論　文

テキスト・音声・画像の協調的処理による放送型スポーツ映像におけるハイライト検出とインデクシング

宮内　進吾[†*]　　馬場口　登[†]　　北橋　忠宏[†**]

Highlight Detection and Indexing in Broadcasted Sports Video by Collaborative Processing of Text, Audio and Image

Shingo MIYAUCHI[†*], Noboru BABAGUCHI[†], and Tadahiro KITAHASHI[†**]

あらまし　本論文では，テキスト・音声・画像の協調的処理による放送型スポーツ映像におけるハイライト検出及びインデクシング法を提案する．提案手法では，まずクローズドキャプションと呼ばれるテキストストリームにおける単語の出現パターンを調べ，ハイライト出現の候補区間を検出する．次に，これらの区間を音声パワーに基づいて再評価し，誤検出と思われるものを棄却する．最後に，得られたハイライト区間をショットに細分割し，音声パワーと大域支配色情報からハイライトショットを同定することによりショットへインデックスを付与する．本手法を実際のアメリカンフットボール中継の映像に適用した結果，再現率 77%，適合率 84%でハイライト区間を効率良く検出し，更に，正しくハイライトを検出できた区間に対して，第 1 候補のみで 75%，第 2 候補までで 97%のショットインデクシングの正解率を得た．また，テキスト，音声，画像と段階的に解析することにより効率的な処理を実現し得ることを実験的に検証した．

キーワード　映像インデクシング，ハイライト検出，放送型スポーツ映像，クローズドキャプション，協調的処理

1. まえがき

近年における映像メディアの利用機会の増大に伴い，その蓄積や検索，編集など，より高度で適切なハンドリング手法が望まれている．映像は時間軸を有するメディアであるため，見たい箇所を探すには，現状では時間軸に沿って目視によりサーチせざるを得ない．これを補完する技術が映像へのインデクシング（索引付け）であり，特に映像の意味内容に基づいたインデックスを付与することは非常に重要であるといえる [1], [2]. 2001 年 10 月に標準化された MPEG-7 [3] においても，どのように映像メディアにインデックスを付与するかについては種々の議論がなされている．

ここでスポーツ映像を対象に上述のようなインデックスの生成を考えた場合，映像内容を特徴づける重要な要素の一つとして，ハイライトに関する情報がある．例えば，サッカーのゴールやアメリカンフットボールのタッチダウンなどのシーンがハイライトと考えられ，検索要求の最も高いものの一つといえる．また，そのようなハイライトを得ることができれば，ダイジェストの生成などにも利用することが可能である．そこで本論文では，テレビ映像などの放送型スポーツ映像（broadcasted sports video）を対象に，ハイライトの出現する時区間の検出法（ハイライト検出），及びその時区間中のショット（映像においてカメラ切換から次のカメラ切換まで間の連続する画像フレーム列）にインデックスを付与する方法（ショットインデクシング）について議論する．

ハイライト検出を利用したインデクシングは，映像インデクシングにおける重要な課題の一つとしてこれまでにもいくつかの提案がなされている．まず，いわば正攻法として考えられるのは画像情報に着目した手法である．これらの多くは，映像を構成する画像列と，ハイライトを表すモデルとを比較することにより検出を図るものである．例えば，Intille ら [4] はアメリカン

† 大阪大学産業科学研究所，茨木市
ISIR, Osaka University, 8-1 Mihogaoka, Ibaraki-shi, 567–0047 Japan
* 現在，松下電器株式会社
** 現在，関西学院大学

フットボールにおける選手の軌跡を追跡し，タッチダウンなどの得点イベントの検出を検討した．丸尾ら [5] は，サッカーにおいてゴールポストや選手，ボールの位置を画像から抽出することによりコーナーキックの検出を試みている．しかしながら，動画像解析によるハイライト検出は一般的に極めて困難な課題であり，モデルの設定方法や処理速度において問題が残る．

このため別のアプローチとして，画像，音声，テキストなどの様々なメディアを協調的・統合的に利用した手法が近年盛んに模索されている [6]〜[11]．Babaguchi ら [6], [7]，新田ら [8] は，このような複数メディア間の関連性に着目した協調的処理をインタモーダル協調と呼び，クローズドキャプションと呼ばれるテキスト情報と例画像を用いてアメリカンフットボールにおける得点イベントやアクションを検出している．Chang ら [9] は，キーワード・スポッティングと歓声検出，モデル画像を利用し，アメリカンフットボールにおけるタッチダウンを検出している．また，Rui ら [10] は特に音声情報に注目し，アナウンサの声の興奮や打球音を利用して野球におけるヒットの検出を試みている．以上をはじめ，これまで様々なメディアに着目したアプローチが試みられているものの，いまだ確立されたものとはなっていない．各種メディアの活用法や，その組合せ方については検討の余地を残している．

そこで本論文では，放送型スポーツ映像のマルチモーダル性に着目し，それを形成するテキスト，音声，画像情報を協調的に解析することによるハイライト検出法，並びにショットへのインデクシング法を新たに提案する [12]．以下，**2.** で提案手法の概要と考え方に言及し，**3.** でテキスト情報によるハイライト区間の検出，**4.** で音声情報によるハイライト区間の検証，**5.** でショットへのインデクシングについて述べる．**6.** において実際のアメリカンフットボールのテレビ中継映像に対する実験を行い，評価を加える．**7.** は本論文のまとめである．

2. 提案手法の概要

本論文では，映像メディアをテキスト・音声・画像というマルチモーダル性を有する情報ストリームの集合と考える．これらのストリーム間には，時間的，意味的に関連がある．このうちテキストは，ここで対象とする放送型映像に特有の情報であり，クローズドキャプション（CC: Closed Caption）テキストを指す．CC テキストは欧米でのテレビ放送や米国のビデオには標準的に入っているデータであり，音声の書き起こし（トランスクリプト）を表す文字情報である．ここでは，映像データから直接デコードし，テキストに時刻印を付すことにより音声や画像ストリームとの時間的対応をとる．スポーツ中継では，試合の内容をアナウンサが実況しているので，CC テキストにはハイライトの出現を示唆する単語が数多く含まれている．

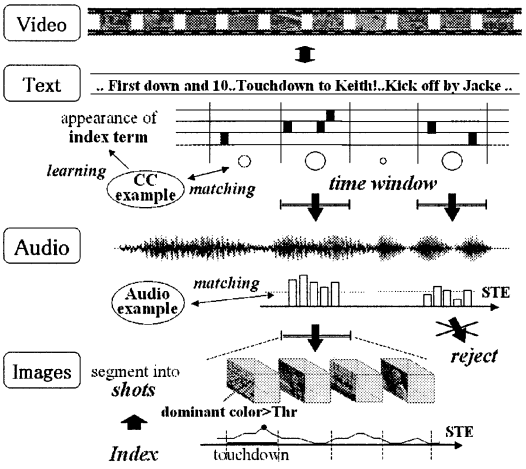

図 1　提案手法の概要
Fig. 1　Outline of the proposed method.

本手法では，テキスト，音声，画像の処理各々の計算負荷を考慮して，負荷の少ないストリームから順に処理し，探索範囲を絞っていくという戦略をとる．提案手法の概要を図 1 に示す．本手法はテキスト（text），音声（audio），画像（image）を解析する三つのプロセスから構成される．第 1 に，ハイライト区間の CC テキストに特徴的に現れるような単語や単語組についての共起パターンを学習サンプル集合から抽出する．そして，これらの出現パターンに基づき，検出対象のストリームにおいてハイライトが存在すると思われる時区間（時間窓（time window）と呼ぶ）を推定する．第 2 に，音声ストリームにおいて，ハイライトでは観客の歓声やアナウンサーの発声のレベルが上昇するという点に注目して，得られた時間窓を音声のパワー特徴に基づき検証する．本論文では，ここまでのプロセスをハイライト検出（highlight detection）と呼ぶ．最後に，画像フレーム間の処理を用いて，時間窓に対応する時区間の画像ストリームを更に細かくショットに分割した後，音声・色情報に基づき，ハイライトが開始するショットにインデックスを付与する．このプロセスをハイライトによるショットインデクシング

（highlight based shot indexing）と呼ぶ．なお本論文では，ハイライトをスポーツのルールで規定された得点イベントが出現するシーンとして考察していく．

3. テキスト情報によるハイライト区間の検出

本手法では，はじめにテキスト（CC）ストリームを解析することにより，検出対象のハイライトが存在すると思われる時区間に相当する時間窓を検出する．

スポーツ映像における CC テキストはアナウンサーの話し言葉が主であるため，構文解析などの自然言語処理の適用は難しい．ここでは，表層的な特徴として単語の出現パターンに着目する．また，CC ストリームを局所的に扱うために CC フレーム（CC frame）という処理単位を定義する．CC フレームとは，一定の時区間に対応する CC の部分テキストを指す．CC 解析は，学習データから CC フレームに対しハイライトの有無が判定可能な分類器（classifier）を生成する学習部と，この分類器を利用して未知ストリームでの時間窓を検出する検出部に大別される．

まず学習部では，学習データから検出対象のハイライトと関連の強い単語や単語組の共起パターンを抽出し，これらをハイライトを特徴づける索引語（index term）とする．そして，CC フレームをフレーム内における各索引語の有無を要素とするベクトルで表す．続いて，ハイライト区間に対応する事例の CC フレームを正例，対応しない CC フレームを負例として，これらのベクトル集合から分類器を構成する．検出部ではテキストストリームを CC フレームごとに走査して，上述の分類器により時間窓を定める．

3.1 索引語の抽出

ハイライト区間の CC テキストには，特徴的に現れる何らかの単語や単語組が存在するはずであり，何らかの手段でこれらを発見しハイライト検出に有効利用する必要がある．例えば，アメリカンフットボールにおけるタッチダウンを検出する場合，最も単純にそれ自身を表す “TOUCHDOWN” という単独のキーワードの出現を調べるという方法を考えよう．この種の主要キーワードは実際のイベント生起とは無関係に出現することが多く，その単独キーワードの出現する時区間を直ちにハイライト区間と同定すると誤検出が頻発することが報告されている [6], [7]．また，Babaguchi ら [7] は重み付きのキーワード列を設定しハイライト検出に利用しているが，キーワード列や重みの設定にはヒューリスティクスが不可欠である．

そこで本手法では，学習用データの CC 事例（CC example）における単語の出現頻度に基づいて，ハイライト部分に特徴的に現れる単語及び単語組，すなわち索引語を自動抽出する．索引語を抽出する際の尺度としては，1）ハイライト区間に高い頻度で出現すること，2）ハイライト区間以外にはあまり現れないこと，という 2 点が重要な要素となる．この観点に基づく評価尺度としては，TF-IDF（Term Frequency - Inverse Document Frequency）などが知られており，CC テキストからのキーワード抽出に適用した例 [13] も報告されているが，大規模な文書コーパスを要するため本研究に直接適用することは困難である．よって本手法では，単語及び単語組の重要度を求める評価式を次のように定義する．

$$w(t) = tf(t, d_{pos}) \times \frac{1}{tf(t, d_{all})},$$

$$w(t \to t_2) = tf(t \to t_2, d_{pos}) \times \log \frac{1}{tf(t \to t_2, d_{all})}$$

すなわち，単語 t の重要度 $w(t)$ は学習データ中のイベント区間に対応する CC フレームの集合 d_{pos} における t の相対出現頻度 $tf(t, d_{pos})$ と，全 CC 事例 d_{all} 中での相対出現頻度 $tf(t, d_{all})$ の比として求められる．単語組の重要度 $w(t \to t_2)$ についても同様であるが，$1/tf(t \to t_2, d_{all})$ が小さくなることが多いため，この値を伸張させるため対数をとっている．なお，ここでの単語組 $t \to t_2$ とは 2 単語 t, t_2 の一定間隔内での共起出現を指し，その順序も考慮する．

この重要度 w を，対象ドメインに関する一般的な用語などが登録された辞書を構成するすべての単語と単語組について計算し，この値の大きい上位 n 個を索引語とする．ただし，順位の計算は単語，単語組のそれぞれで行い，両者を合わせては行わない．そして，CC フレームをフレーム内での索引語 t_i の有無を表す以下のベクトル $\boldsymbol{x}$ で代表させる．

$$\boldsymbol{x} = (x_1, .., x_i, .., x_n), \quad x_i = \begin{cases} 1 & \text{if } t_i \text{ が存在} \\ 0 & \text{otherwise} \end{cases}$$

3.2 分類器の生成

正例・負例の CC フレームを用意し，各事例に対応するベクトルを学習データとしてベクトル空間に登録し，分類器を構成する．ここでの分類器には，k-最近傍法（k-NN 法）を利用する．すなわち，分類対象として，ある CC フレームを表すベクトル $\boldsymbol{x}$ が与えら

れたとき，ベクトル空間上でそれぞれの正例・負例 $\boldsymbol{x_e}$ との距離 d を次式で求める．

$$d = \sqrt{\sum_{i=1}^{n} w_i' \, (\, x_i - {x_i}^e)^2}$$

ただし，$x_i, {x_i}^e$ は各々ベクトル $\boldsymbol{x}$, $\boldsymbol{x_e}$ の i 番目の要素を表し，w_i' は i 番目の索引語の重要度を正規化した値である．そして，この d を最小とするような k 個を正例・負例中から選択し，それらの中での正例数 k_{pos} の割合 $P(\boldsymbol{x})$ としきい値 P^* から分類器 $C(\boldsymbol{x})$ は，

$$P(\boldsymbol{x}) = \frac{k_{pos}}{k} \ , \qquad C(\boldsymbol{x}) = \begin{cases} 1 & \text{if } P(\boldsymbol{x}) > P^* \\ 0 & \text{otherwise} \end{cases}$$

のように 1 あるいは 0 を出力し，出力が 1 のときその CC フレームをハイライト区間と判定する．

3.3 時間窓の検出

検出部では，前述の分類器を利用し，対象とするストリームでの時間窓を検出する．時間窓検出の流れを図 2 に示す．CC ストリームに対し CC フレームを処理単位として，その区間におけるハイライトの有無を判定する．このとき CC フレームの位置をシフトしながら，テキストの先頭から末尾までを走査する．シフト幅は CC フレーム長以下として，CC フレームを重畳させながら移動を繰り返す．

CC フレームにおいて各索引語の有無を調べ，n 次元ベクトル $\boldsymbol{x}$ を生成する．そして分類器 $C(\boldsymbol{x})$ によりハイライト区間と判定されれば，その時区間を時間窓として検出する．なお時間窓が重畳して検出された場合は，得られた時間窓のうち $P(\boldsymbol{x})$ が最大のものを，値が同じ場合は先のものを時間窓として選択する．

提案手法では，最終的に試合の進行（インプレイ）

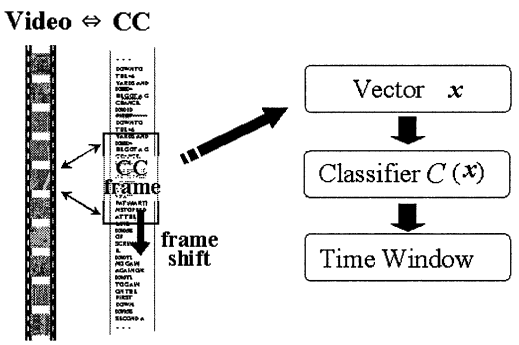

図 2　時間窓の検出
Fig. 2　Detection of time windows.

シーン中でハイライトが開始するショットを定めることを目標とするため，CC フレームの長さは数個のショットを含むように設定する．ハイライトの開始点がわかれば，後続ショットを調べることにより，ある程度長いハイライトにも対処できると考えている．

4. 音声情報によるハイライト区間の検証

CC テキストを利用した時間窓検出の問題点として，テキストと画像の内容は必ずしも一致しないという点が挙げられる．例えば，プレイの合間やハーフタイムなどにそれまで起こったハイライトを振り返って解説しているものなどは，CC 解析のみでは区別できない可能性がある．よってここでは，試合中の得点イベントなどのハイライト区間では歓声やアナウンサーの声が非常に強くなることに注目し，音声解析によるこうした盛り上がりの評価に基づいて，不適当な時間窓を棄却して誤検出の減少を図る．

4.1 音声特徴量

音声特徴量として，音声のパワーの尺度である短時間エネルギー（STE: Short Time Energy）[14] を利用する．$X(m)$ を離散音声信号，$W(m)$ を方形窓としたとき，サンプル数 L の音声フレームにおける STE E_l（l は時間インデックス）を以下に示す．

$$E_l = \sqrt{\frac{1}{L} \sum_m [X(m)W(l-m)]^2} \ ,$$

$$\text{ただし，} \ W(l) = \begin{cases} 1 & \text{if } 0 \leqq l \leqq L-1 \\ 0 & \text{otherwise} \end{cases}$$

ハイライト区間においては STE の値は平均的に高く，また歓声が重畳しているためその分散は低いと考えられる．本手法では時間窓内における STE の平均と分散の二つを特徴量とした 2 次元特徴ベクトル $\boldsymbol{y}$ を用いて時間窓を検証する．

4.2 時間窓の棄却

ハイライト区間か否かの判定は，CC 解析と同じく学習データから作られるモデルと比較することにより実現する．まず，あらかじめ用意したハイライト区間に対応する音声のサンプルを正例，対応しない音声サンプルを負例として，それぞれにおける STE の平均と分散を求める．そして，正例集合，負例集合各々各要素について平均値をとることで，正例・負例のモデルを生成する．

この両モデル $\boldsymbol{y_e}$ と判定対象の特徴ベクトル $\boldsymbol{y}$ を

2 次元ベクトル空間上で比較し，最近傍法に基づいて時間窓か否かを判定する．ベクトル間のユークリッド距離

$$d' = \sqrt{\sum_{i=1}^{2} \frac{1}{\sigma_i} (y_i - y_i{}^e)^2}$$

によりベクトル $\boldsymbol{y}$ が負例のモデルに近ければ，その時間窓を棄却する．上式の $y_i, y_i{}^e$ はそれぞれ，ベクトル $\boldsymbol{y}, \boldsymbol{y_e}$ の i 番目の要素を表す．また各要素は，学習データにおけるそれぞれの標準偏差 σ_i で割ることにより正規化を施す．

5. ショットへのインデクシング

以上で述べたテキスト・音声のストリームの解析処理から，ハイライトを含む可能性の高い時間窓を得ることができる．しかしながら，時間窓は数分の時間幅をもつ時区間であるため，本手法では更に，時間窓の範囲で試合進行（インプレイ）中に得点イベントが開始するショット（ハイライトショット（highlight shot）と呼ぶ）を同定し，得点イベント名のインデックスを付与する．ハイライトによるショットインデクシングの概略を図 3 に示す．

まず，時間窓に対応する部分画像ストリームを更に短いショットに分割する．この分割はショット境界を検出することにより行われ，ここではブロックマッチングと色ヒストグラムを利用する [8]．時間窓をなす画像フレームの列において，ブロックマッチングを用いて隣接フレーム間の非類似度を求めることにより，ショット切換操作の可能性の高い画像列を検出する．ここでは，ブロック間の類似度は色ヒストグラムの絶対値差分和として求める．そして，検出された画像列に対してカットやディゾルブなどの切換操作別の判定を適用することにより，ショット境界を得る．

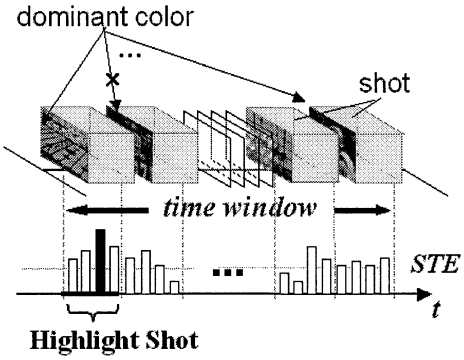

図 3　ハイライトによるショットインデクシング
Fig. 3　Highlight based shot indexing.

次に，分割されたショット群の中からハイライトショットを同定する．この同定に音声特徴である STE（前章参照）と画像特徴である大域支配色を利用し，以下の 2 条件を満たすショットを選択する．

- STE はハイライト発生時の周辺において最も高い値をとると考えられる．そこで，時間窓内における毎秒（1 秒窓 [15]）の STE の値の平均を求め，この最大値を含むようなショットをハイライトショットの候補とする．
- 試合中のプレイ（得点イベントなどを含む）を撮ったショットでは，フィールドなどの背景が大域的に広く映ると考えられる．そこで，このような大域支配色（dominant color）に対応する色領域を RGB 値に対する条件式として定め，各ショットの先頭フレームにおける大域支配色の画素数の割合がしきい値以上であるものを選択する．この制限により，例えばハイライトの直後に挿入される選手のクローズアップといったショットは削除される．

最終的に，ハイライトショットをその開始と終了の画像フレーム番号で記述し，得点イベント名のラベルをそのショットに付与する．

6. 評価実験

提案手法をアメリカンフットボール NFL（National Football League）のテレビ中継の映像に適用し，手法の有効性を検証した．検出対象のハイライトは，得点イベントであるタッチダウン TD（付随イベントのエキストラポイント EP を含む），フィールドゴール FG の 2 種類である．16 試合分の映像を利用し，そのうち 10 試合を学習用データとし，残る 6 試合を実験用データ（未知データ）とした．学習用データ，実験用データの映像の総時間数は各々35，20.5 時間である．それぞれの映像は FOX，CBS，abc の三つの制作会社によるものであり，計 13 名のアナウンサー・解説者による実況を含む．

6.1 テキスト情報によるハイライト検出結果

本実験における CC フレーム，すなわち時間窓の長さは経験的に 60 秒と設定した．これは TD や FG のシーンを内部に包含することができる十分な時区間長である．また CC フレームを移動するときのシフト幅を 10 秒とした（**3.3** 参照）．

ATTACK, ATTEMPT, BACK, BALL, BEHIND, BLOCK, CATCH, CAPTAIN, CHANGE, CONVERSION, DOWN, DEFENCE, END, EXTRA, FIELD, FIRST, FLAG, FOOTBALL, FORMATION, FOUL, FUMBLE, GAIN, GAME, GOAL, GOOD, HALF, INTERCEPT, KICK, LEAD, LINE, LINEBACKER, LONG, MISS, OFFENSE, PASS, PENALTY, PLAY, PLAYER, POINT, PUNT, QUARTER, QUARTERBACK, REFEREE, RETURN, RUN, SCORE, SECOND, SHOT, SHORT, STOP, TACKLE, TEAM, THROW, THIRD, TIME, TOUCHDOWN, YARD, ZONE, *number*, *player*

図 4　アメリカンフットボールの用語辞書（*number*：数字，*player*：選手名）
Fig. 4　Vocabulary for American football.

表 1　抽出された TD の索引語（*kicker*：キッカー名，*QB*：クォータバック名）
Table 1　Extracted terms for TD.

順位	単語	w'	単語組	w'
1	extra	1.00	extra → point	1.00
2	touchdown	0.69	*kicker* → point	0.93
3	point	0.54	point → good	0.81
4	zone	0.40	end → zone	0.71
5	*kicker*	0.33	*QB* → touchdown	0.56
...	...		...	

用意した 10 試合の学習データに含まれる正例の CC フレーム数は，TD，FG ともに 36 個である．負例については，ハイライト区間以外から無作為に抽出した CC フレームと，検出対象ではない方のハイライト（TD を検出する際は FG）をそれぞれ正例と同数用意し，両者を合わせた 72 個を利用した．

まず，学習データからハイライトを特徴づける索引語を抽出した．抽出の際には，辞書として図 4 に示すようなアメリカンフットボールの用語約 60 語と選手名などを登録し，ドメインに関する知識として利用した．抽出された TD の索引語の上位を表 1 に示す．

次に，時間窓の検出結果について述べる．検出のための各パラメータは，以下のような予備実験により決定した．予備実験は学習用データ 10 試合を対象とした．また，評価尺度には再現率=正検出数/正答数，及び適合率=正検出数/検出数を用い，正検出は検出された時間窓内に実際にハイライトが含まれる場合を指す．

- 検出に利用する索引語を選択するために，以下の諸条件のもとでハイライトの検出性能を比較した．単語のみ，単語組のみ，両者を併用の各条件について 5 個，10 個の索引語を選択して比較したところ，表 2 に示す結果が得られた．この結果から，再現率を重視した上で適合率の良好な 10 単語と 10 単語組の併用（計 20 個）を用いるものとした．
- k-NN 法における k の値は，k を変化させて検出性能を比較した結果，最も性能の高かった 17 とした．また $P^* = 1/2$ とした（**3. 2** 参照）．

以上の条件のもと，未知サンプルである実験用データ 6 試合（TD，FG の数は各々28，19）に対する検出結果を表 3 の左（CC 解析欄）に示す．まず未検出の場合には以下のものが存在した．

表 2　索引語選択に関する予備実験
Table 2　Preliminary experiments for term selection.

索引語	再現率	適合率
5 単語	93%	44%
10 単語	93%	38%
5 組	67%	71%
10 組	80%	87%
5 単語+5 組	93%	58%
10 単語+10 組	93%	70%

表 3　ハイライト検出結果
Table 3　Result of highlight detection.

	CC 解析		CC+音声解析	
	再現率	適合率	再現率	適合率
TD	23/28 (82%)	23/32 (72%)	22/28 (79%)	22/24 (92%)
FG	15/19 (79%)	15/26 (58%)	14/19 (74%)	14/19 (74%)
計	81%	66%	77%	84%

- キックオフリターン，2-point conversion（EP）といった例外的な TD：2
- ファウル審議後に TD とされた特殊な場合：1
- TD 後の EP の実況が省略されたもの：2
- “FIELD GOAL” という言葉の省略：2
- そのほか表現の省略や特殊な言い回し：2

一方，誤検出の場合で最も多く見られたのは，これまでに起こったハイライトについて，ハイライトの直後やハーフタイムなどに振り返って解説しているものであった．次に多いのは，TD に至らなかった，あるいは FG に失敗した，といった場合である．

これらの誤りは，CC テキストのみを利用したときにはある程度やむを得ないものと思われる．しかし，各索引語間には冗長性が見られることから，抽出する索引語の選択法については改善の余地があろう．

6. 2　音声情報によるハイライト検出結果

本実験で解析に利用した音声データはサンプリング周波数 11.025 kHz，量子化ビット数 16 bit，モノラルであり，フレーム長を 512，フレーム周期を 128 とし

た．また，時間窓内にコマーシャルが含まれていた場合，その性質が大きく異なることから今回は手動により除去した．学習用の音声の正例には CC 事例の正例と同じ箇所から TD，FG について各々18 個，12 個を用意し，負例についてもそれぞれ同数用意した．

ハイライト検出の結果は表 3 の右（CC+音声解析欄）に見られるとおり，再現率では 4%低下したものの，適合率では 18%もの向上が見られた．これより音声情報による時間窓の棄却が極めて良好に動作していることが確認された．正検出を誤って棄却した場合は，試合の終盤で勝敗を左右しないものなど，いずれも非常に盛り上がりの少ないものであった．逆に，棄却できなかった場合は，惜しくも TD に失敗したものなど，それ自体が大きく盛り上がったためである．

さて，棄却できなかったシーンは，現実には得点イベントのシーンではなかったものの，視聴者の立場から見ると極めて興味深いシーンである可能性が高い．本論文では客観的評価が行いやすいという面から，ハイライトを得点イベントの出現するシーンと定義して手法を設計したが，例えば，僅差の終盤で FG を入れると逆転という状況で最後の FG が外れるというシーンは本手法でのハイライトには含まれないという弊害がある．仮にハイライトの定義を視聴者が興味あるシーンや重要なシーンというように変更すると，処理戦略は大きく変化すると思われる．一案として提案手法のプロセスとは逆に，音声の盛り上がり部分を優先的に探索した後，詳細にテキストの意味解析を行う，若しくは画像中に出現して得点状況を知らせるオーバレイを解析するなどが想定されるが，この場合ハイライトの定義が主観的になるため，評価の点では困難が予想される．映像メディアとハイライトの関係は種々の観点から検討されるべきで，今後の重要な課題であろう．

6.3 ショット インデクシングの結果

ショット分割に関しては，カットやディゾルブによる切換はほぼ正確に分割することができた．ただし，CG を用いた特殊なショット切換については，今回は人手により補完した．この結果，60 秒の時間窓に対し，平均して約 8.8 個のショットが含まれていた．

分割されたショットの中からハイライトショットを同定し，ショット インデクシングを試みた．色情報としては，大域支配色としてフィールドの緑を考え，具体的な RGB の条件式には，$R \leqq G, B \leqq G, 50 \leqq G \leqq 200$ を与えた（**5.** 参照）．ここで，RGB はそれぞれ 256 階調である．また，大域支配色の画素数の割合に対するしきい値を 1/3 と定め，正しいハイライトショットを誤って除外しないよう低めに設定している．

表 4　ショット インデクシング結果
Table 4　Result of shot indexing.

	TD	FG	計
第 1 候補のみ	82%	64%	75%
第 2 候補も含む	96%	100%	97%

結果は表 4 のとおりであり，様々な構図・カメラアングルからなるショットを正しく同定することができた．第 2 候補のショットまで含めた場合は 100% に近い正解率が得られたことから，おおむね良好な結果といえる．誤った例では，ハイライトショットの直前・直後に同定したものが多く，前後の誤差を許容すると正解率は 89%となる．特に FG の正解率が低い原因は，FG 自体がさほど盛り上がらないためであり，サードダウンのショットに誤る例が多かった．

また，色情報を相補的に利用したことにより，TD 直後の選手や監督のアップ，FG 直前のキッカーのアップなどのショットに誤って同定するのを回避できた．

6.4 考　　察

まず，本手法の処理速度について述べる．SGI OCTANE（R12000，300 MHz）を用いて，1 試合（約 3 時間）の映像に対して時間窓の検出に要した時間は，CC 解析に約 2.4 秒，音声解析については平均 9.6 個の時間窓に対し約 11.4 秒であった．一方，ショット分割とハイライトショットの同定については，CC，音声解析により得られた平均 7.1 個それぞれの時間窓について，実時間の半分以下での処理が可能であった．すなわち，1 試合当りの処理時間は合計約 4 分と非常に高速である．これは，CC 解析により以降の負荷の大きな処理の範囲が全体の約 5%に限定されたためであり，段階的な処理の効果といえる．

次に，本手法を類似した従来手法 [7] と比較する．従来手法では，3 試合のアメリカンフットボール映像を対象に，CC テキストのキーワード列と例画像を利用して TD と FG を検出している．結果は再現率 81%，適合率 74%でショットに対してインデックスを付与しており，1 試合当りの処理時間は 340 秒と報告されている．精度，速度の面から従来手法と提案手法に大きな差異は見られないものの，ヒューリスティクスに依存する部分が提案手法の方が少なく，一般性が高いと考える．従来手法 [7], [8] では，本論文での索引語に相当

するキーワード列やその重みの設定法などに，ヒューリスティクスが不可欠であり，またそのための作業労力も無視できない．

一方，提案手法では，図 4 に示す用語辞書を一度用意すれば，索引語やその重みも自動的に導出できる．ここでは，用語辞書は設計者が手作業により準備したが，統計的言語処理の知見 [16] を利用して，このプロセスの自動化が急務である．この問題はフルテキスト文書からのキーワードの自動抽出問題と定式化でき，TF-IDF などの尺度が提案されているが，その利用に必須となる大量コーパスを本問題においてどのように設定すればよいかなど，克服すべき課題は多い．加えて，用語辞書には，ドメイン（ここではアメリカンフットボール）に依存する部分と，試合に依存する部分の双方が存在する．前者は個々の試合の映像には無関係であるため一度作成すると十分であるが，後者は試合の映像ごとに組み込む必要がある．出場選手など毎試合に関するデータについてはインターネット上にあるゲームスタッツの公開サイト（例えば http://www.nfl.com/）を外部メタデータとして参照することが考えられる．

6.5 他スポーツ映像への適用可能性

本節では，本手法を野球映像に適用し，少数データをもとにホームランの検出を試みることによって他のスポーツへの適用可能性を議論する．この実験に利用した映像は，MLB（Major League Baseball）のテレビ中継 5 試合，それぞれ約 3 時間であり，計九つのホームランを含む．これら 5 試合の映像に対し，一つを実験対象，残る四つを学習データとした実験を交互に 5 回繰り返しその精度を評価した．約 60 語の用語辞書を利用し抽出した索引語 10 単語，10 単語組を用い，学習データにおけるすべてのホームランを正例としてハイライト検出したところ，表 5 のような結果が得られた．音声解析併用による検出結果も同時に示す．

ここで CC 解析における検出能力が非常に低いのは，学習データの量，特に正例数の不足に起因するところが大きい．同数の正例を用いてアメリカンフットボールの場合について検証したところ，TD の検出性能は再現率 64%，適合率 39%となり，学習データの不十分さによる性能の低下が裏づけられよう．しかしながら，誤検出の棄却に関しては良好な特性を示していることから，十分な学習データさえ用意すれば，アメリカンフットボールの場合と同等の水準に達することは可能と思われる．

表 5　ホームランの検出結果
Table 5　Result of home run detection.

	再現率	適合率
CC のみ	5/9 (56%)	5/18 (28%)
CC+音声	5/9 (56%)	5/7 (71%)

ショットインデクシングについては，大域支配色の設定をアメリカンフットボールと同様にして実験した結果，検出した五つのハイライト区間すべてにおいて，ホームランを打った直後のショットがハイライトショットに同定された．野球では一般的に，投球ショットから打球を追うショットへの連続する 2 ショットをハイライトショットとするのが妥当であり，同定した直前のショットも含めた 2 ショットに対してインデックスを付与するものとすれば，この結果は十分といえる．

以上を踏まえ本手法の一般性について議論すると，CC 解析については，CC テキストとハイライトとの対応が十分期待できる場合に対しては，用語辞書と十分な学習データを用意すれば，他のスポーツにも対応し得ると思われるものの，これに関しては更なる実験的な検証が必要である．音声の利用については，大きな歓声の起こるようなハイライトに対しては，他のスポーツにも適用でき良好に動作すると推察できるが，音声のパワーによって識別できる種類数などについては，より詳細な実験的検証が必要である．また，色情報としての大域支配色とハイライトとの対応が十分期待できるような対象については，例画像などを用いることなく良好にその候補が選択可能であることが期待できる．

7. むすび

本論文では，テキスト・音声・画像の協調的処理により放送型スポーツ映像から，ある時区間にハイライトが出現するか否かを調べるハイライト検出法，及びハイライト区間中で得点イベントが開始するショットにインデックスを付与するショットインデクシング法を述べた．本手法を実際のアメリカンフットボール中継の映像に適用したところ，再現率 77%，適合率 84%でハイライト検出が実現でき，正しくハイライトを検出できた時区間に対して，第 1 候補のみで 75%，第 2 候補までで 97%のショットインデクシングの正解率を得た．また，テキスト，音声，画像と段階的に解析することにより高速な実行速度を実現した．

今後の課題としては，CC 解析の段階における再現

率の向上が挙げられる．誤検出については音声情報を用いて良好に棄却可能であることが実験的に確認されたため，この点を改善すればいっそうの精度向上が期待できる．更に，提案手法の一般性を多種類のスポーツ映像で実証する必要がある．

謝辞 本研究の一部は，通信・放送機構創造的情報通信技術研究開発制度，及び科学研究費の補助による．

文　献

[1] S.W. Smoliar and H.J. Zhang, “Content based video indexing and retrieval,” IEEE Multimedia, pp.62–72, 1994.
[2] 馬場口登，“メディア理解による映像メディアの構造化，”信学技報，PRMU99-42, July 1999.
[3] “Overview of the MPEG-7 standard (version 6.0),” ISO/IEC JTC1/SC29/WG11 N4509, 2001.
[4] S.S. Intille and A.F. Bobick, “Recognizing planned, multi-person action,” Computer Vision and Image Understanding, vol.81, no.3, pp.414–445, March 2001.
[5] 丸尾二郎，岩井儀雄，谷内田正彦，越後富夫，飯作俊一，“サッカー映像からの特定映像イベントの抽出，”信学技報，PRMU99-41, July 1999.
[6] N. Babaguchi and R. Jain, “Event detection from continuous media,” Proc. 14th ICPR, vol.II, pp.1209–1212, 1998.
[7] N. Babaguchi, Y. Kawai, and T. Kitahashi, “Event based indexing of broadcasted sports video by intermodal collaboration,” IEEE Trans. Multimedia, vol.4, no.1, pp.68–75, March 2001.
[8] 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（D-II），vol.J84-D-II, no.8, pp.1838–1847, Aug. 2001.
[9] Y.L. Chang, W. Zeng, I. Kamel, and R. Alonso, “Integrated image and speech analysis for content-based video indexing,” Proc. IEEE ICMCS’96, pp.306–313, 1996.
[10] Y. Rui, A. Gupta, and A. Acero, “Automatically extracting highlights for TV baseball programs,” Proc. ACM Multimedia 2000, pp.105–116, Oct. 2000.
[11] 佐野雅規，住吉英樹，井上誠喜，“映像版スコアブックの検討，”信学技報，PRMU99-257, March 2000.
[12] 宮内進吾，馬場口登，北橋忠宏，“音声・言語・映像の協調的処理によるスポーツ映像からのイベント検出，”信学技報，PRMU2001-214, Jan. 2002.
[13] M.A. Smith and T. Kanade, “Video skimming and characterization through the combination of image and language understanding techniques,” Proc. CVPR97, pp.775–781, 1997.
[14] T. Zhang and C.J. Kuo, “Heuristic approach for generic audio data segmentation and annotation,” Proc. ACM Multimedia 1999, pp.67–76, Oct. 1999.
[15] L.L.H. Jiang and H.J. Zhang, “A robust audio classification and segmentation method,” Proc. ACM Multimedia 2001, pp.203–211, 2001.
[16] 徳永健伸，情報検索と言語処理，東京大学出版会，1999.

（平成 14 年 3 月 19 日受付，5 月 16 日再受付）

宮内　進吾　（学生員）

平 12 阪大・基礎工・情報卒．平 14 同大大学院前期課程了．同年，（株）松下電器産業に入社．在学中はメディア理解に関する研究に従事．

馬場口　登　（正員）

昭 54 阪大・工・通信卒．昭 56 同大大学院前期課程了．昭 57 愛媛大学工学部助手．大阪大学工学部助手，講師を経て，現在，大阪大学産業科学研究所助教授．平 8〜9 UCSD・文部省在外研究員．工博．メディア処理・人工知能の研究に従事．IEEE，ACM，情報処理学会，人工知能学会各会員．

北橋　忠宏　（正員）

昭 37 阪大・工・通信卒．昭 43 同大大学院博士課程了．同年大阪大学基礎工学部助手．同助教授，豊橋技術科学大学助教授，教授を経て，昭 61 大阪大学産業科学研究所教授．平 14 関西学院大学教授．阪大名誉教授．工博．メディア処理，物体認識，文書画像処理に関する研究に従事．情報処理学会，IEEE，人工知能学会，日本認知学会各会員．