Vol.2013-MUS-99 No.26
2013/5/11

音楽と映像が同期した音楽動画の自動生成システム

平井 辰典[1,2,a)]　大矢 隼士[1,2]　森島 繁生[1,2]

概要：本稿では，著者らがこれまでに研究してきた，任意の入力楽曲を基に既存の音楽動画コンテンツを再利用し音楽と映像が同期した音楽動画を自動生成するシステムを紹介する．本研究では，まずシステムの土台となる音楽と映像の同期手法を主観評価実験により検討した．その結果に基づき，音のエネルギーを表す RMS の変化に，映像のアクセント（明滅や動きなど）を対応させるように既存の音楽動画コンテンツを切り貼りする音楽動画自動生成システムを実装した．また，システムによる生成動画の評価も行った．

1. 研究背景

近年，YouTube やニコニコ動画に代表される動画共有サービスの利用者が爆発的に増えており，動画の視聴，投稿を通じて動画を共有する文化が確立されている．動画全般の視聴者，投稿者が単調増加を続けているなかで，音楽動画コンテンツは最も需要の高い動画ジャンルのひとつとなっている．たとえば，2013 年 1 月の時点で，ニコニコ動画では総動画数の約 23% が音楽に関連した動画であり，その大部分はユーザが制作したコンテンツである．

かつては高価なツールや高度な技術が必要とされてきた動画編集も，近年ではその敷居は低くなってきている．それにより今まで視聴するのみにとどまっていたユーザが自分で新たなコンテンツを制作する機会も増えていった．このようなユーザが生成したコンテンツを指して，CGM（Consumer Generated Media），UGC（User Generated Content）などといった言葉も生まれ，その文化は確立されつつある [1]．なかでも，既存の音楽や動画などのコンテンツを音楽と映像が同期するように組み合わせて制作された動画は「MAD 動画」と呼ばれ，ニコニコ動画を中心とした動画共有サイトで人気を集めている．MAD 動画制作における興味深い現象として，制作された MAD 動画がさらに素材となって新たな動画に利用されるといったコンテンツの連鎖が起きている．既存の動画コンテンツを一次創作物とすると，MAD 動画制作のように既存コンテンツを素材として別の人が 2 次，3 次創作を行うことによるコンテンツの連鎖反応（集団的創造現象）は「N 次創作」と呼ばれている [2]．

CGM 文化は技術の進歩によっていっそう発展し，今まではプロのクリエータにしかできなかったような 3DCG を用いた映像編集が，無償公開されているソフトウェアにより可能となるなど，視聴者がクリエータになるための敷居は下がっており，1 億総クリエータ時代の到来は近づいてきている [3]．しかし，クリエータになるための敷居が下がりアマチュアクリエータの数は増えているが，現状はまだ 1 億総クリエータ時代が到来したとはいい難い．コンテンツ制作への敷居の低下は，創作意欲のある視聴者がクリエータとなることを促進しているが，創作意欲が低い視聴者も含めた**すべての視聴者**がクリエータになるための壁を超えるに至っておらず，1 億総クリエータ時代を前にしての問題点であると考えられる．特に，動画編集では，映像だけでなく，音楽の付与，音楽と映像の同期なども考慮しなくてはならず，制作のためには一定の技術や手間が必要とされており，視聴者がクリエータとなるための第 1 歩への壁は依然として高い．

本研究では，Music Video や Promotion Video などの音楽を主体とした映像作品（以降，音楽動画と呼ぶ）を，任意の動画コンテンツを素材として選ぶことで，専門的な映像編集ツールなどを使わずに自動的に生成するシステムを提案する．それにより，動画編集経験のないユーザや潜在的創作意欲の低いユーザでも，コンテンツを視聴し，好きな動画コンテンツ（素材）を集めるという視聴の範囲内の行動で，動画編集に必要な手間をかけずに新たな N 次創作動画を生成できる．誰もが手軽にコンテンツ生成を体験でき，視聴者の誰もがクリエータとなりうる環境を構築することで，コンテンツ制作の敷居を下げるアプローチとは違ったアプローチで誰もがコンテンツ制作への第 1 歩とな

[1] 早稲田大学
Waseda University, Shinjuku, Tokyo 169–8555, Japan
[2] JST CREST
[a)] tatsunori_hirai@asagi.waseda.jp
本稿は，情報処理学会論文誌 54 巻 4 号に掲載された著者らによる論文「既存音楽動画の再利用による音楽に合った動画の自動生成システム」をまとめなおした研究紹介である．

る体験を得られるようになる．視聴者がクリエータとなるための第 1 歩は重要であり，自動生成結果に不満を持ったり，こだわりを実現したいと考えたりすることが，視聴者がクリエータになるための後押しとなりうる．本研究では，このようなクリエータとしての第 1 歩を支援することにより，誰もがクリエータとなる 1 億総クリエータ時代の到来に貢献することで，CGM 文化の発展を支援することを目標とする．

2. 音楽と映像の同期尺度

音楽動画を自動生成する研究には，いくつかの手法が提案されているが，それらはどれも音楽と映像の知覚的な同期に深く言及していない [4], [5], [6], [7]．本研究では，主観評価実験によって，人が音楽と映像が「合っている」と感じるために必要な音楽と映像の同期手法について考察し，実験結果に基づいた音楽と映像の同期手法を用いて自動的に音楽動画を生成するシステムを構築する．

音楽動画を自動生成するうえで重要となるのは，音楽と映像の同期尺度である．この同期尺度さえ明確になれば，その基準に従って動画を自動生成できる．同期尺度は，音楽動画自動生成システムを構築するうえでの動画の生成ルールとなるため，人の知覚に準じたものである必要がある．音楽と映像の同期を決める基準となる要素として以下のものが考えられる．音楽に関してはテンポやビート，リズムなどのアクセント，映像に関しては画面の明滅やオブジェクトの動きの変化などといったアクセントがあげられる．たとえば，人は音楽のテンポに合わせて手拍子を打ったり，ステージの照明が明滅したりすることで音楽に対する調和を感じ，「気持ちいい」心理状態（情緒的反応）になり，それらのアクセントと音楽のテンポがずれると人は違和感を覚えるといった研究結果が報告されている [8], [9], [10]．また，より広範な音楽と映像の同期にまで言及するために，音楽と映像の時間軸上での調和である「時間的調和」だけでなく音楽と映像のムードの一致による「意味的調和」の両方を考慮した調和度計算手法に関する研究報告もされている [11].

本研究では，音楽と映像の同期をより分かりやすく扱うために，音楽と映像の意味的な調和の要因には言及せず，音楽と映像の時間軸上での調和の実現のみに焦点を置いた同期尺度について検討した．ここで，楽曲の拍節的なアクセントに対して映像の動きのアクセントを一致させると人が同期を感じるという丸山ら [12] や菅野ら [13] による報告を基に，人がさらに同期を感じるような同期尺度について考える．菅野らはこの報告に加え，音列のアクセントの同期（同期要因）と映像の動きの速さと音列のテンポの対応（速度対応要因）の 2 つについて，同期要因の効果の方が大きく，それぞれの要因は独立しているとまとめている [14]．しかし，この実験では，4 分音符に強拍と弱拍を付加したのみのドラム音列しか扱っていない．ここでの音列の拍節的なアクセントの同期とは，音楽の 1 小節に対し，4 つの点（4 拍子の場合）のみにおいて映像のアクセントが付加されている状態を示しており，より詳細なアクセントの変化への言及はなされていない．たとえば 4 分音符のドラム音列の場合とピアノ音列の場合のアクセントでは，拍節的なアクセントの一致を図る場合には両者に同じ映像のアクセントを付加すればよいことになるが，これでは，それぞれの楽器音の減衰の様子などに見られるより詳細な音の変化には対応できない．そのため，楽曲のより詳細な変化を考慮するには拍節的なアクセントの考慮だけでは不十分であると考えられる．

2.1 詳細な音の変化まで考慮した同期手法

音楽と映像において，両者が大きく変化する箇所をマッチングさせると，両者が変化を示す箇所が一致していない動画に比べて音楽と映像の同期の度合いは大きく向上する [15]．そこで，音楽の音の変化に合わせて映像を変化させることで，テンポに基づく拍節的なアクセントに映像を同期させるよりも詳細な，音の変化をも考慮した同期を実現させる尺度を導入する．具体的には，音楽の時間的な変化を表す特徴量の 1 つであり，音のエネルギーを表す RMS（Root Mean Square）に対して，映像のアクセントであるオブジェクトの輝度値とオブジェクトの動きの速さをそれぞれ対応させる．これにより，テンポによって記述可能な 1 拍の長さよりもさらに詳細な時間長での音楽の変化を抽出できる．RMS を表す E は，標本数 n と i 番目の標本値 x_i を用いて，以下の式 (1) のように表せる．

$$E = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \tag{1}$$

2.2 主観評価実験に基づく同期尺度

音楽の RMS と映像のアクセント（映像の動きや明滅）を対応させる同期尺度が人の知覚に準じているかを調査するために，主観評価実験を行った．実験は，テンポに合わせて映像にアクセントを付加する従来の人の知覚に準じた同期手法と，音楽の RMS を用いた提案同期手法を 1 対 1 比較することで行った．

実験に使用した動画は，任意の音楽に対して単純なオブジェクトにそれぞれの同期手法で映像のアクセントを付加することで作成した．まず提案同期手法に基づき，音楽のエネルギーの高さに応じて画面上の単純なオブジェクト（白い正方形の光）の輝度値が変わる映像，オブジェクトが水平方向に動く動きの速さが変わる映像をそれぞれ生成した．同様に，従来の同期手法に基づく映像の生成として，音楽の拍に合わせて映像のアクセント（オブジェクトの明滅，動きの変化）を付加した．

表 1 実験結果（スコア）
Table 1 Result of experiment (score).

使用した楽曲	映像のアクセント			平均値
	動き	明滅	動き＋明滅	
変拍子楽曲 1	4.82	4.82	4.91	4.85
変拍子楽曲 2	4.36	4.64	4.50	4.50
リズム一定楽曲 1	1.77	1.86	1.73	1.79
リズム一定楽曲 2	2.82	2.36	3.41	2.86
リズム一定楽曲 3	3.23	2.86	3.82	3.30
リズム一定楽曲 4	2.64	4.18	2.65	3.16
変拍子ドラム音	4.41	4.55	4.41	4.56
8 ビートドラム音	3.82	2.95	3.23	3.33
平均値	3.48	3.53	3.58	3.53

主観評価実験は，上記の 2 手法で生成した動画を AB 法により 1 対 1 比較することで行った．実験は 20 代の男女 22 名に対して行った．使用した映像は，映像のアクセントとしてオブジェクトの明滅に対応する輝度値の大きさを音楽に対応させたもの，映像の動きの要素として映像の速さを音楽に対応させたもの，さらにその 2 つの要素を組み合わせたものの計 3 種類である．本実験に使用した楽曲は，変拍子の楽曲を含む音楽ジャンルやリズムの異なる楽曲 6 曲とドラムによる単純なリズム音 2 パターンである．各楽曲に対して 2 つの同期手法に基づいて映像にアクセントを付加させた動画の比較を行った．評価は，5 段階の評価項目の AB 法により，提案同期手法を A として，「A の動画の方が合っている」から「B の動画の方が合っている」までのいずれかにより行った．提案手法の方が「合っている」場合のスコアを 5 とし，どちらも同じくらい合っている場合を 3，従来手法の方が合っている場合を 1 となるようにした．また，実験の最後に内観調査として何を基準にして「合っている」と判断したかをアンケート形式で回答させた．

2.3 実験結果

実験結果を**表 1** に示す．スコアは 1 から 5 の値をとり，スコアが 5 に近いほど提案手法の方が合っているといえる．各楽曲のスコアの平均値に注目すると，8 曲中 6 曲の楽曲において，従来の音楽のテンポに合わせて映像のアクセントを付加する同期手法と同等またはそれ以上の評価が得られた．特に，リズムの変化があるような変拍子の楽曲では，本手法に沿って生成した動画の方がより「合っている」という実験結果が得られた．これは，変拍子の楽曲はリズムが一定の楽曲に比べて拍を取るのが困難であることによるものと考えられる．

また，本実験終了後に，被験者が何を基準に「合っている」と判断したかを自由記述形式で回答する内観調査を行った．この結果，「映像が曲のテンポ/リズムに合っているか」といった回答をした人が 22 名中 10 名と最も多く，テンポとの一致の重要性も改めて確認された（**表 2**）．

表 2 内観調査
Table 2 Introspection.

報告内容	報告者数
映像が曲のテンポ/リズムに合っている	10 名
映像が曲の雰囲気に合っている	6 名
映像がドラム音に合っている	5 名
映像がベースに合っている	1 名

以上の結果から，音楽のエネルギーのピークに対して映像のアクセントを一致させる提案同期手法は，人が「合っている」と感じる音楽と映像の同期尺度であるといえる．さらに，人が音楽と映像の同期においてテンポを重要視していることも考えると，テンポを考慮したうえでさらに音楽の RMS と映像のアクセントとの同期を図れば，音楽と映像がより「合っている」同期尺度となると考えられる．

3. 音楽動画自動生成

主観評価実験の結果に基づき，任意の入力楽曲に対して映像を切り貼りすることで音楽と映像が同期している新たな音楽動画を生成するシステムの構築を行った．本システムは，既存の音楽動画コンテンツ群の映像特徴量を抽出してデータベース化する「データベース構築フェーズ」と，入力楽曲を基に，それに合った音楽動画を生成する「音楽動画生成フェーズ」により構成される．

3.1 システム設計

主観評価実験の結果である，『音楽の RMS の変化に映像のアクセント（動き/明滅）を対応させた動画は，人が音楽と映像が「合っている」と感じる』という結果に基づきシステムを設計する．

本システムの処理の流れを**図 1** に示す．まず，データベース構築フェーズではユーザが映像素材として使用したい音楽動画コンテンツ群の映像特徴量の抽出及び，各動画のテンポの推定をする．続いて，音楽動画生成フェーズではユーザが入力した任意の楽曲から RMS を抽出する．次に，データベースの中から，入力楽曲の RMS の推移に最も近い映像特徴量の推移を示す動画を探索する．この際，本研究で定義した音楽と映像のシンクロ率（3.3 節で後述）を計算し，入力楽曲に最も同期する映像素片を探索する．ここで，映像素片を探索する際の各素片の長さは，入力楽曲のテンポに合わせて伸縮される．この映像探索を入力楽曲の全小節に対して行い，選ばれた映像素片どうしをつなぎ合わせ，入力楽曲を貼り付けることで新たな音楽動画が生成される．本システムでは，音響特徴量と映像特徴量を対応させるだけでなく，映像の伸縮によるテンポの一致も図った．これにより，音楽と映像の時間軸上の調和に加えてテンポも同期した音楽動画が生成される．

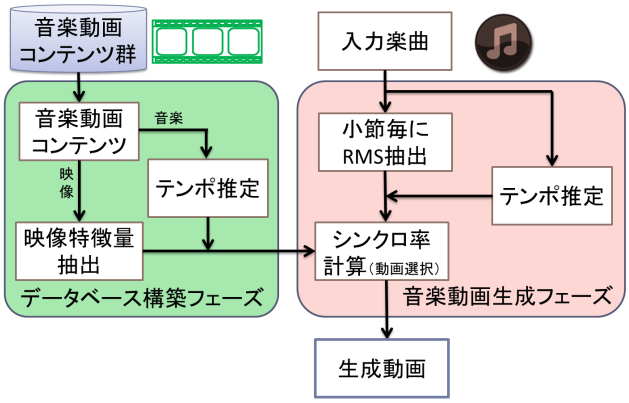

図 1 システムの処理の流れ
Fig. 1 System flow.

3.2 データベース構築フェーズ

データベース構築フェーズでは，既存の音楽動画群から，映像特徴量を抽出する．使用する映像特徴量は，主観評価実験で使用した映像アクセントに基づいたもので，動きの情報である Optical Flow と明滅を表す輝度値である．これらの映像特徴量を全動画の全フレームから抽出する．

3.2.1 映像中の動きの抽出

映像中の動きを表す Optical Flow の値は，前後フレーム間のブロックマッチング法によって抽出した．Optical Flow の抽出は，ブロックのサイズを 5 × 5 ピクセル，ブロックのシフト幅を 2 ピクセルとして映像の前後フレームの全領域における値を求めた．カメラが固定されている場合，Optical Flow の値は動いている物体の領域のみで抽出される．そこで，Optical Flow の値が複数ブロックのまとまりとして現れる箇所をオブジェクトとして定義し，オブジェクトの動きの大きさに注目する．具体的には，Optical Flow の値が現れたブロックが存在するひとまとまりの面積をオブジェクト領域の面積として，領域内の Optical Flow の値の総和を面積で正規化した．これにより，オブジェクトの大小にかかわらず，映像中の動いている物体の速度のみを抽出できる．また，カメラの動きがある場合には，Optical Flow の値は画面全体において現れ，画面サイズと同等のオブジェクトが動く速度が抽出される．これはカメラの動きの大きさに相当する．

これに加え，動き情報において，向きの変化などのアクセントも音楽と映像の同期において重要な要素であるため [8]，速度を時間微分することで加速度を抽出する．これによりオブジェクトの動きの向きや，速さのメリハリが変化する際の映像アクセントを抽出して映像特徴量とした．

3.2.2 映像中の明滅の抽出

映像中の動きの特徴に加え，映像中の明滅の特徴を抽出した．明滅の特徴として，フレームの画面全体の輝度値の平均値を算出した．輝度値の平均値の推移に注目することで，映像中の明滅の特徴が抽出できる．抽出した映像の動き，明滅の各特徴量は，各動画においてその平均値と分散が一致するように標準化する．

3.2.3 映像のテンポ推定

映像特徴量の抽出に加え，データベース中の各動画に対して，動画に付加されている音楽から，その映像のテンポを抽出する．テンポは，ダウンサンプリングした音響信号の包絡線のピークを検出し，ピーク間の距離を基に推定する．ただし，データベース中の動画に音楽が付加されていない場合にはテンポの推定は行わない．

映像のテンポ推定をするうえで，元の動画の音楽と映像のテンポが一致していることが必要となる．データベース動画として音楽と映像のテンポが一致している動画コンテンツを選択することが好ましいが，仮に元の動画の音楽と映像のテンポがずれていた場合でも RMS と映像のアクセントの間の同期には影響を及ぼさない．

3.3 音楽動画生成フェーズ

音楽動画生成フェーズでは，まずユーザが選んだ任意の入力楽曲の RMS を算出する．算出された RMS は，映像特徴量と同様の標準化処理を行った後，テンポ推定を基に 1 小節ごとに切り分ける．ここで，楽曲のテンポと，データベース中の動画コンテンツのオリジナルのテンポの比率に応じて，映像の伸縮を行う．これにより，入力楽曲における 1 小節の長さとデータベース動画における 1 小節の長さを一致させる．これは従来同期手法に沿った音楽と映像の同期にあたるが，提案同期手法は，従来手法を考慮した上でさらに適用できる．それにより，入力楽曲の RMS の推移が乏しい場合など，提案同期手法が効果を発揮しにくい場合でも従来同期手法による音楽と映像の同期がなされていることで音楽と映像のいっそうの同期を図る．

音楽動画生成フェーズでは，これに加え，より詳細な音楽の時間変化と映像のアクセントの一致を図るために，小節中の RMS の推移とデータベース動画の映像特徴量の推移を基に音楽と映像のシンクロ率という概念を定義した．シンクロ率の算出には，入力楽曲の音響特徴量と，データベース動画の映像特徴量との間の相関係数および，値そのものの近さを用いる．シンクロ率は，入力楽曲の 1 小節分の RMS のデータ列$\boldsymbol{x}$と，データベース中の動画 j における前から k フレーム目から 1 小節分の映像特徴量のデータ列$\boldsymbol{y}^{jk}$ との間の相関係数 R^{jk} を用いて，以下の式 (2) で表される．

$$S^{jk} = wR^{jk} + (1-w)\left\{1 - \sqrt{(\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}^{jk}})^2}\right\} \qquad (2)$$

ここで，音響特徴量及び映像特徴量の各データ列の要素数は，映像の伸縮処理及びリサンプリングによって値を揃えている．シンクロ率 S^{jk} において，第 1 項の相関係数 R^{jk} は，$\boldsymbol{x}$と$\boldsymbol{y}^{jk}$ との推移の様子の近さを表す．第 2 項では，音響，映像特徴量の各 1 小節分のデータ列$\boldsymbol{x}$，$\boldsymbol{y}^{jk}$ の平均値

$\overline{\boldsymbol{x}}$，$\overline{\boldsymbol{y}^{jk}}$の差分をとっている．この差分値は$\boldsymbol{x}$と$\boldsymbol{y}^{jk}$ の値そのものの近さを表し，この絶対値を引くことで，相関係数だけでは記述できない値の近さを同期尺度として加えている．それぞれの特徴量は 1 曲ごとに標準化されており，値の近さを考慮することで，局所的な極大，極小値どうしが一致してしまうことが避けられる．音楽と映像のシンクロ率 S^{jk} は，これら 2 つの項に重み w をかけたものである．ここで，重み w は 0 から 1 の間の値をとるが，現在の実装では予備実験の結果 0.5 としている．映像特徴量のデータ列を表す$\boldsymbol{y}^{jk}$ は，速度，加速度，明滅の映像アクセントのいずれかである．アクセントの種類は動画生成の際にユーザが速度，加速度，明滅，または 3 種すべてのいずれかを選択できる．これにより，音楽と動きがシンクロした動画や音楽と明滅がシンクロした動画など，ユーザの選択により異なる動画を生成できる．3 種すべての特徴量を選択した場合には，音楽の各小節において 3 種類のシンクロ率が算出され，その最大値が最も高い映像素片が選択される．シンクロ率の最大値は 1 であり，1 に近いほど両データ列の推移の間に線形の相関があり，さらに両データ列の値そのものが近いといえる．シンクロ率を楽曲の各小節に対し，データベース中の全映像特徴量との間で算出し，入力楽曲の各小節に最も同期した映像素片の探索を行う．

さらに，1 小節ごとに映像が切り替わることで映像への集中が途切れてしまうことを防ぐために，1 度選択された動画素片の続きのシーケンスに重みを付加し，ある閾値以上の S を持つ映像素片が見つかったときにのみ新たなシーンを選択するような処理を加えた．この際，S に付加する重みを時間減衰させることで，シンクロ率の低い映像が長く続くことを防いでいる．

本システムにおけるシンクロ率の算出に基づく映像素片の探索の様子を**図 2** に示す．図 2 下段の「生成動画」より，入力楽曲の RMS の推移と，選択される映像素片の特徴量の推移の様子が近いことが見て取れる．このようにして音楽と映像が同期された音楽動画が自動生成できる．

以上のようにして生成された音楽動画は，主観評価実験の結果を反映した，音楽の詳細な変化に対応して映像が付加された動画であると考えられる．また，本システムの性質上，データベースの動画数が多いほど，入力楽曲の RMS の推移に近い映像特徴量の推移を示す動画の存在確率が高くなり，主観評価実験の結果により忠実な動画が生成されやすくなることが期待できる．

4. 生成動画の評価と考察

4.1 主観評価実験による生成動画の評価結果

本システムにより生成された音楽動画は，音楽の RMS の時間変化に応じて映像中のオブジェクトの動きや明滅の大きさが変わるという主観評価実験の結果を反映した動画

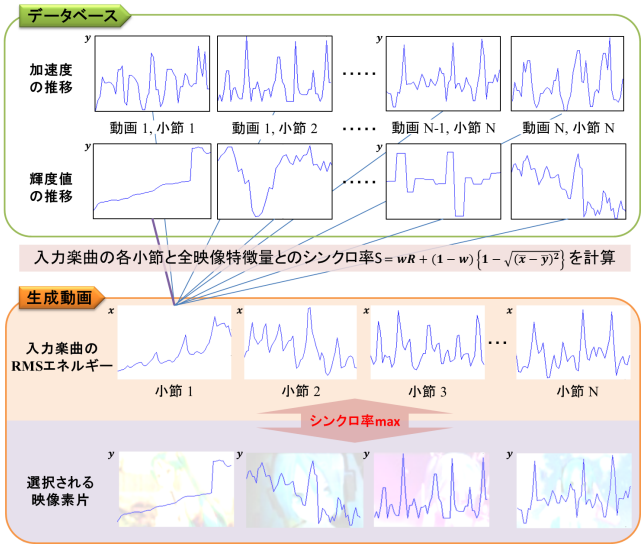

図 2 シンクロ率の計算に基づく動画素片の選択．
Fig. 2 Visualization of selecting music video fragment based on *synchronization rate*.

となった．特に，ドラムの音のように瞬間的なエネルギーの変化が大きい音に合わせて映像中のオブジェクトの動きのアクセントが変化するような音楽動画が生成されやすく，直感的に音楽と映像の同期を感じられる音楽動画が生成できるシステムとなっている．本システムによる生成動画を評価するために主観評価実験を行った．

本システムによる生成動画と音楽に対してランダムに映像を付加することで生成した動画との間で，どちらがより音楽に合っていると感じるかを主観評価実験により判定した．実験に使用したのは，音楽のテンポと映像のテンポが一致していることを確認済みの Michael Jackson のダンス動画 6 作品からなるデータベースと，Lady Gaga の Music Video 6 作品からなるデータベースである．これらのデータベースに対して，「RWC 研究用音楽データベース：音楽ジャンル」[16] の楽曲の中から異なるジャンルの 6 曲をランダムに選択し，1 つのデータベースにつき 3 曲ずつ，計 6 種類の音楽動画を自動生成した．ここで，3.3 節で記述した自動生成の際に音楽に対応させる映像アクセントの種類としては，3 種すべてのアクセントを用いた．さらに，比較用の音楽動画として，同一の 6 曲の楽曲に対して 1 小節ごとに，楽曲と映像の小節長のみが考慮された状態でランダムに映像を付加した 6 種類の動画を生成した．ここで，動画が長すぎることによる評価の困難さを回避するため，楽曲はすべて冒頭の 30 秒のみを用いた．

主観評価実験は，2.2 節で行ったものと同様に，各楽曲に対する 2 種類の方法で生成した動画を AB 法で 1 対 1 比較することで行った．本実験は 20 代の男女 17 名に対して行った．評価は，2.2 節での実験と同様で，本システムにより生成した動画の方が合っている場合のスコアを 5 とし，ランダム生成した動画の方が合っている場合を 1 となるよ

表 3　主観評価実験の結果（スコア）
Table 3　Result of subjective evaluation experiment (score).

楽曲番号	楽曲	スコア
No.3	In Your Arms	3.12
No.9	Waiting for Your Love	2.00
No.13	Guess Again	2.82
No.26	Secret Dreams	3.00
No.30	Kitchen	3.65
No.71	Desperate Little Man	2.76
	平均値	2.89

うにした．評価実験の結果を**表 3** に示す．この結果からは，本システムによって生成された動画も，小節長を考慮してランダムに生成した動画も同等の評価結果となった．17 名の被験者の評価結果の分散は大きく，楽曲ごとの分散の平均値は 1.50 となった．このことから，被験者ごとの評価のばらつきが大きく，本実験による評価結果が有意な結果となっていないといえる．

4.2　考察

本評価実験は，2.2 節で行った主観評価実験と同様の AB 法により生成動画の比較により行ったが，実験に使用した動画は 2.2 節で使用したような単純なオブジェクトが動いたり明滅したりする動画ではなく，複雑なオブジェクトや条件が重なり合った音楽動画コンテンツであった．そのため，被験者ごとに注目するオブジェクトや同期を感じる要因にばらつきが出てしまったことが評価結果のばらつきに影響したものと考えられる．本システムによる生成動画の定量的な評価については今後の課題である．

また，本システムによって自動生成された動画の問題点として，RMS の推移が乏しい楽曲に対して動画を生成すると，音楽と映像の同期の度合いが低くなるなどの問題があった．これに対しては，RMS の推移にメリハリが付加されるような強調処理の追加を検討している．今後，このような自動生成上の問題で生成動画に対する満足度が下がらないようなシステムの改善を加えていく必要がある．

5.　まとめ

本稿では，任意の入力楽曲を基に既存の音楽動画コンテンツを再利用し音楽と映像が同期した音楽動画を自動生成するシステムを紹介した．音楽動画の自動生成は，主観評価実験により検証した音楽と映像の同期手法に基づいて，既存の動画コンテンツの再利用により行い，生成動画は N 次創作物にあたる．また，その生成動画の評価も行った．

音楽と映像を同期させた音楽動画を手動で編集するには一定の技術や手間が必要である．一方，本システムを用いることで，音楽と映像の同期が保障された音楽動画を生成することができる．手動による動画編集ではこだわりを反映させやすいが音楽と映像の同期などの考慮に手間がかかる．本システムの自動生成では，こだわりを反映させにくいが音楽と映像の同期は実現できる．今後，両者の間をつなぐインタフェースを実現するために，本システムにユーザのこだわりを反映させる余地を与えることで，コンテンツの品質，ユーザの望む表現の両面から動画編集を支援できると考えている．

今後，動画のジャンルやユーザの好みに対してより適切な音楽と映像の同期の実現可能性を検討していきたい．それとともに，動画編集インタフェースへの応用などを通じ，本システムを単に動画の自動生成システムにとどまらせず，動画編集経験のない視聴者がクリエータとなり創作活動を行っていくためのきっかけとなるようなシステムとしたい．それにより 1 億総クリエータ時代の到来と CGM 文化の発展に貢献していきたい．

6.　References

[1] 後藤真孝，奥乃　博：CGM の現在と未来:初音ミク，ニコニコ動画，ピアプロの切り拓いた世界 - 編集にあたって，情報処理（情報処理学会誌），Vol.53, No.5, pp.464–465 (2012).
[2] 濱野智史：アーキテクチャの生態系—情報環境はいかに設計されてきたか，NTT 出版 (2008).
[3] 後藤真孝：初音ミク，ニコニコ動画，ピアプロが切り拓いた CGM 現象，情報処理（情報処理学会誌），Vol.53, No.5, pp.466–471 (2012).
[4] Foote, J., Cooper, M. and Girgensohn, A.: Creating Music Videos Using Automatic Media Analysis, *Proc. ACM Multimedia*, pp,553–560 (2002).
[5] Hua, X. , Lu, L. and Zhang, H.: Automatic Music Video Generation Based on Temporal Pattern Analysis, *Proc. ACM Multimedia*. pp.472–475 (2004).
[6] Wang, J., Xu, C., Chng, E., Duan, L., Wan, K. and Tian, Q.: Automatic Generation of Personalized Music Sports Video, *Proc. ACM Multimedia*. pp.735–744 (2005).
[7] Nakano, T., Murofushi, S., Goto, M. and Morishima, S.: DanceReProducer: An Automatic Mashup Music Video Generation System by Reusing Video Clips on the Web, *Proc. Sound and Music Computing Conference*, pp.183–189 (2011).
[8] 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).
[9] 長嶋洋一：音楽的ビートが映像的ビートの知覚に及ぼす引き込み効果，芸術科学会論文誌，Vol.3, No.1 (2003).
[10] 長嶋洋一：音楽的ビートが映像的ビートの知覚に及ぼす引き込み効果（2）—心理学実験システムの開発とレイテンシの計測，情報処理学会研究報告，2003-MUS-51, pp.83–90 (2003).
[11] 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-69, pp.31–36 (2007).
[12] 丸山健夫，安藤明人：音楽と映像のマッチング（1）—テンポと動き，日本心理学会第 60 回大会発表論文集 (1996).
[13] 菅野禎盛，岩宮眞一郎：映像の動きと音楽のリズムの構造的関係が両者の調和感と情緒的印象に及ぼす影響，日本音響学会研究発表会講演論文集，Vol.1999, No.2, pp.559–560 (1999).
[14] 菅野禎盛，岩宮眞一郎：映像と音楽の情緒的印象に対する同期要因と速度対応要因の効果，日本音響学会誌，Vol.56，No.10，pp.695–704 (2000).
[15] 飯塚太郎，Yonghao, Y.，土橋宜典，西田友是：人間の知覚特性を考慮した音と映像の特徴検出および調和の許容時間を考慮したマッチング，情報処理学会研究報告，2008-AVM-63, pp.99–104 (2008).
[16] 後藤真孝，橘口博樹，西村拓一，岡隆一：RWC 研究用音楽データベース：研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情報処理学会論文誌，Vol.45, No.3, pp.728–738 (2004).