E-14

動画の階層性を考慮した手の動きの類似検索

Hand Motion Retrieval Using Video Hierarchy

池田 朋弘†　　野宮 浩揮†　　宝珍 輝尚†
Tomohiro Ikeda　　Hiroki Nomiya　　Teruhisa Hochin

1．はじめに

近年, コンピュータの性能の向上や HDD レコーダ, 動画共有サイトの普及により, 個人が大量の動画データを扱う機会が増加している。そこで, 大量の動画データからうまく目的の動画データが検索できるように, 動画データから多角的に情報を取り出し, 検索に活用する方法が求められている。

このような要求に応えるべく取り組まれている研究の方針として, 大きく２つのアプローチがある。一つ目のアプローチは, 動画の内容を適切に記述することにより動画の意味的な情報を使用する意味アプローチである。これにより, “人が棒を切るシーン”といった検索が可能となる。意味アプローチに該当する研究として, 映像制作の観点からの内容記述[1]やユーザの検索要求に対して柔軟に動画の内容を定義できる内容記述[2]などが提案されている。これらは, 人が映像を観て捉える映像の意味に良く一致するものであり, 直感的な検索となる。しかし, 記述に用いる概念が抽象的であるため, 動画から人手でこの概念を抽出しなければならないという問題がある。

二つ目のアプローチは, 動画の信号的な特徴に注目し情報を取り出す特徴アプローチである。これにより, “赤色の画面領域が右に５画素移動するシーン”といった検索が可能となる。特徴アプローチに該当する研究として, 動体の色情報と移動距離を特徴とした手法[3]や, 動画を時間と空間（平面）からなる３次元の信号と捉え, ３次元周波数成分を特徴とした手法[4]などが提案されている。これらは, 自動的または半自動的に特徴を抽出できるが, 意味アプローチと比べて非直感的である。

そこで本研究では, 両アプローチの欠点を補い, 半自動的な処理で, かつ, 直感的な検索を可能とすることを目的とする。そのために, 特徴アプローチに準じて動画から半自動的に登場オブジェクトの動きを抽出する方針をとる。登場オブジェクトとは動画に映る人物などの物体を示す。さらに入力を動画データとして, その動画データに含まれるオブジェクトと類似した動きをするオブジェクトが含まれるシーンを提示する検索手法の構築を目指す。

この検索手法を実現するシステムの例を図１に示す。この例では, 予め検索対象とする動画データを”動きの取得”機構に入力することによって取得されるオブジェクトの動きをシステムに登録しておく。そして, 検索キーとなる動画データを入力すると, ”動きの取得”機構によってオブジェクトの動きが取得され, ”動きの類似判定”機構によってシステムに登録されたオブジェクトの動きと類似判定を行い, その結果をランキングとして出力する。

†京都工芸繊維大学大学院, Kyoto Institute of Technology

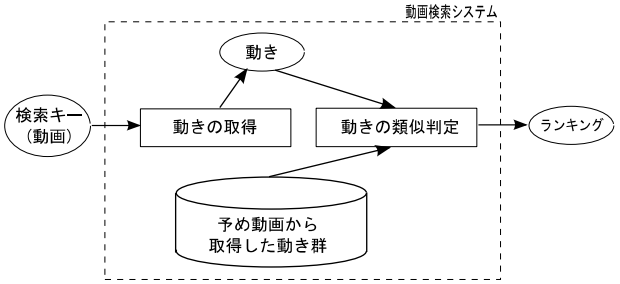

図１　動きに基づく動画検索システムの一例

この手法では登場オブジェクトの動きを類似判定する機構が重要になる。具体的に言うと, 人による視覚的な類似判定と同等な判定結果を出力し, かつ高速な処理を行う機構が求められる。

本研究では, 動きの類似判定に動画の持つ階層性を利用する。これによって, 動きを表すデータをおおまかに比較することができ, 情報へのノイズの影響の低減や, 比較回数の減少につながるので, 類似判定の精度を保ちつつ高速な類似判定が可能になると考えられる。

また, 本研究では動きを時系列データとして扱うが, 時系列データの比較には Dynamic Time Warping（DTW）と呼ばれる手法がよく用いられている[5]。この手法は長さが異なる時系列を比較することが可能であるが, 処理時間が長い, ノイズの影響を受けやすいといった問題がある。

さらに, DTW といった既存の手法は汎用的な時系列データの比較手法なので, 扱う時系列データの性質を考慮した設計はされていない。動画は MPEG-7[6]で導入されているような階層性を持ち, この性質を考慮することによってより精度が高く処理時間が短い手法が設計できる可能性がある。

そこで本稿では, 図１のような動画検索システムを構築するために, 動画の持つ階層性を考慮し, 動きに空間領域と時間領域の階層性を適応した, 動きの類似性判定の手法を提案する。対象とする動画は茶道のお手前の動画で, 点前者の右手の動きを手動で取得し, 類似性判定に用いる。

以後, 2. では動画の持つ階層性および時系列データの比較に関する関連研究について述べる。3. で提案手法について述べ, 4. で実験とその結果を述べる。最後に 5. でまとめを行う。

2．関連研究

2.1　動画の階層性に注目した関連研究

動画は静止画の時系列として構成されるため, 動画には空間領域と時間領域が存在する。さらに, 動画は空間領域と時間領域のそれぞれに対して階層性を持つ。マルチメデ

ィアの高速な内容検索を目的として規格された MPEG-7[6]で導入されている空間領域の階層性を図２に示し，時間領域における階層性を図３に示す。図２では，あるフレームの背景とオブジェクト，オブジェクトの頭と体による階層構造を表している。図３では，動画全体がシーンに分割される階層構造を表している。

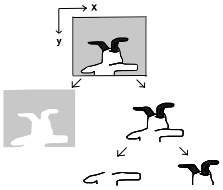

図２　MPEG-7 の空間領域の階層性の例

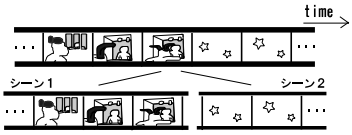

図３　MPEG-7 の時間領域の階層性の例

動画データを含むマルチメディアデータの内容を表す上でこのような階層性を考慮することは，データの内容をより特徴的に捉えることや，データの構造を表すこと，そして高速な処理を行うことなどに対して有効であり，様々な研究で利用されている[7]。本研究においても，動き情報の階層性を考慮する。

3. 提案手法

3.1　動き情報

動画データにおいて，隣接する２フレームを考える。このとき，それぞれのフレーム上の登場オブジェクトの位置の差によって表現される動き，すなわちオプティカルフローが存在する。本稿での動き情報とは，動画データの全てのフレームと対応したオプティカルフローの時系列データのことを言う。

3.2　空間領域上でのオプティカルフローの向きに着目した階層性の利用

空間領域でのオプティカルフローのノイズ対策と向きの段階的な比較のために，動画の空間領域の階層性を動きに導入する。具体的には，図４の(a)のような階層性が導入されていないオプティカルフローの向きを (b)と(c)のように有限個の向きに割り当てる。D_{17}では 0 度から 360 度の向きを 16 分割したものと原点の 17 方向に割り当て，D_9は同様に 8 分割と原点の 9 方向に割り当てている。図４の赤と青で示すオプティカルフローのように割り当てる数が少なくなると，似ている方向のオプティカルフローは同じ方向に割り当てられる。また，(b)の赤と青の割り当てが(c)では１つの割り当てに統合されるように，この割り当ては階層的になっている。

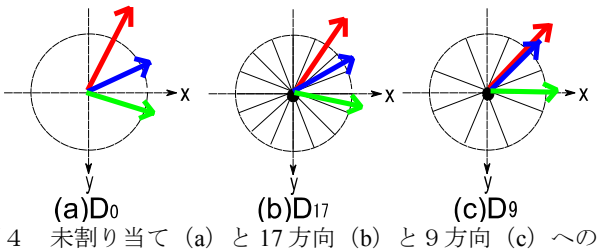

図４　未割り当て（a）と 17 方向（b）と 9 方向（c）への割り当て

3.3　時間領域上での動きの階層化

空間領域でのオプティカルフローのノイズ対策と効率的な動きの比較のために，動画の時間領域の階層性を動きに導入する。具体的には，時間領域上での動き時系列の階層化は隣り合うオプティカルフローの合成によって行う。

オプティカルフローの時系列を $TH_0 = (V_{0,0}, V_{0,1}, \ldots, V_{0,n-1})$とする。ここで，$V_{i,j}$は時間領域上の最下層より$i(\leq n-1)$上の層に対応する時系列のj番目の２次元ベクトルを表す。nはTH_0の時系列長を表す。このTH_0をもとに時系列$TH_i = (V_{i,0}, \ldots, V_{i,n-i})$の各ベクトルは以下のように求まる。

$$V_{i,j} = \sum_{k=j}^{j+i} V_{0,k} \tag{1}$$

例えば，$V_{2,2}$は図 5 のように最下層に対応するTH_0の$V_{0,2}, V_{0,3}, V_{0,4}$の合成によって求まる。

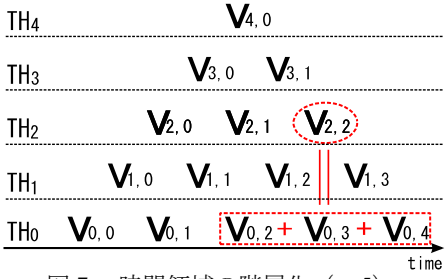

図５　時間領域の階層化（n=5）

3.3　動きの類似性判定方法

検索キーとなる動きの時系列 key とデータベース中の動きの時系列 data を比較して類似度を求め，類似する動き（動画）を決定する。

特にここでは，２つの動きと対応する時系列を比較し，類似度を求めることを考える。各階層の時系列は，類似度取得の際に計算して求める。

ここで$d(TH_a, D_b)$を，動き（オプティカルフロー）の時系列 data に対して，時間領域TH_aのそれぞれのベクトルに空間領域の割り当てD_bを適応させた時系列を表すこととする。さらに$d_{a,i,b}$は $d(TH_a, D_b)$の先頭から i 番目のベクトルとする。$k_{a,i,b}$も$d_{a,i,b}$と同様とする。ただし，$k(TH_a, D_b)$は検索キーとなる動き key に対応する時系列である。

２つの時系列（key, data）の類似度を求める関数 Sim は以下のように定める（ただし，$|k| \leq |d|$で，$|k|$は $k(TH_0, D_b)$の時系列長，$|d|$は $d(TH_0, D_b)$の時系列長である）。

$$Sim(k, d, a, b) = \max_{0 \leq s \leq |d|-|k|} calSim(k, d, a, b, s) \tag{2}$$

$$calSim(k, d, a, b, s) = \frac{1}{|k|} \sum_{i=0}^{|k|-1} cal(k_{a,i,b}, d_{a,i+s,b}) \tag{3}$$

$$cal(k_{a,i,b}, d_{a,j,b}) = \begin{cases} \frac{\langle k_{a,i,b}, d_{a,j,b} \rangle}{|k_{a,i,b}||d_{a,j,b}|} & \text{if } k_{a,i,b} \neq \vec{0} \wedge d_{a,j,b} \neq \vec{0} \\ 0.5 & \text{if } k_{a,i,b} \neq d_{a,j,b} \wedge (k_{a,i,b} \neq \vec{0} \vee d_{a,j,b} \neq \vec{0}) \\ 1 & \text{if } k_{a,i,b} = d_{a,j,b} = \vec{0} \end{cases} \tag{4}$$

4. 実験

4.1 実験方法

提案手法の性能を評価するために実験を行った。ここでは, 茶道でのお点前の動画中の点前者の右手の動きを対象とした（図 6）。カメラと茶道に用いるオブジェクトの位置関係がほぼ同じである２つの動画を使用した。

そのうちの１つの動画から５つの検索用の動き(表 1, 2)を, もう１つの動画から比較用の手の動き(hm)を手動で取得した。さらに, hm の中で key1～key5 と対応する区間(正解区間)を著者の一人が実際に動画を見て表 3 のように定めた。

実際に実験で比較に使用するデータは, 時系列長が 14462 フレーム分である hm に対して, 先頭から 600 フレーム分を 10 フレーム分ずつスライドさせて作成した（data1～data1387）。さらに, その data のうち表２の正解区間に 300 フレーム分以上重なるものを正解 data とした。図 7 では data を作成する様子と正解 data の判定の様子を示している。

図６　使用する動画データの一部

表１　検索用の動きの区間

キー	時間(分:秒)	フレーム	フレーム数
key1	11:13 ～11:30	20386 ～20690	305
key2	12:10 ～12:20	22068 ～22224	157
key3	12:42 ～12:52	22729 ～23149	421
key4	13:00 ～13:11	23318 ～23856	539
key5	15:15 ～15:24	24199 ～24717	519

表２　検索用の動きの内容

キー	動作内容(開始時の動作～終了時の動作)
key1	茶杓をふくさで拭く(茶杓をふくさの上に置く～棗の上に置く)
key2	上下に茶筅を動かす(茶筅を掴む～立てる動作に移る)
key3	お茶碗を茶巾で拭く(茶巾がつく～拭き終わって両手で持つ)
key4	お茶をお茶碗に(茶杓を掴む～棗を元の位置に置く)
key5	窯からお茶碗に水を入れる(柄杓を掴む～離す)

表３　正解区間

対応検索キー	時間(分:秒)	フレーム	正解data
key1	1:11 ～1:20	2156 ～2437	data186 ～data215
key2	2:19 ～2:25	4191 ～4375	data390 ～data409
key3	2:41 ～2:56	4843 ～5307	data455 ～data502
key4	3:05 ～3:26	5572 ～6190	data528 ～data590
key5	3:38 ～3:57	6545 ～7127	data625 ～data684

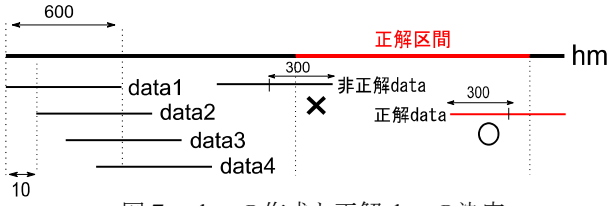

図７　data の作成と正解 data の決定

以上のようにして定めた, key1～key5, data1～data1387, 正解 data によって実験を行った。具体的には, それぞれの key に対して, TH_aの a を 0 から（key の時系列長-1）まで変化させ, かつD_bの b を 0, 17, 9 として data1～data1387 との類似度を求めランキングを作成した。その結果と正解 data の情報より精度を評価した。さらに, 処理時間（プログラムの実行時間）も計測し, 評価した。

また, 2.2 で述べた DTW に関しても key1～key5, data1～data1387, 正解 data によって実験を行い, 精度および処理時間を求め提案手法と比較した。

実験は, CPU: Intel(R) Core(TM)2 Quad, memory: 4GB, OS: Linux(Ubuntu)の PC で行った。

4.2 実験結果

4.2.1 空間領域の階層性の導入結果

空間領域の階層性の有効性を示すため, a を 0 に固定し, b=0, 17, 9 として検索実験を行った。key1～key5 の実験結果を Recall-Precision 曲線で表したものを図 8 から図 12 に示す。

各図において, My は提案手法, DTW は DTW による手法を表し, My の横の数字は b の値を表す。

また, 各検索キーの処理時間を表４に示す。

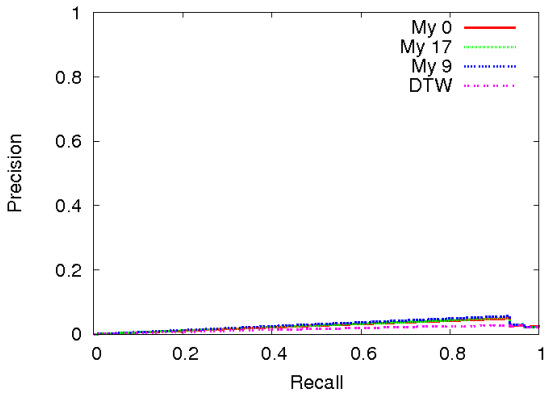

図８　検索キーkey1 の実験結果（a=0）

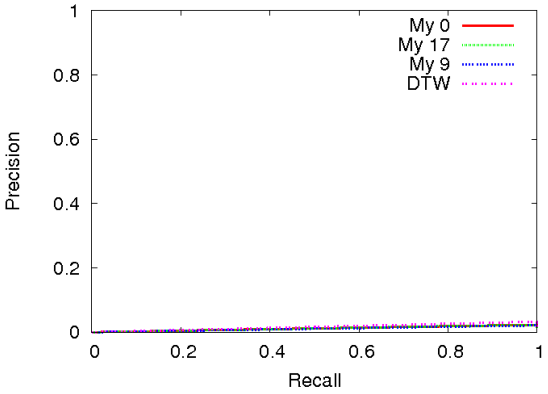

図９　検索キーkey2 の実験結果（a=0）

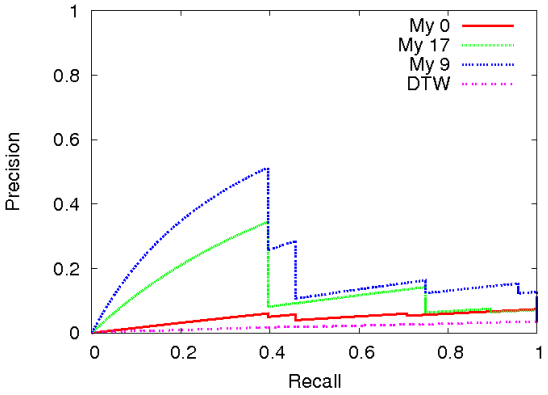

図１０　検索キーkey3 の実験結果（a=0）

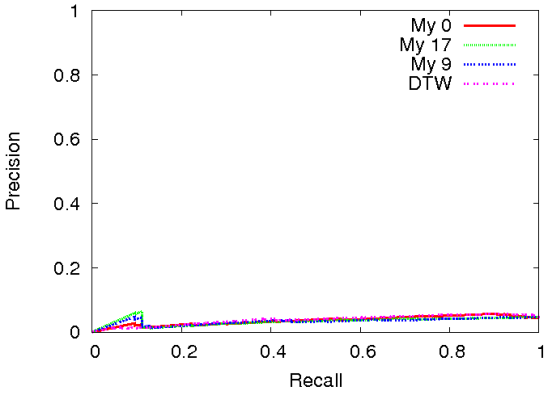

図１１　検索キーkey4 の実験結果（a=0）

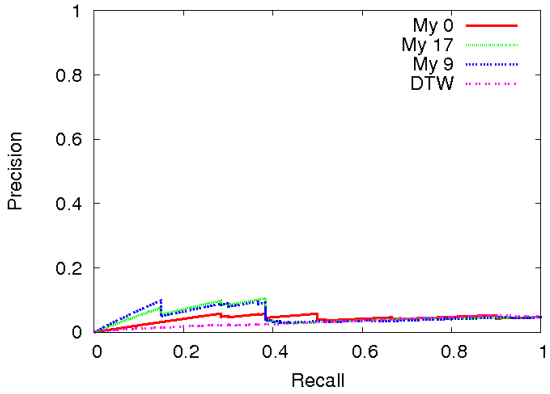

図１２　検索キーkey5 の実験結果（a=0）

表４　a=0 の各検索キーの処理時間[sec]

検索キー	My 0	My 17	My 9	DTW
key1	4.9	7.3	7.3	22.7
key2	5.1	5.9	6.0	12.7
key3	5.5	6.4	6.6	33.0
key4	2.6	3.4	3.4	40.6
key5	3.3	4.1	4.1	39.7

各検索キーを見ると精度は提案手法が DTW と同程度かそれ以上となった。しかし, 全体的に提案手法も DTW も精度が悪い。

key1, key2, key4, key5 のランキングの上位の data は, 手の動きが検索キーと視覚的に類似していないものであった。一方, key3 に関しては, D_0では視覚的に類似していないデータが上位に位置していたが, D_{17}, D_9では視覚的に類似しているデータが上位に位置している。さらに動きへの空間領域の階層性が導入されていないD_0よりもD_{17}, D_9の精度の方が良い。これは, key3 が他の検索キーと比べて空間領域の階層性の導入に合った動きの性質を持っている可能性が考えられる。また, 手作業でデータを取得した際に発生した数ピクセル単位のノイズの混入が key3 のみ少ないといった理由も考えられる。

さらに, 全検索キーにおいてノイズに敏感な DTW の精度が悪いことから全検索キーにはノイズが存在し, 提案手法も DTW と同様に精度が悪いことから空間領域の階層性の導入では前述したノイズの影響を抑えることが出来なかった可能性が考えられる。

処理時間に関しては, 提案手法の方が DTW よりも良い結果となった。動きへの空間領域の階層性が導入されていないD_0よりもD_{17}, D_9の方が長く時間がかかっているが, これは割り当てのための時間が加わっているためである。

4.2.2　時間領域の階層性の導入結果

時間領域の階層性の有効性を示すため, b を 0 に固定し, a を 0 から（key の時系列長-1）に変化させ検索実験を行った。key1～key5 の実験結果を図 13 から図 17 に示す。それぞれの結果に対して, 横軸はTH_aの a, 縦軸はランキングの上位 n 位以内に該当するデータに対する正解率を表す精度 precisionおよび処理時間 time である。ここで, n は各検索

キーの正解データの個数を表し, key1 では 30, key2 では 20, key3 では 48, key4 では 63, key5 では 60 である。また, 例えば n=20 のとき, ランキングの上から 20 番目と 21 番目, 22 番目のデータの類似度が同じ場合, 21 番目と 22 番目のデータも上位 20 位以内に含めている。したがって, この場合上位 20 位以内のデータ数は 22 個となり, 正解データ数が 10 個であれば$\text{presicion} = 10/22 = 0.45$となる。

各図において, My は提案手法, DTW は DTW による手法を表す。

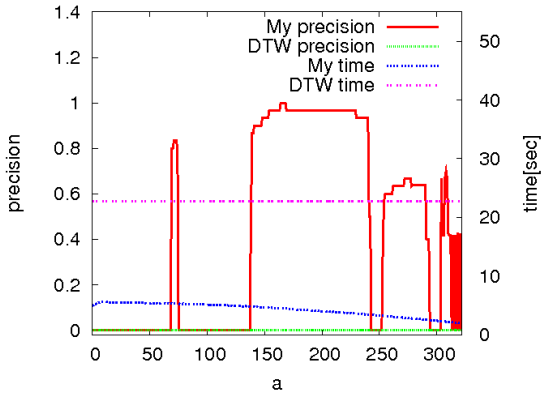

図１３　検索キーkey1 の実験結果（b=0）

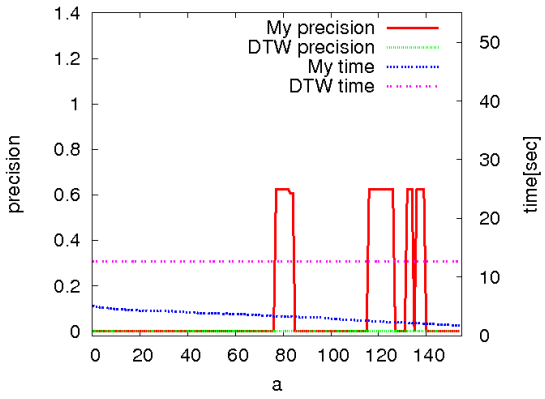

図１４　検索キーkey2 の実験結果（b=0）

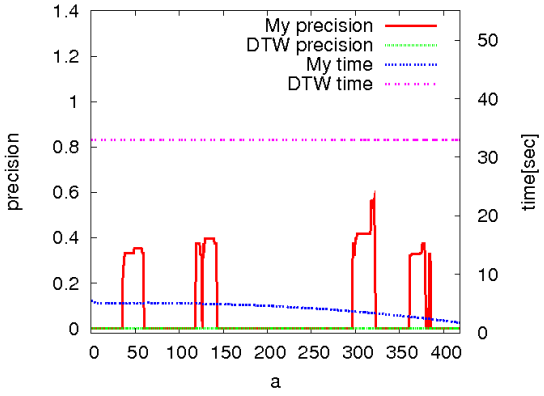

図１５　検索キーkey3 の実験結果（b=0）

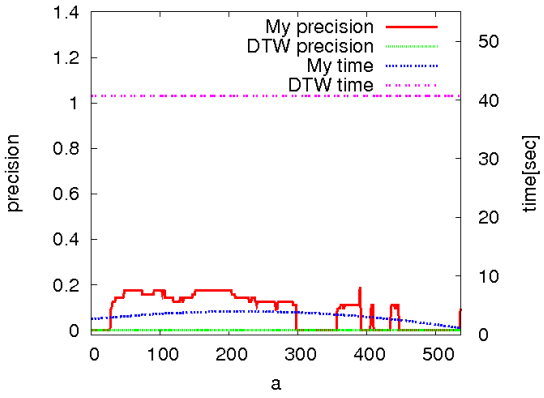

図１６　検索キーkey4 の実験結果（b=0）

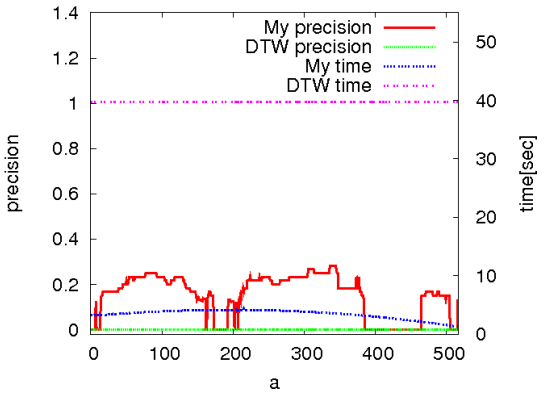

図１７　検索キーkey5 の実験結果（b=0）

precision に関しては, 提案手法の方が DTW よりも良い結果となった。特に key1（図 13）と key3（図 15）に関しては, a の値によっては precision が最大値の 1 となることが確認できる。一方, DTW の precision は非常に低いものであった。原因としては, 今回手作業でデータを取得したため数ピクセルの誤差など, key および data にノイズが多く含まれたためだと考えられる。一方, 提案手法では全ての key である程度正解が得られていることから, 時間領域の階層性の導入によってノイズに強くなったと考えられる。

また, 全ての検索キーにおいて, 動きへの時間領域の階層性が導入されていないTH_0よりも a がある程度の大きさの値の方が良くなっている。このことから, 動きへの時間領域の階層性の導入は有効である可能性が示唆される。

しかし, a をどの値に設定すれば良いかはこの結果からは判断するのは困難である。また, a の値によって precision の値が極端に異なっている。これは, ランキングにおいて隣り合う data は同じか近い類似度をもつことが原因だと考えられる。例えば, key1 の a=210 のランキングの 1 番目は data188 で, 25 番目の data215 まで data189, data190, …と順番にかつ同じ類似度で並んでいる。同様のことが各ランキングでよく見られる。仮に, ランキングの上位にあるこのような data 群が正解であれば precision は大きくなるが不正解であれば precision は小さくなり, 結果的に a の値によって precision が大きく異なる。このような現象が頻繁に起こっていると考えられる。

処理時間に関しては, 提案手法の方が DTW よりも短い。どの key に対しても提案手法は 5 秒程度であるが, DTW に関してはどの key でも 12 秒以上かかり, さらに表 1 の key のフレーム数と正の相関を持つような結果となった。

また, 提案手法において, フレーム数が比較的大きい key4（図 16）と key5（図 17）を見ると time が緩やかに上に凸な曲線であることが分かる。これは, オプティカルフローを合成する回数に依存する処理時間と類似度を求める回数に依存する処理時間がトレードオフの関係にあるからであると考えられる。

5. まとめ

本稿では, 動画の持つ階層性を考慮し, 動きに空間領域と時間領域の階層性を適応した, 動きの類似性判定の手法を提案した。そして, 茶道でのお点前の動画中の点前者の右手の動きを手動で取得し, 検索キー用のデータと比較用のデータに分け類似検索を行う評価実験を実施した。その結果, 時間領域の階層性においては, この階層性を導入しなかった場合と比べて全ての検索キーで精度が向上した。しかし, key4, key5 では precision が 0.5 を満たしていないように, 精度はまだ不十分であると言える。

今後は, 新たな空間領域の階層性の導入の検討, 半自動による手の動き抽出, 類似判定に次元圧縮を用いる手法などとの比較を行っていきたい。

謝辞

本研究は, 一部, 文部科学省科学研究費補助金（課題番号：20300037）による。

参考文献

[1] 柴田正啓：“映像の内容記述モデルとその映像構造化への応用”, 信学論, Vol.J78-D-Ⅱ, No.5, pp.754-764（1995）
[2] 是津耕司, 上原邦昭, 田中克己：“時刻印付きオーサリンググラフによるビデオ映像のシーン検索”, 情処論, Vol.39, No.4, pp.923-932（1998）
[3] 加藤光幾, 石川博：“動画像を対象とする内容検索方式”, 情処研報 DBS, No.111-12, pp.87-94（1997）
[4] 山名信弘, 井辺昭人, 三浦文裕, 前島謙宣, 森島繁生：“動画の３次元周波数成分を用いた顔認証システム”, 信学技報, PRMU2006-22, MI2006-22, pp.13-18（2006）
[5] D.J. Berndt and J. Clifford：“Finding patterns in time series: A dynamic programming approach,” in Advances in Knowledge Discovery and Data Mining, ed. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, pp.229-248, AAAI（1996）
[6] 國枝孝之, 脇田由喜, 高橋望：“MPEG‐7 と映像検索―マルチメディア情報検索の新手法を詳述”, CQ 出版, 第１版（2004）
[7] 柴田正啓：“映像の内容記述モデルとその映像構造化への応用”, 電子情報学会論文誌, Vol.J78-D-Ⅱ No.5, pp.754-764（1995）