LI-19　放送型スポーツ映像の構造解析に基づく意味内容情報の獲得

新田 直子　　馬場口登
Naoko NITTA　　Noboru BABAGUCHI
大阪大学産業科学研究所
Institute of Scientific and Industrial Research, Osaka University

1　はじめに

近年，テレビやビデオ映像などの連続メディアの有効利用への要望が高まり，映像を対象とした内容記述技術が求められている．映像の内容記述形式としてはMPEG7が標準化されているが，各映像に必要な記述内容は，応用アプリケーションに依存する．よって，実際の記述の際には，アプリケーションに即した記述内容を設定し，さらに設定された記述内容に相当する情報の自動的な取得が必要となる．

ここで，映像の利用方法を検索，視聴，編集などに特定すると，記述内容としては，色特徴やオブジェクトなどの物理的内容ではなく，意味内容に基づくものであることが望ましい．これに対して，映像ジャンルをスポーツ映像に特定し，スポーツ映像の構造を考慮した実際のプレイ部分とそれ以外の部分への映像分割法 [1, 2]，スポーツの試合中に起こる得点部分など重要なイベントに対するインデキシング法 [3] などが提案されている．スポーツ映像においては各プレイが重要な意味を持つと考えられるが，映像分割法では分割後の各プレイ部分の意味内容情報の抽出は行っておらず，インデキシング法では，すべてのプレイ部分に対する記述は不可能である．また，プレイと選手に関するアノテーション法 [4] では，スポーツ映像中のアナウンサの発話内容からのキーフレーズの検索によりすべてのプレイ部分の意味内容情報の抽出を試みているが，この手法で用いるすべてのプレイ部分のみを抽出するようなキーフレーズの決定は困難であり，また，スポーツの種類によってキーフレーズを変更しなければならないため，これが汎用性の低下につながっている．

そこで本稿では，スポーツ映像中のアナウンサの発話内容を，スポーツの種類に依存性の低い表層的な特徴を用いて，スポーツ映像の構造に基づいて構造化することにより，映像上の各プレイに意味的に相当する発話部分を特定し，特定された発話内容の利用によりすべてのプレイ部分に対する意味内容情報の獲得を目指す．

2　スポーツ映像の構造に基づく意味内容情報

スポーツ映像は，スポーツの試合，番組の二つの観点から見た構造を持っている．スポーツの試合は一般にスポーツの種類によって形式が定義されており，例えば，アメリカンフットボールであれば，まず前半後半に分割され，次にクォーター部分，1チームの攻撃部分，各ダウンという様に細分化される．また，サッカーの場合，前半後半部分，さらに，1プレイ（スコア，ファール，ボールがフィールド外に出るなど）部分に分割される．ただし，アメリカンフットボールでは各ダウン部分が1プレイに相当する．このように，各スポーツに対して1プレイがスポーツの試合の最小構成要素となり，これを**ストーリーユニット**と定義する．

スポーツ番組は主に，実際に試合が進行するライブシーン，各ライブシーンの繰り返しであるリプレイシーン，その他番組中の試合に関係のないシーン，CMシーンで構成されており，ライブシーンは一般にある特定位置のカメラから撮影された見かけ上類似した画像から始まり，スコア，ファールなど何らかのイベントが発生すると，一度試合が中断され選手のクローズアップ，監督，観客席などの画像に移り変わる．つまり，一つのライブシーンは

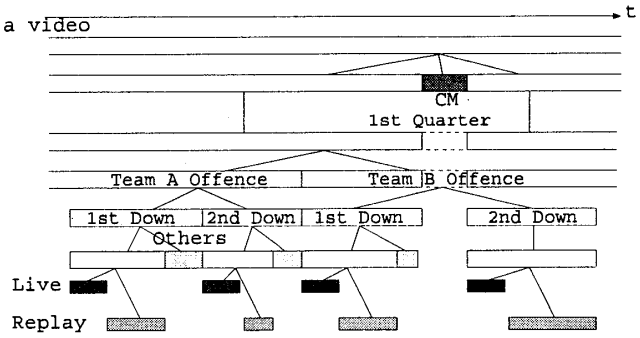

図 1: スポーツ映像の全体構造

1プレイと内容的に一致することになる．よってスポーツ番組上では，一つのライブシーンの始まりから次のライブシーンの始まりまでがストーリーユニットとなる．また，この最小構成要素である各シーンを**シーンユニット**とする．

つまり，スポーツ映像の全体構造は図1のようになり，一本の映像の意味内容情報獲得は，映像をストーリーユニットに分割し，各ストーリーユニットに対して試合上での意味内容情報（プレイ，選手，試合状況，スコア，時間など）を獲得することに相当する．

3　意味内容情報の獲得

2章で議論したように，スポーツ映像の意味内容情報獲得には，映像の各プレイ，つまりストーリーユニットへの分割，及び各ストーリーユニット部分の意味内容情報の獲得が必要となる．映像の分割に関しては，画像ストリーム上の特徴を利用したプレイ部分の検出により実現可能であると考えられる．しかし，画像特徴のみから各部分の意味内容情報を獲得することは困難であり，処理時間の点からも望ましくない．

さて，映像には画像の他にも音声，言語などの情報ストリームが存在するが，スポーツ映像においては，番組内のアナウンサの発話に，各プレイに関する多くの情報が含まれていると考えられる．また，アナウンサの発話は音声の写しである**クローズドキャプション（CC）**により容易に入手可能である．ただし，CCは文の並びにすぎず，各プレイの境界に関する情報は含まれていない．そこで提案手法ではまず，CCに対して各ストーリーユニットへの分割を行う．さらにこれらを，画像ストリームを利用した映像分割結果と対応付けることにより，各プレイの意味内容情報を含むCC部分を特定し，特定されたCC部分の利用により，すべてのプレイに関する意味内容情報を獲得する．

3.1　クローズドキャプションの構造解析

CC内での文章の意味的な区切りとしてはまず，話者の交替が考えられる．ただし，同一の話者が複数シーンにまたがって話し続けた場合，ここに変更の記述は行われない．そこで，文章中の間隔にも着目し，ある一定時間以上空白期間が存在する場合，話題が変化したものとし分割するようにする．ここでは，このように分割した各部分を**CCセグメント**と呼ぶ．

各CCセグメントは，その内容によりスポーツ番組上のシーンのいずれかに属する．ここで，提案手法では，”ライブシーン”，”

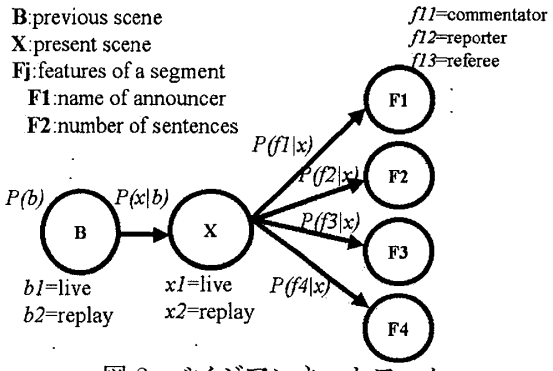

図 2: ベイジアンネットワーク

リプレイシーン，"CM シーン"，"試合に関係しない番組シーン"（観客席，スタジオシーンなど）の 4 種類をシーンカテゴリとする．

本稿では，同スポーツの中のさまざまな映像，また異なるスポーツへの汎用性を考慮し，CC セグメントの表層的な特徴による**ベイジアンネットワーク（BN）**を利用した CC セグメントのシーンカテゴリ分類法を提案する．

各 CC セグメントに対する特徴としては，発話者の名前，文の数，文の長さ，選手名の出現数，試合状況を示すフレーズ（"First down and 10" など）の出現の有無，数字（スコアやヤード数などを表す可能性が高い）の出現の有無の 6 種類としている．さらに，スポーツ番組の性質上，各シーンは並び方に特徴があると考えられる．よって，提案 BN では各 CC セグメントの特徴に加えその直前の CC セグメントのシーンカテゴリを用いる．図 2 に CC セグメントの各特徴とシーンカテゴリの関係を表した BN を示す．ノード X が現在の CC セグメントのシーンカテゴリを表し，親ノード B は直前の CC セグメントのシーンカテゴリを，子ノード F_j は現在の CC セグメントの各特徴を表す．この BN により，現在の CC セグメントがそれぞれのシーンカテゴリに分類される確率は次のように計算される．

$$P(x|e) = [\sum_{alle_B} P(x|e_B)P(e_B)]\prod_{j=1}^{|F|} P(e_{F_j}|x) \quad (1)$$

ここで，e は X 以外のノードの値，e_{F_j} と e_B はそれぞれのノードの値を表す．

上記の BN を用い，各 CC セグメントをシーンカテゴリに分類後，ライブシーンから次のライブシーンまでをストーリーユニットとして同定する．連続した CC セグメントがライブシーンに分類されている場合，それらの CC セグメントは映像上で同一のライブシーンに含まれている場合と，異なるライブシーンに含まれている場合がある．しかし，ライブシーン同士はあまり近接して出現しないと考えられるため，そのような連続した CC セグメントに対しては，セグメント間の時間間隔を計算し，それが閾値以内であった場合は一つのライブシーン，そうでなかった場合は異なるライブシーンとして分割する．このようにして構成される CC セグメント集合を **CC ストーリーユニット**と呼ぶ．

3.2 クローズドキャプションと映像の対応付け

2 章に述べたように，一つのライブシーンは画像上では特徴的な画像列から開始され，選手のアップ，監督などライブシーンと特徴の異なる画像へのカメラ切り替えによって終了する．よって，画像ストリームをショット分割し，各ショットの開始フレーム特徴の利用により映像でのライブシーンを検出し，一つのライブシーンから次のライブシーンまでを**映像ストーリーユニット**とする．

CC にはあらかじめ画像フレームと対応づけられるタイムスタンプを付与しておき，このタイムスタンプを基に映像上でのおおまかな開始時間を計算する．この計算結果と映像ストーリーユニットの開始時間を基に，CC と映像の時間差を考慮に入れ，最も適当なもの同士を対応付ける．

対応付け後は，CC ストーリーユニットの中でもライブシーン，リプレイシーンを主に利用し，対応付けられた映像ストーリーユニットに対する意味内容情報を獲得する．

4 実験結果

ここでは，意味内容情報獲得法の前半部分である映像及び CC のストーリーユニットへの分割に対する実験を行った．実際に放送されたアメリカンフットボール 2 本分（約 5 時間）の映像から抽出した CC に対し，複数の映像から抽出した学習データをサイズを変化させながら用いシーンカテゴリ分類実験を行った結果，平均精度約 60%で収束した．また，この結果を利用し CC ストーリーユニットを生成した結果，再現率は 93%，適合率は 71%であった．シーンカテゴリ中ライブシーンが最も精度良く分類されていたため，ストーリーユニット生成は比較的良好に行えている．さらに，同映像のそれぞれ 1 クォーター分（約 2 時間）の映像に対する映像ストーリーユニット抽出結果は再現率 92%，適合率 78%であり，さらに，再現率 89%，適合率 92%で CC と映像を意味的に対応付けることができた．対応付けの結果，互いの誤検出の削除，CC ストーリーユニットの補完が可能となり，画像ストリームのみを利用した映像分割結果の適合率，及び CC のみを利用したストーリーユニット生成結果を向上させている．

これらの実験により，提案手法の主要部分である，CC，映像のストーリーユニットへの分割，及び各映像部分の意味内容情報を含む CC 部分の特定に関して良好な結果を得ており．この結果は，提案手法の次のステップである，特定された CC 部分から各ストーリーユニットの意味内容情報獲得への十分な可能性を示している．

5 むすび

本稿ではスポーツ映像の構造を踏まえ，映像を記述する際に重要な意味内容情報に関して議論し，映像と，アナウンサの発話情報である CC の両方の構造解析に基づいた意味内容情報獲得法を提案した．また，実験により，CC 及び映像のストーリーユニットへの分割を行い，これらの結果が良好に対応付けられることを示した．これにより，映像での各ストーリーユニットの意味内容情報を含むと考えられる CC 部分を特定でき，すべてのプレイ部分に対する意味内容情報の獲得が実現可能になると考えられる．今後はより多くの映像に対して実験を行い，さらに CC ストーリーユニットからの意味内容情報獲得を試みる．なお，本研究の一部は、日本学術振興会科学研究費の補助を受けている。

参考文献

[1] D.Zhong, and S.F.Chang, "Structure Analysis of Sports Video Using Domain Models", *IEEE ICME'01*,pp.920-923,Aug. 2001.

[2] P.Xu, L.Xie, S.F.Chang, A.Divakaran, A.Vetro, and H.Sun, "Algorithms and System for Segmentation and Structure analysis in Soccer Video",*IEEE ICME'01*,pp.928-931, Aug.2001.

[3] N.Babaguchi, Y.Kawai, and T.Kitahashi, "Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration", IEEE Transaction on Multimedia, vol.4, no.1,pp.68-75, March 2001.

[4] 新田 直子, 馬場口 登, 北橋 忠宏, "放送型スポーツ映像の構造を考慮した重要シーンへの自動アノテーション付け", 信学論, Vol.J84-D-II, No.8, pp.1838-1847, 2001.