データベースシステム 119−46
（1999. 7. 22）

カメラモーションに基づく類似動画像検索

遠藤 斉† 片岡 良治†

本稿では、動画から抽出できるカメラモーションを特徴量として利用したスポーツ映像の内容検索手法を提案する。スポーツ映像にはシーン特有のカメラワークが存在することが多いため、一連のカメラワークを手がかりにシーンの内容に基づいた検索を行えることが期待できる。そこで本稿では検索キーとして指定した映像と検索対象の映像から特徴量としてカメラモーションを抽出し、連続 DP マッチングを適用することによりカメラモーションの類似したシーンを検出する手法を提案する。実際の野球中継の映像を用いて適合率と再現率を評価した結果、その有効性が明らかになった。

Content-based Video Retrieval based on Similarity of Camera Motion

Hitoshi Endoh†, Ryoji Kataoka†

This paper proposes a method for content-based sports video retrieval using camera work information. Since particular camera work for a typical scene exists in sports video, a transition of camera work becomes an effective cue for retrieving a sports scene based on its content. Therefore, the proposing method extracts a series of camera parameters from both a user-specified scene of a retrieval key and a video stream of a retrieval target, and detects scenes having a similar content to the key from the target applying the continuous DP matching. It is evaluated using a video stream of a baseball game. Recall-Precision curves make its effectiveness clear.

1. はじめに

MPEG に代表される映像の圧縮符号化技術の発展に伴い、大量のデジタル映像がデータベースに蓄積･管理されるようになり、利用者が所望の映像を内容に基づき検索できる映像データベースシステムのニーズが高まってきている。映像の内容検索を実現するためには、検索を行うための 2 次情報を個々の映像に付与する必要があるが、映像は一般に単体でも膨大な情報量を有するため、2 次情報を人手で付与することは現実的ではない。そのため、映像を構成する画像から、色情報や輝度情報、カメラワークや被写体の動きなどを特徴量として自動的に抽出し、特徴量の類似性をもとに所望の映像をデータベースから検索する手法が盛んに研究されている[1-5]。

本稿では、動画から抽出できるカメラモーションを特徴量として利用したスポーツ中継の映像の内容検索手法を提案する。スポーツ映像には、シーン特有のカメラワークが存在することが多いため、一連のカメラワークを手がかりにシーンの内容に基づいた検索を行えることが期待できる。例えば、野球中継におけるホームランシーンでは、バッターが打ったボールをカメラで追いかけ、ボールが落下したスタンドをズームアップし、次にダイヤモンドを回るバッターを追う、といった典型的な一連のカメラワークが存在する。サッカー映像においても同様であり、例えばコーナーキックのシーンでは、コーナーからゴール前へ蹴り込まれたボールを追いつつゴール前をズームアップするといった典型的なカメラワークが存在する。

映像を内容に基づいて検索する手法として、これまで以下のような手法が提案されている。1 つは、映像の内容と抽出された特徴量との

†NTT サイバースペース研究所
NTT Cyber Space Laboratories

関係をルール化する手法である。この手法では、映像の内容を特徴量の組み合わせに分解した辞書を用意しておくことによって映像の内容に基づいた検索を実現する。例えば、文献[1]ではサッカー映像に対して、認識技術を用いて選手の位置と短時間動作の内容を特徴量として抽出し、辞書として「ヘディングシュート」＝「選手(ゴール前、ジャンプ)」+「キーパー(ゴール前、ダイブ)」のようなルールを用意することによってプレーの内容に基づいた検索を行うことができる。しかし、ここで必要とされる特徴量の抽出には高度な画像処理技術や認識技術が必要であり、現状の認識技術の精度の問題を考えると、現実的な解とはいえない。

より現実的な解として、直接的に特徴量の内容を検索キーとする手法と、所望の内容を表すシーン映像を検索キーとする手法がある。前者では、ターゲットとなる映像から特徴量として被写体の色･形･移動方向などを抽出しておき、利用者は被写体の色や移動方向を検索キーとして指定することにより、対応するシーンを検索する [2]。しかし、この手法では、検索キーとして指定可能な動きは、例えば、「赤いものが左から右へ」というような単純なものに限られており、先に述べたホームランシーンのような複雑な動きをもつシーンを表現することは難しいという問題がある。後者では、検索キーの映像とターゲットの映像の特徴量マッチングにより検索を行う。文献[4]では、利用者の指定したシーン映像から色情報等を特徴量として抽出し、特徴量ヒストグラムを作成し、ターゲットの映像から抽出した特徴量を順次走査して検索キーと特徴量ヒストグラムのシーンを切り出すことによって、対応するシーンを検索する。しかし、この手法では、特徴量ヒストグラムが動きの順序に関する情報を含まないため、特徴量の変化の順序性を考慮した検索はできない。また、対応するシーンの時間的な長さの違いを考慮していないため、例えば、ホームランのリプレーにおけるスロー再生のように、内容が同じでもシーンの長さが異なる場合には対応できないという問題がある。

本稿では、検索キーとしてシーン映像を用いることにより検索したい映像の内容を指定する手法において、シーンの時間的な伸縮と動きの順序性を考慮した特徴量マッチングを行うことにより、より効果的に映像の内容に基づいた検索を実現する手法を提案する。

2. 提案する類似映像検索手法

2.1. 処理の流れ

本稿で提案する手法の処理の概要を図 1 に示す。本手法では、検索対象の映像と、検索キーのシーン映像から特徴量としてカメラモーションを抽出し、抽出した特徴量に対して時系列のマッチング処理を行うことにより、類似したシーンを検出する。

以下に提案手法の流れを説明する。

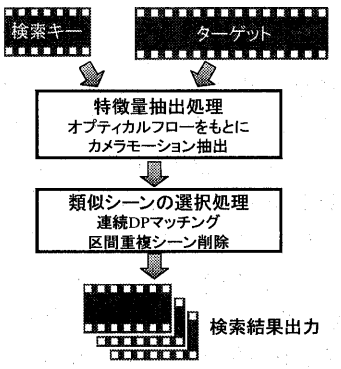

図 1:処理の概要

(1)検索キーの指定

利用者は検索したいシーン映像と同じ内容のシーン映像を用意する。

例：野球のホームランシーン、サッカーのコーナーキックシーン

(2)特徴量の抽出

ターゲットの映像および検索キーとして指定した映像からオプティカルフローをもとに

カメラモーションを抽出する。抽出する特徴量は、カメラの操作に伴う背景の動きを表すグローバルモーションと、被写体の動きを表すローカルモーションとする。

本稿ではオプティカルフローとしてMPEG 符号化情報の動きベクトルを用いて特徴量を抽出する。

(3)類似シーンの選択

検索キーとターゲット間の特徴量マッチングを行い、ターゲットから検索キーに類似したシーンの検出を行う。時系列的な特徴量のマッチングを行う方式はこれまでにいくつか提案されているが、本稿で提案する方式では後述する理由により連続 DP マッチング[6]を用いる。

連続 DP マッチングによる検出では、区間が重複するシーンが多数検出されるため、区間の重複するシーンを削除した上で類似性の高いシーンの出力を行う。

2.2. 特徴量の抽出

オプティカルフローに基づき動画からカメラモーション(パン操作やズーム操作などの度合いをあらわす情報)を求める手法として、本稿では MPEG 符号化情報の動きベクトルを利用してカメラモーションを推定する文献[7]の手法を用いる。紙面の都合上、詳細は省略するが、P ピクチャについて、各マクロブロックの中心画素の位置を (x_i, y_i)、動きベクトルを (u_i, v_i) とするとき、(u_i, v_i) が定点カメラの操作に伴う背景の動き(グローバルモーション)をあらわすならば、次式が成り立つ。

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = G_z \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} G_x \\ G_y \end{pmatrix} \qquad (1)$$

ここで、G_x, G_y, G_z はそれぞれ水平方向のパン操作、垂直方向のパン操作、ズーム操作の度合いを表し、フレームを構成するすべてのマクロブロックに対し式(1)を最小 2 乗法で推定することによって求める。ただし、すべてのマクロブロックがグローバルモーションをあらわすわけではないので、推定値から閾値以上はなれているマクロブロックは被写体の動き(ローカルモーション)をあらわすものとみなし除外する。

ローカルモーションとみなしたマクロブロックの動きベクトルの平均値を (L_x, L_y) とあらわし、特徴量マッチングに活用する。

動画を構成する i 番目の P ピクチャから抽出できる特徴量を

$$f(i) = (G_x, G_y, G_z, L_x, L_y) \qquad (2)$$

とするとき、検索キーS およびターゲット T から抽出した特徴量の時系列 F_S および F_T は次のように表現できる。

$$F_S = f_S(1) \prec f_S(2) \prec \cdots \prec f_S(N_S) \quad (3)$$

$$F_T = f_T(1) \prec f_T(2) \prec \cdots \prec f_T(N_T) \quad (4)$$

ここで、N_S および N_T はそれぞれ S および T に含まれる総 P ピクチャ数である。

2.3. 特徴量マッチング

特徴量の時間的な変化をもとに類似したシーン映像を検出するためには、比較対象を時間的に伸縮させながらマッチングを行う機能と動きの順序性に対応する機能が重要である。本稿ではそのような機能を有する連続 DP マッチングを応用して特徴量マッチングを実現する。

連続 DP マッチングでは、F_T の部分区間に対し、F_S をパターン間距離がもっとも小さくなるように伸縮させながら照合を行い、F_S の終端 $f_S(N_S)$ を $f_T(k)(1 \le k \le N_T)$ に対応付けたときのパターン間距離 $D(k)$ を求める。$D(k)$ は特徴量 $f_S(i)$、$f_T(j)$ 間の距離 $d(i, j)$ に伸縮による重みを掛けたものの積

算であり、以下のようにして求める。

$$D(k)=\frac{1}{k-k'+N_S}g(k,N_S) \quad (5)$$

初期条件

$$g(\tau,0)=\infty(1\le\tau\le N_S) \quad (6)$$

$$g(0,\tau)=2d(0,\tau)(0\le\tau\le N_T) \quad (7)$$

漸化式

$$g(i,j)=\min\begin{pmatrix} g(i-1,j-1)+2d(i,j) \\ g(i-1,j-2)+3d(i,j) \\ g(i-2,j-1)+3d(i,j) \end{pmatrix} \quad (8)$$

ここで、$j=0$に到達したときのiの値をk'とする。また、$d(i,j)$を以下のように定義した。

$$d_{pan}=(G_x(i)-G_x(j))^2+(G_y(i)-G_y(j))^2 \quad (9)$$

$$d_{zoom}=(G_z(i)-G_z(j))^2 \quad (10)$$

$$d_{local}=(L_x(i)-L_x(j))^2+(L_y(i)-L_y(j))^2 \quad (11)$$

$$d(i,j)=\sqrt{\omega_1 d_{pan}+\omega_2 d_{zoom}+\omega_3 d_{local}} \quad (12)$$

ω_1、ω_2、ω_3により、パターン間距離計算におけるグローバルモーションとローカルモーションの影響の度合いを調整できる。パン操作とズーム操作は独立に起こることを考慮し、d_{pan}とd_{zoom}に別々の重みを与えられるようにしてある。パターン間距離は、類似性が高いほど小さい値になる。

検索結果の出力は、$D(k)$をすべての$k(1\le k\le N_T)$について求め、区間の重複するシーンを削除し、パターン間距離の小さい順にシーンを選択することによって行う。

$D(k)$はすべてのPピクチャについて計算されるため、区間の重複するシーンが多数存在する。このため、重複する区間を削除する必要がある。

区間の重複するシーンの削除については、シーンの開始Pピクチャ番号k'の差が閾値thr1未満の区間について、もっともパターン間距離の小さいもののみを採用することによって実現する。

3. 実験と評価

3.1. スポーツ映像への適用

以上に述べた特徴量マッチングによる類似シーン映像検索の効果を検証するために、実際のプロ野球映像を用いて検証を行った。実験のターゲットとして使用した映像情報は、約1時間40分の野球中継をMPEG1符号化したもの(フレーム数171,502、約1GB)であり、46,476個のPピクチャを含む($N_T=46,476$)。また、ターゲットからホームランと内野ゴロのシーンをそれぞれ1つずつ選び、これら2つを検索キーとした。実験に用いた検索キーの詳細は表1の通りである。表1における同一意味内容のシーン(対象シーン)数には、スロー再生によるリプレーシーンも含まれている。

表1:実験に使用した検索キー

検索キー	ホームラン	内野ゴロ
画像サイズ	352 x 240	
長さ(秒)	18	7
Pピクチャ数	148	63
同一内容のシーン数	13	11

ωの値については、$\omega_1:\omega_2:\omega_3$が(i)1:1:1(グローバルモーション、ローカルモーション共に寄与)、(ii) 1:1:0(グローバルモーションのみ)、(iii) 0:0:1(ローカルモーションのみ)の3つの場合について行った。ただし、d_{pan}、d_{zoom}、d_{local}については寄与が同等になるようにあらかじめ分散をそろえておく。また、区間の重複するシーンを削除するための閾値としてthr1=10を用いた。

得られた再現率－適合率グラフを図2、図3に示す。

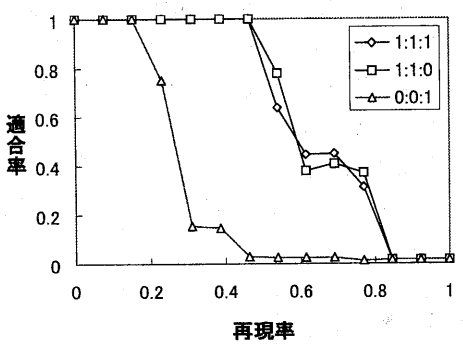

図 2:ホームラン検索

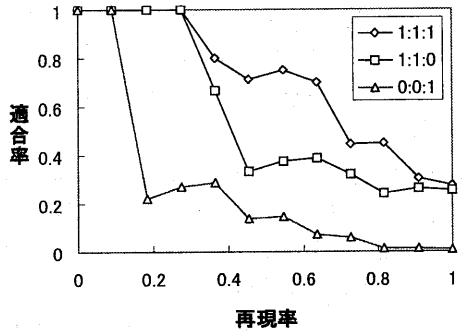

図 3:ゴロ検索

ホームラン検索では、$\omega_1:\omega_2:\omega_3$が 1:1:1 の場合と 1:1:0 の場合に比較的精度のよい結果が得られ、0:0:1 の場合は比較的精度が悪かった。1:1:1 の場合と 1:1:0 の場合のグラフの概形の違いが少ないこと、0:0:1 の場合の検索精度が比較的良くないことから、ホームラン検索についてはグローバルモーションの影響が主体であり、ローカルモーションの影響は小さいということがわかる。実際、ホームランシーンの映像を見ると、特徴的な動きはカメラワークにあり、これは実験結果と一致する。再現率が 0.8 以上で適合率が極端に小さな値になっている。これは、対象となるホームランシーン 13 件のうち、3 件のシーンが検索キーとまったく違うカメラワーク(ホームランが飛び込むスタンド側からボールを追った映像など)で撮影されていることが原因であり、単一の検索キーとのマッチングにより類似シーンを検出しようとする限り避けられない。検索結果の上位には、リプレーのスロー再生のシーンも含まれており、連続 DP マッチングを採用した効果が確認できた。誤検出のシーンには、外野フライが多く含まれていた。外野フライには、バッターが打ったボールをカメラが追いかけ、ボールをキャッチした外野手をズームアップするといったカメラモーションがあり、ホームランのカメラモーションに類似しているためと考えられる。

ゴロ検索では、検索精度の良い順に 1:1:1 の場合、1:1:0 の場合、0:0:1 の場合となった。ホームラン検索と同様にグローバルモーションの影響が主体であるものの、1:1:0 の場合に比べて 1:1:1 の場合の検索精度が良くなっていることから、ローカルモーションも精度向上に寄与していることが分かる。実際のゴロのシーンを見ると、ホームランに比較してカメラの動きは小さく、また、内野手が 1 塁に送球するといった特徴的な被写体の動きがあり、これは実験結果と一致する。

以上の結果から、カメラモーションを特徴量として連続 DP マッチングを行うことにより映像の内容に基づいた検索を行うことができることを確認できた。ホームランと内野ゴロを検索キーとした検索では、グローバルモーションによる影響が主であり、ローカルモーションによる寄与は比較的小さいことがわかった。

3.2. ローカルモーションを用いた検索

上述した野球中継におけるホームラン検索と内野ゴロ検索は映像の特徴量としてグローバルモーションが主体の検索であった。そこで、ローカルモーションの効果を確認するために、固定カメラ(パン、ズームなし)で撮影したシーンを検索キーとして検索を行った。

検索キーとして用意したのは、固定カメラの前で手を大きく2回転させたシーン映像である。検索対象は、3.1.で用いた約1時間40分の野球中継映像である。

$\omega_1:\omega_2:\omega_3$=1:1:1 として検索を行った結果、検索結果の上位には、バッターの素振りシーン(バットを 2 度回転させる)が複数個検出された。

$\omega_1:\omega_2:\omega_3$=1:1:0 とした場合はグローバルモーションの小さいシーンが検出されるものの、被写体の動きが検索キーと類似したシーンは上位には検出されなかった。

この実験はシーン映像のセマンティックに基づいたものではないため、十分な評価を行うことはできないが、本稿で提案した手法を、被写体の動きに基づいた検索にも適用できる可能性があることを示している。

4. まとめ

本稿では、映像間の時系列的な特徴量マッチングにより、検索キーとして指定したシーンに類似するシーンを検索する手法を提案した。特徴量として動画から抽出されるカメラモーションを利用し、スポーツ映像を対象に映像の内容に基づいた検索が可能であることを示した。シーンの時間的な伸縮と動きの順序性を考慮した特徴量マッチングを行うことにより効果的な検索を実現する。利用者は、検索したい内容を表すシーン映像を入力することによって内容に基づいた検索を行うことができる。

実際の野球中継の映像に対する検索を行って提案手法の有効性を評価した。ホームランのシーンと内野ゴロのシーンを検索キーとした検索を通して、映像の内容に基づいた検索を行うことができることを確認した。また、検索する内容によってグローバルモーションとローカルモーションの寄与の大きさが異なることが分かった。

さらに、固定カメラで撮影したシーンを検索キーとした検索を行い、提案手法を被写体の動きに基づいた検索にも適用できる可能性があることを示した。

今後は高速に検索を行うための特徴量のインデックス作成方式について検討を行う予定である。

参考文献

[1]宮森恒ほか：シーン中の短時間動作記述を用いた映像内容検索方式の提案, MIRU'98, I-75-80, 1998

[2]加藤光幾, 石川博：ビデオデータを対象とする異種検索方式の統合システム, 第 10 回データ工学ワークショップ論文集, 4B-3, 1999.

[3]S.Chang et al.：VideoQ: An Automated Content Based Video Search System Using Visual Cues, Proc.ACM Multimedia'97, 1997

[4]Gravin Smith, Hiroshi Murase, Kunio Kashino：QUICK AUDIO RETRIEVAL USING ACTIVE SEARCH, ICASSP1998, 1998.

[5]宮坂圭, 吉田俊之：MPEG ビットストリーム中の動きベクトルの動画像検索への応用, Proc. 1999 IEICE General Conf. D, 1999.

[6]岡隆一, 連続 DP を用いた連続単語認識, 日本音響学会, 音研資, S-78-20, 1978

[7]J. Meng and S.Chang：CVEPS – A Compressed Video Editing and Parsing System, Proc.ACM Multimedia'96, 1996