スペクトログラム画像を用いた楽曲印象分類による時間及び周波数情報と印象の関係分析手法の提案

宮谷大輝[†1]　　中山英樹[†1]

近年インターネットやデバイスの発達により音楽の視聴状況が多様化し，膨大な曲から状況に合った曲を選択することが困難となっている．そこで楽曲の印象による分類の需要が増しているが，ジャンル識別等の他の分類問題とは異なり印象は音楽特徴との関係が明確でないことが知られている．そこで生の音源情報から楽曲の印象を推定し，印象に強く影響を与える要素を抽出する手法を提案する．まず特徴量として音源情報から生成したスペクトログラム画像の画像特徴量を用いる．さらに一定領域の周波数帯ごとに特徴量を抽出しフィッシャーの重みマップを分析することで，各周波数帯の印象判別への影響度合いを明らかにする．また楽曲を一定時間で区切り，その時間での特徴量をまとめて一つの特徴量として表現することで，事後確率から全体の印象への影響が大きい時間帯を分析する．本手法で得られた分析結果については主観評価実験により有効性の検証を行う．

A method for finding the relationship between music mood and the information of time and frequency by music mood classification using spectrogram images

HIROKI MIYATANI[†1]　HIDEKI NAKAYAMA[†1]

Recently, with the development of the Internet and various music devices, it has become harder to choose appropriate music suited to particular scenes from enormous amount of songs because of diverse listening cases. Therefore, demand of music mood classification has been increasing. However, being different from other music classification tasks, mood classification is known to have less-obvious relation to musical features. Thus, in this research we estimate music mood with raw audio information, and propose method to extract influential elements in music impression. Firstly, we use visual features of spectrogram image generated audio information. Additionally, with analysis using the Fisher weight map of features extracted in each frequency, it becomes apparent how much the features effect the mood discrimination. Furthermore, dividing a music into some parts and expression one feature integrated each feature for a length of time, we identify the time period which effects powerfully on whole music mood from each probability. Finally, we verify the validity of analysis with our method through a subjective evaluation experiment.

1. 社会的背景

近年，インターネットにより世界中の音楽を手軽に検索できるようになったり，音楽作成ツールや動画アップロードサービスにより個人が作成した音楽の公開も容易になるなど，視聴可能な楽曲数が膨大に増えている．しかし楽曲は音源そのものが価値の有る商品であるため，検索可能なメタ情報がタイトル，アーティスト，ジャンル，レビューなど少ないことが問題として挙げられる．そのため現在のメタ情報では絞り切れないさらに高度な楽曲を推薦したり，個人が作った曲などでメタ情報が欠落しているものを自動的に補填するなど，音楽推薦や音楽情報検索（MIR：Music Information Retrieval）技術の需要が増加している．

さらに最近ではクラウドサービスやモバイルデバイスの発達により，大量の楽曲をどのような場所にも持ち出せるようになった．そのため音楽の鑑賞の形態も多様化し，泳ぎながら音楽を聴いたり，走るリズムを維持するために音楽を聞くなど，今まででは想定されていないシチュエーションに対する音楽推薦も必要となっている．そこで，楽曲そのものの印象・雰囲気による音楽の推薦や検索の需要が今後も増していくと考えられる．

2. 研究目的

楽曲の印象は文化や個人によっても異なることがある曖昧なものであり，ただ自動的に分類や推薦を行うだけでは納得できない場合も多い．このような曖昧な識別においては，識別した結果のみならず識別理由を抽出することが重要であると考えられる．そこでパターン認識技術を用いた自動楽曲印象分類手法を通じて，楽曲の印象を与える要因を抽出する手法を作成することを本研究の目的とする．本研究では個人で作成した音楽などあらゆる楽曲に適用できることを念頭に置いているため，MIDI データやメロディ・テンポ情報などが欠落している場合などを考慮し，音源情報のみを楽曲の情報として用いる．

楽曲に印象を与える音楽特性を得ることによって，個人や国，年代ごとの嗜好の偏りの要因が抽出できたり，視聴状況に応じて曲の雰囲気を操作して提供できるなどの応用が考えられる．

†1 東京大学
The University of Tokyo

3. 楽曲印象分類について

パターン認識技術を用いた楽曲印象分類の概要は図 1 のようになる．分類器を作成する過程において，音源情報を機械に入力できる形に変換する必要があり，この変換されたものを特徴量という．また一方で印象は人間によってでしか判断ができないため，曲ごとに専門家や一般大衆によって印象ラベル情報が付加されている必要がある．図 1 の Training のようにこれらの特徴量と印象ラベルのセットを大量の楽曲から抽出してパターン認識技術により学習を行うことで分類器を作成する．実際に印象が未知のものに対して分類を行う際は，図 1 の Test のように音源データから先程と同様の方法で特徴量を抽出し，それを分類器に入れることによって印象ラベルが推定できる．一般的には Test で用いる音源データにも印象ラベルが既知のものを用い，分類器で推定したラベルとの比較を行うことで精度を測ったり分類器のパラメータの調整を行う．

本研究では自動楽曲分類で用いる特徴量は音源情報から抽出できるものを考える．音源情報は瞬間の音圧である音波によって観測されるが，そのままの値は扱いにくい．多くの場合は波形の特徴を表現するのに短時間フーリエ変換（STFT）により周波数ごとの振幅の大きさとして変換することで様々な特徴量（Flux, Centroid, Rolloff など）を抽出するのが一般的である．さらに変換された値を変換することで特徴量を得るものもある（MFCCs など）．

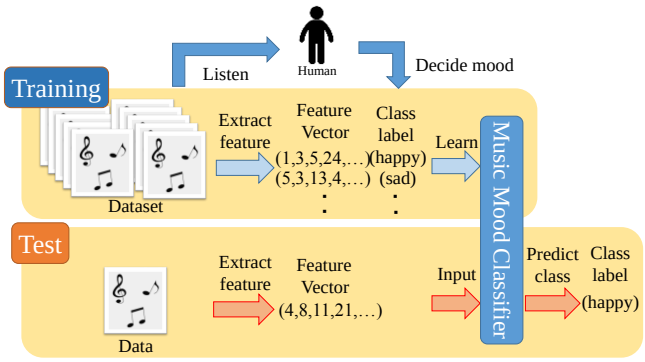

図 1　楽曲印象分類の概要

Figure 1　The outline of music mood classification.

4. 印象に寄与する音楽特性の抽出手法の提案

4.1 要素の選定

音の三要素として音の大きさ，音の高さ，音色があると言われている．音色は印象に影響があり，様々な因子があることが知られている．例えば北村らによる研究[1]では美的因子・金属性因子・迫力因子の 3 つが提示されており，高い周波数で音圧レベルが大きいと鋭い印象になるなどの傾向を示している．ここから音の印象はある範囲の周波数の振幅と関係があると考えられる．そこで本研究では音色が楽曲の印象に関係していると考え，周波数帯の中で特に印象に大きく寄与している部分の抽出を目指す．また，曲調の変化など楽曲内においても印象が変化すると考えられるため，特に印象度合いが強い時間帯の抽出についても検討を行う．

4.2 抽出方法

各特徴量の分類への寄与度合いを抽出する研究としてフィッシャー重みマップが篠原らによって提案されている[2]．これはフィッシャー判別分析の手法を行列に適用させたもので，クラス判別に有効になるようなクラス間分散 $tr\tilde{\sum}_{B}$ が最大，クラス内分散 $tr\tilde{\sum}_{W}$ が最小になるような重みを出力することができる．具体的には式（1）の $J(\omega)$ が最大になるような重み ω が出力され，篠原らの研究においてはこの ω を用いて次元圧縮を行って顔の表情認識を行っている．図 2 のフィッシャー重みマップの黒と白で塗られている部分が顔の表情を識別する上で重要な部分であり，灰色の部分は識別にあまり重要でないことを示している．

$$J(\omega) = tr\tilde{\sum}_{B} \Big/ tr\tilde{\sum}_{W} \qquad (1)$$

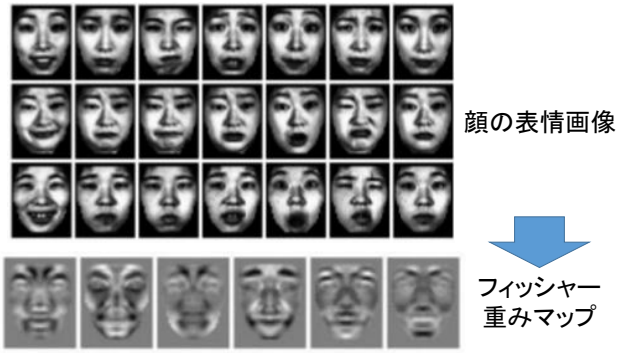

図 2　顔表情識別におけるフィッシャー重みマップによる重みの例[2]

Figure 2　The example of Fisher weight map in facial expression recognition .

これを図 3 のように適用することで印象への寄与度合いが高い周波数帯を抽出できると考える．

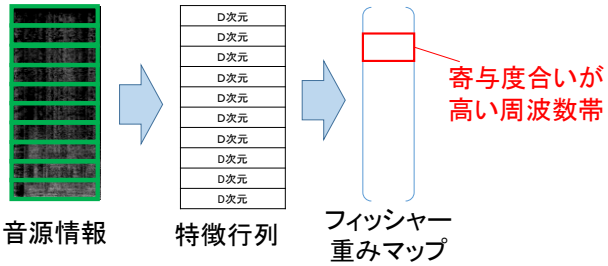

図 3　フィッシャー重みマップによる寄与度合いの高い特徴量の抽出

Figure 3　Extraction of features effective in music mood with Fisher weight map.

また時間帯に関しては，図 4 のように楽曲を一定時間ごとに分割し，各々から抽出した特徴量から，ロジスティック回帰という回帰モデルによりそれぞれの印象クラス属

する事後確率を得ることで，印象の度合いを抽出できると考えた．

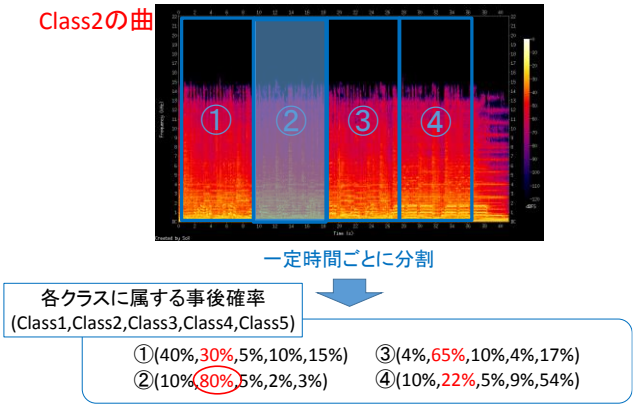

図 4　事後確率による印象の強い時間帯の抽出方法

Figure 4　The extraction method of more impressive time period with posterior probability.

4.3 特徴量の選択

4.2 で述べた抽出を行うためには，一定時間における任意の範囲の周波数帯の特徴を表現できる特徴量が必要である．そこで条件に合致する特徴量として，近年ジャンル分類において音楽特徴量と同等の精度が確認されており，様々な次元の特徴量が抽出できるスペクトログラム画像特徴量を選択した．スペクトログラムとは音波を短時間フーリエ変換したものを時間ごとに並べたもので，縦軸が周波数，横軸が時間，濃淡が振幅の大きさを表している．

スペクトログラムの画像特徴量を利用した楽曲分類の研究として Costa らによる研究がある[3]．Costa らは楽曲のある一定時間のスペクトログラム画像を縦方向(周波数方向)で分割し，各々の周波数帯で抽出した特徴量を用いて分類器を作り楽曲のジャンル分類を行っている．ジャンル分類タスクの識別率は 80%以上とコンペティションである MIREX（the Music Information Retrieval Evaluation eXchange)における最高識別率を超える結果も出している．

4.4 特徴量の設計

印象が強く出ている時間帯を抽出するために，各周波数帯の特徴量をまとめて 1 つの特徴量として分類器を作成する必要がある．また，フィッシャー重みマップを用いて次元を圧縮させた特徴量を用いることでどの程度の性能が出るか評価する必要がある．よって以下の図 5 のような特徴量を用いて実験を行う．

5. 検証実験

5.1 実験概要

本研究で提案した音楽特性と印象の関係分析手法が可能であるのか，また抽出した結果が有効であるのかを検証するため 3 つの実験を行う．まず多クラス識別を行い，本研究に用いる楽曲印象分類器の精度を先行研究との比較を元に確認を行う．次にその印象であるか否かの正否による

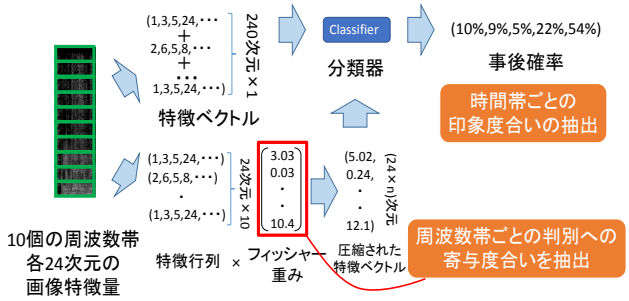

図 5　印象への寄与度合いを抽出する提案手法

Figure 5　The proposed method to extract the degree of contribution to music mood.

二値分類を行い，各クラスの識別精度の確認を行う．最後に各実験から抽出した周波数帯と時間帯の情報が有効なものであるか，主観評価実験を行う．

データセットとしては音源情報と印象のクラスが既知であるものが必要であるため，Mirex-like mood dataset[4]という表 1 のように分けられた 5 クラスのムードで分類された 30 秒の音源が提供されているものを用いた．

表 1　Mirex-like mood dataset クラス詳細情報

Table 1　Class detailed information of Mirex-like mood dataset

Class1	170曲	passionate, rousing, confident, boisterous, rowdy
Class2	164曲	rollicking, cheerful, fun, sweet, amiable/good natured
Class3	215曲	literate, poignant, wistful, bittersweet, autumnal, brooding
Class4	191曲	humorous, silly, campy, quirky, whimsical, witty, wry
Class5	163曲	aggressive, fiery, tense/anxious, intense, volatile, visceral

音源はモノラルに変換し，スペクトログラム画像はモノクロを用いた．スペクトログラム画像は 10 秒ごとに抽出し，上限として 4 k Hz・10kHz まで表現したもの，周波数方向の画像サイズが 513pixel のものと 1025pixel のもので実験を行った．(以下 4k-513，4k-1025, 10k-513, 10k-1025 と表す）また画像特徴量としては先行研究で用いられた GLCM[5]，LBP[6]，シーン画像の識別などでよく用いられる GIST[7]を用いた．GLCM は 0 度，45 度，90 度，135 度のものを用い，距離は 1 と 1,2 両方を用いたものを実験で扱った．LBP は近傍 8 個について，距離は 1,2,3 の 3 種類で実験を行った．また uniform 形式で 59 次元の特徴を抽出した. GIST はスペクトログラムの画像がモノクロであるため，320 次元を特徴量とした．

画像の分割方法としては，時間方向に各曲 3 箇所・9 箇所を抽出したもの，周波数方向は Costa らによる研究[3]を参考に線形尺度，バーク尺度，メル尺度を用いて分割を行った．線形尺度は等間隔に周波数を 10 分割したものである．バーク尺度は聴覚の臨界帯域に対応したものであり，メル尺度は音の高さに対する知覚的尺度である．

5.2 5 クラス分類実験

データセットを訓練とテストに 9:1 の比率で分け，訓練データで作成した分類器によってテストデータが 5 クラス

のどこに属するかを推定する 5 クラス分類を行った．10 回のクロスバリデーションを行い識別率の平均値を評価指標とした．表 2 の“previous”は Costa らの手法のことを指し，周波数帯ごとに分類器を作ってクラス推定を行う方法を用いている．“proposed”は提案手法を指し，周波数帯ごとの特徴量を一つのベクトルにまとめて分類器を作成している．“fisher n”はフィッシャー重みマップで得た重みを n 個使用して次元圧縮を行ったものを特徴量とした手法である．先行研究と提案手法を比べるとほぼ同程度の精度が出ていること，またフィッシャー重みマップで次元圧縮をした特徴量では数パーセントの精度差は生じるが近い精度まで出せることがわかった．またこの精度はデータセット作成者が音楽特徴量を用いて作成した分類器とも同程度の識別率である．

5.3 二値分類実験

各クラスに属するか否かを推定する二値分類を 5 クラス識別とほぼ同様の方法で行った．テストデータとして識別するクラスの 1 割を正答，同数を違うクラスから取り出したためチャンスレートは 50%となっている．フィッシャー重みを 2 つまで使用して次元を圧縮したものを特徴量として使用した結果表 3 のようになった．ここからフィッシャー重みを用いて圧縮した特徴量を用いても class3 については 8 割近い精度で識別できることが確認できる．

表 2　5 クラス識別結果

Table 2　The result of five classes recognition.

	4k-513	4k-1025	10k-513	10k-1025
previous	45.62	46.18	44.27	43.37
proposed	45.73	46.52	43.82	44.61
fisher1	42.13	41.46	44.94	43.37
fisher2	43.93	43.71	44.94	44.04
fisher5	44.72	45.39	46.07	44.49

表内の数値は識別率(%)

表 3　二値識別結果

Table 3　The result of binary classes recognition

	4k-513	4k-1025	10k-513	10k-1025
class1	67.35	67.35	66.18	67.35
class2	67.50	68.44	67.81	66.56
class3	80.71	80.00	80.48	80.24
class4	65.00	65.53	63.68	64.47
class5	77.50	77.19	78.12	77.19

表内の数値は識別率(%)

5.4 主観評価実験

抽出した時間・周波数帯の例は図 6，図 7 のようになる．図 6 は 5 クラス識別で用いた分類器を用い，一つの楽曲内から 5 秒おきに 10 秒ごとのスペクトログラム画像特徴量を抽出し，class3 の事後確率が高い上位 5 位の時間帯を赤，下位 5 位の時間帯を青で表したものである．class3 は日本語では「切ない」などを含むクラスであるので，事後確率が高い時間帯は音が少ない，事後確率が低い時間帯は音が多くなっている傾向が視認できる．

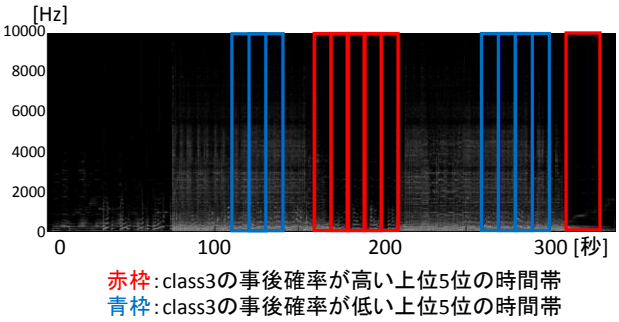

図 6　それぞれの時間帯における印象度合いの抽出

Figure 6　Extract the mood degree in each time period.

時間帯ごとに印象度合いが抽出できていることを確認するため，同じ曲の中の事後確率が高い部分と低い部分を聴き比べる主観評価実験を 9 名に 6 曲ずつ行なった．データとしては FMA（Free Music Archive）[8]の曲を用い，class3 に含まれる“poignant”というタグが付きかつ他のクラスのタグが付いていないもの．その内，5 クラス分類で最も精度が良い分類器で分類を行い class3 に分類された 14 曲からテスト用のデータを選んだ．

全体の正答率は約 70%であり，チャンスレートが 50%であるため事後確率の大小が主観的な印象の強弱とある程度関連があることが確認できる．また各曲の事後確率の差と正答率を分析してみると，同じ曲同士でも事後確率の差が大きいほど正答率が良く，事後確率の差が小さいと正答率も低くなる傾向が確認できた．

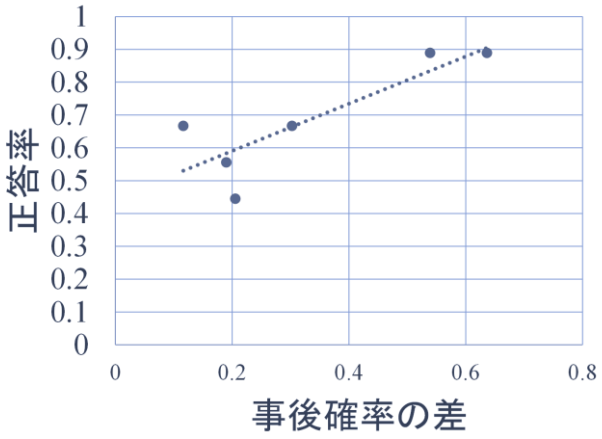

図 7　事後確率の差と正答率についてのグラフ

Figure 7　Graph about the relationship between defference in posterior probability and correct answer rate.

また class3 の二値分類を行った際に識別率が良いパラメータ（分割方法，特徴量，画像の作り方）の分類器で使用したフィッシャー重みマップを絶対値を取って正規化したものを並べたところ，図 8 のようになった．上の図は縦軸

が周波数 Hz であり，下の図は縦軸がピアノの鍵盤の番号になっている．49 番目が A4 と呼ばれるいわゆる 440HZ のラの音である．赤くなっている周波数帯がそれぞれの分類器で識別に重要と抽出された周波数帯であるが，class3 では 200～400Hz が多くの特徴量や分類方法において重要であるということが視認できる．

そこで周波数帯ごとの印象への寄与度合いが有効なものであるのかを確認するため，class3 の曲の印象が強い同じ部分を用い，イコライザで 200～400Hz の振幅を 5～10dB ほど大小にどちらかに加工したものとオリジナルを比べ，印象の変化があったかを主観評価により調べた．データや被験者，曲数は時間帯の実験と同様である．どちらが class3 の印象が強かったか，変化がなかったかの 3 択でアンケートを行ったところ，チャンスレートが 67%ながら，81%が class3 の印象の変化を感じ取った．

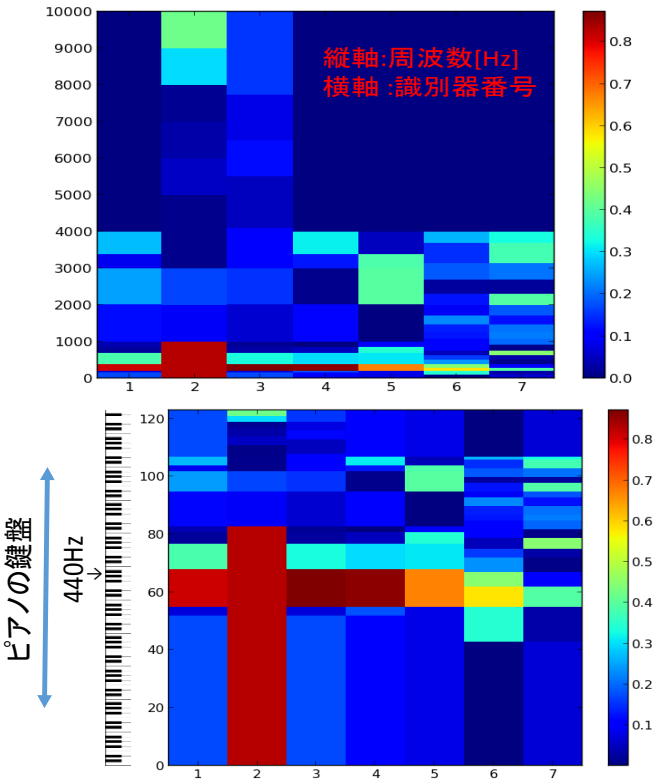

図 8　class3 における周波数帯ごとの印象への寄与度合い

Figure 8　The contribution degree to music mood in each frequency band in class3.

6. 結論と将来展望

本研究では音楽情報処理の分野において，自動楽曲分類手法を用いて識別理由を抽出するという新しいアプローチを検討し，楽曲印象分類技術を用いて印象に寄与する時間帯及び周波数帯を抽出する方法を提案した．また実験を通して提案手法の楽曲印象分類への識別性能を確認し，主観評価から抽出した情報が主観的にも有効なものである可能性を確認した．

将来展望としては，分類器についてより識別精度の高いもの特徴により情報を抽出する必要があること．また主観評価実験の統計的な優位性を確認するべきであり，よりサンプルやクラス数が多いデータセットでの検討をすべきであることが考えられる．

また分類方法や要素の取り方を変えることで，例えば年代別の人気曲の違いという分類で要素として楽器がどの程度楽曲内で弾かれているかという特徴量が取れれば，同様に知見が得られるなどの応用ができると期待している．

参考文献

1) 北村音壱ら,昭和 50 年代の青年に関する音色因子の抽出．音響学会聴覚研資, 1978
2) 篠原雄介ら,フィッシャー重みマップを用いた顔画像からの表情認識．電子情報通信学会技術研究報告パターン認識・メディア理解研究会, Vol. 103, 2004.
3) Y.Costa *et al.*, Music genre classification using LBP textural features. Signal Processing, 92(11), 2012.
4) R Panda *et al.*, Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. In Proc. CMMR, 2013.
5) R M Haralick. Statistical and structural approaches to texture. In Proc. IEEE, 67(5), 1979.
6) T Ojala *et al.*, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(7), 2002.
7) A Oliva *et al.*, Modeling the shape of the scene: a holistic representation of the spatial envelope, IJCV, 42(3), 2001
8) Free Music Archive
http://freemusicarchive.org/