固定カメラ映像からの音声情報を用いた映像コンテンツ生成

Image Content Gneration Using Voice Information from Fixed Camera

足立順　滝口哲也　有木康雄
Jun Adachi　Tetsuya Takiguchi　Yasuo Ariki

神戸大学大学院自然科学研究科
Kobe University Graduate School of Science and Technology

1　まえがき

コンピュータの小型化，記憶デバイスの大容量化により個人の行動記録の入手が簡単になっている．それに伴い，日常生活やパーティ等での映像を常時記録しそれらの冗長な映像から映像編集を行う研究が盛んになっている．これまでの多くの研究では映像処理による編集がなされてきたが，本研究では音声情報を元に，得られた映像から人物の発話区間を抜き出し，発話方向を推定する事で発話者を特定し，それらの情報を用いて興味深い映像コンテンツを作成し提示することを目標とする．

2　映像編集手法

これまで盛んに行われてきた映像編集の研究は画像情報に基づくものが多数であった．しかしその場合，比較的動きの少ない会話部分が不必要な区間として失われてしまう恐れがある．そこで本研究では音声データを用いて人間の発話部分を抽出し，さらに発話方向を推定することによって発話者を特定し，その情報を元に映像を編集し，ユーザーに提示する．

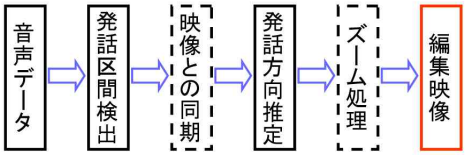

図 1　提案手法における処理の流れ

本システムにおける処理の流れを図 1 に示す．得られた音声データから発話区間を検出し，発話区間のみの映像を抽出する．次に発話方向推定を行い，発話方向に応じてのパン・ズーム等のカメラワークを行う．以下，3 章において発話区間検出，4 章において発話方向推定について述べる．

3　発話区間検出

発話区間検出については，これまでに我々が提案している低 SNR 環境下においても頑健な音声非音声の区間検出が可能な AdaBoost に基づく手法 [1] を用いた．

4　発話方向推定

発話方向の推定については 2 チャンネルマイクロフォンの到来時間差を相互相関の一種である CSP(Cross-Power Spectrum Phase) 係数 [2] に基づいて獲得し，その値から発話方向を推定する．

5　実験

実験にはハイビジョンカメラで撮影された映像と，2 チャンネルマイクロフォンで録音した音声データを使用した．映像 A は 2 人での会話，映像 B は 3 人での会話シーンである．元映像の撮影時間と発話区間との時間差を表 1 に示す．

	元映像	区間検出後の映像
映像 A	5 分	3 分 38 秒
映像 B	5 分	3 分 55 秒

表 1　発話区間検出の結果

発話方向が推定できた場合，そこに発話者がいると仮定して図 2 のようなカメラワーク（ズーム）を行う．発話者が特定できない場合，もしくは複数話者が同時に発話している場合には全体の映像を映す．

図 2　カメラワークの例:発話者へのズーム

6　まとめ

本稿では，固定カメラによって撮影された長時間の映像を自動編集し映像のダイジェストをするシステムを提案し，実験を行った．その結果，冗長な映像からのダイジェスト映像の作成を行うことができた．今後の課題としては，話者の感情判定や，画像情報を利用しての顔画像検出技術等との統合が考えられる．

参考文献

[1] 松田博義，滝口哲也，有木 康雄，“ Real Adaboost による音声区間検出，”日本音響学会 2006 年秋季研究発表会，2-P-12，pp.117-118，2006-09.

[2] M. Omologo，P. Svaizer，“ Acoustic Source Location in Noisy and Reverberant Environment using CSP Analysis，”Proc. ICASSP , Vol. 2, pp.921-924, 1996.