H-081

ニュース映像における話者と被写体の不一致検出

Detection of Mismatching between Speaker and Face in News Videos

小林 尊志† 高橋 友和† 井手 一郎†‡ 村瀬 洋†

Takashi KOBAYASHI Tomokazu TAKAHASHI Ichiro IDE Hiroshi MURASE

1 はじめに

近年放送メディアをめぐる環境の多様化に伴い，放送映像の解析技術に関する研究に注目が集まっている．中でもニュース映像については資料性の点から重要視されており，それに関する研究として，政治家やスポーツ選手のインタビュシーンなど番組関係者以外による“発言シーン”を自動検出する手法 [1] が提案されている．この手法では，画像情報・音声情報・クローズドキャプション (CC) テキスト情報を組み合わせて用いることで，番組関係者による映像区間を除いた発言シーンを自動検出する．しかし，ニュース映像においては画像中の人物とは別に番組関係者による音声がナレーションとして重畳して記録されている映像区間があり，これらの区間の誤検出により検出精度が低下するため，画像と音声の一致を確認する必要がある．画像と音声を同期する手法として，両者の相互情報量を考慮して映像区間を切り出す手法 [2] や監視カメラ映像について画像と音声それぞれに対して前景・背景モデリングを適用することでイベントを認識する手法 [3] が提案されている．本発表では，先行研究によって検出された映像区間を発言シーン候補区間として，画像情報と音声情報の統合的なメディア情報処理を行うことにより，番組関係者による音声区間を除去して候補区間を絞り込む手法を提案する．また，実際に放送されたニュース映像に対して本手法を適用し，有効性の検討を行った．

2 先行研究

先行研究として Ide らの手法 [1] を紹介する．まず，ニュース映像から顔領域を含む発言シーン候補区間を検出する．次にそれらの区間においてクローズドキャプションテキストと音声情報を用いて番組関係者の発話モデルを作成する．その後，発話モデルとの照合によりキャスタショットを除外することで絞り込みを行う．最後に，クローズドキャプションテキスト情報を用いてレポータシーンの前後に含まれる単語の特徴からレポータショットを除去して非番組関係者の発言シーンを検出する．しかし，この手法では画像情報と音声情報を別々に処理するため，画像中の人物とは別の番組関係者によるナレーションが記録されている区間が人物の発言シーンとして誤検出することが多い．

3 提案手法

音声として記録されている人物を“話者”，画像として記録されている人物を“被写体”とする．話者と被写体が一致する区間については，画像内の人物の口唇領域の変化が激しい区間について音声パワーも大きく変化し，不一致区間についてはそれらの間の相関は弱いと考えられる．そこで本研究では，先行研究による発言シーン候補について画像と音声の時系列的相関の値により話者と被写体の一致区間・不一致区間を判断することにより，精度の向上を図る．図 1 に処理の流れを示し，以降で詳細を述べる．

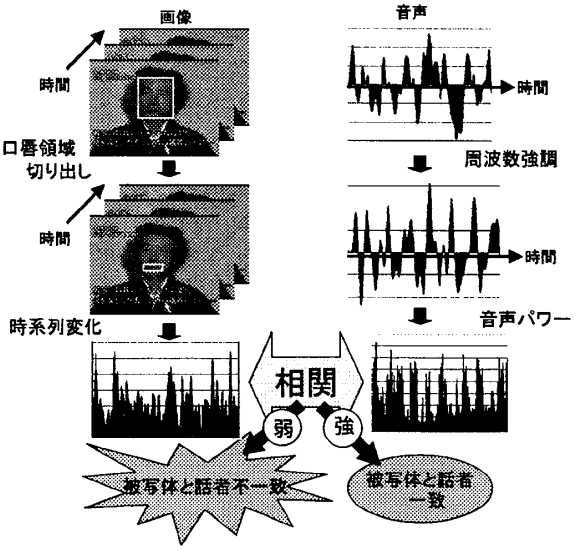

図 1 話者と被写体の不一致検出処理の流れ

3.1 画像特徴の時系列変化

発言シーン候補区間から口唇領域を矩形領域として切り出し，矩形領域の縦画素数の変化量の絶対値 I_1

$$I_1(t) = |\, v(t) - v(t-1) \,|. \qquad (1)$$

と，縦横比の変化量の絶対値 I_2

$$I_2(t) = \left| \frac{v(t)}{h(t)} - \frac{v(t-1)}{h(t-1)} \right|. \qquad (2)$$

を抽出する．ここで t はフレーム番号，$v(t)$ は矩形領域の縦方向の画素数，$h(t)$ は矩形領域の横方向の画素数を表す．

3.2 音声特徴の時系列変化

画像のフレームレートに合わせるように再サンプリングしながら変換窓 T 点について高速フーリエ変換を適用して音声パワーを取り出す．音声特徴の表現方法として，音声パワー A_1 と音声パワーの対数 A_2 をとったものの 2 種類を用いる．

† 名古屋大学 大学院情報科学研究科

‡ 国立情報学研究所

3.3 相関の計算

特徴間の相関は次式で定義される相互相関係数を用いる．

$$C(I,A)=\frac{\sum_t \left(I(t)-\hat{I}\right)\left(A(t)-\hat{A}\right)}{\sqrt{\sum_t \left(I(t)-\hat{I}\right)^2}\sqrt{\sum_t \left(A(t)-\hat{A}\right)^2}}. \quad (3)$$

ここで $\hat{I}$, $\hat{A}$ はそれぞれ映像区間における画像特徴と音声特徴の平均値を表す．ただし音声と画像の時系列変化の間にフレーム単位で同期が取れているとは必ずしも言えないため，前後に F フレームずつずらしながら相関係数を計算し，最も大きい値を波形間の相関値とする．

4 実験

4.1 実験条件

2004 年 1 月に放送されたニュース番組「NHK ニュース 7」の映像から，Ide らの手法 [1] により話者と被写体の一致 50 区間・不一致 50 区間を取り出し，用いる特徴の組み合わせを検討した．これらの 100 映像区間について画像特徴と音声特徴の間の相関値を計算し，全体の相関の平均値を閾値として映像区間の一致・不一致の判別を行い，判別率を求めた．映像の仕様については表 1 の通りである．

表 1 実験に用いた映像の仕様

フレームレート	30fps	オーディオ形式	PCM
圧縮形式	MPEG-1	サンプリング	48 kHz
区間長	1.5〜15 秒	量子化数	16 bit
解像度	352 × 240 pixel	チャンネル	モノラル

また実験におけるパラメータは (表 2) の通りである．

表 2 実験に用いたパラメータ

窓関数	ハミング窓関数
フーリエ変換窓幅 T	1,024 点
ずらし幅 F	7 frame

画像特徴 $I_1(t)$ と $I_2(t)$, 音声特徴 A_1 と A_2 の組み合わせを変えたとき判別率は表 3 のようになった．本実験では人手により口唇領域を切り出した．

表 3 特徴の組み合わせによる判別率

画像特徴＼音声特徴	$A_1(t)$	$A_2(t)$
$I_1(t)$	58.0 %	61.0 %
$I_2(t)$	59.0 %	57.0 %

4.2 実験結果

表 3 より，$I_1(t)$ と音声特徴 A_2 において最も高い判別率 61.0 %が得られた．誤判別された映像区間を見ると，顔領域が横や下を向いていて (図 2) 正しく画像特徴を抽出できなかったものや，現場音によるノイズの影響が大きいため音声特徴が抽出しづらい映像区間が多かった．様々な撮影環境と豊富なカメラワークが存在するニュース映像については，より有効な特徴量やその抽出方法を検討する必要がある．

図 2 失敗した検出例

5 まとめ

本発表では，話者と被写体の不一致区間の検出により発言シーン検出精度を向上させる手法を提案した．実験の結果，様々な撮影環境が想定されるニュース映像への適用については課題があるものの，ある程度の有効性を確認した．また本発表中の実験では人手により口唇領域を切り出したが，口唇領域の切り出しと追跡を自動で行う必要がある．今後はより有効な特徴量の検討，およびノイズ除去について検討していきたい．

謝辞

本研究の一部は科学研究費補助金による．また，本発表の実験では映像情報処理に MIST ライブラリ (http://mist.s.m.is.nagoya-u.ac.jp/trac/) を使用した．

参考文献

[1] Ichiro Ide, Naoki Sekioka, Tomokazu Takahashi, Hiroshi Murase, "Assembling personal speech collections by monologue scene detection from a news video archive", Proc. Eigth ACM SIGMM Intl. Workshop on Multimedia Information Retrieval (MIR2006), pp.223-229, Oct. 2006.

[2] Trevor Darrell, John W.Fisher III, Paul Viola, William Freeman, "Audio-visual segmentation and "The Cocktail Party Effect"", ICMI 2000, Third International Conference Proc, pp.32-40, Oct. 2000

[3] Marco Chritani, Manuele Bicego, Vittorio Murino, "Audio-visual event recognition in surveillance video sequences" ,IEEE Transaction on Multimedia, Vol.9, No.2, pp.257-267, Feb. 2007