Vol.2011-MUS-91 No.20
2011/7/29

アノテーション情報を付加した画像内容推定結果に基づく自動ダンス動画生成システム

長谷川裕記[†]　前島謙宣[†]　森島繁生[†]

本研究では動画に付随するアノテーション情報とユーザーが指定した情報に基き、画像に描写されているターゲット要素の特徴を機械学習することによって、データベース内の動画選択を行い音楽にマッチしたダンス動画を自動生成するシステムを構築した。画像内の輪郭特徴を表す特徴量、アノテーション情報を表す動画コンテンツに割り振られたタグ情報を用いて画像内容推定を行っており、先行研究より画像内の構図を考慮したダンス動画生成ができ、ユーザーがシステムを利用する際の自由度を上げる事が可能となった。

An automatic dance video creation system based on comprehension of image using annotation

Yuki HASEGAWA[†]　Akinobu MAEJIMA[†]
and　Shigeo MORISHIMA[†]

This paper presents a system that automatically generates a dance video clip appropriate to music by segmenting and concatenating existing dance video clips. This system is based on machine learning for annotation and features of image. We create system can consider what object draw in the image, so user can control system more flexible than prior study. Because we use features express shape of object in image, and annotation attended videos in internet to guess what draw in the image.

1. はじめに

本研究の背景は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たことにある。近年では、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、また誰でも手軽に動画を作り創作活動に参加できるシステムを制作することを目的とする。

関連研究として、音楽と視覚における人間の感性を扱う研究として、動画と音楽の関連性を調査した研究[1)]や、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、能動的にシステムが人間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)]。映像を推薦するものとして、ビデオカメラで撮られた映像を音楽の小節線単位で自動的に組み合わせる研究がある[6)]。この研究では一般人の撮影した映像の内、手振れや感光によって不鮮明になってしまった動画の一部を除き、残った動画を分割し組み合わせ音楽に付与しホームビデオを自動的に生成する。このような関連研究を踏まえ、室伏らは音楽と動画の関連性を分析し、その分析結果に基づき自動的にダンス動画を生成するシステムを制作した[7)]。この研究で分析対象としたものはCGMの作品であり、作品を分割、組み合わせることによってダンス動画を生成した。

この先行研究をベースとして、我々は画像内に描かれている物体（以降オブジェクトと呼称する）の大まかな輪郭形状を表す特徴量を追加し、システムに取り入れた。また、先に挙げた関連研究[1)]において、映像と音楽の関係性では、音楽のテンポや演奏の様子と映像の変化の一致（時間的調和）だけでなく、映像に描画される話の内容と音楽の雰囲気の一致（意味的調和）が必要と指摘されている。そこで、動画コンテンツ群に付加されたアノテーション情報も加味することによって、より映像の内容を反映したダンス動画を生成することを目指す。

[†] 早稲田大学
Waseda University

2. 画像内容推定方法

室伏らの研究 [7]において、色相、彩度、明度、画面全体から算出されたオプティカル・フローなどを用いて画像の特徴量空間を形成している。これらの特徴量は画像の構図が反映されないものである。そのため、ユーザーが望む要素に応じて動画を切り変えることができなかった。そこで本研究においては、物体形状を認識し画像中の要素を探る指標となる特徴量として Histogram of Oriented Gradient（HOG）特徴量をシステムに入力することを提案する。この特徴量を注目したい画像の一部の要素から抜き出し、他の画像と比較することによって欲しい要素の入った動画を優先的にダンス動画生成システムに反映することができる。

2.1 Histogram of Oriented Gradient(HOG)特徴量

Histogram of Oriented Gradient 特徴量は、画像内の局所領域における輝度の勾配強度、勾配方向に基づいたヒストグラムを形成して算出される特徴量であり、画像中からの人物検出に用いられている [8)-9)]。類似した特徴量として Scale Invariant Feature Transform(SIFT)特徴量があるが [10)]、SIFT が対象の点に対する特徴量を記述するのに対し、HOG は一定領域に対する特徴量の記述を行う。これによって画像内の注目要素の大まかな物体形状を表すことができる。

HOG 特徴量算出に当たり、まず画像のピクセル数を横方向に 180 ピクセル縦方向に 120 ピクセルとする。また、画像内の 5×5 ピクセルの領域をセルと定義する。

2.1.1 輝度の勾配方向と勾配強度の算出

画像内のピクセルの隣り合う要素を比較し勾配強度と勾配方向を算出する。輝度を L、ピクセル座標を x、y して f を

$$f_x(x,y) = L(x+1,y) - L(x-1,y) \qquad (1)$$

$$f_y(x,y) = L(x,y+1) - L(x,y-1) \qquad (2)$$

と定義した際に、勾配強度 m、勾配方向 θ は

$$m(x,y) = \sqrt{f_x(x,y)^2 + f_y(x,y)^2} \qquad (3)$$

$$\theta(x,y) = \tan^{-1}\frac{f_x(x,y)}{f_x(x,y)} \qquad (4)$$

となる。

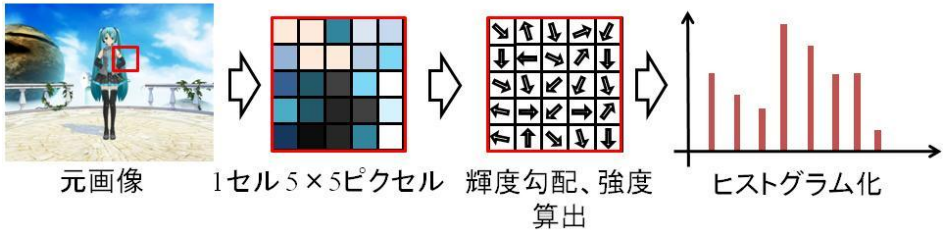

図 1 HOG 特徴量

2.1.2 ヒストグラムの作成

勾配方向を 0 から 180 度まで 20 度づつ区分し、セル内における勾配強度を合計したヒストグラムを作成する。*i* 行 *j* 番目のセルの持つヒストグラムを *Fij = [f1 f2 …f9]* とすると、一画像内においてセルは 864 個存在し、セル内に 9 つの要素を持つヒストグラムを持つことから、一画像中に 7776 次元の特徴量を持つこととなる。

2.2 HOG 特徴量に基づく探索方法

以上の方法によって抽出した特徴量を利用し、指定されたオブジェクトを含む画像を探索する方法について述べる。

2.2.1 オブジェクトの特徴抽出

オブジェクトが入ったセル領域（ターゲットブロックと呼ぶことにする）を長方形型に設定し、縦、横に入るセル数を指定する。このターゲットブロック内に含まれる個々のセルの持つヒストグラムの値を各要素毎に合計し、オブジェクトの特徴量とする。ここで、ターゲットブロックを B とすると構成要素は

$$B[b_k] = \sum F_{ij}[f_k] \qquad (5)$$

となり、ターゲットブロックに含まれる全ての勾配強度の合計 m_s は

$$m_s = \sum(F[s_k]) \qquad (6)$$

となる。

2.2.2 探索ウィンドウの設定

ターゲットブロックと縦横の長さが等しく同数のセルを含む領域を探索ウィンドウと呼ぶことにする。この探索ウィンドウ内のセルの各要素もターゲットブロックと同様に合計する。よって、探索ウィンドウを S とすると要素は

$$S[s_k] = \sum F_{ij}[f_k] \qquad (7)$$

と表される。

2.2.3 HOG 特徴量比較法

ターゲットブロック B と探索ウィンドウ S の各要素の差分をまとめた差分 h は

$$h = \sum(B[b_k] - S[s_k]) \qquad (8)$$

と表され、これを m_s によって割った値 h/m_s が 0.2 を超えない場合、注目するオブジェクトに近い形状をもったが画像に含まれると判断する。なお、0.2 という値は予備実験としてキャラクター探索を行った際に、比較的誤りが少なく該当するオブジェクトが描画された画像を選択できたため実験的に定めた値である。

2.3 動画コンテンツの取得

システムにおいて使用する動画群を取得する。動画群取得の際に注目することとして、

・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

などが挙げられる。

これらの条件を満たす動画群として MIKUMIKU DANCE(MMD)による動画群が挙げられる。MMD とは樋口優氏によって作成された 3DPV 制作ツールであり、樋口氏のホームページ内で無償公開されている [11]。この MMD 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 [12]。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 113 件を取得した。

2.4 画像探索

2.4.1 動画コンテンツからの画像データベース作成

各動画はフレームレート（fps）30fps 前後で再生されるが、連続するフレーム毎の内容の差異は大きくなく、全てのフレームでデータベースを形成すると同じような画像内容の中で探索を行ってしまう。そこでシーンチェンジを全くしない動画を除けば、多くの動画が 10 秒以内に一度はシーンチェンジするため、各動画 10 秒につき 1 フレームの割合で画像を抽出し、合計で 2263 フレームの画像を抽出しデータベースとした。

2.4.2 画像探索結果

着目するオブジェクトとして、手動でキャラクターの体型をターゲットブロックに指定し探索を行った（図 2）。このキャラクターはデータベース中 95 の動画に現れる。結果、38 フレームがデータベースから選ばれ、その内同じキャラクター、vocaloid シリーズ [13]として関連のあるキャラクターなどが 30 フレーム含まれていた。また、同じキャラクターの顔を手動でターゲットブロックに指定した探索も試みた（図 3）。結果、110 フレームがデータベースから選ばれ、18 フレームは顔の含まれる画像であったが、複数の人物を含む画像なども見られた。オブジェクトを含むセル数が増え、ほぼ画面を覆う領域をターゲットブロックとして指定してしまうと、ターゲットブロック内の顔とは違った部分から顔と同じ勾配方向、強度を検知してしまい、全く異なった形状であってもヒストグラムの値が似てしまうことが原因と考えられる。

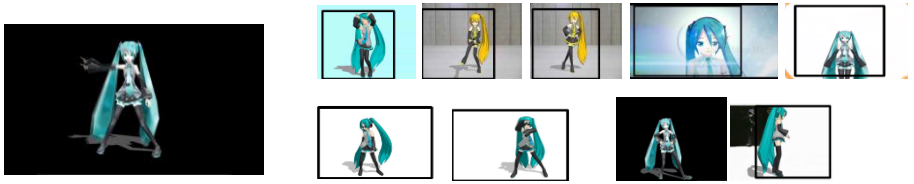

図 2　探索例　左　探索対象 右　探索結果例

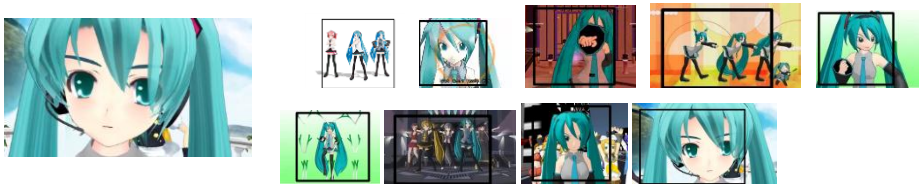

図 3　探索例　左　探索対象　右　探索結果例

2.5 アノテーション情報の付加

探索結果を向上させる方法として、事前にユーザーが注目すると思われるキャラクター、背景などから特徴量を抽出し機械学習を行い、探索精度を上げる方法が考えられる。しかし事前に学習させた場合、学習させたキャラクターや背景のみの探索しか行うことができないため、動画の持つ様々な要素に着目した探索はできない。そこで探索の補助として、動画に付随したアノテーション情報を利用することとする。アノテーション情報として具体的にインターネット上の動画共有サイト「ニコニコ動画」における「タグ」情報を用いることとする。この「タグ」は一般ユーザーが動画に合わせて編集可能な要素であり、動画の含まれるカテゴリーや、内部に写るキャラクターなどの要素を表す言葉を当てはめ、動画共有サイト内のグループ分けや検索を行うためのものである（図 4）。似た機能として、動画の時系列中のある区間を指定して短いフレーズを流す「コメント」機能もあり、音響情報との関連性が調査されている [14]。この研究において不特定多数のユーザーが付ける「コメント」には動画内部の情報を表している妥当性があることが指摘されている。その指定を行うことによって動画の意味的な解釈をシステムに反映することができる。例えば特定のキャラクターが特定の衣装を着た場合に着目したい場合には、キャラクター名を示す「タグ」情報と HOG 特徴量を合わせることで、より高精度な探索を行うことができる。

図 4　インターネット動画共有サイトにおける「タグ」情報

3. ダンス動画群から音楽に合わせた動画を選択するシステム

3.1 システム概要

以上の画像内容推定方法を用い、ユーザーにとって最適と思われる動画を生成するシステムを構築する。動画生成法は先行研究にならい、ダンス動画をデータベースとしその動画素片の選択による。本研究におけるシステムにおいてユーザーから入力が必要なものは

・楽曲情報
・注目したいオブジェクトの含まれる画像とそのオブジェクトの存在する領域指定
・注目したい「タグ」情報

である。まずデータベースの傾向を分析し音響特徴と画像特徴の関係性を分析する。次に楽曲情報から、分析された傾向に基づきその楽曲にふさわしい画像特徴を算出する。次にアノテーション情報による探索する動画コンテンツの選択を行う。最後にユーザーの指定したオブジェクトの写る可能性を HOG 特徴量により算出し、この二つの指標を鑑みた上で動画の素片を選ぶ。(図 5)

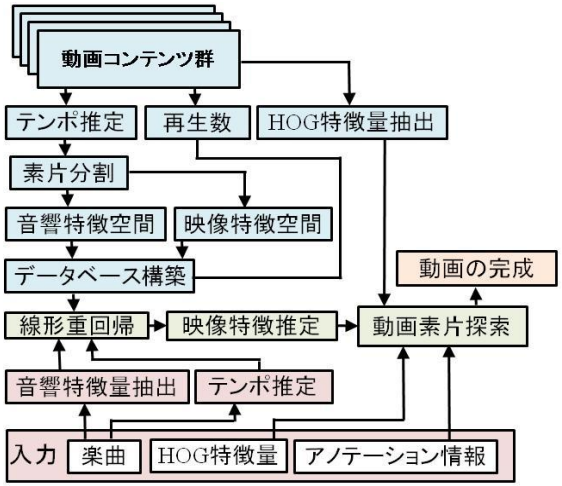

図 5　システムフロー

3.1.1 動画コンテンツの取得

2 章「動画コンテンツ取得」に述べた MIKUMIKU DANCE を用いて作られた動画コンテンツ群をデータベースとする。

3.1.2 楽曲の小節線及びテンポ推定法

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

$$R_a(\tau) = \frac{1}{T}\sum_{t=1}^{T}\big(E(t)\cdot E(t+\tau)\big) \qquad (9)$$

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 60～120bpm 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 $P(t+\tau)$と $E(t)$との相互相関関数 $Rc(\tau)$は

$$R_c(\tau) = \frac{1}{T}\sum_{t=1}^{T}\big(E(t)\cdot P(t+\tau)\big) \qquad (10)$$

と計算され、$Rc(\tau)$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。

3.1.3 音響特徴量抽出

特徴量算出に当たり、ジャンル分類の研究、音楽と動画の関連性を調査した研究[1)2)5)15)]、色彩との関連性を調査した研究などを参考に音響特徴量を抽出する。特徴抽出は、まずサンプリング周波数が 44.1 kHz のモノラル音響信号を、音響信号の振幅が 0.9 となるように正規化してから行った。分析窓のシフト幅は 1470 点 (30 fps) とし、窓幅はシフト幅に合わせて 1470 点とする。アクセントの特徴量としては、主に楽曲のパワーとその短時間での変化を表現するために、フィルタバンク毎の 4 次元のパワー出力(1.- 4.)と Spectral Flux(5.)を用いた。周波数に現れる楽曲の盛り上がり関連する特徴量として、Spectral Centroid(6.)、Spectral Roll-off(7.)を算出した。印象に関する特徴量としては、楽曲の音色に関連した Zero-crossing rate(8.)と MFCC(Mel-Frequency Cepstral Coefficients)の直流成分と低次 12 項、合わせて 13 次元の特徴量を用いた（ 9. - 21.)。以上のように、音楽のアクセント、盛り上がり、印象に関する計 21 次元のフレーム特徴量を抽出した。印象に関する特徴量としては、

表 1　音楽と映像のフレーム特徴

次元	音響特徴量
1.-4.	envelope
5.	Spectral Flux
6.	Spectral Centroid
7.	Spectral Roll-off
8.	Zerocross-rate
9.-21.	MFCC

次元	映像特徴量
1.	オプティカル･フローの微分値
2.	輝度値ヒストグラムの微分値
3.4.	色相の平均と分散
5.6.	彩度の平均と分散
7.8.	明度の平均と分散

3.1.4 HOG 特徴量以外の映像特徴量抽出

映像特徴量には、音楽と動画の関連性を調査した研究 [1]、を参考に、アクセントおよび印象に関する特徴量を決定した（表 1)。アクセントに関する特徴量としては、画面の動きやダンス動作とそれらの時間変化や画面の切り替わりを表現するために、オプティカル・フロー(1.)と輝度値(2.)の時間微分の平均値を用いた（各 1 次元)。オプティカル･フローはブロックマッチング法を用い、ブロック数 64×48、シフト幅 1、最大シフト幅を 4 として計算している。また，画面全体に対して細かい範囲でブロックマッチング法を適用した後，各フレームのオプティカル・フローの全ブロックの値に対してメディアンフィルタをかけることで，局所的なブロックの値をなくし，まとまった物体が動いた場合の加速度が抽出されるようにした．さらに，物体の大小に限らず，注目すべき移動物体の動きの大きさを特徴量として抽出するために，メディアンフィルタをかけた後の移動範囲で加速度の値を正規化した．

印象に関する特徴量としては、映像の雰囲気を表現するために、全画素における色相、彩度、明度、それぞれの値の平均と標準偏差を用いた（全 6 次元）(3. – 8.)。これにより、映像のアクセントと印象に関する計 8 次元のフレーム特徴量を抽出した。なお、前処理として、画面サイズを 128×96 にリサンプリングを行っており、分析窓のシフト幅は、映像のフレームレートである 1 フレーム毎（1/30 s）としている。

3.1.5 小節毎の特徴量のまとめ

小節線内の特徴量の時系列を考慮するために抽出された音響特徴量に DCT 変換を行う。フレーム特徴を 16 点でリサンプリングし、この 16 点に対して DCT をかけ、そのうちの低 4 次元を使用する。これにより音楽特徴量は 84 次元、映像特徴量は 32 次元となる。次に主成分分析を行い、特徴量の直行化、及び次元削減を行う。この時に、累積寄与率は　95%で次元数を削減する。

3.2　動画選択法

ユーザーが指定したオブジェクトの HOG 特徴量とデータベース中の画像の HOG 特徴量の近似値と、データベースから抽出した音響特徴量と映像特徴量の関係性、アノテーション情報から入力楽曲にふさわしい動画素片を選択する。選択に際し、入力されたアノテーション情報と同様の「タグ」を持つ動画素片とする。その動画素片内の選択方法について以下に述べる。

3.2.1 特徴量回帰

入力楽曲の音響特徴から楽曲にふさわしい映像特徴量を推定する。学習は、音楽の小節特徴量を説明変数､映像の小節特徴量を目的変数とした線形重回帰によって行う。回帰にあたり、動画コンテンツ内の内容差を加味するため、再生数による重み付けを行う。重み付けは再生数 s、重み ω とすると

$$\omega = \alpha \cdot \lfloor \log_{10} s \rfloor + \beta \quad (11)$$

という関数で表すこととし、ω を自然数に直した数の特徴量を新たにデータベースに追加することとする。また、動画コンテンツ群の再生数の分布に即した重み付けとするため、$\alpha=1.0$、$\beta=2.0$ と設定する。

また、今回のデータベース内においては、同一の楽曲に対して違った動画を合わせる創作物が含まれる。このような多様な音楽の解釈がある動画素片に対しては、音響特徴量と映像特徴量に対して一対一の対応関係があるとは言い難い。これにより特徴量の近似する要素はまとめられ､矛盾が少ないデータベースとすることが可能となる。この動画群の特徴量のクラスタリングする方法として、特徴量空間のユークリッド距離に基づく *k-means* クラスタリングを行う。この際にクラスタ数は 10 とする。以上の方法によって、サンプリング周波数 f の入力楽曲の n 番目の小節線 i_n から推定された映像特徴量を $\nu_n(i_n,f)$ を算出する。

3.2.2 コスト設定及び探索動画指定

選択する際に選ぶ基準となるコストを定義付け、このコストを最小化するという形で動画コンテンツ群の動画素片の選択を行う。

実際のデータベース中の楽曲 m の t 番目の素片 d の持つ映像特徴量 $\nu(d_{(t,m)},f)$ との距離 $d(i_n,d_{(t,m)})$ を

$$d(i_n, d_{(t,m)}) = \sqrt{\sum_f \left(v(d_{(t,m)}, f) - v_n(i_n, f)\right)^2} \quad (12)$$

とする。次に動画素片内のフレームにおいて、2 章 2.2「HOG 特徴量に基づく探索方法」で述べた HOG を比較する方法によって算出した HOG 特徴量の差分 h を算出する。探索された小節の持つフレーム内で、最も小さい差分 h を $h(i_n,d_{(t,m)})$ とし、選ばれた素片 i_n までの累計コスト $c(i_n,d_{(t,m)})$ を

$$c(i_n, d_{(t,m)}) = \begin{cases} d(i_n, d_{(t,m)}) + h(i_n, d_{(t,m)}) + c(i_{n-1}, d_{(\tau,\mu)}) & if \mu = m, \tau = t-1 \\ p_c \times d(i_n, d_{(t,m)}) + h(i_n, d_{(t,m)}) + c(i_{n-1}, d_{(\tau,\mu)}) & otherwise \end{cases} \quad (13)$$

として、素片の時系列を加味し、選ばれた素片の次に当たるものが選ばれやすいような設定を行う。また、p_c=5.0 と設定する。
累積コストは最終小節 N において最も累積コストが小さい素片 d_{min} を式(14)でもとめた後、バックトレースによって得る。

$$d_{min} = arg\min_{t,m} c(i_N, d_{(t,m)}) \qquad (14)$$

4. おわりに

本研究において、先行研究の自動的にダンスを生成するシステムを踏まえ、より画像内容の構造を反映できるよう新たな画像特徴量を追加し、ダンス動画自動生成システムを作製した。また、動画選択時に人為的に動画コンテンツ群に割り振られたアノテーション情報も加味することによって、更に画像内容推定の精度を上げることが可能となった。

今回のシステムにおいては最初にユーザーが設定する要素として音楽、キーワードの他に注目したいキャラクターや背景などの要素を手動で入力する必要があり、キーワードも動画に付随した中の限られた要素に限られた。

ユーザーの求める更に抽象的なキーワードを入力するだけで欲しい動画を作製するためには、アノテーション情報と画像内に描画された特徴を学習する仕組みが必要となる。そのためには、現在より更にアノテーションの言語的な構造や、画像に描画される要素を自動的に追う特徴量の利用方法を考える必要がある。今後このような点を踏まえ、更に手軽に求める動画を作ることが可能となるシステムを作製していきたい。

参考文献

1) 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).
2) 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), 1219-1230, 2003-11-01
3) 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).
4) 茂出木 敏雄:映像コンテンツ解析による BGM サウンドトラックの自動生成 電気学会論文誌. C, 電子・情報・システム部門誌　125(7), 1004-1010, 2005-07-01
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) Foote, J., Cooperand, M. and Girgensohn, A.: Creating music videos using automaticmedia analysis, Proceedings of the tenth ACM international conference on Multimedia,pp. 553{560 (2002).
7) 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.
8) N.Dalal, B.Triggs,"Histograms of oriented gradients for human detection",Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),pp.886-893,2005.
9) 大戸 和博　土肥 慶亮　柴田 裕一郎 [他] : HOG 特徴と AdaBoost による人検出処理の FPGA への実装 (リコンフィギャラブルシステム)　電子情報通信学会技術研究報告 110(362), 117-122.
10) D. G. Lowe, “Object recognition from local scaleinvariantfeatures”, Proc. of IEEE .InternationalConference on Computer Vision (ICCV), pp.1150-1157, 1999.
11) Vocal Promotion Video Project　http://www.geocities.jp/higuchuu4/index.htm.
12) ニワンゴ：ニコニコ動画，http://www.nicovideo.jp/.
13) Vocaloid http://www.vocaloid.com/.
14) 吉井和佳, 後藤真孝 :MusicCommentator:音楽に同期したコメントを自動生成するシステム 情報処理学会研究報告，Vol.2009-MUS-81 No.20
15) Gillet, O. and Richard, G.: Comparing Audio and Video Segmentations for Music VideosIndexing, Proceedings of IEEE International Conference on Acoustics, Speech and SignalProcessing (ICASSP 2006), pp. V{21{V{24 (2006).