論文 / 著書情報
Article / Book Information

論題	時系列データに対するデータ駆動型アプローチに基づく野球放送の頑健なシーン認識
Title	
著者	安藤 亮一, 篠田 浩一, 古井貞熙, 望月 貴裕
Author(s)	Koichi Shinoda, SADAOKI FURUI, Takahiro Mochizuki
出典	画像の認識・理解シンポジウム（MIRU 2007）IS-1-17, Vol. , No. , pp. 570-575
Citation	, Vol. , No. , pp. 570-575
発行日 / Pub. date	2007, 7

画像の認識・理解シンポジウム(MIRU 2007) IS-1-17, pp.570-575 (2007-7)

時系列データに対するデータ駆動型アプローチに基づく野球放送の頑健なシーン認識

安藤 亮一† 篠田 浩一† 古井貞熙† 望月貴裕††

† 東京工業大学 情報理工学研究科 計算工学専攻 〒 152-8552 東京都目黒区大岡山 2-12-1
†† NHK 放送技術研究所 〒 157-8510 東京都世田谷区砧 1-10-11
E-mail: †ando@ks.cs.titech.ac.jp, {shinoda,furui}@cs.titech.ac.jp, ††mochizuki.t-fm@nhk.or.jp

あらまし 本論文では, 野球放送ビデオのための頑健なシーン認識システムを提案する. 本システムは, 連続音声認識などでしばしば用いられる時系列データに対するデータ駆動型アプローチに基づく. 各シーンをマルチストリーム HMM を用いてモデル化し, さらに, 試合ごとの特徴の違いに対応するために教師なしゲーム適応を行う. また, シーン間のコンテキストを表現するために統計的言語モデルの一つである n-gram モデルを用いる. 加えて, シーンごとの時間長の違いに対し, シーン時間長モデルを用いる. 提案するシステムの評価として, 大リーグ野球放送 25 試合分の評価データを用い, 16 種類のシーンに対するシーン認識実験を行い, 平均の F 値において 54.9%の性能を達成した.

キーワード CBVIR, スポーツ ビデオ, インデクシング, 隠れマルコフモデル, n-gram モデル, 適応

A Robust Scene Recognition System for Baseball Broadcast Using Data-Driven Approach

Ryoichi ANDO†, Koichi SHINODA†, Sadaoki FURUI†, and Takahiro MOCHIZUKI††

† Department of Computer Science, Tokyo Institute of Technology 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
†† NHK Science & Technical Research Laboratories 1-10-11 Kinuta, Setagaya-ku Tokyo 157-8510 Japan
E-mail: †ando@ks.cs.titech.ac.jp, {shinoda,furui}@cs.titech.ac.jp, ††mochizuki.t-fm@nhk.or.jp

Abstract We propose a robust scene recognition system for baseball broadcast videos. This system is based on the data-driven approach which has been successful in continuous speech recognition. It uses a multi-stream hidden Markov model to model each scene and an unsupervised adaptation method to achieve robustness against differences in environmental conditions among games. It also employs an n-gram language model to represent the contexts among scenes, and a model for scene length information. The proposed system was evaluated in scene recognition experiments with 16 scene types acquired from video data of 25 baseball games. The system reduced errors in scene recognition by 6.3 % absolute.

Key words CBVIR, sports video, indexing, HMM, n-gram model, adaptation

1. はじめに

近年のコンピューター技術, 特にストレージ技術の進歩により, マルチメディアコンテンツが急増している. マルチメディアコンテンツの検索や要約のためにはインデックスを付与することが必要であるが, 現状ではこの作業は人手によって行われているためコストが大きい. そのためパターン認識技術を用いて自動でインデックスを付与する技術 (Content-Based Video Information Retrieval: CBVIR) の研究が盛んに行われている [1–6]. 本研究では, 野球放送を対象としたシーン認識を行う.

野球放送の最小単位はフレームと呼ばれ, 1 枚の静止画像である. そして, 一つの固定されたカメラで撮影された複数のフレームからショットが構成される. さらに, ショットのシーケンスからシーンが構成される. 本研究ではピッチングショットから次のピッチングショットまでをシーンと定義する. ショットの遷移の情報はシーンの特徴を表し, シーン認識において重要な情報となる (図 1). 野球のシーン認識において, Chang ら [1] は隠れマルコフモデル (Hidden Markov Model: HMM) に基づく手法を提案している. 彼らの手法では, ビデオデータをショット単位に区切り, ショット系列に対して HMM に基づきシーン認識を行う. この手法では, ショットの種類とショット間の遷移に関するドメイン依存の知識を用いてシステムの性能を改善して

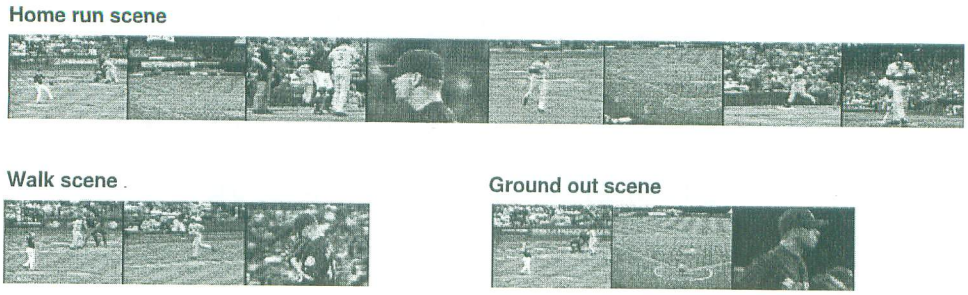

図 1　ホームラン，四球，内野ゴロのショットシーケンスの例

いる．しかし，現実のデータではショット間の遷移の種類が多数存在することや，ショットの分類が困難であるといった問題があり，このシステムは一般的なアプリケーションを考慮した場合，頑健性に欠ける．

Nguyen ら [2] の研究では，野球放送に対するシーン認識を行った．この研究では，各々のシーンに対しマルチストリーム HMM を用いてシーンをモデル化する．これに加えて，シーン認識性能を改善させるために教師なしモデル適応を用いる．この研究では，試合の流れとは直接関係の少ないシーンを除いたダイジェストデータを用いて評価をしている．Ando ら [3,4] の研究では，シーンコンテキストを表現するために n-gram モデルを用い，通常の試合を評価データとしシーン認識を行い，その効果を確認している．本論文では，それらの手法を組み合わせた統合的なシステムを提案する．これに加えて，シーン時間長情報をモデル化し，それを用い認識率の改善を行う．

本論文の構成は以下の通りである．第 2 章で提案システムについて説明する．第 3 章で特徴量を説明し，第 4 章でマルチストリーム HMM とモデル適応の手法を説明する．第 5 章でシーンコンテキストを表現する n-gram モデルについて説明し，第 6 章でシーン時間長情報をモデル化したモデルについて説明する．第 7 章で評価実験の結果を報告し，第 8 章で全体をまとめ，今後の課題を述べる．

2. シーン認識

2.1 フレームワーク

本論文では以下で説明するような時系列データに対するデータ駆動型アプローチに基づく頑健なシーン認識システムを提案する．連続音声認識とのアナロジーから，ショットは音素，シーンは単語にそれぞれ対応させることができる．シーン認識問題は次のように定式化することができる．観測された特徴ベクトルの時系列 $\boldsymbol{O}$ が与えられたとき，シーン列 $\boldsymbol{S}$ の出現確率は次式となる．

$$P(\boldsymbol{S} \mid \boldsymbol{O}) \propto P(\boldsymbol{O} \mid \boldsymbol{S})P(\boldsymbol{S}) \qquad (1)$$

ここで $P(\boldsymbol{O} \mid \boldsymbol{S})$ はシーン列 $\boldsymbol{S}$ から観測ベクトル時系列 $\boldsymbol{O}$ が出現する確率を表し，また，$P(\boldsymbol{S})$ はシーン列 $\boldsymbol{S}$ の出現確率を表す．ここで，$P(\boldsymbol{O} \mid \boldsymbol{S})$ を計算するモデルをビデオモデルと呼び，$P(\boldsymbol{S})$ を計算するモデルを言語モデルと呼ぶ．本論文では，ビデオモデルとしてマルチストリーム HMM を用い [2]，言語モデルとして n-gram モデルを用いる [3,4]．

これに加えて，それぞれのシーンの時間の長さもシーン認識に重要な特徴の一つと考え，そのような情報をモデル化したシーン時間長モデルを本システムに追加する．ビデオモデル，言語モデル，シーン時間長モデルから計算された尤度は次式となる．

$$\begin{aligned} P(\boldsymbol{S} \mid \boldsymbol{O}, \boldsymbol{L}) &\propto P(\boldsymbol{O}, \boldsymbol{L} \mid \boldsymbol{S})P(\boldsymbol{S}) \\ &= P(\boldsymbol{O} \mid \boldsymbol{S})P(\boldsymbol{L} \mid \boldsymbol{S})P(\boldsymbol{S}) \end{aligned} \qquad (2)$$

ここで，$\boldsymbol{L}$ はシーン列 $\boldsymbol{S}$ に対するシーン時間長列であり，$\boldsymbol{S}$ を与えたとき，$\boldsymbol{O}$ と $\boldsymbol{L}$ は互いに独立であると仮定する．シーン時間長モデルは，式 (2) の $P(\boldsymbol{L} \mid \boldsymbol{S})$ を計算する．$P(\boldsymbol{S} \mid \boldsymbol{O}, \boldsymbol{L})$ を最大化するシーン列 $\boldsymbol{S}$ を認識結果とする．

2.2 システム概要

図 2 に提案するシーン認識システムの概略図を示す．シーン認識システムは，学習フェーズと認識フェーズの 2 つに分けることができる．

学習フェーズ: 学習フェーズではビデオモデル，言語モデル，シーン時間長モデルのパラメータを推定する．予め学習データの動画像に対して人手によりラベル付けを行う．これらのラベルデータはシーンの開始時間と終了時間を含む．次に学習データの動画像から静止画像を取り出し，特徴量を計算する．ビデオモデルとして用いるマルチストリーム HMM のパラメータは，この特徴量とラベルデータを用いて推定する．また，ラベルデータを用いて，言語モデルとして用いる n-gram モデルのパラメータを推定する．さらに，ラベルデータのシーンの開始時間と終了時間より，シーンの時間長を求めシーン時間長モデルを作る．

認識フェーズ: まず，学習時と同様にテストデータの動画像から静止画像を取り出し，特徴量を計算する．次に，この特徴量を用いて学習時にモデル化したビデオモデル，言語モデル，シーン時間長モデルを用い，シーン認識を行う．得られた認識結果はシーン列とシーンの開始時間と終了時間を含む．さらに，この結果をもとに第 4.2 節で説明する適応手法を用い，試合ごとにモデルを適応する教師なしゲーム適応を行う．最後に，ゲーム適応を行ったビデオモデルを用いて認識処理をもう 1 度行い，得られた結果を最終的な認識結果とする．

3. 特 徴 量

本論文では，システムの汎用性を維持するため野球に強く依存しない特徴量として，静止画像における主成分特徴量 (PF)，差分画像における主成分特徴量 (DPF)，カメラワーク特徴量 (CF) を用いる．

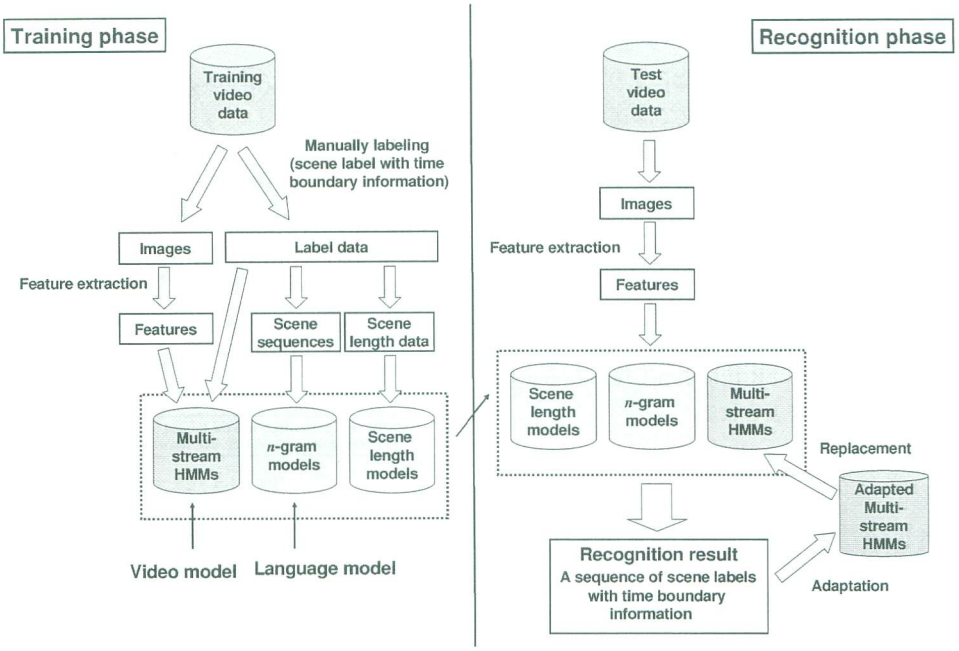

図 2　システムの概略図

3.1　静止画像における主成分特徴量 (PF)

各フレームの画像に対して, 主成分分析 (PCA) を用い次元圧縮することによって, シーンの特徴と関係ない雑音を排除し, 全体的な特徴のみを抽出する [2]. まず, 計算量を削減するために, 720×480 ピクセルの画像を 72×48 ピクセルまで圧縮する. 次に, RGB 画像からグレースケール画像を作り, さらに, それを列ベクトルへ変換する. その列ベクトルに対し PCA を行い, 最初の 10 基底を採用する. 入力フレームを列ベクトルに変換して上記の基底へ射影し, 10 次元の特徴ベクトルを作る.

3.2　差分画像における主成分特徴量 (DPF)

PF は静止画像の特徴を表すが, シーンの特徴としては移動物体の情報も重要である. 移動物体の情報として, 連続フレーム間の差分情報のみに注目し, PF で行っていた処理を差分画像に対して同様に行い, 特徴量とする [3].

3.3　カメラワーク特徴量 (CF)

野球放送ビデオにおいて, 移動物体の情報は重要であるが, カメラの動きもシーン認識に有用な情報となる. それを表現するためにオプティカルフローを用いる [3, 4]. まず, 計算量を削減するために, 720 × 480 ピクセルの画像を 240 × 160 ピクセルのサイズに圧縮する. 次に RGB 画像からグレースケール画像を作る. その画像に 20 ピクセル間隔でサンプル点を配置し, L-K 法 [7] を用いてオプティカルフローを計算する. まず, カメラのパン, チルトを表現するためにオプティカルフローベクトルの x, y 成分それぞれの平均を求める. また, 選手のアップなど物体が大きく映っているとき, オプティカルフローのばらつきは大きくなる傾向があり, そのような特徴を表現するためにオプティカルフローの x, y 成分それぞれの標準偏差を計算する. そしてズームの度合いを求めるためにオプティカルフローが全体的に内向きか外向きかを求め, ズームの度合いとする. オプティカルフローの x, y 成分それぞれの平均と標準偏差, それとズームの度合いの 5 次元の値を特徴量とする.

4.　ビデオモデル

ビデオモデルは式 (1) の $P(\boldsymbol{O} \mid \boldsymbol{S})$ を計算するモデルである. 本論文ではビデオモデルとしてマルチストリーム HMM を用いる.

4.1　マルチストリーム HMM

HMM は時間的に変化するパターンのモデル化に広く用いられ, 多くのシーン認識の研究で用いられている [1–5]. 本論文では, 音声認識の分野でしばしば用いられるマルチストリーム HMM を用い, シーンのモデル化を行う. マルチストリーム HMM は, 特徴ベクトルを複数のストリームに分割し, ストリームごとに重み付けが可能な HMM である. このストリーム重みを最適化することで認識率の向上が期待できる.

また, 従来手法 (e.g., [1]) では, シーン HMM の各状態は 1 つのショットタイプに対応し, 許される状態間遷移 (トポロジー) は, シーンごとに異なっている. このトポロジーは多くの場合, 何らかのヒューリスティックスに基づき, 経験的に決められている. しかし, データの量が増えるにつれて, 各シーンに含まれるショットの遷移は多様になり, 可能なショットタイプの遷移全てを表現するシーン HMM の構築は困難になる. それに対し本論文では, 全てのシーンに対し同じトポロジーをもつ HMM を用いる. これにより, シーン HMM の作成が容易になり, 未知のデータに対しても頑健性をもつことができる. すなわち, 新しいシーンを加えるときやデータが増加したとき, モデル設計をやり直す必要がない.

4.2　ゲーム適応

多数の試合のデータから学習されたモデルは, 異なるゲーム間で共通の特徴を表現している. このモデルは, 任意の試合のデータに対するシーン認識において使用可能である. しかし, 各試合は異なる特徴をもっている. 例えば, このような特徴として, 試合が行われる球場やチームのユニフォームの色などが

考えられる. そこで, 認識対象の試合のデータを用いて, ビデオモデルを試合固有の特徴に適応させることにより認識性能の向上をはかるゲーム適応を行う [2]. 本論文では適応手法として MLLR+MAP 推定法を用い, HMM の正規分布の平均ベクトルに対してのみ適応を行う.

4.2.1 MLLR 法

Maximum Likelihood Linear Regression (MLLR) 法は, HMM のパラメータを適応データに対して最尤となるように線形変換する手法である [8]. HMM の正規分布の平均ベクトルを $\mu = [\mu_1 \cdots \mu_n]^\top$ とすると, MLLR 法によって得られる正規分布の平均ベクトル $\hat{\mu}$ は, 次式で表される.

$$\hat{\mu} = A\mu + b. \tag{3}$$

ここで A は, $n \times n$ の行列, b は n 次元のベクトルである.

4.2.2 MLLR+MAP 推定法

Maximum A Posteriori (MAP) 推定法では HMM のパラメータもある確率密度分布に従って分布する確率変数とみなす [9]. そして, その分布はデータの観測により変化すると考える. すなわち, データを観測した後の方が観測前よりも HMM のパラメータ分布についてより正確な知識が得られていると期待する. 事前分布である HMM の正規分布の平均ベクトルを $\mu = [\mu_1 \cdots \mu_n]^\top$ とし, 観測されたデータを $x_i (i = 1, \cdots, N)$ とすると MLLR 法によって得られる正規分布の平均ベクトル $\hat{\mu}$ は, 次式となる.

$$\hat{\mu} = \frac{\tau\mu + \sum_{i=1}^{N} x_i}{\tau + N} \tag{4}$$

ここで, τ は事前分布に対する重みである. この場合の MAP 推定量は事前分布のパラメータと最尤推定量との重み付け平均である. サンプル数 N が大きくなるに従い, MAP 推定量 $\hat{\mu}$ は最尤推定量に近づく. すなわちデータ量が極めて少ない場合には事前分布のパラメータの重みが大きく, データ量が増えるにつれてデータから得られる最尤推定量の重みが徐々に大きくなる. なお, 本研究では MLLR によってモデルのパラメータの変換を行い, これを事前知識として MAP 推定を行う MLLR+MAP 推定法を用いる. すなわち, 式 (4) の μ として MLLR 法で得られた式 (3) の $\hat{\mu}$ を用いる.

5. 言語モデル

言語モデルは式 (1) の $P(S)$ を計算するモデルであり, シーンのコンテキストを表現する. 自然言語処理の分野で, 単語間のコンテキストを表現する手法として, 文法に基づく手法と統計的言語モデル (Statistical Language Model: SLM) を用いる手法が挙げられる. 文法を用いたシーン認識の研究として, Liang らの手法 [6] がある. 彼らの手法では, 画面上の試合のステータス表示から得られたアウトの数, 得点, ランナー位置の三つの情報より, 野球ルールに基づく文法を用いシーン認識を行っている. しかし, 一般にスポーツ番組は文法に沿わないような例外的なシーン列の出現も多い. また, スポーツのルールは一般的に複雑なため人手によって記述しなければならない. そこで, 本論文では SLM の一つである n-gram モデルを用い, シーンコンテキストを表現する. n-gram モデルではある時点でのシーンの出現確率を直前の $(n-1)$ シーンにのみ依存するとみなす. 与えられたシーン列 $s_1^N = s_1, \cdots, s_N$ に対して, その出現確率 $P(s_1^N)$ は次式となる.

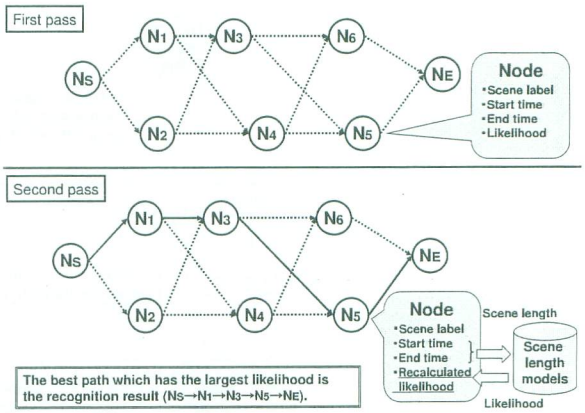

図 3 2 パスアルゴリズムの概略図

$$P(s_1^N) = \prod_{i=1}^{N} P(s_i \mid s_{i-n+1}^{i-1}) \tag{5}$$

なお, $n = 1, 2, 3$ の場合をユニグラム (unigram) , バイグラム (bigram) , トライグラム (trigram) と呼ぶ.

6. シーン時間長モデル

それぞれのシーンの時間長情報はシーンの重要な特徴の一つとして考えられる. 例えば, ホームラン, タイムリー, 長打などの重要なシーンではシーンの時間は比較的長くなるが, 一方, ストライク, ボール, ファールなどのシーンではシーンの時間は短くなる傾向がある. このようなシーン時間長情報はシーン認識に有用であると考えられる. シーン時間長モデルは式 (2) での $P(L \mid S)$を計算し, 以下のように表される.

$$P(L \mid S) = \prod_{i=1}^{N} P(l_i \mid s_i) \tag{6}$$

ここでは, $L = l_1, \cdots, l_N$ はシーン列 $S = s_1, \cdots, s_N$ に対するシーン時間長列とする. また, それぞれの$P(l_i \mid s_i)$ $(i = 1, \cdots, N)$ は互いに独立であると仮定している. 本論文では, 正規分布の確率密度関数を用いて$P(l_i \mid s_i)$を表現する. l をシーン時間長とし, μ_s をシーン s のシーン時間長の平均, σ_s^2 をシーン s のシーン時間長の分散とすると, それぞれのシーン s に対する確率密度関数 $f_s(l)$ は以下のように表すことができる.

$$f_s(l) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{(l-\mu_s)^2}{2\sigma_s^2}} \tag{7}$$

なお, μ_s と σ_s^2 は学習データより求めることができる.

シーンの時間長情報を考慮したシーン認識システムを簡単に実現するために, 本論文では以下で説明する 2 パスアルゴリズムを用いる. 図 3 に 2 パスアルゴリズムの概略図を示す.

第 1 パス: まず, ビデオモデルと言語モデルを用いてシーン認識を行い, 高い尤度を持ったシーン列のシーングラフを出力す

表 2　提案手法を用いたときの結果の F 値 (%). "+"は現在の行の手法を上の行の手法に加えシーン認識を行ったことを表す. (Baseline: シングルストリーム HMM, MHMM: マルチストリーム HMM, TM: トライグラム, SM: シーン時間長モデル, GA: ゲーム適応)

Method	hr	ch	bh	ebh	wk	st	s	b	f	po	so	fo	go	ef	rp	op	avg.
Baseline	54.9	37.4	44.4	12.7	52.3	4.7	32.2	46.9	49.1	53.8	59.8	59.6	65.0	87.6	62.6	55.7	**48.6**
+MHMM	73.8	40.7	41.2	28.0	57.3	8.5	36.0	45.7	50.8	62.2	53.6	60.5	64.9	86.7	57.2	63.3	**51.9**
+TM	74.8	45.8	44.5	22.0	55.6	8.5	33.2	52.4	51.2	58.1	59.7	62.4	67.5	88.6	68.5	66.5	**53.7**
+SM	74.7	45.8	44.0	25.0	55.3	8.5	33.5	52.6	51.6	58.0	59.4	62.5	67.8	88.6	68.1	68.6	**54.0**
+GA	74.9	45.8	45.7	24.6	55.8	8.5	34.2	53.7	50.1	58.6	59.1	63.1	68.2	90.3	69.9	75.0	**54.9**

表 1　シーンラベルとその出現数

シーン	ラベル	出現数	シーン	ラベル	出現数
ボール	b	1701	三振	so	223
リプレイ	rp	1578	ヒット	bh	203
ストライク	s	1062	牽制	po	139
アウトオブプレー	op	937	四球	wk	136
ファール	f	895	タイムリー	ch	59
ゴロ	go	380	長打	ebh	50
フライ	fo	352	ホームラン	hr	38
CG エフェクト	ef	272	盗塁	st	24

る. このグラフではノードはシーンを表し, シーンの開始, 終了時間とその尤度を保持している. また, エッジはシーンからシーンへの遷移を表す.

第 2 パス: 第 1 パスで得られたシーングラフのノードが保持するシーンの開始時間と終了時間よりシーンの時間長を求め, シーン時間長モデル, ビデオモデル, 言語モデルから式 (2) で表されるような尤度を計算する. 最後に, その尤度が最大となるようなパスを探索し, そのパスに含まれるシーン列を認識結果とする.

7. 実　　験

7.1 実験条件

本実験では評価データとして, NHK 放送技術研究所より提供された 25 試合分 (83 時間) の大リーグ野球放送ビデオを用いた. なお, これらのデータ全てに対して人手によってラベル付けを行った. 評価データを 5 試合ずつの 5 つのグループに分割し, 交差検定 (cross validation) を行い, それぞれのグループの結果を平均したものを認識結果とした. 表 1 に認識対象となるシーンラベルを示す. アウトオブプレー (op) とは選手交代のシーンなどの試合とは関係の無いシーンを指し, CG エフェクト (ef) は選手の成績表示などの CG エフェクトを指す. 評価データに含まれるシーンラベルとその出現数を表 1 に示す.

シーン認識システムの評価方法として Precision (適合率) と Recall (再現率) の調和平均である F-measure (F 値) を用いる. これらはフレーム単位で計算する.

学習フェーズにおいて, シーン HMM は表 1 に示すシーンラベルごとに Hidden Markov Model Toolkit (HTK) [10] を用いてモデル化を行った. 全てのシーン HMM は同じトポロジー, 状態数, 混合数とし, トポロジーは left-to-right, 状態数は 50, 混合数は 2 とした. これらの条件は予備実験によって決定した. また, マルチストリーム HMM のストリーム重みも同様に予備実験により最適化を行った. 特徴量は第 3 節で説明した静止画像における主成分特徴量 (PF), 差分画像における主成分特徴量 (DPF), カメラワーク特徴量 (CF) を用いた. 言語モデルに関しては, 予備実験より, ユニグラム, バイグラムと比較するとトライグラムが最も性能が高かったため, 本論文では言語モデルとしてトライグラム用いた. また, トライグラムのモデル化には CMU-SLM-Toolkit [11] を用いて行った. シーン時間長モデルに関しては, シーンごとにシーン時間長の平均と分散を学習データより求めた.

7.2 実験結果

まず, マルチストリーム HMM の有効性を確かめるために, シングルストリーム HMM とマルチストリーム HMM の比較実験を行った. シングルストリーム HMM の結果は PF, DPF, CF の特徴量を一つの特徴量として連結し, 通常の HMM を用いて認識した結果であり, 本論文のベースラインとした. 認識には音声認識エンジン Sphinx4 [12] を用いた. なお, この実験においては, 入力データが任意のシーン列を含むことを許す単純な文法を用いた. 提案手法を用いたときの結果の F 値を表 2 に示す. シングルストリーム HMM を用いたときの F 値の平均は 48.6%となった (表 2 の Baseline) . マルチストリーム HMM を用い, CF, DPF, PF の重みの組み合わせをそれぞれ (0.3, 0.6, 0.1) としたときの F 値の平均は 51.9%となり (表 2 の+MHMM) , F 値の平均は 3.3 ポイント改善された. これらの結果よりマルチストリーム HMM の有効性を確認できた.

次に, シーンコンテキストの有効性を確認するために, 単純な文法を用いたときとトライグラムを用いときの比較実験を行った. 単純な文法を用いたときの F 値の平均は 51.9% (表 2 の+MHMM) であるのに対し, トライグラムを用いた時の F 値の平均は 53.7% (表 2 の+TM) となり, F 値の平均は 1.8 ポイントの改善を得た. この結果より, シーンコンテキストを考慮することはシーン認識に有効であることが確認できた. シーンコンテキストの有効性を説明するために F 値の改善が 11.3 ポイントと大きかったリプレー (rp) を例に考える. リプレー (rp) はホームラン (hr) など重要なシーンの後に出現しやすい. トライグラムを用いることでそのようなシーンコンテキストを表現できたため, 重要なシーンのあとに出現したリプレー (rp) の認識誤りを減らすことができたと考えられる. 一方, 長打 (ebh) や四

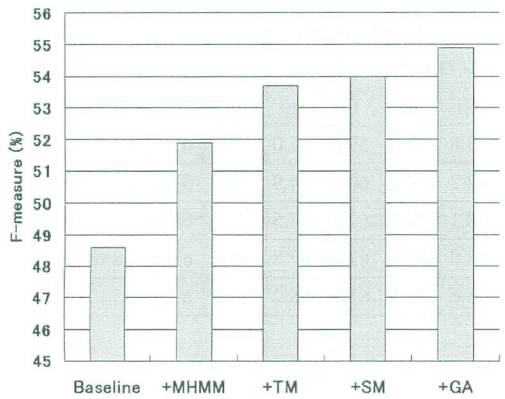

図 4　提案手法を用いたときの平均の F 値 (%) (Baseline: シングルストリーム HMM, +MHMM: マルチストリーム HMM, +TM: トライグラム, +SM: シーン時間長モデル, +GA: ゲーム適応)

球 (wk) などのシーンはシーンコンテキストを考慮することで F 値が低下してしまった. これはトライグラムをモデル化した際の学習データが少なかったため, 正しいシーン列の出現確率が推定できず, 認識率が低下したと考えられる. この問題は学習データを増やすことで改善できると期待される.

次に, シーン時間長モデルの有効性を確認するためにシーン時間長モデルを用いた実験を行った. シーン時間長モデルを用いることで F 値の平均は 54.0% (表 2 の+SM) となり, 0.3 ポイントの改善を得た. この結果より, シーンの時間長情報はシーン認識に有効であることが確認できた. 特に, 長打 (ebh) やアウトオブプレー (op) のような比較的シーンの時間が長いシーンに対しては効果的であった.

最後に, ゲーム適応の有効性を確認するために, ゲーム適応を行った. ゲーム適応を行うことにより F 値の平均は 54.9%となり, 0.9 ポイントの改善を得た. この結果によりゲーム適応の有効性が確認できた.

各手法を組み合わせたときの F 値の平均をまとめたものを図 4 に示す. マルチストリーム HMM を用い, ストリーム重みを最適化することで 3.3 ポイント, それに加え, トライグラムを用い, シーンコンテキストを考慮することで 1.8 ポイント, さらに, シーン時間長モデルを用いることで 0.3 ポイント, 最後にゲーム適応を用いることで 0.9 ポイントの改善を得た. そしてシステム全体として 6.3 ポイントの改善を得ることができた. 最終的に得られた結果において, シーンごとの F 値を比較すると盗塁 (st) の F 値は 8.5%と低い. これは, 表 1 からわかるように, 盗塁 (st) はシーンのサンプル数が少ないため, 学習が正しく行われず, このような結果になったと考えられる.

8. おわりに

本論文では, 野球放送ビデオのための頑健なシーン認識システムを提案した. 本研究ではマルチストリーム HMM を用いシーンをモデル化し, n-gram モデルを用いてシーンコンテキストをモデル化した. さらに, シーン時間長を考慮する手法と試合ごとにモデルを適応させる手法を用い, 認識率の改善を行った. 特徴量として静止画像における主成分分析特徴量 (PF), 差分画像における主成分分析特徴量 (DPF), カメラワーク特徴量 (CF) を用いた. 評価データとして大リーグ野球放送 25 試合分を用いた. システム全体で 6.3 ポイントの改善を得ることができ, 提案手法の有効性を確認できた.

今後は, 音響情報やテキスト情報など, 他のモーダルを用いたマルチモーダル認識への拡張を計画している. 例えば, 音響情報を用いることで盛り上がりの起こるシーンに対しては認識率の改善が期待される. 本研究で比較的認識率の低かったタイムリーや長打などの歓声が起こるシーンに対して有効であると考えられる. 最後に, 本研究の提案手法を他のコンテンツに対して用いてみたい. 野球のシーンのようにはっきりとシーンを定義できるスポーツコンテンツに対しては, 本提案手法が有効であると思われる.

謝　辞

本研究は 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の援助を受けた.

文　献

[1] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden Markov models," *Proc. IEEE International Conference on Image Processing*, vol. 1, pp. 609–612, 2002.

[2] H. B. Nguyen, K. Shinoda, and S. Furui, "Robust scene extraction using multi-stream HMMs for baseball broadcast," *IEICE Transactions on Information and Systems*, vol. E89-D, no. 9, pp. 2553–2561, 2006.

[3] R. Ando, K. Shinoda, S. Furui, and T. Mochizuki, "Robust scene recognition using language models for scene contexts," *Proc, the 8th ACM international workshop on Multimedia information retrieval*, pp. 99–106, 2006.

[4] 安藤 亮一, 篠田 浩一, 古井 貞熙, and 望月 貴裕, "動画像インデクシングのためのシーン時系列の確率的言語モデル," **第 *12* 回画像センシングシンポジウム 予稿集**, pp. 513–518, 2006.

[5] T. Mochizuki, M. Tadenuma, and N. Yagi, "Baseball video indexing using patternization of scenes and hidden Markov model," *Proc. IEEE International Conference on Image Processing*, vol. 3, pp. 1212–1215, 2005.

[6] C.-H. Liang, W.-T. Chu, J.-H. Kuo, J.-L. Wu, and W.-H. Cheng, "Baseball event detection using game-specific feature sets and rules," *Proc. IEEE International Symposium on Circuits and Systems*, vol. 4, pp. 3829–3832, 2005.

[7] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proc. 7th International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.

[8] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[9] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[10] "Hidden Markov model toolkit," `http://htk.eng.cam.ac.uk`.

[11] "Cmu statistical language modeling toolkit," `http://www.speech.cs.cmu.edu/SLM_info.html`.

[12] "Sphinx4," `http://cmusphinx.sourceforge.net/sphinx4`.