3A3-04

シーン推定と漫画技法を用いた体験要約システム

Experience Summarization By Scene Presumption And Using Comic Format

小関 悠[*1*2] 角 康之[*1*2] 西田 豊明[*1*2] 間瀬 健二[*2*3]
Yu Koseki Yasuyuki Sumi Toyoaki Nishida 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 Infomatics, Kyoto University ATR Media Information Science Laboratories

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

In this paper, we proposed the system which automatically summarize experience data. First, system presume scenes, meaningful combinations of interactions, by using sensor data. Second, it calculate how important each scene is. With this result, it can remove trivial scenes. Next, it capture a good snapshot from a image of each scene and decorate it with comic format which allow readers of summary to understand easily. Last of all, it determinate the alignment of scenes by using their importance.

1. はじめに

近年、ビデオカメラや赤外線センサーといった機器が小型化し、また安価に利用出来るようになったことにともない、これらを組み合わせた体験記録の手法が複数提案されている。例えばインタラクション・コーパス収集システム [1][2] は、博物館や学会などの展示場で用いられることを想定した体験記録システムであり、参加者が装着するウェアラブルカメラとセンサー、環境に設置する小型カメラとセンサー、得られたセンサー情報からインタラクションの解釈を行うサーバーなどで構成されている。

こうした複数のカメラを用いた体験記録システムでは、容易に膨大な量の映像データを得ることが出来る。しかし一方、このようにして得た映像データはあまりに大量なので、例えばこれらの中から目的の場面、あるいは何らかの重要なシーンを見つけ出すというような作業は、非常に困難になる。また映像データの要約を作成する場合でも、従来のビデオ編集のように人間がその作業を行うには、扱う映像データがあまりに大量なので、やはり困難である。

以上のような背景により、体験記録システムが簡単に構築されるようになり、扱うデータが大量になるほど、そうしたシステムに対応した自動的な体験要約システムの構築が求められる。

そこで本稿ではセンサー情報を用いて、自動的に体験要約を行うシステムを提案、実装する。体験記録システムにはインタラクション・コーパス収集システムを用い、まずこれによって得たインタラクション情報から、意味のあるインタラクションの列なり、「シーン」を推定する手法について述べる。続いて、そうして得たシーンの重要度を求める手法について簡単に述べる。これにより体験要約の作成のために必要な、重要なシーンの推定や、不要なシーンの省略が可能になる。続いてシーンの中から適切な静止画を切り出し、漫画技法を用いた修飾を行うことで、より分かりやすい要約が可能であることを述べる。最後に、複数のシーンをそれぞれの重要度に応じて配置し表示する手法について述べる。

なお今回は体験記録データとして、2 時間ほどの長さの、一実験室で行われた展示会を対象とした。参加者は 6 人、展示ブースは 3 つ、映像データはのべ 20 時間強である。本稿中に示した幾つかの定数は、このデータに対応したものである。

連絡先: 小関 悠 (Yu Koseki), 京都大学大学院情報学研究科, 京都市左京区吉田本町工学部 10 号館 223 号室, koseki@ii.ist.i.kyoto-u.ac.jp

2. シーン推定

本稿では、一日は複数のシーンの列なりであると仮定する。ある人の朝は「起きる」「歯を磨く」「ごはんを食べる」「外に出る」…という複数のシーンから構成されており、また別の人の朝は「起きる」「車に乗り込む」「店で買い物をする」…という複数のシーンから構成されている、といった具合である。シーンは一日の体験を時間的に、その意味に応じて適切に分断したもの、と定義出来る。

例えば「起きる」というシーンは「目を覚まし」「体を起こす」という二つのシーンと捉えることが出来るかもしれないし、あるいは逆に二つのシーンをまとめて「起きて歯を磨く」というシーンと見なすことが出来るかもしれないが、そのどれが最も適切な体験の分断であるのかは、求める要約結果によって変化する。しかし一般的には「起きる」「歯を磨く」というように分断するのが、最も単純な見方で、良いと言えるだろう。

今回、実装に用いるインタラクション・コーパス収集システムでは、センサー情報から自動的に解釈されるインタラクションの種類として「見る」「見られる」「見つめ合う」というものと「話す」「話される」「話し合う」というものがある。これらのインタラクションは非常に断片的であり、個々の長さも短いため、体験要約システムの構築にはこうしたインタラクションの組み合わせから、ある程度の長さで意味を成すようなシーンを推定する手法が求められる。

例えば、図 1 のようなインタラクションの組み合わせを考える。このように「A と話す」というインタラクションが 2 回あり、それから「B と話す」というインタラクションが 2 回ある場合、まず「A と話すシーン」(シーン 1)、それから「B と話すシーン」(シーン 2) を推定するのが妥当である。全てのインタラクションをまとめて「A や B と話す」という 1 つのシーンにすることや、インタラクションごとに「A と話すシーン 1」「A と話すシーン 2」などとばらばらのシーンにすることも出来るが、それぞれシーンの意味が複雑になる、あるいはシーンが短くなりすぎるので、妥当とは言えない。

続いて図 2 のような組み合わせを考える。図 1 の 3 つ目の

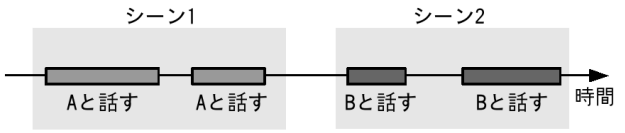

図 1: インタラクションの組み合わせ例 1

インタラクションが「B と話す」から「A と話す」に変化している。この場合もやはり「A と話すシーン」と「B と話すシーン」を推定するのが妥当と考えられる。図 1 と図 2 から、シーンの長さはインタラクションの組み合わせに応じて柔軟に変化する方が好ましいことが分かる。

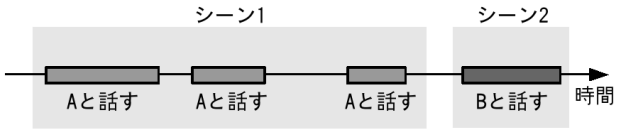

図 2: インタラクションの組み合わせ例 2

図 1 や図 2 のような組み合わせでは「同種、同対象のインタラクションが続いている間は同じシーンとする」という方針で良かった。しかし実際には、誰かと話している間に少しだけ別の誰かに声をかけられる、というようなことがままある。このような場合あまりに多くのシーンに分割し過ぎると、シーン数が多過ぎて扱いに困ることになり、また個々のシーンがあまりに短か過ぎて意味を成さなくなってしまう。

図 3 はそのような例である。ここでは「A と話す」というインタラクションの合間に「B と話す」というインタラクションが存在している。これを「A と話すシーン」「B と話すシーン」「再び A と話すシーン」「C と話すシーン」と 4 つのシーンに推定することも出来るが、「A と話している合間に B とも話すシーン」（シーン 1）「C と話すシーン」（シーン 2）と 2 つのシーンに推定する方が、前述のような理由により妥当であると考えられる。

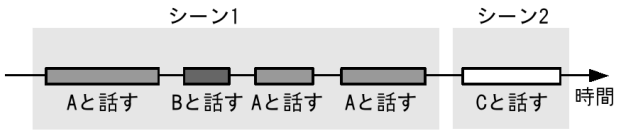

図 3: インタラクションの組み合わせ例 3

また図 4 は別の意味で「同種、同対象のインタラクションが続いている間は同じシーンとする」という方針にためらいを感じる例である。ここでは全てのインタラクションが「A と話す」であるが、2 つ目と 3 つ目のインタラクションの間に大きな時間的余白があるため、全てを 1 つのシーンにまとめるのは妥当ではないと考えられる。ある程度以上の時間的空白があった場合は同じ相手と話をしたのだとしても、別の話題になるなど、以前の対話とは関連性が薄くなる可能性が高いからである。

以上のようなシーン推定のアルゴリズムを実装するため、まず「距離」という概念を導入する。この「距離」とはインタラクション間の関連性のことであり、「距離」が短いほど同じシーンに入る可能性が高いことを意味する。基本的に「距離」はインタラクション間の時間差で代用するが、インタラクションの種類や対象が異なる場合は時間差に 1 以上の定数をかけたものとする。定数は体験記録の対象がどのような場であるかに応じて変化すべきだろうが、ここでは 10 とする。これにより、図 3 のような異対象でも時間差の小さなインタラクションが同一シーンと推定出来る一方、図 4 のような同対象でも時間差の大きなインタラクションは別のシーンと推定出来るようになる。

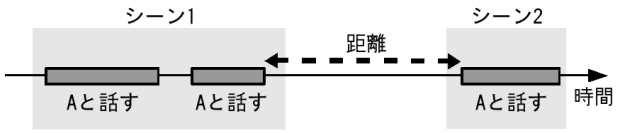

図 4: インタラクションの組み合わせ例 4

このようにして個々のインタラクション間の「距離」を求めた後に、その「距離」を用いて階層的クラスタリング・アルゴリズムである群平均法を導入する。クラスタリングは、全てのクラスター間の距離があらかじめ定めた値を越えるまで何度も繰り返す。ここでは、その値を 900 としている。これは同種同対象のインタラクションは、基本的には、900 秒（15 分）以内であれば同一シーンに含まれることを意味する。また異種、あるいは異対象のインタラクションは、基本的には、その 1/10、90 秒以内であれば同一シーンに含まれる。この値もやはり体験記録の対象に応じて考える必要がある。

以上のようにインタラクションの集合に対してクラスタリングを行うと、複数のクラスターを形成出来る。そして、こうして得られた個々のクラスターは、ある程度の長さを持った、何らかの意味を持つシーンと見なすことが出来る。

3. シーンの重要度

優れた体験要約を作成するためには、シーンに分割するだけでは不十分で、個々のシーンの重要度を求める必要がある。これにより重要なシーンを目立たせたり、あまり重要でないシーンを省くことが可能になる。

ここではそれぞれのシーンの重要度を、そのシーンが含む全てのインタラクションを点数化し、合計したものとして考える。それぞれのインタラクションにどのような点数を与えるかは、やはり体験記録の場がどのようなものであるか、あるいはどのような体験要約を作成したいか、といったことに大きく左右されるが、ここでは図 5 のように定めている。

例えば図 3 のようなデータが与えられた場合、シーン 1 は「A と話す」「B と話す」「A と話す」「A と話す」というインタラクションから構成されているので、2 点が 4 つ、合計 8 点になる。シーン 2 は「C と話す」だけから構成されているので 2 点である。

こうして得た点数、重要度は、後にそれぞれのシーンをどれくらい大きく表示するか、あるいは表示させないか、という判断に用いる。

インタラクションの種類	点数
見る	1 点
見られる	0 点
見つめ合う	2 点
話す	2 点
話される	2 点
話し合う	4 点

図 5: インタラクションの点数化

4. 漫画技法の導入

映像データは基本的に、撮影時間と同じだけの視聴時間が必要になる。もちろん重要ではないシーンを省略するというような要約手法を用いることで、視聴時間をある程度短くすることは可能である。しかし膨大な映像データの内容を短時間でつかむためには、映像データという形式にこだわらない方が、大幅な要約が可能となる。

例えば漫画では、吹き出しや擬音語、擬態語を用いることで、その場面の音の情報を視覚的に示したり、あるいはコマの形状や大きさを変化させたり、ハイライト処理を行うことで、その場面の意味の情報を視覚的に示す手法が多用されている。こうした手法により、漫画は様々な情報をコマやページの中に含めながら、それらを容易に見てとることが可能になっている。

本稿では映像データの要約手法として、このような漫画技法を用いたパラパラアニメを用いる。これは、個々のシーンから適切なカットを複数、静止画形式で切り出し、漫画技法を用いてその内容を修飾した上で、パラパラアニメのように次々と表示するというものである。

この手法ではまず初めに、一つのシーンとして得られる映像データから、そのシーンが含む個々のインタラクションについて、センサーの発火時間に合わせた静止画を自動的に切り出す（図 6）。インタラクション・コーパス収集システムでは通常、一つのインタラクションは複数のセンサー情報から意味付けられているので、ここではなるべくそのインタラクションの中心にあったセンサーの発火時間（図中◆のタイミング）から静止画を切り出す。

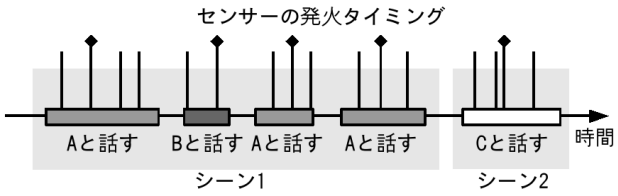

図 6: シーンからの静止画切り出し

切り出しをセンサーの発火時間に合わせるのは、その瞬間、センサーを持ったインタラクション対象はカメラに捉えられている可能性が高いからである。

図 7 の左はそのようにして映像データから切り出された静止画の例である。この例ではインタラクションの対象である人物が大きく画像の中心に表示されているが、より明示的に示すためハイライト処理を行う。図 7 の右はハイライト処理を行った結果である。センサーの位置情報から、インタラクションの対象が画像のどのあたりに存在するかを類推し、その周辺以外はモノクロにし、またコントラストを強めて曖昧にさせている。これにより大勢の人物が画像に映し出されているような場合でも、どの人がインタラクションの対象であるのか、簡単に理解出来る。例ではハイライト処理により、インタラクションの対象でない右側の人物がモノクロになって潰れていることが見てとれる。

図 7: ハイライトの例

また「話す」「話される」といったインタラクションが起きた場合は、対応する吹き出しを画面に表示する（図 8）。画面の端から表示される吹き出しは、その映像を撮影している本人が話をしたことを示している。会話の中身については、映像データから得られる音声の品質が良くない上に、周囲の雑音が多く、発音も不明瞭であるため、音声認識は出来なかった。しかしそのかわりに、インタラクションの長さに応じて記号列を吹き出しの中に書き加えているため、その会話がどれだけ長かったかは簡単に見てとることが出来る。記号列は偶然意味を持つことがないよう、日本語やアルファベットの使用を避けている。

図 8: 吹き出しの例

このように漫画技法を用いて修飾した、個々のインタラクションの内容を示した静止画を、シーンごとにまとめてパラパラアニメとする。これにより映像データの意味を失わず、時間的に大幅な圧縮が可能となる。

5. 重要度に応じた配置

これまでに述べた手法により完成したシーンごとのパラパラアニメを、1280 × 1024 ドットのディスプレイに一度に表示出来るように配置する。パラパラアニメは、そのシーンの重要度が全てのシーンの重要度合計のどれだけを占めるかに応じて、大・中・小のいずれかのサイズで表示、あるいは表示しないようにする。この基準と、サイズの対応は図 9 の通りである。

図 9 に記した基準は代表的な例で、実際にはシーン数や他のシーンの重要度に応じて変化する。例えばシーン数が 4 し

サイズ	画像の大きさ	基準
大	600 × 450	約 25%、4 つまで
中	400 × 300	約 12%、4 つまで
小	200 × 150	ページを埋められるまで
表示なし		それ以外

図 9: シーンのサイズとその基準

かないような体験記録の場合は、その重要度によらず全て大サイズになる。

また、他のシーンに比べて著しく重要度が劣るシーンは表示させない。これは全てのシーンを網羅して表示することよりも、一画面内におさめることを重視した結果である。

なおシーンの配置アルゴリズムには、あらかじめ定めた 1296 通りのパターンから、どれくらい重要度が特定のシーンに偏っているかに応じて、最も適切なものを選ぶという手法を用いている。またシーンはその時間に合わせて、なるべく体験の序盤のものが左上方向に、体験の終盤のものが右下方向になるように配置する。

このようにして、一人一日ごとの体験要約が作成される。アウトプットは GIF アニメと HTML の汎用的な組み合わせであり、ほとんどのウェブブラウザで閲覧可能である。図 10 はアウトプットの例である。重要と判定されたシーンが右上に大きく配置されている。これは、ある展示ブースの前でこの日最も長く会話が交わされたシーンである。

図 10: ある人の体験要約

6. おわりに

本稿では、大量の映像データに対し、センサー情報を用いて自動的にシーンを推定し、またシーンごとの重要度を求めるシステムについて述べた。また映像データから適切な静止画を切り出して、漫画技法を用いた修飾を行い、それらを重要度に応じて配置することにより、短い時間で理解出来る要約を作成するシステムについても述べた。今後はより多様な体験記録に対しての本システム評価や、複数人の体験記録についての横断的な要約手法について考えたい。

7. 謝辞

本研究は情報通信研究機構の委託研究「超高速知能ネットワーク社会に向けた新しいインタラクション・メディアの研究開発」により実施した。

参考文献

[1] 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, 情報処理学会論文誌, Vol.44, No.11, pp.2628-2637, 2003 年 11 月

[2] Y. Sumi, S. Ito, T. Matsuguchi, S. Fels, K. Mase: Collaborative capturing and interpretation of interactions, Pervasive 2004 Workshop on Memory and Sharing of Experiences, pp. 1-7, 2004.

[3] 坂本 竜基, 角 康之, 中尾 恵子, 間瀬 健二, 國藤 進：コミックダイアリ：漫画表現を利用した経験や興味の伝達支援, 情報処理学会論文誌, Vol.43, No.12, pp.3582-3595, 2002 年 12 月