4X-4

# 音楽と映像の調和度計算モデルを用いたクロスメディア検索

斎藤 博己† 糸山 克寿‡ 吉井 和佳‡ 駒谷 和範‡ 尾形 哲也‡ 奥乃 博‡

† 京都大学 工学部情報学科 ‡ 京都大学 大学院情報学研究科 知能情報学専攻

## 1. はじめに

音楽のプロモーションビデオは，音の明るさや激しさなどの雰囲気に合うような映像が付与されており，ユーザは音楽を聴く，映像を見るといった単一のメディアとしての楽しみだけでなく，音楽と映像の調和による楽しみを得られる．計算機が音楽と映像の調和を判断できれば，音楽同士，映像同士といった同一モダリティでの検索だけでなく，音楽から映像，映像から音楽といったクロスメディア検索が可能になり，カラオケ背景映像の自動選択や，音楽プレーヤのビジュアライザ，初心者によるコンテンツ製作支援などへの応用が期待できる．

音楽と映像の調和に基づくクロスメディア検索を実現するためには，これらが調和する関係を定量的に扱う枠組みが必要となる，西山らは，音楽と映像の調和度計算モデル [1] を開発し，音楽と映像の調和に関する定量的な評価の可能性を示唆したが，クロスメディア検索へ適用するには問題が残されていた．我々は，この問題点を解決する新たな調和度計算モデルを構築し，クロスメディア検索を実現した．

## 2. 音楽と映像の調和度計算手法

一般に音楽と映像の調和は，意味的な側面と時間的な側面に分類され [2]，意味的な調和には，音楽と映像のムードの一致による調和，音楽の持つシンボリックな意味と映像内容の一致による調和があり，時間的な調和には，音楽リズムと映像の動きのアクセントの同期による調和がある．本研究では音楽と映像のムードの一致による意味的な調和を扱うモデルをクロスメディア検索に利用し，音楽に調和する映像の検索を行なった．

本研究では，音楽 $M$ と映像 $V$ とが調和しているとき，両者のムードは互いに連想可能，すなわち相互に変換が可能な状態であると考える．$M$ と $V$ に対して，$M$ から連想される映像 $\hat{V}$ が $V$ と類似し，$V$ から連想される音楽 $\hat{M}$ が $M$ と類似しているとき，両者は調和しているとみなす．そこで，音楽と映像それぞれのムードを表す特徴空間を用意し，両空間の間の写像を既存のコンテンツを基に構築する．2.1 節以降に具体的な手法を説明する．

### 2.1 特徴量の抽出

音楽と映像におけるムードを表現する特徴量を，ムード検出に関する先行研究 [3, 4] を参考に，表 1 のように設計する．これらの特徴量はフレームごとに抽出を行い，特徴量ごとに平均が 0，分散が 1 となるように正規化し，主成分分析を用いて次元を圧縮する．実験的に最適な次元数を求めたところ，音楽特徴量は 15 次元（累積寄与率 0.95），映像特徴量も 15 次元（同 1.00）となった．

表 1: 使用する特徴量（括弧内の数字は特徴量の個数）

| | |
|---|---|
| 音楽特徴量 (34) | |
| 音量 | 全体およびサブバンド * ごとの音量 |
| 音色 | スペクトルの特徴量（重心，幅，ロールオフ，フラックス，コントラスト（サブバンド * ごとのピーク値，バレー値，それらの差)） |
| 映像特徴量 (15) | |
| 色 | 明度の重心・分散，CIELUV 色空間におけるヒストグラム |
| モーション | オプティカルフローの時間微分 |
| ショット | YUV 色空間におけるヒストグラムの時間微分 |

* サブバンドにはバンク数 7 のオクターブフィルタバンクを使用.

### 2.2 ムードの相互変換

コンテンツごとの「音楽と映像の調和のしかた」を学習するため，フレームごとに抽出された音楽特徴量 $M$ を，同じくフレームごとに得られた映像特徴量 $V$ へと変換する線形写像 $P$ を以下の線形方程式を解くことで構築する．

$$MP = V \tag{1}$$

音楽と映像の組み合わせごとに写像を学習するので，「類似した音楽が大きく特徴の異なる映像と調和する」という場合があったとしても，「写像の違い」として説明することができる．

### 2.3 調和度の計算

本研究ではミュージックビデオのような音楽と映像が必ず対になっているコンテンツを対象とする．また．その音楽が映像と調和しているとみなし，前述の方法でコンテンツごとにムードを相互変換する写像を作成した．その上で，ある音楽と映像の調和度を計算するには，映像に与えられた線型写像を用いて音楽のムード特徴量を映像特徴量に変換した先で距離を計算する．

## 3. 手法の評価

本手法の妥当性を検証するため，評価実験を行った．本実験では，音楽プロモーションビデオやクラシックやロックのライブ映像など，合計 50 本のコンテンツを対象とした．これらのコンテンツでは，音楽と映像の調和が計られていることが多い，また音楽のジャンルごとに映像も類似していることが多いため，人手による大まかなグルーピングが容易である，一つの楽曲の長さが数分

Cross-media Retrieval Using a Congruency Model between Music and Video in Multimedia Content: *Hiroki Saito, Katsutoshi Itoyama, Kazuyoshi Yoshii, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

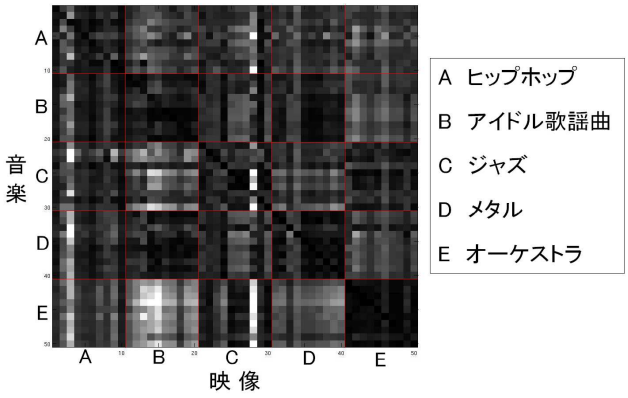


図 1: 音楽と映像の調和度

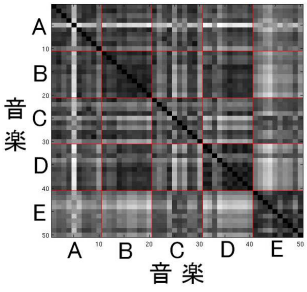


図 2: 音楽と音楽の調和度

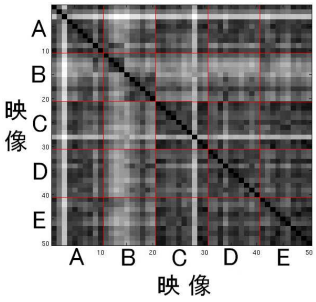


図 3: 映像と映像の調和度

程度で，その間に大きなムードの変化が起こりにくい，等の理由による．これらの映像作品を対象として，

1. 音楽と映像の調和度の計算（本手法）
2. 音楽と音楽の調和度の計算
3. 映像と映像の調和度の計算

を行った．

### 3.1 実験結果

音楽と映像，音楽同士，映像同士の調和度計算の結果をそれぞれ図 1，図 2，図 3 に示す．映像はコンテンツのジャンルごとに 5 種類に分類している．要素の色が黒に近づくほど調和度が高い．

図 2 の音楽と音楽の調和度では，アイドル歌謡曲 (B) とメタル (D) とが調和していること，これらのジャンルの楽曲では楽器構成が類似しており，そのことを反映していると考えられる．ヒップホップ (A) は他のどのジャンルともあまり調和しておらず，ジャズ (C) とクラシック (E) とはやや調和しているものの，それぞれのジャンル内部の方がより調和している．また，図 3 の映像と映像の調和度では，基本的に暗い背景でのライブ演奏であるジャズ，メタル，オーケストラが比較的調和していることが分かる．明るいステージでの映像が中心のアイドル歌謡曲はどのジャンルとも調和しない．

図 1 の音楽と映像の調和度では，基本的に図 2 と似た傾向が見られるが，ジャズとオーケストラの調和度がそれぞれのジャンル間でほぼ等しい点，ヒップホップの音楽が他のジャンルの映像とも調和している点などが異

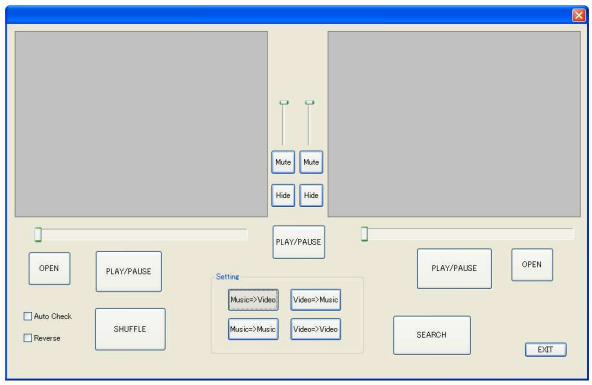


図 4: インタフェース

なっている．音楽と映像の調和を考えることで，一方だけの特徴による調和では扱えなかった関係が扱えていることが分かる．

## 4. クロスメディア検索インタフェース

本稿で述べた調和度計算手法を用いて，図 4 に示す音楽と映像のクロスメディア検索インタフェースを構築した．左右 2 つのウインドウに映像コンテンツを表示でき，それぞれのコンテンツの音量や映像のオン・オフはユーザが自由に制御できる．

左右のウインドウ下部の「検索」ボタンを押すと，反対側のウインドウにそのコンテンツに調和した新たなコンテンツが表示される．どのような調和に基づく検索を行うかは．中央のパネルで設定することができる．

## 5. おわりに

本稿では，ムードの一致に基づいて音楽と映像の調和度を計算するモデルを構築し，その妥当性を評価した．また，この調和度計算手法に基づくクロスメディア検索インタフェースを実装した．今後は，特徴量の洗練化や時間的なムードの変化を扱えるようモデルを拡張すると共に，検索インタフェースに関する評価を行う予定である．

**謝辞** 本研究の一部は，科研費，グローバル COE，Crest-Muse の支援をうけた．

## 参考文献

[1] 西山他，“マルチメディアコンテンツにおける音楽と映像の調和度計算モデル”，情処研報，2007-MUS-69, pp. 31–36, 2007.

[2] 岩宮眞一朗，“音楽と映像のマルチモーダル・コミュニケーション”，九州大学出版会，2000.

[3] L. Lu, D. Liu, and H. J. Zhang, “Automatic Mood Detection and Tracking of Music Audio Signals,” *IEEE Trans. on Audio, Speeck, and Language Process*, Vol. 14, No. 1, pp. 5–18, 2006.

[4] H. B. Kang, “Affective Content Detection using HMMs,” *Proc. of the 11th ACM Multimedia*, pp. 259-262, 2003.