平成21年度 情報工学コース卒業研究報告要旨

村瀬 研究室	氏　　名	熊 谷　章 吾
卒業研究題目	モノローグシーン抽出のための 被写体と話者の不一致検出	

本研究では，モノローグシーン抽出のための被写体と話者の不一致検出手法を提案する．

近年，インターネット上での映像配信サービスの普及に伴い閲覧可能な映像が増大し，重要なシーンやユーザの所望するシーンの検索技術が求められている．本研究ではニュース映像中の重要なシーンである記者会見や選挙演説，インタビューなど，被写体自身が肉声で発話しているシーン（図 1）に注目する．本研究ではこのようなシーンをモノローグシーンと呼ぶ．モノローグシーンには，話者の表情や態度，声のトーンなどの情報が豊富に含まれているため，発言集や要約映像の生成などの支援において重要な役割を果たすと考えられる．

モノローグシーンでは，映像中の顔領域が中央付近に大きく映ることが多いため，抽出の際には映像中の顔領域の位置と面積の情報が利用できる．しかし，顔領域が中央付近に大きく映っていても，実際に流れている音声はアナウンサのものであるといった，被写体と話者が一致していないシーン（図 2）も存在する．そのため，モノローグシーンの抽出には画像情報と音声情報の統合利用が必要である．小林らは，被写体の口唇動作と話者の音声波形の共起性に着目したモノローグシーンの抽出手法を提案した [1]．彼らは，画像特徴として口唇領域の縦横比，音声特徴として音声信号の平均パワーに注目し，それらの相関に基づく被写体と話者の一致・不一致の識別を試みた．しかし，口唇領域の縦横比および音声信号の平均パワーのみでは，被写体と話者の一致・不一致を識別することは難しく，識別精度は不十分であった．

そこで本研究では，複数の画像特徴と音声特徴の相関に着目した手法を提案する．本手法では，画像特徴として口唇領域の縦横比と口内領域の面積およびそれらの変化量，音声特徴として音声信号のスペクトル包絡を表す線形予測係数と平均パワーおよびそれらの変化量を利用する．これら画像特徴と音声特徴の組み合わせの相関値により表現される特徴空間を用い，被写体と話者の一致・不一致の識別を行う．これにより，不一致区間の高精度な抽出を目指す．

本手法の有効性を確認するため，小林らの手法に基づく比較手法と提案手法とで一致・不一致の識別精度を実験により評価した．室内で撮影した 10 名分の映像と音声から作成した合計約 3,000 秒の評価用セットを用いた．その結果，5 秒間の映像を入力とした場合は，比較手法では 62.5 %に対し提案手法では 68.1 %の識別精度が得られた．これにより，提案手法の有効性が確認できた．

図 1: 被写体自身が発話しているシーン

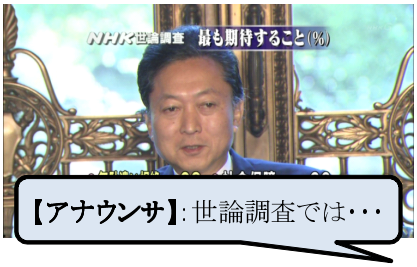

図 2: 被写体と話者が一致していないシーン

[1] 小林ら, “ニュース映像における話者と被写体の不一致検出”, FIT2007.