2007－MUS－69 (6)
2007／2／23

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

西山正紘† 北原鉄朗† 駒谷和範†
尾形哲也† 奥乃博†

概要 本稿では，アクセント構造およびムードの一致に基づいて，音楽と映像の調和の度合い（調和度）を計算する枠組を提案する．一般に，音楽と映像の調和要因としては，時間的なアクセントの一致による時間的調和と，ムードの一致による意味的調和の２つが存在する．従来の研究では，それぞれの要因のみしか扱っておらず，両要因を統一的に扱った事例は存在しない．そこで本稿では，音楽と映像の調和度を，アクセント構造の一致に基づいて定量化した調和度とムードの一致に基づいて定量化した調和度の重み付き線形和で表現する．アクセント構造の一致は音楽と映像それぞれの特徴量系列間の相関に基づいて，ムードの一致はそれぞれの特徴部分空間内における相互の特徴量の連想に基づいて定量化する．実映像作品を対象とし，本手法の有効性を実験により評価した．

A Computational Model of Congruency between Music and Video in Multimedia Content

Masahiro Nishiyama ,† Tetsuro Kitahara ,†
Kazunori Komatani ,† Tetsuya Ogata † and Hiroshi G. Okuno†

Abstract In this paper, we propose a framework that understands congruency between music and video based on similarity of accent structure and mood. There are two types of congruency between music and video: temporal congruency related to synchronization of accents and semantic congruency related to similarity of mood. Previous works, however, have dealt only with either congruency. We model the temporal congruency based on the correlation between accent feature sequences extracted from audio and visual content, and the semantic congruency based on mutual mapping between two feature spaces representing music and video respectively. Then, we integrate the two types of congruency as a weighted linear sum. Our experiments with real-world content show the effects of our method.

1. はじめに

一般に，映画やテレビのようなマルチメディアコンテンツでは，映像に合うような音楽がBGMとしてコンテンツをより印象的なものにするために用いられている．また，ミュージックビデオのようなマルチメディアコンテンツでは，音楽に合うように映像が付加されており，音楽の印象をより鮮明なものにするために役立っている．このように，マルチメディアコンテンツにおける音楽と映像は，組み合わされることで互いの印象を強め合っている．しかし，映像と音楽をただ組み合わせればよいというものではない．両者を適切に組み合わさなければ，コンテンツをより印象的なものにするどころか，かえってその印象を悪くしてしまうだろう．つまり，映像と音楽を組み合わせる際には，それらを適切に組み合わせる必要がある．映像と音楽がどのようなパラメタで関わりあっているかというデザインが解明できれば，マルチメディアコンテンツの創作支援などの応用が期待できる．

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]．音楽と映像の調和に関する従来研究には，時間的調和を扱ったもの[2),3)]，意味的調和を扱ったもの[4),5)]があり，それぞれの要因から音楽と映像の調和を扱っているが，両要因を統一的に扱った事例は存在しない．一方，心理学の分野では，時間的調和を扱ったもの[6)]，意味的調和を扱ったもの[7)]に加え，両要因を統一的に扱ったものが事例として存在する[8),9)]．文献8)，9)では，両要因が調和に及ぼすモデルが提案されており，そのモデルの有効性が文献7)によって証明されている．よって音楽と映像の調和を理解する上では，両要因を扱う必要があると考える．

そこで，本稿では，時間的調和および意味的調和に基づいて，マルチメディアコンテンツにおける音楽と映像の調和の度合い（調和度）を計算するための枠組

† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

を提案する．時間的調和に関しては両者の時間的なアクセント（アクセント構造）の一致に基づいて，意味的調和に関しては両者の持つムードの一致に基づいて定量化する．そして，音楽と映像の調和度を，これら2種類の調和度の重み付き線形和として表現する．

ここで，アクセント構造の一致をどのように判断するのか，またムードの一致をどのように判断するのかが重要となる．アクセント構造の一致に関しては，音楽と映像のアクセントが同期する時に両者に調和が感じるとされている[1]．よって，時間軸上で音楽と映像それぞれのアクセントが表現される特徴を用意し，それら特徴量系列間の相関に基づいてアクセント構造の一致を判断する．ムードの一致に関して，本稿では，音楽と映像が調和している状態を，両者を相互に連想可能な状態と仮定する．音楽と映像それぞれのムードを表現する特徴空間を用意し，特徴量からボトムアップに構築された写像に基づいて相互の特徴量を連想することでムードの一致を判断する．

以下に本稿の構成を述べる．2章では，本稿で扱う調和要因の定義を行う．3章では，本稿で提案する音楽と映像の調和度計算手法について述べる．4章では，提案手法の有効性を実映像作品を対象とした実験により評価する．最後に5章では，本稿のまとめを行う．

2. 時間的調和と意味的調和

本稿では，図1のように音楽と映像との調和を時間的調和と意味的調和に分けて考え，与えられた作品に対してシーン毎にこの2種類の調和度を計算する手法を設計する．なお，作品からシーンへの切り出しは予め行っておくものとする．

2.1 時間的要因

一般に，音楽と映像の時間的調和とは，時間的に進行する音事象と映像における動きのアクセントの，時間軸上での調和を意味する[1]．例として，音楽の拍節構造と動きのアクセントが一致すると調和が感じられることが挙げられる．本研究ではこのような両者の時間的なアクセント（アクセント構造）の一致に基づく時間的調和を扱う．

アクセント構造に基づいてその調和を判断する過程を次のように考える．一般に，音楽と映像のアクセントが同期する時に両者に調和が感じるとされている[1]．そこで，時間軸上で音楽と映像それぞれのアクセントが表現される特徴を用意し，両特徴量系列間のピーク値の同時性を判断することで調和の度合いを定量化する．特徴空間は3.1.1項で述べるように，既存のアクセント検出等の研究を参考に設計し，特徴量系列間のピーク値の同時性は3.1.2項で述べるように，特徴量系列間の相関に基づいて判断する．

2.2 意味的要因

一般に，音楽と映像の意味的調和とは，音楽と映像

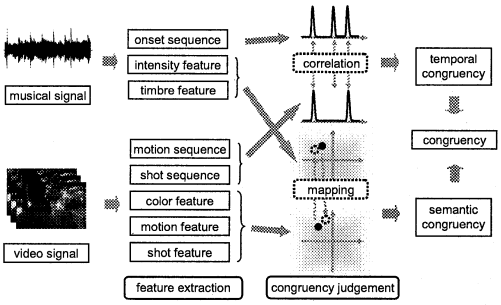


図1 音楽と映像から調和を理解する過程．

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する．

ムードに基づいてその調和を判断する過程を次のように考える．まず，音楽と映像が調和している状態を，両者を相互に連想可能な状態と仮定する．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$\hat{\mathrm{V}}$がVと類似し，Vから連想される音楽$\hat{\mathrm{M}}$がMと類似する時，MとV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 (上述の「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，3.2.1項で述べるように，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3.2.2項で述べるように主成分分析に基づいて構築する．

3. 音楽と映像の調和度計算手法

本章では，2章で考察したことを元に，音楽音響信号と映像信号の調和度を計算する手法を提案する．

3.1 アクセント構造の一致に基づく調和度計算手法

3.1.1 アクセント構造を表現する特徴量の抽出

音楽と映像におけるアクセント構造を表現する特徴量を，先行研究[2),3)]等を参考に，表1のように設計した．これらの特徴量により，音楽のアクセント音や映像の動きのアクセントやショットの切り替わりを表現する．これらの特徴量はフレーム毎（30fps）に抽出を行い，特徴量系列として表現される．よって抽出される音楽特徴量系列 $f_{\mathrm{m},t}$ および映像特徴量系列 $f_{\mathrm{v},t}$ の次元はそれぞれ1，2次元となる．

3.1.2 アクセント構造に基づく調和度計算モデル

特徴量系列間のピーク値の同時性を，以下で述べる

表 1 アクセント構造に基づく調和度計算で使用する特徴量.

音楽特徴量	
1	オンセット系列
映像特徴量	
モーションに関する特徴	
1	オプティカルフローの平均値系列の時間微分
ショットに関する特徴	
2	YUV 色空間におけるヒストグラム系列の時間微分

特徴量系列間の相関に基づいて判断する．まず音楽特徴量系列 $f_{\mathrm{m},t}$，映像特徴量系列 $f_{\mathrm{v},t}$ に対して 0 から 1 の間の値を取るよう正規化を行い，$f_{\mathrm{v},t}$ に対してはさらに各時刻毎に成分の平均を取り，その次元数を 1 にする．そして，特徴量系列中のノイズを除去し，ピークを強調させるための処理を行う．具体的には，各特徴量系列の時間方向における平均および標準偏差を計算し，各時刻において特徴量時系列の値からそれぞれ平均と標準偏差の定数倍の値を減算する（本実験では定数を 2.0 に設定した）．以上の処理後の特徴量系列をそれぞれ $f'_{\mathrm{m},t}$, $f'_{\mathrm{v},t}$ とする．これらの特徴量系列を用いて，音楽 M と映像 V のアクセント構造に基づく調和度 $Cor(\mathrm{M}||\mathrm{V})$ を，

$$Cor(\mathrm{M}||\mathrm{V}) = \frac{\sum_t f'_{\mathrm{m},t} f'_{\mathrm{v},t}}{\sqrt{\sum_t (f'_{\mathrm{m},t})^2 \sum_t (f'_{\mathrm{v},t})^2}} \qquad (1)$$

で定義する．アクセント構造に基づく調和度 $Cor(\mathrm{M}||\mathrm{V})$ は -1 から 1 の範囲の値で与えられる．

3.2 ムードの一致に基づく調和度計算手法

3.2.1 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[10)~13)] 等を参考に，表 2 のように設計した．これらの特徴量により，音楽の激しさや明るさ，映像の色調や明るさ，動きの激しさを表現する．これらの特徴量はフレーム毎 (30fps) 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 g_{m} および映像特徴量 g_{v} の次元はそれぞれ 66，34 次元となる．

3.2.2 ムードに基づく調和度の計算

音楽特徴空間と映像特徴空間の間の写像を，以下の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 g_{m}，映像特徴量 g_{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 g'_{m}, g'_{v} とする．これらに対してそれぞれ主成分分析を行うと，g'_{m}, g'_{v} は次式で近似できる．

$$g'_{\mathrm{m}} \simeq A_{\mathrm{m}} x_{\mathrm{m}},\ g'_{\mathrm{v}} \simeq A_{\mathrm{v}} x_{\mathrm{v}} \qquad (2)$$

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 に対して主成分分析を行うと，x は次式で近似できる．

$$x = \begin{pmatrix} x_{\mathrm{m}} \\ x_{\mathrm{v}} \end{pmatrix} \simeq Pc = \begin{pmatrix} P_{\mathrm{m}} \\ P_{\mathrm{v}} \end{pmatrix} c \qquad (3)$$

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 $\hat{x_{\mathrm{v}}}$，および画像特徴量 x_{v} から連想される音楽特徴量 $\hat{x_{\mathrm{m}}}$ は，式 3 を変形させることにより，

$$\hat{x_{\mathrm{v}}} = P_{\mathrm{v}} P_{\mathrm{m}}^{-} x_{\mathrm{m}},\ \hat{x_{\mathrm{m}}} = P_{\mathrm{m}} P_{\mathrm{v}}^{-} x_{\mathrm{v}} \qquad (4)$$

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 のムードに基づく調和度 $Dist(\mathrm{M}||\mathrm{V})$ を，

$$Dist(\mathrm{M}||\mathrm{V}) = \{d(x_{\mathrm{m}}, \hat{x_{\mathrm{m}}}) + d(x_{\mathrm{v}}, \hat{x_{\mathrm{v}}})\}/2 \qquad (5)$$

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．ゆえにムードに基づく調和度 $Dist(\mathrm{M}||\mathrm{V})$ は -1 から 1 の範囲の値で与えられる．

3.3 アクセント構造とムードの一致に基づく調和度の計算

以上で定義したアクセント構造に基づく調和度 $Cor(\mathrm{M}||\mathrm{V})$ とムードに基づく調和度 $Dist(\mathrm{M}||\mathrm{V})$ を用いて，アクセント構造とムードの一致に基づく調和

表 2 ムードに基づく調和度計算で使用する特徴量.

音楽特徴量	
音量に関する特徴	
1	全体の音量
2 - 7	サブバンド * 毎の音量
音色に関する特徴	
8	スペクトル重心
9	スペクトル幅
10	スペクトルロールオフ
11	スペクトルフラックス
12 – 18	サブバンド毎のピーク値
19 – 25	サブバンド毎のバレー値
26 – 33	サブバンド毎のピーク値とバレー値の差
映像特徴量	
色に関する特徴	
1 – 2	CIELUV 色空間における L 値の重心，分散
3 – 8	CIELUV 色空間におけるヒストグラム
9 – 10	YUV 色空間における Y 値の重心，分散
11 – 12	YUV 色空間における U 値の重心，分散
13 – 14	YUV 色空間における V 値の重心，分散
モーションに関する特徴	
15	オプティカルフローの時間微分
ショットに関する特徴	
16	CIELUV 色空間におけるヒストグラムの時間微分
17	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.

表 3 使用した映像作品 (アニメ「ファンタジア 2000」より).

(1-1) 交響曲第 5 番ハ短調 op.67「運命」より
(1-2) 交響詩「ローマの松」より
(1-3)「ラプソディ・イン・ブルー」より
(1-4) ピアノ協奏曲第 2 番ヘ長調 op.102 より
(1-5) 交響詩「魔法使いの弟子」
(1-6)「威風堂々」第 1, 2, 3, 4 番より

表 4 使用した映像作品 (映画).

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

度 $Con(\mathrm{M}||\mathrm{V})$ を,

$$Con(\mathrm{M}||\mathrm{V}) = (1-\alpha)Cor(\mathrm{M}||\mathrm{V}) + \alpha Dist(\mathrm{M}||\mathrm{V}) \quad (6)$$

で定義する．ここで，α は 0 から 1 の間の値を取る各要因の重みを表現する定数である．よって，アクセント構造とムードに基づく調和度 $Con(\mathrm{M}||\mathrm{V})$ は -1 から 1 の範囲の値で与えられる．

4. 評価実験

本章では，提案手法の有効性を確認するために実映像作品を用いた実験を行う．音楽と映像が効果的に組み合わされている作品は，アニメやミュージックビデオのような音楽に対して映像が付加された作品（音楽ベースの作品）と，映画やドラマのような映像に対して音楽が付加された作品（映像ベースの作品）の 2 種類に分けられると考える．そしてその調和の判断に関して，音楽ベースの作品においてはアクセント構造に基づく調和が，映像ベースの作品においてはムードに基づく調和が優位に働くと仮定する．そこで本実験では各種類の作品の例として，アニメーションと映画の作品を対象として実験を行う．実験 1 では音楽ベースの作品を対象にアクセント構造の調和に関する実験を行う．実験 2 では映像ベースの作品を対象にムードの調和に関する実験を行う．最後に実験 3 では両作品を対象に両要因に基づく調和に関する実験を行う．

4.1 実験 1：アクセント構造の調和に関する実験

実映像作品を対象として，アクセント構造の一致に基づき音楽と映像の調和を判定する実験を行った．評価データとして，表 3 の映像作品から各々5 シーン (約 20 秒/シーン) を切り出し，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，3.1 節で述べた手法により計算した．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．

実験結果の妥当性を，被験者による結果と比較することで評価した．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，実験結果の識別率，精度，再現率を調べた．被験者実験の結果の閾値は 0 とし，実験結果の閾値 θ は経験的に求め，本実験では 0.1 とした．その結果を表 6 の左列に示す．最高で 76.0 %，平均 61.3 %の識別率が得られており，提案手法の有効性を確認できた．なお，作品 (1-3) の評価値が他の作品に比べて低くなっている．これはオリジナルの組み合わせ以外において抽出されたアクセントが偶然にも一致したため相関値が高い値を取るデータが多く，その一方で被験者による評定ではそれらのデータに対して低い値が付与されていたためである．このことから人間が調和を判断する際には，各時刻におけるアクセントの一致以外の要因も考慮していると考えられる．被験者実験において，一致の判断に関して，直前までの一致の傾向に基づき未来の一致時刻を予測するという力学系の性質が存在するという意見を得た．よって，これを考慮すれば提案手法の性能がさらに向上すると期待される．

4.2 実験 2：ムードの調和に関する実験

実映像作品を対象として，ムードの一致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 4 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 2 の特徴量を抽出し，3.2.2 項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 (2-1) のデータの評価をする際には，作品 (2-1) 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．

4.2.1 被験者実験による結果との比較

実験結果の妥当性を，被験者による結果と比較することで評価した．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，実験結果の識別率，精度，再現率を調べた．被験者実験の結果の閾値は 0 とし，実験結果の閾値 θ は経験的に求め，本実験では 0.2 とした．オープン学習における結果を表 7 の中央列に示す．最高で 88.0 %，平均 68.0 %の識別率が得られており，提案手法の有効性を確認できた．次にクローズド学習とオープン学習における識別率の結果を表 5 に示す．本実験で使用

表 5　クローズドとオープン学習における識別率の比較 (単位%).

#	(2-1)	(2-2)	(2-3)	(2-4)	(2-5)	(2-6)
closed	92.0	80.0	64.0	68.0	64.0	68.0
open	88.0	80.0	52.0	68.0	64.0	56.0

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．

4.2.2　主成分空間上でのシーンの分布

各主成分の因子負荷量を調べたところ，第一主成分は音量やピーク値，バレー値，各色空間ヒストグラムの時間微分と相関が高いことが分かった．また第三主成分はスペクトル重心やスペクトルロールオフ，L 値（明度），Y 値（輝度）と相関が高いことが分かった．この知見に基づき，各シーンの主成分空間上での分布を調べるために，各シーンに対して激しさと明るさの観点からその程度を人手により 3 段階 SD 法（激しい—穏やか，明るい—暗い）で付与した．主成分空間上でのシーンの分布を図 2 に示す．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．

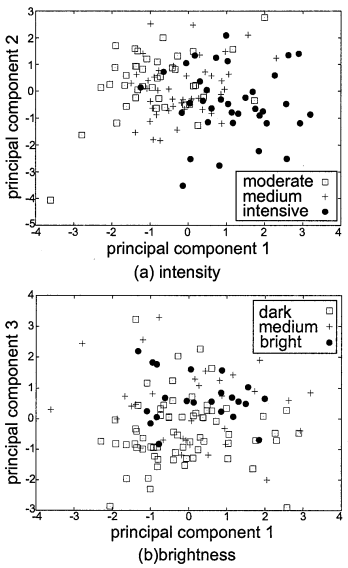


図 2　主成分空間内での各シーンの分布．

4.3　実験 3：アクセント構造とムードの調和の統合に関する実験

実映像作品を対象として，アクセント構造とムードの一致に基づき音楽と映像の調和を判定する実験を行った．評価データは実験 1，2 で使用したデータを利用する．まず，実験 1 で使用した表 3 のデータのムードの一致に基づく調和度を実験 2 における手順と同様に計算した．次に，実験 2 で使用した表 4 データのアクセント構造の一致に基づく調和度を実験 1 における手順と同様に計算した．最後に 3.3 節で述べた手法でアクセント構造とムードの一致に基づく調和度を計算した．実験結果の評価の比較対象として使用する人間による評定は，実験 1，2 で収集したものを使用した．

4.3.1　被験者実験による結果との比較

実験結果の妥当性を，被験者による結果と比較することで評価した．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，実験結果の識別率，精度，再現率を調べた．被験者実験の結果の閾値は 0 とし，実験結果の閾値 θ は経験的に求め，本実験では表 3 のデータに関しては 0.1，表 4 のデータに関しては 0.2 とした．表 3 のデータの結果を表 6 に，表 4 のデータの結果を表 7 に示す．なお，表中の太字は統合により値の向上もしくは維持が確認できたものを表す．アクセント構造のみに基づいたもの（左列），ムードのみに基づいたもの（中央列）に比べ，両者に基づいたもの（右列）の評価値（特に調和に関する精度，不調和に関する再現率）が向上しており，提案手法の有効性が確認できる．

4.3.2　重みの変化による評価値の変化

上記の被験者実験による結果との比較では，評価値が高い値を取る際の各調和要因の重みが表3，4のデータの間で大きく異なる値を取っていた．そこで，α を変動させた時の各評価値の変化を図 3 に示す．図 3 より，表 3 の作品においては，α の値が増加するに従い評価値が減少し，一方，表 4 の作品においては，α の値が増加するに従い評価値が増加している．これより各調和要因の重み α はコンテンツ依存であると考察される．つまり，表 3 のような音楽ベースのコンテンツ（アニメーション等）においてはアクセント構造の一致による調和判定が優位であること，表 4 のような映像ベースのコンテンツ（映画等）においてはムードの一致による調和判定が優位であると考えられ，本章冒頭で立てた仮説と一致する結果が得られた．

5.　おわりに

本稿では，アクセント構造および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実映像作品を対象とした評価実験により確認した．また，音楽ベースの作品においてはアクセント構造に基づく調和が，映像ベースの作品においてはムードの一致に基づく調和が優位に働くことを確認した．

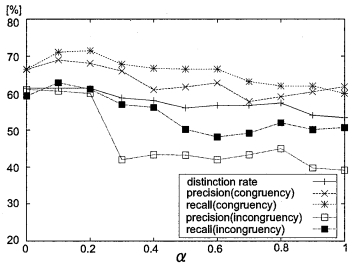
今回はムードの一致の判断において音楽と映像の特徴量の線形な関係を扱ったが，両者の間には非線形な関係も存在するはずである．また，シーンの前後関係という文脈的な要因が調和の判断に影響を与えること

表 6　表 3 の映像作品による評価（D.R.:識別率，P.R.:精度，R.R.:再現率，単位は%）．

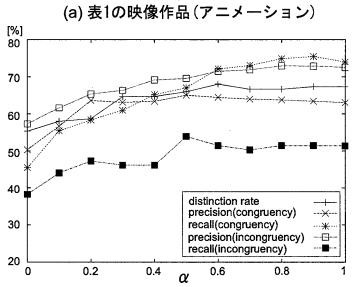
	temporal congruency $\theta = 0.1$					semantic congruency $\theta = 0.2$					temporal and semantic congruency $\alpha = 0.1, \theta = 0.1$				
		congruency		incongruency			congruency		incongruency			congruency		incongruency	
#	D.R.	P.R.	R.R.	P.R.	R.R.	D.R.	P.R.	R.R.	P.R.	R.R.	D.R.	P.R.	R.R.	P.R.	R.R.
(1-1)	60.0	66.7	57.1	53.8	63.6	52.0	60.0	42.9	46.7	63.6	56.6	63.6	50.0	50.0	**63.6**
(1-2)	76.0	80.0	66.7	73.3	84.6	60.0	62.5	41.7	58.8	76.9	72.0	72.7	**66.7**	71.4	76.9
(1-3)	44.0	33.3	40.0	53.8	46.7	44.0	25.0	20.0	52.9	60.0	**44.0**	**33.3**	**40.0**	**53.8**	46.7
(1-4)	76.0	60.0	75.0	86.7	76.5	56.0	33.3	37.5	68.8	64.7	**76.0**	**62.5**	62.5	82.4	**82.4**
(1-5)	52.0	63.6	46.7	46.2	60.0	36.0	44.4	26.7	31.3	50.0	48.0	60.0	40.0	40.0	**60.0**
(1-6)	60.0	66.7	57.1	53.8	63.6	52.0	66.7	28.6	47.4	81.8	**72.0**	**81.8**	**64.3**	**64.3**	**81.8**
ave.	61.3	61.7	57.1	60.7	65.8	50.0	48.7	32.9	51.0	66.2	**61.3**	**62.3**	53.9	60.3	**68.6**

表 7　表 4 の映像作品による評価（D.R.:識別率，P.R.: 精度，R.R.:再現率，単位は%）．

	temporal congruency $\theta = 0.1$					semantic congruency $\theta = 0.2$					temporal and semantic congruency $\alpha = 0.9, \theta = 0.2$				
		congruency		incongruency			congruency		incongruency			congruency		incongruency	
#	D.R.	P.R.	R.R.	P.R.	R.R.	D.R.	P.R.	R.R.	P.R.	R.R.	D.R.	P.R.	R.R.	P.R.	R.R.
(2-1)	60.0	16.7	16.7	73.7	73.7	88.0	80.0	66.7	90.0	94.7	**88.0**	**80.0**	**66.7**	**90.0**	**94.7**
(2-2)	60.0	40.0	50.0	73.3	64.7	80.0	71.4	62.5	83.3	88.2	**80.0**	**71.4**	**62.5**	**83.3**	**88.2**
(2-3)	48.0	36.3	40.0	57.1	53.3	52.0	41.7	50.0	61.5	53.3	48.0	36.3	40.0	57.1	**53.3**
(2-4)	52.0	28.6	22.2	61.1	68.8	68.0	55.6	55.6	75.0	75.0	**68.0**	**55.6**	**55.6**	**75.0**	**75.0**
(2-5)	56.0	46.7	70.0	70.0	46.7	64.0	55.6	50.0	68.8	73.3	**64.0**	**55.6**	50.0	68.8	**73.3**
(2-6)	32.0	44.4	25.0	25.0	44.4	56.0	100.0	31.3	45.0	100.0	**56.0**	**100.0**	**31.3**	**45.0**	**100.0**
ave.	51.3	35.4	37.3	60.0	58.6	68.0	67.4	52.7	70.6	80.8	67.3	66.5	51.0	69.9	**80.8**



(a) 表1の映像作品（アニメーション）



(b) 表2の映像作品（映画）

図 3　α を変動させた時の評価値の変化．

が被験者実験を通して確認された．今後はこれらの問題に取り組むとともに，より多くの評価データから提案手法の有効性を確認していく予定である．

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) Gillet, O. and Richar, G.: Comparing Audio and Video Segmentations for Music Videos Indexing, *Proc. of ICASSP*, Vol. 5, pp. 21–24 (2006).
3) Shiratori, T., Nakazawa, A. and Ikeuchi, K.: Synthesizing Dance Performance Using Musical and Motion Features, *Proc. of IEEE Int. Conf. on Robotics and Automation*, pp. 3654–3659 (2006).
4) 茂出木敏雄: 映像コンテンツ解析による BGM サウンドトラックの自動生成, 電気学会論文誌, Vol. 125, No. 7, pp. 1004–1010 (2005).
5) 矢倉規光, 梶川嘉延, 野村康雄: 動画像に合う音楽の検索手法に関する一検討, 信学技報, Vol. HIP2005-99, No. 479, pp. 109–114 (2005).
6) Lipscomb, S. D.: Cognition of Musical and Visual Accent Structure Alignment in Film and Animation, *Proc. of the Int. Conf. of Music Perception and Cognition*, pp. 309–313 (1996).
7) Bolivar, V.J., Cohen, A.J. and Fentress, J.C.: Semantic and formal congruency in music and motion pictures: Effects on the interpretation of visual action, *Psychomusicology*, Vol. 13, pp. 28–59 (1994).
8) Marshall, S. K. and Cohen, A. J.: Effect of musical soundtracks on attitudes toward animated geometric figures, *Music Perception*, Vol. 6, pp. 95–112 (1988).
9) Boltz, M., Schulkind, M. and Kantra, S.: Effect of background music on the remembering of filmed events, *Memory and Cognition*, Vol. 19, pp. 593–606 (1991).
10) Lu, L., Liu, D. and Zhang, H. J.: 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).
11) Synak, P. and Wieczorkowska, A.: Some Issues on Detecting Emotions in Music, *Proc. of Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, Vol. 2, pp. 314–322 (2005).
12) Kang, H. B.: Affective Content Detection using HMMs, *Proc. of the 11th ACM Multimedia*, pp. 259–262 (2003).
13) Wei, C. Y., Dimitrova, N. and Chang, S. F.: Color-Mood Analysis of Films Based on Syntactic and Psychological Models, *Proc. of IEEE Int. Conf. on Multimedia and Expo.*, Vol. 2, pp. 831–834 (2004).