ニュース映像における発言シーン検出手法の提案

The proposal of a method for detecting speech scenes in news videos

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

† 名古屋大学大学院 情報科学研究科
Graduate School of Information Science, Nagoya University

‡ 国立情報学研究所
National Institute of Informatics

1 はじめに

ニュース映像は，社会情勢からスポーツにわたる幅広い分野の情報を提供するだけでなく，映像による情報提示は受け手の視覚にうったえ，直感的な理解を容易にする．特に映像中の登場人物が発言しているシーン (発言シーン) は，発言内容はもちろんのこと，話者の表情や態度など活字では表現しきれない情報を含む．そこで本研究では，番組関係者 (アナウンサ，レポータなど) 以外の登場人物による発言シーンの自動検出手法を提案する．抽出した発言シーンは，映像の要約や検索など様々な用途に利用することを考えている．

2 発言シーンの検出

2.1 処理手順の概要

一般的に，特定人物による発話を検出するためには，事前にその人物の音声特徴を学習し，音声照合すれば良い．しかし，ニュース映像においては，日々刻々と新しい話題が提供され，登場人物も未知である場合が多いため，学習サンプルを網羅的に収集する上記の手法は現実的ではない．そこで，本研究ではニュース映像特有のシーン構造を利用して，アナウンサの発話区間を除去する消去法により，発言シーンを検出する (図 1 参照)．

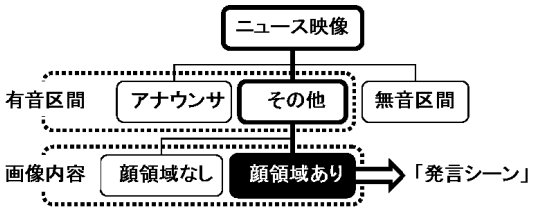

図 1 シーン構造と発言シーンの検出過程

2.2 アナウンサ発話区間の除去と顔領域の検出

はじめに，0.5 秒毎の FFT パワー平均により有音区間を検出する．次に，画像を手がかりにアナウンサショットを検出し，それらのショットに対応する音声データから得られる 25 次の LPC ケプストラム係数を長時間平均することにより，アナウンサ発話モデルを作成する．そして，検出した有音区間に対して発話モデルとの比較を行い，アナウンサ発話区間を検出する．最後に，検出したアナウンサ発話区間を全有音区間から除去し，残ったその他の有音区間に対して顔領域を検出することで発言シーンを得る．なお，顔領域の検出には，[1] のオブジェクト検出手法を用いた．

3 実験と考察

「NHK ニュース 7」(約 30 分間) を 5 日分用いて発言シーンの検出を試みた．画像を手がかりとしたアナウンサショット検出は，先行研究 [2] などで高精度に実現されている．本研究では，各日ごとに 10 秒程度のアナウンサショットを人手で切り出してアナウンサ発話モデルを作成するものの，それ以外は自動検出した．顔領域は占有面積が全画像に対して 7.6 %以上の正面顔を検出条件とした．その結果，5 日分のニュース映像に対するアナウンサ発話区間を，適合率 90.0 %，再現率 81.5 %と比較的高い精度で検出することができた．

次に，検出したアナウンサ発話区間を除去した有音区間から発言シーンを検出した．検出性能を表 1 に示す．結果は，現時点においてまだ十分な精度は得られていない．これは，アナウンサ発話区間が完全に除去できていないことの他に，音楽・環境音などの有音非発話区間や，アナウンサ以外の番組関係者の発話区間が除ききれていないためであり，今後このような区間の検出・除去が必要となる．

表 1 発言シーンの検出性能

検出率	サンプル No.					平均
	1	2	3	4	5	
適合率 (%)	50	50	39	20	43	40
再現率 (%)	71	64	58	33	75	60

4 おわりに

本稿では，ニュース映像特有のシーン構造を利用した発言シーンの検出手法を提案した．結果として，アナウンサ発話区間は比較的高い精度で検出できたが，発言シーンの検出についてはまだ十分な精度は得られていない．今後は，クローズドキャプションなどの利用や，顔領域の検出において唇の動きの有無などを検出条件に加えることにより，さらなる改善を目指す．

謝辞

本研究の一部は、文部科学省科学研究費補助金および 21 世紀 COE 研究費による。

参考文献

[1] Alexander Kuranov, Rainer Lienhart, and Vadim 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.

[2] 井手一郎, 山本晃司, 浜田玲子, 田中英彦, “ ショット分類に基づく映像への自動的索引付け手法 ”, 信学論 (D-II), vol.J82-D-II, no.10, pp.1543-1551, Oct. 1999.