtyři hlavní části – detekce osoby, detekce obličeje, výpočet poměru a klasifikace.



Obrázek 10 – Blokový diagram detekce dítěte a dospělého v obrázku [22]

K detekci osoby je využita Haarova vlnka a algoritmus AdaBoost, které jsou popsány v předchozí kapitole.

V dalším kroku je nalezena hlava osoby. K tomu je využit stejný způsob jako u prvního kroku.

Ve třetím kroku je vypočten poměr hlavy a těla podle vzorce:

$$r = \frac{l_H}{l_T} \quad 0 < r < 1$$

l_H je výška hlavy a l_T je výška trupu. Z obrázků samozřejmě není možné získat absolutní výšku osob. Používá se relativní velikost z obrázků. Po aplikování vzorce lze podle výsledku osobu přiřadit do jedné ze skupin. Pro dospělé osoby vychází tento poměr kolem 0,15 a pro dítě je to kolem 0,2.

Algoritmus nejlépe funguje na obrázky stojících postav a chodců. Platí i pro video v reálném čase. U sedících osob není možné správně určit jejich relativní velikost.

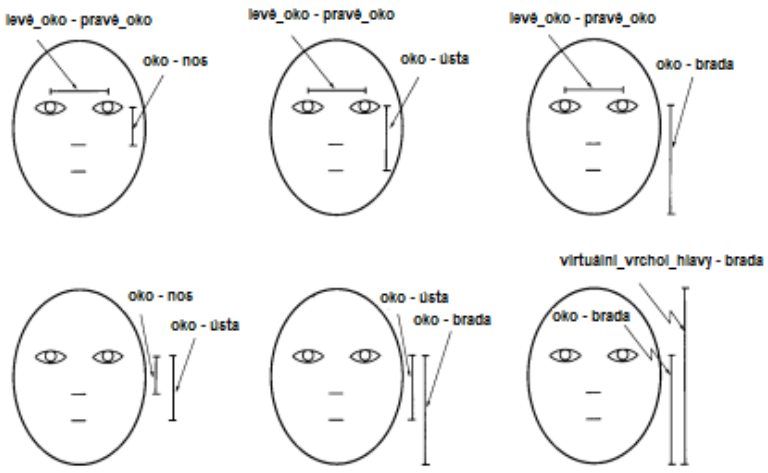
5.2 METODA 2 – RYSY V OBLIČEJI OSOBY

Druhá metoda je v jistém pohledu robustnější. Aplikuje se jen na obličej osoby. Není nutné mít obrázek celého člověka, zato ale obrázek musí mít lepší rozlišení než u metody 1, kde postačuje rozeznat hrubou velikost částí lidského těla.

Algoritmus metody [23]:

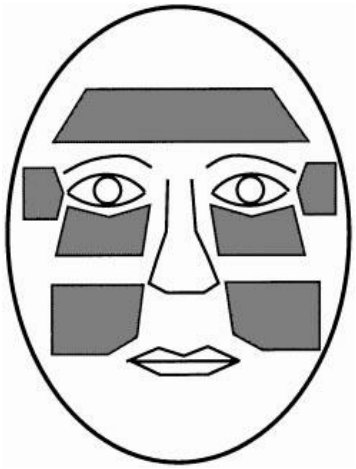
A. Najít rysy obličeje
    1. Najít hrubý ovál
    2. Najít bradu; upravit ovál
    3. Najít strany tváře
    4. Vypočítat virtuální vrchol hlavy
    5. Najít oči
    6. Najít ústa
    7. Najít nos
B. Vypočítat poměry rysů obličeje
C. Vypočítat analýzu vrásek
D. Spojit B a C; určení věkové kategorie

Z níže uvedeného obrázku je patrné, jaké poměry rysů obličeje jsou počítány. Pomocí těchto poměrů lze odlišit dítě od dospělého a seniora.



Obrázek 11 – Vypočítávané poměry rysů obličeje [23]

Dále se zaměřujeme na místa v obličeji s nejčastějším výskytem vrásek – čelo, kolem očí atd. Tato místa jsou podrobena analýze a detekci vrásek.



Obrázek 12 – Místa pro analýzu vrásek [23]

Pomocí analýzy vrásek lze odlišit seniora od dospělého a dítěte.

Spojením výsledků vypočítaných poměrů rysů obličeje a analýzy vrásek, lze s jistou přesností klasifikovat digitální obrázek obličeje osoby do jedné ze tří kategorií – dítě, dospělý, senior.

6 POŽADAVKY A NÁVRH APLIKACE

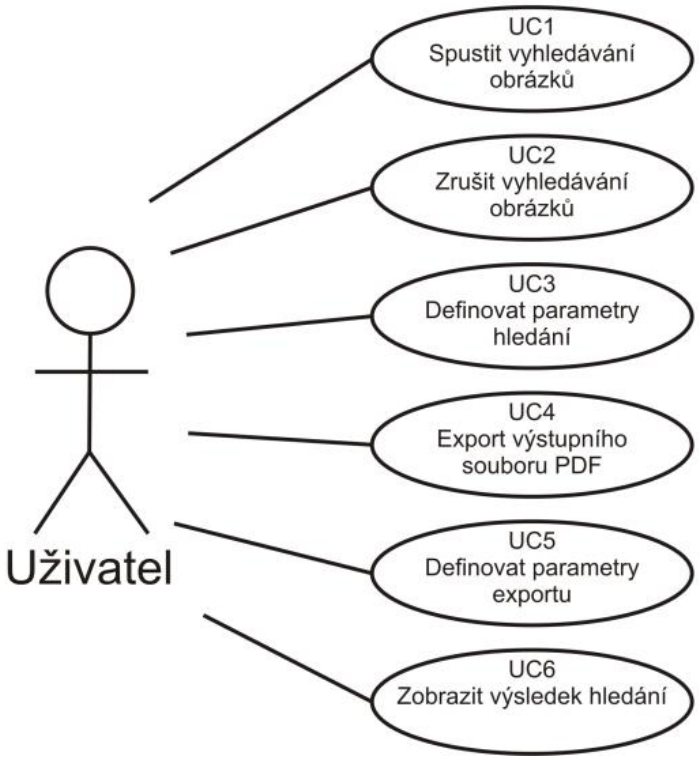
Tato kapitola se zabývá požadavky na aplikaci a jejím návrhem.

6.1 POŽADAVKY NA APLIKACI

- Do aplikace budou implementovány metody analýzy grafických souborů.
- Vstupem pro aplikace bude adresář.
- Aplikace bere v potaz soubory JPG, PNG, TIF a BMP.
- Aplikace umožní export výstupu do souboru PDF.
- Aplikace bude vyhledávat potenciálně závadné obrázky.

6.2 USECASE DIAGRAM

Use Case diagram (diagram případů užití) – vychází ze zadání a cílů (obecně z požadavků zákazníka). Jedná se o tzv. mapování uživatelských požadavků na jednotlivé Use Case. Diagram vypovídá o tom, co systém bude umět.



Obrázek 13 - UseCase diagram

6.3 POUŽITÉ TECHNOLOGIE

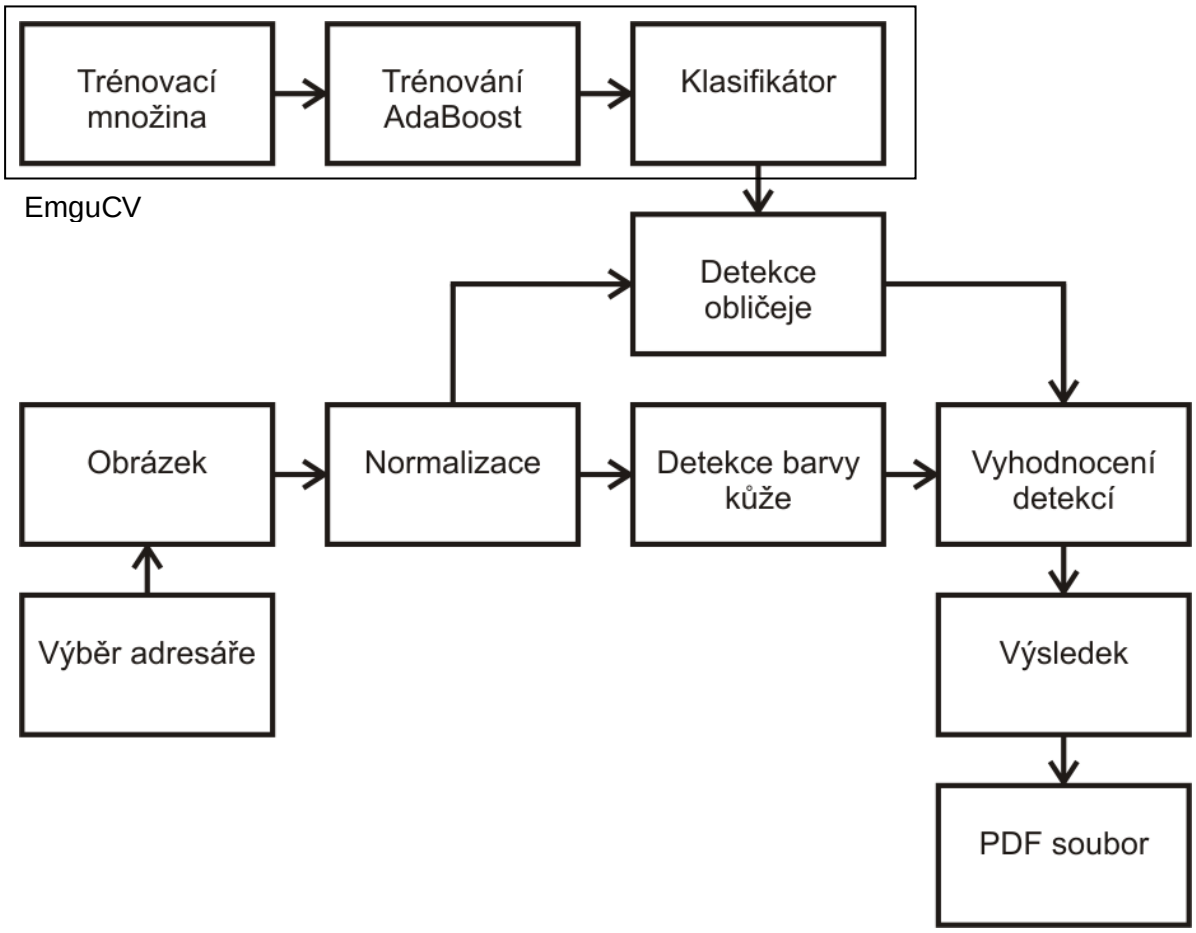
Aplikace je vytvořena na platformě Microsoft .NET Framework 4.0 a naprogramována v jazyku C#.

Jako vývojové prostředí bylo využito nástroje společnosti Microsoft – Visual Studio 2010.

Pro vytváření uživatelského prostředí Windows aplikací nabízí platforma .NET jmenný prostor Windows.Forms.

Pro detekci obličeje je využito EmguCV (emgucv-windows-x86-gpu 2.4.2.1777). Díky němu lze využít funkce knihovny OpenCV na platformě .NET.

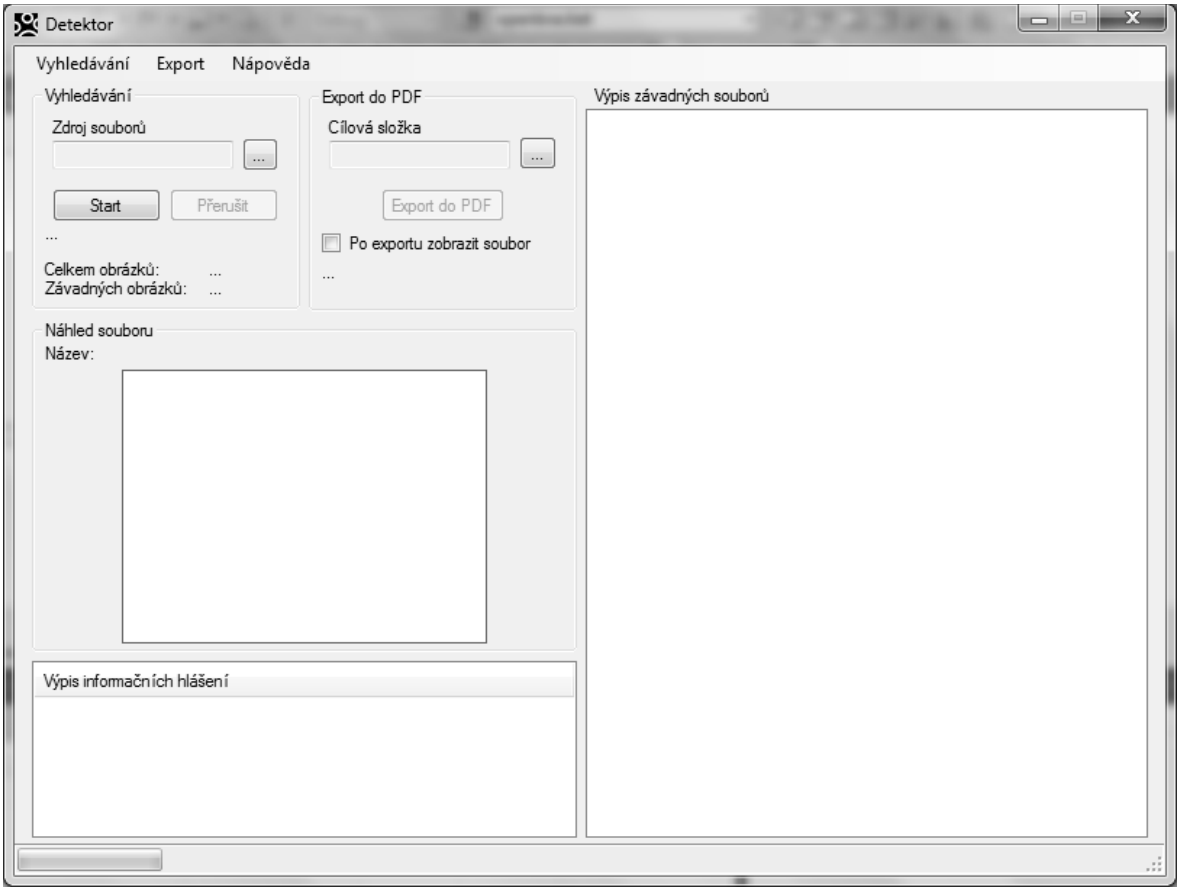
Dále pro export výsledků do PDF souboru je využita otevřená knihovna PDFsharp.



Obrázek 14 - Fáze programu

6.4 NÁVRH GRAFICKÉHO PROSTŘEDÍ

Při návrhu grafického rozhraní bylo využito výše zmíněné Windows.Forms ve Visual Studiu 2010. Snahou bylo udělat jej co možná nejintuitivnější a jednoduché na ovládání, ale neopomenout při tom funkčnost. Okno aplikace má pevnou velikost 800x600 pixelů. Grafické rozhraní lze rozdělit na menu a dvě hlavní části.



Obrázek 15 - Grafické prostředí aplikace

Pravou část grafické rozhraní zabírá okno na zobrazování výpisu závadných souborů. Jejich náhledy lze zobrazovat v levé části.

Levá část grafického rozhraní aplikace poskytuje ovládací prvky k vyhledávání závadných obrázků a k exportu výsledků aplikace do přehledného PDF souboru. Část pro vyhledávání obsahuje výběr adresáře, který poskytuje aplikaci zdroj

souborů. Dále pak ovládací prvky vyhledávání a zobrazení počtu nalezených souborů. Vedle se pak nachází výběr adresáře pro export PDF souboru a ovládací prvky.

Levá střední část grafického rozhraní obsahuje okno pro zobrazení náhledu obrázků. Pod ním se nachází tabulka pro výpis informačních hlášení aplikace.

7 IMPLEMENTACE

7.1 NAČTENÍ SOUBORŮ

Jak už bylo uvedeno, načítání souborů se provádí z vybraného adresáře. V principu to funguje tak, že uživatel vybere složku a program v ní vyhledá obrázky. V programu se jedná o metodu *NacteniJmenObrazku()*, která má jako vstupní parametr řetězec. Tento řetězec je cesta k vybrané složce. Metoda vyhledá obrázky a společně s informacemi o obrázku zapíše jejich názvy do pole, které metoda vrací.

7.2 DETEKCE KŮŽE, DETEKCE OBLIČEJE A ROZHODOVACÍ ALGORITMUS

Jak už je v předchozích kapitolách popsáno, k detekci závadných obrázků je využita funkce knihovny EmguCV a algoritmus detekce kůže.

Oba tyto algoritmy jsou navíc implementovány v samostatných aplikacích a jsou přiloženy na CD jako příloha. Byly vytvořeny po rešerši zdrojů pro testování samotných algoritmů.

Po načtení jmen souborů se prochází postupně jednotlivé soubory podle názvu a určuji pomocí metody *Zavadnost()*, jestli se jedná o citlivé obrázky. Pokud je tomu tak, název souboru a informace o něm se uloží do pole *zavadneObrazky()*. Zároveň se k němu vypočte hash SHA1 a uloží se do stejného pole. Poté, co se zkontrolují všechny obrázky, vyhledávání se ukončí.

Celé toto zpracování obrázků je vyřešeno pomocí kontrolky Backgroudworker. Jedná se o kontrolku, která vyřeší problém zamrzávání uživatelského prostředí, pokud se zůstane déle v metodě *Zavadnost()*. V principu se operace vykoná v jiném vlákně.

Celý tento problém je řešen třemi metodami. Jedna pro provedení operace – *zpracovaniBackgroundWorker_DoWork()*, druhá dává zpětnou vazbu s průběhem operace – *zpracovaniBackgroundWorker_ProgressChanged()* a třetí pro závěrečné operace – *zpracovaniBackgroundWorker_RunWorkerCompleted()*.

Nyní se podíváme detailně na metodu *Zavadnost()*. Obsahuje dva způsoby detekce a výsledné vyhodnocení detekcí. Vrací logickou hodnotu závadnosti souboru – true nebo false. Vstupem metody je název souboru. Metoda si nahraje obrázek, který se poté normalizuje do jednotné velikosti. Soubory jsou různých velikostí, například fotografie dnes dosahují velikostí v řádech tisíců pixelů, zatímco obrázky z internetu bývají menších rozměrů. Zpracování velkých obrázků detekčními algoritmy by zabralo zbytečně mnoho času, proto je vhodné obrázky normalizovat do jednotné velikosti a uspořit tím čas běhu programu. U jednoho obrázku je to minimální úspora, ale u stovek obrázků jde o podstatnou úsporu času.

Detekce kůže je popsána v jedné z předešlých kapitol. Nyní bude popsáno praktické naprogramování algoritmu. Základem jsou dva *for* cykly, díky nim se prochází obrázek pixel po pixelu. Každý bod se načte a rozdělí do složek RGB modelu. Dále v těle cyklů je samotný převod složek RGB do modelu YCbCr. Nás zajímají jen hodnoty pixelů, které odpovídají lidské kůži. Na to slouží podmínka *if*, která tyto body zachycuje a počítá v proměnné *pocetPozitivni*. Následně po průchodu celým obrázkem, je vypočteno procento zastoupení pixelů s kůží, které se následně vyhodnotí na základě prahu. Určení jeho velikosti je velice složitá otázka, jelikož detekce není stoprocentní, velikost prahu určuje zařazení do skupiny. Z obecného hlediska se musíme zamyslet, jestli jsme ochotni připustit některé obrázky nesprávně zařazené do kategorie závadných (false positives). Druhou možností je nastavit přísnější práh, ovšem v tom případě se objeví závadné obrázky, které budou určeny jako nezávadné. Z požadavku na aplikaci vychází, že v tomto případě je vhodné nastavit benevolentnější práh a tím získat v celkovém objemu detekovaných obrázků větší počet závadných.

Detekce obličeje je vyřešena pomocí otevřené knihovny EmguCV, která poskytuje širokou škálu nástrojů pro zpracování obrazu. V předešlé kapitole je již popsán princip detekce. Nejprve je nutné převést obrázek do odstínů šedi. Následně je využito funkce knihovny a použita metoda *DetectHaarCascade()*, která vrací nalezené tváře. Dále je nutné zmínit, tuto metodu lze využít pro vyhledání jakéhokoli objektu (tvář). Vstupním parametrem metody je objekt, který specifikuje, co je předmětem vyhledávání. Konkrétně je to specifikováno v XML souboru, který se načítá – *haarcascade_frontalface_default.xml*. Poté je do proměnné oblicejPocet uloženo počet nalezených tváří.

Po průchodu oběma detekcemi je potřeba získané průběžné výsledky vyhodnotit a klasifikovat obrázek do jedné z kategorií (závadný, nezávadný). Tuto hodnotu metoda vrací zpět. Nejprve následují dvě podmínky *if*, které vyhodnotí počet nalezených tváří. To znamená, když klasifikátor nenašel žádnou tvář, *oblicejHodnota* se nastaví na hodnotu *flase*. Naopak, pokud nastane situace, že se našla jedna nebo více tváří, nastaví se proměnná do *true*. Proměnná *procentoPozitivni* závisí také na počtu nalezených tváří. Výsledné zpracování vychází z tabulky na obrázku 13.

Nastavení podmínek u výsledného zpracování není jednoduché a je pak otázkou preference, na jaké kategorie obrázků bude jakým způsobem vyhodnocení reagovat. Například v situaci, kdy *kuzeHodnota* = *true* a *oblicejHodnota* = *false*. Jedná se o kategorii obrázků, kde bylo identifikováno tělo, ale osoba může být otočená a není ji vidět do tváře. Ale zároveň to může být chybně detekovaná kůže a může se jednat o čistě náhodné prostředí s podobným odstínem. Postup, jak na některé situace reagovat, je navržen v kapitole Možnosti rozšíření.

		oblicejHodnota	
		true	false
kuzeHodnota	true	true	true
	false	false	false

Tabulka 1 - Vyhodnocení detekcí

7.3 EXPORT VÝSLEDKŮ DO PDF

Export reportu je důležitou součástí aplikace. Výstupní soubor má být přehlednou reprezentací potenciálně závadných obrázků. PDF soubor obsahuje na začátku celkové shrnutí a dále detailní informace o souborech (název obrázku, autor, pokud je dostupný z EXIF, cesta, datum vytvoření, datum pořízení, pokud je dostupné z EXIF, velikost obrázku, rozměry obrázku, výrobce zařízení a model, pokud je dostupný z EXIF, hash SHA1) a náhled. Pro export PDF souboru je využita otevřená knihovna PDFsharp.

8 VYHODNOCENÍ APLIKACE

Pro otestování přesnosti aplikace byla vytvořena sada obrázků. Ta obsahovala 2653 obrázků. Z celkového počtu jich bylo 665 závadných (na obrázku byla osoba s odhaleným tělem, jsou započteny i osoby v plavkách a jiném ne zcela zahalujícím oblečení). Nezávadných obrázků bylo 1988.

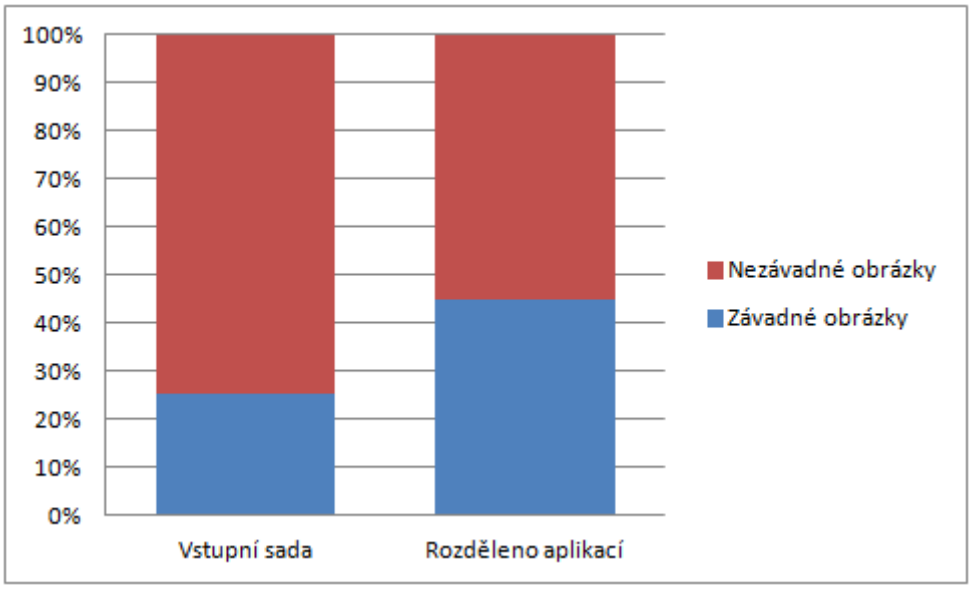
Celkem obrázků	2653
Závadných	665
Nezávadných	1988

Tabulka 2 – Vstupní sada obrázků

Po zpracování sady obrázků programem bylo dosaženo následujících výsledků. Aplikace označila 1187 obrázků jako potenciálně závadné a 1466 obrázků jako nezávadné.

Označeno jako závadné	1187
Označeno jako nezávadné	1466

Tabulka 3 - Rozdělení obrázků aplikací



Obrázek 16 - Rozdělení dat na vstupu a výstupu

Po překontrolování výstupního souboru aplikace bylo zjištěno, že z 1187 potenciálně závadných obrázků jich bylo 638 závadných a 549 nezávadných.

Správně zařazeno jako pozitivní	638
Nesprávně zařazeno jako pozitivní	549

Tabulka 4 - Rozdělení obrázků aplikací (pozitivní)

Z počtu 1466 obrázků, které aplikace označila jako nezávadné, bylo skutečně nezávadných 1439. Tedy jen 27 závadných obrázků aplikace nezachytila.

Správně zařazeno jako negativní	1439
Nesprávně zařazeno jako negativní	27

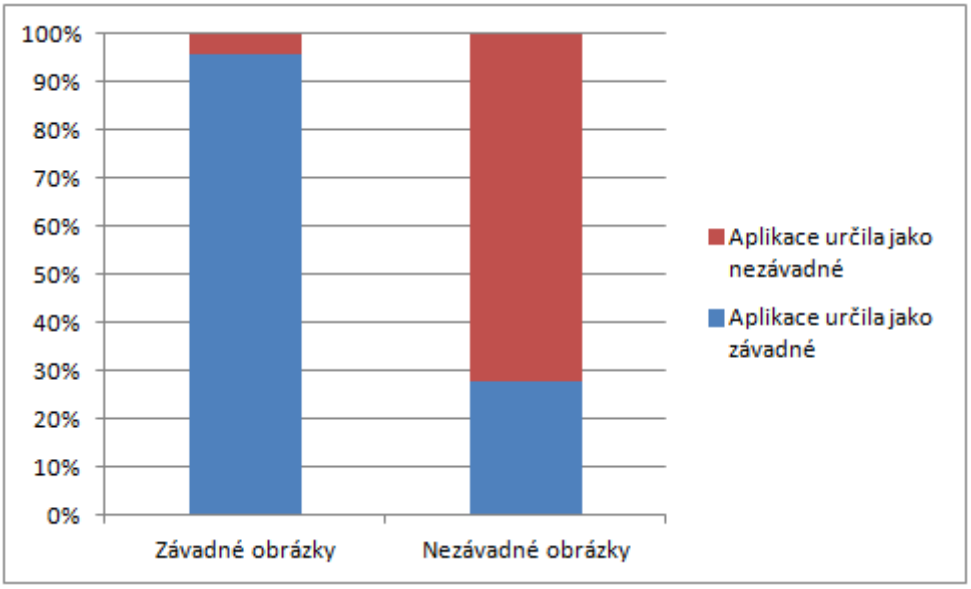
Tabulka 5 - Rozdělení obrázků aplikací (negativní)

Aplikace nesprávně zařadila 549 obrázků jako pozitivní. Je to poměrně značné množství, ale odpovídá nastavení v programu (výsledné vyhodnocení obou metod detekcí).

Cílem bylo odhalit co nejvíce závadných obrázků, toto se nepodařilo pouze u 27 obrázků.

Detekováno závadných	638 / 665 * 100 = 95,94%
Nedetekováno závadných	27 / 665 * 100 = 4,01%
Chybná detekce	549 / 2653 = 0,21 /na obrázek

Tabulka 6 - Přesnost aplikace



Obrázek 17 - Přesnost aplikace

8.1 STATISTICKÉ VYHODNOCENÍ VÝSLEDKŮ APLIKACE

Máme skupinu vzorků (odpovídá počtu testovaných obrázků – 2653). Díváme se na to tak, že na jedné skupině vzorků byla prováděna dvě vyhodnocení: První odpovídá reálnému rozdělení vzorků na závadné a nezávadné. Druhé bylo prováděno pomocí vytvořené aplikace.

8.1.1 TESTOVÁNÍ ZÁVISLOSTI OBOU VYHODNOCENÍ

Testujeme platnost hypotézy, že obě vyhodnocení jsou na sobě nezávislá, neboli rozdělení aplikací a reálného rozdělení vzorků se liší.

Níže v tabulce Sledování je to, co jsme naměřili, ať už na vstupu nebo na výstupu.

	aplikace		
data	ano	ne	
ano	638	27	665
ne	549	1439	1988
All Grps	1187	1466	2653

Tabulka 7 - Sledování

Za platnosti nulové hypotézy (nezávislosti obou sledování) je pravděpodobnost, že náhodně vybraný jedinec patří do buňky tabulky, zapsána v tabulce Očekávané pravděpodobnosti.

	aplikace		
data	ano	Ne	
ano	0,112	0,139	0,251
ne	0,335	0,414	0,749
All Grps	0,447	0,553	1,000

Tabulka 8 - Očekávané pravděpodobnosti

Když tyto pravděpodobnosti vynásobíme celkovým počtem jedinců (tj. číslem 2653), dostaneme tabulku Očekávání (tj. očekávané počty za platnosti nulové hypotézy).

	aplikace		
data	ano	ne	
ano	297,533	367,467	665,000
ne	889,467	1098,533	1988,000
All Grps	1187,000	1466,000	2653,000

Tabulka 9 – Očekávání

Například pravděpodobnost, že náhodně vybraný obrázek byl vyhodnocen jako závadný na vstupu i aplikací (tj. ve sloupci „ano“ i v řádku „ano“) je rovna za předpokladu nezávislosti pravděpodobnosti, že patří danému řádku * pravděpodobnosti, že patří danému sloupci:

Například:
P(ano vstup, ano výstup) = p(ano vstup)*p(ano výstup) = 665/2653*1187/2653 = 0,112

Pomocí χ^2 testu pak testujeme, zda se tabulka Sledování rovná tabulce Očekávání. $\chi^2(1) = 940.891$, $P < 10^{-30} < 0.05$, neboli zamítáme nulovou hypotézu o nezávislosti obou vyhodnocení. Tímto je vyvrácena možnost, že by aplikace nedetekovala závadné obrázky.

V tabulce Reziduály, což je (sledování – očekávání) pak lze odvodit, v jakém směru se obě tabulky liší.

	aplikace		
data	ano	ne	
ano	340,467	-340,467	0,00
ne	-340,467	340,467	0,00
All Grps	0,000	0,000	0,00

Tabulka 10 – Reziduály

Například u výstupů z aplikace je zachyceno mnohem více závadných z těch, kteří byli závadní na vstupu, než by se očekávalo. Tím se potvrzuje vysoká míra úspěšnosti aplikace v odhalení závadných obrázků.

8.1.2 McNemarův test

Pomocí McNemarova testu se testuje:

a. H01: pravděpodobnost stejného počtu záchytů u obou metod = pravděpodobnost stejného počtu nezáchytů u obou metod.

   Neboli testuje se, že obě metody zachytí totéž: $\chi^2(1) = 308.14$, $P < 10^{-30} < 0.05$, neboli zamítám nulovou hypotézu, metody nezachytí totéž. Pravděpodobnost záchytu je nižší, než pravděpodobnost nezáchytu.

b. H02: pravděpodobnost, že aplikace nezachytí to, co se zachytilo na vstupu je stejná jako pravděpodobnost, že aplikace zachytí to, co nezachytil vstup.

   $\chi^2(1) = 471.25$, $P < 10^{-30} < 0.05$, neboli zamítám nulovou hypotézu. Pravděpodobnost chybného záchytu aplikace je nižší, než pravděpodobnost neodhalení záchytu provedením aplikace.

$\chi^2(1)$	940,89	p=0,0000
McNemar (A/D)	308,14	p=0,0000
McNemar (B/C)	471,25	p=0,0000

Tabulka 11 - McNemar 1

	aplikace ano	aplikace ne	
data ano	638	27	665
	24,048%	1,018%	25,066%
data ne	549	1439	1988
	20,694%	54,240%	74,934%
All Grps	1187	1466	2653
	44,742%	55,258%	100%

Tabulka 12 - McNemar 2

8.2 PODMÍNKY SPUŠTĚNÍ A PROHLÍŽENÍ

Pro spuštění aplikace je podmínkou .NET Framework 4.0.

Aplikace exportuje výstup do PDF dokumentu verze 1.4. Pro zobrazení souboru je podmínkou verze Adobe Acrobatu 5.0 a vyšší.

8.3 REPORT APLIKACE

Aplikace umožňuje export výsledků do PDF souboru. Níže je zobrazen příklad takového souboru. Report obsahuje celkové výsledky detekce a dále informace k jednotlivým souborům. Každý obrázek má i svůj náhled.

Report aplikace Detektor

Celkový počet obrázků: 99
Počet závadných obrázků: 59
Procento závadných obrázků: 59,6

Název obrázku: 020.jpg (autor:)
Cesta: C:\Desktop\Testovací sada jpg\finální\020.jpg
Datum vytvoření: 10.4.2014 15:00:21 (datum pořízení(EXIF): 26.8.2005 9:45:54)
Velikost: 27823 B
Rozměry: 247 x 370 pixelů
Výrobce zařízení(EXIF): Canon
Model(EXIF): Canon EOS 20D
SHA1: 8F9639A12CD71A476950FF26A184AD782CA60491

Název obrázku: 060709-holky-travnik.jpg (autor: System.Collections.ObjectModel.ReadOnlyCollection`1[System.String])
Cesta: C:\Testovací sada jpg\finální\060709-holky-travnik.jpg
Datum vytvoření: 10.4.2014 14:35:04 (datum pořízení(EXIF): 9.7.2006 14:08:55)
Velikost: 123803 B
Rozměry: 480 x 640 pixelů
Výrobce zařízení(EXIF): Panasonic
Model(EXIF): DMC-LZ1
SHA1: 73680D19D7F9231B56795DDD17E40E5B244005DE

Název obrázku: 1016713-img-emma-griffiths-emma-rvacka.jpg (autor:)
Cesta: C:\Testovací sada jpg\finální\1016713-img-emma-griffiths-emma-rvacka.jpg
Datum vytvoření: 14.4.2014 9:41:09 (datum pořízení(EXIF):)
Velikost: 74535 B
Rozměry: 978 x 733 pixelů
Výrobce zařízení(EXIF):
Model(EXIF):
SHA1: D4A8D0B98BC65AA3A3CB6DC2986C7FDF024359B1

Název obrázku: 114839.jpg (autor:)
Cesta: C:\Testovací sada jpg\finální\114839.jpg
Datum vytvoření: 14.4.2014 9:39:27 (datum pořízení(EXIF):)
Velikost: 37120 B
Rozměry: 600 x 250 pixelů
Výrobce zařízení(EXIF):
Model(EXIF):
SHA1: B7265DD7474E111A561AA873CB7555531F743674

Název obrázku: 1370822.jpg (autor:)
Cesta: C:\Testovací sada jpg\finální\1370822.jpg
Datum vytvoření: 30.3.2014 15:19:38 (datum pořízení(EXIF):)
Velikost: 50415 B
Rozměry: 680 x 382 pixelů
Výrobce zařízení(EXIF):
Model(EXIF):
SHA1: 23C98518DC8DFFE9BCD00B91EF58A2B520F3829C

Obrázek 18 - Report aplikace

ZÁVĚR

V rámci této práce byla vytvořena aplikace pro detekci pornografie. Z vybraného adresáře a podadresářů program vybere obrazové soubory, klasifikuje u nich závadnost a zobrazí výsledky, které je dále možno exportovat do přehledného souboru PDF.

Aplikace klasifikuje závadnost na základě dvou vybraných algoritmů (detekce obličeje a detekce kůže), které jsou popsány v kapitole 4 Metody zpracování obrazu. Dále byl testován algoritmus na detekci bradavek. Ten ale není součástí aplikace. V termínu se nepodařilo dosáhnout uspokojivých výsledků.

Ze zhodnocení výsledků aplikace v kapitole 8 Vyhodnocení aplikace vyplývá, že se podařilo vytvořit program na detekci pornografie. Program Detektor vykazuje vysokou míru detekování závadných obrazových souborů. Více než 95% ze skupiny závadných vzorků.

V kapitole Možnosti rozšíření aplikace jsou stručně popsány vlastnosti, které lze u aplikace vylepšovat. Dále v kapitole 4.3 je přehled veřejně dostupných algoritmů a vylepšení pro detekci nahoty.

MOŽNOSTI ROZŠÍŘENÍ APLIKACE

- natrénování klasifikátoru pro hledání dalších objektů (bradavky)
- propracovanější grafické rozhraní a větší výběr možností vyhledávání
- možnost zásahu do exportu PDF a výběru vypisovaných dat
- rozšíření typů vyhledávaných grafických souborů, přidání video souborů a jiných typů (např. PDF atd.)
- vyhledávání klíčových slov v názvech souborů
- vyhledávání na základě hlavičky souboru, ne podle koncovky typu souboru (pachatel může souborům upravit koncovky)

CITOVANÁ LITERATURA

1. **Porada, Viktor a Rak, Roman.** Teorie digitálních stop a její aplikace v kriminalistice a forenzních vědách. *Karlovarská právní revue.* 2006, 4.

2. **Svetlík, Marián.** Digitální forenzní analýza a bezpečnost informací. *Data security management.* Forenzní analýza, 2010, 1.

3. **kolektiv autorů.** *A-Ž Malý encyklopedický slovník, Academia 1972, s. 920).* Praha : Academia, 1972. ISBN: 21-082-72.

4. **Microsoft.** News. [Online] [Citace: 11. listopad 2014.] http://news.microsoft.com/download/presskits/photodna/docs/photodnafs.pdf.

5. **Sarris, Nikos, Grammalidis, Nikos a Strintzis, Michael G.** *Detection of Faces and Facial Features in Images using a Novel Neural Network Technique.* místo neznámé : In Proc. of WSEAS Int. Conf. on Neural Network and Applications , 2001. pp. 6361-6366.

6. **Marcial-Basilio, Jorge A., a další.** Detection of Pornographic Digital Images. *INTERNATIONAL JOURNAL OF COMPUTERS.* 2011, Sv. Vol. V, Issue 2.

7. **Neagoe, Victor Emil a Neghina, Mihai.** *Face Detection Using a Dual Cross-Validation of Chrominance/Luminance Channel Decisions and Decorrelation of the Color Space.* místo neznámé : In Proc. Of 14th WSEAS Int. Conf. on Computers, 2010. pp. 391-396.

8. Emgu CV: OpenCV in .NET. [Online] [Citace: 19. březen 2014.] http://www.emgu.com/wiki/index.php/Main_Page.

9. *OpenCV.* [Online] [Citace: 19. březen 2014.] http://opencv.org/.

10. *Robust Real-Time Face Detection.* **Viola, Paul a Jones, Michael J.** Vancouver, Canada : autor neznámý, 2001.

11. **Lienhart, Rainer a Maydt, Jochen.** *An Extended Set of Haar-like Features for Rapid Object Detection.* místo neznámé : IEEE ICIP, 2002. 900-903.

12. **Mahmoud, Tarek M.** A New Fast Skin Color Detection Technique. *World Academy of Science.* Engineering and Technology, 2008, 19.

13. **El-Hafeez, Tarek Abd.** A new system for extracting and detecting skin color regions from PDF documents. *International Journal on Computer Science and Engineering.* Department of Computer Science, Faculty of Science, Minia University, El-Minia, Egypt, 2010.

14. **Phung, Son Lam, Chai, Douglas a Bouzerdoum, Abdesselam.** Adaptive skin segmentation in color images. School of Engineering and Mathematics - Edith Cowan University, Perth, Australia, 2003.

15. **Duan, Lijuan, a další.** Adult image detection method base-on skin color model and support vector machine. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 2002.

16. **Ap-apid, Rigan.** An algorithm for nudity detection. College of Computer Studies, De La Salle University, Manila, Philippines.

17. **Chi-Yoon, Jeong, Jong-Sung, Kim a Ki-Sang, Hong.** Appearance-based nude image detection. Republic of Korea, 2004, 0-7695-2128-2/04.

18. **Flores, Pedro Ivan Tello, Guillén, Luis Enrique Colmenares a Prieto, Omar Ariosto Niño.** Approach of RSOR algorithm using HSV color model for nude detection in digital images. *Computer and Information Science.* Apartado postal J-32, Ciudad Universitaria, Puebla, México, 2011.

19. **Fleck, Margaret M., Forsyth, David A. a Bregler, Chris.** Finding naked people.

20. **Lee, Jiann-Shu, a další.** Naked image detection based on adaptive and extensible skin color model. Taiwan, 2006.

21. **Qin, Wu a Guodong, Guo.** Age classification in human body images. [Online] 2013. http://dx.doi.org/10.1117/1.JEI.22.3.033024 .

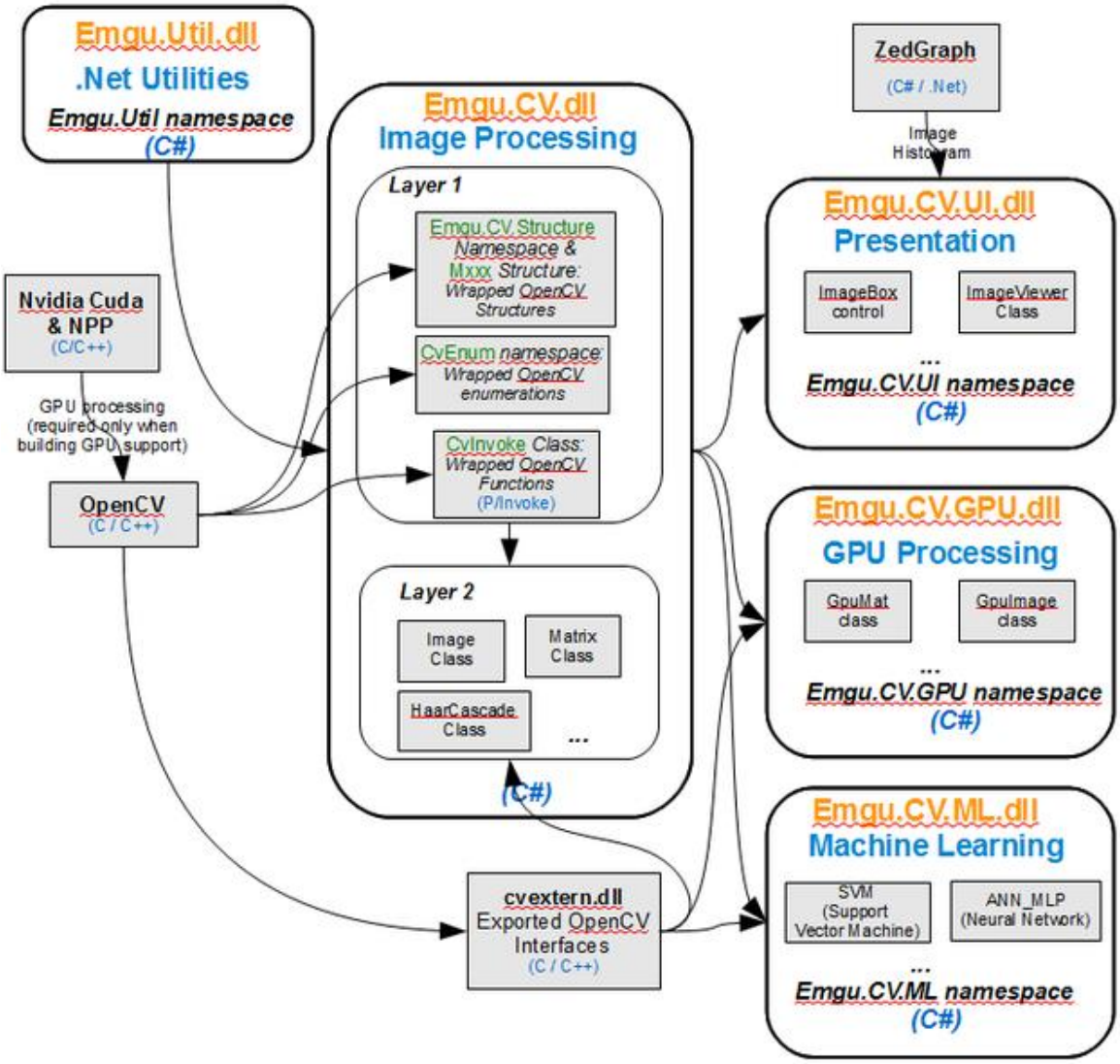
22. **Ince, Omer F., a další.** Child and Adult Classification Using Ratio of Head and Body Heights in Images. *International Journal of Computer and Communication Engineering.* 2014, Sv. Vol. 3, No. 2.

23. **Kwon, Young H. a Lobo, Niels da Vitoria.** Age Classification from Facial Images. *Computer Vision and Image Understanding.* 1999.

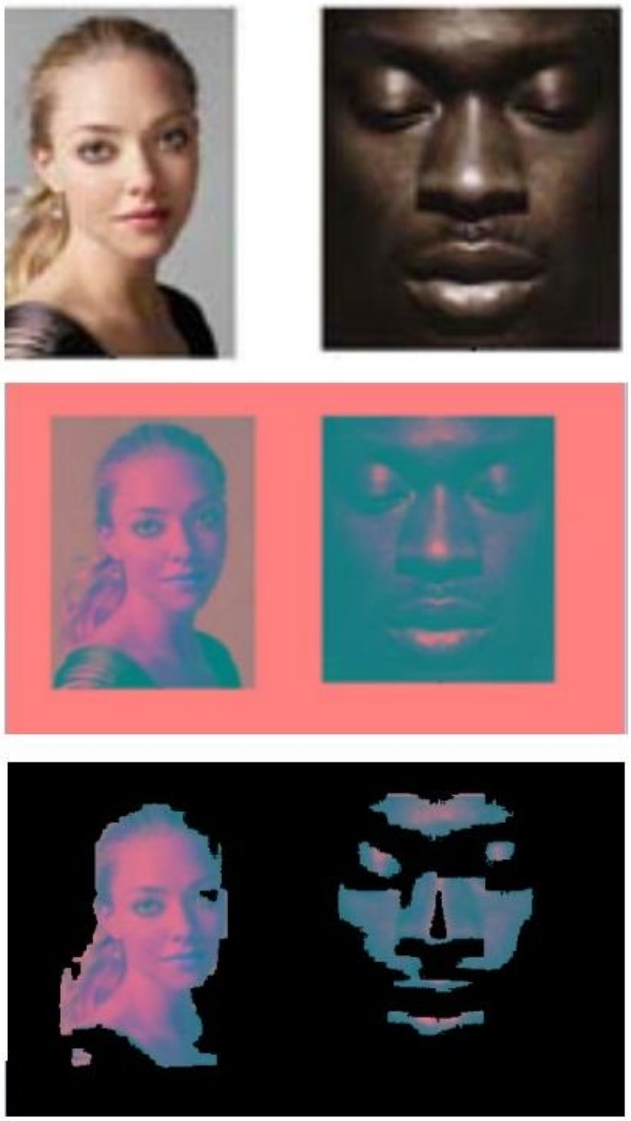
SEZNAM OBRÁZKŮ

SEZNAM TABULEK

PŘÍLOHA 1 – ARCHITEKTURA EMGUCV



PŘÍLOHA 2 – PŘÍKLAD VÝSTUPU SEPARACE OBRÁZKU



PŘÍLOHA 3 – ALGORITMUS ADABOOST

1. Vstup:

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\}, \text{počet iterací T}$$

2. Inicializace vah:

$$D_1(i) = \frac{1}{m}$$

3. Cyklus pro t = 1, …, T:

   a. Výběr klasifikátoru na základě vážené trénovací chyby

$$\varepsilon_j = \sum_{i=1}^{m} D_t(i) I\left[y_i \neq h_j(x_i)\right]$$

$$h_t = \arg\min_{h_j \in H} \varepsilon_j$$

   b. Pokud

$$\varepsilon_t = 0 \text{ nebo } \varepsilon_t \geq \frac{1}{2}, \text{pak konec cyklu}$$

   c. Nastavení

$$\alpha_t = \frac{1}{2} \log\left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right)$$

   d. Úprava vah

$$D_{t+1}(i) = \frac{D_t(i) e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$

$$\text{kde } Z_t = \sum_{i=1}^{m} D_t(i) e^{-\alpha_t y_i h_t(x_i)}$$

4. Výstup

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$