I-69

音情報を用いたイベント検出による映像要約

Video Summarization based on Event Detection using Audio Information

小河誠巳† 相澤清晴‡
Tomomi OGAWA Kiyoharu AIZAWA

1 はじめに

近年，大量の映像コンテンツを効率よく利用するための諸技術に注目が集まっている [4]．その中の技術の一つに映像要約がある．映像要約の利用により，長時間の映像を視聴する前に短時間で内容を確認することができる [1]．要約映像の作成は，放送映像だけでなく未編集映像である個人体験映像や，ホームビデオにおいてより有効であると考えられる．ホームビデオや個人体験映像は編集が行われていない，もしくは行われていたとしても編集点が非常に少ないといった特徴があり，タイトル付けがされていない場合などは内容の確認にも手間がかかってしまう．そこで，映像要約を用いることで全体の映像を簡単に確認することができるようになる．

映像要約の重要な課題としては，どのようにして個人の好み，重要なイベントを要約映像に反映するかが挙げられる．本稿では，音声情報を利用したイベントの検出と要約映像の作成法について提案する．

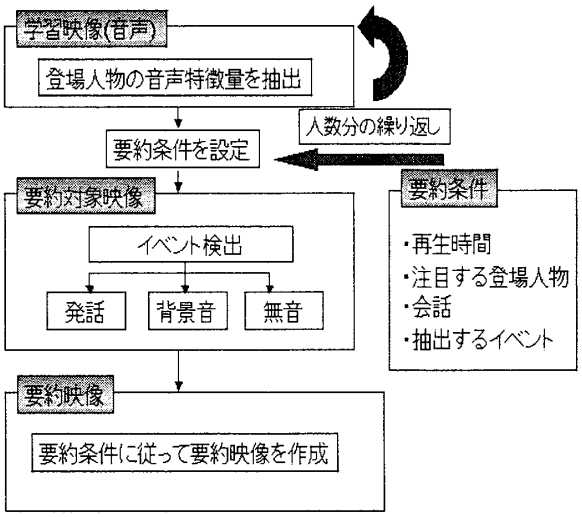

Fig. 1: Flow chart

2 音情報を用いたイベント検出

2.1 音情報から得られるイベント

本手法では，音情報を用いてイベント検出を行い，検出されたイベントを要約映像に反映させる．イベントとしては様々なものが考えられるが，今回は大きく分けて”発話”，”背景音”，”無音”の３種類をイベントとして定義する．”発話”区間ではさらに話者認識を行うことで登場人物の特定も行う．Fig.1 に全体の処理の流れを示す．

話者認識に用いる登場人物の音声情報はあらかじめ他のデータから求めておく．音声特徴量には LP ケプストラム係数 $c_m, m = 1, ..., p$ を用いた．

$$c_m = \alpha_m + \sum_{k=1}^{m-1} (\frac{k}{m}) c_k \alpha_{m-k}, \quad 1 \leq m \leq p \qquad (1)$$

ここで，$\alpha_m, 1, ..., p$ は LP パラメータ，p はモデルの次数である．

次に要約映像にどういった内容を反映させたいのか要約のための条件を設定する．要約対象映像ではイベント検出を行い，最後に要約条件で指定された区間を要約映像として出力する．

2.2 話者認識の概要

本手法で用いる話者認識手法は [2] に準じている．

今回用いる話者認識の手法では，まず元データを話者空間に射影した後 GMM(Gaussian Mixture Model) を用いて話者モデルを構築し，次に編集対象映像から得た特徴量と GMM との対数尤度を求め話者認識を行う．本手法では用意してあるモデルのうちで最も高い対数尤度を得られたモデルの ID を希望話者とする．

$$SpeakerID = \arg \max_{1 \geq i \geq Nspeakers} P(X|\lambda_i) \qquad (2)$$

ここで N は話者数 (モデルの数)，λ_i はモデルを表す．

2.3 イベントの検出

イベントは”発話”，”背景音”，”無音”の３種類の検出を行う．まず，発話区間の検出には次の短時間パワーを用いる．m はフレーム番号，N はフレームのサンプル数を示す．

$$p(m) = \frac{1}{N} \sqrt{\sum_{n=1}^{N} s_m(n)^2} \qquad (3)$$

Fig.2 に音響信号の短時間パワー例を示す．Fig.2 のように得られた短時間パワーから，閾値以上の値を持つフレームを発話フレームとして設定し，閾値未満の値を持つフレームを無音フレームとする．しかし，このままのデータでは発話区間が細切れになってしまうので，

・ある閾値以下の長さの無音フレーム (発話フレーム) が，ある閾値以上の発話フレーム (無音フレーム) に挟まれている場合，その無音フレームは前後の発話フレームに含まれる

という基準を設け，データの補正を行う．

背景音は発話区間の検出時に無音区間と判定された区間と，発話区間において話者の特定ができなかった区間から求められる．背景音の検出には [3] の手法を用い，短時間パワーと ZCR(ero-crossing rate) が閾値以下であれば無音区間として判定する．

†東京大学大学院工学系研究科
Dept. of Elec. Eug., The University of Tokyo
‡東京大学大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo

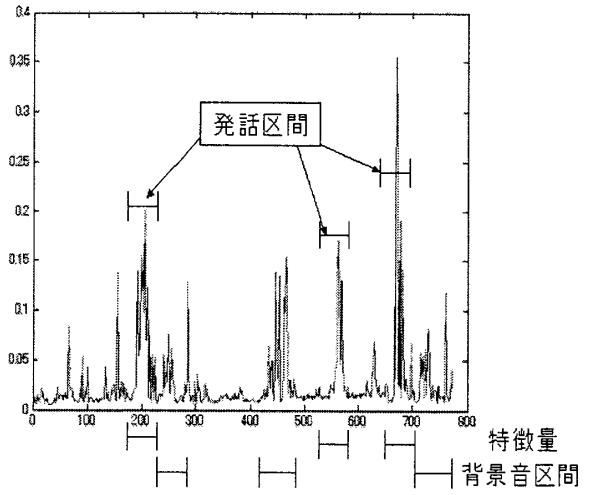

Fig. 2: Short time power

2.4 要約映像の作成

各種編集条件に従った要約映像の作成について述べる．編集条件には以下のようなものが考えられる．

・再生時間
要約映像の時間．長すぎては要約にならないし，短すぎては内容が不十分である可能性がある．
・登場人物
ユーザの希望する登場人物を中心に要約映像を作成することができれば，より個人の興味に沿った映像要約が可能となる．
・イベント
様々なイベントを定義することにより幅広い要約映像を作成することが可能となる．今回の手法では，イベントとして３種類を定義した．
・会話
特定の人物のみに注目するのではなく，特定の人物の前後で起こった発話まで抽出することで映像の内容理解が進むことが期待できる．

要約映像に抽出するイベントは，発話区間，背景音区間，無音区間を組み合わせて作成される．発話区間を抽出する場合，単独の話者のみに注目するか，複数の話者の抽出も可能である．複数の話者の抽出を行う場合，希望話者の発話区間の前後に閾値以下の範囲で他の話者の発話区間が隣接していた場合”会話”として抽出を行うことができる．

3 実験

3.1 実験データ

実験映像には約 30 分間の個人体験映像を用いた．メガネ型のカメラを装着し，本郷構内を撮影した未編集の映像である．個人体験映像は一般的に長時間の未編集映像であり，映像要約を行うことにより無駄な区間を省いて視聴することが可能になる．音声データはサンプリングレート 22050Hz，量子化ビット 8 である．

3.2 実験結果

前節での条件に従って映像要約を行った．　Fig.3 に３つのモデルの場合の話者認識結果を示す．発話区間はほぼ検出された．Fig.3 のデータから６つのの発話区間が検出され，結果が確認できた．しかし，体験映像の音声には雑音成分が強く混入するため認識程度は 50 % 以下となった．

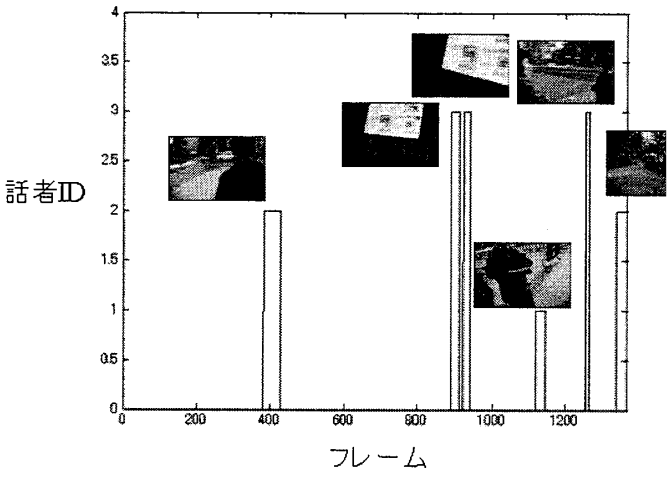

Fig. 3: Sample of video abstraction

4 まとめと今後の課題

音声情報を用いてイベント検出を行い要約映像を作成する手法を提案した．話者認識を行うことで登場人物毎の要約映像の作成が可能となった．

今回はイベントを３種類としたが今後は背景音の分析を行い，さらに分類を行う予定である．また対雑音性を持つ話者認識の手法についても検討し，認識率の向上を図りたいと思う．

参考文献

[1] 中村裕一、外村佳伸，”見たい部分を簡単に短時間で ”，電子情報通信学会誌 Vol.82 No.4 pp.346-353.

[2] 南憲一，阿久津明人，浜田洋，外村佳伸，”音韻性を抑えた話者空間への射影による話者認識”，信学誌，D-II，Vol.J85-D-II，No.4，pp.554-562 April 2002.

[3] Lie Lu, Hao Jiang and HongJiang Zhang, ”A Robust Audio Classification and Segmentation Method”, ACM Multimedia 2001.

[4] Rao Wang, Zhu Liu, and Jin-Cheng Huang, ”MULTIMEDIA CONTENT ANALYSIS”, IEE SIGNAL PROCESSING MAGAZINE NOV 2000.

東京大学大学院工学系研究科電子情報工学専攻相澤研究室
〒 113-8656 東京都文京区本郷 7-3-1
TEL:03-5841-6761
E-mail:t-ogawa@hal.t.u-tokyo.ac.jp