投球の次ショットに重きを置いたシーンのパターン化と離散隠れマルコフモデルを用いた野球放送映像の自動イベント分類

Automatic Event Classification in Baseball Broadcast Videos Using Scene Patternization Focusing on Post-Pitch Shot and Discrete Hidden Markov Models

正会員 望月 貴裕[†], 正会員 藤井 真人[†], 正会員 八木 伸行[†], 篠田 浩一^{††}

Takahiro Mochizuki[†], Mahito Fujii[†], Nobuyuki Yagi[†] and Kouichi Shinoda^{††}

Abstract A method has been developed for automatically classifying baseball video scenes into some events that describe their content. The baseball scenes are patternized using a set of rectangles with image features and motion vectors. The basic unit for patternization is a shot. For the second shot of each scene which includes significant information for event-classification, a partial shot generated by dividing the shot is used as a processing unit. The scenes used for training are expressed as sequenced symbols based on the patternized data for shots and partial shots. “Event-unknown” baseball scenes are assigned “event-indexes” (i.e., homerun, single, walk, etc.) using discrete hidden Markov models that have been trained with the training symbol sequences for each kind of event. An experiment using videos of seven Major League Baseball games produced good results, demonstrating that this method can automatically classify events with high accuracy.

キーワード：映像分類，ショット，パターン化，離散 HMM，野球映像

1. ま え が き

近年のデジタル技術の発達により，デジタイズされた映像データをハードディスクなどに大量に蓄積することが可能となった．それに伴い，映像を内容に基づいて管理および検索する技術が求められている¹⁾．特に，スポーツ映像は生起するイベント種が比較的区別しやすく，ユーザの「何がしたいか」という要求も明確であるため，検索技術の適用が強く望まれている．よって，スポーツ映像のシーンを内容によって自動的に分類する，ハイライトを抽出する，あるいは映像を要約するための様々な研究が行われている．馬場口らはアメリカンフットボールを処理対象とし，映像に付与されているクローズドキャプションと簡易な色特徴を用いて，各シーンを数種類のイベント種へ分類する技術

を提案している²⁾．この手法は，クローズドキャプションの正確さに強く依存しており，その信頼度が低い映像，あるいは全く付与されていない一般映像への適用は難しい．渡辺らは，野球映像の各ショットを，フレーム画像を分割したサブブロックにおける動きベクトルのマハラノビス距離の学習に基づいて，投球ショットとそれ以外のショットに分類している³⁾．しかし，映像から抽出する画像情報量および学習データ量が充分でないため，分類対象として扱うことのできるイベント種が大きく制限される．Ekin らは，フレーム画像中の芝生領域の割合およびペナルティボックスの位置を画像情報としてサッカー映像のシーンの分類を試みている⁴⁾．しかし，簡易な特徴のマッチング処理のみに基づいた分類処理であるために分類精度が低く，それを向上させるためには，学習処理を強化する必要がある．

隠れマルコフモデル (HMM) は，時系列で推移する事象の解析に非常に適したモデルであり，映像から抽出した特徴の推移に基づく HMM を利用した様々なシーン分類，イベント認識手法が提案されている．Xu らは，ブロック単位の動きデータを画像特徴として HMM に学習させ，バスケットボール映像のイベント認識を行っている⁵⁾．しかし，この手法は，大部分が水平方向のカメラの動きで構成されているような，バスケットボール特有の条件を満たす映像のみを処理対象としている．Xie らが提案している手法は，

2006 年 12 月 25 日受付，2007 年 4 月 16 日再受付，2006 年 5 月 22 日採録

[†] NHK 放送技術研究所

(〒 157-8510 世田谷区砧 1-10-11, TEL03-5494-5173)

^{††} 東京工業大学 情報理工学専攻 計算工学専攻, TEL03-5734-3526

(〒 152-8552 目黒区大岡山 2-12-1)

[†] NHK (Japan Broadcasting Corporation) Science and Technical Research Laboratories

(1-10-11, Kinuta, Setagaya-ku, Tokyo 157-8510 Japan)

^{††} Department of Computer Science, Tokyo Institute of Technology (2-12-1, Ookayama, Meguro-ku, Tokyo 152-8552 Japan)

フレーム画像の主要色領域の比率や動き情報の推移に基づいて構築した HMM を用いてサッカー映像のシーン分類を試みているが⁶⁾、「プレイシーン」と「ブレイクシーン」の2種類のみの判別にとどまっている。Rca らは、テニス映像を処理対象とし、コート上のラインおよび選手の位置をフレーム画像から抽出し、その推移を学習した HMM を用いて、得点シーンなどのシーン分類を行っている⁷⁾。しかし、この手法は、プレイ中の映像がフィールドのルーズショットのみで構成されていることを想定しており、選手のアップや観客席の映像などが挿入される一般の放送映像への適用は困難である。

HMM によるスポーツ映像のシーン分類については、野球放送映像を対象とした手法も多く提案されている。野球放送映像は、ホームラン、ツーベース、シングルヒット、四球などのイベント毎にカメラの切り替えなどによる画の流れのパターンが異なる傾向があり、HMM を用いた学習および認識処理に非常に適しているためである。その手法のひとつとして、各シーンにおけるフレーム単位での画像特徴の推移を学習した HMM を用いてシーンのイベント種を判定する手法が提案されている⁸⁾。フレームを推移の単位とするこの手法は、時間方向の伸縮（例：投手が一旦プレートを外して投球間隔が伸びる場合など）に対して影響を受けやすい。また、カメラのスイッチングミスや物体の横切りなど想定外の短区間映像の挿入により、分類精度が低下すると考えられる。さらに、HMM のモデルとして連続 HMM を採用しているため、学習時の計算コストが比較的高くなる。一方、連続 HMM に比べて計算コストの低い離散 HMM を用いた手法も提案されている。各シーンを構成するショット毎にフレーム単位の画像特徴を計算し、その画像特徴系列に基づいてショットを7種類のシンボルのいずれかに割当て、シーンをシンボル列で表現し、シーンの分類を試みる手法が提案されている⁹⁾。この手法は、シンボル化のための画像特徴抽出単位がフレームであり、ショットが持つ画像特徴系列のバリエーションが広がると考えられる。それにもかかわらず、ショットに割当てたシンボルの種類が固定でかつ少ないため、多種のイベント分類が難しい。

本論文では、フレームを推移の単位とするのではなく、イベントシーンの各ショットを基本単位とする。特に、投球の次のショット（以下「第2ショット」）について、ショットを一定間隔で分割した「部分ショット」を推移の単位とすることを特徴とする。そしてショットおよび部分ショットのパターン化によりイベントシーンをシンボル列表現し、イベント毎のシンボル列集合を学習した離散 HMM を用いて、野球放送映像のシーン分類を行う手法を提案する。パターン化処理は、計算コストおよび分類精度に影響するコードブックのシンボル数を抑えながら、より多くのイベント種の分類を可能にする。さらに、イレギュラーな時間軸方向の伸縮や区間映像の挿入などの「ノイズ」が生じた映像に対しても、精度を低下させることなく、ロバストなイベン

ト自動分類を実現する。また、第2ショットを分割することにより、イベントを区別するための重要な情報を含む映像区間の画の推移やカメラワークのパターン表現が詳細となり、分類精度の向上が期待できる。第2ショットを分割しない場合との比較実験によりそれを実証する。

2. イベントシーンの表現手法と学習処理

本手法では、HMM による学習処理を用いる。学習用イベントシーンは野球放送映像から抽出する。イベントシーンは複数のショットから構成される区間映像で、イベントシーンの開始ショットは投手の投球ショットとし、終了ショットは学習用映像の分析により、以下の条件のいずれかを満たすものとする。

- リプレイシーンの直前ショット。
- イニングの終わりに「スコアボード CG」がスーパーされたショット。
- 上記以外の場合、次の投球ショットの開始フレームから T_e フレーム遡った時点でのショット。本論文では、野球放送映像のリプレイシーンの一般的な長さを考慮し、 $T_e = 450$ とした。

そして、抽出された学習用イベントシーンを、そのイベント内容によって、可能な限り次の七つのイベント種に振り分ける。

- ホームラン
- ツーベース
- シングルヒット
- フライアウト
- 内野ゴロ（内野安打含む）
- 四球
- 三振

この七種を採用した理由は、画の推移パターンの傾向が比較的掴みやすいイベント種であると予想されるからである。このいずれにも該当しないシーン（例：タイムリーヒット、三塁打、盗塁など）は、出現頻度が低くかつ画の推移パターンのバリエーションが多いため、精度の高い学習データの構築が難しい。よって現段階では自動分類は難しいと考え、本論文では処理対象とせず、今後の課題とする。図1に、イベントシーンの例を示す。(a)がツーベース、(b)がホームランの例である。これらの学習用イベントシーンをパターン列で表現し、それらを用いて HMM の学習を行う。学習処理の流れを以下に示す。各処理の詳細は、以下の各節で述べる。

- (1) 各イベント種のすべてのイベントシーンのショットチェンジを検出し、さらにその第2ショットを固定長の「部分第2ショット」系列に分割する。
- (2) それらのショットおよび部分第2ショットをパターン化処理によりパターン化データで表現する。

図 1 イベント毎の画の推移例
Example of event-dependent picture transitions.

- (3) パターン化データのクラスタリング処理によりパターン化データとシンボルを対応付けるためのコードブックを作成し、コードブックに基づき各イベントシーンをシンボル列へ変換する。
- (4) イベント種毎にシンボル列集合を HMM に学習させる。

2.1 ショットチェンジ検出

本論文の提案手法では、ショット、あるいはショットの中をさらに分割した部分ショットを推移の単位とするため、ショットチェンジ検出処理が必要となる。映像処理の分野において、様々なショットチェンジ検出手法が提案されている¹⁰⁾¹¹⁾が、本論文では、フレーム画像を分割した各ブロックから計算されたフラクタルシーケンス (FSQ)¹²⁾の変化を検出の尺度としたショットチェンジ検出¹³⁾を採用する。この手法を採用した理由は、観客席などの複雑なテクスチャの早い動きによる誤検出を抑えることができ、スポーツ映像に非常に適しているためである。図 2 にショットチェンジ検出手法の概要を示す。Step1 では、入力フレームを $D_x \times D_y$ 個のサブブロックへ分割し (図 2 では $D_x = 4, D_y = 3$)、各ブロックにおける FSQ $v_1, v_2, \dots, v_D$ ($D = D_x \times D_y$) を計算する。ここで、 v_i は i 番目のブロックから計算された FSQ を表す。Step2 では、入力フレームと前フレームの i 番目のブロックに関する類似度 S_i を、 v_i と V_i (前フレームの i 番目のブロックから求めた FSQ) とのユークリッド距離に基づいて求める。そして Step3 では、 $S_i < S_{th}$ であるブロック数を n とし、FSQ が大きく変わったブロックの占有率を表す $R = n/D$ が閾値 R_{th} を超えた場合、入力フレームをショットチェンジ点として判断する。本論文では、予備実験により、 $R_{th} = 0.25$ と定めた。

この手法では、ショットチェンジがディゾルブ、フェード

図 2 ショットチェンジ検出
Shot change detection.

イン、フェードアウト、あるいはワイプなどの特殊なショットチェンジを高精度で検出することは難しい。フレームを随時処理していくリアルタイムショットチェンジ検出の研究分野においては、上記のようなケースにもロバストに働く検出手法がないのが現状である。しかし、本論文で用いたイベントシーンには、そのような特殊なショットチェンジはほとんど含まれておらず、本論文が提案するイベント分類手法へ与える影響はないと考えている。

また、ショットとしてみなすフレーム数の下限値 T_s を設定した。ミススイッチングなどによる短い映像区間の出現や、プレイに無関係なオブジェクト (審判など) が横切るなど、ショットチェンジとして検出すべきでないケースに対処するためである。本論文では、使用するイベントシーン集合中に出現する上記のケースを分析し、 $T_s = 20$ と設定した。

すべての学習用イベントシーンにおける第 1 ショットチェンジ (投球ショットからの切替わり) と第 2 ショットチェンジの検出精度はほぼ 100 % である。野球放送映像においては、これらのショットチェンジはフレーム画像の絵柄が大きく変化する傾向があり、検出が容易なためである。それ以降のショットチェンジについては若干検出精度が低下するが、学習処理により吸収できる程度のエラー量であり、イベント分類精度への影響はほとんどないと考えている。

2.2 第 2 ショットの分割

イベントシーンにおける第 2 ショットは、打球の方向、野手の捕球および送球など、イベントを区別するための重要な情報を多く含んでいる。そこで本論文では、第 2 ショットに含まれる情報を学習データに強く反映させるために、第 2 ショットについては、ショットを T_2 フレーム間隔で分割した「部分第 2 ショット」を処理単位とする。 T_2 は、「内野

図 3 第 2 ショットの分割
Partition of the second shot.

ゴロイベントシーンの第 2 ショット」において、野手が捕球したタイミングで区切ることを主目的として設定する。野手の送球によりカメラの動く方向が大きく変わり、その変化のタイミング周辺で分割するのが望ましいと考えるからである。本論文では、学習用映像の分析により、 $T_2 = 45$ と定めた。第 2 ショットの分割による部分第 2 ショットの生成例を図 3 に示す。

2.3 イベントシーンのパターン列表現

学習データは、イベントシーンの各ショットあるいは各部分第 2 ショットのパターン化処理によって構築される。「パターン化」は、物体と背景を区別するオブジェクトベースの映像表現の一つであり、各フレームの HSV (色相, 彩度, 明度) ヒストグラムに基づいた手法¹⁴⁾などが提案されている。本論文で採用した「パターン化」は、単一色の領域だけではなく、従来手法で扱うことが難しかった複雑なテクスチャ領域 (例: 球場の観客席, ベンチの選手の集まり) をひとつのオブジェクトとしてみなすことが可能である。この手法は、映像内容の簡略表現の一種であり、映像の要約のために用いられることもある。

パターン化処理された区間映像 (ショットあるいは部分第 2 ショット) をパターン化データと呼ぶこととし、以下「PD」と記述する。PD は複数の「矩形構造体」(後述)で構成される。パターン化処理は以下の六つの処理から成る。処理の流れを図 4 に示す。

前処理 処理中の区間映像が含まれている映像ファイル全体から、フレーム間隔 T_{lax} でフレーム画像を切り出してそれぞれをブロック分割する。そして、それらのブロックを画像特徴ベクトルの類似度に従ってクラスタリングし、「テクスチャクラスタ」を形成する。ここで、画像特徴ベクトルは HSV 平均値ベクトルおよび FSQ により構成し、その類似度はベクトル同士のユークリッド距離に基づいて定める。本論文では、 $T_{lax} = 100$ とし、各フレーム画像を $240 \times 180(\text{pixel})$ に正規化し、ブロックの大きさを $20 \times 20(\text{pixel})$ とする。

Step1 区間映像の先頭フレームに、初期「追跡点」を格子状に設置する。本論文では、 $240 \times 180(\text{pixel})$ に正規化されたフレーム画像に対して $192 (= 16 \times 12)$ 個

図 4 区間映像のパターン化処理
Patternization of video.

の追跡点を用いることとする。

Step2 区間映像の最終フレームまで、追跡点近傍から切り出したブロック画像のブロックマッチングによる追跡処理を行う。ブロック画像の大きさは、Step3 の処理においてテクスチャクラスタとの整合性が必要となるため、隣接追跡点とのオーバーラップを許容して $20 \times 20(\text{pixel})$ とする。

Step3 各フレームの全追跡点に、その近傍ブロック画像の画像特徴ベクトルと最も類似度が高い中心ベクトルを持つテクスチャクラスタを割当てて。ただし、閾値以上の類似度のクラスタが存在しない追跡点は除去する。そして各フレームにおいて、同じテクスチャクラスタが割当てられた複数の追跡点同士を「矩形グループ」にまとめる。ここで、グループを組めない「孤立追跡点」はノイズとして除去する。

Step4 矩形グループの位置および割当てられたテクスチャクラスタに基づいて、先頭フレームから時間軸方向に、可能な限り隣接フレーム間で矩形グループを対応付ける。

Step5 矩形構造体から構成される PD を出力する。矩形構造体は、以下の要素を持ったデータである。

位置情報 先頭フレームにおける矩形グループの追跡点群に外接する矩形の頂点座標
画像特徴 割当てられたテクスチャクラスタの中心画

図 5 パターン化データを視覚的に表現した例
Examples of visual expression of patternized data.

像特徴ベクトル

動きヒストグラム 先頭フレームから Step4 で対応付けることができたフレームまでの、グループに属する追跡点の動きベクトル (Step2 のブロックマッチングによる) の方向を 6 種類に量子化して算出したヒストグラム

図 5 に PD の例を示す。ここでの表示は、一目で PD が持つ情報の概要を見られるようにするために、数値データである PD を画像にして視覚的に表現したものである。矩形に敷き詰められたテクスチャは、割当てられたテクスチャクラスタの中心ベクトルにもっとも近い画像特徴ベクトルを持つブロック画像である。また矢印は、動きヒストグラムの中でのもっとも高い数値およびその方向を表したものである。矩形でない形状の領域は、画像描画時の重なりによるものであり、実際のデータは矩形である。

パターン化は、1 章で述べたような想定外の時間方向の伸縮や短区間映像の挿入にほとんど影響を受けない強さを持つ。図 6 は人工的に作成した映像を用いて、パターン化の長所を表したものである。図 6(a) では、下段の投球ショットが上段の約 2 倍のフレーム数であるにもかかわらず、それぞれのショットから生成された PD はほぼ同じである。また、図 6(b) および (c) では、下段の映像には、上段には出現しない想定外短映像区間 (15 フレーム) が挿入されている。しかし、(b)(c) どちらの場合も、上段と下段で極めて類似した PD の生成が実現されている。

また、本パターン化は画像特徴の若干異なる区間映像をほぼ同じ数値データの PD へ変換することができるため、イベントシーン分類を高精度に行うための学習データ量およびコードブックのサイズを抑えることが可能である。

2.4 パターン列のシンボル列への変換

本論文では、離散 HMM を用いた学習を行うために、2.3 節の処理でパターン列表現されたイベントシーンをさらにシンボル列へ変換する。そのためには、PD とシンボルを対応付けるコードブックを作成する必要がある。まず、学習用イベントシーンにおける各 PD を、シーン中での出現位置に基づき、以下の四つの「時間的 PD グループ (以下 TPG)」に自動分類する。ここで、投球ショット直後の部分第 2 ショット (先頭部分第 2 ショット) は「打球の方向」に関する情報を多く含むため、それ以降の部分第 2 ショットと区別して扱うこととし、先頭部分第 2 ショットのみの

図 6 パターン化データの長所
Merit of patternized data.

TPG を構成する。

- A 投球ショットに対応した PD
- B 先頭部分第 2 ショットに対応した PD
- C 2 番目以降の部分第 2 ショットに対応した PD
- D 第 3 ショット以降のショット (後半ショット) に対応した PD

コードブックは、TPG 毎に作成する。異なる TPG 間でのシンボルの種類を分離することで、各シーンにおけるシンボル推移がより確実に表現され、その結果、イベント分類精度が向上すると考えられるためである。図 7 に、コードブック作成の全体の流れを示す。学習用イベントシーンのすべての PD を TPG 毎に自動的に振り分け、各 TPG において PD の階層的クラスタリングを行いコードブックを作成する。この各 TPG におけるコードブック作成処理の概念図を図 8 に示し、詳細を以下に述べる。

各 TPG におけるコードブック作成処理

Step1 TPG 中の全 PD について、パターン化前の区間映像の先頭フレームを切り出し、先頭フレームの画像特徴類似度に基づいてクラスタリング処理を行う。ただし TPG-A については、すべて酷似した画像 (センターカメラから撮影した画像) であるため、このステップにおける分類を行わない。画像特徴としては、画像検索の分野で使用されている、画像を分割した各ブロックから計算した HSV 平均値ベクトルおよび FSQ を採用し、類似度 S_0 は、ブロック毎に計算した HSV 平均値ベクトルおよび FSQ のユークリッド距離に基づき定める¹⁵⁾。 S_0 が S_{th0} 以上である PD 同士を同じクラスタにまとめていく。作られたクラスタのうち、以降の処理の簡易化およびシンボル数抑制のため、要素数が閾値以下のクラスタは除去する。

図 7 コードブック作成の全体の流れ
Outline of generating codebooks.

図 8 各 TPG における階層的クラスタリングによるコードブック作成

Generating codebooks by hierarchic-clustering in each TPG.

図 9 手動で除去すべきクラスター例
Example of clusters to be removed manually.

Step2 本来ならば自動的に次ステップに進むのが望ましいが、現状、Step1での画像クラスタリングの精度が充分ではない。そこで、できるだけ質の良いコードブックを作るために、手動によるクラスターの絞り込み処理を行うこととする。Step1で作成された各クラスター

について、クラスター中心画像特徴ベクトルに近い上位10枚の画像を目視し、コードブックの質を低下させ、分類精度へ悪影響を及ぼすと予想される「素性の悪いクラスター」を手動で除去する。具体的には、別クラスターへ属するのが望ましい異種の画像が混在するが、それらを画像特徴のみで分離することが難しいと思われるクラスターである。その例を図9に示す。図9(a)のクラスターは、外野と内野の画像が混在しているが、画像の類似性の観点ではこれらは酷似しており、画像特徴のみでの分離が難しい。しかしこれらが同一クラスターのまま同じシンボルが割当てられた場合、「内野ゴロ」と「打球が外野へ飛ぶシーン」との分類精度を低下させる恐れがある。また、図9(b)のような、一塁を駆け抜けたバッターランナーの画像と野手の画像が混在するクラスターも、内野ゴロと外野フライの分類に関してノイズとして作用すると推測される。さらに、図9(c)のように、明らかに場面の異なる複数の画像で構成されているクラスターも、各イベント種のシンボル列構成に影響を与え、分類精度全体の低下を招くと考えられる。ここで、本ステップにおいて、正確な処理による「素性の悪いクラスター」の完全な除去を目指す必要はない。下の階層で正確にクラスタリングされるケースが考えられる上に、学習処理により少量の「素性の悪いクラスター」の影響を吸収することが可能だからである。このStep2までで形成されたクラスターを「第0層PDクラスター」とする。

Step3 第0層PDクラスターのそれぞれについて、PDにおける「背景矩形構造体」（面積最大の矩形構造体であり、ほとんどの場合背景領域に該当する）の類似度 S_1 に基づいたクラスタリングを行う。 S_1 が閾値 S_{th1} 以上であるPD同士をまとめていく。本処理においても、要素数が閾値以下のクラスターは除去する。このステップにおいて形成されたクラスターを「第1層PDクラスター」とする。そして、第1層PDクラスターのそれぞれについて、「第1層中心矩形構造体」（クラスターに属するPDの背景矩形構造体の各要素について平均を取ったもの）を計算しておく。本ステップは、主に背景領域の種類（芝、フェンス、観客席など）とその動き（カメラワーク）をコードブックに反映させるものである。

Step4 第1層PDクラスターのそれぞれについて、PDにおいて、「人物矩形構造体」（中央付近に存在する矩形構造体の中で最も面積の大きいものであり、人物領域に該当する可能性が高い）の類似度 S_2 に基づいたクラスタリングを行う。 S_2 が閾値 S_{th2} 以上であるPD同士をまとめていく。ただし、本処理においても、要素数が閾値以下のクラスターは除去する。このステップにおいて形成されたクラスターを「第2層PDクラスター」とする。そして、第2層PDクラスターそれぞれについて

て、「第2層中心矩形構造体」(属するPDの人物矩形構造体の各要素について平均を取ったもの)を計算しておく。本ステップは、人物の画像特徴(ユニフォーム、審判など)やその動きをコードブックへ反映させるものである。また、Step2で例えば、人物のアップ画像と人物の写っていない芝生画像が混在するクラスタが除去されずに残ってしまった場合に、それらを分離する役割も併せ持つ。

Step5 形成された第2層PDクラスタのすべてのクラスタに、他のTPGのクラスタと重複がないようにシンボル(数値ラベル)を割当てる。

どの第2層PDクラスタにも属さないPD用に「unknown」に該当するシンボルも用意する。こうして各TPDで構築されたクラスタとシンボルの対応がコードブックとなる。Step3およびStep4で計算される矩形構造体の類似度 $S_i(i=1,2)$ は、以下の式(1)で定める。

$$S_i = \begin{cases} 1 & (\text{双方のPDに該当矩形構造体なし}) \\ 0 & (\text{片方のPDのみ該当矩形構造体あり}) \\ \frac{(w_{ri}S_{ri} + w_{fi}S_{fi} + w_{mi}S_{mi})}{(w_{ri} + w_{fi} + w_{mi})} & (\text{それ以外}) \end{cases} \quad (1)$$

i : $i=1$ のときが背景に該当する矩形構造体の場合を意味する。 $i=2$ のときが人物に該当する矩形構造体の場合を意味する。

S_{ri} : 矩形構造体の重なり度。矩形の中心を合わせ、重なっている領域の割合として算出。値の範囲は0-1に正規化。

S_{fi} : 画像特徴ベクトルに関する類似度。ベクトルのユークリッド距離に基づき算出。値の範囲は0-1に正規化。

S_{mi} : 動きヒストグラムに関する類似度。動きヒストグラムをベクトルとして考え、ユークリッド距離に基づき算出。値の範囲は0-1に正規化。

w_{ri}, w_{fi}, w_{mi} : 各類似度に関する重み。本論文では、 $i=1$ の場合は「背景」であり、重なり度は重要ではなく、かつ動きが重要であると考え、予備実験により、 $w_{r1} = 0.0, w_{f1} = 1.0, w_{m1} = 3.0$ と設定した。また、 $i=2$ の場合は、どの要素も等しく重要と考え、 $w_{r2} = w_{f2} = w_{m2} = 1.0$ と設定した。

こうして構築されたコードブックを用いて、学習用イベントシーンの各PDをシンボル化する。学習用PDのうち、除去されることなく、第2層PDクラスタのいずれかに属しているものは、そのクラスタへ割当てられたシンボルへ変換する。それ以外のPD、つまりStep1-4のいずれかで除去されたPDについては、以下の処理の流れでコードブックの階層(ツリー)をたどって行き、シンボルを割当てる。

シンボル割当処理

Step1 処理中のPDと同じTPGのコードブックを選択する。

Step2 処理中のPDのパターン化前の区間映像における先頭フレームの画像特徴ベクトル(コードブック作

成のStep1で用いたものと同じ)と、選択したコードブックの第0レイヤPDクラスタそれぞれの中心画像特徴ベクトルとの類似度を計算し、その類似度が最大でかつ閾値 S'_{th0} 以上であるクラスタを選択する。そのようなクラスタが見つからなかった場合は、処理中のPDに「unknown」シンボルを割当て処理を終了する。ここで、「unknown」シンボルが多く割当てられてイベント分類が低下するのを避けるため、 S'_{th0} は S_{th0} よりも低く設定する。

Step3 処理中のPDにおける「背景矩形構造体」と、Step2で選択されたクラスタの子クラスタ(第1層PDクラスタ)それぞれの第1層中心矩形構造体について、式(1)と同様の類似度 S_1 を計算する。そして、 S_1 が最大でかつ閾値 S'_{th1} 以上であるクラスタを選択する。そのようなクラスタが見つからなかった場合は、処理中のPDに「unknown」シンボルを割当て処理を終了する。ここでも、Step2と同様の理由で S'_{th1} は S_{th1} よりも低く設定する。

Step4 処理中のPDにおける「人物矩形構造体」と、Step3で選択されたクラスタの子クラスタ(第2層PDクラスタ)それぞれの第2層中心矩形構造体について、式(1)と同様の類似度 S_2 を計算する。そして、 S_2 が最大でかつ閾値 S'_{th2} 以上であるクラスタを選択する。そのようなクラスタが見つからなかった場合は、処理中のPDに「unknown」シンボルを割当て処理を終了する。ここでも、Step2と同様の理由で、 S'_{th2} は S_{th2} よりも低く設定する。

Step5 Step4で選択された第2層PDクラスタに割当てられているシンボルを、処理中のPDに割当てる。

以上の処理により、学習用イベントシーンのすべてがシンボル列へ変換される。

2.5 HMMの構築

イベント毎に準備した複数の離散HMMに2.4節で説明した方法で生成したシンボル列を学習させ、各HMMのパラメータを決定する。本論文では、フリーの離散HMMプログラム¹⁶⁾を用いた。図10に、HMMによるイベント種毎の学習処理の流れを示す。まずStep1では、イベント種に属するシンボル列に近い長さのシンボル列のグループに振り分ける。具体的には、長さ $L-1, L, L+1$ のシンボル列を束ねて「長さ L のシンボル列グループ」とする(隣接した長さグループでのシンボル列の重複は許容)。シンボル列の長さはイベントシーンを構成するショット数を反映しており、イベント分類の重要な手がかり(例えば、ホームランシーンは構成ショット数が多い傾向、三振や四球は少ない傾向)と考えられ、分類処理において入力イベントシーンと近い長さのシンボル列を学習したHMMのみを用いることで、可能性の著しく低いイベント種の尤度計算を回避することができる。そしてStep2において、それぞれの長さのシンボル列グループに対して、個別に離散HMM

図 10 HMM の構築
Setup of HMM.

図 11 イベント分類処理の流れ
Outline of event classification.

を準備し、グループの属するシンボル列を学習させてパラメータを決定する。ここで、使用する HMM のモデルは、最も一般的な left-to-right モデルとし、シンボル列の長さと同じ状態数とした。離散 HMM を用いた本処理は、数十秒から数分という非常に低い計算コストで実行可能である。

3. イベント分類処理

イベント種が未知のシーンを、イベント毎に学習した HMM を用いて、イベント分類する。分類処理は、学習処理と同様に、第 2 ショットに重きを置いたシーンのシンボル列化に基づいて行われる。図 11 に分類処理の流れを示す。

Step1 において、イベント種未知シーンのショットチェンジを検出し、学習処理と同様に第 2 ショットを部分第 2 ショットへ分割する。そして、ショットあるいは部分第 2 ショット

を、学習時と同じくパターン化処理する。Step2 で、ショットおよび部分第 2 ショットのそれぞれを、対応する TPG のコードブック（学習時に作成）を用いて 2.4 節の手法でシンボルを割当て、シンボル列を生成する。Step3 では、Step2 で生成されたシンボル列について、各イベント種における HMM 群の中からそのシンボル列と同じ長さのシンボル列グループを学習したもののみを用いて認識処理を行い、それぞれのイベント種である確からしさ（尤度）を算出する。そして Step4 で、尤度が最大であるイベントを、この未知イベント種シーンのイベント種として決定する。

4. 分類実験

本論文では、18 試合分の映像ファイルを三つのセットに分け、1 セットを実験用、2 セットを学習用として相互検証実験を行う。各セットの映像ファイルの詳細データを表 1 に示す。この表が示すように、これらの映像ファイルは、球場、時間帯、現地での放送チャンネルの各条件が同一なもののみで構成されているのではなく、いくつかのバリエーションを持つ。また、コードブック作成時の手動処理に要する時間についてだが、本論文の実験では、クラスあたり 2~10 秒程度の判断時間で、Step1 で形成されるクラス（すべての TPG での合計）が 120~150 個程度、手動により除去されたクラス数は 3~4 個であった。今回は相互検証実験を行ったため、実験の度にコードブックを作り直したが、本来は、一旦良質なコードブックが作成されればそれを更新する必要がほとんど生じないことを考慮すると、問題のない作業時間であるといえる。また、この作業時間は今後、処理に特化して作成された GUI を用いることでさらに短縮可能であると思われる。

実験結果の数値データに触れる前にまず、2.3 節で述べたパターン化処理のメリットが本実験においても生かされていることを図 12 の各例によって示す。図 12(a) は上段と下段は酷似した映像であるが、上段には「人の横切り」が生じている。しかしそれには影響されず、上下段ともにほぼ類似したパターン化データが出力され、同じシンボルの割当てが実現している。また、上段と下段で「横切り」の回数が異なる図 12(b) についても、同様に同じシンボルが割当てられている。また、図 12(c) の上段は、投手が一旦プレートを外すことにより、下段に比べて著しく長い映像となっているが、これについても、上段と下段で同じシンボルが割当てられていることがわかる。

次に、表 2 に本提案手法によるイベント分類結果を示す。Set1 が実験用で Set2, 3 を学習用としたものが実験 1, Set2 が実験用で Set1, 3 を学習用としたものが実験 2, Set3 が実験用で Set1, 2 を学習用としたものが実験 3 である。これらの実験では、すべてのイベントシーンが必ず 7 種類のイベントのいずれかに分類されるため、合計欄において再現率と適合率が同一の値となり、約 70~76% の精度が得られた。さらに、表 3 は、18 試合分の実験用映像ファイ

ルの中で、最も高い分類精度を得た映像ファイル V04 での実験結果のみを示したものである。また、表 1 を見ると、この実験における学習用映像ファイル (Set2, 3) の中に、V04 と同じ球場で行われたものが非常に少ない。それにもかかわらず高い精度が得られている。これは、すべての条件での野球映像を網羅するような大量の学習データを必要とせず、数十試合程度の扱いやすい学習データ量で高精度のイベント分類を可能とする本提案手法の長所を示したものである。従来手法においては、本手法の平均精度よりも高い分類精度を得ている結果⁸⁾が存在するが、一組のデータセットによる結果のみが示され相互検証実験を行っていないなど、検証が不十分である。一方本実験では、より多くのデータで相互検証実験を行い、かつ一部の映像ファイルについては、その従来手法を上回った精度が得られている。それらの点を加味すると、本提案手法は、信頼度の高いロバストなイベント分類を実現しているといえる。

ただし、イベント種毎の精度を評価すると、ツーベースおよび四球については、十分な分類精度が得られていない。これについて考察する。まずツーベースについては、前述の従来手法⁸⁾においても分類精度が低く、学習データが不足していることに起因していると考えられる。ツーベースは、打者の一塁から二塁への進塁について画の推移のバリエーションが比較的広い。しかし、その他の部分 (投球から打球が飛ぶショットまでと打者の二塁到達ショット以降) の画の推移についてはツーベース特有の傾向がしっかりと現れているため、学習データを増加させることによって上記の部分的な画の推移バリエーションを網羅することが可能であり、精度の向上が期待できると考える。

また四球については、打者が一塁へ移動する四球シーンと打者が一塁ベンチへ帰る三振シーンの先頭ショットのカメラワークが酷似しているのが主原因と考える。今回の実験では三振シーンの学習シーン数が四球の約 2 倍であり、かつ上記のような三振シーンが多かったため、四球シーンが三振へと誤分類されるケースが多く生じた。対策としては、上記のようなカメラワークの酷似した先頭ショットについても、その終盤においては撮影対象に違いが生じる (一塁ベース上かベンチか) 傾向を考慮に入れ、PD のデータ構成をショット終了状態がより強く反映されるように改良することが考えられる。

ここでさらに、第 2 ショットの分割による効果を検証するために、比較対象として第 2 ショットを分割せずに同様の実験を行った。表 4 にその結果を示す。表 2 と比較して明らかなように、いずれの実験についても、主に打球が飛ぶイベント (三振、四球以外) について本提案の効果が現れており、約 7~9% の精度向上が実現している。

5. む す び

本論文では、第 2 ショットに重きを置いたシーンのパターン列表現と離散 HMM を用いて、野球映像の各シーンを自

表 1 各セットの映像ファイルの詳細
Detail of videofiles in each set.

(a) Set1					
ファイル	対戦カード	球場ラベル	夜 (N) / 昼 (D)	放送 Ch	
V01	SEA - OAK	1	N	A	
V02	SEA - ANA	3	N	A	
V03	MIN - NYY	4	N	B	
V04	NYY - TRB	6	D	C	
V05	ANA - SEA	2	N	B	
V06	MIN - SEA	2	D	B	
(b) Set2					
ファイル	対戦カード	球場ラベル	夜 (N) / 昼 (D)	放送 Ch	
V07	SEA - OAK	1	N	A	
V08	ANA - SEA	2	N	B	
V09	MIN - SEA	2	N	B	
V10	NYY - TRB	4	N	C	
V11	BAL - NYY	4	D	B	
V12	BAL - NYY	4	D	B	
(c) Set3					
ファイル	対戦カード	球場ラベル	夜 (N) / 昼 (D)	放送 Ch	
V13	MIN - SEA	2	D	B	
V14	NYY - SEA	2	N	B	
V15	NYY - SEA	2	D	B	
V16	BAL - NYY	4	D	B	
V17	NYY - TRB	6	D	C	
V18	NYY - BOS	7	D	A

図 12 ノイズに影響されずにシンボルが割当てられている例
Examples that symbols can be assigned without influences by noises.

動的にイベント分類する新しい手法を提案した。パターン列は、シーン中の各ショットおよび第 2 ショットを分割した部分ショットをパターン化することにより生成される。パターン化処理により、コードブックのシンボル数を抑えることができ、少ない学習データで高い分類性能が得られる。さらに、長い投球間隔、人が横切る、あるいはミススウィッチングなどの「揺らぎ」に対してロバストなイベント分類が可能となる。また、第 2 ショットを分割することにより、イベントを区別するための重要な情報 (打球の方向、野手の捕球および送球など) を含む映像区間の画の推移やカメラワークのパターン表現が詳細となり、分類精度が向上する。そして、実際の野球放送映像を用いた実験結果により、提案手法がシーンを高い精度で数種類のイベント種に自動

表 2 本手法による実験結果
Test results using proposed method.

(a) 実験 1 (実験用: Set1, 学習用: Set2, 3)			
イベント種	実験シーン数	再現率	適合率
ホームラン	13	69.2	64.3
ツーベース	12	33.3	44.4
シングルヒット	48	66.7	64.0
四球	28	46.4	48.1
フライアウト	96	69.8	71.3
内野ゴロ	122	96.7	95.2
三振	61	77.0	75.8
計	380	76.3	76.3
(b) 実験 2 (実験用: Set2, 学習用: Set1, 3)			
イベント種	実験シーン数	再現率	適合率
ホームラン	13	69.2	90.0
ツーベース	3	33.3	12.5
シングルヒット	55	58.2	72.7
四球	38	55.3	46.7
フライアウト	88	71.6	70.0
内野ゴロ	85	94.1	87.0
三振	57	57.9	66.0
計	339	70.5	70.5
(c) 実験 3 (実験用: Set3, 学習用: Set1, 2)			
イベント種	実験シーン数	再現率	適合率
ホームラン	13	84.6	55.0
ツーベース	16	6.2	33.3
シングルヒット	48	68.8	50.8
四球	25	56.0	43.8
フライアウト	111	62.2	73.4
内野ゴロ	81	90.1	83.9
三振	73	75.3	83.3
計	367	69.8	69.8

表 3 映像ファイル毎の全実験の中で最も精度の高かった実験結果

Test result with the highest accuracy in all videofile-by-videofile tests.

映像ファイル V04 (球場ラベル: 6)			
イベント種	実験シーン数	再現率	適合率
ホームラン	1	100.0	100.0
ツーベース	0	-	-
シングルヒット	10	90.0	75.0
四球	6	33.3	40.0
フライアウト	13	76.9	90.9
内野ゴロ	29	100.0	100.0
三振	7	57.1	50.0
計	66	83.3	83.3

的に分類可能であることが示された。本手法は、大量の野球映像データの中から素早く目的のシーンを探し出すことを可能とし、放送現場における映像編集作業を支援する。また、本手法を発展させることにより、野球の試合映像からスコアブックを自動で生成する手法の実現につながるものと期待している。また、本論文では野球を対象としたが、例えば体操競技など、イベントや種目によるカメラワークや画の切り替わりパターンに強い傾向があると推測される他のスポーツへの適用も可能であると考えられる。

今後の課題としてはまず、学習データを増やす必要がある。本論文では、大量の学習データは必要としないが、本実験におけるツーベースのような、画の推移のバリエーションがやや広目であるイベント種に関しても常に安定した分類結果を得るためには、ある程度の学習データの増加は不

表 4 第 2 ショットを分割せずに行った実験結果
Test results using the method without the second shot partitioning.

(a) 実験 1 (実験用: Set1, 学習用: Set2, 3)			
イベント種	実験シーン数	再現率	適合率
ホームラン	13	46.2	60.0
ツーベース	12	8.3	33.3
シングルヒット	48	58.3	54.9
四球	28	71.4	47.6
フライアウト	96	65.6	62.4
内野ゴロ	122	84.4	81.7
三振	61	63.9	83.0
計	380	68.4	68.4
(b) 実験 2 (実験用: Set2, 学習用: Set1, 3)			
イベント種	実験シーン数	再現率	適合率
ホームラン	13	61.5	61.5
ツーベース	3	33.3	14.3
シングルヒット	55	27.3	38.5
四球	38	65.8	59.5
フライアウト	88	52.3	63.0
内野ゴロ	85	88.2	67.0
三振	57	70.2	75.5
計	339	61.9	61.9
(c) 実験 3 (実験用: Set3, 学習用: Set1, 2)			
イベント種	実験シーン数	再現率	適合率
ホームラン	13	53.8	58.3
ツーベース	16	0.0	0.0
シングルヒット	48	50.0	32.4
四球	25	60.0	36.6
フライアウト	111	64.0	73.2
内野ゴロ	81	84.0	81.9
三振	73	64.4	82.5
計	367	63.2	63.2

可欠である。また、ショット終了状態がより強く反映されるように PD のデータ構成を改良することも課題として挙げられる。この改良により、本実験で分類精度の低かった四球だけでなく、その他のイベント種の分類精度向上も期待できる。さらに、パターン化データの精度や、コードブック作成時のパターン化データのクラスタリング精度は、コードブックの質に強く関わる要素であり、それらの向上も重要な課題である。それに関連して、本論文では、コードブックを構成するための各閾値、パラメータ類を予備実験により手動で設定し、それらの値を固定してすべての実験を行ったが、学習データに対する統計処理などの導入により、最適なパラメータ値を自動的に設定する必要がある。

また本論文では、タイムリーシーンやエラーシーンなど、画の推移パターンの傾向が掴みにくいであろうイベントを扱わなかったが、これらのシーンも含めた多種イベントの高精度分類を実現するために、本手法を改良していくとともに、野球のルールイベント生起確率への反映、スーパー領域の画像特徴としての導入や、映像だけでなくクローズドキャプションや音声情報などを利用した他手法との融合も視野に入れていかなくてはならないと考えている。

〔文 献〕

- 1) N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang and A. Zakhov: "Applications of Video-Content Analysis and Retrieval", IEEE Multimedia 7(3), pp.42-55 (2002)

- 2) N. Babaguchi, Y. Kawai and T. Kitahashi : "Event Based Video Indexing by Intermodal Collaboration", Proceedings of First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99), Orlando, FL, USA, pp. 1-9 (1999)
- 3) A. Watanabe, K. Komiya, J. Usuki, K. Suzuki and H. Ikeda : "Effective Designation of Specific Shots on Video Service System Utilizing Mahalanobis Distance", IEEE Transactions on Consumer Electronics, **51**, 1, pp.152-159 (2005)
- 4) A. Ekin, A. M. Tekalp, and R. Mehrotra : "Automatic Soccer Video Analysis and Summarization", IEEE Transactions on Image Processing, **12**, 7, pp.796-807 (2003)
- 5) G. Xu, Y.-F. Ma, H.-J. Zhang, S. Yang : "Motion Based Event Recognition Using HMM", Proc of. the Int. Conf. on Pattern Recognition (ICPR 2002), **2**, pp.831-834 (2002)
- 6) L. Xie, S. F. Chang, A. Divakaran, and H. Sun : "Structure Analysis of Soccer Video with Hidden Markov Models", In Proc. ICASSP 2002, **4**, pp13-17 (2002)
- 7) N. Rea, R. Dahiya and A. Kokaram : "Classification and Representation of Semantic Content in Broadcast Tennis Videos", Proc of the Int. Conf. on Image Processing 2005 (ICIP2005), pp. III-1204-1207 (2005)
- 8) H. B. Nguyen, K. Shinoda and S. Furui : "Robust highlight extraction using multi-stream Hidden Markov Models for baseball video", Proc of. the Int. Conf. on Image Processing 2005 (ICIP 2005), pp. III-173-176 (2005)
- 9) P. Chang, M. Han and Y. Gong : "Extract highlights from baseball game video with hidden markov models", Proc of. the Int. Conf. on Image Processing 2002 (ICIP 2002), **1**, pp.609-612 (2002)
- 10) J. S. Boreczsky and L. A. Rowe : "Comparison of video shot boundary detection techniques", Proc. SPIE, **2664**, pp.170-179 (1996)
- 11) R. Lienhart : "Comparison of automated shot boundary detection algorithms", Proc. SPIE, bf 3656, pp.290-300 (1999)
- 12) 望月, 藤井, 伊藤 : "新しいフラクタル特徴とロバストな構図情報を用いた画像検索", 映情学誌, **57**, 6, pp.719-72 (2003)
- 13) 望月, 蓼沼, 八木 : "静止画と音声によるドラマ番組配信のための要約画像系列生成", 信学技報, IE2005-21 (MVE2005-21), pp.19-24 (2005)
- 14) T. Lin, H.-J. Zhang : "Automatic Video Scene Extraction by Shot Grouping", Proc of. the Int. Conf. on Pattern Recognition (ICPR2000), **4**, pp.39-42 (2000)
- 15) 望月, 蓼沼, 藤井, 伊藤 : "データベース中の代表的なテクスチャと色を用いて作成した問合せ画像による画像検索", 信学論, **J88-D-II**, 8, pp.1729-1739 (2005)
- 16) ICOT Free Software Web Site (on AITRG Project) : <http://www.icot.or.jp/ARCHIVE/Museum/IFS/abst/099-J.html>

もちづき たかひろ

望月 貴裕 1996 年, 東京工業大学大学院物理情報工学専攻修士過程修了。同年 NHK 入局。同放送技術局報道技術センター中継制作部を経て, 1998 年より, 放送技術研究所にて画像認識の研究に従事。現在, 同 (人間・情報) 知能処理グループ研究員。正会員。

ふじい まひと

藤井 真人 1983 年, 名古屋大学工学研究科電気工学修士過程修了。同年 NHK 入局。1987 年, NHK 放送技術研究所。1991 年, CMU 客員研究員。1998 年, エイ・ティ・アール人間情報通信研究所室長。2001 年, NHK 放送技術研究所に復帰。現在, 同 (人間・情報) 知能処理グループ主任研究員。主に画像認識, 視覚のバイオサイバネティクス, 立体画像情報処理に関する研究に従事。正会員。

やぎ のぶゆき

八木 伸行 1980 年, 京都大学大学院電気工学専攻修士課程修了。同年 NHK 入局。甲府放送局, 放送技術研究所, 技術局, 編成局を経て, 現在, 放送技術研究所 (人間・情報) 知能処理グループリーダー。2005 年～東京工業大学特任教授 (兼任)。画像・映像・メディア情報処理, コンピュータアーキテクチャ, コンテンツ制作技術, デジタル放送などの研究開発に従事。工博。正会員。

しの だいち

篠田 浩一 1987 年, 東京大学理学部物理学科卒業。1989 年, 同大学院修士課程修了。同年日本電気 (株) 入社。以後, 音声認識の研究に従事。その間, 1997 年から 1998 年にかけて, 米国ルーセントテクノロジー・ヘル研究所客員研究員。2001 年から東京大学情報理工学系研究科助教授。2003 年から東京工業大学情報理工学研究科助教授。現在に至る。研究分野は, 音声・動画像などに対する統計的パターン認識。博士 (工学)。