E-033

音楽のムード分類結果を利用したホームビデオへの自動 BGM 付与・同期手法

Automatic Home Video BGM Selection and Synchronization Based on Music Mood Classification

小野 佑大† Yudai Ono　石先 広海‡ Hiromi Ishizaki　帆足 啓一郎‡ Keiichiro Hoashi　滝嶋 康弘‡ Yasuhiro Takishima　甲藤 二郎† Jiro Katto

1. はじめに

近年，動画共有サイトが普及し，ホームビデオに BGM を付与して他者と共有するユーザが増えている．より魅力的なコンテンツを制作するために，映画やドラマの様に，動画のムードに合った楽曲を選定する事や，ミュージックビデオの様に，動画と楽曲とを連動させる事が重要である．しかしその作業は，多大な労力を要し，高い編集技術も必要なため，一般ユーザには難しい．そこで，その様なコンテンツの制作を支援する技術が必要である．

従来研究[1,2]では，動画編集ルールに基づき楽曲を選択したり，音楽構造と動画構造を合わせてミュージックビデオを自動生成したりする事で，ユーザのコンテンツ制作を支援している．しかし，これらには，ユーザの付加したいムードの楽曲を選定できないと言う課題がある．また，楽曲のサビ区間などを BGM として使用しているが，それ以上に動画を効果的に演出する区間を付与する事で，より魅力的なコンテンツを生成できると考えられる．

そこで本稿では，この課題に対し，ホームビデオへの BGM 付与の効率化を図るために，音楽のムード分類結果から得た印象語に基づき楽曲を選定し，動画と同期する区間を楽曲から自動抽出する BGM 付与・同期手法を提案する．さらに本手法の有効性を被験者評価実験により検証する．

2. 提案手法

2.1 概要

本手法では，BGM として使用する楽曲を選定するために，音楽のムード分類結果で得た印象語を利用する．そして，ユーザはその印象語を入力し，絞り込んだ楽曲群から選曲を行う．

次に，選択された楽曲から動画を効果的に演出する区間を抽出するために，動画の動きと楽曲の音量の変化するタイミングが一致する区間を楽曲から抽出して動画へ付与する．本手法では，心理学で定義される動画と音楽の調和に関する要因である時間的なアクセントの一致[3]に着目し，動画と調和する区間を楽曲から抽出する．

2.2 音楽のムード分類結果に基づく楽曲選定方法

本節では，文献[4]に基づいて楽曲をクラスタリングし，生成されたクラスタへ印象語を付与する．そして，ユーザが選択した印象語と一致するクラスタを選定し，楽曲を絞り込み，選曲を行う．音楽のムードは心理学で定義される Russell 空間[5]で表現する．これは人間の感情を活性-不活性(Arousal)，快-不快(Valence)の二軸(以下，AV 軸)から成る 2 次元平面上で定義し，その空間上の座標値(以下，AV 値)で印象語を表現する．Arousal は音楽の音量やリズムの強さなど，Valence は音楽の音色やリズム変動などに相関する．

まず初めに，音楽のムードに関連する 29 次元特徴量を抽出し，主成分分析を行う．そして，第一主成分に-1 を乗じたものを Arousal 値，第二主成分を Valence 値として Russell 空間へ楽曲を写像する．次に，AV 値を-1~1 で正規化後，AV 軸に対し階層的に k-means 法を適用し，空間上の各象限に均一にクラスタを生成する．そして，各クラスタの重心に近い印象語でラベル付けする．最終的に，入力された印象語を含むクラスタを選定し，そのクラスタに属する楽曲群から BGM となる楽曲を選択する．

2.3 動画を効果的に演出する BGM 区間抽出法

本節では，動画を効果的に演出する BGM 区間の抽出法について述べる．まず，動画と音楽から時系列の特徴量を抽出する．動画特徴はフレーム内の動きベクトルの総和で計算される Motion Activity を，音楽特徴は信号の二乗平均平方根のフレーム間差分を用いる．

次に，各特徴量の変化を捉えやすくするために平滑化し，時系列解析で利用される特異スペクトル変換[6]によって変化を検出する．これは，時系列のある時点の変化に対して過去と未来の部分系列から特徴抽出をし，その非類似性で変化の度合い(変化度)を求める手法である．

最後に，検出した各特徴の時系列の変化度に対して式(1)で定義される相互相関関数を計算する．

$$(f * g)(\tau) = \sum_{t} f(t)g(\tau - t) \qquad (1)$$

f と g は時刻 t における動画と音楽の時系列の変化度を示し，ラグタイム τ 毎に f を g の先端から終端まで移動させる．そして，(1)の計算結果が高い区間を抽出する．

3. 評価実験

3.1 実験概要

提案手法の有効性を検証するために，2 つの評価実験を行う．実験 1 では，クラスタリング結果に基づく楽曲選定方法を評価する．また，実験 2 では動画を効果的に演出する BGM 区間抽出法を評価する．

3.2 実験 1~楽曲選定方法の評価~

3.2.1 実験方法

本実験では，Russell 空間上の印象語を検索クエリとして Flickr[7]から収集した 500 ファイルの動画と，文献[4]にて使用されている USPOP の 206 曲を使用した．

†早稲田大学　理工学術院

‡株式会社　KDDI 研究所

表 1. 各クラスタが示す Russell の印象語[5](一部抜粋)

クラスタ1	クラスタ2	クラスタ3	クラスタ4
リラックスした	悲しい	楽しい	怒った
優しい	ほろ苦い	明るい	怖い
穏やかな	憂鬱な	嬉しい	イライラした
甘い	冷たい	興奮した	緊迫した
満足した	退屈な	驚いた	不安な

楽曲群に対して，クラスタ数を 4 に設定し，2.2 節に記載される方法で，各クラスタにラベリングした．表 1 に各クラスタに付与された印象語を示す．そして，入力となる動画に，各クラスタから無作為に選定した楽曲のサビ部分を BGM として付与する事で実験用の動画を作成した．これらと BGM のない動画の計 5 種類を 1 セットとし，計 5 セット用意した．

次に実験の流れについて述べる．まず，動画に付与された印象語を提示し，被験者は 1 セットずつ視聴する．そして，各動画から受けるムードが，提示した印象語にどれだけ近いか評価する．その評価基準は，10: とても近い～0: 全く近くない，の 11 段階を用いた．尚，被験者は大学生 19 名である．

3.2.2 実験結果

表 2 に被験者から得た評価値の平均を示す．これを見ると，動画の印象語が含まれるクラスタの楽曲を BGM として用いた動画は，BGM なしの動画と比べ評価値が高い．この結果から，動画の印象語に合った楽曲を選定している事が分かり，本手法は有効であると言える．

しかし，「ほろ苦い」を見ると，動画の印象語を含むクラスタ 1 の評価値よりも，動画の印象語を含まないクラスタ 2 の評価値の方が僅かに高い．実際に，クラスタ 1 とクラスタ 2 それぞれの楽曲の AV 値を見てみると，クラスタ 1 の楽曲の Arousal 値は-0.39，Valence 値は 0.13，クラスタ 2 の楽曲の Arousal 値は-0.5，Valence 値は-0.11，であった．この 2 曲は異なるクラスタに属しているが，Russell 空間上での距離が比較的近く，ムードが類似している可能性が高い．したがって，クラスタ 1 とクラスタ 2 のムードの違いが分かり辛く，被験者によって意見が分かれてしまったと言う事が原因と考えられる．

3.3 実験 2~BGM 区間抽出法の評価~

3.3.1 実験方法

本実験では，実験 1 と同じ動画と音楽を用い，音楽は，動画の印象語が含まれるクラスタの楽曲のみ使用した．そして楽曲のサビ部分を付与した動画と，提案手法で BGM を付与した動画の 2 種類を 1 セットとして，計 5 セット用意した．

次に，実験の流れについて述べる．まず，被験者は 1 セットずつ視聴し，BGM が動画にどれだけ同期しているか評価する．その評価基準は，10: とても同期している～0: 全く同期していない，の 11 段階を用いた．そして全ての動画の評価が終了後，「動画の完成度を高めるために BGM を同期させる事は効果的か」，と言うアンケートを行った．その評価基準は，2: 効果的だ~0: 効果的でない，の 3 段階を用いた．尚，被験者は実験 1 と同様で，手法における順序効果は考慮してある．

表 2. 実験 1 の評価値の平均

提示した印象語	BGMなし	クラスタ1	クラスタ2	クラスタ3	クラスタ4
怒り	5.28	3.89	3.28	4.56	**6.00**
冷たい	5.50	4.44	**4.78**	3.00	2.72
明るい	6.17	5.00	6.22	**7.94**	6.56
甘い	6.11	**7.17**	4.17	4.89	2.67
ほろ苦い	3.00	5.61	**5.39**	2.83	3.83

表 3. 実験 2 の評価値の平均

動画の印象語	サビ部分	提案手法
怒り	6.00	**7.72**
冷たい	5.39	**6.39**
明るい	6.22	**8.11**
甘い	6.28	**7.00**
ほろ苦い	6.28	**6.44**

3.3.2 実験結果

初めに全体アンケートの結果を見ると，19 人中 15 人が効果的であると答えた．この事から，動画と BGM の同期を図る事は重要であると言える．次に，表 3 で示される被験者から得た評価値の平均を見ると，サビ部分を BGM として加えた動画と比べ，提案手法の方が高い評価値を示している．この結果から，提案手法で付与した BGM は動画を効果的に演出できていると言える．

しかし，「ほろ苦い」の様に，評価値に差の無い結果も見られる．実際に「ほろ苦い」の楽曲の AV 値を見てみると，Arousal 値は-0.5，Valence 値は-0.11 で，Russell 空間上の第 3 象限に存在している事が分かった．この様な楽曲は，リズムや音量変化等が小さい．そのため，動画と同期させた場合も被験者がこれらを認知し辛く，評価値に差が出なかったと考えられる．

4. まとめ

本稿では，ホームビデオへの BGM 付与の効率化を図るために，音楽のムード分類結果から得た印象語で楽曲を選定し，動画と同期する区間を楽曲から自動抽出する BGM 付与・同期手法を提案した．そして，被験者による主観評価実験により提案手法の有効性を示した．今後は，BGM 選曲の自動化と BGM 区間抽出法の改善を試みる．

参考文献

[1] Philippe Mulhem *et al.*, "Pivot Vector Space Approach for Audio-Video Mixing," *IEEE Multimedia*, Vol.10, No.2, pp.28–40, 2003.
[2] Foote J *et al.*, "Creating music videos using automatic media analysis," *Proceedings of ACM multimedia*, New York, pp.553–560, 2002.
[3] 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.
[4] 小野佑大ほか, "ホームビデオへの自動BGM付与のための心理学に基づく音楽分類手法," 第72回情処全大, 1T-02, 2010.
[5] J. A. Russell, *Journal of Personality and Social Psychology*, (6).
[6] T. Id´e *et al.*, "Knowledge discovery from heterogeneous dynamic systems using changepoint correlations," *In Proc. SIAM Intl. Conf. Data Mining*, pp.571–575, 2005.
[7] Flickr, available at http://www.flickr.com/