VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

ROZPOZNÁVÁNÍ HUDEBNÍCH STYLŮ

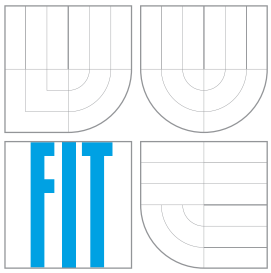
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

KAMIL BEHÚŇ

BRNO 2011

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

ROZPOZNÁVÁNÍ HUDEBNÍCH STYLŮ

MUSIC STYLE RECOGNITION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

KAMIL BEHÚŇ

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. MICHAL HRADIŠ

BRNO 2011

Abstrakt

Tato práce se zabývá rozpoznáváním hudebních stylů. V úvodu je přehled aktuálních metod používaných při rozpoznávání hudebních stylů. Další kapitoly jsou věnovány vytvořenému systému pro rozpoznávání hudebních stylů. Výsledný systém obsahuje dvě metody extrakce příznaků. První využívá extrakci Mel-frekvenčních kepstrálních koeficientů z nahrávek a druhá extrakci příznaků ze spektrogramu nahrávek. Pro klasifikaci výsledný systém využívá Support Vector Machine.

Abstract

This thesis deals with the music style recognition. The introduction is an overview of current methods used in the music style recognition. Next chapters deals with the system created for the music style recognition. The final system is consists of two feature extraction methods. The first uses the Mel-frequency cepstral coefficients extraction from records and the second uses feature extraction from spectrograms of records. The final system uses Support Vector Machine for classifying.

Klíčová slova

Rozpoznávaní hudebních stylů, extrakce příznaků, klasifikace, Support Vector Machine, Toolkit Skrytého Markovovho Modelu, Mel-frekvenční kepstrální koeficienty, lokální příznaky.

Keywords

Music style recongization, feature extraction, classification, Hidden Markov Model Toolkit, Mel-frequency cepstral coefficients, local features.

Citace

Kamil Behúň: Rozpoznávání hudebních stylů, bakalářská práce, Brno, FIT VUT v Brně, 2011

Rozpoznávání hudebních stylů

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Ing. Michala Hradiše

.....
Kamil Behúň
16. května 2011

Poděkování

Děkuji vedoucímu mé práce Ing. Michalovi Hradišovi za odbornou pomoc při problémech, které vznikli při řešení této práce a za čas, který mi věnoval.

Obsah

Kapitola 1

Úvod

Pokrok zaznamenaný v posledných desaťročiach v oblastiach informačných, komunikačných a mediálnych technológiách je srpevádzaný s vyprodukovaním veľkého množstva dát. Ako je uvedené v [5, 14] to práve obzvlášť platí pre hudbu a hudobné databázy, ktoré neustále exponenciálne rastú. Táto skutočnosť si teda vyžaduje vhodné nástroje pre vyhľadávanie a manipuláciu s takýmto množstvom dát, pretože ručné indexovanie a radenie takéhoto množstva dát by bolo neúnosné a zdĺhavé. Samozrejme je možné pri tejto činnosti využiť metadata, ktoré pesničky obsahujú a udržujú tak informácie nielen o autorovi, ale aj o hudobnom štýle. Avšak ako je uvedené v [5] sú často nekompletné a občas aj nekonzistentné. Nekonzistentnosť pri hudobných štýloch sa môže vyskytnúť už len z toho dôvodu, ako je spomenuté v [5], tak pre žánre neexistuje presná definícia vlastností, podľa ktorých možno piesne do nich presne zaradiť. Okrem toho moderné piesne obsahujú prvky z viacerých žánrov zároveň, čo úlohu o to sťažuje. Toto je len jeden z príkladov, kde by automatické rozpoznávanie hudobných štýlov mohlo nájsť svoje uplatnenie.

Pri rozpoznávaní je okrem dobrého klasifikátora, ktorý sa na trénovacej množine dát natrénuje aby potom dobre zvládol danú úlohu, dôležitá aj vhodná metóda extrakcie príznakov. To je spôsob akým vybrať zo vzorku trénovacej alebo nevidenej sady dôležité informácie, ktoré ho na rozumnej úrovni popisujú a na základe ktorých je ho možné pre danú triedu dobre klasifikovať. Teda cieľom tejto práce je urobiť krátky prehľad o súčasných metódach klasifikácie hudobných štýlov a extrakcií príznakov. Nasledne na základe týchto poznatkov vytvoriť klasifikačnú metódu a vhodnú sadu príznakov pre rozpoznávanie hudobných štýlov. A nakoniec túto metódu natrénovať a otestovať na vhodnej dátovej sade.

Pre extrakciu príznakov v tejto práci boli vyskúšané dve metódy. Prvá získava príznaky priamo zo zvuku pomocou Mel-frekvenčných kepstrálnych koefcientov a druhá, ktorá využíva možnosť previesť signál z 1D (zvuku) na signál 2D (obrázok) a až z tohto prevedeného signálu (obrázka) extrahovať príznaky. Ako klasifikátor sa používa Support Vector Machine s rôznymi jadrovými funkciami.

Rozdelenie tohto dokumentu je nasledovné. V nasledujúcej kapitole 2 je krátky prehľad existujúcich prístupov extrakcie príznakov a klasifikácie, ktoré sa pre rozpoznávanie hudobných štýlov používajú. Keďže jedna z metód extrakcie príznakov využíva extrakciu príznakov z obrázka, je tento prístup popísaný v kapitole 3. V kapitole 4 sú následne opísané navrhnuté prístupy ,ktoré som použil a v kapitole 5 je opísaný postup pri implementácií týchto prístupov. Okrem toho je v tejto kapitole aj prehľad nástrojov, ktoré som pri práci použil. Kapitola 6 okrem experimentov, ktoré som s vytvoreným systémom rozpoznávania hudobných štýlov vykonal, obsahuje aj popis dátových sád, na ktorých tieto experimenty prebehli a porovnanie jednotlivých dosiahnutých výsledkov experimentov. V záverečnej ka-

pitole 7 je zhrnutie a diskusia nad dosiahnutými výsledkami ako aj zamyslenie o možnom ďalšom pokračovaní a rozšírení tejto práce.

Kapitola 2

Existujúce prístupy

Táto kapitola obsahuje prehľad súčasných prístupov rozpoznávania hudobných štýlov na základe existujúcich prác. Obsahuje prístupy pri extrakcií príznakov ako aj klasifikačné metódy, ktoré sa momentálne pre túto úlohu používajú. Ďalšie odstavce budú rozdeľovať jednotlivé prístupy na základe prístupov k extrakcií príznakov.

V [20] extrakciu príznakov rozdeľujú do troch skupín a to na *príznaky farby tónu*, *príznaky výšky tónu* a *príznaky rytmického obsahu*. Toto rozdelenie je používané aj v [8]. Cieľom príznakov rytmického obsahu je v [20] nájdenie najnápadnejších periodicít signálu. K tomuto účelu sa v tejto práci používa beat histogram, ktorého výpočet je založený na vlnkovej transformácií a reprezentuje váhu rôznych rytmických periód v signále. Pre príznaky výšky tónu sa v [20] používajú dva histogramy výšok, v ktorých každý interval zodpovedá hudobnému tónu so špecifickou výškou frekvencie (napríklad pre komorné A = 440Hz). Tieto dva histogramy výšiek sa odlišujú v tom, že jeden zlučuje tóny jednotlivých oktáv a druhý to nerobí. Takže histogram výšiek, ktorý zlučuje tieto oktávy obsahuje informáciu o harmonickom obsahu a histogram výšiek, ktorý nezlučuje oktávy obsahuje informácie o výškovom rozsahu. V [20] sú príznaky rytmického kontextu a príznaky výšky tónu počítané cez celú pesničku, čo môže byť pri veľmi nehomogénnych signáloch problém, pretože tak môže dôjsť ku skresleniu extrahovaných príznakov lokálnymi nepravidelnosťami. Treťou skupinou sú príznaky farby tónu a ako je uvádzané v [20], tak sú využívané aj pri rozpoznávaní reči a hudba-reč diskriminácií. Cook et al., ďalej uvádzajú, že výpočet týchto príznakov je založený na krátkodobej analýze ukázanej aj na obrázku 2.1 a tieto príznaky sú extrahované zo spekrtogramu spočítaného pomocou krátkodobej Fourierovej transformácie. Do tejto skupiny príznakov patria nasledujúce príznaky, ktorých popis je čerpaný z [20]:

- **Spektrálne ťažisko** je ťažiskom rozsahu spektra krátkodobej Fourierovej transformácie, teda určuje v akom frekvenčnom rozsahu sa nachádza hlavná časť signálu.
- **Spectral Rolloff** je príznakom definovaným ako hodnota (prah) frekvencie f, pod ktorou sa nachádza určité percento energie (v [20] je to 85 % energie).
- **Spektrálny tok** je mierou ako rýchlo sa mení energetické spektrum signálu. Spektrálny tok je v [20] definovaný ako rozdiel medzi dvomi po sebe idúcimi normalizovanými spektrálnymi snímkami.
- **Miera prechodu nulou** definuje sa ako početnosť prechodov signálu zo zápornej oblasti do kladnej.

- **Hlasitosti** je príznakom, ktorý sa snaží modelovať subjektívny vnem hlasitosti človekom. Prvým krokom je teda modelovanie frekvenčnej charakteristiky ľudského vonkajšieho a vnútorného ucha. A táto frekvenčná charakteristika je potom používaná ako váhová funkcia, ktorá buď zdôrazňuje alebo zoslabuje spektrálne zložky práve na základe ľudského počutia.

- **Mel-frekvenčné kepstrálne koeficienty (MFCC)** sú nelineárnou reprezentáciou spektra a bližšie sú popísané v sekcií 2.1, kde je opísané ako sa tieto koeficienty získavajú a kde všade sú používané.

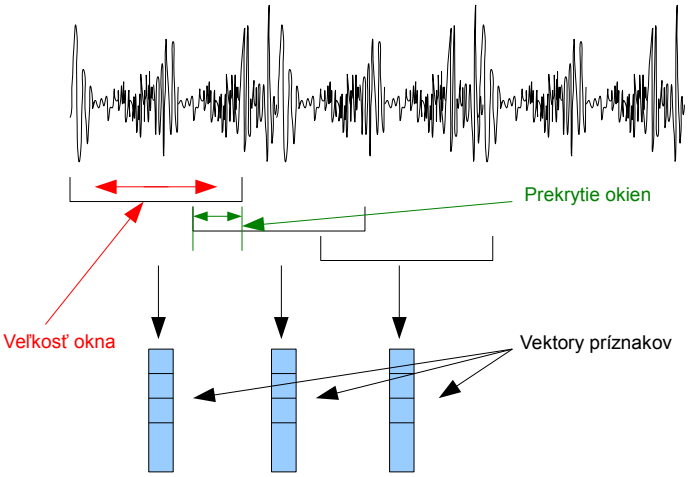
Z hore uvedených príznakov farby tónu v [20] dosahujú najlepšie výsledky Mel-frekvečné kepstrálne koeficienty. Celkovo príznaky farby tónu v [20] dosahujú lepšie výsledky než príznaky rytmického obsahu a príznaky výšky tónu. Cook et al., ďalej uvádzajú, že príznaky rytmického obsahu a príznaky výšky tónu možno použiť ako doplnkové príznaky k príznakom farby tónu, čo ako môžete vidieť v [20] dosahuje ešte lepšie výsledky, než samostatné využitie príznakov farby tónu.

V [20] sa s príznakmi výšky tónu, príznakmi farby tónu a príznakmi rytmického obsahu používa ako klasifikátor Gaussov model zmesí (GMM), ktorého pravdepodobnostné rozloženie každej triedy (žánru) je popísané špecifickým počtom Gaussových funkcií rozloženia pravdepodobnosti. Pre inicializáciu GMM Cook et al., využívajú K-means a Gaussové funkcie rozloženia pravdepodobnosti sú na triedy mapované pomocou iteratívneho EM algoritmu. Čo sa týka príznakov farby tónu, sú používané pre rozpoznávanie hudobných štýlov aj v [5, 8, 12]. V [5] sa ako klasifikačná metóda používa metóda referenčných vektorov. Táto metóda používa pri klasifikácií referenčné vektory, ktoré sú vlastne vhodnými vektormi príznakov trénovacej sady pre najlepšie popísanie triedy, do ktorej tieto vektory spadajú. Tieto referenčné vektory sa teda získavajú pri trénovaní a sú určované samostatne pre každé dve triedy (páry). Pre 5 žánrov by to teda boli 4 sady referenčných vektorov pre každý žáner. Klasifikácia pomocou tejto metódy referenčných vektorov potom prebieha tak, že klasifikovaná vzorka je klasifikovaná pre každé dva žánre zvlášť a žáner s najväčším počtom dielčích „víťazstiev“ bude triedou kam spadá táto trénovacia vzorka. V [12] sa ako klasifikátor využíva Support Vector Machine, ktorý je bližšie popísaný v 2.2.

Zatiaľ, čo hore uvedený prístup bol viac zameraný na príznaky farby tónu, nasledujúci prístup, ktorý je využívaný v [13, 12, 14], je zameraný viac na príznaky rytmického obsahu a ide o príznaky založené na rytmickom vzore. Tieto príznaky sú extrahované zo spektrálnej reprezentácie (spektrogramu) zvukového signálu rozdeleného do segmentov, napríklad pre [12] sa použili 6 sekundové segmenty, čo je rozdiel oproti príznakom rytmického obsahu, ktoré boli spomínané vyššie a boli extrahované z celého signálu nahrávky. Ako je uvedené v [12] zameriavajú sa na 24 frekvenčných pásiem rozhodujúcich pre ľudský sluchový systém. Medzi tieto príznaky patria nasledujúce príznaky, ktorých popis čerpá z [13, 12, 14]:

- **Rytmický vzor** je súbor príznakov založený na psychoakustickom modely, zachytávajúcich výkyvy na frekvenčných pásmach zásadných pre ľudský sluchový systém. Proces extrakcie sa skladá z dvoch krokov. V prvom sa vytvára psychoakusticky modifikovaný spektrogram a v druhom kroku sa zo spetrogramu získava matica reprezentujúca rytmusový vzor s uvedením výskytu rytmu, ale tiež opisujúca menšie výkyvy na všetkých frekvenčných pásmach ľudského sluchového rozsahu.

- **Rytmický histogram** vzniká zhlukovaním hodnôt modulovaných amplitúd pre zásadné pásma ľudského sluchového systému, ktoré boli počítané pri *rytmickom vzore* a sú deskriptorom všeobecnej rytmickej charakteristiky pre daný kus audia.



Obrázek 2.1: Krátkodobá analýza pozostávajúca z extrakcie príznakov z malých prekrývajúcich sa úsekov signálu, respektíve okien signálu. V [20] sa táto krátkodobá analýza vykonáva pomocou krátkodobej Fourierovej transformácie.

- **Štatistický deskriptor spektra** sa počíta v dvoch krokov, pričom prvý je obdobný ako v *rytmickom vzore*. Následne sú z výsledku extrahované štatistické údaje (priemer, medián, rozptyl, šikmosť, špicatosť, minimum aQmaximum) pre každé z kritických pásiem. V [12] sa uvádza, že dobre popisujú výkyvy na kritických pásmach a sú schopné veľmi dobre zachytiť a popísať akustický obsah.
- **Časový štatistický deskriptor spektra** sa získava spočítaním štatistických údajov (priemer, medián, rozptyl, šikmosť, špicatosť, minimum a maximum) pre *štatistický desktriptor spektra* jednotlivých kúskov audia. V [12] sa uvádza, že takto získané príznaky zachytávajú zmeny v čase v spektre na jednotlivých kritických frekvenčných pásmach.
- Pri výpočte **modulačno frekvenčného deskriptoru rozptylu** sa využíva matica spočítaná v *rytmickom vzore* a to tak, že pre každú modulačnú frekvenciu v jednotlivých kritických frekvenčných pásmach sú vypočítané štatistické údaje (priemer, medián, rozptyl, šikmosť, špicatosť, minimum a maximum).

V [12] sú tieto vyššie uvedené metódy extrakcie založené na rytmickom vzore testované s tým, že najlepšie výsledky pri rozpoznávaní hudobných štýlov dosahujú príznaky *štatistického deskriptora spektra*, nasledované príznakmi *rytmického vzoru* a *časovým štatistickým deskriptorom spektra*. Ešte lepšie výsledky sa však dosahujú pri použití *hybridných príznakov*, ktoré sú inšpirované z [20], kde spojením príznakov *farby tónu*, *výšky tónu* a príznakmi *rytmického obsahu* do jedného vektoru príznakov sa dosahujú lepšie výsledky. V [12] však sa namiesto *príznakov rytmického obsahu* použitých v [20], použijú práve niektoré z vyššie uvedených príznakov *rytmického obsahu* založených na *rytmickom vzore* a tie sú spojené s príznakmi farby tónu a výšky tónu, čím teda vzniká viac druhou *hybridných metód*. Najlepšie výsledky pri rozpoznávaní hudobných štýlov v [12] z týchto *hybridných metód* dosahuje spojenie s príznakmi *štatistického deskriptora spektra*.

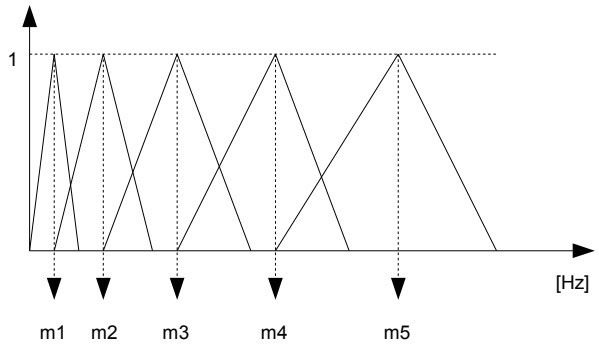
Príznaky založené na rytmickom vzore boli v [13] použité v kombinácií s rôznymi klasifikátormi, najlepšie výsledky s týmito príznakmi dosiahol *Support Vector Machine*. So slabšími výsledkami ako *Support Vector Machine* bol v [13] použitý klasifikátor *K-najbližších susedov*, ktorý si pamätá všetky trénovacie vzorky a na základe K-najbližších susedov potom pri klasifikovaní určuje, do ktorej triedy jednotlivé nevidené dáta patria. Aj keď mal *K-najbližších susedov* slabé výsledky, v [13] boli vyskúšané aj klasifikátori, ako napríklad *Naivný Bayes*, *Rozhodovací strom* a dalšie, s horšími výsledkami (viz. [12]). *Support vector machine* bol pre klasifikáciu hudobných štýlov taktiež použitý aj v [14, 16]. V [12] bol *Support Vector Machine* použitý v kombinácií s príznakmi založenými na rytmickom vzore, ako aj s hybridnými príznakmi.

Doposiaľ boli spomenuté len príznaky extrahované priamo z audia. Avšak ako je uvedené v [13, 14] existujú aj takzvané *symbolické príznaky*, ktoré v prvom kroku prevádzajú audio do MIDI formátu. MIDI využíva definovanú paletu hudobných nástrojov, pre ktoré sa dajú nastaviť rôzne parametre (hlasitosť, výška, tempo, atď.). V druhom kroku sa z tohto formátu extrahujú symbolické deskriptory. Takýmito príznakmi môžu byť napríklad počet nôt, počet významných mlčaní, atď. Hlavným nedostatkom tejto metódy je fakt, že prevod na MIDI funguje dosť neiste a pre normálnu hudbu nie je uspokojivo vyriešený. Tieto príznaky sa teda nepoužívajú samostatne, ale používajú sa v kombinácií s niektorými príznakmi popísanými vyššie.

2.1 Mel-frekvenčné kepstrálne koeficienty (MFCC)

Ako je uvedené na začiatku kapitoly 2, tak Mel-frekvenčné kepstrálne koeficienty patria do skupiny príznakov farby tónu. Spomedzi príznakov farby tónu práve tieto koeficienty dosahujú pri rozpoznávaní hudobných štýlov v [20] najlepšie výsledky. Pre rozpoznávanie hudobných štýlov boli tiež použité aj v [5, 8]. MFCC sa kvôli dobrým výsledkom používajú nielen pri rozpoznávaní hudobných štýlov, ale aj pri rozpoznávaní reči, napríklad v [21]. Výpočet Mel-frekvenčných keptrálnych koeficientov pozostáva z niekoľkých krokov, ktoré budú teraz vysvetlené. Existuje viacero variant výpočtu týchto koeficientov, tento popis je inšpirovaný z [23, 20].

Extrakcia MFCC je založená na krátkodobej analýze (viz. obrázok 2.1), pričom sa ako okno používa Hammingové okno. Používa sa preto, aby sa utlmil začiatok a koniec signálu a výpočet sa zameral na stred. Ďalším krokom, ktorý je použitý v [23], je použitie vysoko pásmového filtru, čím sa zvýšia nižšie frekvencie signálu. Ďalej je vypočítané spektrum pomocou Diskrétnej Fourierovej transformácie. Na výsledné spektrum sa použije Melová banka filtrov, teda sa použijú nelineárne rozmiestnené trojuholníkové filtre hustejšie rozložené v nižších frekvenciách tak, aby sme získali energie v Melovéj mierke. Táto Melová banka filtrov je zobrazená na obrázku 2.2. Melová banka filtrov sa používa preto, lebo ľudské ucho má lepšiu rozlišovaciu schopnosť práve pre nižšie frekvencie. Takto získané Melové energie sa následne zlogaritmujú, pretože ľudské ucho počuje logaritmicky. Nakoniec sa z týchto Melových logaritmických energií získajú pomocou Diskrétnej kosinusovej transformácie výsledné Mel-frekvenené kepstrálne koeficienty, ktoré sú vlastne amplitúdami výsledného spektra signálu.



Obrázek 2.2: Trojuholníkové filtre určené k prevedeniu energií spektra na energie v Melovéj mierke. Na obrázku možno vidieť, že v nižších frekvenciách sú trojuholníkové okná hustejšie ako vo vyšších a je tak kôli vlastnostiam ľudského ucha.

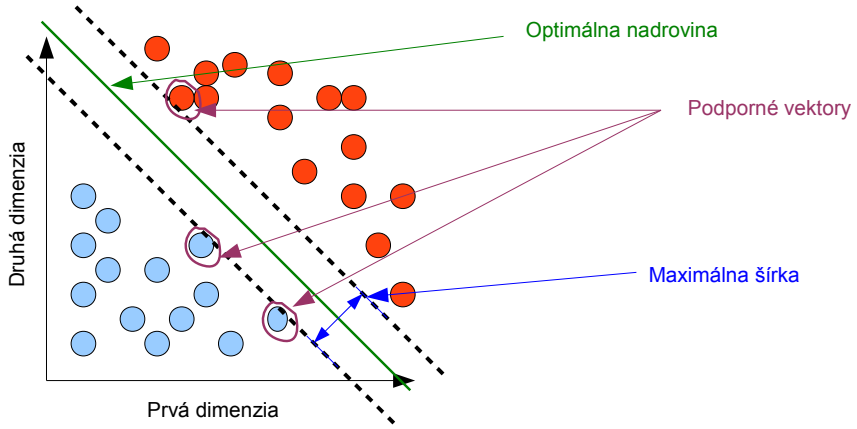
2.2 Support vector machine (SVM)

Support vector machine je lineárny klasifikátor, ktorý dokáže riešiť nelineárne separovateľné úlohy. Pre rozpoznávanie hudobných štýlov ako bolo uvedené v kapitole 2, je často využívaný a dosahuje veľmi dobré výsledky, čo môžeme vidieť v [13], kde je porovnávaný s ďalšími klasifikátormi. Informácie uvedené ďalej o tom, ako SVM funguje a čo je jeho cieľom, sú čerpané z týchto zdrojov [7, 2].

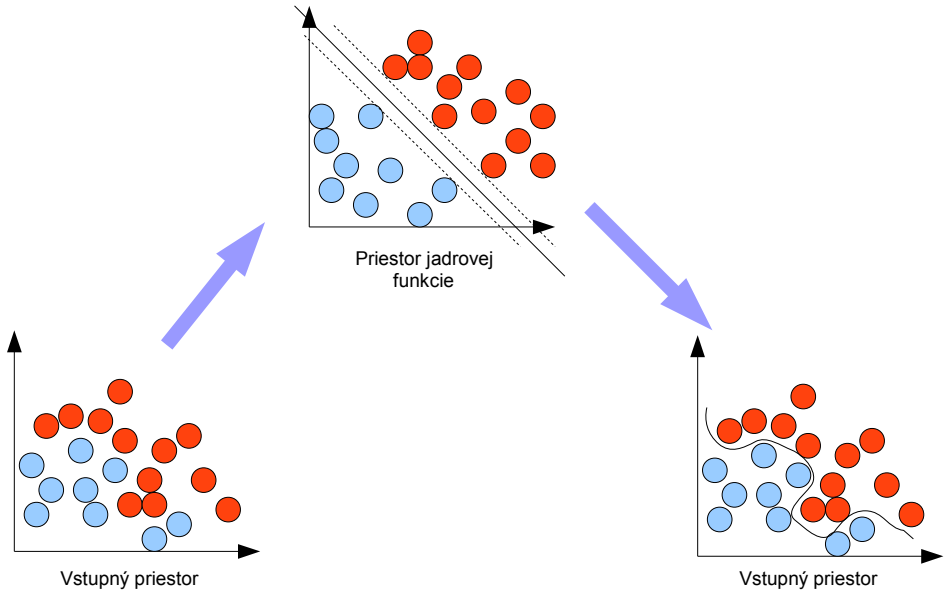
Cieľom SVM je lineárne oddeliť rozpoznávané triedy maximálnou možnou lineárnou hranicou, nadrovinou (viz. obrázok 2.3). Pre popis tejto oddeľovacej nadroviny sa používajú podporné vektory (support vectors). Tieto podporné vektory sú vlastne vektory príznakov trénovacej sady, ktoré sa nachadzajú najbližšie k oddeľovacej nadrovine. Takže SVM je schopné nájsť tie trénovacie príklady, ktoré sú pre nájdenie oddeľovacej nadroviny podstatné a ostatné trénovacie vzorky sú ďalej pre tento klasifikátor nepotrebné.

Bežne však rozpoznávané triedy vo vstupnom priestore nie sú lineárne separovateľné. SVM preto využíva prevod týchto tried zo vstupného priestoru do viac-dimenzionálneho priestoru, kde je túto lineárnu hranicu možno efektívnejšie nájsť. Práve pre tento účel SVM používa jadrové funkcie, ktoré fungujú tak, že rozpoznávané triedy sú zo vstupného priestoru prevedené do priestoru príslušnej jadrovej funkcie, kde sa lineárne separujú tak, že sa nájde oddeľovacia nadrovina a táto lineárna hranica je naspäť na mapovaná do vstupného priestoru, čo možno vidieť na obrázku 2.4. Takýmto spôsobom práve môžu vzniknúť rôzne nelineárne hranice medzi rozpoznávanými triedami.

Nie vždy je však dobré triedy presne oddeliť. Hlavne, keď máme k dispozícií konečnú množinu trénovacích dát, ktoré môžu obsahovať nielen reprezentatívne vzorky. Z toho dôvodu SVM obsahuje parameter C, ktorý reprezentuje mieru tolerancie chýb pri oddeľovaní tried. Hľadanie správnej hodnoty parametru C, môže byť spojené s hľadaním hodnôt parametrov príslušnej jadrovej funkcie. V [9] používajú pre túto úlohu mriežkové vyhľadávanie (grid search).



Obrázek 2.3: Nadrovina medzi dvoma rozpoznávanými triedami. Nadrovina je popísaná podpornými vektormi, ktoré sú vlastne vhodnými vektormi príznakov rozpoznávaných tried. Nadrovina je vždy natočená tak aby mala najväčšiu možnú šírku medzi rozpoznávanými triedami.



Obrázek 2.4: Tu je zobrazené použitie jadrovej funkcie. Triedy sú vo vstupnom priestore lineárne neseparovateľné. Použitím jadrovej funkcie sa tieto triedy prevedú zo vstupného priestoru do priestoru použitej jadrovej funkcie, kde je možné tieto triedy lineárne separovať. Nájdený lineárny separátor sa naspäť prevedie z priestoru použitej jadrovej funkcie do vstupného priestoru.

Kapitola 3

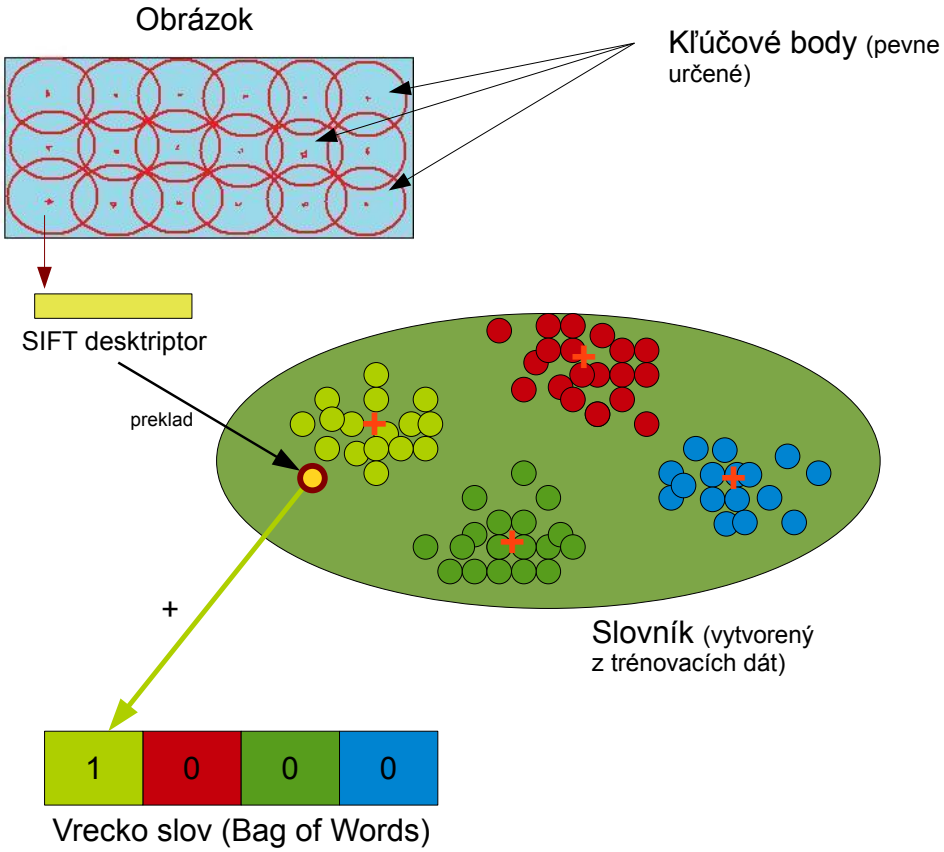
Obrazové príznaky

V tejto kapitole bude opísaná extrakcia príznakov z obrázka založená na lokálnych príznakoch a deskriptoroch, ktoré tieto lokálne príznaky popisujú. Lokálnym príznakom je časť obrazu so špecifickými vlastnosťami a deskriptorom je vektor, ktorý tento príznak popisuje. Lokálne príznaky a ich deskriptory majú široké využitie a sú využívané napríklad pri detekcií známych objektov v obraze alebo vo videu, pri vyhľadávaní obrázkov podľa obsahu, pri detekcií pohybu vo videu, pri zarovnávaní obrazov v panorámach, atď. Dôvodom ich širokého využitia, ako je uvedené v [15], sú vlastnosti týchto deskriptorov lokálnych príznakov. Dekriptory lokálnych príznakov sú všeobecne invariantné voči zmene veľkosti, otočeniu a čiastočne aj voči zmene osvetlenia. Vďaka týmto vlastnostiam je práve možné tieto deskriptory dobre porovnávať a nachádzať podobné vzory v rôznych obrázkoch. Medzi tieto deskriptory lokálnych príznakov patria napríklad SIFT deskriptory, SURF dekriptory, BRIEF deksriptory a iné.

Ďalej bude práve popísaný prístup použitý v [11, 17], zaoberajúci sa extrakciou príznakov z obrázka a ich spracovaním do použiteľnej podoby, založený na týchto deskriptoroch lokálnych príznakov. V tomto prístupe ide o vytváranie Vrecka Slov (Bag of Words) alebo tiež len BoW reprezentácie, pre príslušné obrázky. Získavanie BoW reprezentácie sa skladá z dvoch krokov, za prvé z extrakcie lokálnych príznakov z obrázka a za druhé prevedenie deskriptorov lokálnych príznakov obrázka pomocou slovníka na BoW reprezentáciu obrázka.

Pre extrakciu príznakov z obrázka, ako je uvedené v [11, 17], sa používajú deskriptory lokálnych príznakov. V [11, 17] používajú SIFT dekriptory lokálnych príznakov. V [15] sa tieto deskriptory počítajú pre kľúčové oblasti, ktoré sa v obrázku lokalizovali pomocou Rozdielov Gaussových funkcií (Diferential of Gaussion), avšak pre prístup opisovaný v tejto časti práce a použitý v [11, 17], sa používa hustý odber príznakov z homogénnej mriežky. To znamená, že kľúčové oblasti, pre ktoré sú SIFT deskriktory extrahované, sú napevno rozmiestnené v obrázku vzdialené od seba v určitých intervaloch (v [11] bol tento interval osem pixelov a SIFT deskriptory o veľkosti 16x16 pixelov). SIFT dekriptory sú v kľúčových oblastiach získavané na základe gradientov v okolí a sú histogramom týchto gradientov [15]. Jeden SIFT deskriptor je vlastne 128 rozmerný vektor.

Druhou fázov získavania BoW reprezentácie obrázka je spracovanie lokálnych príznakov a ich desktriptorov a vytvorenie z nich BoW reprezentácie. Pre tento účel, ako je uvedené v [11, 17], sa používa slovník, ktorý je vytvorený pomocou K-means z lokálnych príznakov a ich desktriptorov trénovacích obrázkov a každý zhluk vytvorený pomocou K-means reprezentuje jedno kódové slovo. K-means sa používa pre určenie stredov jednotlivých zhlukov a potom na základe minimálnej vzdialenosti k stredu sa určí príslušné kódové slovo, ktoré bude deskriptor lokálnych príznakov reprezentovať. Počet zhlukov reprezentuje počet kó-



Obrázek 3.1: Princíp akým sa zo SIFT deskriptorov získava prostredníctvom slovníka Vrecko slov (BoW). Ukázaná je extrakcia prvého SIFT deskriprota obrázka z lokálnej oblasti, jeho prevod na kódové slovo a následná inkrementácia príslušnej položky v histograme (BoW).

dových slov. V [11] bolo použitých 400 kódových slov a v [17] vyskúšali použiť až cez 4000 kódových slov. Po tom ako sa deskriptory lokálnych príznakov obrázka preložia na kódové slová, sa z týchto kódových slov urobí histogram početnosti a tento histogram je BoW reprezentáciou príslušného obrázka. BoW reprezentácia obrázka je teda vektor, ktorého veľkosť závisí na počte kódových slov slovníka. Postup pri vytváraní BoW reprezentácie obrázka je ilustrovaný na obrázku 3.1.

Vyššie uvedený postup pre získanie BoW reprezentácie obrázka je metódou extrakcie príznakov a s touto metódou extrakcie príznakov sa ako klasifikátor pri obidvoch prácach [11, 17], z ktorých táto kapitola najviac čerpala, použil Support Vector Machine. V [17] bol SVM použitý s chi-kvadrát jadrovou funkciou a s novými multi-jadrovými metódami.

Kapitola 4

Návrh

Táto kapitola popisuje aké metódy extrakcie príznakov a klasifikácie som pre rozpoznávanie hudobných štýlov navrhol a využil v tejto práci. Pri návrhu som vychádzal z teoretických znalostí opísaných v kapitole 2 už existujúcich prístupov pri rozpoznávaní hudobných štýlov, ale aj z kapitoly 3, ktorá opisuje extrakciu príznakov z obrázka a pri jednej z metód extrakcie príznakov navrhnutej pre túto prácu som práve tieto poznatky využil.

Navrhol som dve metódy extrakcie príznakov a prvá sa nazýva Metóda Histogramov Mel-fekvenčných kepstrálnych koeficientov, ktorá je bližšie popísaná v 4.1. Táto prvá metóda extrakcie príznakov je založená na extrakcií Mel frekvenčných kepstrálnych koeficientov, ktoré ako je možné vidieť v kapitole 2, dosahujú pri rozpoznávaní hudobných štýlov veľmi dobré výsledky. Druhou navrhnutou metódou extrakcie príznakov je nový prístup, ktorý som nazval Metóda Obrázovej Spektrálnej analýzy a ktorý je bližšie popísaný v 4.2. Tento druhý prístup navrhnutý pre extrakciu príznakov je založený na prevode zvuku na spektrogram, z ktorého sa extrahujú príznaky, ktoré sa používajú pre rozpoznávanie obrazu.

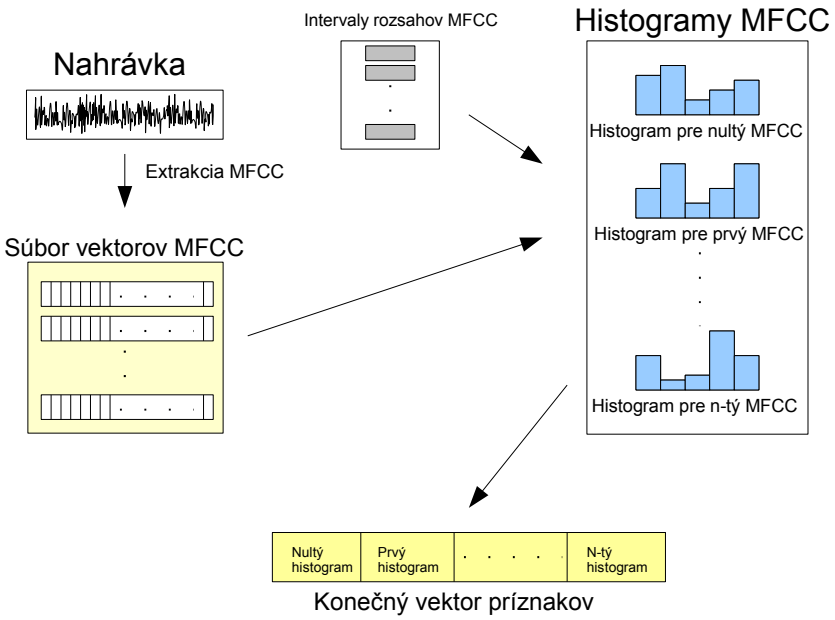
Pre obidve tieto metódy extrakcie príznakov som ako klasifikátor navrhnutý Support Vector Machine, ktorý je bližšie popísaný v sekcií 2.2 a ako je uvedené v kapitole 2 je pri rozpoznávaní hudobných štýlov často využívaný a dosahuje pre túto úlohu dobré výsledky.

Implementačné detaily navrhnutých metód sú bližšie opísané v kapitole 5.

4.1 Metóda Histogramov Mel-frekvenčných kepstrálnych koeficientov

Metóda Histogramov Mel-frekvenčných kepstrálnych koeficientov je založená na extrakcií Mel-frekvenčných kepstrálnych koeficientov (MFCC) zo signálu (nahrávky). MFCC sú bližšie popísané v sekcií 2.1 a ako možno vidieť v kapitole 2, pri rozpoznávaní hudobných štýlov dosahujú dobré výsledky. Práve kvôli ich dobrým výsledkom som MFCC pre rozpoznávanie hudobných štýlov použil aj v tejto práci.

Princíp extrakcie príznakov metódou Histogramov Mel-frekvenčných kepstrálnych koeficientov je zobrazený na obrázku 4.1. Postup je taký, že najprv sú z nahrávky extrahované pomocou krátkodobej analýzy MFCC. Po extrakcií MFCC je k dispozícií súbor vektorov týchto Mel-frekvenčných kepstrálnych koeficientov. Zo vzniknutého súboru vektorov MFCC sa urobia histogramy pre každý jeden koeficient. To znamená, že pre nultý, prvý, druhý a ostatné MFCC sa urobí príslušný histogram. Aby ale bolo možné tieto histogramy vytvoriť je nutné ešte pred ich vytvorením získať informáciu o tom, aký je rozsah hodnôt pre jednotlivé Mel-frekvenčné kepstrálné koeficienty. Tieto rozsahy jednotlivých Mel-frekvenčných



Obrázek 4.1: Princíp vytvorenia príznakov pomocou Metódy Histogramov Mel-frekvenčných kepstrálnych koeficientov. Najprv sú z nahrávky extrahované MFCC, z ktorých sú vytvorené jednotlivé histogramy pre jednotlivé Mel-frekvenčné kepstrálne koeficienty. Tieto histogramy MFCC sú následne skonkatenované do výsledného vektoru príznakov reprezentujúceho príslušnú nahrávku.

kepstrálnych koeficientov je dobré zisťovať tak, aby tieto rozsahy MFCC boli normalizované voči nahrávkam rôznych hudobných štýlov, pretože rôzne hudobné štýly môžu mať tento rozsah rozdielny. Tu sa ponúka možnosť získať tieto rozsahy MFCC z celej trénovacej sady, ktorá obsahuje nahrávky rôznych hudobných štýlov alebo použiť reprezentatívnu vzorku nahráviek trénovacej sady obsahujúcu nahrávky z každého hudobného štýlu.

Po získaní rozsahov pre jednotlivé Mel-frekvenčné kepstrálne koeficienty sa teda môže pristúpiť k vytvoreniu histogramov pre tieto jednotlivé Mel-frekvenčné kepstrálne koeficienty. Vytváranie histogramov MFCC sa vykonáva pre každý súbor vektorov MFCC extrahovaného z nahrávky zvlášť, čím sa získa údaj o tom, v akých rozsahoch a v akých hodnotách je najväčšie zastúpenie príslušných MFCC v tejto nahrávke. Po vytvorení MFCC histogramov sa tieto histogramy skonkatenujú do výsledného vektoru, ktorý bude príznakovým vektorom reprezentujúci príslušnú nahrávku.

4.2 Metóda Obrazovej Spektrálnej analýzy

Metóda Obrazovej Spektrálnej analýzy, ako bolo uvedené v kapitole 1 využíva možnosť previesť signál z 1D na 2D signál. Je to teda metóda, ktorá využíva možnosť urobiť z audia spektrogram, na ktorý sa dá pozerať ako na 2D signál, a teda je na neho možné použiť postupy ako na obrázok. Táto metóda extrakcie príznakov sa teda skladá z dvoch krokov a to z prevedenia audia nahrávky na spektrogram (obrázok) a druhým krokom je extrakcia príznakov z tohto spektrogramu pomocou postupov pre extrakciu príznakov z obrázka.

Postup pre vytvorenie spektrogramu (obrázka) zo zvuku je založený na krátkodobej

Fourierovej transformácií (STFT). Pomocou STFT sa získa súbor spektier, konkrétne Mel-frekvenčných spektier, získaných z veľmi malých úsekov signálu. Spektrum získané prostredníctvom STFT sa na Mel-frekvenčné spektrum prevedie pomocou Melových baniek filtrov, podobne ako v sekcií 2.1 pri MFCC. Aby sa znížil dynamický rozsah týchto Mel-frekvenčných spektier sú tieto Mel-frekvenčné spektrá prevedené na Logaritmické Mel-frekvenčné spektrá. K tomuto účelu som navrhol vzorec:

$$x = \left(\frac{\log(e)}{maxL}\right) * maxV \tag{4.1}$$

, kde e je jedna hodnota Mel energie spektra, $maxL$ je maximálna logaritmická hodnota Mel energie spektra v celom súbore a $maxV$ je výledná logaritmická hodnota Mel energie spektra. Týmto postupom sa získa súbor Logaritmických Mel-frekvenčných spektier, ktorý bude výsledným spektrogramom.

Z vytvoreného obrázka sa potom pomocou prístupu opísaného v kapitole 3 extrahujú príznaky vo forme vektora BoW. Pre popis lokálnych príznakov sa využijú SIFT dekriptory a použije sa hustý odber príznakov z homogénnej mriežky, ktorá určí kľúčové oblasti, pre ktoré sa SIFT dekriptory budú vytvárať. Z reprezentatívnej vzorky nahráviek sa vytvorí slovník a na pomocou slovníka sa potom SIFT deskriptory preložia na kódové slová. Z kódových slov sa následne vytvorí BoW reprezentácia príslušného obrázka a teda príslušnej nahrávky.

Kapitola 5

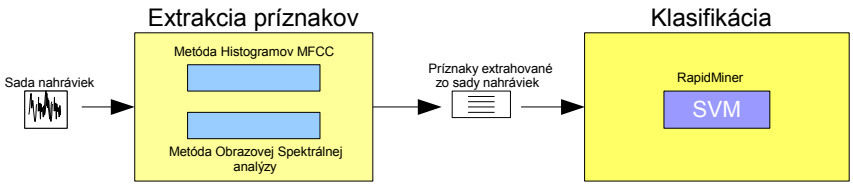
Implementácia

Táto časť práce vychádza z kapitoly 4 a je určená k popisu implementačných detailov jednotlivých častí systému, ktoré som pre rozpoznávanie hudobných štýlov navrhol. Keďže pri práci som tiež využil už existujúce nástroje, obsahuje táto kapitola aj prehľad týchto použitých nástrojov a ich nastavení.

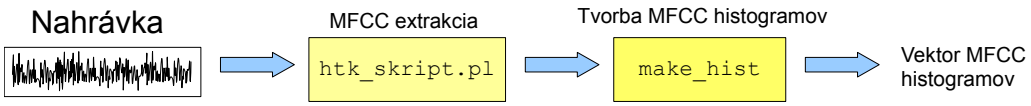
Finálny systém pre rozpoznávanie hudobných štýlov, ktorý som pre túto prácu vytvoril sa skladá z dvoch nezávisle pracujúcich častí, čo je možné vidieť na obrázku 5.1. Prvou časťou je extrakcia príznakov a druhou klasifikácia.

Pre extrakciu príznakov som, ako je aj uvedené v kapitole 4, navrhol dve metódy extrakcie príznakov. Prvou metódou extrakcie príznakov je *Metóda Histogramov Mel-frekvenčných kepstrálnych koeficinetov*. Metóda Histogramov Mel-frekvenčných kepstrálnych koeficinetov využíva viacero vytvorených nástrojov pre získanie výsledného vektoru príznakov popisujúceho príslušnú nahrávku, čo je zobrazené na obrázku 5.2. Postup pri extrakcií príznakov metódou Histogramov MFCC je teda taký, že najprv sa z nahrávky extrahujú Mel-frekvenčné kepstrálne koeficienty použitím skriptu `htk_skript.pl`, ktorý je popísaný v sekcií 5.2 a ktorého použitie pre túto metódu je bližšie opísané v podsekcií 5.2.1. Pre extrakciu príznakov skript `htk_skript.pl` využíva Toolkit Skrytého Markovovho Modelu (HTK), ktorého bližší popis a prostriedky, ktoré som v tejto práci využil, sú popísané v sekcií 5.1. V ďalšom kroku extrakcie príznakov pomocou metódy Histogramov MFCC sa použije program `make_hist`, ktorého činnosť je bližšie opísaná v sekcií 5.3. Program `make_hist` zo získaných Mel-frekvenčných kepstrálnych koeficientov získaných pomocou skriptu `htk_skript.pl` vytvorí pre každý koeficient histogram a tieto histogramy potom skonkatenuje do výsledného vektoru príznakov popisujúceho príslušnú nahrávku.

Druhou metódou extrakcie príznakov je *Metóda Obrazovej Spektrálnej analýzy*, ktorá využíva pri extrakcií príznakov podobne ako metóda Histogramov Mel-frekvenčných kepstrálnych koeficientov viacero vytvorených nástrojov pre získanie výsledného vektoru príznakov popisujúceho príslušnú nahrávku, čo je možné vidieť na obrázku 5.3. Prvým krokom extrakcie príznakov pomocou metódy Obrazovej spektrálnej analýzy je extrakcia Mel energií spektra pomocou skriptu `htk_skript.pl`, ktorého bližší popis použitia pre túto metódu extrakcie príznakov obsahuje podsekcia 5.2.2. Pre extrakciu príznakov, ako bolo tiež spomenuté vyššie, skript `htk_skript.pl` využíva Toolkit Skrytého Markovovho Modelu (HTK). V druhom kroku extrakcie príznakov metódou Obrazovej Spektrálnej analýzy sa použije program `make_PGM`, ktorého bližší popis obsahuje sekcia 5.4 a z extrahovaných Mel energií spektra získaných pomocou skriptu `htk_skript.pl`, `make_PGM` vytvorí obrázok vo formáte PGM, ktorý spektrogramom nahrávky. V poslednom kroku extrakcie príznakov pomocou metódy Spektrálnej Obrazovej analýzy sa použije sada nástrojov vyvinutá na UPGM FIT



Obrázek 5.1: Štruktúra finálneho systému pre rozpoznávanie hudobných štýlov, ktorý som vytvoril. Tento systém oddeľuje extrakciu príznakov a klasifikáciu do dvoch nezávisle pracujúcich častí. Pre extrakciu príznakov boli vyskúšané dve metódy (Metóda Histogramov MFCC a Metóda Obrazovej Spektrálnej analýzy). A ako klasifikátor sa použil Support Vector Machine v nástroji RapidMiner.



Obrázek 5.2: Extrakcia príznakov Metódou Histogramov Mel-frekvenčných kepstrálnych koeficientov, ktorá je vykonávaná vo viacerých krokoch. Každý krok má na starosti jeden nástroj (program alebo skript) s príslušnou funkciou, ktorú vykonáva.

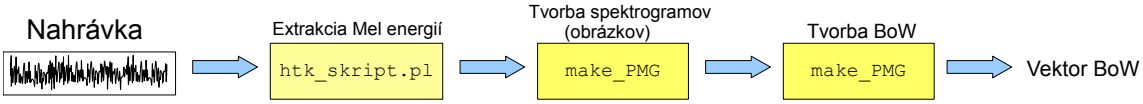
VUT pre účely klasifikácie obrazu, ktorá sa používá primárne pre účely súťaží TRECVID a PASCAL [10]. Tieto nástroje umožňujú jednoducho zo sady obrázkov extrahovať ich BoW reprezentáciu, v mojom prípade BoW reprezentáciu nahrávky.

Druhou časťou finálneho systému pre rozpoznávanie hudobných štýlov je klasifikátor a pri obidvoch metódach extrakcie príznakov popísaných vyššie som použil Support Vector Machine. Support Vector Machine som však pre túto prácu neimplementoval, ale využil som už existujúce implementácie tohto klasifikátora.

Teda pre prístup založený na extrakcií príznakov metódou Histogramov Mel-frekvenčných kepstrálnych koeficientov a pre prístup založený na extrakcií príznakov metódou Obrazovej Spektrálnej analýzy, som pri klasifikácií využil už existujúci nástroj RapidMiner obsahujúci klasifikátor Support Vector Machine. RapidMiner je open-source systém pre dolovanie dát, strojové učenie, dolovanie textov plus na ďalšie úkony a v najnovších verziách obsahuje aj prehľadné užívateľské rozhranie [3, 1]. V tejto práci som použil RapidMiner verzie 5 dostupný na [3]. Support Vector Machine je len jeden z rady klasifikátorov, ktoré RapidMiner obsahuje. Okrem klasifikátora SVM som využili v RapidMineri aj prostriedky pre cross-validáciu a prostriedky pre optimalizáciu parametrov SMV a parametrov príslušných jadrových funkcií, s ktorými bola klasifikácia pri experimentoch vykonávaná. Jednotlivé nastavenia, s ktorými sa experimenty v RapidMineri prevádzali sú uvedené v kapitole 6.

5.1 Hidden Markov Model Toolkit (HTK)

Toolkit Skrytého Markovovho modelu, ktorého návod na použitie a popis možno nájsť v [22], je nástroj pre vytváranie a manipuláciu so Skrytými Markovovmi Modelmi (HMM). HTK je primárne určený k rozpoznávaniu reči. Obsahuje nástroje pre analýzu reči (všeobecne 1D signálu), trénovanie pomocou HMM, testovanie a analýzu výsledkov. Je zložený z rady



Obrázek 5.3: Extrakcia príznakov Metódou Obrazovej Spektrálnej analýzy, ktorá je vykonávaná vo viacerých krokoch. Každý krok má na starosti jeden nástroj (program alebo skript) s príslušnou funkciou, ktorú vykonáva.

knižníc a nástrojov, ktoré sú stránke [4] dostupné vo forme zdrojových kódov v jazyku C.

HTK som v tejto práci využil pri extrakcií príznakov a informácie som čerpal z [22]. Extrakcia príznakov v HKT je založená na krátkodobej analýze (viz. obrázok 2.1) a v HTK existuje špeciálny nástroj `HCopy`, ktorý na základe zadaných parametrov extrahuje zo vstupného súboru požadované príznaky. K zadávaniu týchto parametrov, určujúcich aké príznaky majú byť zo vstupného súboru extrahované, slúži konfiguračný súbor a spustenie `HCopy` potom vyzerá nasledovne:

```
HCopy -C nazov_konfiguracneho_suboru vstupny_subor vystupny_subor_priznakov
```

Zadávanie jednotlivých parametrov v konfiguračnom súbore závisí od typu príznakov, ktoré chceme získať. Ďalej sú teda spomenuté a vysvetlené tie, ktoré som využil, a teda sú pre účel tejto práce relevantné.

Prvým spomenutým parametrom je `SOURCEKIND`, ktorý určuje druh súboru, z akého sa budú príznaky extrahovať (Pre WAV súbory je nastavovaný na `WAVEFORM`). Ohľadom vstupného súboru sa ešte zadáva aj parameter `SOURCEFORMAT` určujúci, akého formátu je vstupný súbor, teda napríklad či ide len o čistý signál (NOHEAD) alebo ešte súbor obsahuje hlavičku. Keďže extrakcia príznakov je založená na krátkodobej analýze, je možné si prostredníctvom parametru `WINDOWSIZE` zvoliť veľkosť okna a pomocou parametru `TARGETRATE` veľkosť posunu tohto okna. Obidva parametre `WINDOWSIZE` ako aj `TARGETRATE` sa zadávajú v jednotkách 100-kách ns. Pre zadanie vzorkovacej periódy sa používa parameter `SOURCERATE`, zadávaný v jednotkách desiatich mikrosekúnd. Prostredníctvom parametru `USEHAMMING` možno zadať, či sa má použiť alebo nepoužiť Hammingové okno. Ak je prítomný v signále DC offset nastavuje sa parameter `ZMEANSOURCE` a pre normalizáciu energie v zvukových nahrávkach sa používa parameter `ENORMALISE`. Pre zvýraznenie vstupného signálu slúži parameter `PREEMCOEF`, ktorého hodnota je z intervalu 0 až 1. Pri rečových signáloch, ako možno vidieť v príkladoch z [22] sa väčšinou táto hodnota nastavuje na 0,97. Druh výsledných extrahovaných príznakov určuje parameter `TARGETKIND` a v tejto práci som využil dva druhy extrahovaných príznakov.

Prvým druhom extrahovaných príznakov, ktoré som použil, sú Mel-frekvenčné kepstrálne koeficienty, pre ktoré sa parameter `TARGETKIND` nastavuje na hodnotu `MFCC`. V prípade, že chceme do výsledku zahrnúť aj nultý koeficient, tak sa k `MFCC` pripojí prípona `_0`, takže `TARGETKIND` sa bude rovnať `MFCC_0`. Pre výpočet Mel-frekvenčných kepstrálnych koeficientov v HTK, ako je uvedené v [22], sa využíva podobný princíp ako je uvedený v sekcií 2.1, a teda na spektrum signálu sa aplikujú logaritmické banky filtrov a výsledné kepstrálne koeficienty sa získajú pomocou Diskrétnej Kosinusovej transformácie. O tom koľko kanálov filterbaniek sa použije rozhoduje parameter `NUMCHANS`. Počet požadovaných Mel-frekvenčných kepstrálnych koeficientov (okrem nultého, ktorý do tohto parametru zahrnutý nie je) sa určuje pomocou parametru `NUMCEPS`.

Druhým extrahovaným druhom príznakov, ktoré som použil, sú príznaky získavané analýzou pomocou baniek filtrov. Pre extrakciu týchto príznakov sa `TARGETKIND` nastavuje na `FBANK`. Vypočítajú sa tak, že na frekvenčné spektrum signálu sú aplikované banky filtrov, presnejšie Melové banky filtrov, čím sa získajú Melové energie spektra. Počet použitých filtrov a teda získaných energií určuje parameter `NUMCHANS`. V prípade, že analýza pomocou baniek filtrov nemá byť vykonaná na celé frekvenčné spektrum, tak je možné nastaviť vrchnú alebo spodnú frekvenČú hranicu. Vrchná frekvenčná hranica sa nastavuje pomocou parametra `HIFREQ` a spodná pomocou parametra `LOFREQ`.

Všetky tieto príznaky sú však extrahované a interpretované tak aby sa mohli ďalej využívať pri rozpoznávaní práve v HTK. K interpretácií do prijateľnej podoby v HTK slúži nástroj `HList`, ktorý prevádza tieto príznaky pri použití parametra `-r` do prijateľnej podoby tak, že jeho výstup obsahuje na každom riadku jeden vektor príznakov pre jedno okno signálu.

5.2 htk_skript.pl

Táto sekcia obsahuje popis skriptu `htk_skript.pl` a popis jeho použitia pre jednotlivé metódy extrakcie príznakov, ktoré som pre túto prácu navrhol.

Skript `htk_skript.pl` je skriptom vytvoreným pre túto prácu v jazyku Perl. Hlavnou úlohou tohto skriptu je extrakcia príznakov založená na využívaní nástrojov (`HCopy` a `HList`) Toolkitu Strytého Markovovho Modelu (viz. sekcia 5.1). Okrem nástrojov HTK využíva skript `htk_skript.pl` aj ďalšie nástroje, ktoré sú pre jeho činnosť potrebné. Týmito nástrojmi sú nástroje na konverziu nahráviek do formátu WAV, kde prvým je `mpg123` využívaný pre prevod MP3 na WAV a druhým `sox` využívaný pre prevod AU formátu na WAV. Okrem prevodu z AU na WAV sa `sox` v `htk_skript.pl` používa aj na prevzorkovanie nahrávky.

Principiálne skript `htk_skript.pl` pracuje tak, že sa mu zadá vstupný adresár, rekurzívne sa vo vstupnom adresári zanoruje a vyhľadáva nahrávky zadaného formátu. Je schopný pracovať s formátmi MP3, WAV a AU. Po nájdení nahrávky túto nahrávku prevedie na formát WAV a to v prípade, ak je táto nahrávka vo formáte MP3 alebo AU. V prípade, že sa to od `htk_skript.pl` vyžaduje, prevzorkuje túto nahrávku na zvolenú vzorkovaciu frekvenciu. Posledným úkonom `htk_skript.pl` je extrakcia zadaných príznakov z WAV nahrávky na základe konfiguračného súboru a vytvorenie výstupného súboru z týchto príznakov. Výstupný súbor s extrahovanými príznakmi je vytvorený v zadanom výstupnom adresári, v ktorom `htk_skript.pl` vytvorí rovnakú adresárovú štruktúru ako je vo vstupnom adresári a výstupný súbor s extrahovanými príznakmi sa teda vytvorí v príslušnom adresári na základe pozície, ktorú má nahrávka vo vstupnom adresári, z ktorej sa príznaky extrahovali. Keďže `HCopy` extrahuje príznaky na základe konfiguračného súboru je nutné tento konfiguračný súbor skriptu `htk_skript.pl` zadať.

Skript `htk_skript.pl` ako už bolo spomenuté v úvode tejto kapitoly využívaný pri obidvoch metódach extrakcie príznakov, ktoré som pre túto prácu navrhol. Použitie skriptu `htk_skript.pl` pre metódu Histogramov Mel-frekvenčných kepstrálnych koeficientov je opísané v podsekcií 5.2.1 a použitie pre metódu Obrazovej Spektrálnej analýzy je opísané v podsekcií 5.2.2.

5.2.1 Použitie `htk_skript.pl` pri Metóde Histogramov MFCC

Ako je uvedené v úvode tejto kapitoly a možno to vidieť aj na obrázku 5.2, tak `htk_skript.pl` sa používa ako prvý krok extrakcie príznakov metódou Histogramov MFCC, takže táto podsekcia opisuje použitie tohto skriptu pri tejto metóde. Nachádza sa tu opis použitých parametrov v konfiguračnom súbore nástroja `HCopy`, s ktorým `htk_skript.pl` pracuje, ako aj budú uvedené hodnoty týchto parametrov, s ktorými som experimenty na dátových sadách vykonal.

Keďže úlohou `htk_skript.pl` pri Metóde Histogramov MFCC je extrakcia Mel-frekvnčných kepstrálnych koeficientov z nahráviek, tak v konfiguračnom súbore som parameter `TARGETKIND` nastavil na `MFCC`. Využíval saom však aj nultý Mel-frekvenčný koeficient, takže parameter `TARGETKIND` som nastavil na `MFCC_0`. Použitých bolo 31 baniek filtrov. Parameter pre zvýraznenie signálu som nastavil na hodnotu $0,97$ a použil som Hammingové okno. Tiež bol nastavený aj parameter `ENORMALISE` a `ZMEANSOURCE`. Posun okna som nastavil na 10 ms a experimenty v kapitole 6 som potom vykonal s takto nastavenými parametrami, pričom som experimentoval s veľkosťou použitého okna a počtom Mel-frekvenčných keprálnych koeficientov. V konfiguračnom súbore sa tiež menila vzorkovacia frekvencia a to v závislosti na vzorkovacej frekvencií príslušnej dátovej sady.

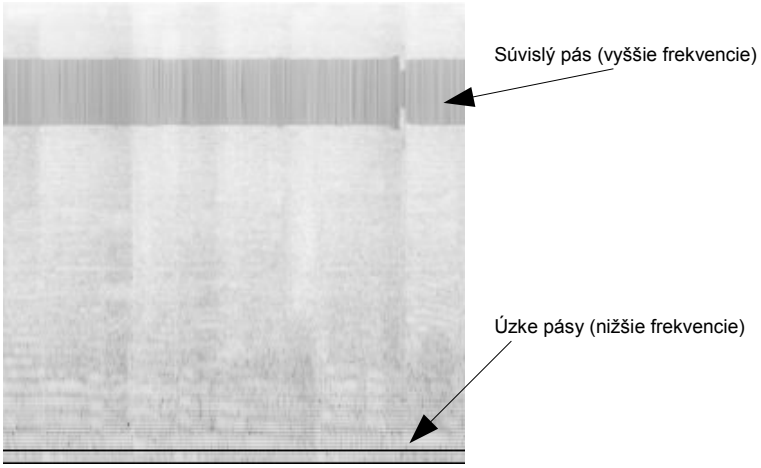
Výstupom tejto fázy metódy Histogramov MFCC je súbor obsahujúci na každom riadku vektor Mel-frekvenčných kepstrálnych koeficientov pre jedno okno signálu. Hodnoty jednotlivých MFCC sú od seba oddelené medzerou.

5.2.2 Použitie `htk_skript.pl` pri Metóde Obrazovej Spektrálnej analýzy

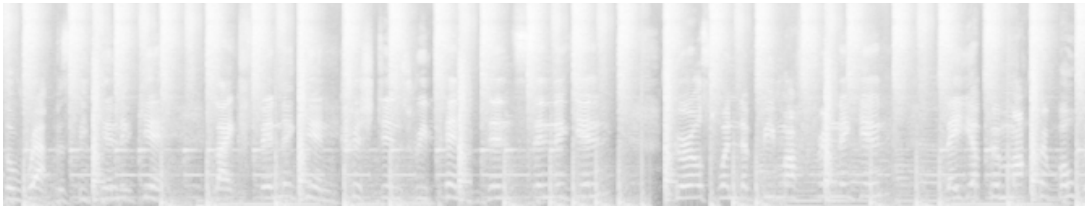
Tak ako aj pri metóde Histogramov MFCC, tak aj pri metóde Obrazovej Spektrálnej analýzy sa pre prvý krok extrakcie príznakov používa skript `htk_skript.pl` (viz. obrázok 5.3). Tento skript extrahuje príznaky z nahrávky na základe konfiguračného súboru a parametrov v ňom, preto sa v tejto časti nachádza opis týchto použitých parametrov.

Keďže úlohou `htk_skript.pl` pri metóde Obrazovej Spektrálnej analýzy je extrakcia Mel energií spektra z nahráviek, tak v konfiguračnom súbore som parameter `TARGETKIND` nastavil na `FBANK`. Tak ako pri metóde Histogramov MFCC, tak aj pri Metóde Obrazovej Spektrálnej analýzy som parameter pre zvýraznenie signálu nastavil na hodnotu $0,97$ a tiež som použil Hammingové okno. Okrem toho boli nastavené aj parametre `ENORMALISE` a `ZMEANSOURCE`. Vzorkovacia frekvencia bola nastavená v závislosti na vzorkovacej frekvencií príslušnej dátovej sady. Tieto parametre zostali pri experimentovaní nemenné. Nastavenie správnej veľkosti okna, nastavenie posunu okna a počet extrahovaných Mel energií však bolo komplikovanejšie, pretože rozlíšenie a vzhľad výsledného obrázka (spektrogramu), sa práve menil v závislosti na týchto parametroch a preto sa ich správne nastavenia som určoval až testovaním.

Pri extrahovaní väčšieho množstva energií spektra bol výsledný obrázok ostrý a mal väčšie rozlíšenie, avšak v častiach obrázka patriacim nižším frekvenciám vznikali viditeľné pásy pravdepodobne spôsobené tým, že trojuholníky (filtre) sú v nízkych frekvenciách príliš tenké. Okrem toho sa v časti obrázka patriacej vyšším frekvenciám vyskytoval zvláštny súvislý „pás“, ktorého pôvod sa v tejto práci nepodarilo vysvetliť, je možné, že za to môže MP3 kompresia vstupu. Pri znížení počtu extrahovaných energií sa tieto pásy, či už z oblasti obrázka patriacej nižším frekvenciám alebo súvislý pás v oblasti obrázka patriacej vyšším, neodstránili a výsledný obrázok sa znížením počtu extrahovaných energií rozmazal. Vyskúšal som prevzorkovať nahrávku na väčšiu frekvenciu a síce sa pásy v oblasti patriacej nižším frekvenciám čiastočne odstránili, pás v oblasti patriacej vyšším frekvenciám sa



Obrázek 5.4: Ukážka časti obrázka PGM vzniknutého pri extrahovaní príznakov pri Metóde obrazovej Spektrálnej analýze. Pri vytvorení som použil 256 Mel energií spektra, okná signálu mali veľkosť 25 ms, posun okna 10 ms a určil som žiadnu vrchnú frekvenčnú hranicu.



Obrázek 5.5: Ukážka časti obrázka PGM vzniknutého pri extrahovaní príznakov pri Metóde obrazovej Spektrálnej analýze. Pri vytvorení som použil 128 Mel energií spektra, okná signálu mali veľkosť 100 ms, posun okna 25 ms a určil som hornú frekvenčnú hranicu na 12000 Hz.

úmerne vyššej vzorkovacej frekvencií zväčšoval. Pri odstraňovaní pásov z oblasti patriacej nižším frekvenciám najlepšie výsledky dosahovalo zväčšovanie okna signálu, z ktorého sa Mel energie extrahovali. Pás v oblasti patriacej vyšším frekvenciám sa síce zväčšením okna neodstránil, ale ani nezväčšil. Tento pás v oblasti patriacej vyšším frekvenciám som nakoniec odstránil pomocou určenia vrchnej frekvenčnej hranice, po ktorú sa Mel energie počítali. Určením vrchnej frekvenčnej hranice sa okrem odstránenie pásu v oblasti patriacej vyšším frekvenciám obrázok zostril a tak som mohol použiť aj menší počet extrahovaných Mel energií, pretože ako už bolo uvedené, pri použití menšieho množstva extrahovaných Mel energií sa obrázok rozmazal, čo som však práve určením vrchnej frekvenčnej hranice vyriešil.

Výsledná konfigurácia, pre ktorú sa Mel energie spektra z nahrávky extrahovali je teda nasledovná: veľkosť okna bola nastavená na 100 ms, posun tohto okna bol nastavený na 25 ms a bolo extrahovaných 128 Mel energií spektra. Na obrázku 5.5 možno vidieť, ako vyzerá obrázok pri použití tejto konfigurácie (128 Mel energií, 100 ms okno, 25 ms posun).

Výstupom `htk_skript.pl` a teda prvej fáze extrakcie príznakov metódou Obrazovej Spektrálnej analýzy je súbor Mel energií obsahujúci na každom riadku vektor 128 Mel energií oddelených medzerou.

5.3 make_hist

Ako možno vidieť na obrázku 5.2, tak použitie programu `make_hist` je druhou a poslednou fázou pri extrakcií príznakov metódou Histogramov Mel-frekvenčných kepstrálnych koeficientov. Hlavnou úlohou programu `make_hist` je vytvorenie vektoru histogramov Mel-frekvenčných kepstrálnych koeficientov, čo sú výsledne extrahované príznaky Metódy Histogramov MFCC, ktoré reprezentujú príslušnú nahrávku. V tejto sekcií je stručne popísané ako tento program funguje, jeho využitie v tejto práci ako aj jeho vstupy a výstupy.

Program `make_hist` som vytvoril pre účely tejto práce a je na implementovaný v jazyku C++. Pri svojej činnosti využíva už existujúci nástroj na parsovanie XML súborov vytvorený v C++ a nazývaný TinyXml viz. [18].

Vstupom `make_hist` je adresárová štruktúra obsahujúca pre každý žáner (žáner je reprezentovaný adresárom) súbory obsahujúce vektory MFCC získané krátkodobou analýzou. Jeden vektor MFCC je na jednom riadku súboru, pričom jednotlivé koeficienty sú od seba oddelené medzerou. Je možné naraz zadať ako vstup adresárovú štruktúru trénovacej sady aj nevidenej sady.

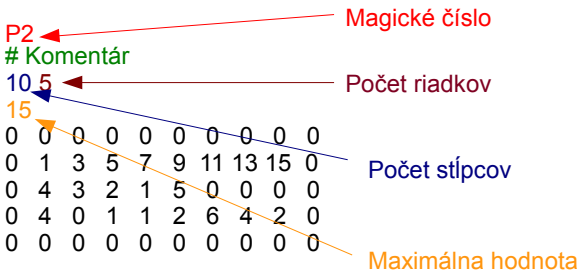
Prvým krokom `make_hist` je získanie rozsahov (intervalov) pre jednotlivé MFCC. Program `make_hist` umožňuje tieto rozsahy buď načítať z XML, vyhľadať v zadanej sade a uložiť do XML. Takže rozsahy možno získať pre trénovaciu sadu a uložiť si ich do XML pre nevidenú sadu alebo rovno získať rozsahy aj pre nevidenú a trénovaciu sadu zároveň. V tejto práci som získanie rozsahov jednotlivých histogramov MFCC vykonával aj pre trénovaciu aj pre nevidenú sadu.

Po získaní rozsahov jednotlivých MFCC program `make_hist` vytvorí pre jednotlivé MFCC príslušné histogramy. Je možné si zvoliť na koľko intervalov budú histogramy rozdelené. V tejto práci som využíval 16 intervalov pre histogram. Histogramy sa robia postupne pre každý vstupný súbor zvlášť a vždy keď sa urobia pre nejaký vstupný súbor, sú hodnoty jednotlivých intervalov histogramov znormované počtom vektorov MFCC obsiahnutom v príslušnom vstupnom súbore. Tieto znormované histogramy MFCC príslušného vstupného vektoru potom program `make_hist` skonkatenuje do výstupného vektoru. Výstupný vektor histogramov MFCC je vytvorený vo formáte LibSVM. Formát LibSVM ako možno vidieť v [6] je nasledovný:

```
label index:hodnota index:hodnota index:hodnota index:hodnota ...
```

, kde `label` je číslo označujúce triedu (žáner). Hodnota `index` začína od hodnoty 1 a postupne sa inkrementuje. K indexom sú intervaly hitogramov priraďované od prvého intervalu histogramu nultého MFCC až po šestnásty interval histogramu n-tého MFCC. Výsledný vektor teda bude obsahovať číslo triedy a 16 (intervalov) * n (počet histogramov MFCC) hodnôt s príslušným indexom.

Tento výstupný vektor histogramov MFCC vo formáte LibSVM je ukladaný do výstupn=eho súboru, kde každý riadok bude reprezentovať jeden výstupný vektor. Každej vstupnej adresárovej štruktúre sa vytvorí jeden súbor obsahujúci vektory príznakov pre všetky nahrávky v tejto adresárovej štruktúre s priradeným príslušným číslom žánru. Okrem toho je aj pre každú adresárovú štruktúru vygenerovaný textový súbor, ktorý obsahuje informáciu o tom aké čísla boli priradené ku ktorým žánrom.



Obrázek 5.6: Formát PGM súboru. Na základe zadaných hodnôt sa vytvorí obrázok v stupni šedi. Magické číslo určuje typ súboru a kódovanie. Komentáre začínajú znakom #. Počet stupňov šedi určuje maximálna hodnota v súbore.

5.4 make_PGM

Využitie programu `make_PGM` je druhou fázou pri extrahovaní príznakov Metódou Obrazovej Spektrálnej analýzy, čo možno vidieť aj na obrázku 5.3. Úlohou programu `make_PGM` je vytvorenie obrázkov (reprezentujúce výsledný spektrogram) z Mel energií spektra extrahovaných z nahrávok. Mel energiám spektra extrahovaných z nahrávok ešte pred ích prevedením na obrázok, `make_PGM` zníži dynamický rozsah Mel-frekvenčných spektier tak, ako som navrhol v sekcií **??**. V tejto sekcií je stručný popis ako tento program funguje, jeho využitie v tejto práci ako aj jeho vstupy a výstupy.

Program `make_PGM` som vytvoril pre účely tejto práce a je na implementovaný v jazyku C++.

Vstupom `make_PGM` je adresárová štruktúra obsahujúca pre každý žáner (žáner je reprezentovaný adresárom) súbory, obsahujúce vektory Mel energií spektra získané krátkodobou analýzou. Jeden vektor Mel energií spektra je na jedom riadku súboru, pričom jednotlivé Mel energie sú od seba oddelené medzerou. Každá nahrávka teda bude popísaná takýmto súborom Mel energií. V tejto práci to bude výstup získaný pomocou skriptu `htk_skript.pl` (sekcia 5.2).

Vstupné súbory Mel energií spektra potom `make_PGM` prevádza na obrázok vo formáte PGM, ktorého formát možno vidieť na obrázku 5.6 a na základe zadaných hodnôt v PGM súbore vznikne príslušný obrázok v stupni šedi. PGM obrázok sa potom uloží do zadaného výstupného adresára. V zadanom vstupnom adresári sa vytvorí obdobná adresárová štruktúra ako je vo vstupnom a na základe pozície súboru obsahujúceho vektory Mel energií spektra sa vo výstupnom súbore na obdobnú pozíciu uloží aj vytvorený obrázok vo formáte PGM.

Pred vytvorením PGM obrázka z Mel-frekvenčných spektier, prevedie `make_PGM` tieto spektrá na Logaritmické Mel-frekvenčné spektrá a z nich vytvorí PGM obrázok. Avšak pri testovaní som prišiel na drobný nedostatok vzorca 4.1, ktorý som navrhol pre prevod Mel energií spektra na Logaritmické Mel energie spektra a to taký, že Mel energie spektra uložené v súbore a získavané v predchádzajúcom kroku skriptom `hkt_skript.pl` majú hodnoty aj v intervale od 0 po 1. A keďže logaritmus čísla od 0 po 1 dáva záporné hodnoty, tak výsledná hodnota ukladaná do PGM by bola tiež záporná. Ako som zistil pozorovaním, tak všetky záporné hodnoty v obrázku sa potom interpretujú čiernou farbou bez rozdielu o akú veľkú zápornú hodnotu ide. Okrem toho výslednú hodnotu treba previesť na celé číslo, čo bolo pri návrhu vo vzorci 4.1 zanedbané. Výsledný vzorec, ktorý `make_PGM` na prevod Mel energií

spektra na výstupné hodnoty obrázka PGM formátu používa, vyzerá nasledovne:

$$x \quad = \quad (\text{integer}) \left(\left(\frac{\log (e+1)}{maxL} \right) * maxV \right) \tag{5.1}$$

, kde ku každej premennej e, čo je jedna Mel energia spektra vstupného súboru, sa pripočíta hodnota 1 čím sa zabráni aby výsledok logaritmu bol záporný. Hodnotu $maxV$, ktorá určuje výstupnú maximálnu vstupnú hodnotu a teda aj v PGM formáte počet stupňov šedi, ktoré bude mať výsledný obrázok, som nastavil na 255. Výsledná hodnota vzorca je nakoniec pretypovaná na integer kvôli odstráneniu desatinných miest.

Kapitola 6

Experimenty

V tejto kapitole sú popísané experimenty, ktoré som so systémom pre rozpoznávanie hudobných štýlov vytvoreným a popísaným v kapitole 5 vykonal. V experimentoch som porovnával hlavne úspešnosť klasifikácie dvoch navrhnutých a vytvorených prístupov pri rozpoznávaní hudobných štýlov a to prístup založený na extrakcií príznakov metódou Histogramov MFCC a prístup založený na extrakcií príznakov metódou Obrazovej Spektrálnej analýzy. Pre porovnanie týchto prístupov som pri experimentoch použil dátová sadu, ktorú som vytvoril práve pre túto prácu. Táto dátová sada je zložená z trénovacej sady a nevidenej sady. Pri experimentoch som tiež použil aj druhú dátovú sadu, nazývanú GTZAN žánrová kolekcia [19], ktorá bola použitá pri experimentoch aj v [20], čím som mohol porovnať nielen prístupy navrhnuté pre túto prácu, ale aj tieto výsledky experimentov, ktoré som dosiahol v tejto práci s výsledkami experimentov z [20]. Obe tieto dátové sady použité pri experimentoch sú bližšie popísané v sekcií 6.1.

Výsledky jednotlivých prístupov extrakcie príznakov s klasifikátorom SVM a príslušnými jadrovými funkciami sú pre dátovú sadu, vytvorenú pre túto prácu, opísané a porovnané v sekcií 6.2. Pre dátovú sadu používanú pri experimentoch v [20], nazývanú GTZAN žánrová kolekcia, sú výsledky jednotlivých prístupov extrakcie príznakov s klasifikátorom SVM a príslušnými jadrovými funkciami opísané v sekcií 6.3. V sekcií 6.3 sú tiež výsledky experimentov, ktoré som dosiahol v tejto práci na GTZAN žánrovej kolekcií, porovnané s výsledkami experimentov z [20].

6.1 Dátové sady

Pri experimentovaní som použil dve dátové sady, ktoré sú opísané práve v tejto sekcií. Prvú z týchto sád som vytvoril priamo pre túto prácu a druhú získal z [19].

Dátová sada, ktorú som pre túto prácu vytvoril sa skladá z dvoch sád a to z trénovacej a nevidenej. Obidve tieto sady, trénovacia aj nevidená, obsahujú nahrávky rozdelené do piatich hudobných štýlov, ktorými sú *Rock*, *Hiphop*, *Country*, *Reggae* a *Klasická (inštrumentálna) hudba*. Sada trénovacích a nevidených nahrávok neobsahuje žiadne rovnaké nahrávky. Trénovacia sada obsahuje pre každý žáner priemerne 1200 tridsať-sekundových nahráviek. Nevidená sada obsahuje pre každý žáner presne 1000 tridsať sekundových nahráviek. Všetky nahrávky, či už trénovacej alebo nevidenej sady, sú vo formáte MP3 so vzorkovaciou frekvenciou 44100 Hz.

Druhou dátovou sadou je dátová sada získaná z [19] a nazývaná GTZAN žánrová kolekcia. GTZAN žánrová kolekcia je použitá pri experimentoch aj v [20]. GTZAN žánrová

kolekcia nie je rozdelená na dve sady trénovaciu a nevidenú, ale je tvorená len jednou sadou. Nahrávky tejto sady sú rozdelené do desiatich žánrov, ktorými sú *Rock*, *Reggae*, *Pop*, *Metal*, *Jazz*, *Hiphop*, *Disco*, *Country*, *Classical* a *Blues*. Každý žáner tejto sady obsahuje presne 100 tridsať-sekundových nahráviek. Všetky nahrávky tejto sady sú vo formáte AU a majú vzorkovacou frekvenciu 22050 Hz.

6.2 Výsledky pre dátovú sadu vytvorenú pre túto prácu

V tejto sekcií sú opísané a porovnané výsledky experimentov pre dva prístupy použité pri rozpoznávaní hudobných štýlov tvoriace systém, ktorý som pre rozpoznávanie hudobných štýlov vytvoril. Experimenty opísané v tejto sekcií som vykonal na dátovej sade vytvorenej pre túto prácu, ktorá sa skladá z trénovacej a nevidenej sady (bližší popis v sekcií 6.1). Trénovaciu sadu som použil pre natrénovanie klasifikátora a nevidenú na otestovanie.

Prvým testovaným prístupom bol prístup založený na extrakcií príznakov metódou Histograov Mel-frekvenčných keprálnych koeficientov. S touto metódou extrakcie príznakov som použil klasifikátor SVM a boli vyskúšané dve jadrové funkcie, ktorými boli lineárna jadrová funkcia a RBF jadrová funkcia. Pri trénovaní som použil cross-validáciu so šiestimi validáciami a taktiež som použil prostriedky pre optimalizáciu `C` parametru SVM a parametru γ pri použití RBF jadrovej funkcie. Prvým cieľom experimentov s metódou Histogramov Mel-frekvenčných keptrálnych koeficientov bolo zistiť optimálny počet Mel-frekvenčných kepstrálnych koeficientov extrahovaných z nahrávok pre rozpoznávanie hudobných štýlov. Pre nájdenie optimálneho počtu MFCC bolo vyskúšaných 5 MFCC s nultým MFCC, 11 MFCC s nultým MFCC a 22 MFCC s‘nultým MFCC. Pre tento účel som používal okno o veľkosti 25 ms s posunom 10 ms a ako možno vidieť v tabuľke 6.1, tak najlepšia úspešnosť klasifikácie na nevidenej sade bola dosiahnutá pre 22 MFCC s nultým MFCC a SVM s RBF jadrovou funkciou (táto úspešnosť klasifikácie bola $79,2\,\%$). V tabuľke 6.1 je tiež pre porovnanie ukázané, že pri použití rovnakého počtu MFCC, ale pri použití SVM s lineárnou jadrovou funkciou je dosiahnutá úspešnosť klasifikácie horšia (72,94 %). Výsledky teda ukazujú, že z dvojice jadrových funkcií SVM dosiahla najlepšie výsledky RBF jadrová funkcia. Ďalším cieľom experimentov bolo určiť najlepšiu veľkosť okna a pre tento účel som vyskúšal okná o veľkosti 25 ms, 50 ms a 100 ms, pričom experimenty som vykonával teda pre 22 MFCC s nultým MFCC a SVM s RBF jadrovou funkciou. Najlepšia dosiahnutá úspešnosť klasifikácie na nevidenej, ako možno vidieť v tabuľke 6.1, bola 79,42 % a to pre okno veľkosti 50 ms. Najlepšia úspešnosť klasifikácie pri prístupe založenom na extrakcií príznakov metódou Histogramov MFCC, bola teda úspešnosť klasifikácie $79,42\,\%$ dosiahnutá pre 22 MFCC s nultým MFCC, 50 ms okno a SVM klasifikátor s RBF jadrovou funkciou. Pre tento najlepší dosiahnutý výsledok reprezentuje tabulka 6.2 príslušnú maticu zámien (confusion matrix).

Druhým testovaným prístupom bol prístup založený na extrakcií príznakov metódou Obrazovej Spektrálnej analýzy, s ktorou som použil ako klasifikátor SVM s chi-kvadrát jadrovou funkciou. Ako už bolo spomenuté v časti 5.2.2 pri extrakcií príznakov metódou Obrazovej Spektrálnej analýzy som použil 128 extrahovaných Mel energií spektra, vrchný frekvenčný prah 10000 Hz, veľkosť okna 100 ms a posun okna 25 ms. Experimenty som zameral na fázu metódy Obrazovej Spektrálnej analýzy, pri ktorej sa extrahujú príznaky z obrázka pomocou lokálnych oblastí popísanými deskriptormi. Experimentoval som s veľkosťami lokálnych oblastí a to pre polomery 8x8, 16x16 a 32x32 pixelov. Pre rozostupy medzi oblasťami na obrázku bola použitá homogénna mriežka s rozostupmi 8x8 pixelov a použil som slovník s veľkosťou 4096 slov. Najlepšie výsledky experimentov pre rôzne veľkosti lo-

Metóda Histogramov MFCC		SVM		Úspešnosť klasifikácie
Počet MFCC	**Veľkosť okna**	**Jadrová funkcia**	**Parametre SVM**	
5 + 0-tý MFCC	25 ms	RBF	$C = 46,416$, $\gamma = 0.999$	69,7 %
11 + 0-tý MFCC	25 ms	RBF	$C = 255$, $\gamma = 2.154$	74,26 %
22 + 0-tý MFCC	25 ms	RBF	$C = 5011$, $\gamma = 1$	79,2 %
22 + 0-tý MFCC	25 ms	lineárna	$C = 2,05$	72,94 %
22 + 0-tý MFCC	50 ms	RBF	$C = 25$, $\gamma = 0,999$	**79,42 %**
22 + 0-tý MFCC	100 ms	RBF	$C = 255$, $\gamma = 0,999$	79,22 %

Tabulka 6.1: Výsledky experimentov pre prístup založený na extrakcií príznakov metódou Histogramov MFCC. Vykonané boli s posunom okna 10 ms a najlepšia úspešnosť klasifikácie na nevidenej sade bola 79,42 % pre 22 MFCC s nultým MFCC, 50 ms oknom a použitým SVM s RBF jadrovou funkciou ($C = 25$, $\gamma = 0,999$).

	Rock	**Klasická hudba**	**Hiphop**	**Country**	**Reggae**
Rock	**680**	9	108	182	65
Klasická hudba	15	**988**	21	22	6
Hiphop	37	2	**765**	52	96
Country	248	1	23	**737**	32
Reggae	20	0	83	7	**801**

Tabulka 6.2: Matica zámien (confusion matrix), ktorá reprezentuje najlepšiu dosiahnutú úspešnosť rozpoznávania hudobných štýlov (79, 42 %) pomocou extrakcie príznakov metódou Histogramov Mel-frekvenčných kepstrálnych koeficientov a klasifikátorom SVM s RBF jadrovou funkciou ($C = 25$, $\gamma = 0,999$). Riadky v tejto matici reprezentujú názov žánru a stĺpce určujú kam boli jednotlivé nahrávky pre daný žáner naozaj zaradené. Tento výsledok bol dosiahnutý na dátovej sade, ktorú som pre túto prácu vytvoril (na nevidenej sade). Pri extrakcií príznakov metódou Histogramov MFCC bolo použitých 22 MFCC s nultým MFCC a oknom veľkým 50 ms s posunom 10 ms.

kálnych oblastí, som pri rozpoznávaní hudobných štýlov, ako možno vidieť v tabuľke 6.3, dosiahol pre veľkosť lokálnych oblastí 16x16 pixela, kde úspešnosť klasifikácie na nevidenej sade bola 91, 86 %.

Lepšie výsledky experimentov pri rozpoznávaní hudobných štýlov, na dátovej sade vytvorenej pre túto prácu, som dosiahol pre prístup založený na extrakcií príznakov metódou Obrazovej spektrálnej analýzy využívajúci klasifikátor SVM s chi-kvadrát jadrovou funkciou, kde tento prístup dosiahol najlepšiu úspešnosť klasifikácie na nevidenej sade 91,86 %, čo je o viac ako 12 % väčšia úspešnosť klasifikácie, ako bola dosiahnutá pre prístup založený na extrakcií príznakov metódou Histogramov Mel-frekvenčných kepstrálnych koeficientov využívajúci klasifikátor SVM s RBF jadrovou funkciou, pre ktorý bola táto úspešnosť klasifikácie 79,42 %.

Veľkosť lokálnych oblastí	Úspešnosť klasifikácie
8x8 pixelov	89,81 %
16x16 pixelov	**91,86 %**
32x32 pixelov	89,8 %

Tabulka 6.3: Výsledky experimentov pre prístup založený na extrakcií príznakov metódou Obrazovej Spektrálnej analýzy na dátovej sade, ktorú som pre túto prácu vytvoril (konkrétne na nevidenej sade). Bolo pre ne použitých 128 Mel energií, vrchná frekvenčná hranica 12000 Hz, veľkosť okna 100 ms a posunom 25 ms.

6.3 Výsledky pre GTZAN žánrovú kolekciu

Experimenty popísané v tejto sekcií som vykonal na dátovej sade získanej z [19], nazývanej GTAZAN žánrová kolekcia (bližší opis tejto sady je v sekcií 6.1). V tejto sekcií sú opísané a porovnané výsledky experimentov pri rozpoznávaní hudobných štýlov dvoch prístupov tvoriacich systém, ktorý som navrhol a vytvoril pre rozpoznávanie hudobných štýlov pre účely tejto práce. Okrem toho v tejto sekcií porovnávam mnou dosiahnuté výsledky na GTZAN žánrovej kolekcií s výsledkami, ktoré na tejto dátovej sade dosiahli v [20]. Vzhľadom na to, že som experimenty na GTZAN žánrovej kolekcií vykonával až pri dokončovaní tejto práce, z dôvodu nedostatku času neboli experimenty komplexnejšie a riadil som sa výsledkami získanými a opísanými v sekcií 6.2.

Experimenty v [20] sa vykonávali tak, že sa použila cross-validácia s 10-timi validáciami, a teda GTZAN žánrová kolekcia sa rozdelila rovnomerne pre každý žáner na 10 častí, pričom 9 častí sa vždy použilo na trénovanie a jedna časť na testovanie a to tak, že sa postupne tieto časti pri testovaní a trénovaní obmieňali. Tento experiment v [20] opakovali 100-krát a výsledok vznikol spriemerovaním priebežných výsledkov pre jednotlivé iterácie, pričom najlepšie tieto spriemerované výsledky dosahovali v [20] úspešnosť klasifikácie 61 %.

Pri experimentovaní s GTAZAN žánrovou kolekciou v tejto práci som použil rovnaký experiment ako sa použil v [20], a teda bola použitá cross-validácia s 10-timi validáciami, ktorá GTZAN žánrovú kolekciu rozdelila rovnomerne pre každý štýl rozdelila na 10 častí, pričom sa vždy 9 častí použilo na trénovanie, jedna na testovanie a postupne sa tieto časti na testovanie a trénovanie obmieňali.

Prvé experimenty boli vykonané na prístupe založenom na extrakcií príznakov metódou Histogramov MFCC. S touto metódou extrakcie príznakov som ako klasifikátor použil SVM s RBF funkciou, ktorá dosiahla pri experimentoch prevádzaných na dátovej sade, ktorú som pre túto prácu vytvoril, lepšie výsledky ako lineárna jadrová funkcia (viz. sekcia 6.2). Použil som tiež pri trénovaní ďalšiu vrstvu cross-validácie so šiestimi validáciami a prostriedky pre optimalizáciu parametru `C` a patametru γ, ktoré RapidMiner poskytuje. Pri extrakcií príznakov metódou Histogramov MFCC som pri experimentoch použilo 22 MFCC s nultým MFCC, pretože práve pre tento počet som pri experimentoch opísaných v 6.2 dosiahol najlepšie výsledky. Vyskúšané boli extrakcie MFCC pre okná o veľkosti 25 ms a 50 ms, pričom pre obe varianty okien som použil posun okna 10 ms. Ako bolo napísané vyššie, tak experiment z [20] bol prevádzaný 100-krát a výsledok sa spriemeroval. Z časových dôvodov som vyskúšal len 10-násobnú iteráciu experimentu a to pre najlepšiu variantu veľkosti okna pre jednu iteráciu. Pre jednu iteráciu som najlepšiu úspešnosť klasifikácie dosiahol pre okno o veľkosti 25 ms a táto úspešnosť klasifikácie bola 74 % (matica zámien pre túto úspešnosť je reprezentovaná tabuľkou 6.4). pre okno o veľkosti 50 ms táto úspešnosť

	Blues	Metal	Classical	Rock	Disco	Hiphop	Country	Jazz	Reaggae	Pop
Blues	**87**	4	0	7	2	1	4	4	6	0
Metal	3	**88**	0	2	1	4	1	1	0	0
Classical	0	0	**92**	0	1	0	0	5	0	1
Rock	2	4	1	**57**	12	3	11	3	3	2
Disco	1	1	0	16	**60**	7	6	1	2	4
Hiphop	2	1	0	3	9	**63**	0	0	11	5
Country	4	2	1	9	6	1	**68**	1	6	4
Jazz	1	0	6	4	0	1	4	**81**	2	3
Reggae	0	0	0	2	2	17	3	3	**67**	4
Pop	0	0	0	0	7	3	3	1	3	**77**

Tabulka 6.4: Matica zámien (confusion matrix), ktorá reprezentuje najlepšiu úspešnosť rozpoznávania hudobných štýlov (74 %) pre extrakciu príznakov metódou Histogramov Mel-frekvenčných kepstrálnych koeficientov a klasifikátor SVM s RBF jadrovou funkciou, na GTZAN žánrovej kolekcií. Riadky v tejto matici reprezentujú názov žánru a stĺpce určujú kam boli jednotlivé nahrávky pre daný žáner naozaj zaradené. Pri extrakcií príznakov metódou Histogramov MFCC bolo použitých 22 MFCC s nultým MFCC a oknom veľkým 25 ms s posunom 10 ms.

klasifikácie bola 72, 8 %. Experiment pre desať iterácií som teda vykonal pre okno o veľkosti 25 ms. Výsledná úspešnosť klasifikácie pre 10 iterácií experimentu využívajúceho cross-validáciu o desiatich validáciach pre rozdelenie sady na desať častí a postupné kombinovanie pri trénovaní a testovaní, dosiahla 73, 88 %, teda porovnateľnú pre jednu iteráciu, kde bola úspešnosť klasifikácie 74 %.

Experimenty na GTZAN žánrovej kolekcií som vykonal aj pre prístup založený na extrakcií príznakov metódou Obrazovej Spektrálnej analýzy. Pre túto metódu extrakcie som použil tento-krát klasifikátor SVM s RBF jadrovou funkciou, podobne ako pri prístupe založenom na extrakcií príznakov metódou Histogramov MFCC. Ako už bolo spomenuté v podsekcií 5.2.2, pri extrakcií príznakov metódou Obrazovej Spektrálnej analýzy som použil 128 extrahovaných Mel energií spektra, veľkosť okna 100 ms, posun okna 25 ms, ale na rozdiel od experimentov opísaných v 6.2, som nepoužil žiaden vrchný frekvenčný prah. Vrchný frekvenčný prah som nepoužil preto, lebo výsledné obrázky v častiach obrázkoch patriacim vyšším frekvenciám neobsahovali žiadny súvislý pás ako bolo opísané v podesekcií 5.2.2. Tento pás sa pravdepodobne neobjavoval preto, lebo nahrávky neobsahovali frekvencie vyššie ako 12000 Hz, kde práve tento súvislý pás vznikal. Experimenty som pre tento prístup, podobne ako v sekcií 6.2, zameral na extrakciu príznakov z obrázka a to konkrétne na veľkosti lokálnych oblastí, z ktorých sa deskriptory potom extrahovali. Boli vyskúšané lokálne oblasti o polomere 8x8, 16x16 a 32x32 pixela. Pre rozostupy medzi oblasťami na obrázku sa použila homogénna mriežka s rozostupmi 8x8 pixelov a použil sa slovník s veľkosťou 4096 slov. Pri prístupe založenom na extrakcií príznakov metódou Obrazovej Spektrálnej analýzy som experiment z časových dôvodov nevykonával viac-krát ako tomu bolo v [20], a tak výsledky pre tento prístup boli vykonané pre jeden experiment (tento experiment je popísaný vyššie). Ako možno vidieť v tabuľke 6.6, tak najlepšie výsledky pre rôzne veľkosti lokálnych oblastí boli dosiahnuté pre lokálne oblasti o veľkosti 32x32 pixelov a úspešnosť klasifikácie pre túto veľkosť bola 86, 4 %. Práve pre túto úspešnosť klasifikácie 86, 4 % reprezentuje

	Blues	**Metal**	**Classical**	**Rock**	**Disco**	**Hiphop**	**Country**	**Jazz**	**Reaggae**	**Pop**
Blues	**92**	0	1	2	0	2	2	5	2	2
Metal	0	**98**	1	0	0	2	0	0	0	0
Classical	2	0	**83**	3	0	3	0	6	2	10
Rock	4	1	3	**88**	3	1	1	6	3	2
Disco	0	0	0	1	**86**	0	1	3	3	1
Hiphop	0	1	2	0	0	**91**	0	0	1	2
Country	0	0	0	1	2	0	**94**	0	1	3
Jazz	0	0	0	1	3	0	0	**77**	3	3
Reggae	2	0	2	1	3	0	0	2	**83**	5
Pop	0	0	8	3	3	1	2	1	2	**72**

Tabulka 6.5: Matica zámien (confusion matrix), ktorá reprezentuje najlepšiu úspešnosť rozpoznávania hudobných štýlov (86, 4 %) pre prístup založený na extrakcií príznakov metódou Obrazovej Spektrálnej analýzy a klasifikátor SVM s RBF jadrovou funkciou, na GTZAN žánrovej kolekcií. Riadky v tejto matici reprezentujú názov žánru a stĺpce určujú kam boli jednotlivé nahrávky pre daný žáner naozaj zaradené. Pri extrakcií príznakov metódou Obrazovej Spektrálnej analýzy som použili pri extrakcií príznakov z obrázka lokálne oblasti s polomerom 32x32 pixelov.

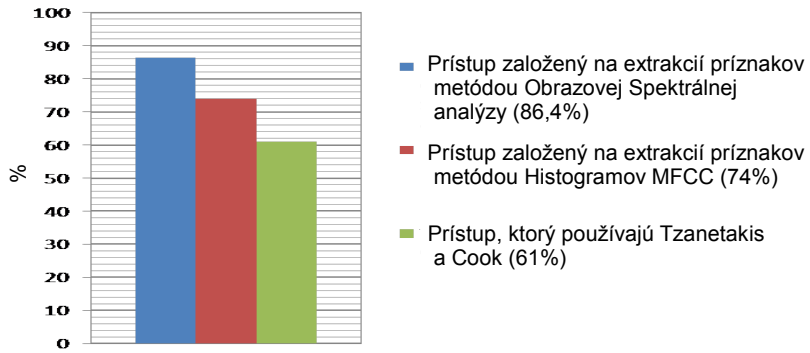
Veľkosť lokálnych oblastí	**Úspešnosť klasifikácie**
8x8 pixelov	83,4 %
16x16 pixelov	84,2 %
32x32 pixelov	**86,4 %**

Tabulka 6.6: Výsledky experimentov pre prístup založený na extrakcií príznakov metódou Obrazovej Spektrálnej analýzy na GTZAN žánrovej kolekcií. Bolo pre ne použitých 128 Mel energií, veľkosť okna 100 ms a posunom 25 ms.

tabuľka 6.5 maticu zámien (confusion matrix).

Keďže z časových dôvodov som nevykonal požadovaný počet experimentov ako tomu bolo v [20] a pre 10 iterácií, pri prístupe založenom na extrakcií príznakov metódou Histogramov MFCC, sa výsledok len veľmi mierne zhoršil, budem pri porovnávaní výsledkov predpokladať, že pre 100 iterácií by boli jednotlivé výsledky približne rovnaké.

Dosiahnuté výsledky experimentov ukazujú, že prístupy, s ktorými som experimentoval na GTZAN žánrovej sade v tejto práci, majú lepšie úspešnosti klasifikácie než výsledky experimentov dosiahnuté v [20]. Porovnanie týchto výsledkov reprezentuje graf na obrázku 6.1. Najlepšie výsledky dosiahol prístup založený na extrakcií príznakov metódou Obrazovej Spektrálnej analýzy, ktorý dosiahol na GTZAN žánrovej kolekcií najlepšiu úspešnosť klasifikácie 86, 4 %. Prístup založený na extrakcií príznakov metódou Histogramov MFCC dosiahol úspešnosť klasifikácie 74 % a najlepšie dosiahnuté výsledky experimentov z [20] dosiahli úspešnosť klasifikácie 61 %. Výsledky experimentov teda ukazujú, že najlepšia dosiahnutá úspešnosť klasifikácie, ktorú som v tejto práci dosiahol je od najlepšej úspešnosti klasifikácie dosiahnutej v [20] o približne 25 % lepšia.



Obrázek 6.1: Graf, ktorý reprezentuje najlepšie dosiahnuté úspešnosti pre jednotlivé prístupy, ktoré som v tejto práci na GTZAN žánrovej kolekcií dosiahol, v porovnaní s najlepšou úspešnosťou klasifikácie, ktorú Tzanetakis a Cook na GTZAN žánrovej kolekcií v [20] dosiahli.

Kapitola 7

Záver

Cieľom tejto práce bolo na základe naštudovaných existujúcich prístupov pri rozpoznávaní hudobných štýlov vytvoriť systém pre rozpoznávanie hudobných štýlov a nakoniec tento systém otestovať na vhodnej dátovej sade. Tento systém na rozpoznávanie hudobných štýlov mal obsahovať vhodnú metódu extrakcie príznakov a klasifikačnú metódu, ktorá pri úlohe rozpoznávania hudobných štýlov dosahuje dobré výsledky.

Výsledkom mojej práce je systém pre rozpoznávanie hudobných štýlov skladajúci sa z dvoch prístupov pri extrakcií príznakov a využívajúci klasifikátor Support Vector Machine s jadrovými funkciami RBF a chi-kvadrát. Prvým prístupom extrakcie príznakov je metóda Histogramov Mel-frekvenčných kepstrálnych koeficientov, ktorá je založená na extrakcií MFCC prostredníctvom krátkodobej analýzy z nahrávok. Druhým prístupom extrakcie príznakov je metóda Obrazovej Spektrálnej analýzy, ktorá je založená na prevádzaní signálu z 1D (zvuku) na 2D (spektrogram) a extrahovaní príznakov z tohto spektrogramu pomocou postupov, ktoré sa využívajú pri extrakcií príznakov z obrázka.

Obidva tieto prístupy extrakcie príznakov a SVM s príslušnou jadrovou funciou som otestoval na dvoch dátových sadách. Prvou dátovou sadou bola sada, ktorú som pre účely tejto práce vytvoril, ktorá obsahuje 30-sekundové nahrávky piatich žánrov a je rozdelená na trénovaciu a nevidenú sadu. Avšak kvôli autorským právam použitých nahráviek nie je ďalej táto sada poskytnutá. Druhou sadou, ktorú som pri testovaní použil je GTZAN žánrová kolekcia získaná z [19] a použitá pri experimentoch v [20]. GTZAN žánrová kolekcia obsahuje 30-sekundové nahrávky rozdelené do desiatich žánrov.

Výsledky experimentov na obidvoch dátových sadách ukázali, že prístup extrakcie príznakov metódou Obrazovej Spektrálnej analýzy dosahuje lepšie výsledky než prístup extrakcie príznakov metódou Histogramov Mel-frekvečných kepstrálnych koeficientov. Výsledky experimentov na GTZAN žánrovej kolekcií ďalej ukázali, že obidva prístupy, ktoré som v tejto práci vytvoril, dosahujú lepšie výsledky na tejto dátovej sade ako boli na rovnakej dátovej sade dosiahnuté v [20]. V [20] tiež uvádzajú, že bol vykonaný výskum, za pomoci študentov vysokých škôl, cieľom ktorého bolo zistiť akú úspešnosť pri rozpoznávaní hudobných štýlov má človek. Študenti mali zaradiť poskytnuté nahrávky do desiatich žánrov a túto úlohu dokázali splniť správne na 70 %. Keďže obidva prístupy, ktoré som vytvoril v tejto práci dosahujú lepšiu úspešnosť ako 70 %, pri rovnakom počte žánrov a pri podobnom zložení týchto žánrov, možno tieto výsledky považovať za uspokojivé. Pri prístupe Obrazovej Spektrálnej analýzy a SVM s RGB jadrovou funkciou dosiahla úspešnosť klasifikácie na GTZAN žánrovej kolekcií až $86,4$ %.

Experimenty, ktoré som v tejto práci prevádzal, však z časových dôvodov neboli dostatočne komplexné, či už pri GTZAN dátovej sade, kde nebolo vykonávané rovnaké množstvo

experimentov ako v [20], ale aj čo sa týka hľadania optimálnych parametrov, či už pri parametroch SVM alebo pri parametroch obidvoch vytvorených prístupov extrakcie príznakov. Tieto experimenty by bolo vhodné v budúcnosti dokončiť aby zodpovedali experimentom vykonaným v [20] a aby boli nájdené najoptimálnejšie parametre, či už pri parametroch klasifikátora SVM s príslušným jadrom alebo pri parametroch vytvorených prístupoch extrakcie príznakov, tak aby úspešnosť klasifikácie bola čo najvyššia.

V ďalšom pokračovaní tejto práce okrem komplexnejších experimentov, by bolo vhodné pre jednotlivé prístupy extrakcie príznakov ako aj pre celý systém rozpoznávania vytvoriť uživatelské rozhranie, čo by značne zjednodušilo prácu s týmto systémom.

Ďalej by bolo vhodné pokúsiť sa vylepšiť prístupy extrakcie príznakov vytvorené v tejto práci a hlavne zamerať sa na prístup extrakcie príznakov metódou Obrazovej Spektrálnej analýzy, ktorá dosahovala najlepšie výsledky v tejto práci. Vylepšením týchto prístupov extrakcie príznakov môže byť napríklad doplnenie o extrahované príznaky rytmu, podobne ako je to vyskúšané v [12], kde tento prístup vylepšil dosiahnuté výsledky.

Literatura

[1] RapidMiner [online]. 2011-04-20 [cit. 2011-04-28].
URL http://en.wikipedia.org/wiki/RapidMiner

[2] SVM - Support Vector Machines [online]. [cit. 2011-04-20].
URL http://www.dtreg.com/svm.htm

[3] RapidMiner [online]. [cit. 2011-04-28].
URL http://rapid-i.com/content/view/181/190/lang,en/

[4] HTK – What is HTK? [online]. [cit. 2011-04-29].
URL http://htk.eng.cam.ac.uk/

[5] Barbedo, J. G. A.; Lopes, A.: Automatic genre classification of musical signals. *EURASIP J. Appl. Signal Process.*, ročník 2007, January 2007: s. 157–157, ISSN 1110-8657.
URL http://dx.doi.org/10.1155/2007/64960

[6] Chang, C.; Lin, C.: LIBSVM FAQ [online]. 2011-04-21 [cit. 2011-05-01].
URL http://www.csie.ntu.edu.tw/~cjlin/libsvm/faq.html

[7] Cristianini, N.; Shawe-Taylor, J.: *Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000, ISBN 0521780195.

[8] Deutscher, M.: Rozpoznávač hudobního stylu z mp3. diplomová práce, FIT VUT v Brně, 2009.

[9] Hastie, T.; Rosset, S.; Tibshirani, R.; aj.: The Entire Regularization Path for the Support Vector Machine. *Journal of Machine Learning Research*, ročník 5: s. 1391–1415.
URL http://academic.research.microsoft.com/Publication/1243429/the-entire-regularization-path-for-the-support-vector-machine

[10] Hradiš, M.; Beran, V.; Řezníček, I.; aj.: Brno University of Technology at TRECVid 2010. In *TRECVID 2010: Participant Notebook Papers and Slides*, National Institute of Standards and Technology, 2010, str. 11.
URL http://www.fit.vutbr.cz/research/view_pub.php?id=9444

[11] Lazebnik, S.; Schmid, C.; Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, Washington, DC, USA: IEEE Computer Society, 2006, ISBN 0-7695-2597-0, s. 2169–2178.
URL http://dx.doi.org/10.1109/CVPR.2006.68

[12] Lidy, T.; C. N. Silla, J.; Cornelis, O.; aj.: On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-Western and ethnic music collections. *Signal Process.*, ročník 90, April 2010: s. 1032–1048, ISSN 0165-1684.
URL http://dx.doi.org/10.1016/j.sigpro.2009.09.014

[13] Lidy, T.; Mayer, R.; Rauber, A.; aj.: A Cartesian Ensemble of Feature Subspace Classifiers for Music Categorization. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, editace J. S. Downie; R. C. Veltkamp, International Society for Music Information Retrieval, Utrecht, Netherlands: International Society for Music Information Retrieval, August 2010, ISBN 978-90-393-53813, s. 279–284.

[14] Lidy, T.; Rauber, A.; Pertusa, A.; aj.: MIREX 2008: Audio Music Classification Using A Combination Of Spectral, Timbral, Rhythmic, Temporal And Symbolic Features. In *MIREX 2008 - Music Information Retrieval Evaluation eXchange, MIREX Genre Classification Contest.*, Philadelphia, Pennsylvania, USA, 2008.

[15] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, ročník 60, November 2004: s. 91–110, ISSN 0920-5691.
URL http://portal.acm.org/citation.cfm?id=993451.996342

[16] Schuller, B.; Eyben, F.; Rigoll, G.: Tango or Waltz?: putting ballroom dance style into tempo detection. *EURASIP J. Audio Speech Music Process.*, ročník 2008, January 2008: s. 6:1–6:12, ISSN 1687-4714.
URL http://dx.doi.org/10.1155/2008/846135

[17] Snoek, C. G. M.; Rooij, O. D.; Huurnink, B.; aj.: The MediaMill TRECVID 2009 Semantic Video Search Engine. In *Proceedings of the TRECVID Workshop*, 2009.
URL
http://www.science.uva.nl/research/publications/2009/SnoekPTRECVID2009

[18] Thomason, L.; Berquin, Y.; Ellerton, A.: Tinyxml [online]. [cit. 2011-05-01].
URL http://www.grinninglizard.com/tinyxml/

[19] Tzanetakis, G.: Data Sets [online]. [cit. 2011-04-20].
URL http://marsyas.info/download/data_sets

[20] Tzanetakis, G.; Cook, P. R.: Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, ročník 10, č. 5, 2002: s. 293–302.
URL http://dx.doi.org/10.1109/TSA.2002.800560

[21] Viikki, O.; Laurila, K.: Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Commun.*, ročník 25, August 1998: s. 133–147, ISSN 0167-6393.
URL http://dx.doi.org/10.1016/S0167-6393(98)00033-8

[22] Young, S.; Evermann, G.; Gales, M.; aj.: The HTK Book (for HTK Version 3.4). 2006.
URL http://htk.eng.cam.ac.uk/docs/docs.shtml

[23] Černocký, H.: Zpracování řečových signálů – studijní opora. Technická zpráva, FIT VUT v Brně, 2006.