I-046

音響信号処理に基づくサッカー映像のインデクシング手法
Soccer Video Indexing based on Acoustic Signal Processing

塩崎 崇 † **大平 茂輝** ‡ **誉田 雅彰** † **白井 克彦** †
Takashi Shiozaki Shigeki Ohira Masaaki Honda Katsuhiko Shirai

1. はじめに

近年のネットワーク技術の発展と情報の多様化、また専門性の高い放送局の増加やブロードバンドの普及によるデジタル放送の本格化により、映像コンテンツを扱う機会が今後、ますます増加していくと考えられる。映像コンテンツを始めとするマルチメディアコンテンツの増加は我々の選択肢を広げるメリットがある反面、その複雑性のために内容が把握しきれなくなる恐れがある。そのため、音声や画像を含むマルチメディアコンテンツの認識・検索といった高度なマルチメディア情報処理技術が求められてきている。マルチメディアコンテンツは時間的・空間的に複雑であるため、内容理解や、認識が極めて困難である。近年では、スポーツ映像・ニュース映像などの、比較的限定された映像を用いて、内容理解や検索の研究が盛んに行われている。動画像の要約を生成する際、動画像処理によるインデクシングなどは行われてきているが、音声・音響的特徴を用いた研究は少ない[1]-[3]。

本研究では、対象をサッカーの放送映像に限定している。サッカー映像の検索・要約作成には、音声や動画像情報の特徴を解析しインデクスとして利用することが必要となる。そのためにイベントの特定、インデクシングの効率化が求められる。そこで主要な音声・音響情報を用い、インデクシング、処理対象区間の限定といった前処理及び動画像処理の精度向上を図る。またイベント及びシーンの重要度に関しても分析を行い、要約作成の補助を目的とする。

2. 本研究のアプローチ

サッカー映像に対してイベントの特定、インデクシングを行うにあたり、これらの音声情報の中で効果が高いと考えられるのは以下の二つである。

1. 実況音声の言語的情報
2. 歓声等の音響的情報

このうち 1 に関しては、連続 DP マッチングを用いたワードスポッティング法等を用い、サッカーにおいて重要である「ゴール」、「シュート」といった語句を抽出する方法が考えられる。しかしこの手法を用いるにあたり、雑音の問題、また実際ゴールシーンで「ゴール」と、シーン名を言っているかといった問題があるため、本研究では用いない。

そこで本研究では 2 の歓声等の音響的情報に着目し研究を行った。

3. 重要区間の検出

サッカー映像要約における重要区間の定義は、要約の目的によって様々であるが、本研究ではニュース等の放送要約に代表されるような一般性の高い要約を作成するために、音声パワーを用い、観客の盛り上がりに着目した重要区間の検出を試みる。このことによりインデクシングの処理区間の限定といった前処理として利用できると考えられる。

音声パワーに対し分析窓を窓長 5 秒で 0.5 秒ずつシフトさせ、各区間のパワーの平均を求め、閾値を超える区間を重要区間とした。閾値としてパワーの全体平均を与えた。重要区間では盛り上がりが瞬間的ではなく一定区間続くと考え、区間長が短いもの（5[sec] 以下）は入れないことにした。7 試合分のデータに対し重要区間の検出を行ったところ、区間数は 1 試合平均 71 区間、区間長の平均は 20[sec] となった。また、ゴール、シュートがどの程度検出できたかを表 1 に示す。

表 1: 音声パワーによるゴール及びシュートシーンの検出率

項目	抽出数	抽出率
シュート	98/185	53.0 %
ゴール	24/26	92.3 %

表 1 からもわかるように、要約において最も重要と考えれられるゴールシーンに関して、高い精度で検出できており、音声パワーを用いることにより重要区間が検出されていることがわかる。

4. 音声情報を用いたインデクシング

音声パワーにより大まかな盛り上がりの区間が検出できたが、各区間における具体的なイベントを捉えるには至っていない。今回は歓声の音響的特徴である音声パワー、スペクトルに着目して分析を行い、インデクシングを試みた。

4.1 プレイイベントカテゴリの定義

インデクシングを行うにあたり、サッカーのイベントを構成する要素として、基本的なプレイであるシュート、クロス、ドリブル、パスに、ゴールを加えた 5 つをプレイイベントカテゴリとして定義する。

4.2 プレイイベントカテゴリの識別

定義した 5 つのカテゴリについて、カテゴリごとの音声的な特徴を用い、識別を試みた。

既にラベルがふられているデータから、カテゴリごとにデータを集め繋げたデータに対して、窓長 1[sec] の分析窓を 0.1[sec] ずつずらしながら分析、学習した。特徴量として、音声パワーの平均、パワー差分、対数パワースペクトルの重心を用い、識別にはマハラノビス距離を使用した（表 2）。

音響的特徴のみでプレイイベントを分類することは困難であるという予想通り、全体として識別率は低いものとなったが、得点カテゴリとそれ以外のカテゴリとは区

†早稲田大学, Waseda Univercity
‡名古屋大学, Nagoya Univercity

表 2: 識別実験結果

		カテゴリ別					全体
		Goal	Shoot	Cross	Dribble	Pass	
学習データ	Goal	**76.92** %	**23.08** %	0.00 %	0.00 %	0.00 %	
	Shoot	**7.55** %	**49.06** %	18.87 %	11.32 %	13.21 %	47.70 %
	Cross	0.00 %	20.37 %	**31.48** %	**21.30** %	**26.85** %	
	Dribble	0.86 %	3.29 %	**14.66** %	**22.53** %	**58.66** %	（**94.54** %）
	Pass	0.58 %	3.97 %	**11.43** %	**16.99** %	**67.02** %	
学習データ	Goal	**46.15** %	**46.15** %	7.69 %	0.00 %	0.00 %	
(4-fold	Shoot	**5.66** %	**45.28** %	26.42 %	9.43 %	13.21 %	44.21 %
cross-validation)	Cross	0.93 %	25.00 %	**27.78** %	**25.00** %	**21.30** %	
	Dribble	0.79 %	3.58 %	**18.74** %	**24.32** %	**52.58** %	（**93.67** %）
	Pass	0.48 %	4.98 %	**13.98** %	**20.75** %	**59.82** %	
評価データ	Goal	**14.29** %	**66.67** %	14.29 %	4.76 %	0.00 %	
	Shoot	**9.38** %	**34.38** %	25.00 %	15.63 %	15.63 %	42.61 %
	Cross	1.27 %	21.52 %	**39.24** %	**15.19** %	**22.78** %	
	Dribble	0.60 %	3.40 %	**22.30** %	**15.12** %	**58.58** %	（**94.17** %）
	Pass	0.87 %	4.53 %	**19.76** %	**12.92** %	**61.91** %	

() 内の数値は得点カテゴリとそれ以外のカテゴリの識別率を示す。

別できると考えられ、音声パワー、スペクトルを用いることで粗いインデクシングが可能であることが示された。

5. 要約の作成

現状の音声情報のみを用いて得られたインデックスを使用して、どの程度の要約が作成可能か、実際に要約を作成した。

3 章で提案した手法により検出された重要区間に対し、4.2 節で行ったカテゴリの識別を行い、得点カテゴリ（ゴール、シュート）であると識別された区間にイベントがあると考え、その前後区間を抽出した。また、ゴールまたは得点カテゴリであると識別された区間が多い順に要約に加えた。このことにより重要区間の長さや、イベント区間の音声パワー等が重要度として要約に反映されると考えられる。試合開始直後は重要なイベントが無くても盛り上がりが続く傾向にあるので除外した。1 シーン 15 秒とし、10 シーンからなる要約を作成した（図 1）。

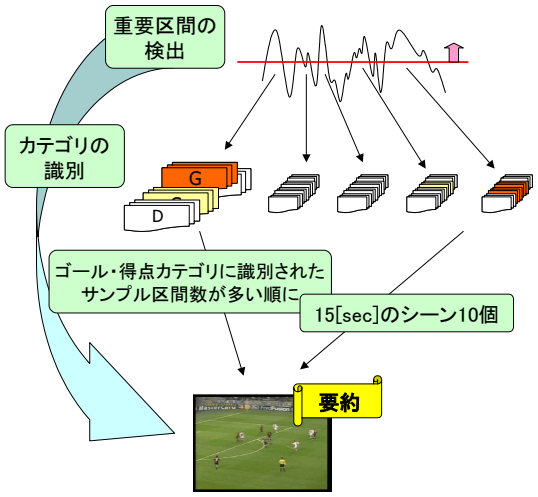

図 1: 要約作成手順

結果として、学習データに対しては 4 ゴール中 4 ゴール、評価データに対しては 4 ゴール中 3 ゴールと、一般性の高い要約を作成する際、最も重要だと考えられるゴールシーンについては高い精度で検出できた。ゴール以外で抽出したシーンとして、シュートシーン、ゴール前シーン、ファウルシーンなど重要シーンが検出された。このように、音声情報のみから得られたインデックスを用いることで、放送されているような一般性の高い要約に近いものが作成可能となった。

6. まとめ

歓声の音響情報を用いることで、観客の盛り上がりを捉えることができ、ゴールやシュートに分類される得点に絡むプレイカテゴリとそれ以外のカテゴリという粗いインデックスが得られる見通しがついた。

このことにより、動画像処理を用いたインデクシングの効率化に大きな効果が期待でき、放送されているような一般性の高い要約に近いものが作成可能となった。

今後は盛り上がりの立ち上がりや継続長などの時間的特徴量の検討、オプティカルフローによるカメラワークの検出結果を用いた補正処理、またシーン単位での分析を進める予定である。

参考文献

[1] 小河 誠巳, 相澤 清晴 : “音声情報を用いたイベント検出による映像要約”, 情報科学技術フォーラム,FIT2002,I-69,pp.137-138,2002-7.

[2] 重森 猛, 金子 剛志, 緒方 淳, 藤本 雅清, 有木 康雄, 塚田 清志, 濱口 伸, 清瀬 基 : “音響・言語適応処理を用いたスポーツ実況中継音声の認識 ～ハイライトシーン検出への応用～”, 信学技報, SP2002-166,pp.33-40,2003-1.

[3] 須場 康貴, 浜田 玲子, 井手 一郎, 坂井 修一, 田中 英彦 : “料理映像の索引付けのための音響解析手法の検討”, 第 64 回情報処理学会全国大会,No.2L-4,Vol.2,pp.17-18,2002-3.