社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2007－HCI－126　(9)
2007／11／12

感性キーワードと映像特徴を用いたBGM自動生成

藤江 哲也 , 相川 清明

東京工科大学バイオ情報メディア研究科
〒192-0982　東京都八王子市片倉町 1404-1
Email　g31070252a@gss.teu.ac.jp

あらまし　近年インターネットによる動画配信、地上デジタル放送による番組に対応する技術にはめまぐるしいものがあり、又その中で番組制作へのスピードも求められてきている。こうした背景において今回MA作業という音付け作業に注目し、これを自動化できないかと考えた。今回映像のBGMが流れる部分の特徴を洗い出しそれをもとにソフトウェアを開発した。また、映像と音のメディア変換といった視点からみても新たな視点からソフトウェアを開発した。本方法はバラエティー、ニュース番組等の自動効果音付加システムへの応用やデジタルコンテンツ等のコンテンツの自動生成に応用が考えられる。

Automatic BGM Generation using Emotional Keywords and Movie Features

Tetsuya Fujie　Kiyoaki Aikawa
†Tokyo University of Tecknology
1404-1 katakuracho, Hachioji-shi, Tokyo, 192-0982,Japan

Abstract　Recently, it is requested to boost speed for the works on TV program, content production. This report describes a method for automatic BGM generation for Multi Audio works. This method uses both emotional keywords and signal features obtained from the picture sequence. This method can contribute to automation to producing variety and news programs for TV or Internet

1. はじめに

現在のメディア変換を用いた映像コンテンツ、音楽コンテンツへのアプローチは主にシンセサイザーの発する音からきらびやかな映像を自動生成する、映像の特徴量から音を生成するといったリアルタイム性のあるものがほとんどであり、使用法のほとんどが楽器自身として使われているのが現状である。こういったものの多くがユーザーにとって可能性を追求するものにとどまり[1]、ヘビーユーザーしか使用していないような現状をみると、こうしたメディア変換技術がまだ発展途上にあり、誰にでも使いやすくもっと簡潔な方法でツールを使用できるようになったり、不必要なパラメータ値をコンテンツ自身から取り込めるようになるのではと考えている。従来の研究について映像から音を出すという研究に関して述べていくと映像の周波数から音を出すと言うことはもはや実現されている[1]。しかし、問題としてそれらは意図的に音を出すものではなく、ただ連続的に音を出すものがほとんどである(映像を意図的なものにすれば別だが)。本研究ではこういった音、ミュージックをある意図された時間のみで再生することができないだろうかと考えた。そのため本研究で注目した点は映像が音、ミュージックなどを再生すべきタイムライン上でどのような特徴を持っているか、又それ特徴をどうやって抽出するのかといったところに注目し、効果音自動的にその映像に合った効果音付をできるツールを提案する。またこうした選曲付けの自動化を進めることにより、現状のメディア変換をより効率的なものへと変え、よりインテリジェントな映像制作ツールを構築し、制作の生産効率を上げることに直結し、コンテンツ量を増やしながらも質を落とさないような番組作りが増えると考えた。また、今回映像から人間の服の色を抽出してくることにより、その服が季節感に合った服であるかどうかということを春らしい、夏らしい、秋らしい、冬らしいといった感性語(季節感)にパラメータを変更できるようにした。

2. ソフトウェアに関して

今回開発した自動効果音付加ツールはコンボボックスから音をつけたい感性語(楽しい、さわやかなど)のキーワードを選び書き出しボタン[2]を押すことでよくバラエティーの番組にあるような説明部分に的確にBGMを書き出すことができるツールである。

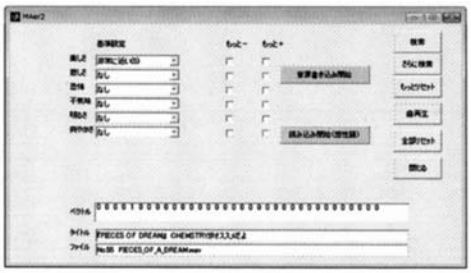

図1　自動効果音付加ツール

ここでは、ニュース番組やバラエティなどの毎週使われるような映像を扱うことを想定している[2]。

3.自動効果音付加ツール処理の流れ

UIのほとんどは[2]から流用したものだが、今回の研究は効果音書き出しの自動化ということで作った部分は主に書き出し部分、映像読み込み部分である。映像を分析するに当たって、今回は映像の速度場を求め、それが限りなく0に近いときに映像が止まっている(番組の説明部分である)と考えた。本研究ではオプティカルフロー(ブロックマッチング法)を参考に速度場が0に近いフレームを求め速度がある際のベクトルを作ることは省略し、画像全体を動きテンプレートを巡回させることで1フレー

ム前の画像の差分を求めることで速度場を求めていった[3][4]。
今回映像の速度場をＮＴＳＣの規格を参考に映像信号を 720×480 ドット,色信号ＲＧＢを r,g,b とし、速度場を求めるために巡回させるテンプレートを動きテンプレートとする。

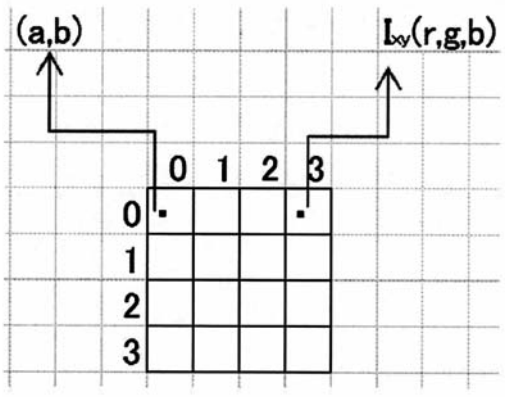

図２　動きテンプレートの説明

動きテンプレートの最小の単位は 4×４pixel とし,（1，1）の部分を毎回（a, b）とし、一つの画素 I の中の RGB 成分を(r,g,b)と表現し、これを巡回させていった。また、映像から読み込んだ各フレームは frame, 1 つ前のフレームを frame-1 とし、映像の 1 フレーム中の r,g,b のそれぞれの動きを *Tr, Tg, Tb* とした。

$$S(x,y)=\sum_{x=0}^{3}\sum_{y=0}^{3}\mathrm{I}_{a+x,b+y}(r,g,b) \quad (1)$$

$$T(r,g,b)=\sum_{x=4}^{740}\sum_{y=4}^{480}\mathrm{frame}\times S_{xy}-\sum_{x=4}^{740}\sum_{y=4}^{480}(\mathrm{frame}-1)S_{x-1,y-1} \quad (2)$$

Tr+Tg+Tb が限りなく 0 に近い場合これを動きのないフレームとみなし、１０フレーム動きがない場合音を生成するようにした。実際に software を使う場合は音を何度も上書きするように動作する。また今回文字キーワードだけでは表現者の意図した BGM を選択できないという観点から映像のパラメータから『季節感』というキーワードを抽出してくるようにした。

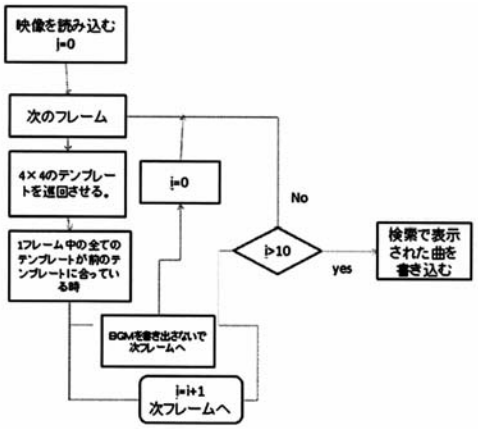

図３　動きテンプレートを巡回させる際の映像の速度場を求めるフローチャート

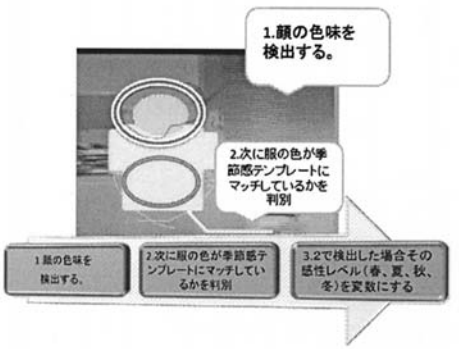

図４　色味テンプレートを巡回させるための方法

技術的に言及していくと、まず図４にあるように１フレーム中の人間の顔の色味をキャプチャする[5]。これに関しては顔の大きさがある程度の大きさ（映像の 1/4~1/32）であることを想定し、8×8pixel の色味テンプレートを毎フレーム中全て巡回させこれがマッチングし

たときにそれを顔と捕えるようにした[3][6]。次にその顔と認識した8×8pixelの色味テンプレートと大きさが同じ大きさで顔の下の部分の画像をキャプチャし、それをRGBからHSVに変換することで人間味のある色の識別を可能にした。

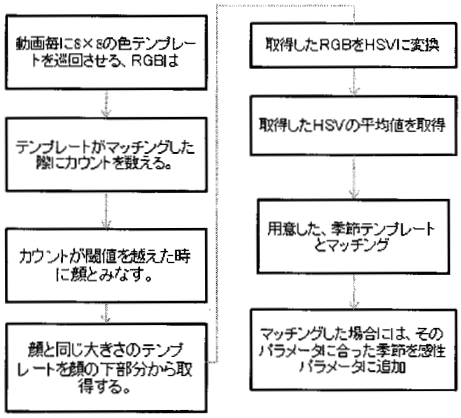

図5　色味テンプレートを巡回させた際における、映像から感性キーワードを抽出するためのフローチャート

次に『季節感』というキーワードと結びつけるためにあらかじめ用意した色味を参照することでこの感性語のパラメータを変化させた。ポップアップにあるキーワードと映像から抽出したキーワード（パラメータ）を両方同時に使用することによりマルチモーダルな選曲を可能にしている。この機能は全ての映像中の１フレームで起きればその服に合ったキーワードを取得してくるようにしている。

また、映像フレームにカット検出の技術を用い従来使われているカットを検出したらカットポイントを生成するような使い方でなく、映像が、違うショットに切り替わってもこれをBGMを書き込むポイントだと認識させないようにした[7]。

次に音のＢＧＭを生成することについて技術的に言及していくと、映像の速度場が０に限りなく近い映像が続くときに音源をキーワードから選ばれた音源を参照し、それまでの映像に動きがある部分には空白トラックを作成した[5][6]。これをつなぎ合わせて書き出すことでＤＡＷや映像制作ツールに書き出したトラックをインポートしたときに何も手を付け加えなくても音源が的確な位置かつ、その映像に適した音を配置することができる。

3. 実験

次は実際に日本語教材用に作られた映像を使いこのツールがどれくらい正確にＢＧＭを生成できるのかということを最小二乗法を用いた単回帰分析行うことで検証した。

映像のサンプルは30種類用意し、それぞれの映像が音を生成し始める時間（startと表記）、生成し終わる時間(endと表記)し、ここで書き込むべきと想定した時間を推定時間、実際に音を生成し始めた時間、し終わる時間を実測時間とし、横軸ラベルを推定時間、縦軸ラベルを実測時間とし、最小二乗法を用いた単回帰分析を用いて行った。ラベルの単位は秒であるとする。なお本論文の全てのグラフは最小二乗法を用いた単回帰分析を用いているものとする。

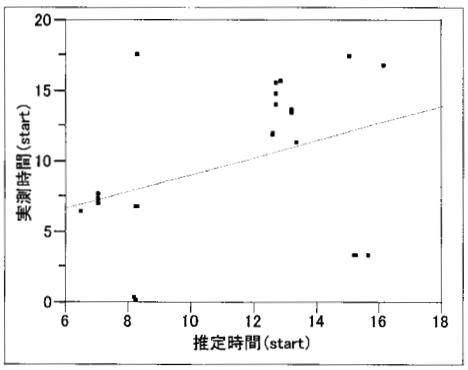

図6　ＢＧＭを生成し始めるタイミングにおける推定時間と実測時間との時間的誤差

表1　図６におけるあてはめの要約

R2乗	0.142847
自由度調整R2乗	0.10988
誤差の標準偏差(RMSE)	5.229676
Yの平均	10.07107
オブザベーション(または重みの合計)	28

図６は音を生成し始めた推定時間と実測時間の分析である。図６の場合はずれ値が極端に直線から離れていてこれによりグラフ上ではあまり正確に音を生成しているようには見えないが、推定時間が 13 秒、実測時間 15 秒周辺に大きな固まりがある。この部分は正確にＢＧＭを生成できている部分である。またこれだけ書き込む時間がずれる理由として、今回の映像では映像の速度場が０に限りなく近くなるフレームが多々あり、この部分と間違えて音を生成してしまうケースが図６で多々見られた。この部分の精度の悪さに関しては図７で別の視点から考察している。

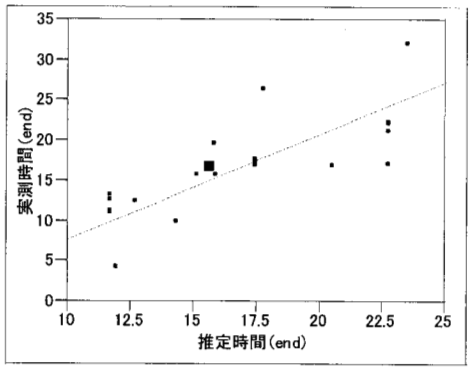

図7　BGMを生成し終わるタイミングにおける時間と実測時間との時間的誤差

表2　図７におけるあてはめの要約

R2乗	0.640851
自由度調整R2乗	0.627038
誤差の標準偏差(RMSE)	3.977483
Yの平均	15.61179
オブザベーション(または重みの合計)	28

次に図７は音を生成し終わる時間関係を示したものである。図７は全体的にいえばまとまっているようなグラフだが、正確性に関して言えば、図６で示している音を生成し始める実測時間のほうが正確に生成できているものが前述の図６の説明で固まりの部分を示したように多かった。しかし図６，７のグラフは回帰直線の精度自体はそれほど良い精度ではなかった。この理由として映像の速度場が０に近い時間上に音を生成する仕様のため特に図６のような相関をもたないようなグラフになってしまうのだと分析した。

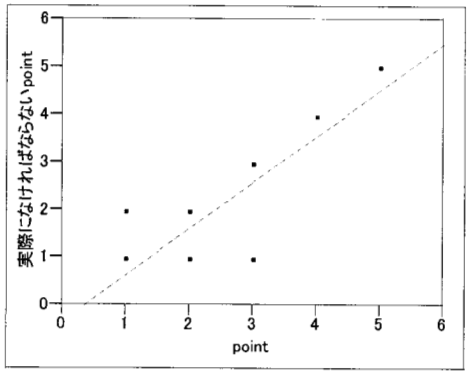

図8　実際にBGMを書き込んだ point と実際にBGMがあるべき場所におけるの point の誤差のグラフ

実際に図８ではプログラム上で音を生成するタイミングをすべて網羅し、その部分に point と名付けた映像のセグメンテーション作業を施し[3][4]、これをグラフ化した。図６の推定

時間と実測時間の平均誤差は実測時間のほうが推定時間よりも1.46秒速く、図8で、正しいポイントに生成するＢＧＭの平均は、0.27秒遅い。

表3 図8におけるあてはめの要約

R2乗	0.633629
自由度調整R2乗	0.619537
誤差の標準偏差(RMSE)	0.89509
Yの平均	2.571429
オブザベーション(または重みの合計)	28

今回の実験で、pointといったセグメント化されたデータにおいては48％という確率でＢＧＭを正確な位置に生成できるようになった。

4.結論

この報告では人間の発するキーワードからだけでなく映像から感性を取得し、それを人間の発想するキーワードとミックスすることでマルチモーダルな効果音自動付加システムを提案した。また、効果音付の自動化を提案した。正確な音付けはまだ実現できていないが、これから精度向上を行いたい。

参考文献

[1]池田徹志、室田健吾、石黒浩,“全方位映像から音楽情報へのメディア変換に基づく資格情報の伝達,”情報処理学会研究報告, Vol.48.No1, pp274-282, 2007

[2]相川清明、谷島加奈子,“ベクトル空間法を用いた相対的感性表現による音検索,”情報処理学会研究報告,2006-ＳＬＰ-65,pp.5-10,Februaly 2007

[3]奥野洋平、朱青、富永英義,”同画像インデキシングを目的とした人物顔追跡に関する検討,”情報処理学会研究報告,ＩＰＳＪ（ＡＶＭ）,pp.145~150

[4]服部しのぶ、亀山渉、富長英義,”映像特徴空間に基づく同系映像の分類,”情報処理学会研究報告,2003-ＡＶＭ-43, pp.145~150 , 2003

[5]小渡悟、星野聖 ,”オプティカルフローと色情報に基づく掌の検出と追跡によるジェスチャ認識,”情報処理学会研究報告,ＩＰＳＪ Vol4 No.SIG9（ＣＶＩＭ７）,pp-47-54

[6]入谷勝、茂野聡登志、前原隆、金丸隆志、関根優年,“テンプレートマッチングにおける主要成分分布の解析,”映像情報メディア学会技術報告

[7]池添剛、梶川嘉延、野村康雄 ,“音楽感性空間を用いた感性語によるデータベース検索システム,”情報処理学会研究報告 Vol.42, No.12 ,pp.3201-3212

[8]中神央二、渡辺裕、富永英義 ,“差分画像を利用したアニメーション映像からのオブジェクト抽出,”情報処理学会オーディオビジュアル複合情報処理,40-6, pp.31~35,2003