ニュース発言集作成のためのモノローグシーンの検出

Detecting Monologue Scenes in News Videos for Making a News Speech Archive

関岡直城[†] 高橋友和[†] 井手一郎[†‡] 村瀬洋[†]
Naoki Sekioka Tomokazu Takahashi Ichiro Ide Hiroshi Murase

1 はじめに

近年，大容量 HDD の普及により映像を大量に蓄積して利用する機会が増えつつある．これに伴い，大規模なニュース映像群からユーザの所望するシーンを自動的に抽出して提示する技術が求められている．そこで本研究では，ニュース映像から演説やインタビューなどのモノローグシーンを自動収集することにより，登場人物の発言集を作成する．関心のある人物の過去数年間の発言を瞬時に閲覧したり，番組製作者側の編集作業を支援するなど様々な用途への利用を考えている．本稿では，その前処理段階として，画像・音声・テキストを用いた統合メディア処理によるモノローグシーン候補の自動検出について述べる．

2 従来研究

モノローグシーンの検出は，TREC Video 2003[1] の高次特徴抽出タスクで「News subject monologue」として課題になるなど，注目されている．[2] では，Video OCR を映像中のキャプションに適用することにより人名を検出し，CNN，ABC 各放送局の人名データベースと照合することで，モノローグシーンを検出している．また [3] では，ショット中の各フレームから抽出した音声特徴ベクトルを 10 個のクラスタに分類し，そのクラスタ間の平均，中央値，標準偏差により，モノローグショットと会話ショットとを識別している．しかし，TREC Video で対象としている CNN や ABC といった海外のニュース映像と日本のニュース映像では大きな違いがある．それは，前者には放送局から提供されるクローズドキャプションと呼ばれる音声の書き下しテキスト (日本の文字放送と字幕に相当) に，番組関係者とそれ以外の人物の発言を区別する情報が含まれていることである．[4] では，この特徴を利用し，画像情報とテキスト情報の対応付けによるセマンティックな映像セグメントの検出手法を提案している．

これらの従来研究に対し，本研究では，日本のニュース映像 (NHK ニュース 7) を対象としたモノローグシーンの検出を試みる．

3 モノローグシーンの検出

番組関係者 (キャスタやレポータなど) 以外の人物による演説やインタビューなどのモノローグシーンを検出する手法について述べる．

3.1 予備実験：画像情報による検出

モノローグシーンを検出するうえで，顔領域の検出は最も重要な手がかりとなる．そこで，RGB カラーヒストグラムによるカット検出後，各シーンに対して，[5] のオブジェクト検出手法を用いて顔領域を検出し，モノローグシーンを取得する．なお，顔の大きさが画像全体の 8%以上の正面顔かつ，フレームの中央付近 (図 1 の灰色領域) に顔領域の重心が含まれる場合を検出条件とした．表 1 に画像情報のみを利用した検出結果を示す．全体の適合率は 37%，再現率は 79%という結果が得られた．検出漏れはどのサンプルにおいても少数で，その主な原因は，顔の向きと帽子などによるオクルージョンが挙げられる．これに対し検出誤りが数多く見られたが，その原因として以下の 2 つが挙げられる．

- キャスタシーンやレポータシーンの誤検出
- 映像中の人物と音声の不一致

前者は，キャスタやレポータの顔のクローズアップシーンを誤って検出してしまった場合，後者は，キャスタなどが映像中の人物に関する情報や，その発言内容を間接的に言及する場合であり，どちらもニュース映像に頻繁に見られるシーンである．

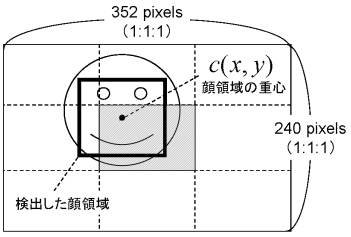

図 1 顔領域の検出条件 (位置)

表 1 モノローグシーンの検出結果 (画像情報のみ)

	サンプル No.				
	1	2	3	4	平均
適合率 (%)	26	38	67	16	37
再現率 (%)	83	71	86	75	79

3.2 統合メディア処理による検出精度の向上

画像情報のみを用いた上記の予備実験で得られた結果に対し，ここでは，統合メディア処理によるモノローグシーンの検出手法について述べる．

画像情報のみを利用した検出における誤検出を低減させるためには，画像情報に加え音声情報を利用する必要がある．ここでは，番組関係者 (キャスタ，レポータ) の話者モデルをそれぞれ作成し，3.1 で検出したモノローグシーンに対応する音声情報とのモデル照合を行う．こ

[†]名古屋大学大学院 情報科学研究科
[‡]国立情報学研究所

れにより，誤って検出されたキャスタシーンやレポータシーン，映像中の人物と音声が一致しないシーンを除去できると考える．

3.2.1 音声情報による同一人物の発話区間検出の検討

ここでは，モデル照合によるキャスタの発話区間の検出実験により，話者モデルの性能を評価する．話者モデルを作成する際のパラメータ等の詳細を表 2 に示す．なお，話者モデル作成のための学習サンプルは，入力映像から人手で約 20 秒の音声データを切り出して用いた．また，モデル照合に用いる類似度 (ベクトル量子化歪) の判定閾値は，適合率が優先されるように経験的に設定した．実験結果を表 3 に示す．これからも分かるように，画像情報を補助的に併用する分には十分な精度が得られたと考える．

表 2　話者モデル作成の詳細

フレーム長	20(ms)
フレーム間隔	10(ms)
音声特徴量	25 次 LPC ケプストラム係数
モデル作成手法	ベクトル量子化

表 3　キャスタ発話区間検出の実験結果

	サンプル No.				
	1	2	3	4	平均
適合率 (%)	98	92	95	92	94
再現率 (%)	89	81	90	88	87

3.2.2 テキスト情報を用いた話者モデル作成の自動化

本研究では，モノローグシーン候補の自動検出を目的としている．そのためには，話者モデルを作成する際の学習サンプルの収集も自動化する必要がある．そこで，クローズドキャプションに記述された発話時刻を手がかりに，各番組関係者の学習サンプルを入力映像から自動取得する．この発話時刻を取得するためには，各番組関係者の発話する周辺の文を検出する必要がある．キャスタは，ニュース映像の放送開始後，最初に発話する人物と考えられるので，クローズドキャプションの冒頭に記述された文の発話時刻を取得すればよい．一方，レポータの発話時刻の取得には，キーワード検索と形態素解析を行い，発話直前の文を検出する．20 日分のニュース映像において，レポータが発話する直前の文に含まれる語の出現頻度を表 4 に示す．ここでは，この上位 3 つの語，“記者”，“取材”，“中継” を検索の際のキーワードとして採用する．次に，以下の 2 つの文法的な条件を満たす文を形態素解析により判定する．

条件 1. 文の語尾が「固有名詞＋ “さん”」

条件 2. 文が過去形でない

条件 1 は，キャスタによるレポータへの呼びかけであり，この条件 1 を満たす場合には，無条件にレポータの発話直前の文として採用する．また，条件 2 を満たす場合には，上記の 3 つのキーワードのいずれかが含まれていれば採用する．

3.2.3 モノローグシーン検出の自動化

音声情報の話者モデル作成の自動化により，本研究の目的であるモノローグシーンの検出も完全自動化できる．

表 4　語の出現頻度

語	“記者”	“取材”	“中継”	“その他”
出現頻度	31(3)	9(0)	7(3)	4(1)

( ) 内の数値はレポータの発話直前の文以外の箇所で検出された数

4　実験と考察

画像・音声・テキスト情報を用いた統合メディア処理により，モノローグシーンの検出実験を行った．結果を表 5 に示す．全体の適合率は 77%，再現率は 79%という結果が得られた．画像情報のみを用いたモノローグシーンの検出結果 (表 1) と比べ，誤検出の数が大幅に低減されていることが分かる．除去できなかった誤検出の大半はレポータシーンであったが，その原因は，突発的なレポータの発話によりクローズドキャプションにテキスト特徴が表れず，話者モデル作成の際に発話時刻が取得できなかったためと考えられる．この突発的な発話は，主に録画中継に頻繁に見られる特徴で，テキスト情報から判断するのは困難であるが，もうひとつの特徴として，比較的発話時間が長いことが挙げられる．そのため，話者セグメンテーションにより，各話者の発話時間が得られれば，この特徴を用いて除去できるのではないかと考えている．また，検出漏れの原因は，3.1 で述べたように，顔領域の検出精度にあるため，様々な向きへの対応などの改善が今後必要である．

表 5　モノローグシーンの検出結果 (メディア統合)

	サンプル No.				
	1	2	3	4	平均
適合率 (%)	71	71	92	75	77
再現率 (%)	83	71	86	75	79

5　おわりに

本稿では，ニュース発言集の作成のための前処理段階として，統合メディア処理によるモノローグシーン候補の自動検出について報告した．複数のメディアを用いることによる検出精度の向上は確認できたものの，モノローグシーンの候補としての検出精度は，まだ十分ではなく，検出漏れを最低限抑えることが今後の目標となる．

参考文献

[1] http://www-nlpir.nist.gov/projects/tv2003/

[2] A. Hauptmann, R.V. Baron, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, N. Moraveji, N. Papemick, C.G.M. Snoek, G. Tzanetakis, J. Yang, R. Yan, and H.D. Wactlar, “Informedia at TRECVID 2003:Analyzing and Searching Broadcast News Video”, Online Proc. TRECVID 2003.

[3] L. Wu, Y. Guo, X. Qiu, Z. Feng, J. Rong, W. Jin, D. Zhou, R. Wang, M. Jin, “Fudan University at TRECVID 2003”, Online Proc. TRECVID 2003.

[4] Y. Nakamura, T. Kanade, “Semantic Analysis for Video Contents Extraction － Spotting by Association in News Video”, Proc. ACM Multimedia ’97, pp.393-401, 1997.

[5] A. Kuranov, R. Lienhart, and V. Pisarevsky, “ An Empirical Analysis of Boosting Algorithms for Rapid Objects with an Extended Set of Haar-Like Features ”, Intel Tech. Rep. MRL-TR-July02-01, 2002.