Vol.2012-EC-23 No.21
2012/3/27

音楽動画コンテンツにおける類似性評価尺度の提案

長谷川裕記[†]　　森島繁生[†]

動画投稿サイトに存在する動画群に対する類似度を測る研究を試みた。この類似尺度を提案することにより，二次創作された動画群における関係性を明らかにすることが本研究の目的である。対象として動画サイト内で同一のアノテーション情報をユーザーから付与された動画群を選別する。このデータベース内部の動画群の動画特徴，画像特徴などを比較し，類似動画を作成し，主観評価を行う。主観評価から得られた結果などを用いて，最終的に動画間の類似度を測る尺度を定義した。

Proposal of Music Video Similarity Measuring scale

Yuki Hasegawa[†]　and Shigeo Morishima[†]

We propose the similarity measuring scale of music video. Our purpose is analyzing the relation of videos created by internet website users. We pick up videos annotated same words from website to make database, and calcurate picture and movie features from database. Finally we make similar music video created scenes from database.

1. はじめに

近年，プロモーションビデオとして音楽に合わせた動画を作ることが一般化し，動画と音楽を同時に楽しむ文化が普及しつつある。更にプロのみではなく，インターネット上の動画共有サイトに一般ユーザーが自身で制作した動画コンテンツを投稿する能動的な文化へと成長している。このような文化を消費者生成メディア，CGM（Consumer Generated Media）と称し認知が進んでいる。この作品群は互いに影響し合い，さらに作品が派生し，そのモデルはN次創作と呼ばれている。また，作品の多くは従来のプロモーションビデオ，映像作品などの影響を受け，演出などで作品の再現なども行われている。

関連研究として，音楽に合わせて動画が作られるという文化を背景に，その関係性を分析する試みが．岩宮らによって行われた 1)．その研究を受け，色彩感覚と音や音楽の関連性を調査した研究 2)-3)があった．また，システムが人間に対して能動的に推薦を行うシステムとして，動画に最適な音楽を推薦する研究 4)や、色彩から音楽を推薦する研究がある 5)。これらの音楽と動画の関係性をモデル化し数値として示す可能性に着目し，室伏らは動画を推薦するだけでなく動画の切り貼りによってダンス動画を生成するシステムを示した 6)．そして自動的に生成されたダンス動画を用いて主観評価を行い，その動画が動画投稿サイトにおける再生数 5000 程度の品質があることを示した．このような音楽と動画の関係性をモデル化し，数値的に扱う際に，その類似度が重要となるが，関連研究では，扱う動画群に実際にどの程度の類似性があるか，またそれらがどのような尺度において測られるべきであるのかという点について調査が不十分である．動画間の繋がりを画像特徴量などによって調査する試みは物体認識などの分野においてなされており 10)-11)，特に動画投稿サイトの動画の繋がりを示す研究として，久保らの研究がある 12)．久保らは SIFT 特徴点軌跡を用い，機械学習などを行わずに動画の類似性を評価したが，背景のない動画において誤認識が見られた．

このような背景から，本研究では，動画コンテンツ群の傾向分析を通じた人間の感性の探求を目的とする。

その手法として，音楽に合わせて創作された動画間の類似度を測る尺度を提案する．その類似性を測る尺度を設けることにより，二次創作における関係性を明らかにする．また類似度の有効性を示すため，類似尺度に基づいて類似動画を探索し動画の切り貼りによって類似動画を作成する．

[†] 早稲田大学
Waseda University

2. 比較手法

本手法において，類似動画を比較する際に，まず動画の条件をそろえ，次に画像内の特徴を比較し，動画の類似性を測る。画像の特徴として，局所の形状，画像の色合いに着目し，特徴量を元に動画を比較する．人間の主観なども反映させるために，類似動画を生成し主観評価実験を行い，最終的に人間の印象を加味した類似動画を生成する尺度を定義する．具体的には、以下に示す手順で行う(図 1).

(1) データベースの動画内のシーンチェンジを検出し，シーン毎にまとめ，これを最小単位にする
(2) シーン毎にカメラの動きを推定する
(3) 楽曲のテンポを推定し，推定されたテンポに基づき一小節間隔でフレームを抽出する
(4) 抽出したフレーム毎に画像特徴量を算出する．本手法において形状を表す特徴量として SURF 特徴量，色合いを表す特徴量としてカラーヒストグラムを算出し，データベースの動画の特徴量空間を形成する．
(5) 比較対象とする入力動画に対しても同様の処理を行う．
(6) 算出した画像特徴量に基づき，シーン毎に動画の類似度を測る．
(7) 入力動画の各シーンと最も類似したシーンを切り貼りすることによって，形状に基づく類似動画，色合いに基づく類似動画を生成する．
(8) 生成した 2 つの類似動画に対し主観評価を行い，主観評価結果に基づき，人間の印象を加味した類似動画を生成する．

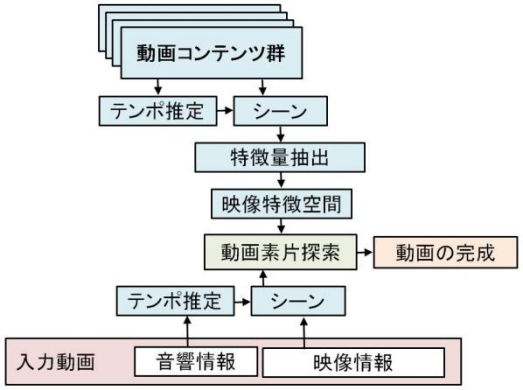

図 1　本研究の手法のフロー

2.1 データベース取得

システムにおいて使用する動画群を取得する。本研究では動画群取得の際に注目する際の条件として、以下の 3 点を設けた.
・動画コンテンツ内容がダンスを中心に作られた PV を意識したものであること
・動画の編集により構成された内容で各々の動画の関連性が高いこと
・不特定多数の人間から一定の評価を得ていること
これらの条件を満たす動画群として，バンダイナムコゲームスから販売されているゲーム「idol m@ster」,「アイドルマスター　Live for You!」を素材として二次創作された動画が考えられる 14)。このゲームによる動画が多数投稿されているインターネット上の動画投稿サイト「ニコニコ動画」から動画群を取得する 13)。取得に当たり動画の再生数を評価の指標とし，10,000 回以上の再生があり，内容がダンスを中心とした動画である上位 103 件を取得した。また、取得する際にフレームレートを 30fps とし、縦 384 ピクセル，横 512 ピクセルに統一した.

2.2 動画要素抽出

フローの手順にのっとり，まずデータベースのシーンを分け，シーン内に含まれるフレームを一小節区間ごとに抽出する。また，シーン毎にカメラの方向を検出し，シーン毎にカメラの動きを割り振る. このシーンを最小単位として, その類似度を測る.

2.2.1 シーンチェンジ検出

シーンチェンジを判定する手法として, 輝度値ヒストグラムの χ 二乗検定を用いる。フレーム毎に輝度値ヒストグラムを算出し，前後するフレームの χ 二乗検定によって算出した値に閾値を設け，一定の閾値を越えた箇所でシーンチェンジが行ったものとする。また，7 フレーム連続で閾値を越える場合，初めのフレームからシーンチェンジが起こったものとする。本手法において，閾値は 0.05 とする.

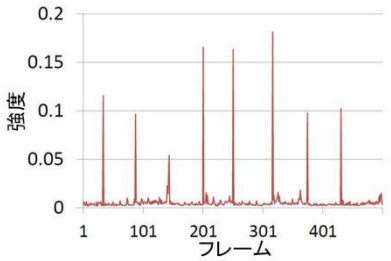

図 2　フレーム毎のショット検出

2.2.2 テンポ推定

データベースに対して行った予備実験において比較的良い結果が得られたため，本システムにおいても先行研究 6)と同様の音響信号のパワーの相関関数に基づく簡易的な方法でテンポ推定を行う。即ち音響信号のパワーの自己相関関数，及びそのパワーと推定されたテンポで生成されたパルス列との相互相関関数を計算し，それぞれピークピッキングによってテンポと小節線の位置を推定する。
まず楽曲の音響信号をモノラル音響信号として，16kHz にダウンサンプリングしてその絶対値を取り, さらに 1kHz にダウンサンプリングして音響信号のパワーに相当する関数 E(t)を得る。それによって時間長 T のパワーE(t)の自己相関関数は

$$R_a(\tau) = \frac{1}{T}\sum_{t=1}^{T}\left(E(t)\cdot E(t+\tau)\right) \quad (1)$$

となる。これを一拍の時間長とする。また，データベースとする楽曲の含まれやすい 100～200bpm をテンポの範囲とすることで倍テンポ，半テンポの誤りをなくす。
一拍毎にピークを持つパルス関数 P(t+τ)と E(t)との相互相関関数 Rc(τ)は

$$R_c(\tau) = \frac{1}{T}\sum_{t=1}^{T}\left(E(t)\cdot P(t+\tau)\right) \quad (2)$$

と計算され，Rc(τ)のピーク時刻は，楽曲中の一拍目の時刻を表わしている。この一拍目を小節線の開始位置とし，4/4 拍子としたうえで小節線を決定した。

2.2.3 カメラの動き検出

動画を取る際に使われたカメラの動きに関して多くの研究が行われており 7)，詳細なカメラワークを認識するためにはそれらの研究を参照する必要があるが，データベースに対して行った予備実験において有効な結果を出した、より簡易的な方法によってカメラの動きを算出する。
まず，動画の時系列順に並んだフレームの内，前後するフレームに着目する。前フレームに対して，後フレームを 45 度毎に右，右上，上，左上，左，左下，下，右下に 1 ピクセルごとに移動した画像，また移動させなかった画像の二乗和（SSD: Sum of Squared Differences）を算出する。本手法において二乗和は画像に対してピクセル単位で算出し，RGB 値を差し引き，その値を二乗し合計する。このうち，二乗和の最も低い方向へカメラが移動したとする。

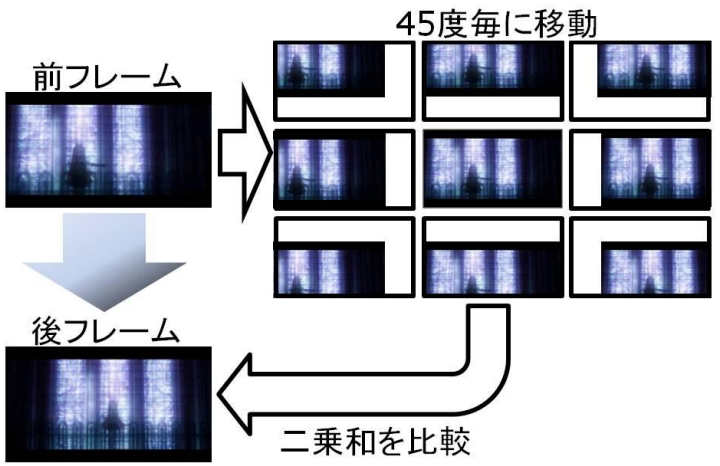

図 3　カメラの動き検出のフロー

2.3 画像特徴量

2.3.1 SURF 特徴量

SURF（Speed Up Robust Feature）特徴量はスケール変化，回転変化に普遍な特徴量で，Haar-wavelet 変換から算出される 64 次元の特徴量である 9)。同様の特徴量として SIFT 特徴量がある 8)。SIFT 特徴量に比べ SURF 特徴量は短時間で算出可能であり，ノイズに頑強であるとして提案されている．本研究に置いて，フレーム毎にヘジアン値 4500 以上の特徴量を用いる。

2.3.2 k-means クラスタリング

本手法において，SURF 特徴量を強度順に算出しているため，その特徴量の総数はフレーム毎に異なる．数百の特徴点を算出した場合，特徴量の次元の総数は何万という単位になり，他フレームとの比較が困難になる．そこで，物体認識の際に使われる手法である bag of features7)-8)にならい，k-means クラスタリングによって，特徴量のクラスタリングを行い次元削減する. Bag of features は言語解析のための手法として成立した後に物体認識に転用されるようになったものであり，その手順として，局所特徴量の集合によって形成される特徴量空間を k-means クラスタリングを用いてクラスタ毎に分け，そのクラスタの中心を visual words として量子化し，その頻度をヒストグラム化する．その結果，画像毎に作られたヒストグラムに物体名などでラベル付け

し SVM などの手法によって学習し，物体認識を行う（図 4）．本手法においては量子化などを行わず，visual words となるクラスタ中心の特徴量をそのまま用いる。また，クラスタ数は 16 とする．

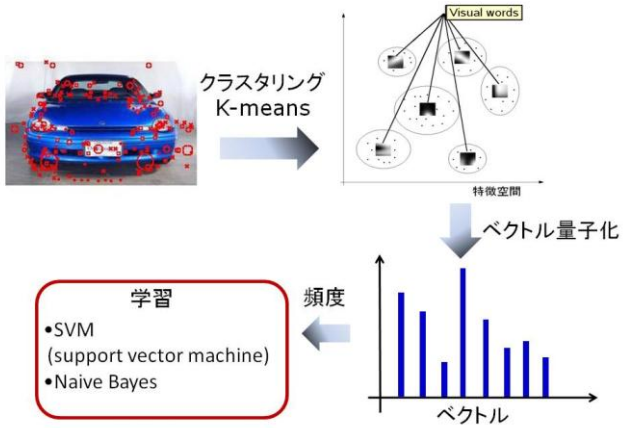

図 4　bag of features のフロー

2.3.3 SURF 特徴量によるフレームの比較法

上記の処理により，フレーム毎に 16 点，64 次元の特徴量が抽出され，次元数は 1024 次元となる．この特徴量を用いてフレーム間の類似度を比較する際に，比較対象のフレームが持つ特徴量ベクトル S_t、データベースのフレームの持つ特徴量ベクトル S のユークリッド距離 h を算出する（式 3）．そして，最も距離の近かった特徴点の距離を可算し，それをフレーム間の類似度の尺度 H とし、最も値の小さいものが最も類似しているフレームとする（式 4）．図 5，6 に例としてデータベースとなる動画群から抽出した 3024 フレーム内において，本手法において最も類似していると判断された画像の対を載せる．

$$h = \sum_{k=1}^{64} (St[b_k] - S[s_k])^2 \quad (3)$$

$$H = \sum_{i=1}^{16} \min h_i \quad (4)$$

図 5　SURF 特徴量による画像探索例 1　左クエリ　右：探索結果

図 6　SURF 特徴量による画像探索例 2　左：クエリ　右：探索結果

2.3.4 カラーヒストグラム特徴量

色合いやその構図を表す特徴量として，カラーヒストグラム特徴量を用いる．本手法ではカラーヒストグラムを得る際に，RGB 値を 64 色に減色し，16 分割した画面からその領域内のカラーヒストグラムを得る。これにより，1 フレームにつき 16 のヒストグラムを得，1 ヒストグラムにつき 64 次元を得て，合計 1024 次元となる．

2.3.5 カラーヒストグラム特徴量によるフレームの比較法

SURF 特徴量比較における式 3 と同様に，フレームごとに算出された特徴量のユークリッド距離を比較することによってフレーム間の類似度を測る．しかし，カラーヒストグラムにおいては，各領域の位置ごとに距離を算出し，その距離を合計したものをフレーム毎の距離とする．図 7，8 に例としてデータベースとなる動画群から抽出した 3024 フレーム内において，本手法において最も類似していると判断された画像の対を載せる．

Vol.2012-EC-23 No.21
2012/3/27

図 7　カラーヒストグラム特徴量による画像探索例 1　左クエリ　右：探索結果

図 8　カラーヒストグラム特徴量による画像探索例 2　左クエリ　右：探索結果

2.4 特徴量比較による類似動画生成

画像特徴量に基づく類似動画探索の有効性を検証するため，データベース動画のシーンを切り貼りすることによって類似動画を生成する．まず比較対象とする動画のシーン毎に，一小節線区間ごとにフレームを抜き出し，2 種類の画像特徴量を抽出する．比較動画のシーンより長く，同じカメラの向きのシーンに限定し，画像特徴量の比較を行う．それぞれの特徴量に基づきフレーム間の類似度 H_f を算出し，時間順にフレーム毎に算出した距離を足し，フレーム間の距離とする（式 5）．それによってシーン毎の距離が算出され，最も距離の短かったシーンを抜き出す．

$$M = \sum H_f \quad (5)$$

3. 主観評価実験

主観評価の詳細を述べる。評価はＡＢ法において行う。被験者は 20 代の男女 13 名で，どちらの手法で合成された動画か区別がつかないよう配慮して声質評価を行った。被験者には「どちらの動画の印象がより元の動画に近いか」という項目によって，A の方が当てはまる，A の方がやや当てはまる，どちらともいえない，B の方がやや当てはまる，B の方が当てはまる，の 5 段階で評価をおこなった。

3.1. 結果

回答として「カラーヒストグラムを用いた類似動画の方が当てはまる」が 13 回答，「カラーヒストグラムを用いた類似動画の方がやや当てはまる」が 22 回答，「どちらともいえない」が 10 回答，「SURF 特徴量を用いた類似動画の方がやや当てはまる」が 4 回答，「SURF 特徴量を用いた類似動画の方が当てはまる」が 3 回答，という結果を得た。

4. 評価実験結果を加味した類似動画の作成

主観評価結果に対し「カラーヒストグラムを用いた類似動画の方が当てはまる」，「SURF 特徴量を用いた類似動画の方がやや当てはまる」，に対して 2 点，「カラーヒストグラムを用いた類似動画の方がやや当てはまる」，「SURF 特徴量を用いた類似動画の方が当てはまる」に対して 1 点，を配し，重みを決定する。カラーヒストグラムに基づいた動画に対して 48 点，SURF 特徴量に基づいた動画に対して 10 点が配された。この得点に基づいて，画像特徴量に対する重みを決定する．本手法において，距離を可算しその距離が短いものを探索する手法を取っているため，より点数の高い特徴量に低い重みをつける算出を行う．総合点数で各特徴量の点数を割った後に、その数で 1 を引いた後の数を重みとする．それによりカラーヒストグラムに対する重み ω_c は 0.172414，SURF 特徴量に対する ω_s 重みは 0.827586 となった。

今回の手法において，画像特徴の 1 フレームに対する次元数は一致するので，この重みを直接フレームの距離に掛けた値を比較することによって動画を形成する（式 6）．

$$M = \sum \omega_s \times H_f + \omega_c \times H_f \quad (6)$$

5. おわりに

本稿では，相互に影響し合い形成される動画投稿サイトにある動画群の類似性を測る尺度を提案し，人間の印象を加味した上で類似動画を生成する試みを示した．今後，音響特徴との関係性なども調査することにより，音楽―動画コンテンツ間の関係性をより深く分析することが可能となるだろう．

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) 西山正紘，北原鉄朗，駒谷和範，尾形哲也，奥乃　博：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，情報処理学会研究報告，2007-MUS-069, pp. 111{118 (2007).
3) 長田典子　岩井大輔　津田学　和氣早苗　井口征士　音と色のノンバーバルマッピング -色聴保持者のマッピング抽出とその応用-電子情報通信学会論文誌. A, 基礎・境界 J86-A(11), 219-1230, 2003-11-01
4) 藤澤隆史，谷　光彬，長田典子，片寄晴弘：和音性の定量的評価モデルに基づいた楽曲ムードの色彩表現インタフェース，情報処理学会論文誌， Vol. 50, No. 3, pp. 1133{1138 (2009).
5) 中西 崇文, 芳村 亮, 北川 高嗣:色彩の印象からの楽曲自動生成方式の実現(マルチメディア, 夏のデータベースワークショップ DBWS 2006), 情報処理学会研究報告. データベース・システム研究会報告 2006(78), 1-8, 2006-07-13
6) 室伏空 中野倫靖 後藤真孝 森島繁生 ダンス動画コンテンツを再利用して音楽に合わせた動画を自動生成するシステム Vol.2009-MUS-81 No.21 情報処理学会研究報告. [音楽情報科学] 2011-MUS-89(15), 1-6, 2011-02-04.
7) Chong-Wah Ngo, Ting-Chuen Pong and Hong-Jiang Zhang "Motion-Based Video Representation for Scene Change Detection" International Journal of Computer Vision 50(2), 127–142, 2002
8) D.G.Lowe, "Distinctive image features from scale-invariant keypoints" International Jornal of Computer Vision, 60, 2(2004), pp.91-110,2004.
9) H.Bay, T.Tuytelaars, and L. Van Gool, "SURF:Speeded-Up Robust Features" IN Proc. of ECCV, 2006
10) J.Yang, Y.-G. Jiang, A.g.Hauptmann, and C.-W.Ngo,"Evaluating Bag-of-Visual-words Representations in Scene Classification" In Proc. Of MIR, pp.197-206, Sep.2007.
11) J. Sivic and A. Zisserman:"Video Google: a text retrieval approarch to object matching in videos," Proc. Int. Conf. Comp. Vis., Vol.2,pp.1470-1477, 2003
12) 久保裕明・斎藤英雄・佐藤真一　SIFT 特徴点軌跡を用いた周辺ヒストグラムによる動画類似度ベクトル量子化を用いない非学習型の動画検索法　電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 110(414), 43-48, 2011-02-10
13) ニワンゴ：ニコニコ動画，http://www.nicovideo.jp/.
14) バンダイナムコゲームス：THE IDOLM@STER OFFICIAL WEB http://www.bandainamcogames.co.jp/cs/list/idolmaster/.