H-021

音楽TV番組における特定ダンスシーンの符号化検索

Compressed video search of specific dance scene in music TV program

ZHANG XUEQI*　　森田 啓義 *
Hiroyoshi Morita

1　はじめに

近年，TV チャンネル数の増加に伴い，ユーザーが視聴できるコンテンツも益々多様化になってきており，膨大なコンテンツから興味のあるシーンを自動的に抽出する技術が強く望まれている．

本研究では，様々なコンテンツの中から，特に音楽ライブ番組において，歌手が歌いながら行う特定のダンスシーンの自動抽出に着目した．ダンスシーンは同一グループによる同じ振り付けであっても，番組ごとに様々なカメラアングルで撮影されることが多く，ファンにとってはどれも見逃したくないだけでなく，従来の顔検出などのビデオ処理技法を適用してもダンスシーンを特徴づけることは困難である．

そこで本研究では，圧縮動画の符号化パラメータを特徴量とし，平滑化 z-スコアアルゴリズムと，BSpline 曲線を組み合わせることによって，音楽ライブ番組における特定ダンスシーンの検索手法を提案する．

2　関連研究

音楽番組のシーン解析に関しては，顔認識手法により出演者変化の検出と，音響信号が属するオーディオクラスの帰属確率により無音、音楽、音声、雑音の分類を用いて音楽番組における司会シーンと歌唱シーンの分割を行った研究 [1] や，顔認識手法に基づく口の動き検出手法と，歌声区間と非歌声区間の状態を行き来する隠れマルコフモデルによって楽曲を表現することで，混合音中の歌声区間の推定手法を組み合わせることによって歌唱シーン検出をする研究 [2] が報告される．しかし，画像と音響特徴のみでは，出演者のからだ全体の動きを把握できないため，検出された歌唱シーンにおけるダンスパフォーマンスの有無までを判断できない．

*電気通信大学大学院情報理工学研究科，Graduate School of Informatics and Engineering, The University of Electro-Communications

3　圧縮動画におけるダンスシーンの特徴

3.1　圧縮符号化パラメータ

MPEG -2（地上デジタル放送の標準規格である符号化法）では 1 秒間に約 30 枚のフレーム（ピクチャとも言う）が I、P、B と いう 3 種類のピクチャに分類され，それぞれのピクチャに定められた符号化法によって高効率なデータ圧縮が実現されている．特に B ピクチャでは, 直前と直後の I/ P ピクチャを利用した順方向、逆方向、双方向予測符号化を行っている．各ピクチャは, マクロブロック (以下 MB) と呼ばれる単位に分割され，MB ごとに符号化処理が行われる．さらに，MB のサイズには 16x16 と 16x8 画素の 2 種類があり，それらの中から符号化効率が尤も良くなるように決定される．

圧縮動画データを観察すると，16x8 画素の MB（以下 16x8 MB）は移動体の境界領域に集中して発生することや，画面の動きの激しい部分に発生することが分かっている [3]，そこで，本研究では，B ピクチャにおいて上述した 3 種類の予測符号化を行っている 16x8MB（以下予測 16x8MB）の数に着目した．

3.2　音楽番組映像における 16x8 MB の考察

まず予備実験として，9 本の音楽番組映像（再生時間は 9 分 ~15 分）を使用し, GOP ごとに予測 16x8MB の数を調べた．ここで，GOP とは MEPG-2 で定められた時間的に連なったピクチャを一定枚数まとめる単位であり，通常は 15 枚のピクチャからなり，その中に B ピクチャは 10 枚含まれている．

実験に用いる音楽番組映像 9 本のうち，3 本には少なくとも一つ以上のダンスシーンが含まれていて，残りの 6 本はダンスシーンは含まれていない．図 1(a) はダンスシーンを含むある 1 本の映像の結果であり，一方，ダンスシーンのないある 1 本の映像の結果は図 1(b) に示す．両図とも，横軸は GOP 番号を表し, 縦軸は予測 16x8MB

の総数を表す. 図 1(a) においては GOP 番号 153~486 の区間 (黒枠で囲まれる部分) はダンスシーンである.

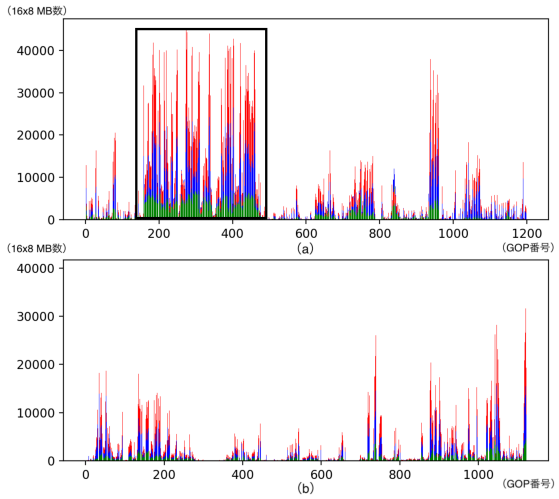

図 1 : GOP ごとに予測 16x8MB の数. 積み上げ棒について, 緑の部分は双方向予測符号化 16x8MB であり, 青の部分は逆方向予測符号化 16x8MB であり, 赤の部分は順方向予測符号化 16x8MB である.

9 本の映像を用いて得た知見をまとめると, ダンスシーンでは，予測 16x8MB が発生しやすく、20000 以上の数の 16x8MB が発生した GOP が多いことと，ダンスシーンの始まりと終わりの GOP 番号において，急激な 16x8MB 数の変化が発生することである.

4 従来手法

離散信号のピーク検出については，過去の一定の長さの信号の平均と標準偏差に基づいて，現時点のピークを検出する平滑化 z-スコアアルゴリズムが報告されている [4]. 同アルゴリズムでは，移動平均を用い，過去の信号を平滑化するので，次のピークを検出する際に，現時点のピークからの影響は程よく減ることが可能であると考えられる.

4.1 平滑化 z-スコアアルゴリズム

平滑化 z-スコアアルゴリズムを [Algorithm 1] に示す 10 ステップからなる.

4.2 平滑化 z-スコアを用いたダンスシーン検出の実験

平滑化 z-スコアアルゴリズをダンスシーン検出に適用するにあたって，入力データや各パラメータを以下のように定めた：

Algorithm 1 平滑化 z-スコアアルゴリズム

入力: データ列 $\boldsymbol{X} = X_1X_2X_3\ldots X_n$
出力: 判定結果 $\boldsymbol{Y} = y_1y_2y_3\ldots y_n$
● 内部変数：ℓ : スライド窓の長さ（$\ell \geq 2$）
w : 重み係数
τ : 閾値
1. **for** i=1, 2, 3$\cdots n, n+l-1$ **do**
2. $\mathrm{S}_i = wX_i + (1-w)\,\mathrm{S}_{i-1}$ (1)
を求める. ここで, $\mathrm{S}_0 = 0$.
3. 移動平均 $\overline{S_i}$ を求める.

$$\overline{S}_i = \begin{cases} \frac{1}{i}\sum_{k=1}^{i}\mathrm{S}_k & 1 \leq i \leq l-1 \\ \frac{1}{l}\sum_{k=i-l+1}^{i}\mathrm{S}_k & l \leq i \leq n \\ \frac{1}{n-i+l}\sum_{k=i-l+1}^{n}\mathrm{S}_k & n+1 \leq i \leq n+l-1 \end{cases} \quad (2)$$

4. 移動標準偏差 σ_i を求める.

$$\sigma_i = \begin{cases} 0 & 1 \\ \sqrt{\frac{1}{i-1}\sum_{k=1}^{i}\left(\mathrm{S}_k - \overline{S}_i\right)^2} & 2 \leq i \leq l-1 \\ \sqrt{\frac{1}{l-1}\sum_{k=i-l+1}^{i}\left(\mathrm{S}_k - \overline{S}_i\right)^2} & l \leq i \leq n \\ \sqrt{\frac{1}{n-i+l-1}\sum_{k=i-l+1}^{n}\left(\mathrm{S}_k - \overline{S}_i\right)^2} & n+1 \leq i \leq n+l-1 \end{cases} \quad (3)$$

5. **end for**
6. $z_1 = 0, z_2 = 0, y_1 = 0, y_2 = 0$ とおく
7. **for** i=3, 4$\cdots n$ **do**
8. $\overline{S_{i-1}}$ と σ_{i-1} を用いて，X_i を正規化する.

$$z_i = \frac{X_i - \overline{S}_{i-1}}{\sigma_{i-1}} \quad (4)$$

9. 判定

$$y_i = \begin{cases} 1 & \text{if } |z_i| \geq \tau \\ 0 & \text{if } |z_i| < \tau \end{cases} \quad (5)$$

10. **end for**

入力データ列は，GOP 単位でまとめた B ピクチャにおける予測 16x8MB の総数 $\boldsymbol{X} = X_1X_2X_3\ldots X_n$ であり，n は GOP 数である．さらに，ℓ=30，w=0.5 とおく．特に，閾値 τ の決め方は，データに依存するので，本研究では式 (4) で得たデータ列 z_i の累積分布関数を cdf(t) とおき，0.969≤ cdf(t) ≤0.981 を満たす t の区間の中点を閾値 τ と定めた．再生時間は 9 分 ~15 分の 10 本のダンスシーンありの音楽番組映像を用いて実験を行った. 結果は表 1 に示す．表 1 において，ダンスシーンの正解（始点-終点）および結果（検出された始点-検出された終点）の単位は GOP 番号である.

映像 No.	GOP 数	ダンスシーンの正解	結果	誤差（Δ始点/Δ終点）
1	1200	153-486	159-462	-6/-24
2	1663	398-758	377-762	+21/+4
3	1928	122-424	73-356	+ 49/-68
		654-847	x	x
		848-1042	x	x
4	1829	1232-1554	1250-1623	-18/+69
5	1797	326-646	x	x
6	1771	1396-1768	x	x
7	1751	16-440	1346-1632	*
8	1879	324-636	342-652	-18/+16
9	1657	336-676	321-600	+15/-76
10	1471	258-682	279-762	-21/+80

表 1 : 平滑化 z-スコアを用いたダンスシーン検出の結果.“x”はダンスシーンを検出できないこと（以下未検出）を表す.“＊”は結果の誤差がかなり大きいこと（以下検出外れ）を表す.

誤差における“Δ始点”は検出された結果の始点と正解のダンスシーンの始点の差であり，“＋”は正解のダン

スシーンの始点より前の GOP で検出されたことを表し，“－”は正解のダンスシーンの始点より後の GOP で検出されたことを表す．誤差における“Δ終点”は検出された結果の終点と正解のダンスシーンの終点の差であり，“＋”は正解のダンスシーンの終点より後の GOP で検出されたことを表し，“－”は正解のダンスシーンの終点より前の GOP で検出されたことを表す．

表 1 より，平滑化 z-スコアアルゴリズムを用いたダンスシーンの検出手法では，映像 3、5、6 に対して未検出が発生し，映像 7 に対して検出外れが発生したことが分かった．

5　提案手法

平滑化 z-スコアアルゴリズムを用いた実験の結果を踏まえて，閾値の τ の決め方による検出精度への影響を減らすため，平滑化 z-スコアアルゴリズムの後半で行っている X_i の正規化と，閾値を用いた判別を行わないで，移動平均の部分 (式 (1) ∼ (2)) と，新たに BSpline 曲線 [5] による平滑化を組み合わせる手法を提案する．

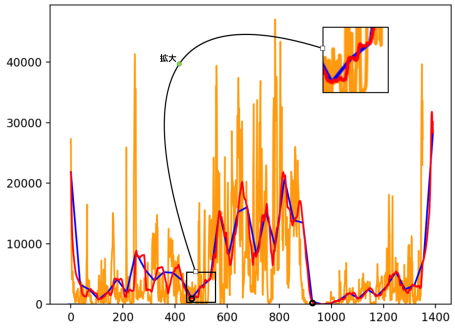

図 2：ダンスシーンありの映像に 16x8MB 数の移動平均折れ線と BSpline 曲線．予測 16x8MB の総数はオレンジ線であり，16x8MB 総数の移動平均折れ線は赤い線であり（スライド窓の長さ $\ell = 30$），移動平均折れ線の BSpline 曲線は青い線である．k =3, BSpline 係数 c の数=移動平均の長さ（1381），ノット t の数=1385，近似点数（以下 N_B）= 50.

5.1　BSpline 曲線の定義

BSpline 曲線は，BSpline 係数とノットより定義される区分的多項式曲線である。具体的な定義は [定義 1] に示す．点列を補間する BSpline 曲線を求める際に，与えられた点列の横座標定義域の閉区間から，一定数の等間隔 x 座標を取得し，各 x 座標を用いて，式（6）によって $S(x)$ を求める．最後に，それらの $(x, S(x))$ が座標となる一定数の点（以下近似点）を繋げて生成した曲線は BSpline 曲線である．予備実験の観察からダンスシーンの始まりと終わりの GOP 番号において，急激な 16x8MB 数の変化

定義 1　BSpline 曲線の定義

基本式：$\mathrm{S}(x) = \sum_{j=0}^{n-1} c_j B_{j,k;t}(x)$　(6)

内部変数：

（i）BSpline 多項式の次数：k

（ii）BSpline 係数：$\mathrm{c} = c_0 c_1 c_2 ... c_n, n \geq k+1$,
c は与えられた点列の縦座標値により求める．

（iii）ノット：$\mathrm{t} = t_0 t_1 t_2 ... t_u, u = n+k+1$,
t は与えられた点列の横座標の閉区間により求める．

NOTES：

$$B_{i,0}(x) = \begin{cases} 1 & \text{if } t_i \leq x \leq t_{i+1} \\ 0 & otherwise \end{cases} \quad (7)$$

$$B_{i,k}(x) = \frac{x - t_i}{t_{i+k} - t_i} B_{i,k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(x) \quad (8)$$

が発生することが分かっているので，図 2 において，黒い丸でマークする極小値に対する横軸の座標（極小点と呼ぶ）がダンスシーンの始点と終点であると考えられる．拡大された部分の赤い線と青い線を観察すると，多くの極小値を持つ赤い線から生成した BSpline 曲線上に，極小値の数は減少したことが分かった．

5.2　BSpline 曲線によってダンスシーン検出の手順

BSpline 曲線上におけるダンスシーンの始点と終点である可能性の大きい極小点の検出する際，以下のものを用いる．

- k=3
- BSpline 曲線上における近似点の x 座標の集合：
$\xi = \{\xi_1, \xi_2, \xi_3, ..., \xi_n\}$,
$\xi_i = \frac{GOP\ 数+l-2}{n-1} \times (i-1) + 1 \quad (1 \leq i \leq n, n = N_B)$.
- 極小点の集合：$\mathcal{S}_{\min} = \{\underline{x}_1, \underline{x}_2, \underline{x}_3, ... \underline{x}_k\}$（$0 \leq k \leq n$），$\mathcal{S}_{\min} \subset \xi$.
- BSpline 曲線上の極大値に対する横軸の座標を極大点と呼ぶ．極大点の集合：$\mathcal{S}_{\max} = \{\overline{x}_1, \overline{x}_2, \overline{x}_3, ... \overline{x}_m\}$（$0 \leq m \leq n$），$\mathcal{S}_{\max} \subset \xi$.
- 極点（極小点と極大点）の集合：$\mathcal{M} = \mathcal{S}_{\min} \cup \mathcal{S}_{\max} = \{x_1, x_2, x_3, ... x_{k+m}\}$（$x_i < x_{i+1}, 0 \leq k+m \leq n$），$\mathcal{M} \subset \xi$.

BSpline 曲線によってダンスシーン検出の手順は [Algorithm 2] に示す．検出手順においては，$\mu_1 = 7000, \mu_2 = 4500$ とおく．

5.3　提案手法を用いたダンスシーン検出の実験

4.1.1 に使用した映像データを利用して提案手法を用いて予備実験を行った，結果は表 2 に示す．

提案手法を用いた予備実験の結果を観察すると，映像 3 に対して未検出が発生した．未検出の映像本数については，平滑化 z-スコアアルゴリズムを用いた実験の結果

Algorithm 2 提案アルゴリズム

入力: ダンスシーンがある音楽番組映像における予測 16x8MB 移動平均の BSpline 曲線 $S(x)$
出力: 始点集合 $\mathcal{X}_i^s$, 終点集合 $\mathcal{X}_i^e$

for i=1, 2, 3$\cdots k+m$ **do**
  もし，$x_i \in \mathcal{S}_{\min}$， $x_{i+1}\ x_{i+2} \in \mathcal{M}$ であり，さらに，2 つの条件
  1) $B(x_i) \leq \mu_1$
  2) $\min\{B(x_{i+1}) - B(x_i), B(x_{i+2}) - B(x_i)\} \geq \mu_2$
  が全て真であるならば，x_i を $\mathcal{X}_i^s$ に代入する．
end for
$\mathcal{X}_i^s$ を $\{x_1^s, x_2^s, ..., x_l^s\}$ とおき，
for i=1, 2, 3$\cdots l$ **do**
  for j=1, 2, 3$\cdots k+m$ **do**
    もし，$x_j \in \mathcal{S}_{\min}$， $x_{j-1}\ x_{j-2} \in \xi$ であり，さらに，2 つの条件
    1) $250 \leq x_j - x_i^s \leq 510$
    2) $\min\{B(x_{j-1}), B(x_{j-2})\} > B(x_j)$
    が全て真であり，x_j の数は 1 であるならば，x_j を $\mathcal{X}_i^e$ に代入する．もし，x_j の数は 1 以上なら，それらの平均を $\mathcal{X}_i^e$ に代入する．
  end for
end for
$\mathcal{X}_i^s$ と $\mathcal{X}_i^e$ の中から，GOP 区間の端点に対応したペアについて，16000 以上の予測 16x8MB 数をもつ GOP の数が 60 個以上あれば，そのペアをダンスシーンの端点でとして出力する．

映像 No.	1	2	3	4	5	6	7	8	9	10
Δ始点	+28	+53	+3, +16, x	+20	+66	+75	-56	+13	-7	+16
Δ終点	-40	-12	+55, +75, x	+37	-26	-67	+31	+103	+28	-25

表 2： BSpline を用いたダンスシーン検出の予備実験結果.ℓ=30, w=0.8, N_B=50.

（以下従来手法）より少なくなっている．さらに，検出外れの映像はなかった．提案手法を検証するため、ほかの 10 本のダンスシーンありの音楽番組映像（映像の情報は表 3 に示す）を用いて本実験を行った．w の値は 0.5 以上なら，大きな変化が出ないため，w を 0.8 に設定した．本実験の結果を表 4 に掲げる．

映像 No.	1	2	3	4	5	6	7	8	9	10
GOP 数	1352	1145	1795	1797	1269	1797	1729	1375	1557	1258
ダンスシーンの区間	516	661	1111	296	935	107	876	51	276	21
	|	|	|	|	|	|	|	|	|	|
	892	1042	1455	614	1256	425	1384	420	583	474

表 3：提案手法を用いた本実験用の映像の情報

ダンスシーンを見逃さないため，誤差については，“－”より“＋”の方が良いと考えられる．本実験の結果を見ると，$\ell = 30$，$N_B = 40$ と，$\ell = 40$，$N_B = 50$ の実験において，映像 5、6、9 に対して誤検出が発生した．$\ell = 30$，$N_B = 50$ の方には，映像 5 に対して誤検出が発生した．誤検出を除いてそれぞれの誤差の平均は表 5 に示す．$\ell = 30$，$N_B = 50$ の方は，始点誤差の平均は一番小さく，終点誤差は正数である．以上を踏まえて，ℓ=30，w=0.8，N_B=50 の方がより良い結果を得られることが分かった．

映像 No.	$\ell = 30$、$N_B = 50$	$\ell = 40$、$N_B = 50$	$\ell = 30$、$N_B = 40$
1	+66/-57	+34/+72	+53/+70
2	+39/-49	+10/+12	+24/-57
3	-79/+144	-11/+78	△ -17/+49
4	-1/+130	-3/+134	+155/-97
5	△ +36/+14	△ +29/+24	△ +31/+17
6	+107/-16	+107/-28 △	+107/-96 △
7	-20/-40	+11/-176	+61/-184
8	-6/+38	-6/+12	+15/+14
9	-15/+128	△ +146/-95	△ +154/-93
10	+21/-81	+21/-78	+21/-76

表 4：BSpline を用いたダンスシーン検出の本実験結果.“△”はダンスシーンでないシーンをダンスシーンとして検出されたこと（以下誤検出）を表す.

	$\ell = 30$、$N_B = 50$	$\ell = 40$、$N_B = 50$	$\ell = 30$、$N_B = 40$
始点	+14.8	+33.8	+60.4
終点	+21.1	-4.5	-45.3

表 5：BSpline を用いたダンスシーン検出の本実験誤差の平均

6 まとめ

本研究では，圧縮動画の符号化パラメータ 16x8MB と，平滑化 z-スコアアルゴリズムアルゴリズムにおける移動平均の部分と，BSpline 曲線を組み合わせる手法を提案し，音楽番組映像からのダンスシーン検出を実現した。今後，既存のダンス映像を検索キーとし，符号化パラメータを用いてマッチングすることによって音楽番組映像における同一のダンスシーンの検出を目指す.

参考文献

[1] 吉田 壮, 小川 貴弘, 長谷山美紀,“歌謡番組における映像の構造に注目したシーン分割手法 ”, 電子情報通信学会論文誌, Vol.J97-D , No.7 (2014).

[2] 平井 辰典, 中野 倫靖, 後藤 真孝, 森島 繁生,“歌手映像と歌声の解析に基づく音楽動画中の歌唱シーン検出手法の検討 ”, 情報処理学会研究報告 , Vol.n, No.n (2014).

[3] 単鴻,“MPEG2 動きベクトルを用いた複数移動体の検知・追跡システム ”, 電気通信大学大学院情報システム学研究科修士論文 (2008).

[4] Patrick Perkins, Steffen Heber,“Identification of Ribosome Pause Sites Using a Z- Score Based Peak Detection Algorithm ”, 2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS) (2018).

[5] Tajima Robotics, https://tajimarobotics.com/basis-spline-interpolation-program/ (2018/8/5).