1B1-02

体験要約のためのビデオ自動編集手法

Automatic video editing for experience summarization

熊谷賢[*1*2] 中原淳[*2] 角康之[*1*2] 間瀬健二[*2*3]
Ken Kumagai Atsushi Nakahara Yasuyuki Sumi Kenji Mase

[*1]京都大学情報学研究科 [*2]ATR メディア情報科学研究所
Graduate School of Informatics, Kyoto University ATR Media Information Science Laboratories

[*3]名古屋大学情報連携基盤センター
Information Technology Center, Nagoya University

This paper proposes a method to summarize personal experiences by automatically editing video recorded in social events. Our approach is to automatically edit video images by two or more of environmentally installed camera/microphones and wearable camera/microphones which capture scenes at the same time. In order to summarize such scenes into video clips, we need a method to edit two or more viewpoint images (extract of an image and change of a viewpoint). First, we propose a method that we summarize an experience with an image. Second, we also propose a method to extract highlight scenes of experience with still pictures. These contents create a sense of presence – of actually being there.

1. はじめに

近年，ホームビデオやカメラ付き携帯電話の普及，環境カメラの普及から，日常の体験を記録したビデオ映像が氾濫している． こうしたビデオ映像を編集，加工するのは，一般的に煩雑な作業である．我々の興味は，そういったビデオ映像を利用した人の体験要約を実現することであり，そのための手法としてビデオ映像の自動編集手法を提案する．

ビデオ映像を自動編集する際に重要となるのは，映像データを意味のある単位に分割するために，構造を与えることである． 従来，このような構造を与えるための手法として，画像処理を用いた手法が研究されてきた．一例として，人体の簡単なモデルを仮定して，ビデオ内の人の検出や動きを理解する試み [Wren 97] があった．こうした手法を用いることで，一定の成果をあげることができるが，与えられる構造に限界がある．その一方で，あらかじめ用意された構造を用いる手法も研究されてきた．例えば，放送されるニュース映像は一般的にあらかじめ決まったシナリオがあり，画像処理という手法に加えてこうしたシナリオを併用することで，より精度の高い構造を与えることができる．しかし，日常の体験にあらかじめシナリオが存在することはまれであるため，我々が考える問題点にこうした手法を用いることはできない．本研究では，赤外線 ID システムを用いることとした．これは，センサをビデオカメラと同時に用いて，タグ認識データを映像データと同時に記録する．このタグ認識データを解析することで，映像データに構造を与える．

我々の興味と関連の深い研究として，ウェアラブルなビデオ収集システムを利用して，個人の記録を行う試みもなされてきた [Kawamura 02]．しかし，この研究は基本的に単一視点でとられた映像を加工，編集することに重点が置かれている．これに対し，我々は人のインタラクションに興味があり，環境に埋め込まれたカメラや，複数の装着型カメラから得られる多視点の映像を協調的に用いて体験を要約する．

この際に，問題となるのが，視点の選択である．例として，装着型カメラを装着した複数の人が会話している体験を想定する．このように複数の視点映像がある場合，それぞれの瞬間で，どの視点を選択するかが問題となる． 次に，体験をいくつかの静止画を用いて要約することを考える． 静止画を用いて要約する場合，どの視点を選択するかに加え，どの瞬間を選択するかが問題となる．本研究では，こうした体験を映像，静止画に要約する際に生じる問題の解決法について述べる．

連絡先: 熊谷賢，京都大学情報学研究科，京都市左京区吉田本町，kumagai@lab1.kuis.kyoto-u.ac.jp

2. 体験キャプチャルーム

日常の体験を要約するのに先立ち，学会等の展示会といった場面での体験を要約することを考える．こうした展示会における見学者の体験を要約した静止画及び, 音声を含む映像を自動的に生成し，それらを見学者に提供する．本研究では，体験キャプチャシステム [角 03] を利用して，タグ情報を取得し，それを用いる方法をとった．体験キャプチャシステムでは，センサ群を用いて，見学者の注視状況，移動情報，発話状況等のデータを得る．

赤外線 ID システムによるタグ情報の組み合わせから,「〜を見学した」,「〜に滞在した」,「〜と話した」といった，シーンの切出しと解釈を行うことができる．本研究では，それらの解釈の抽象度をボトムアップに上げていき，展示会場における他のユーザとのインタラクションに関するハイライトシーン（以下，event と呼ぶ）を抽出する手法 [高橋 03] を用いた．ビデオ自動編集手法は，こうして得られたインデックス情報を分析し，それらをもとに映像リソース・音声リソースを加工して，event を要約した静止画（以下，サムネイルと呼ぶ）・要約ビデオ（以下，サマリビデオと呼ぶ）を自動生成する手法である．以下，ビデオ自動編集手法をサマリビデオ・サムネイルのそれぞれの項目に分けて説明する．

本手法の評価に，テストベットとして 2003 年 11 月 6，7 日に ATR 研究所で行われた研究発表会のポスター展示会場における展示員と見学者の体験を要約する試みを行った．この展示会場は，さまざまなセンサ群を用いることで，人の視線状況や発話状況を検出することができるようになっている．以後，この展示会場を体験キャプチャルームと呼ぶことにする．

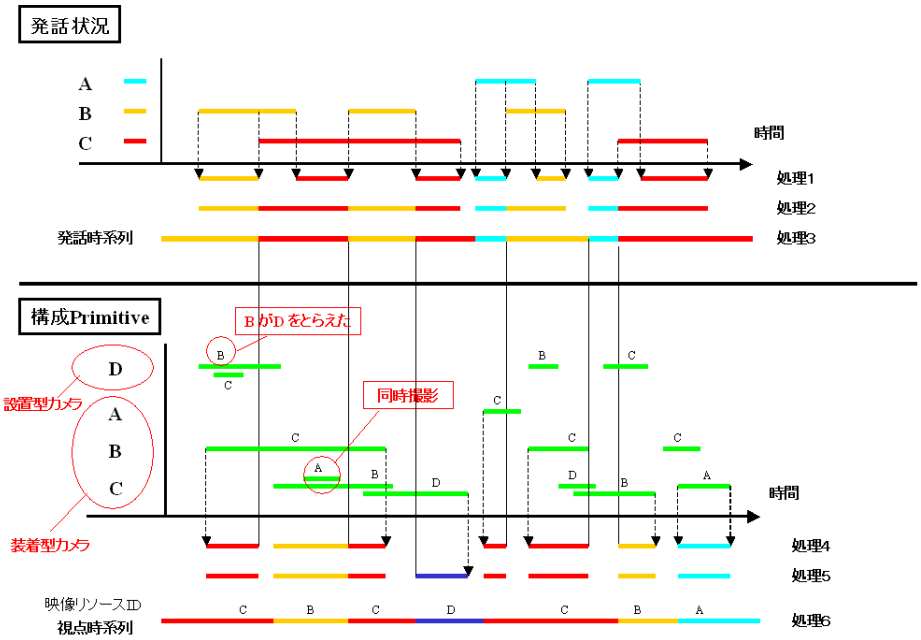

図 3: 発話時系列，視点時系列の生成例

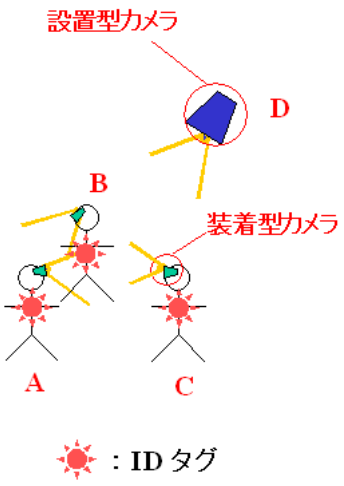

図 1: event の例（GROUP-DISCUSSION）

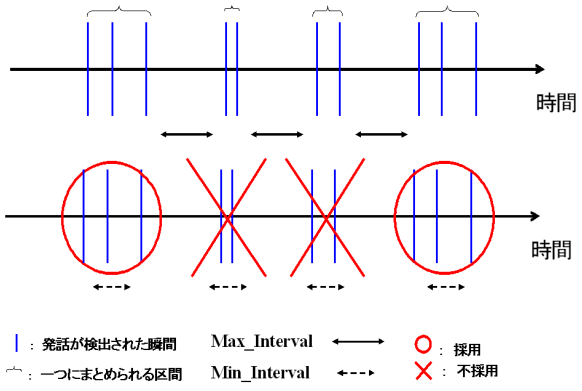

図 2: 音声データのクラスタリング

3. サマリビデオ

サマリビデオを生成する際，問題となるのは次の 2 点である．

1. どの映像リソースを選択するか
2. どの音声リソースを選択するか

例として，図 1 の event（GROUP-DISCUSSION）を考える．この場合，それぞれの時間帯において，A，B，C そして，D の映像リソースのどれを選択するか，計 4 通りの選択方法がある．仮に， この event を A の体験として要約する場合，一つは，A の映像リソースのみを用いる方法が考えられるが，GROUP-DISCUSSION（複数人で議論した）という event の映像を，常に単独の視点の映像リソースで構成することが適切ではないし，こうした理由から，単独の映像リソースのみを用いるのではなく，複数の映像リソースを協調的に用いる必要がある．

次に 2 つ目の問題について，映像の場合と同様，A の event と考える場合，A の音声リソースのみを用いるという方法が考えられる．しかし，議論している相手の音声が全くないビデオが生成された場合，それを見返した時に，何について議論されていたのかを知ることは難しい．したがって，A，B，C の音声リソースをうまく用いる方法が重要となる．以下，それぞれ映像・音声リソースの選択方法について説明していく．

3.1 映像リソースの選択方法

複数の映像リソースを協調的に用いるという問題に対し，本研究では，下記に注目するという方針をとった．

- 発話者をとらえている映像リソースを優先して用いる．

まず，event に参加した人の発話状況を調べる．発話は，断続的に検出されるが，中にはノイズが含まれていたり，一つ一つの発話記録がとぎれているため，意味のある単位となっていない．そこで，クラスタリング処理を行うことで，意味のあるオブジェクトの発話状況を検出する．処理の流れを図

2 に示す．まず，断続的な発話記録を意味のある区間として分割するために，Max-Interval を導入する．Max-Interval 以上の間隔を空けずに発話記録が検出され続けた区間を抽出する．次に，発話がノイズによるものか，そうでないかを調べるために Min-Interval を導入する．抽出された区間のうち，Min-Interval 以下の区間はノイズとして除去する．

次に，クラスタリング処理の結果，図 1 の発話状況が，図 3 のようになっているとする．始めに，発話者が一人の時間帯は，その人をその時間帯の発話者とする（処理 1）．発話者が複数いる時間帯は，その時間帯に発話を開始した人を発話者とする（処理 2）．発話者がいない時間帯は，その一つ前の時間帯に発話している人を発話者とする補正を行い，例外的に最初の時間帯だけは，一つ後ろの時間帯の発話者による補正を行う（処理 3）．この結果，発話時系列を得る．

そして，event の primitive を参照して，装着型カメラの映像リソースから，各時間帯の発話者をとらえているものを探す（処理 4）．ここで，図 3 の同時撮影のように，発話者をとらえている映像リソースが 2 つ存在する場合は，その中から任意に 1 つを選ぶ．適切な映像リソースがない時間帯がある場合，次に，設置型カメラで発話者をとらえている映像リソースを探す（処理 5）．それでも映像リソースがない時間帯は，発話時系列の時と同様に，一つ前の時間帯の映像リソース（例外的に最初の時間帯だけは，一つ後ろの時間帯の映像リソース）を用いて補正する（処理 6）．この一連の処理の結果，event を要約する際，それぞれの時間において，どの映像リソースを採用するかの時系列（視点時系列）を得て，これを元に，無音ビデオを生成する．

なお，不測の事態等で，必要な映像データが存在しない場合，それ以外で，event をとらえた任意の映像リソースを用いることで対応する．

3.2 音声リソースの選択方法

無音ビデオに対応する音源の生成方法として，次の方針を採用した．

- event に参加した全員の音声リソースを取得し，それらを合成する．

この方針を採用したのは，次の理由による．すなわち，発話として検出されなかった音声の中には，あいづちといったものもあり，ある区間で発話者とされなかったオブジェクトの音声も，参加者の間のやりとりや，場の雰囲気といったものを知るためには重要だからである．

上記の方針により生成した合成音と，無音ビデオを重ね合わせて，サマリビデオを生成する．

4. サムネイル

サムネイルを生成する際，問題となるのは，次の 2 点である．

1. どの映像リソースに含まれる静止画を採用するか（映像リソースの選択）
2. どの瞬間を採用するか（時間の選択）

サマリビデオの場合と同様に，どの映像リソースを採用するかが問題となる．ただし，サマリビデオと異なるのは，静止画は，時間幅を持たないので，さらにどの瞬間を採用するか，が問題となる．図 1 の場合，A，B，C，D の 4 つの映像リソースがあるので，最大で，event 時間長の 4 倍の数だけ候補があることになる．

視点の優先度 / Event の種類	最優先	準優先
TOGETHER-WITH	設置型カメラ	装着型カメラ
GROUP-DISCUSSION	設置型カメラ	装着型カメラ
LOOK-SAME-OBJECT	装着型カメラ	設置型カメラ
JOINT-ATTENTION	装着型カメラ	設置型カメラ
TALK-ABOUT	装着型カメラ	設置型カメラ

図 4: event の種類に応じた映像リソースの優先度

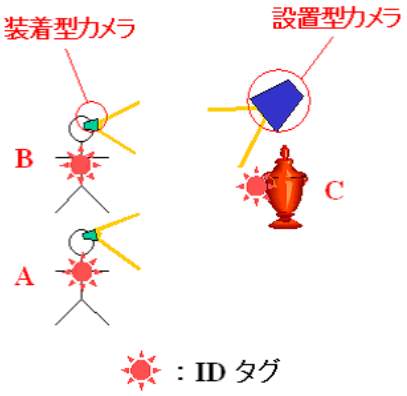

図 5: event の例（TALK-ABOUT）

以下，サムネイルの選択の項で，映像リソースの選択を説明し，サムネイルの生成の項で，時間の選択を説明する．

ここで注意しておくべきことは，サムネイルの選択と生成は，独立のプロセスということである．すなわち，最良のサムネイルを 1 つだけ生成する，というわけではなく，候補となりうるサムネイルをすべて生成した後で，その中から状況に応じて選択して用いる．これは，event を複数のサムネイルで，要約したい場合や，システムの不測の事態で，最良（と考えられる）のサムネイルが得られない場合でも，次点のもので代用することができるといった頑健性のある処理を想定しているためである．

4.1 サムネイルの選択方法

event を要約する際に，映像リソースを選択する原則として，図 4 の映像リソース優先度を設定した．このように映像リソースを設置型・装着型カメラに大別した理由は，装着型カメラからは，人の目で見えるような局所的な映像を得られるのに対し，設置型カメラからは，全体を俯瞰的にとらえた映像を得ることができるというそれぞれのカメラの特性による．

また，各 event ごとに視点の優先度をこのように設定した理由は次のとおりである．すなわち，LOOK-SAME-OBJECT（展示物を一緒に見た），JOINT-ATTENTION（共同注視），TALK-ABOUT（展示物について話し合った） といった展示物に関連する event は，展示物をとらえる可能性のある装着型カメラを優先し，他の参加者との関連が強い TOGETHER-WITH（誰かと一緒にいた），GROUP-DISCUSSION（複数人で議論した），といった event は，全員を俯瞰的にとらえる可能性のある設置型カメラを優先した．例えば，図 1 の場合は，D の映像リソースのうち，A，B，C を俯瞰的にとらえたサムネイルの優先度を最も高くし，次に，A，B，C の映像リ

ソースのうち，お互いをとらえたサムネイルを優先する．また，図 5 の場合は，A，B の映像リソースのうち，議論の対象となっている展示物をとらえたサムネイルの優先度を最も高くし，次に，C の映像リソースのうち，議論をしている A，B を俯瞰的にとらえたサムネイルを優先する．

4.2 サムネイルの生成方法

event は，複数の primitive で構成されており，primitive が，意味のあるインタラクションの最小単位となるので，event を代表するサムネイルを生成するという問題は，primitive のサムネイルを生成するという問題に帰着する．すなわち，それぞれの primitive のサムネイルが event のサムネイルの候補となる．以下，primitive のサムネイルを生成する方法について説明していく．

primitive には，開始時間，終了時間，primitive 名，動作主 ID，相手 ID の属性があり，開始時間から終了時間までのどの時間を採用するかの問題が残っている．この問題を解決する際に，次の方針を採用した．

1. 相手 ID をとらえている時間を優先する．

2. ID が中心にとらえられている時間を優先する．

1 について，primitive は，動作主 ID と，相手 ID 間のインタラクションと定義されており，相手 ID とは異なる ID は，ノイズであり，primitive のサムネイルを考える場合，当然相手 ID をとらえているもので無ければならない．2 は自明であるとする．そして，primitive ごとに，上記のルールを満たす時間を算出し，静止画を生成する．

なお，状況によっては複数人をとらえているサムネイルが重要である場合もある．例えば，図 1 の場合，D の映像リソースで，A，B，C を同時にとらえているサムネイルがあれば，それぞれを個別にとらえている場合よりも，重要である．本研究では，単独の人が写っているサムネイルだけでなく，そうしたサムネイルも生成できるようにした．

5. 実験と考察

2003 年 11 月 6，7 日に ATR 研究所で行われた研究発表会において，体験キャプチャルームで，本手法を実装したシステムを稼動させたところ，サマリビデオを 572，サムネイルを 6336，それぞれ自動的に生成した．これらを展示員・見学者に提供して感想を聞いたところ，次のような意見が得られた．

まず，サマリビデオに関して，肯定的なものとしては，「それぞれの時間における発話者をとらえた映像を見ることができるので，すべての参加者の映像リソースを見なくても，場の雰囲気のようなものは，伝わってくる」というものがあった．これは，発話者に注目した映像リソースの選択により，臨場感のあるサマリビデオ生成を実現することができたためであると考えられる．否定的なものとしては，「自分の話したことが他人に聞かれてしまうので，音声が残されることに抵抗がある」というものがあった．個人的な話をしていて，他の人に話を聞かれたくないような場合，音声の扱い方が重要となってくる．状況を理解しやすくするために音声を利用することと，その結果，プライバシーを侵害しないようにすることの両立が，今後，より社会的な場面で，要約技法を用いる際の課題である．

次に，サムネイルに関しては，「サマリビデオに出てこない静止画がサムネイルとして採用されていることに違和感がある」という意見があった．我々は，サマリビデオに比べて，サムネイルは，時間的な幅を持たないため，適切な静止画を生成するためには，状況に応じた（event の種類に応じた）視点・時間の選択が必要であると考えた．そのため，サマリビデオと別々のものとして生成することとなり，サムネイルで表される場面が，サマリビデオに登場しない場合もある．今回は，サマリビデオ・サムネイルを提示する際に，サマリビデオの一場面という認識を与えてしまったことが，上記のような違和感を与えることにつながってしまったと考えられるので，今後，サマリビデオ・サムネイルの提示方法を改良する必要があると考えられる．

6. おわりに

サマリビデオに関して，今回は，発話者をとらえた映像リソースを優先して用いるという手法を用いた．今後，さらに他の適切なルールがあるかどうかを，映像のリソースと primitive を見比べて，ボトムアップ的に考察していきたい．

また，展示会場以外でのサマリビデオ・サムネイル生成を考察していきたい．

7. 謝辞

本研究を進めるにあたり，多分のご意見，ご協力を賜りました伊藤禎宣氏を始めとする ATR メディア情報科学研究所の皆様，高橋昌史氏を始めとする西田研究室の皆様に感謝します．本研究は情報通信研究機構の研究委託により実施した．

参考文献

[Wren 97] Christopher Wren, Ali Azarbayejani, Trevor Darrell, Alex Pentland: Pfinder Real-Time Tracking of the Human Body, IEEE Transactions on Pattern Analysis and Machine Intelligence , 19 No. 7 (1997).

[Kawamura 02] Tatsuyuki Kawamura, Yasuyuki Kono, Msatsugu Kidode: Wearable Interfaces for a Video Diary: towards Memory Retrieval, Exchange, and Transportation, The 6th International Symposium on Wearable Computers (ISWC2002) , 31-38 (2002).

[角 03] 角康之，伊藤禎宣，松口哲也，Sidney Fels，間瀬健二: 協調的なインタラクションの記録と解釈, , 2628-2637, 情報処理学会論文（2003）.

[高橋 03] 高橋昌史，伊藤禎宣，角康之，間瀬健二: 複数センサを利用したインタラクション・パターンの自動抽出, , ユビキタスコンピューティングシステム，情報処理学会研究報告（2003）.