インターモーダル協調の薦め
－マルチモダリティに着目したメディアのセマンティック解析－

馬場口　登*　（大阪大学）

Why Do We Need Intermodal Collaboration? -Semantic Analysis of Media Focusing on Multimodality-

Noboru Babaguchi (Osaka University)

１．まえがき

筆者は，映像などのマルチメディアデータから，インデクシングやアノテーションに有用なセマンティック情報を自動獲得するために，**インターモーダル協調**[1](intermodal collaboration)なるマルチメディア（動画，音声，テキスト，図形など）解析戦略の有効性を指摘し，種々の応用を試みてきた．本稿では，放送映像，サーベイランス映像，ミーティング映像に対する応用例を示すと共に，今後の展開を探っていく．

２．インターモーダル協調

映像は，動画，音声，テキストなど，マルチモダリティを有する時系列データの集合体である．映像のコンテンツに基づく検索や要約，編集などの高度なアプリケーションを実現するには，映像のセマンティックスの考慮が不可欠である．しかしながら，信号レベルの映像データと記号レベルの概念との間には，**セマンティックギャップ**と呼ばれる障壁が立ちはだかっている．セマンティックギャップの克服に全世界の研究者がしのぎを削っているのが，マルチメディア・コンテンツ処理研究の趨勢である．

元来，映像の中で圧倒的なデータ量を占める動画に対して種々の解析手法が提案されてきたが，完璧にセマンティックスを抽出するにはパターン認識技術は未だ未成熟である．そこで，動画とは情報組成を異にする音声やテキストの同期性を考慮してセマンティック解析を試みようとする手法，すなわちインターモーダル協調が増えてきた．音声やテキストには本質的にセマンティックスを表現した記号がメディアに埋没しているため，これを動画に対応付けようという考え方である．マルチモダリティを考慮することによって，信頼性やロバスト性の向上も期待される．

３．放送映像への応用

３．１　動画，テキストの利用

Babaguchi ら[2]はインターモーダル協調の具体化として，スポーツ映像における**イベントによるインデクシング**を議論した．その手法は，音声のトランスクリプトであるクローズドキャプション（CC）と動画の協調解析によってホームランやタッチダウンなどの各スポーツ固有のイベントを検出し，イベント名をインデクスとして付与するものである．CC ストリームから目標イベントに関するキーワード列の抽出を行い，イベントが出現する可能性の高い時区間を求める．続いて，その時区間内のショット（連続性を有する一連の画像フレーム列）に対し，色情報の時間軸上での類似性を基にショットにインデクスを付ける．アメリカンフットボールの TV 中継における，得点イベント（イベント総数 40 個）に対するインデクシング結果は再現率 81%，適合率 74%であった．インターモーダル協調により，CC のみを用いた場合に比べ適合率に関して 25% の精度向上が見られた．尚，得点イベントによるインデクシングを応用して，映像要約法を提案している[3].

一方，Nitta ら[4]は上と同様の考え方で，スポーツ映像の中の各ライブシーンに対し**選手とプレイによるインデクシング**を試みた．アメリカンフットボール映像に対して実験を行った結果，再現率 86%，適合率 95%で正確なインデクシングが可能となった．さらに，スポーツ映像を 1 プレイに相当する**ストーリーユニット**に分割し，その意味内容を含む CC 部分からテキスト形式の**アノテーション**を生成する手法を提案した[5]．ベイジアンネットワークを用い，CC セグメントを 4 種類のシーンに分類して CC ストーリーユニットを発見する．次に，動画特徴に基づき映像ストーリーユニットを抽出し，これを CC ストーリーユニットと時間的に対応付け，最終的にアノテーションを生成するもので良好な結果が報告されている．

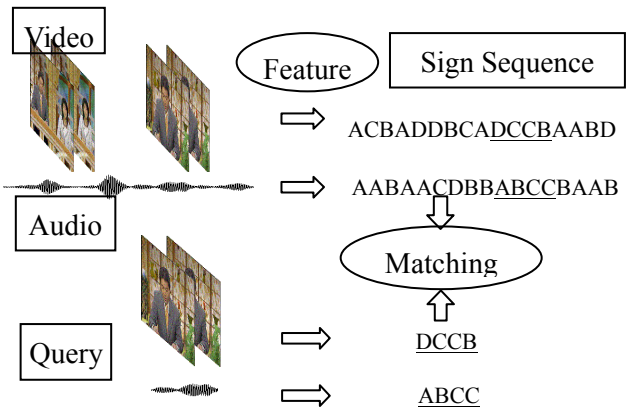

Fig.1　記号列マッチングによる類似シーン検索

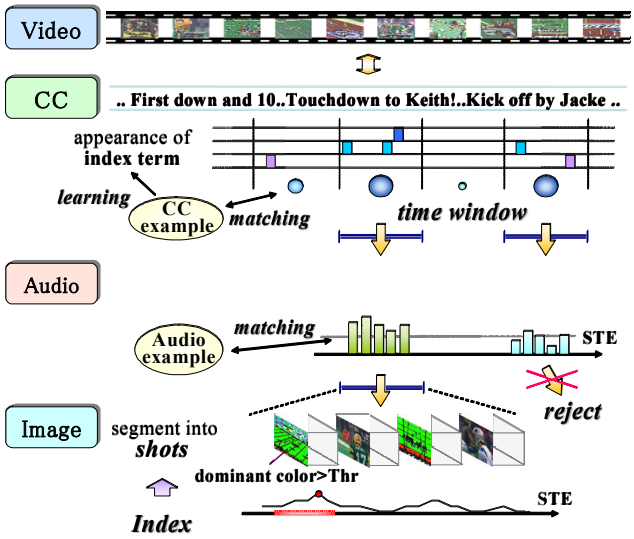

Fig.2　インターモーダル協調によるハイライト検出

３．２　動画，音声の利用

筆者ら[6]は，比較的長い時間のクエリを許容できる，**類似シーン検索**の実現に動画と音声の協調処理を検討した．動画と音声を各々固定長のセグメント（パケットと呼ぶ）に分割し，パケットごとに特徴抽出を行う．動画パケットでは動きと色の特徴を，音声パケットでは音源分類に有用な特徴を用いる．これらの特徴ベクトルを特徴空間でクラスタリングし，クラスタを表現する記号に変換し，パケット列を記号列に変換する．この記号列はパケットの性質を表すもので，あたかも DNA 配列のようなデータとなる．クエリも同様に記号列に変換し，記号列間で DP マッチングを施すことにより時間軸上の類似性を検出し，クエリに類似したシーンを検索する．Fig.1 にこの様子を示す．動画や音声単独で検索する場合より，両者を併用する場合に良い結果が得られることを実証した．

３．３　動画，音声，テキストの利用

宮内ら[7]は先行手法[2]のドメインやヒューリスティックに依存する面を克服するためにキーワードの学習機構および音声解析を導入して，**ハイライトによるインデクシング**の拡張を試みた．ハイライトシーンで音声パワーが大きくなることに着目し，音声特徴として短時間エネルギーを用いて CC ストリームの解析結果を検証することが改良点である．Fig.2 に提案手法の概要を示す．アメリカンフットボール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できた．

４．サーベイランス映像への応用

我々は，サーベイランス映像に対し，映像と音響の同期的観測系および解析系を設計し，**協調的イベント検出**を実現した[8]．サーベイランス映像が放送映像と異なる点は，定点観測のためシーン変化がないということであり，コンテンツに立脚したイベント抽出が求められる．イベントとして，環境内の部分エリアへの入出，エリア間移動，エリア内停留を想定して，人物の追跡及び環境モデルとのマッチングによるイベント同定法を具体化し，特に環境内への入出に関して，ドアの開閉音などの環境音に着目している．

タイムラインと空間マップのマーキングによるサーベイランス映像可視化システムのプロトタイプを試作し，90%程度のイベント抽出精度が得られること，メディア時間の約半分で処理が終了することを実験的に確かめた．

５．ミーティング映像への応用

ミーティング映像を対象に，**マルチメディア・ログ**の自動作成を図るシステムを現在，構築中である．複数の人間の円卓形式ミーティング風景を，忠実に記録し，利用するというニーズは潜在的に多い．特に，発言者の正面顔を記録したいという場合には，人物の背後にあるカメラからはオクリュージョンが生じ，うまくいかない可能性が高い．そこで，筆者らは単一の全方位カメラとアレーマイクを用い，可搬性のセンサを検討している．発言者に対して発話の音響的トリガーによりセンサからの方位を計測し，その方位の透視投影画像を記録することにより正面顔の方向の動画並びに同期した音声を保存する．

６．おわりに

マルチモダリティに着目する重要性は，我々人間の情報獲得やコミュニケーションの営みを観察すれば，容易に認識できよう．つまり，ロバスト性や精度の向上を目指すなら，マルチモダリティに頼る以外の手立てはないとも言える．筆者らのこれまでの研究は，映像のセマンティックスに立脚したインデクシング，アノテーション，さらには要約にマルチモダリティを重視した処理戦略の有効性を示唆するものである．最近開催された ICME2004 では，セマンティックギャップの解決に，マルチモーダル解析と学習（SVM など）を強調する一方，音響の有効利用を図る試みが目立っていた．マルチモダリティに着目すると処理コストは高くなる傾向があるが，人間は極めて巧妙に並列処理しているものと思われ，効率のよいマルチモーダル処理パラダイムの開発が今後重要となるであろう．

文　献

(1) N. Babaguchi, et al: Proc. ICIP2003, 2003
(2) N.Babaguchi, et al: IEEE Trans. Multimedia, Vol.4, No.1, pp.68-75, 2002.
(3) N. Babaguchi, et al:　IEEE Trans. Multimedia, Vol.6, No.4, 2004.
(4) N. Nitta, et al: Multimedia Tools and Applications (to be published).
(5) 新田，他：電子情報通信学会論文誌(D-II), Vol.J86-D-II, No.8, pp.1222-1233, 2003.
(6) N.Babaguchi, et al: Proc. ICME2004, 2004.
(7) 宮内，他：電子情報通信学会論文誌(D-II), Vol.J85-D-II, No.11, pp.1692- 1700, 2002.
(8) N. Babaguchi, et al:　Proc. MIS2002, pp.18-27, 2002.