動画像に対応した背景音楽の自動生成システム

5Z－2　　尾上　直之　張　汀汀　橋本　周司　田中　章喜
早稲田大学理工学部　　松下技研株式会社

1．はじめに

近年、テレビ・ビデオ等のAudio Visualメディアの充実にともなって、われわれは映像を見て、同時に音を聞いている。音のない映像を見ると、違和感さえ感じるものである。背景音楽（ＢＧＭ）の多くは、作曲家によって作曲されており、簡単なＢＧＭについても製作の自動化はほとんど行われていない。

我々は、動画像から背景音楽を自動生成するシステムの制作を行っている[1]。動画像から単純な構造的特徴を抽出し、これらをリアルタイムで音楽的特徴と結び付け、適当なメロディーを付加するシステムである。BGMを生成する試みは、これまでにも報告されている[2]が、ここでは画像の内容に立ち入らず物理的な特徴のみに注目し、最終的に画像受信側に小さなアダプタを付加して、個人の好みに合わせた音楽を生成することを目的としている。本稿では、システム構成の概要とサンプル動画像を用いた音楽生成の結果を報告する。

2．システム構成

2－1．システムの概要

システムは、図１のように画像処理・統合処理・音楽処理の３つの部分により構成される。

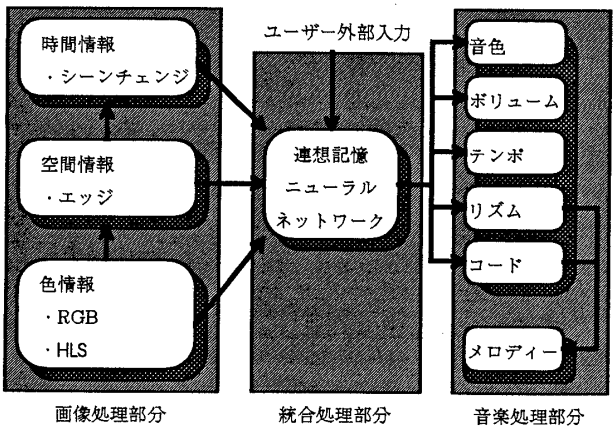

図１　システムの概要

Background Music Generation System Based on Scene Analysis

Naoyuki Onoe, Dingding Chang, Shuji Hashimoto, Akiyoshi Tanaka
Waseda University, Matsushita Research Institute Tokyo, Inc.

2－2．画像処理部分

動画像の１フレームを３×３のエリアに分割し、このエリアごとに画像の特徴パラメータを計算する。

エリア (1,1)	エリア (2,1)	エリア (3,1)
エリア (1,2)	エリア (2,2)	エリア (3,2)
エリア (1,3)	エリア (2,3)	エリア (3,3)

図２　エリア分割

画面を図２のようなエリアに分割することにより、画像全体の平均的な特徴だけでなく、局所的な特徴も得られる。各エリア（n×m）ごとに計算する画像の特徴パラメータは以下のとおりである。

a）カラー情報：

$$L_R(i,j,t)=\sum_{x_i=1}^{n}\sum_{y_j=1}^{m} l_R(x_i,y_j,t) \quad (1)$$

ただし、$L_R(i,j,t)$は時刻 t のエリア(i,j)における Red の輝度総和であり、$l_R(x_i,y_i,t)$は時刻 t のエリア内の点(x_i,y_i)における Red の輝度を表わす。同様にして Green, Blue, Hue, Lightness, Saturation も計算する。

b）エッジ密度：

$$E_R(i,j,t)=\sum_{x_i=1}^{n-1}\sum_{y_j=1}^{m-1}\{|l_R(x_{i+1},y_j,t)-l_R(x_i,y_j,t)| + |l_R(x_i,y_{j+1},t)-l_R(x_i,y_j,t)|\} \quad (2)$$

ただし、$E_R(i,j,t)$は時刻 t のエリア(i,j)における Red のエッジ数総和を表わす。同様にして Green, Blue についても計算する。

c）シーンチェンジ：

$$SqErr=\frac{\sum_{i=1}^{3}\sum_{j=1}^{3} L_{R,G,B}(i,j,t+1)\times\sum_{i=1}^{3}\sum_{j=1}^{3} L_{R,G,B}(i,j,t)}{\sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}\{L_{R,G,B}(i,j,t+1)\}^2\times\sum_{i=1}^{3}\sum_{j=1}^{3}\{L_{R,G,B}(i,j,t)\}^2}}$$

$$\begin{cases} SC = 1 (if\ SqErr < \theta) \\ SC = 0 (other) \end{cases} \quad (3)$$

ただし、*SqErr*は時刻 *t* と *t+1* のフレーム間の類似度であり、*SC*はシーンチェンジを表わす。すなわち、２つのフレーム間の類似度で θ 以下であるときをシーンチェンジとみなすわけである。

２－３．統合処理部分

統合処理部分は、このシステムで中核をなす部分である。画像処理部分での出力を入力とし、音楽処理部分への入力を出力とする多入力多出力の連想記憶ニューラルネットワークを用いる。ユーザーが連想記憶ニューラルネットワークを学習させることにより、動画像をユーザー好みの音楽に対応させることができる。統合処理部分は、シーンチェンジを検出したときには、すべての出力を更新し、それ以外のときには、特定の出力のみを更新する。つまり、シーンチェンジが生じると基本コード列とリズムパターンが更新され新しい音楽が生成される。また、同一シーン内ではひとつの曲の中でテンポ・リズム・音色・ボリューム等が画像的特徴と結びついた形で変化する。

２－４．音楽処理部分

音楽処理部分では、コードとリズムパターンに基づいて、以下のような簡単な確率的規則により、メロディーが作曲される。

１．リズムパターンに合わせてメロディーの拍子（音符・休符）を決める。

２．コードを構成する音階を、メロディーの中に５０％以上含む。

３．２．以外の音階は経過音（コードの構成音の間を順次進行する）・刺繍音（コードの構成音から２度上がり、または下がって構成音に戻る）倚音（コードの構成音から２度上か下から順次構成音に戻る）などの音階によって構成する。

メロディーは、小節分毎にほかの音楽パラメータとともにMIDI音源に出力される。

３．サンプル動画像に対する背景音楽の考察

約３分程のサンプル動画像を用いて、シーンチェンジを検出した結果を図３に示す。図３において類似度が0.9を下回ったところをシーンチェンジとしてみなしている。サンプル動画像において、シーンチェンジはかなり正確に検出できた。

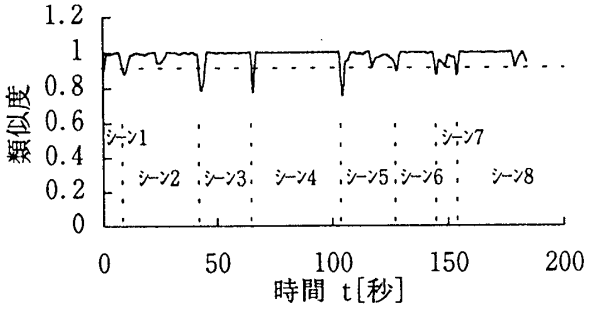

図３　シーンチェンジの検出結果

背景音楽は、メロディーがコード列とリズムパターンにしたがって作曲されているので、コードやリズムが似ている場合には、あまり変化が感じられなってしまうが、単純なシステムの割には、違和感のない背景音楽が生成できた。

４．おわりに

本稿では、動画像に対応した背景音楽の自動生成システムの概要を述べた。ここでは、リアルタイムでの画像入力から背景音楽を生成することを述べたが、さらに一貫性のとれた背景音楽の生成が可能となるように、録画などによって２パスでの背景音楽の生成も検討している。また、ユーザーインタフェース及び感性的なマッチングについては、今後の課題である。

[参考文献]

[1]Shuji Hashimoto, DingDing Chang: Music Generation from Moving Image, ICMC Proceedings 1995 pp.361-362 (1995)

[2]Nakamura J. et. al: Automatic Background Music Generation based on Actor's Mood and Motions, The Journal of Visualization and Computer Animation, vol.5 pp.247-264 (1994)

[3]Gong Y. et. al: Image Indexing and Retrieval based on Color Histograms, Multimedia Modeling, World Scientific, pp.115-126 (1995)

[4]Kiyoki Y. et. al: A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning, SIGMOD RECORD, vol.23 No.4 pp.34-41 (December, 1994)

[5]吉川　厚、他：ファジー演算を使った作曲支援システム，信学論 A vol.J74-A No.5 pp.735-742 (1991)

[6]織田　英子：作曲のしかた，成美堂出版 (1994)