音楽および動画における特徴構造のマッチングによる自動構成

飯塚 太郎† Yue Yonghao† 土橋 宜典‡ 西田 友是†
東京大学† 北海道大学‡

1. はじめに

映画やテレビなどのマルチメディアコンテンツにおいて、コンテンツをより印象的なものにするため、映像と音楽とが組み合わされている。心理学的見知によると、一般的に音楽と映像の調和は、両者の時間的な変化箇所の一致によると考えられている[1]。音楽には音の強弱や高低から感じとれるリズムがあり、そのリズムに映像が同期していると、人間は心地良いと感じる。特にミュージックビデオやアニメのオープニング・エンディング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いる。

これら音や映像において変化の大きくなる箇所を端点とする区間を特徴構造と呼ぶ。映像と音楽とを単純に組み合わせるだけでなく、両者が変化する箇所を一致させること、すなわち特徴構造をマッチングすることで、そのコンテンツの印象をより高めることができる。

従来のコンテンツ作製においては、動画を部分的に追加切除するマッチング操作を人間が手作業により行っていたため、データを編集・管理する技術者にとって大きな負担となっていた。そこで本研究では、この負担を軽減するため、動画の追加切除を行わず、動画の再生速度の調節を行うことにより、動画内容を維持したまま両者を同期させる手法を提案する。本研究は、映画やアニメ、ミュージックビデオ等のコンテンツの創作支援に応用が期待される。

2. 音楽・映像の特徴構造の検出

本稿では、ミュージックビデオやアニメのオープニング・エンディングビデオのように、音楽の特徴構造を基準とし、映像のカットや色調の大きな変化といった特徴構造を同期させることにより編集を行うようなコンテンツを対象とする。それぞれの特徴構造を音信号のエネルギー変化の時間微分、および、映像の各フレームにおける RGB ヒストグラム変化の時間微分から検出した。

2.1. 音楽情報からの特徴構造の検出手法

様々な音楽において、楽器はリズムのタイミングに合わせて発音させる場面が非常に多い[2]。楽器による発音時刻が音楽の時間的な特徴を強く表していることを意味している。

ある音が発せられたとき、その音に対応する周波数のパワースペクトルは増加する。全周波数を同時に考慮したパワースペクトルについて、時間軸方向変化分を求めることにより、特徴構造の検出を行う。

図 1 に検出の概要図を示す。図 1 では紙面奥および手前方向に時刻と周波数の軸を、上方向にパワースペクトルの軸を示している。まず、音楽信号を高速フーリエ変換し、周波数軸にて離散化したある周波数 f について、時刻 t および 1 タイムステップ前の時刻 $t-1$ におけるパワースペクトル $p(t,f)$、$p(t-1,f)$を算出し、変化分 $d(t,f)$を式(1)より求める。

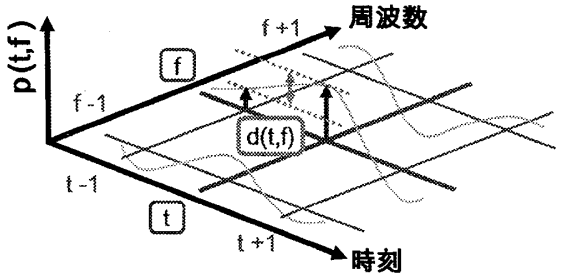

図1　パワースペクトルの変化分の算出

$$d(t,f) = p(t,f) - p(t-1,f) \qquad (1)$$

次に、式(1)の計算を全ての周波数ついて行い、全周波数領域における時間増分$D_{music}(t)$ を式(2)により算出する。

$$D_{music}(t) = \sum_f d(t,f) \qquad (2)$$

そして、ある閾値よりも大きくなる時刻を検出し、このときi番目に検出されたものをt_iとする。

2.2. 動画情報からの特徴構造の検出手法

人間の視覚においては、色や明るさが重要な要素を占めていることから、映像中のカットの切り替わりや、カット以外の色彩や明るさの大きな変化を動画における特徴構造として検出する[3]。

まず、動画中の j および $j-1$ 番目のフレームにおける全ピクセルについて、RGB 各要素を調べ、両フレームのヒストグラム(RGB 各 256 階調)を計算する。式(3)より、両ヒストグラムの RGB 各要素 $X_r(j)$、$X_r(j-1)$ の差分を求め、ヒストグラム変化 $D_{histo}(t)$ を全フレームについて算出する。そして $D_{histo}(t)$ について 256 階調の総和 $D_{video}(t)$ を式(4)より求める。

図 2 に検出の概要図を示す。横軸は時刻、縦軸は

Automatic composition by the matching of the characteristic structure in music and animation: Taro Iizuka†, Yue Yonghao†, Yoshinori Dobashi‡, Tomoyuki Nishita†
(† University of Tokyo, ‡ Hokkaido university)

$D_{video}(t)$ を表しており、$D_{video}(t)$ がある閾値よりも大きくなるフレームを検出し、このとき k 番目に検出されたフレームの時刻を t_k' とする。

$$\mathbf{D}_{histo}(j) = \sum_{(y=R,G,B)} \left| \mathbf{X}_y(j) - \mathbf{X}_y(j-1) \right| \quad (3)$$

$$D_{video}(j) = \sum \mathbf{D}_{histo}(j) \quad (4)$$

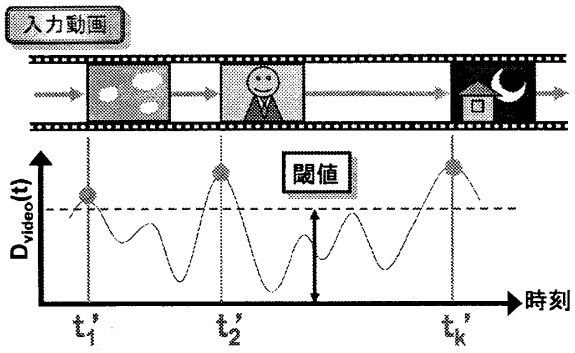

図2　動画特徴構造の検出

3. 音楽・映像の特徴構造のマッチング

動画再生速度の調整により、動画全てを用いたまま、動画の特徴構造を音楽の特徴構造に一致させたコンテンツの構成を行う。

まず、検出された音楽と動画それぞれの特徴構造の中から、マッチングさせるものを決定する。音楽と動画とで特徴構造の数が異なる場合、隣り合う動画の特徴構造の中で、音楽の特徴構造に対し時間的に最も近いものを用いる。

図 3 に特徴構造のマッチングにおける概要図を示す。横軸は時刻を示しており、基準となる音楽の特徴構造の開始時刻 t_i について、その前後一定時間内に存在している動画の特徴構造の中で、最も近接しているもの(開始時刻 t_k')をマッチングさせる。

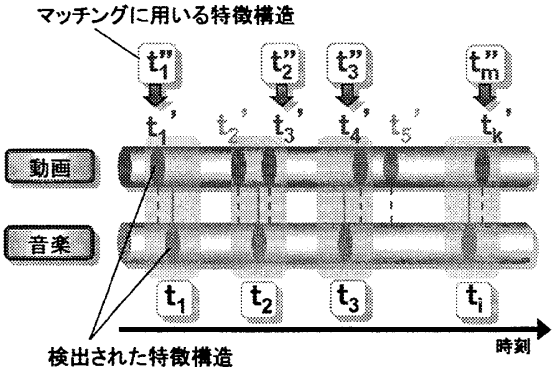

図3　特徴構造のマッチング

選び出された動画の特徴構造の開始時刻 t_k' (濃い文字で表記)を順番に t_m'' としたとき、t_m'' および t_{m-1}'' 間の時間、そして、t_i および t_{i-1} 間の時間を求め、その区間における動画の再生速度の比率 S を式(5)により計算する。

$$S = \frac{t_m'' - t_{m-1}''}{t_i - t_{i-1}} \quad (5)$$

そして算出した速度にて動画を再生し、音楽と合成することによりコンテンツを構成した。

4. 結果

提案手法の有効性を確認するために、入力音楽と動画を単純に重ね合わせた場合と、マッチングを行った場合とについて、音楽と動画の特徴構造が同期し、調和度の高いコンテンツになっているか、被験者による評価実験を行った。

入力動画中に含まれる色彩の大きな変化やカットの数に対し、音楽に調和していると感じられる特徴構造がいくつあるかをカウントし、マッチングの前後における一致度(%)を図 4 に示す。

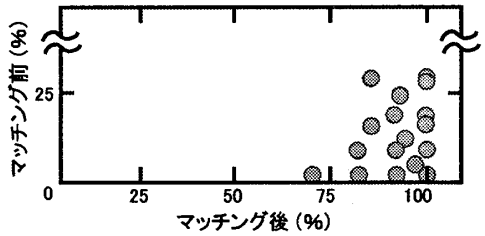

図4　マッチング前後における一致度

マッチング前では最低 0%、最高で 30%の一致度だったのに対し、マッチング後では最低 72%、最高 100%と、全体的に 70%近い向上が確認された。

しかし、動画の特徴構造についてはヒストグラムの変化に注目して検出を行っていたため、物体の動きによる変化はあるがヒストグラムの変化がない場合、例えば、ダンスなど人物の動きに対しては、音楽の同期は未解決であった。

5. まとめと今後の課題

動画および音楽における特徴構造のマッチングを行うことにより、フレームの追加切除を行わず全ての動画を用いて、両者が調和したコンテンツを作製した。今後の課題として、映像中における人物などの大きな動きに対し、特徴構造の検出を可能にすることへ向けて、検出法の最適化などが挙げられる。また、あらかじめ音楽に付加することを前提として用意された動画を対象としていたが、複数の動画のデータベースから、音楽の特徴構造に似た箇所を選び出し、自動でコンテンツの構成を行うシステムの実現を試みたい。

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] M. Goto “An audio-based real-time beat tracking system for music with or without drum-sounds” Journal of New Music Research, 30, 2, pp. 159-171

[3] Olivier Gillet, Gael Richard “Comparing Audio and Video Segmentations for Music Video Indexing” International Conference on Acoutsics, Speech, and Signal Processing, 2006, vol. 5, pp. 21-24