論文 / 著書情報
Article / Book Information

論題	野球放送のためのデータ駆動型アプローチを用いた得点シーン検出
著者	石原 一樹, 安藤 亮一, 篠田 浩一, 古井 貞熙, 望月 貴裕
出典	第13回 画像センシングシンポジウム 予稿集, Vol. , No. , pp. 513-518
発行日 / Issue date	2007, 6
Note	第13回 画像センシングシンポジウム講演論文集より転載

野球放送のためのデータ駆動型アプローチを用いた得点シーン検出

A Score Scene Detection for Baseball Broadcast Using Data-Driven Approach

石原一樹 †, 安藤 亮一 †, 篠田 浩一 †, 古井貞熙 †, 望月貴裕 ‡
Kazuki Ishihara† Ryoichi Ando† Koichi Shinoda† Sadaoki Furui† Takahiro Mochizuki‡
† 東京工業大学 情報理工学研究科 計算工学専攻, 東京都目黒区大岡山 2-12-1
‡NHK 放送技術研究所 東京都世田谷区砧 1-10-11
†Department of Computer Science, Graduate School of Information Science and Engineering
Tokyo Institute of Technology
‡NHK Science & Technical Research Laboratories
E-mail: ishihara@ks.cs.titech.ac.jp

Abstract

野球放送のインデクシングにおいて, 画像情報に加え, 音響情報を加えたデータ駆動型アプローチを用いる手法を提案する. 動画像のモデルとして, マルチストリーム HMM を用いた. 25 試合分の野球放送データを用いシーンの認識の評価実験を行った. 画像情報のみを用いた場合と比較して, 16 種類のシーン認識における F 値の平均は 3.2%改善した. また, 得点シーン検出において, Recall が 90.4%となった. これらの結果により提案手法の有効性が確認された.

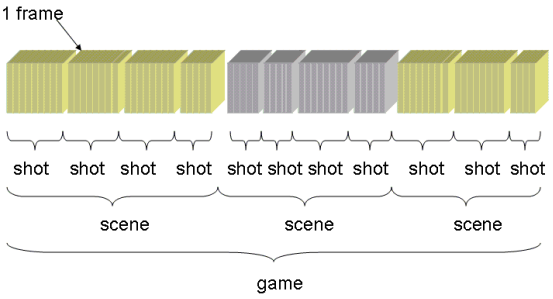


図 **1**　野球放送の構造

1　まえがき

近年, 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を効率よく利用するためには, 検索や要約が必要となる. そのため, 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るものとなり, コストが大きい. そのような背景から, 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 (Contents Based Video Infomation Retrieval;CBVIR) が研究対象として注目されおり, ニュース, スポーツ, 映画など様々なコンテンツに対して研究が行われている [1, 2, 3, 4, 5]. この論文では野球放送のシーンの認識を対象とする.

野球放送のデータ構造の最小単位はフレームで, 一枚の静止画像である. 1 つの固定カメラで撮影された多数フレームからショットが構成される. 更に, ショットのシーケンスからシーン (イベント) が構成される (図 1). ショットの遷移の情報はシーンの特徴を表し, シーンの認識において重要な情報となる (図 2). ゲームの内容を理解するために重要となるシーンはハイライトと呼ばれる.

インデクシングに用いられるマルチメディア情報としては, 映像情報, 音響情報などがあり, それぞれを用いた認識手法が多く提案されている. また, それらの情報を統合して認識を行うことで, より高精度なシステムの構築が期待できる. 我々の先行研究 [1, 2] では映像情報のみでシーンの認識を行ってきた. 映像情報を用いたシーンの認識手法として, 映像情報から PCA による主成分特徴量やフレーム間の差分, カメラワークなどの複数の特徴を抽出し, それらを組み合わせることでシーンの認識を行う手法が提案されている. それ以外にも, シーンを構成するショットをパターン化し, パターン化したデータをもとに認識する手法も提案されている [3]. 映像情報のみを用いる手法では認識が難しいシーンが存在するが, 音響情報を用いることで性能向上を試みた研究もある. 音響情報を用いたシーンの認識手法としては, 例えば, アナウンサーの発話を用いるものがある [7, 8]. しかし, スポーツ中継における発話内容の認識は, 1) 背景音が大きい, 2) 発話の自発性が高い, 3) クロストークが発生しやすい, 4) 映像情報と時間的なずれがある, などの問題があり, 認識性能が低い. また, アナウンサーの興奮度を用いてハイライトシーン検出を行った研究もあるが [9], そこでは, 音響情報のみで映像情報は用いられていない.

本論文では, 従来提案してきたマルチストリーム隠れマルコフモデル (Hidden Markov Model:HMM), 混合ガ

図 2　ホームラン, 四球, 内野ゴロのショットシーケンスの例

ウス分布モデル (Gaussian Mixture Model:GMM) の枠組に音響情報のモードを追加したマルチモーダルシーン認識システムを構築し, シーンの認識性能を評価する. 音響情報として発話内容を直接用いずに, Mel-Frequency Cepstrum Coefficient (MFCC) で表現されるスペクトル特徴を用いる. MFCC を用いることで, 球場の歓声や, アナウンサーの興奮度をモデル化することができる [10]. 特に, 動画像検索する際, ユーザからの要求が高いシーンと音響的に盛り上がるシーンが一致していることが多い. 音響情報を用いることで, それらのシーンの認識率の向上が期待できる.

野球放送などのスポーツ映像では, ユーザの需要が大きいものの１つに得点シーンの検索がある. そこで, 得点シーンの検出の性能向上を目的とする. 得点シーンは, ハイライトとなるシーンのひとつである. テレビ局などのコンテンツホルダーが, ハイライトを作成する際, 従来では人が目視でハイライトとなるシーンを探していたが, コストや時間が非常にかかる. そこで予め, 得点シーンの候補を自動的にある程度絞りこむ手法を検討する. 得点シーンでは, 音響的盛り上がりが高くなるので, 特に音響情報が有効であると考えられる.

本論文の構成は以下の通りである. 2 章ではシーン認識システムに付いて説明する. 3 章で特徴量を説明した後, 4 章でシーンモデルの説明し, 5 章で音響特徴量と画像特徴量の融合について述べる. 6 章でシーン認識システムについて述べ, 7 章で評価実験結果を報告し, 8 章で全体をまとめ, 今後の課題を述べる.

2　シーン認識システム

2.1　シーン認識問題の定式化

本論文では, 入力された映像から自動的にシーンの認識を行い, インデックスを作成することを目的とする. 連続音声認識とのアナロジーから, シーンの認識問題を次のように定式化する. 観測された特徴ベクトルのシーケンス $\boldsymbol{O}$ が与えられたとき, シーン時系列 $\boldsymbol{H}$ の出現確率は次のようになる.

$$P(\boldsymbol{H}|\boldsymbol{O}) \propto P(\boldsymbol{O}|\boldsymbol{H})P(\boldsymbol{H}) \qquad (1)$$

ここで $P(\boldsymbol{O}|\boldsymbol{H})$ はシーン時系列 $\boldsymbol{H}$ から観測ベクトルの時系列 $\boldsymbol{O}$ が出現する確率を表し, また, $P(\boldsymbol{H})$ はシーン時系列 $\boldsymbol{H}$ の出現確率を表す. ここでは, $P(\boldsymbol{O}|\boldsymbol{H})$ を表現するモデルをシーンモデルと呼ぶ, $P(\boldsymbol{H})$ を表現するモデルを言語モデルと呼ぶ. 本論文ではシーンモデル $P(\boldsymbol{O}|\boldsymbol{H})$ のみを対象とし, 言語モデル $P(\boldsymbol{H})$ は全てのシーンで同一の値をもつと仮定する. すなわち, $P(\boldsymbol{O}|\boldsymbol{H})$ を最大にするシーン時系列 $\boldsymbol{H}$ が認識結果となる. 連続音声認識とのアナロジーにおいて, ショットは音素, シーンは単語にそれぞれ対応する.

2.2　シーン認識手法

シーンの認識を行う手法について説明する. 図 3 に示すフレームワークを用い, シーンの学習及び認識を行う. シーン認識システムは, 大きく分けて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, 各シーンをモデル化する. まず, これらのモデルを作るために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, 音響情報を抽出し, 特徴量を計算する. シーンモデルとして用いるマルチストリーム HMM は, これらの特徴量とラベルデータによりモデル化される.

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, 音響情報を抽出し, 特徴量を計算する. 次に, これらの特徴量を用いて学習時にモデル化したシーンモデルを用い, 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.

3　特徴量

3.1　MFCC(Mel-Frequency Cepstrum Coefficient)

MFCC は音声認識に特化した特徴量であるが, 文献 [9, 11] などで, 音声以外の一般の音響分類に用いた場合の有効性が報告されている. そこで, 本論文においても, 観客の歓声, 音楽といった会場音とアナウンサーの音声の興奮度を MFCC により特徴量化して認識を行う.

人間の聴覚は周波数成分に対し, メル尺度と呼ばれる, 対数に近い非線形な特性を示すことが知られている. 音声認識においても, 音響特徴量として, メルスケール変換したケプストラムを用いることによって, メル変換を

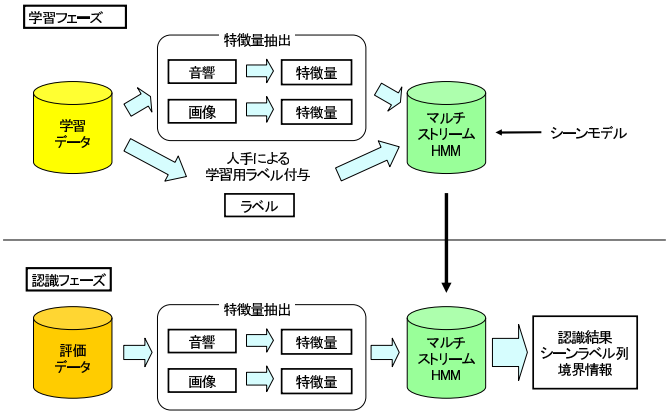


図 3　シーン認識フレームワーク

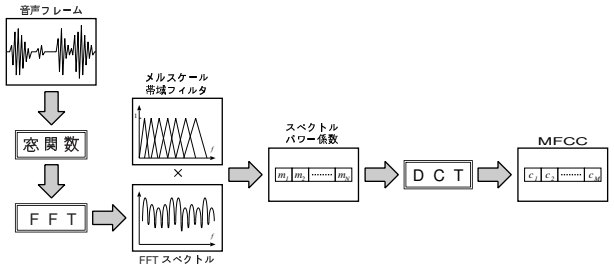


図 4　MFCC の生成過程

行わなかった場合に比べ, 認識性能が向上することが確かめられている.

音響情報からケプストラムを求める過程を図 4 に示す. まず音声フレームに対して, ハミング窓によりフレーム端の処理を行った後, FFT を施して周波数成分に変換する. 音声信号処理として, 一定時間の音声信号を切り出してフレームとし, ハミング窓によってフレーム端の処理を行った後, 得られた周波数成分を式 (2) で示される変換式に基づいてメルスケールの帯域フィルタにかけ, 各帯域内でのスペクトルパワー係数 $m(l)$ を計算する.

メル周波数は,

$$f' = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

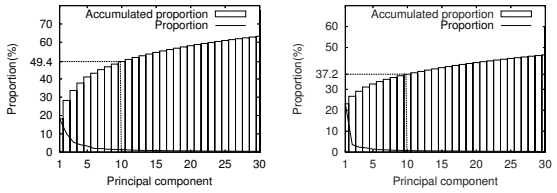
により計算される. f はメル変換前の, f' は変換後の周波数である. f の単位は [Hz] である.

スペクトルパワー係数 $m(l)$ は以下のように決められる.

$$m(l) = \sum_{k=lo}^{hi} W(k,l)|X(k)| \qquad (l = 1, \ldots, L) \tag{3}$$

$$W(k,l) = \begin{cases} \frac{k - k_{lo}(l)}{k_c(l) - k_{lo}(l)} & \{k_{lo} \le k \le k_c(l)\} \\ \frac{k_{hi}(l) - k}{k_{hi}(l) - k_c(l)} & \{k_c \le k \le k_{hi}(l)\} \end{cases} \tag{4}$$

ただし, L はフィルタバンクのチャンネル数, $k_{lo}(l)$, $k_c(l)$, $k_{hi}(l)$ は, それぞれ l 番目のフィルタの下限, 中心, 上限



(a) PF における寄与率　　(b) DPF における寄与率

図 5　寄与率

のスペクトルチャンネル番号であり, 隣り合うフィルタ間で

$$k_c(l) = k_{hi}(l-1) = k_{lo}(l+1) \tag{5}$$

なる関係がある. さらに $k_c(l)$ はメル周波数軸上で等間隔で配置される.

最後に, フィルタバンク分析によって得られた L 個の帯域におけるパワーを離散コサイン変換することで MFCC を求めることができる.

$$c_{\mathrm{mfcc}}(i) = \sqrt{\frac{2}{N}} \sum_{l=1}^{L} \log m(l) \cos\left\{\frac{\pi i}{N}(j - 0.5)\right\} \tag{6}$$

3.2　静止画像における主成分特徴量 (PF)

静止画像における主分分析特徴量 (PF) は各フレームの画像に対し PCA を用い次元圧縮を行うことによって, シーンとは関係ない雑音を排除し, 全体的な特徴を抽出できる. 以下の手順により PF を計算する.

1. 計算量を減少させるために, 720×480 ピクセルの画像を 72×48 ピクセルのサイズに圧縮する.
2. RGB 画像からグレースケール画像を作り, それを列ベクトルに変換する.
3. ビデオデータからランダムに 5000 フレーム分の画像を選択する.
4. 選択した 5000 フレームの画像を用いて PCA により固有ベクトルを得る. 求められた固有値の寄与率を図 5 に示す. 本論文では, 予備実験の結果から第 10 主成分 (累積寄与率 49.4%) までの固有ベクトルを用いる.
5. 得られた固有ベクトルを用い入力画像より 10 次元の特徴ベクトルを計算する.

3.3　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ことができるが, シーンの特徴としては移動物体の情報も重要である. 差分画像における主分分析特徴量 (DPF) は, 移動物体の情報として連続フレームの差分情報に注目する. DPF は PF と同様に PCA により差分画像の次元圧縮を行い特徴量として用いる. 以下の手順により DPF を計算する.

1. 計算量を減少させるために, 720×480 ピクセルの画像を 72×48 ピクセルのサイズに圧縮する.

2. 2 つの連続するフレーム画像の差分をとり, RGB 画像からグレースケール画像を作り, それを列ベクトルに変換する.
3. ビデオデータからランダムに 5000 フレーム分の差分画像を選択する.
4. 選択した 5000 フレームの差分画像を用いて PCA により固有ベクトルを得る. 求められた固有値の寄与率を図 5 に示す. 本論文では, 予備実験の結果から第 10 主成分 (累積寄与率 37.2%) までの固有ベクトルを用いる.
5. 得られた固有ベクトルを用い, 入力された差分画像より 10 次元の特徴ベクトルを計算する.

3.4 カメラワーク特徴量 (CF)

野球放送の場合, 各ショットタイプは固有のカメラワークを有することが多い. 例えば, ピッチングショットの場合, カメラはほとんど動かない. 四球など, 打者がホームベースからファーストベースへ移動するショットでは, カメラは打者を追って水平方向の動きをする. また, 打者が打ってフライが上がるショットでは, カメラは垂直方向の動きをする. このように, カメラワークはショット及びシーンを表す特徴として考えられ, 多くの研究に使われている (e.g. [12]) . 野球の試合を撮影するカメラはその位置が固定されているので, 野球放送におけるカメラワークはパン (水平方向への移動) , チルト (垂直方向への移動) , ズームに制限されている. 本論文ではこれらのカメラワークを表すためにオプティカルフローを用い, 特徴量を計算する. まず, 現在のフレームの画像に N 個のサンプル点を配置し, i 番目の点を, $\vec{p}_i = [p_{x_i}\ p_{y_i}]^\top$ とする. そして, それらの点に対応する次のフレームの画像での点 $\vec{p'}_i = [p'_{x_i}\ p'_{y_i}]^\top$ を階層的 Lucas - Kanade 法を用いて求める. i 番目の点のオプティカルフローベクトルを $\vec{\nu}_i = [\nu_{x_i}\ \nu_{y_i}]^\top$ とすると次式で表される.

$$\vec{\nu}_i = \vec{p'}_i - \vec{p}_i \tag{7}$$

ここで, カメラのパンやチルトを表すために, 全ての点のオプティカルフローの平均 $\vec{\mu} = [\mu_x\ \mu_y]^\top$ を計算する.

$$\vec{\mu} = \frac{1}{N}\sum_{i=1}^{N}\vec{\nu}_i \tag{8}$$

選手のアップなど物体が大きく映っているときに, オプティカルフローのばらつきは大きくなる傾向がある. そこで, 選手のアップシーンなどを表現するためオプティカルフローの標準偏差 $\vec{\sigma} = [\sigma_x\ \sigma_y]^\top$ を計算する.

$$\vec{\sigma} = \frac{1}{N}\sum_{i=1}^{N}(\vec{\nu}_i - \vec{\mu})^2 \tag{9}$$

次に, カメラのズームを表すためにオプティカルフローが全体的に内向きか, 外向きかどうかを求め, それをズームの度合いとする. ズームの度合い z は次式で表される.

$$z = \frac{1}{N}\sum_{i=1}^{N}\{(\nu_{x_i} - \mu_x)(p_{x_i} - c_x) + (\nu_{y_i} - \mu_y)(p_{y_i} - c_y)\} \tag{10}$$

ここで, c_x, c_y を現在のフレームの画像での中心位置の座標とする.

以上より求めた $(\mu_x, \mu_y, \sigma_x, \sigma_y, z)$ を特徴ベクトルとする. ここで CF の計算手順を以下に示す.

1. 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.
2. RGB 画像から輝度成分を取り出しグレースケール画像にする.
3. 20 ピクセル間隔でサンプル点を配置し, 次のフレームと現在のフレームよりオプティカルフローを計算する.
4. 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と標準偏差, それとズームの度合いを計算し, 特徴ベクトルとする.

4 シーンモデル

シーンモデルにはマルチストリーム HMM, Gaussian Mixture Model(GMM) を用いる. 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る. マルチストリーム HMM は, 複数のストリームを重み付けして利用できるように拡張した HMM である (図 6). HMM のパラメータは, 学習データの画像, 音響から特徴ベクトル系列を抽出し, その特徴ベクトル系列を用い学習を行うことにより推定される. 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. このようなデータ駆動型のアプローチは, 音声認識の分野でよく用いられており, 高い認識性能と頑健性を得ている. GMM は, 音声の定常的な特徴をモデル化するため用いられる. 実際, GMM を用いた話者認識や雑音識別の研究が広く行われている. GMM は, HMM で状態数 1 とした場合と同じである. 本論文では, マルチストリーム GMM を用い, 得点シーン検出を行う.

5 画像特徴量と音響特徴量の融合

マルチモーダルシーン認識には, 音響情報と画像情報をどの時点で融合するかという重要な問題がある. これについて多くの手法が考え出されているが, feature fusion 法による手法と, マルチストリーム HMM による融合と 2 種類の手法がある.

5.1 Feature fusion 法

Feature fusion 法は, 音響情報から抽出した音響特徴量と映像情報から取り出した画像特徴量を, パラメータのレベルで融合して音響–画像特徴量とし, 通常の HMM などにより学習・認識を行う方法である. この方法では, 音響と画像をまとめて単一のベクトルとして扱うため, 従来の音声認識で用いられてきたモデルの作成法や学

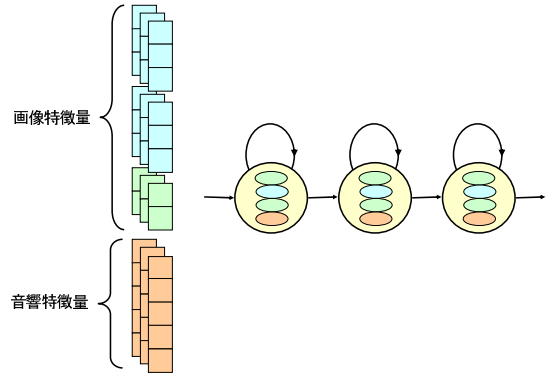


図 6　マルチストリーム HMM

習・認識アルゴリズム, さらに適応化といった手法を利用することが容易である. しかし, 音響と画像という異なった情報を単一のベクトルとして扱っているため, 各情報の信頼性の違いに対して柔軟に対応できないという問題がある.

5.2　マルチストリーム HMM による融合

マルチストリーム HMM は, 音響と画像情報を出力尤度計算に用いるモデルのレベルで融合する手法である. この手法は, 各ストリームの重みを変えることで, 雑音環境などに応じた頑健な認識を行うことができる. また, 映像と音響で, より重要なストリームの重みを相対的にあげることにより認識率の向上が期待できる. また, 音響と映像とで各情報の信頼性が異なる場合でも, ストリーム重みを変化させることで, より頑健で認識率の高いシーンの認識を行うことができる.

本論文では, マルチストリーム HMM を用いたシーンの認識を行う.

6　評価実験

6.1　実験条件

提案手法の評価データとして, NHK 放送技術研究所より提供された 25 試合分 (75 時間) の大リーグ野球放送データを用いた. 評価データを 5 試合づつの 5 つのグループに分割し, 交差検定 (Cross Validation) を行った結果を平均したものを認識結果とした.

実験は, シーンを細かく 16 種類に分類し, 各シーンの結果の平均を評価するもの (シーン認識) と, シーンを得点シーン, 非得点シーンに分類し, 得点シーンの結果を評価するもの (得点シーン検出) の 2 つ行った. 表 1 にシーン認識で認識するシーンラベルを示す. 5 つに分割したグループに含まれるシーンラベルの内訳を表 2 に示す.

得点シーン検出では先行研究 [10] で, 得点シーンは, ホームチームの得点シーンと, アウェイチームの得点シーンの 2 つに分類すると有効であることが分かっている. 本論文でも, 得点シーンを「ホームチーム得点」と「アウェイチーム得点」に分類することにし, 得点シーン検出でのシーンラベルを「ホームチーム得点 (hsc)」, 「アウェイチーム得点 (asc)」, 「非得点 (nsc)」, 「リプレイ (rp)」, 「アウトオブプレイ (op)」, 「CG エフェクト (ef)」の 6 種類に分類した. 同様に, シーンラベルの内訳を表 3 に示す.

表 1　シーンラベル

シーンラベル	内容
base hit(bh)	シングルヒット
extra-base hit(ebh)	長打
clutch hit(ch)	タイムリーヒット
home run(hr)	ホームラン
ground out(go)	内野ゴロ
fly out(fo)	フライアウト
strike out(so)	三振
strike(s)	ストライク
ball(b)	ボール
fall(f)	ファール
walk(wk)	四球 (デッドボールを含む)
pickoff(po)	牽制
steal(st)	盗塁 (盗塁失敗を含む)
out of play(op)	アウトオブプレー (試合とは関係のないシーン)
replay(rp)	リプレイ
effect(ef)	エフェクト (選手の成績表示などの CG)

6.2　評価方法

認識の評価のために, Recall, F 値を用いた. F 値は Precision(P) と Recall(R) の調和平均である. 本論文では各シーンに対する, Precision(P) と Recall(R) は以下のように計算する.

$$P = \frac{C}{S}, \quad R = \frac{C}{T} \tag{11}$$

ここで, C は認識結果に含まれる正解のシーン, S は認識結果におけるそのシーンの全体, T はデータに含まれるそのシーンの全体を示す.

6.3　シーンモデル

シーンモデルとして, 各々のシーンに対し 1 つの HMM を準備した. ここで用いた HMM のトポロジーは自状態遷移と次の状態への遷移のみを許す単純な構造 (left-to-right HMM) を採用した. HMM の状態数は予備実験の結果より, シーン認識では 40 とした. 得点シーン検出では状態数を 1 (GMM) とした. 学習及び認識に用いた特徴量は 3 章で説明した PF, DPF, CF, MFCC の組合せである. HMM の学習には HTK [13] を用い, 認識には Sphinx [14] を用いた.

以上のようなシーンモデルを用いて, 認識を行った.

表 2　５つに分割したグループに含まれるシーンラベルとその出現数

ラベル	グループ ID					
	1	2	3	4	5	Total
b	271	371	353	308	398	1701
rp	235	351	270	356	366	1578
s	192	192	221	194	263	1062
op	185	197	181	199	175	937
f	160	166	199	200	170	895
go	78	67	73	86	76	380
fo	70	81	66	60	75	352
ef	58	72	52	52	38	272
so	41	36	48	49	49	223
bh	38	49	38	41	37	203
po	24	19	25	35	36	139
wk	25	38	23	23	27	136
ch	6	17	10	9	17	59
ebh	10	9	12	11	8	50
hr	9	9	4	10	6	38
st	4	4	6	9	1	24

6.4　シーン認識

本実験では, イニング境界を既知とし, イニングごとに認識を行う. 評価はフレームベースで行い, 各シーンの F 値の平均を評価とする. 表 4 は, マルチストリーム HMM を用い, 画像と音響のストリーム重みを変えて認識を行ったときの結果である.

PF, DPF, CF の画像特徴量のストリーム重みは, 予備実験の結果よりそれぞれ 0.45, 0.45, 0.10 と設定した. この比を固定したまま, 画像と音響の重みの比を変えて認識を行った. 画像, 音響のストリーム重みを変化させることによって, 認識率も変化する. 画像のみを用いた時の F 値の平均は 49.1%. 画像と音響のストリーム重みの組み合わせを 0.8, 0.2(PF, PDF, CF, MFCC をそれぞれ 0.36, 0.36, 0.08, 0.2) としたときの F 値の平均は 52.3%となり, 3.2 ポイント改善された.

次に, それぞれのシーンについて考察する. マルチストリーム HMM を用い, 画像と音響を最適な重みで融合した場合, 「ホームラン (hr)」が 7.6 ポイント, 「タイムリーヒット (ch)」が 7.5 ポイント, 「長打 (ebh)」が 6.8 ポイントと, F 値が大きく改善した. これらのシーンは, 他のシーンと比べて, 観客やアナウンサーの興奮度が高まるシーンである. 音響情報では, これらの特徴をモデル化することができるので, 音響情報を用いたことによって認識率が向上した.

6.5　得点シーン検出

本実験では, シーン境界を既知とし, シーンごとに認識を行う手法と, シーン認識で得られたシーン境界を用いて認識を行う手法の 2 つ行った. 評価はシーンベースで行い, 得点シーンの Recall を評価する.

表 3　得点シーン検出におけるシーンラベルとその出現数

ラベル	グループ ID					
	1	2	3	4	5	Total
hsc	12	28	8	14	9	71
asc	9	25	18	18	16	86
nsc	1446	1748	1662	1612	1776	8244
rp	288	442	360	433	441	1964
op	237	222	241	119	22	1186
ef	72	84	63	61	49	329

表 5 に, シーン境界既知で混合数を変えたときの得点シーン検出結果を示す. 音響のみ, 画像のみ, 音響＋画像の全てで, 混合数 16 の時に再現率が一番高くなっている. また, 音響のみと画像のみを同じ混合数で比較すると, 再現率は音響のみの方が高く, 適合率は画像のみのほうが高いことがわかる.

次に, 音響と画像を融合させた時の結果をみると, F 値では混合数 128 の時に 45.6%となっているが, 再現率が 65.0%と低い. しかし, 混合数 16 の時には, 再現率は 90.4%となっており, 非常に高い精度で検出できた. また, この時の適合率は 15.2%となった. 得点シーンは全シーン中約 1.5%程度しかない. 得点シーン検出を行うことによって, 9.6%の取りこぼしはあるものの, 探索効率が 10 倍近く向上したことを意味している.

表 7 に画像＋音響の混合数 16 のときの confusion matrix を示す.「ホームチーム得点 (hsc)」と「アウェイチーム (asc)」間の認識誤りを無視し,「得点シーン (hsc+asc)」にまとめて評価する.

表 7 をみると,「得点シーン」であるのに他のシーンと認識されたのが, 157 シーン中 15 シーンあり, そのうち 4 シーンが「リプレイ (rp)」となっている. また,「リプレイ (rp)」が「得点シーン (hsc+asc)」と認識されたものが 81 シーンもある. つまり「得点シーン (hsc+asc)」と「リプレイ (rp)」で認識過りが多い.「リプレイ」は, 前後に CG エフェクトが入ったり, 効果音が入り, その間に前のシーンのリプレイが流れるので時間的変化の特徴が顕著である. GMM は, 定常的な特徴をモデル化し, 時間的変化の特徴はモデル化できないため, 認識誤りが多くなったと考えられる.

次に, ホーム得点とアウェイ得点それぞれについて考察する. 表 6 に結果を示す. 画像と音響を組み合わせた場合, Recall は両方とも約 12%改善, Precision は「ホーム得点 (hsc)」が 8.2%,「アウェイ得点 (hsc)」が 0.4%改善している. このことから, 音響を組み合わせた場合, ホーム得点の方がより改善していることが分かる. これは, ほとんどの場合, 球場の大多数をホームチームファ

表 4　マルチストリーム HMM を用いたシーン認識時の F 値 (%)

	hr	ch	bh	ebh	wk	st	s	b	f	po	so	fo	go	ef	rp	op	平均
画像のみ	62.8	35.8	44.7	24.7	52.2	0.0	34.3	47.5	47.5	51.6	58.3	60.2	63.6	87.7	55.5	58.6	**49.1**
画像+音響 (1.0, 0.0)	68.4	34.3	41.9	27.4	54.3	16.8	35.5	44.6	47.1	50.9	57.2	57.2	62.2	86.9	57.2	56.4	49.9
画像+音響 (0.0, 1.0)	37.8	12.0	11.9	6.9	7.9	0.0	4.2	5.0	12.3	7.5	7.5	15.0	8.2	11.1	3.8	21.6	10.8
画像+音響 (0.5, 0.5)	64.3	40.5	43.4	31.0	51.7	8.5	38.9	47.0	48.8	50.8	49.5	58.9	62.7	84.8	55.9	59.8	49.8
画像+音響 (0.8, 0.2)	**70.4**	43.3	48.9	31.5	54.3	4.7	37.6	47.1	49.5	52.5	59.1	60.8	64.5	89.5	60.9	62.3	**52.3**

表 5　GMM を用いて混合数を変えたときの結果 (%)

混合数		1	2	4	8	16	32	64	128	256	512
音響のみ	Recall	81.5	84.7	86	86	86.6	82.8	77.7	63.7	37.6	17.2
	Pecision	4.8	5.4	7.1	7.2	7.9	8.2	9.1	12.2	14.6	26.2
	F 値	9.1	10.1	13.2	13.3	14.4	15	16.2	20.5	21	20.8
画像のみ	Recall	63.1	82.8	80.9	78.3	79	79	72	56.1	40.8	15.3
	Pecision	5.7	4	5.7	9.2	11.3	13.6	18.2	23.9	31.8	40.7
	F 値	10.5	7.6	10.7	16.5	19.7	23.2	29	33.5	35.8	22.2
音響＋画像	Recall	71.3	84.7	87.9	89.8	**90.4**	88.5	80.3	65	33.8	8.9
(重み 0.8:0.2)	Pecision	7.5	5.3	7.5	12.7	15.2	19.1	23.9	35.1	49.5	77.8
	F 値	13.6	10	13.9	22.3	26.1	31.4	36.8	**45.5**	40.2	16

表 6　ホーム得点とアウェイ得点の結果 (ホーム得点:Hsc, アウェイ得点:Asc)

	画像のみ		画像+音響	
	Hsc	Asc	Hsc	Asc
Recall	78.9	79.1	90.1	90.7
Precision	9.8	13.0	18.0	13.4
F 値	17.43	22.33	30.01	23.35

表 7　画像＋音響の cofusion matrix (単位はシーン数であり, 行は認識ラベル, 列は正解ラベルを表す)

	op	rp	ef	hsc	asc	nsc	Total
op	118	12	0	0	1	27	158
rp	108	1246	5	36	45	117	1557
ef	7	8	96	0	1	1	113
hsc	1	3	0	**56**	**8**	3	71
asc	0	1	0	**9**	**69**	7	86
nsc	440	672	13	260	448	6411	8244
total	678	1948	114	361	575	6742	10418

ンが占めており, ホームチームの得点の時とアウェイチームの得点の時とで歓声の大きさが違い, ホームチーム得点の時の方が歓声が大きいことから,「ホーム得点 (hsc)」の改善が大きかったと考えられる.

次に, シーン認識で得られたシーン境界を用いて認識をおこなった場合の認識結果を表 8 に示す. 混合数 4 の時に, 再現率が 87.2%となり, シーン境界既知と比べた場合 3.2 ポイント低下したが, ほとんど差がないことがわかった. これは, シーン認識で得られたシーン境界が, 実際のシーン境界とほとんど同じであったためと考えられる. 実際に得点シーン検出を行うアプリケーションなどを考えた場合, 認識時にはシーン境界はわからないため, この手法が有効であることがわかった.

7　まとめと今後の課題

本論文では, 野球中継番組を対象とした, データ駆動型アプローチを用いたシーンの認識を行う手法を提案した. 画像特徴量 PF, DPF, CF と音響特徴量 MFCC をマルチストリーム HMM を用いて融合し, 各シーンをモデル化し, 認識を行った. 特にユーザからの要求が大きいと思われる得点シーンの検出を行った.

画像情報を用いた場合と比較して, 画像と音響を組み合わせた場合, 16 種類のシーン認識における F 値の平均は　改善した. また, 得点シーン検出において, Recall が 90.4%になった. これらの結果により, 提案手法の有効性を確認した.

提案手法では, 画像情報のみでは表せない, 歓声やアナウンサーの興奮度という, 実際に野球の試合を見た人の感情を特徴量として使うことができるため, 見ている人にとって感情が高まりやすいシーンの認識精度が高くなっている. また, そのように見ている人の感情が高まりやすいシーンは, 多くの場合, 映像検索においての要求が多いシーンと合致すると思われる. そのため, 提案手法はシーンの認識のためのインデクシング手法として効果的であるものと思われる.

本論文では野球放送についてインデクシングを行っ

表 **8**　シーン認識で得られたシーン境界を用いた場合 **(%)**

混合数		1	2	4	8	16	32
音響＋画像	R	73.1	80.8	**87.2**	84.6	82.1	79.5
(重み 0.8:0.2)	P	10.2	7.8	10.5	15.9	18.7	22.2
	F	18.0	14.2	18.7	26.7	30.5	34.7

たが, 本論文の提案手法を他のスポーツでも用いてみたい. 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. 例えば, サッカーは野球と違い, 連続したプレーの連続なのでシーンの定義が難しいが, 得点シーン検出に限れば, ゴールシーン前後を得点シーンとし, その他のシーンを非得点シーンと定義すれば本提案手法を用いることができると思われる.

今回は画像特徴量として, PF, DPF, CF, 音響特徴量として, MFCC を用いたが, スポーツ中継の音・映像情報には, ほかにもシーンの認識に有効な特徴量が含まれているものと思われる. そこで, 新たな特徴量抽出手法を検討する必要がある. 本論文では, 画像特徴量と音響特徴量の融合に, マルチストリーム HMM を用いて融合した. 評価データを用いてストリーム重みを最適化を行ったが, この重みは未知のデータに対して必ずしも最適ではない. そのため, 自動的に最適な重みを決定する手法の検討が必要である. また, 他の融合手法についても検討する必要がある.

本論文では, シーンコンテキストを表現する言語モデルを考慮しなかった. つまり, 全てのシーンの出現確率は同一のものとした. 今後, 言語モデルを用いることによって, 認識率の改善が期待できる.

得点シーン検出では, シーン認識と同じフレームワークを用いたが, サポートベクターマシーン (Support Vector Machine:SVM) を用いた手法など, 他の手法も検討していきたい.

謝辞

この研究は 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.

参考文献

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] S. Eickeler and S. Müller, “Content-based video indexing of TV broadcast news using hidden Markov models,” *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, pp. 2997–3000, 1999.

[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] Y. Gong, M. Han, W. Hua, and W. Xu, “Maximum entropy model-based baseball highlight detection and classification,” *International Journal of Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 181–199, 2004.

[8] 佐古 淳，有木 康雄, “知識を用いた音声認識による野球実況中継の構造化,” **第 *6* 回音声言語シンポジウム**, vol. SP2004-136, pp. 85–90, 2004-12.

[9] Yong Rui, Anoop Gupta, and Alex Acero, “Automatically extracting highlights for tv baseball programs,” *ACN Nultimedia*, pp. 105–115, 2000.

[10] T. Miyazaki, H. Nakagawa, R. Nakagawa, K. Iwano, K. Shinoda, and S. Furui, “野球中継番組を対象とした音響情報を用いたシーン認識,” **日本音響学会**, vol. 1, no. 11-9, pp. 19–20, 2006.

[11] Perfecto Herrera, Alexandre Yeterian, Fabien Gouyon, “Automatic classification of drum sound: a comparison of feature selection methods and classification techniques,” *Proc. of the int. conf. on Music and artificial intelligence*, pp. 69–80, 2002.

[12] S. Takagi, S. Hattori, K. Yokoyama, A. Kodate, and H. Tominaga, “Sports video categorizing method using camera motion parameters,” *Proc. IEEE International Conference on Multimedia and Expo*, vol. 2, pp. 461–464, 2003.

[13] *Hidden Markov Model Toolkit*, `http://htk.eng.cam.ac.uk`.

[14] *Sphinx4*, `http://cmusphinx.sourceforge.net/sphinx4`.