音声言語ストリームに着目したスポーツ映像からのイベント検出

5Q−5

宮内 進吾　馬場口 登　北橋 忠宏
大阪大学 産業科学研究所

1 はじめに

近年における計算機の大容量化や放送・通信技術の発展に伴い，映像メディアに対する期待が一層の高まりを見せている．この膨大な情報量を持つ連続メディアを効率よく利用するためには，映像コンテンツに関する情報を付与し，コンテンツに基づいて構造化しておく必要がある．しかしながら，画像解析によるコンテンツ情報の抽出は困難であり，現状では精度・処理速度において問題が残る．このため，画像と音声・言語をはじめとする種々のストリーム間の関連性に注目したアプローチが現在盛んに模索されている [1]．そこで，本稿では特に放送型映像メディアにおける音声言語ストリームに着目し，これを利用した時区間イベントの検出法について検討する．

なお，本稿では検討対象のドメインをスポーツ映像とする．スポーツ番組は，コンテンツに基づいた検索や要約が望まれるドメインの一つである．

2 言語情報を利用したイベント検出法

本稿では，音声のトランスクリプトとして得られるクローズドキャプション (Closed Caption:CC) と呼ばれる言語 (テキスト) ストリームを利用し，スポーツ映像から特定の時区間イベントを検出する手法について述べる．CC を利用したイベント検出は，テキストの検索・分類などの問題に帰着させることができる．ここでは，この情報検索の分野で広く用いられるベクトル空間モデルの適用を試みる．

提案手法は，イベント区間に対応する CC のセグメントに特徴的に現れる語や語の組を抽出し，これらを各次元に対応させたベクトル空間を生成する学習部と，このベクトル空間を利用して未知映像におけるイベント 発生区間を検出する検出部に大別される．

2.1 イベントを特徴付ける索引語の抽出

まず学習部では，あらかじめ用意されたサンプルの CC から，検出対象のイベントと相関性の強い語および語の組を抽出する．以下ではこれらを併せて索引語と呼び，そしてこの索引語を各次元の特徴量に対応させたベクトル空間を考える．

2.1.1 索引語の重み計算

本手法では，イベントを特徴付ける索引語を抽出するにあたり，その候補としてドメインに関する用語や人名をあらかじめ関連語リストという形で登録しておく．そして，サンプルの CC におけるこれらの出現状況を調べることにより，イベント区間の CC セグメントに特徴的に出現する索引語を選択する．

このような索引語を抽出する際の評価尺度としては，

・イベント区間に高頻度で出現する (網羅性)

・イベント区間以外にあまり現れない (特定性)

という 2 点が重要な要素となる．情報検索の分野においては，この条件に基づいた尺度として $tf \cdot idf$ などの有用性が認められている [2]．本手法では，これらを参考に語，および語の組の重みを求める評価式を次のように定義する．

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

$$w(t_1+t_2) = tf(t_1+t_2, d_{event}) \times log \frac{1}{tf(t_1+t_2, d_{all})}$$

すなわち，語 t の重み $w(t)$ はイベント区間の CC セグメントの集合 d_{event} における t の相対出現頻度 (t の出現数 / d に含まれる語数) と全 CC データ d_{all} 中での相対出現頻度の比として求められる．語の組 t_1+t_2 の重み $w(t_1+t_2)$ についても同様である．なお，ここでの語の組とは 2 語 t_1, t_2 が一定間隔以内に共起して出現することを指す．

2.1.2 ベクトル空間による表現

この重み w を，登録された全ての語およびその組合せについて計算し，値の大きい上位 n 個をイベントを特徴付ける索引語とする．そして，CC セグメントをこれら索引語の出現頻度を特徴量としたベクトルで表現することを考える．すなわち，いま n 個の索引語を選択した場合，CC セグメントはそれぞれの出現頻度からなる n 次元ベクトルで表すことができ，n 次元空間上の 1 点として表現される．

さらに，正例・負例となるサンプルの CC セグメントそれぞれについてこれらの特徴量を算出し，あらかじめベクトル空間上に登録しておく．この空間を利用することにより，未知セグメントにおけるイベントの出現判定が可能となる．

2.2 ベクトル空間モデルに基づくイベントの検出

イベントの検出処理は，対象の映像ストリームを一定の時区間，すなわち時間窓ごとに区切り，この時間窓におけるイベントの有無を順次判定していくことで実現される．イベント検出の流れを**図 1** に示す．

2.2.1 時間窓のベクトル表現

まず，区切られた時間窓に対応する CC セグメントにおいて，選択された特徴量，すなわち各索引語の出現頻度をそれぞれ算出し n 次元ベクトルを生成する．これにより，時間窓もまたベクトル空間上の一点として

Detection of Events from Sports Video Using Auditory and Linguistic Streams

Shingo MIYAUCHI, Noboru BABAGUCHI and Tadahiro KITAHASHI, *I.S.I.R.* Osaka University

表 1: touchdow に関連する索引語の抽出結果

rank	term	w (w')	term (set)	w (w')
1	EXTRA	13.0 (1.00)	EXTRA+POINT	0.158 (1.00)
2	TOUCHDOWN	6.58 (0.50)	EXTRA+*kicker*	0.122 (0.77)
3	POINT	5.44 (0.41)	GOOD+POINT	0.118 (0.74)
4	*kicker*	4.77 (0.36)	POINT+*kicker*	0.113 (0.71)
5	ZONE	3.71 (0.28)	EXTRA+GOOD	0.106 (0.67)
...	...		...	

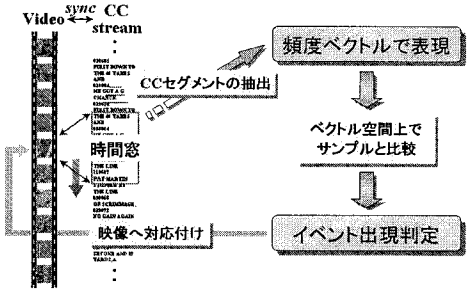

図 1: 検出システムの概観

表すことができる．そして，空間上でのサンプルデータとの類似度に基づいて，この時間窓におけるイベントの有無が判定される．

2.2.2 k-最近傍法によるイベント出現判定

ベクトル空間における正例・負例サンプルとの類似度は，空間上でのベクトル間の Euclid 距離により評価可能である．しかしながら，ベクトルの各要素がイベントの判定に果たす重要度は異なり，この重要度はすなわち求められた各索引語の重みにあたる．そこで，索引語の重みを考慮した距離尺度 d を次のように定義する．

$$d = \sum_{i=1}^{n} w_i' \left({x_{window}}^i - {x_{sample}}^i \right)^2$$

${x_{window}}^i$ は時間窓を表す n 次元ベクトルの i 番目の要素であり，${x_{sample}}^i$ はサンプルのそれを指す．また，w' は各索引語の重み w を正規化した値である．

この類似度評価を用い，k-最近傍法 (k-Nearest Neighbor rule) に基づいて時間窓におけるイベントの有無を判定する．すなわち，空間上で d を最小とする k 個のサンプルにおいて正例数＞負例数となれば，この時間窓をイベント発生区間と判定する．

3 評価実験

本手法をアメリカンフットボールの TV 中継に適用し，基礎実験として touchdown イベントの検出を試みた．サンプルの CC データは 9 試合・約 30 時間分用意し，正例として touchdown に対応する CC セグメントを 30，負例として 50 の CC セグメントをそれぞれ用いた．また，時間窓の長さを 90 秒 とし，k-最近傍法における k の値は 3 とした．

まず，約 50 語の関連語リストに基づいて touchdown を特徴付ける索引語を抽出した．抽出された語，および語の組を**表 1** に示す．次に，これら上位 5 つの索引語の出現頻度をそれぞれ特徴量とし，5 試合に対して touchdown の検出実験を試みたところ**表 2** に示す結果を得た．また，1 試合 (約 3.5 時間) あたりの平均処理速度は，SGI O_2 上で 0.9 秒と非常に高速であった．なお，ここでの正検出とは検出した時間窓内に実際のイベントが含まれることを指す．

表 2: touchdown の検出結果

特徴量	再現率	適合率
5 単語の頻度	17/23 (74%)	17/36 (47%)
5 単語+5 組	19/23 (83%)	19/34 (56%)

$$\text{再現率} = \frac{\text{正検出数}}{\text{正答数}}, \quad \text{適合率} = \frac{\text{正検出数}}{\text{検出数}}$$

4 考察

未検出の原因は，実況の省略や口語特有の簡略表現により，抽出された索引語があまり現れなかったためである．また，特殊なケースの touchdown が 1 つ含まれていた．一方で誤検出には，直前のイベントを振り返っているものが多く見られた．いずれの場合も，言語情報のみから正しく検出するのは困難と考えられる．

解決策としては，音情報の協調的な利用が考えられる．歓声やアナウンサの声 [3] などの音特徴も考慮することで，特に適合率の改善が期待される．また**表 1** から明らかなとおり，選択される各特徴量は相関が強い．このような冗長な次元は精度低下の原因となるため，次元削減についても検討する必要がある．

5 まとめ

本稿では，言語ストリームにおける語の頻度特徴を利用し，時区間イベントを検出する手法について述べた．今後は提案手法を改善し，多数の評価実験により手法の有効性を検証する予定である．

なお，本研究の一部は日本学術振興会科学研究費・基盤 (B)(代表:馬場口) の補助による．

参考文献

[1] 馬場口 登, “インターモーダル協調による放送型スポーツ映像の構造化と要約生成”, 人工知能学会研究会資料, SIG-CII-NOV-7, pp.26-34, Nov.2000.

[2] 徳永 健伸, 岩山 真, “重み付き IDF を用いた文書の自動分類について”, 情報処理学会自然言語処理研究会, Vol.100,No.5, pp.33-40, Mar.1994.

[3] Yong Rui, Anoop Gupta, Alex Acero, “Automatically Extracting Highlights for TV Baseball Programs”, Proc. ACM Multimedia 2000, pp.105-116, Oct.2000.