RK-007

事例映像に基づくシーンに対する適応的音楽選択
Adaptive music selection for video scenes based on example videos

金 壯一 † 新田 直子 † 馬場口 登 †
Jangil Kim Naoko Nitta Noboru Babaguchi

1. まえがき

近年，ディジタルビデオカメラ等の普及により，一般ユーザにとって映像編集が身近になっている．映像編集とは素材映像から選択した適切な部分映像を並べ替え，結合したものに，音楽，効果音などの音響や，カットやディゾルブなどのトランジション効果といった要素メディアを付与することで新たな映像を生成するプロセスである．本稿では，映像編集において特に，部分映像列に対する音楽の付与に着目する．現在まで，音楽の付与された映像中の動画と音楽の関係について考察した研究がいくつかあり，例えば Zettl[5] は，動画の明るさや動き等の特性は，曲の明るさやテンポ等の音楽の特性に関係があると述べている．また，岩宮 [6] は，被験者に，動画のみ，音楽のみ，動画と音楽を合成した映像をそれぞれ視聴させ，暗い音楽には暗い映像が適応しているといった視覚と聴覚の相互作用の印象評価について考察している．さらに，Brownrigg[10] は，映画を分析し，戦争映画には悲しい音楽が多い等，映画のジャンルによる音楽の違いについて考察している．このように，動画の特性に応じて，適切な音楽は異なると考えられる．

しかし，動画に適切な音楽を巨大な音楽データベースから選択することは一般ユーザにとって非常に難しく，時間と労力を要する．そこで，自動的に適切な動画と音楽を選択し合成する，映像編集手法が提案されている．古賀ら [8] は，事前に被験者に動画，音楽を視聴させ，各々の印象のイメージ語をアンケートで得た後，同じイメージの動画と音楽に対し高くなるよう設定した一致度に基づき，動画及び画像に対して適切な音楽を選択している．ここでは，一つの動画や画像は印象が一定であると仮定し，適切な音楽を一曲選択することを目的としている．これに対し，Hua ら [9] は，音楽の構造に着目し，一曲の音楽において印象が変化すると考え，まず，音楽をセグメントに分割し，各セグメントに適切な映像をホームビデオから選択することにより，ミュージックビデオを生成している．また，Mulhem ら [7] は，動画と音楽それぞれの特徴量をダイナミクス，ピッチ，モーションの 3 つのカテゴリに分け，ホームビデオ中，印象が一定となる部分映像（以下，シーンと呼ぶ）に対し，各カテゴリにおいてもっとも動画と相関性の高いと思われる音楽を選択している．これらの例に見られるように，一本の映像は印象の異なる複数のシーンで構成されることが一般的であり，各シーンに対して適切な音楽を付与する必要がある．しかし，これらの研究で設定された動画と音楽の適切さの評価基準は，人手で決定したものであるため妥当性に疑問が残る．

そこで本研究では，映像編集の専門家により音楽が付与された映像では，技法の一つとして，各シーンに対して，動画の特性に応じて適切な音楽が付与されていると仮定し，このような事例映像に従った基準に基づき，与えられたシーン列に対し，各シーンの動画特性に適応的に，適切な音楽を選択する手法を提案する．

†大阪大学大学院工学研究科, Graduate School of Engineering, Osaka University

2. 提案手法

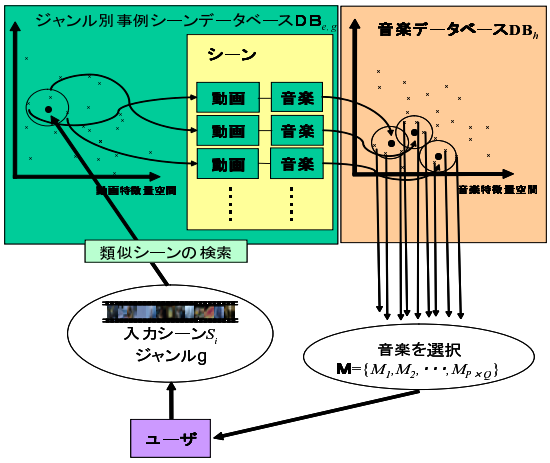

図 1: 提案手法の概要

図 1 に提案手法の概要を示す．まず，映像編集の専門家により音楽の付与されたシーンが事例シーンデータベースとして与えられているものとする．これらの事例シーンに対しては，専門家の技法により，動画の特性が類似したシーンには特性が類似した音楽が付与されていると考える．提案手法は，ユーザからシーン列が与えられると，各シーンに対し動画特性の類似するシーンを事例シーンデータベースから検索し，検索された事例シーンに実際に付与されていた音楽に類似する音楽を音楽データベースから検索し，ユーザに提示する．ただし，適切な音楽は 1 曲とは限らないため，複数の音楽候補を提示する．

ここで，動画特性は色や動きなどの画像の低レベル特徴量により表現するが，同じような特徴量を持つ動画でも音楽により印象を変えることができる．例えば，同じように輝度値が高いシーンであっても，コメディでは明るい音楽，ホラーでは暗めの静かな音楽が付与されていることが多い．そこで，シーンと共に，アクション，コメディ，ホラーといった，作りたい映像のジャンルもユーザからの入力とし，事例シーンデータベースはジャンル別に構成されたものを用いる．また，音楽データベースは，ROCK や POP 等のジャンルを問わない音楽により構成されるものとする．ただし，一曲の音楽においては印象が変化すると考えられるため [9]，曲調の変化点で分割した音楽セグメントとして登録する．

提案手法は以下の 3 つのプロセスで構成される．

表 1: シーン及び音楽の特性と特徴量

	特徴量	特性
動画	HSV(明度)	明るさ
	HSV(色相，彩度)	色合い
	映像乱雑度 [11]	動き
	平均ショット長	速さ
音楽	ZCR(零点交差率)	粗さ
	Spectral Cendtroid	明るさ
	BPM(Beat per Minute)[4]	速さ
	ダイナミックレンジ 音量の標準偏差	騒がしさ

Step 1：特徴量抽出

入力シーン $\mathbf{S} = \{S_1, S_2, ..., S_N\}$ とジャンル g が与えられたとき，S_i の動画特徴量 $\mathbf{F}_{v,\,i}$ を抽出する．ただし，N はシーンの総数である．

Step 2：類似シーンの検索

入力シーンの特徴量 $\mathbf{F}_{v,\,i}$ に最も類似する特徴量 $\mathbf{F}_{v,\,w}$ を持つ P 個の事例シーン $\mathbf{C} = \{C_1, C_2, ..., C_P\}$ をジャンル別事例シーンデータベース $\mathbf{DB}_{e,g} = \{E_1, E_2, ..., E_W\}$ から検索する．ただし，W は事例シーンデータベースに含まれる事例シーン数である．

Step 3：類似音楽の検索

検索された事例シーン $C_p = E_w$ に付与されている音楽の特徴量 $\mathbf{F}_{a,\,w}$ と最も類似する特徴量 $\mathbf{F}_{a,\,z}$ を持つ Q 個の音楽セグメント $\mathbf{M} = \{M_1, M_2, \cdots M_Q\}$ を音楽データベース $\mathbf{DB}_h = \{H_1, H_2, ..., H_Z\}$ から選択し，入力シーン S_i に適切な音楽候補としてユーザに提示する．ただし，Z は音楽データベースに含まれる音楽セグメント数である．

2.1 特徴量抽出

映画のジャンルへの分類 [11] や音楽のジャンルへの分類や音源分類 [12][13] などの研究によると，動画の印象には，動画中の色の情報，動画中の動き，動画の移り変わりの速さ，音楽の印象には，曲の粗さ，明るさ，速さ，騒がしさなどの特性が関連する．提案手法では，各々の特性を表現するため，表 1 に示す動画特徴量，音楽特徴量を用いる．ただし，映像乱雑度 [11] とは，カメラの動きなどによる映像中の全体的な動きと異なる動きをする画素の数であり，映像中の動きのばらつきを表す．また，ダイナミックレンジは，音量の変動の範囲を表す．

事例シーンデータベースの各事例シーン E_w に対して，動画特徴として，HSV 各値の平均と分散，映像乱雑度の平均，平均ショット長の 8 次元の特徴ベクトル $\mathbf{F}_{v,w}$，音楽特徴として，ZCR の平均，Spectral Cendroid の平均，BPM，ダイナミックレンジの平均，音量の標準偏差の 5 次元の特徴ベクトル $\mathbf{F}_{a,w}$ を抽出する．また，入力シーン S_i に対しては，上記の動画特徴ベクトル $\mathbf{F}_{v,i}$，音楽データベース中の音楽セグメント H_z に対しては，上記の音楽特徴ベクトル $\mathbf{F}_{a,z}$ を抽出する．

Input S_i (入力シーン)
$\mathbf{DB}_{e,g}$ ={E_1,E_2,･･･,E_W} (事例映像データベース)
Output $\mathbf{C}$={C_1,C_2,･･･,C_P}
1. for k=1,2,･･･,P
2. w=argmin(D($\mathbf{F}_{v,\,w}$, $\mathbf{F}_{v,i}$))
3. C_k=E_w
4. $\mathbf{DB}_{e,g}$ ← $\mathbf{DB}_{e,g}$-E_w

図 2: 類似シーンの検索

Input $\mathbf{C}$={C_1,C_2,･･･C_P}
$\mathbf{DB}_h$ ={H_1,H_2,･･･, H_Z} (音楽データベース)
Output $\mathbf{M}$={M_1,M_2,･･･,M_β}
1. for p=1,2,･･･,P
2. for q=1,2,･･･,Q
3. ϕ=argmin$_z$(D($\mathbf{F}_{a,z}$, $\mathbf{F}_{a,w}$)) (ただし C_p=E_w)
4. $M_{(p-1)\times Q+q}$=H_ϕ
5. $\mathbf{DB}_h$← $\mathbf{DB}_h$-M_ϕ

図 3: 類似音楽の検索

2.2 類似シーンの検索

入力シーン S_i と動画特性の類似したシーンを P 個，事例シーンデータベースから検索する．図 2 に示すように，ジャンル別事例シーンデータベース $\mathbf{DB}_{e,g} = \{E_1, E_2, ..., E_W\}$ に含まれる事例シーン E_w の映像特徴ベクトル $\mathbf{F}_{v,w}$ に対し，入力シーン S_i の映像特徴ベクトル $\mathbf{F}_{v,i}$ とユークリッド距離 $D(\mathbf{F}_{v,w}, \mathbf{F}_{v,i})$ の最も近い P 個のシーン $\mathbf{C} = \{C_1, C_2, ..., C_P\}$ を検索する．P はユーザが任意に設定することが可能である．

2.3 類似音楽の検索

次に，検索された各事例シーンに付与されている音楽に類似した音楽セグメントを Q 個，音楽データベースから検索する．図 3 に示すように，音楽データベース $\mathbf{DB}_h = \{H_1, H_2, ..., H_Z\}$ に含まれる音楽セグメント H_z の音楽特徴ベクトル $\mathbf{F}_{a,z}$ に対し，検索された各事例シーン $C_p = E_w$ の音楽特徴ベクトル $\mathbf{F}_{a,w}$ とユークリッド距離 $D(\mathbf{F}_{a,z}, \mathbf{F}_{a,w})$ の最も近い Q 個の音楽セグメント $\mathbf{M} = \{M_1, M_2, ..., M_Q\}$ を検索する．すべての $C_p \in \mathbf{C}$ について上記の処理を行うことにより，最終的に，計 $\beta = P \times Q$ 個の音楽セグメントがユーザに提示される．ただし，Q はユーザが任意に設定することが可能である．

3. 実験

3.1 実験概要

事例映像として，複数のシーンの並びで構成され，各シーンに対して専門家により異なる音楽が付与された映画予告映像を用いた．事例シーンデータベースを作成するにあたり，まず映画予告映像をジャンルごとに分類する必要がある．そこで，ジャンルをアクションとホラーに限定し，制作者の異なる映画の予告映像をイン

表 2: 映画予告映像のジャンル及びシーン数

アクション	本数	22
	シーン数	52
ホラー	本数	21
	シーン数	49

ターネットサイトの BLAIRWITCH.DE[3] 等を用いて収集した．各映画のジャンルについては，The Intermet Movie Database[1] を参考とした．ただし，ホラーとしては，ホラー以外にスリラー，サスペンス，サイコ，クライム，スプラッタとタグのついたものも選択した．

次に，各映画予告映像を音楽の切り替えに基づきシーンに分割し，各シーンの動画と音楽をセットとして持つジャンル別事例シーンデータベースを作成した．ここで，事例映像の音ストリームには付与された音楽の他，映像中の環境音，会話等が含まれているため，音楽の特性のみを表現する特徴量の抽出が困難となる．そこで，各シーンに対する音楽として，soundtrack.net[2] を参考に映画のサウンドトラックなどから人手で抽出した，対応する音楽セグメントを用いた．しかし，サウンドトラックに含まれていない，映画予告映像のためのみに作成された，といった理由から入手できなかった音楽については，予告映像の音ストリームから人手で切り出した，映像中の環境音，会話音などによる影響が少ない部分を用いた．

表 2 に事例映像に用いた映画のジャンル及び各ジャンルの映画予告映像の本数とシーン数を示す．各映画予告映像のシーン数は 1〜5 シーンである．

また，音楽データベースとしては，映画のサウンドトラック，クラシック，J-POP など様々なジャンルの音楽を 123 曲収集し，曲調の変化点で人手で分割した後，曲調の違いを考慮して選択した 180 個の音楽セグメントを登録した．

3.2 客観評価実験

まず，事例シーンデータベースから 1 つの事例シーンを取り出し，この事例シーンに付与されていた音楽を音楽データベースに追加する．取り出した事例シーンを入力シーンとし，実際に付与されていた音楽が音楽データベースから選択されるかを検証することで客観評価とする．ただし，$P = 3$，$Q = 3, 4, 5$ とした．またこの時，事例映像データベースには，取り出した事例シーンは含めない．取り出す事例シーンを変更した交差検定法により，以下に定義する正解率により評価した結果を表 3，4 に示す．

$$\text{正解率} = \frac{\text{正しい音楽が付与された事例シーン数}}{\text{事例シーン数}} \quad (1)$$

表より，動画特性が類似した事例シーンに付与されていた音楽に特性が類似した音楽を選択する，という単純な手法で，ジャンルがアクションの映像に対しては，約 60〜70%，ホラーの映像に対しては，55%程度の事例シーンに対し，正しい音楽が選択されており，専門家が制作した事例映像においては，実際に，動画特性が類似した映像には特性が類似した音楽が付与されており，提案手法によりそれを模倣することが可能であるといえる．

表 3: 客観評価結果 (アクション)

P	Q	β	正解率
3	3	9	59.6%(31/52)
3	4	12	65.4%(34/52)
3	5	15	67.3%(35/52)

表 4: 客観評価結果 (ホラー)

P	Q	β	正解率
3	3	9	53.1%(26/49)
3	4	12	55.1%(27/49)
3	5	15	57.1%(28/49)

図 4: シーン検索の失敗例

正解音楽が選択されなかった場合の多くにおいて，類似シーンの検索に失敗が見られた．類似シーンの検索失敗の例を図 4 に示す．図 4 は入力シーン，検索されたシーンをそれぞれ構成する各ショット（一つのカメラにより連続して撮影された部分映像）の最初のフレームを並べている．入力シーンでは，シーン全体を通して暗いショットが多いが，検索されたシーンには明るいショットや暗いショット等がまばらに存在している．このように，ショットごとの彩度や色相は全く違うが，シーンにおける平均値を用いているために特徴量が類似し，特性の異なるシーンが検索されたと考えられる．また，類似音楽の検索においても用いている特徴量が少ないために，特性の異なる音楽が検索される場合が見られた．よって今後は，音楽特徴量としてコード進行などの時間的に変化する特徴量，動画特徴量としてショット単位での特徴量などを考慮する必要がある．

3.3 主観評価実験

次に，主観評価実験として，事例データベースに含まれる映画予告映像 3 本について，音楽データベースから選択された音楽を付与した MOVIE1 と音楽データベースから選択した類似度が低い音楽を付与した MOVIE2 を 11 人の被験者に提示し，どちらが適切な音楽が付与されているか選択させた．ただし，MOVIE1 を生成する際には，映画予告映像を構成するシーンを一つずつ事例シーンデータベースから取り出し，提案手法により音楽を選択した上で，すべてのシーンを連結させた．また，$P = 3$，$Q = 3$ とし，各シーンに対して選択された 9 個

表 5: 主観評価結果

	MOVIE1	MOVIE2
Star Wars: Episode III	100%(11/11)	0%(0/11)
Darkness	90.9%(10/11)	9.1%(1/11)
Aeon Flux	90.9%(10/11)	9.1%(1/11)

図 5: アクション映画のシーン例

の音楽セグメントから編集者が適切と思われるものを 1 つ選択した．表 5 に被験者の評価結果を示す．

表 5 より，被験者のほぼ全員が提案手法により選択された音楽の方が動画に適切であると判断したことが分かる．しかし，どちらの映像も遜色ないという意見もいくつか見られた．これは，本来入力シーンに含まれているはずの会話音，環境音などを映像から除去したためと考えられる．今後は，会話音，環境音を含む映像を用いた実験による考察も必要である．

また，入力となる映画予告映像とは別ジャンルの事例シーンデータベースを用いて音楽を付与した映像 MOVIE3 と MOVIE1 を被験者に提示し，どちらが元のジャンルの映像として適切な音楽が付与されているか選択させたところ，すべての被験者が MOVIE1 を選択した．図 5 に入力としたアクション映画のシーンの一例を示す．アクションの事例シーンデータベースからは曲調が速い音楽が多く選択され，ホラーの事例シーンデータベースからは曲調のゆったりした音楽が多く選択された．この結果から，同じ特性を持つ動画であってもジャンルによって適切な音楽は異なり，ジャンルに応じた音楽選択を実現する上で，事例シーンデータベースをジャンル別に用意することが重要であるといえる．

4. まとめ

本稿では，専門家により音楽が付与された映像を事例映像とし，事例映像に基づき，シーンの動画特性に応じて適切な音楽セグメントを選択する手法を提案した．ジャンルをアクションに限定した映画予告映像を事例映像とし，映画予告映像から抽出した事例と異なるシーンに対し，約 180 個の音楽セグメントが登録された音楽データベースから，実際に付与されていた音楽が選択されるか実験した結果，適切な音楽セグメントの候補数を 9 個にした場合，59.6% のシーンに対し正しい音楽が選択された．また，ジャンルをホラーに限定した場合も同等の結果が得られた．この結果から，専門家により制作された映像においては，動画の特性と音楽の特性に関連があり，これらの事例映像に基づいて，映像に音楽を付与することにより，専門家の技法を模倣することができるといえる．また，実際に提案手法により選択された音楽，及び類似度の低い音楽を付与した映像をそれぞれ提示したところ，90% 以上の被験者が提案手法により選択された音楽のほうが適切と評価した．この結果から，事例映像に基づいて選択した音楽は，主観的に見ても動画と適合しているといえる．しかし，提案手法の要素技術である類似シーン検索，類似音楽検索の精度はまだ低いと考えられる．今後は，より精度を上げるために，シーンを構成するショットの特性のばらつきを考えた特徴量や，音楽のコード進行などの時間的変化を考えた特徴量を用いる．また，今回は専門家の制作した映像を入力映像としたため，ある程度良好な結果が得られた．しかし，ホームビデオ等の一般ユーザが撮影した映像を入力とする場合，専門家が制作した事例映像との品質の差を考慮した特徴量についても検討する必要がある．その上で，数千曲の大規模データベースの実験・考察し，また入力シーンに環境音，会話等が含まれている映像を用いた実験も行う予定である．

なお，本研究は科学研究費補助金若手研究（B）（課題番号 20700087）によるものである。

参考文献

[1] “ The International Movie Database, ” *http://www.imdb.com/*

[2] “ Soundtracknet, The Art of Film and Television Music, ” *http://www.soundtrack.net/*

[3] “ Blairwitch.de, Horror Movie Entertainment, ” *http://www.blairwitch.de/index.php?seitenid=1*

[4] ”MixMeister BPM Analyzer, ” *http://www.mixmeister.com/bpmanalyzer*

[5] Herbert Zettl, “ Sight, Sound, Motion: Applied Media Aesthetics, ”Wadsworth Pub Co，1998.

[6] 岩宮眞一郎，“ オーディオ・ヴィジュアル・メディアによる音楽聴取行動における視覚と聴覚の相互作用 ”，日本音響学会誌４８巻３号, pp.146-153, 1992.

[7] Philippe Mulhem, Mohan S.Kankanhalli, Ji Yi and Hadi Hassan, “ Pivot Vector Space Approach for Audio-Video Mixing,” *IEEE Multimedia 2003*, Vol.10, No.2, pp.28–40, 2003.

[8] 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，Vol.23, No.59, pp. 25–32, 1999.

[9] Xian-Sheng Hua, Lie Lu and Hong-Jiang Zhang, “ Automatic Music Video Generation Based on Temporal Pattern Analisis, ”*ACM International Conference on Multimedia*, pp.472–475, 2004.

[10] Brownrigg Mark, “ Film Music and Film Genre, ”*Doctoral Dissertation, University of Stirling.*

[11] Zeeshan Rasheed, Yaser Sheikh and Mubarak Shah, “ On the Use of Computable Features for Film Classification, ”*IEEE Transactions on Circuits and Systems for Video Technology*, Vol.15，pp.52–64, 2005.

[12] Martin F.Mckinney and Jeroen Breebaart, “ Feature for Audio and Music Classification, ”*International Society for Music Information Retrieval 2003*, pp.151–158, 2003.

[13] Changsheng Xu, N.C.Maddage, Xi Shao, Fanq Cao and Qi Tian,“ Musical Genre Classification Using Support Vector Machines, ”*Acoustics，Speech，and Signal Processing 2003*, Vol.5, pp.429–432, 2003.