2N-6

マルチメディアコンテンツにおける音楽と映像の調和に関する分析

西山 正紘　　北原 鉄朗　　駒谷 和範　　尾形 哲也　　奥乃 博

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

映画やドラマのようなマルチメディアコンテンツでは，BGM として音楽が用いられ，コンテンツをより印象的なものにしている．しかし，音楽と映像をただ組み合わせればよいわけではなく，映像に調和した音楽が組み合わされて初めてコンテンツは印象的なものになる．心理学的知見によると，一般に，音楽と映像の調和に関する要因としては，時間的なアクセントの一致による時間的調和と，ムードやシンボリックな意味の一致による意味的調和の 2 つが存在する [1]．計算機が音楽と映像の調和を判断できるようになれば，初心者によるコンテンツ制作支援等への応用が期待できる．そこで本稿では，音楽と映像の調和を理解できる計算機の実現の第一段階として，意味的調和に基づいた音楽と映像の調和度計算モデルを提案する．

関連研究として，印象語を介して音楽と映像の調和度を判断し，それに基づき映像に合う音楽の検索を実現した事例がある [2][3]．しかしこれらの研究では，印象語は逐一人手で付与されており，ラベリングの信頼性や大規模 DB に対しては付与にかかるコストの点で問題がある．そこで，本研究では，印象語を介さずに音楽と映像の調和度を自動計算する手法を提案する．具体的には，音楽音響信号と映像信号からそれぞれボトムアップに特徴量を自動抽出し，それぞれの特徴量間の写像を実現することで調和度を計算する．

2. ムードの一致に基づく音楽と映像の調和度計算手法

一般に，音楽と映像の意味的調和とは，音楽と映像のムードの一致による調和と，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に分けられる [1]．前者に関しては激しい場面には激しい音楽が合うこと，後者に関しては日本人には別れの場面に「蛍の光」の音楽が合うことが例として挙げられる．ただし，音楽や映像のシンボリックな意味理解は文化的背景等の高次の知識処理を必要とするため対象外とし，本研究ではムードの一致に基づく意味的調和のみを扱う．また，簡便のため同一シーン内ではムードは時間的に変化しないと仮定し，シーンの切り出しは予め行っておくものとする．

音楽と映像からその調和を理解する過程を以下のように仮定する（図 1）．本研究では，音楽と映像の調和している状態を，両者を相互に連想可能な状態と考える．つまり，ある音楽 M と映像 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 と類似する時，V と M 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 (上記の「連想」に相当) を構築することで調和の度合いを定量化する．特徴空間は，2.1 節で述べるように，既存のムード検出等の研究を参考に設計し，特徴空間の間の写像は 2.2 節で述べるように主成分分析に基づいて構築する．

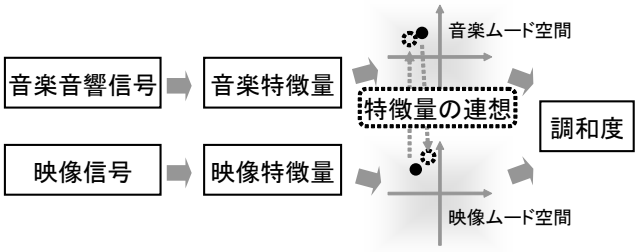


図 1: 音楽と映像から調和を理解する過程.

表 1: 使用する特徴量 (括弧内の数字は特徴量の個数).

	音楽 (33 個)
音量	全体およびサブバンド * 毎の音量
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 * 毎のピーク値，バレー値，それらの差)[4])
	映像 (17 個)
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.

2.1 ムードを表現する特徴量の抽出

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 [4][5] 等を参考に，表 1 のように設計する．これらの特徴量はフレーム毎に抽出を行い，最終的にはその平均と標準偏差で特徴量を表現する．よって抽出される音楽特徴量 f_m および映像特徴量 f_v の次元はそれぞれ 66，34 次元となる．

2.2 ムードの一致に基づく調和度計算モデル

音楽特徴空間と映像特徴空間の間の写像を，以下で述べる主成分分析に基づくモデルにより構築する．なおモデルは，音楽と映像の調和した組を用いて学習する．

モデルの学習

まず音楽特徴量 f_m，映像特徴量 f_v に対して標準化を行う．標準化後の音楽特徴量，映像特徴量をそれぞれ u_m, u_v とする．これらに対してそれぞれ主成分分析を行うと，u_m, u_v は次式で近似できる．

$$u_m \simeq A_m x_m,\ u_v \simeq A_v x_v \tag{1}$$

ここで，A_m, A_v は主成分を並べた行列，x_m, x_v は主成分の係数である．主成分は固有値が 1 以上のものを採用する．(本実験では，x_m, x_v はそれぞれ平均 10，5 次元となった．) 音楽と映像の調和を考慮して制作された作品から x_m, x_v を抽出した時，これらには相関があると考えられるので，これらを連結した特徴量 x に対してさらに主成分分析を行うと，x は次式で近似できる．

Analysis of Congruency between Music and Video in Multimedia Content: *Masahiro Nishiyama, Tetsuro Kitahara, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

表 2: 使用した映像作品.

(1) パイレーツ・オブ・カリビアン
(2) スターウォーズ・エピソード 1
(3) スターウォーズ・エピソード 2
(4) キャッチ・ミー・イフ・ユー・キャン
(5) バック・トゥ・ザ・フューチャー
(6) オペラ座の怪人

$$x = \begin{pmatrix} x_m \\ x_v \end{pmatrix} \simeq Pc = \begin{pmatrix} P_m \\ P_v \end{pmatrix} c \qquad (2)$$

ここで，P は主成分を並べた行列，c は主成分の係数である．(本実験では，c は平均 7 次元となった．)

モデルを用いた調和度の計算

音楽特徴量 x_m から連想される画像特徴量 $\hat{x_v}$，および画像特徴量 x_v から連想される音楽特徴量 $\hat{x_m}$ は，式 2 を変形させることにより学習したパラメタ P_m, P_v を用いて，

$$\hat{x_v} = P_v P_m^- x_m,\ \hat{x_m} = P_m P_v^- x_v \qquad (3)$$

と表現できる．ここで，P_m^-, P_v^- はそれぞれ P_m, P_v の一般化逆行列である．これらの連想された特徴量を用いて，音楽 M と映像 V の調和度 $Dist(M||V)$ を，

$$Dist(M||V) = \{d(x_m, \hat{x_m}) + d(x_v, \hat{x_v})\}/2 \qquad (4)$$

で定義する．ここで，$d(x, \hat{x})$ は x, $\hat{x}$ のコサイン距離 $\langle x, \hat{x}\rangle / \|x\| \cdot \|\hat{x}\|$ である．ゆえに調和度 $Dist(M||V)$ は -1 から 1 の範囲の値で与えられる．

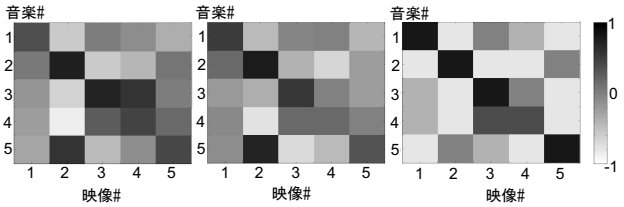
3. 評価実験

実映像作品を対象として，提案手法に基づき音楽と映像の調和を判定する実験を行った．入力信号として，表 2 の映像作品から各々20 シーン (約 20 秒/シーン) を切り出し，得られた計 120 組の音楽音響信号と映像信号を用いた．それらの信号から表 1 の特徴量を抽出し，2.2 節で述べたモデルの学習を行った．学習はクローズド，オープンの 2 通りの方法で行った．オープンは作品単位での評価を行い，例えば，作品 (1) のデータの評価をする際には，作品 (1) 以外のデータを用いて学習を行った．次に評価データとして，各映像作品から 5 シーンを選び，得られた 5 つずつの音楽音響信号と映像信号に対し，それら全ての組み合わせである 25 組 (その中の 5 組は元々の信号の組み合わせ) のデータを作成し，計 150 組のデータを用意した．これらのデータの調和度を，学習したモデルを用いて計算した．また実験結果の評価の比較対象として，人間による評定を被験者実験により収集した．5 人の被験者には各データを視聴した後，その調和度を 5 段階 SD 法を用いて付与してもらった．

3.1 被験者実験による結果との比較

実験結果の例として，作品 (1) の評価データの調和度評価の結果 (クローズド学習，オープン学習，被験者) をそれぞれ図 2(a)，(b)，(c) に示す．縦軸は音楽のシーン番号，横軸は映像のシーン番号，図の i 行 j 列の要素はシーン i の音楽とシーン j の映像の調和度を表す．要素の色が黒色に近づくほど調和度が高いことを表す．

実験結果の妥当性を，被験者による結果との一致度をもって評価した．閾値処理により結果の調和・不調和への 2 値化処理を行い，被験者実験の結果を正解とした時，実験結果の一致度を調べた．閾値を 0.4 とした時の結果



(a) クローズド (b) オープン (c) 被験者

図 2: 被験者実験による調和度評価との比較.

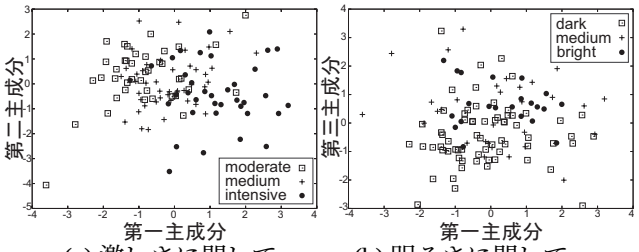
表 3: 被験者実験結果を正解とした時の一致度 (単位%).

作品番号	(1)	(2)	(3)	(4)	(5)	(6)
クローズド	84	80	84	72	68	64
オープン	84	80	64	72	60	64

を表 3 に示す．最高で 84 %，最低でも 60 %の一致度が得られており，提案手法の有効性が確認できる．また本実験で使用した作品は，作品毎に制作者が異なるため音楽と映像の組み合わせ方に個性が反映されると考えられるが，クローズドとオープンの結果にあまり差がないことから，作品に依存しない音楽と映像の一般的な写像がモデルに学習されていることが示唆される．

3.2 主成分空間上でのシーンの分布

主成分空間上でのシーンの分布を図 3 に示す．各シーンへのラベルは人手により三段階で付与した．なお，このラベルは主成分空間上での分布を調べるために付与したものであり，調和度計算自体には必要ない．図より，シーン同士の類似度が空間内の距離に反映されていることが確認でき，第一主成分が激しさ，第三主成分が明るさを表現する軸であると解釈できる．



(a) 激しさに関して (b) 明るさに関して

図 3: 主成分空間内での各シーンの分布.

4. 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提案し，提案手法の有効性を評価実験により確認した．今後は，時間的要因も扱う調和度計算モデルへ手法を拡張する予定である．

謝辞: 本研究の一部は，日本学術振興会科学研究費補助金，21 世紀 COE プログラムの支援を受けた．

参考文献

[1] 岩宮眞一郎. 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.

[2] 古賀広昭, 下塩義文, 小山善文. 画像に合った音楽の選定技術. 映像情報メディア学会技術報告, pp. 25–32, 1999.

[3] 宝珍輝尚, 井田俊博, 都司達夫. 印象に基づく映像と音楽の相互検索に関する一考察. 情処研報, pp. 97–104, 2002.

[4] L. Lu, D. Liu, and H. J. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Trans. on Audio, Speech, and Language Process.*, Vol. 14, No. 1, pp. 5–18, 2006.

[5] H. B. Kang. Affective content detection using HMMs. *Proc. of the 11th ACM Multimedia*, pp. 259–262, 2003.