情報処理学会論文誌 Vol.57 No.12 2516–2525 (Dec. 2016)

ミュージアムにおける一人称動画短縮のための場面抽出 ——自身での振返りと他者との共有

長徳 将希[1,a)] 小泉 直也[1,b)] 苗村 健[1,c)]

受付日 2016年1月20日, 採録日 2016年9月6日

概要：個人が撮影した動画は，そのままでは冗長で見返しにくいという問題がある．本稿では，自身が撮影した動画を，自身での振返りと他者との共有という 2 つの使用用途に合わせて，適切に場面抽出する方法を検討する．体験型展示の鑑賞で説明員との対話がなされるような，インタラクティブな要素の多い作品展示における鑑賞体験を撮影対象とし，ミュージアム体験の流れの要素である「見る・話す・聞く」場面をピックアップするよう，オプティカルフローと発話検出器を用いたシーンカットを行った．以上の仕組みを用いて，メディアアートの展示会である東京大学制作展においてユーザスタディを実施した．その結果，自身での振返りには，作品を見つめているシーンと自身が発言しているシーンが，他者との共有には，作品を見つめているシーンと説明員が発言しているシーンがダイジェストとして適切であることが明らかになった．

キーワード：鑑賞体験，一人称動画，自動要約，振返り，共有

Automatic Scene Selection of First-person Video for Recalling and Sharing Museum Experiences

Masaki Chotoku[1,a)] Naoya Koizumi[1,b)] Takeshi Naemura[1,c)]

Received: January 20, 2016, Accepted: September 6, 2016

Abstract: It is tedious to watch long raw video clips. In this paper, we assumed two types usages of digest video clips: those to recall users' own experiences and those to share users' experiences with others, and examined the appropriate way to edit each of them. We set video clips in which museum visitors can act interactively as the object of our study, and generated video digests using optical flow and speech detector to retrieve scenes associated with users' flow of experiences: "watching" and "talking, listening". Moreover, we conducted a user study utilizing our system at the iiiExhibition2015, a media art exhibition held by the University of Tokyo. We found that scenes of "watching" and "talking" tended to achieve higher evaluation scores in the case of recalling users' own experiences, whereas scenes of "watching" and "listening" tended to get higher evaluation scores in the case of sharing users' experiences with others.

Keywords: Museum Experience, First-Person Video, Automatic Summarization, Recall, Share

1. はじめに

人は記録する生き物である．歴史上数多く存在する壁画や絵画が示すように，古来より人は目に映る現実の様子を保存してきた．その目的の 1 つに，人の営みや自然現象といった，そのときその場所でしか体験できない出来事を保存し，自分自身で過去の体験を省みたり，他者に自身の体験を共有したりすることがあげられる．現代においてはデジタルカメラによって，より手軽で安価に自身の体験を記録・保存することができる．さらに近年では，身につけて撮影を行うウェアラブルカメラの普及により，自身の見たものをそのまま撮影することが個人でも容易に行えるようになった．さらに，映像を Instagram や Facebook，Twitter などの SNS で共有することが広く一般的に行われるよう

[1] 東京大学
The University of Tokyo, Bunkyo, Tokyo 113–8654, Japan
[a)] chotoku@nae-lab.org
[b)] koizumi@nae-lab.org
[c)] naemura@nae-lab.org

© 2016 Information Processing Society of Japan

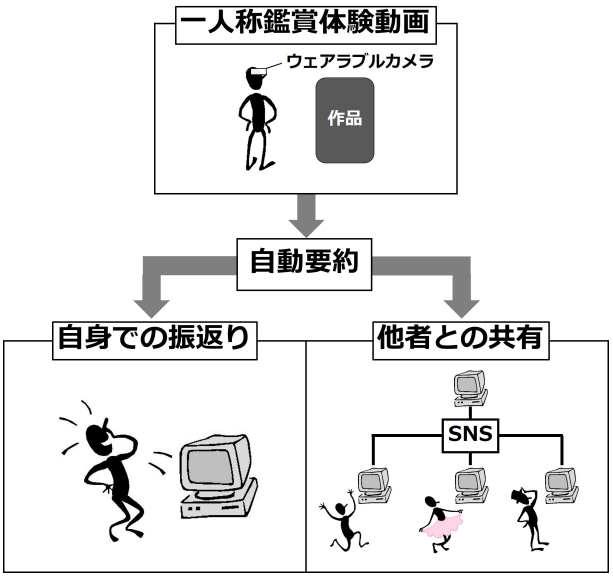


図 1 本システムのコンセプト
Fig. 1 Concept of This System.

になってきており，Instagram では 1 日に平均約 8,000 万枚もの写真や動画が投稿されている．ここから，自身の体験を他者と共有したいというニーズの高さが分かる．

しかしながら，個人が撮影した動画は，その場の状況をありのまま残すためしばしば冗長で，そのままでは見返しにくいという問題が指摘されている [1], [2]．そこで本研究では，個人が撮影した動画の自動要約にむけた場面抽出について検討する．このとき，同じ要約動画であっても使用用途によりユーザの評価が異なることがいわれており [3]，自動要約を行う際には，誰がどういう場面で使用するダイジェスト動画なのかを明確に設定する必要がある．

本稿では，ミュージアムにおいて，作品と対話的に鑑賞を行っている体験動画を対象とし，以下の 2 つの使用用途において，適切なシーンカットの違いを明らかにする．

- 自身の体験を思い出すために個人的に見る動画
- 展示会の様子を他者に紹介するために一般公開する動画

本システムのコンセプトを**図 1** に示す．

2. 関連研究

2.1 動画の自動要約

既存動画の自動要約を行う研究として，テレビ番組や映画，ネット上に公開された動画などの自動要約を行う研究がさかんに行われている [4], [5], [6]．さらに，自身で撮影した動画を自動要約する研究として，脳波を利用する手法 [1] や位置情報と地理情報を用いる手法 [2]，赤外線 ID タグシステムを用いる手法 [7]，音量情報および加速度センサを用いる手法 [8] などが提案されている．これらの手法は，主にシステムを利用するユーザ自身が振り返ることを目的とした自動要約手法である．

一方で，動画を他者と共有することを想定して自動要約を行う研究が Buschek らによってなされている [9]．彼らは加速度の変化を基に動画の自動要約を行い，人が編集したものと比較してユーザ評価の違いを報告している．自動カットが 1 シーン 7 秒，人がカットしたものが 1 シーン平均 10 秒だったが，ユーザからはどちらもカットの頻度が高いという指摘がされた．さらに，編集された動画に関して，それを一般公開・友人と共有・個人視聴の 3 つの用途で考えた際，公開範囲が広がるにつれて，動画に必要なクオリティが高くなることが報告されている．しかし，それぞれの用途で具体的にどのようなシーンが望まれるのかに関しては明らかになっていない．そこで本研究では，異なる目的に対して，ユーザが必要とするシーンにどのような違いがあるかを明らかにする．本稿では，動画編集の目的に関しては，個人視聴を目的としたプラベートな動画と，一般公開による不特定多数との共有を目的としたパブリックな動画の 2 つの用途のみを考えた．友人と共有するという目的は，Youtube や Facebook の動画の公開設定では存在するが，友人などを対象とした場合は個人間や関係性に依存するため，一般化した議論が難しいと考えた．そのため本研究では，人間関係などのコンテキストに依存しない部分として個人視聴と一般公開の 2 つの目的を取り上げ，それぞれで必要とされるシーンの違いを研究対象とした．そこで本研究では，異なる目的に対して，ユーザが必要とするシーンにどのような違いがあるかを調査する．

2.2 ミュージアムにおける鑑賞体験の記録

ミュージアムにおける鑑賞体験を記録し，持ち帰る仕組みが提案されてきた．ボタン押下から得られた嗜好情報によって，オリジナルのリーフレットが貰える Peaflet がソンらによって提案されている [10]．Durrant らは，テーマパークにおいて，個人が撮影した画像とテーマパーク側の機材により撮影した画像をマージしてお土産のリーフレットを作成するシステム，Automics を提案している [11]．小関らは，ウェアラブル型と設置型の端末を用いてユーザ体験を記録し，それを漫画的レイアウトを組んだぱらぱらアニメに自動要約するシステムを提案している [12]．また，インスタグラムに投稿されたミュージアムでの写真を対象として，写真を自動で整理することでユーザ体験の物語を作る研究が Weilenmann らによって行われている [13]．

これらの先行研究は，静止画で体験を持ち帰る仕組みであったが，本研究は場面抽出された動画を持ち帰る仕組みである．場面抽出する際に重要視すべき点としては，東京国立近代美術館のスクール・プログラムで述べられている，「見る」「考える」「話す」「聞く」という，鑑賞体験におけるユーザの行為を参考にした [14]．

© 2016 Information Processing Society of Japan

3. 提案システム

3.1 目的

ミュージアム体験において撮影された動画から，体験の要点となるシーンを抽出することを目的とする．ミュージアム体験の要点に関しては，対話型鑑賞 [15] を参照した．この対話的鑑賞の指針は「見る」「考える」「話す」「聞く」であり，本研究ではそれを参考に研究を設計した [14]．本研究では，これらの行為を抽出することで，ユーザが体験した内容の要点を抽出できると考えた．一方で，「考える」行為は非感覚的な行為であり，抽出が難しいため，本システムでは残りの 3 つの行為がなされる場面を抽出対象とした．3 つの場面を「見る」と「話す」「聞く」の 2 つに分け，「見る」をまとめて抽出する「静視カットアルゴリズム」および，「話す」「聞く」をまとめて抽出する「発話カットアルゴリズム」を提案する．

3.2 システム設計要件

システムの設計要件として，以下を設定した．

- 要件 1：「見る」および「話す」「聞く」場面の動画内からの抽出
- 要件 2：撮影，編集および公開の手軽さ

3.2.1 要件 1 に対する解決手段

(1)「見る」場面の抽出（静視カット）

「見る」振舞いは身体の動きが少ない場面と考えられ，オプティカルフローベクトルの大きさで抽出できると考えた．そこで，オプティカルフローベクトルの大きさの画面全体での平均が小さい，静視している場面をピックアップする．

(2)「話す」「聞く」場面の抽出（発話カット）

「話す」「聞く」振舞いは，動画音声から発話検出を行うことで抽出できると考えた．具体的には，音声認識エンジン Julius [16] の機能の 1 つである，adintool [17] を用いて音声波形データ中の発話区間の検出を行う．

3.2.2 要件 2 に対する解決手段

体験を阻害せず，自然なインタラクションの中でピックアップシーンの選択を行う必要がある．撮影時には，ハンズフリーで撮影が可能なウェアラブルカメラを 1 つのみ使用することで手軽さを実現する．また要件 1 で抽出する静視および発話という行為を利用することで，追加デバイスを必要とせずハンズフリーで行える行為により場面抽出を行う．さらに公開の手軽さのため，切り出された 1 シーンを SNS にそのままアップロード可能となるよう，Instagram [18] のアップロード制限を基準に，1 シーンの最長時間を 15 秒とした．また，文献 [9] では，カットされた動画を視聴する際に，10 秒という長さでは短いと感じるユーザが多いことが指摘されているため，最低でも 10 秒を超えるよう，1 シーンの最短時間を 11 秒として動画の長さを確保することを設計指針とした．

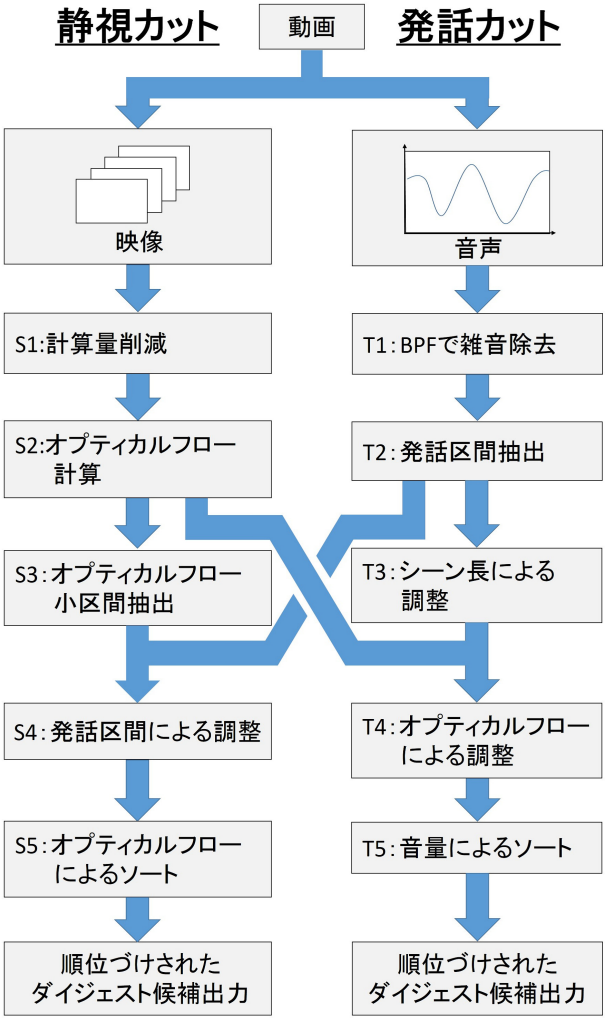


図 2 静視カットと発話カットの処理の流れ
Fig. 2 Processing Flow of Gaze-Cut and Speech-Cut.

3.3 抽出方法

「静視カット」では，オプティカルフローの値から映像揺れが小さい場面を抽出する．ただし，動画区間の切れ目で発言が途切れてしまうシーンカットを行うと，違和感を感じてしまう．そこで，発話区間の途中からシーンが開始したり，発話区間の途中でシーンが終了したりすることがないように調整を加えた．「発話カット」では，発話検出器により発話区間を抽出する．ただし，歩行しているシーンやあまりに映像揺れが大きいシーンは振返りや共有には不適切な場面であると考えた．そこで，オプティカルフローの値を用いてそれらのシーンを除外する．

「静視カット」と「発話カット」の処理の流れを図 2 に示す．それぞれのシーンカットについて以下で説明する．

3.3.1 静視カット

S1 計算量削減

動画から得られた映像から，各フレームごとに画面全

© 2016 Information Processing Society of Japan

体のオプティカルフローの合計値を計算する．本システムで使用を想定しているカメラの解像度とフレームレートは，1920 × 1080 pixel，30 fps だがそのままオプティカルフローの計算を行うと計算コストが高く，処理に時間がかかるため，映像を 160 × 90 pixel，10 fps に変換する．

S2 オプティカルフロー計算
オプティカルフローの計算は，今回扱うウェアラブルカメラで撮影した映像のような，フレーム間での変化量が大きい映像に対してもロバストな結果が得られる Liu らが提案した手法を用いた [19]．

S3 オプティカルフロー小区間抽出
3.2.2 項より，1 シーンの時間は 11–15 秒である．ここではまず，1 シーン 11 秒での切り出しを行い，S4 での区間の切れ目の調整で，11 秒の前後 2 秒をどこまで延長するかを決定する．

(i) オプティカルフローのフレーム全体の合計値 $F_{raw}(t)$ を平滑化し，ノイズの影響を小さくする．

$$F(t) = \frac{\int_0^{t_w} F_{raw}(t+\tau)d\tau}{t_w} \quad (1)$$

ここで，$t_w = 11(\mathrm{sec})$ とする．

(ii) 閾値を定義し，$F(t)$ が閾値を下回る区間を静的区間とする．閾値は，動画全体の $F(t)$ の平均とした（T_{end} は動画全体の長さから t_w を引いた時間）．

$$Threshold = \frac{\int_0^{T_{end}} F(t)dt}{T_{end}} \quad (2)$$

(iii) 各静的区間から，最もオプティカルフローが小さい区間を抽出する．

S4 発話区間による調整
発話検出器によって検出された発話区間が，区間の切れ目にならないように調整する．S3 の処理で切り出してきた 11 秒のシーンの前後 2 秒間において，発話がなされていない区間を探索する．発話がなされていない区間が複数ある場合は，最も長い区間を選択する．選択された非発話区間でシーンが開始あるいは終了するように，S3 の処理で切り出してきた区間を拡張する．なお非発話区間がなかった場合は，シーンを 2 秒延長する．

S5 オプティカルフローによるソート
各静的区間から得られたシーン群を，各シーンの $F(t)$ の平均値でソートし，小さい順に並べてダイジェスト候補として出力する（区間の開始位置を t_{start}，終了位置を t_{end} とする）．

$$Priority = \frac{\int_{t_{start}}^{t_{end}} F(t)dt}{t_{end} - t_{start}} \quad (3)$$

3.3.2 発話カット

T1 バンドパスフィルタ（BPF）で雑音除去
100–900 Hz の BPF を通すことで，人の声の周波数領域外の雑音を除去する．

T2 発話区間抽出
adintool を用いて音声波形データ中の発話区間の検出を行う [17]．

T3 シーン長による調整
adintool によって検出された発話区間に対して，隣り合った区間どうしの間隔が 1.85 秒以内なら，1 つの発話あるいは対話と見なし，区間を結合する処理を行った．1.85 秒というパラメータに関しては，文献 [20] において 1.85 秒の沈黙を発話や対話の切れ目としていたため，その値を参考にした．ただし，1 区間が 15 秒以上となる場合には，最も発話区間どうしの間隔が広い箇所から再帰的に二分していき，1 区間が 15 秒を超える区間がないようにする．そのうえで，1 区間が 11 秒未満のシーンを除去する．

T4 オプティカルフローによる調整
歩行シーンやあまりに映像揺れが大きいシーンを取り除くため，区間全体が静視カット (II)–(ii) で定義した静的区間にないシーンを除去する．

T5 音量変化の積分値でソート
より大きな声やより長く発話・対話した箇所の優先度が高くなるよう，以下の手順で優先度を定義した．

(i) 音声波形 $s(t)$ を 10 ms ごとに区切り，その区間の総和をエネルギー関数 $E(t)$ と定義する [21]．

$$E(t) = \int_{-5\mathrm{ms}}^{5\mathrm{ms}} s(t+\tau)d\tau \quad (4)$$

(ii) 環境音による影響を減らすため前フレームからの差分をとり，その絶対値を区間全体で足し合わせたものを，その区間の優先度と定義する（区間の開始位置を t_{start}，終了位置を t_{end} とする）．

$$Priority = \int_{t_{start}}^{t_{end}} |(E(t) - E(t-\Delta t))|dt \quad (5)$$

(iii) 各区間に対して優先度を計算し，大きい順に並べてダイジェスト候補として出力する．

4. 実証実験

本実験では，以下の 3 点を明らかにする．

(1)「自身での振返り」に適切なシーンと「他者との共有」に適切なシーンの相関関係．
(2) 発話カットと静視カットで，それぞれどのようなシーンが切り出されたのか．
(3)「自身での振返り」と「他者との共有」で，それぞれどのようなシーンが適切か．

4.1 実験手順

2015 年 11 月 12 日から 16 日の 5 日間開催されたメディアアートの展示会である，第 17 回東京大学制作展でユー

© 2016 Information Processing Society of Japan

図 3 実験時の様子
Fig. 3 Situation of the Experiment.

図 4 実験第 2 回時に使用した UI
Fig. 4 User Interface of the Second Experiment.

ザスタディを行った [22]．この展示会は，展示のほとんどが体験型であり，本システムの利用シーンと一致していたためユーザスタディの場として適していた．5 日間で計 25 名の実験参加者を募集し，実験を行った．実験協力者は男性 12 名，女性 13 名，平均 22.8 歳であった．撮影機材として，Panasonic 製のウェアラブルカメラ HX-A500 を用いた [23]．

実験は 1 人ずつ実施し，2 回に分けて行った．実験第 1 回は，ウェアラブルカメラを身に着けた状態で東京大学制作展の 9F 展示室の展示を体験してもらい，体験の様子をユーザ視点の一人称動画で撮影した．9F 展示室での実験の様子を**図 3** に示す．展示室には 9 つの作品があり，そのうちヘッドホンをかけて体験するなど，ウェアラブルカメラとの物理的干渉のため体験が難しかった 2 作品を除く 7 作品を体験してもらった．撮影前参加者には，この実験はミュージアムでの体験をダイジェスト動画にする際の最適な編集方法を調査するための実験であることが伝えられた．またミュージアムにおける理想的な鑑賞の流れとして，「見る」「考える」「話す」「聞く」の 4 つの行為が説明された．さらに「見る」場面を抽出するためにじっと見ているシーンを，「話す」「聞く」場面を抽出するために喋っているシーンをダイジェストにピックアップする仕組みであることが説明され，「作品をじっと見ることと，喋ることを意識すると良いダイジェスト動画となります」という教示が与えられた．

実験第 2 回は，第 1 回から約 1 週間後に行った．実験内容は，自身が撮影した動画から本システムで切り出した 20 シーンを見返し，以下の質問項目に回答するというものである．

Q1 各シーンに関して，A，B 2 つの観点から 5 段階で評価をお願いします（5：そう思う，4：ややそう思う，3：どちらともいえない，2：あまりそう思わない，1：そう思わない）
A. 自身の体験を思い出すために個人的に見る動画として適切か
B. 展示会の様子を他者に紹介するために一般公開する動画として適切か
Q2 各シーンに関して，シーンの種類をお答えください（複数回答可）
a. 作品を見つめているシーン
b. 作品に関して自身が発言しているシーン
c. 作品に関して他者が発言しているシーン
d. その他
Q3 各シーンに関して，シーンの良い点・悪い点をお答えください
Q4 A，B それぞれの用途に関して，今回提示したシーンを適切だと感じる順に並べ替えてください
A. 自身の体験を思い出すために個人的に見る動画
B. 展示会の様子を他者に紹介するために一般公開する動画

ユーザに提示した 20 のシーンは，発話カットアルゴリズムで切り出した上位 10 シーンと，静視カットアルゴリズムで切り出した上位 10 シーンを合わせて，時系列順に並べ替えたものである．なお，じっと見つめながら対話する場面など，発話カットと静視カットで似たシーンが切り出されることがあったが，その場合も各切り出されたシーンの絶対的な評価を得るために，ユーザには，同一のシーンが存在した場合は，それぞれ独立したものとして評価を行うよう指示した．

実験時に使用したユーザインタフェースを**図 4** に示す．シーン選択領域で項目を選択すると自動的にシーンが再生される仕組みである，ユーザはこの UI を用いて自身の動画から切り出された 20 シーンを視聴し，評価を行った．

4.2 結果

(1)「自身での振返り」に適切なシーンと「他者との共有」に適切なシーンの相関関係

2 つの動画使用用途によって好まれるシーンがどの程度異なっているのかを確認する．Q4 に関して，A の観点で

© 2016 Information Processing Society of Japan

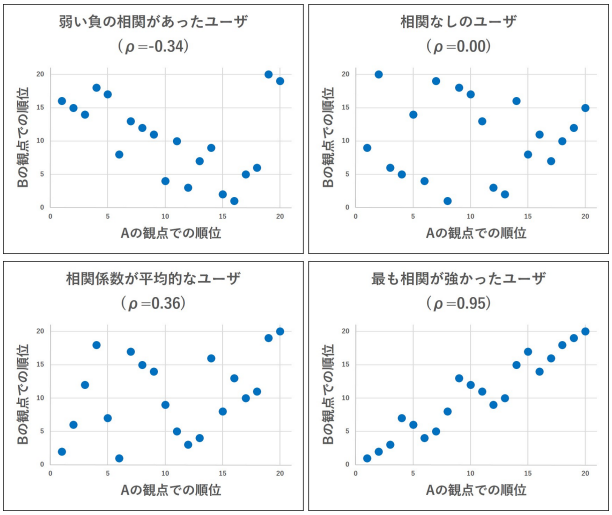


図 5 A の観点での順位と B の観点での順位のプロット図（代表的なユーザを抜粋）（A：自身の体験を思い出すために個人的に見る動画，B：展示会の様子を他者に紹介するために一般公開する動画）

Fig. 5 Prot of Rank Order in Application A and Application B.

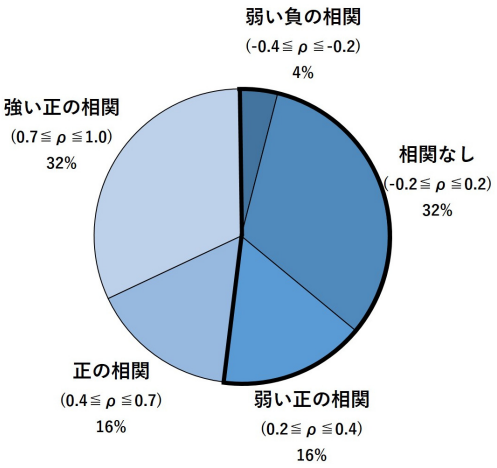


図 6 A の観点での順位と B の観点での順位のスピアマン順位相関係数検定結果のユーザ割合（A：自身の体験を思い出すために個人的に見る動画，B：展示会の様子を他者に紹介するために一般公開する動画）

Fig. 6 Participant Ratio of Spearman's Rank Order Correlation Coefficient in Application A and Application B.

の順位を横軸に，B の観点での順位を縦軸にプロットしたものを，代表的なユーザを抜粋して**図 5** に示す．さらに，A の観点での順位と B の観点での順位のスピアマンの順位相関係数を算出した．相関の度合いごとにユーザをまとめ，その割合を**図 6** に示す．52%のユーザが，相関が弱いか相関なしという結果となった．A，B 2 つの動画使用用途に対して，別々の編集を行うことが望ましいといえる．

(2) 発話カットと静視カットで，それぞれどのようなシーンが切り出されたのか

シーンカットアルゴリズムが適切に機能したことを検証

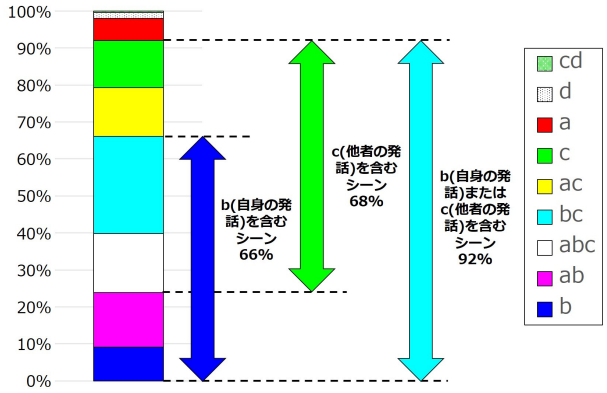


図 7 発話カットアルゴリズムで切り出されたシーンに対して，ユーザが回答したシーンの種類（a：作品を見つめているシーン，b：作品に関して自身が発言しているシーン，c：作品に関して他者が発言しているシーン，d：その他）

Fig. 7 The Relation of Cut-out Scenes by the Speech-Cut Algorithm and Participants' Scene Labeling.

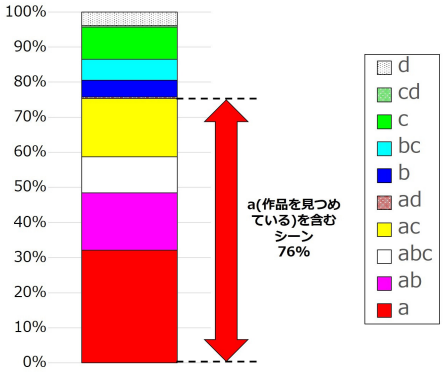


図 8 静視カットアルゴリズムで切り出されたシーンに対して，ユーザが回答したシーンの種類（a：作品を見つめているシーン，b：作品に関して自身が発言しているシーン，c：作品に関して他者が発言しているシーン，d：その他）

Fig. 8 The Relation of Cut-out Scenes by the Gaze-Cut Algorithm and Participants' Scene Labeling.

するため，発話カットで「見る」シーンが，静視カットで「話す・聞く」シーンが，それぞれどの程度抽出されていたかをユーザアンケートをもとに調査した．Q2 でユーザが行ったラベリングの割合を，2 種類のシーンカットアルゴリズムごとに**図 7**，**図 8** に示す．発話カットで切り出されたシーンのうち，「b. 作品に関して自分が発言しているシーン」，「c. 作品に関して他者が発言しているシーン」のいずれかが含まれているのは，92%であり，カットが正しく機能していた，つまり「見る」シーンを抽出できていたことが分かる．一方 b，c が含まれなかったシーンは，ユーザの笑いや感嘆詞，周囲の作品のサウンドエフェクトなどのノイズが抽出された場面が多かった．また，静視カットで切り出されたシーンに関しては，「a. 作品を見つめているシーン」が含まれているものが 76%であり，こちらもおおむねカットが正しく機能していた，つまり「話す・聞く」シー

© 2016 Information Processing Society of Japan

表 1 ユーザが回答したシーンラベルに対する動画使用用途ごとの評価の分散分析結果（A：自身の体験を思い出すために個人的に見る動画として適切か，B：展示会の様子を他者に紹介するために一般公開する動画として適切か，a：作品を見つめているシーン，b：作品に関して自身が発言しているシーン，c：作品に関して他者が発言しているシーン）

Table 1 ANOVA of the Evaluation of Each Video Applications for the Participants' Scene Labeling.

			「A」評価		「B」評価		各種の*n*
a	*b*	*c*	*M*	*SD*	*M*	*SD*	
有	有	有	4.30	0.83	3.29	1.37	14
		無	4.04	1.02	3.10	1.32	96
	無	有	3.45	0.98	3.55	1.23	35
		無	3.53	1.11	3.03	1.29	55
無	有	有	3.93	1.11	2.96	1.32	78
		無	4.00	0.89	2.46	1.15	75
	無	有	3.11	1.10	3.23	1.37	81
		無	2.71	1.44	2.21	1.42	66
検定結果※							
*a*の主効果			12.11***		13.66***		
*b*の主効果			65.19***		*n.s.*		
*c*の主効果			*n.s.*		13.18***		
a×*b*			*n.s.*		*n.s.*		
a×*c*			*n.s.*		*n.s.*		
b×*c*			*n.s.*		*n.s.*		
a×*b*×*c*			2.86*		*n.s.*		

※数値は*F*値を示す。***：p<.001，**：p<.01，*：p<.05を示す。

ンを抽出できていたことが分かる．一方 a が含まれなかったシーンは，映像の中心に作品は映っているが，説明員との対話や体験に集中しており「見つめている」とユーザが判断しなかったと考えられるものが多かった．映像の中心に作品が映っていないシーンは全 250 シーン中 16 シーンで，6.4%であった．

(3)「自身での振返り」と「他者との共有」で，それぞれどのようなシーンが適切か

「自身での振返り」と「他者との共有」で，それぞれどのようなシーンが適切かを検討するため，Q2 でのラベリングと，そのラベルが付けられたシーンの Q1 での 5 段階評価の対応を見た．ユーザが各シーンに対して行ったラベリングの結果を，a 有り/a 無し，b 有り/b 無し，c 有り/c 無しのように，ラベル回答の有無で二値データ化し，これらの 3 つの変数（a の有無・b の有無・c の有無）を独立変数，ユーザが回答した A，B それぞれの観点からの 5 段階評価を従属変数とした，対応のない 3 要因分散分析を実施した．結果を**表 1** に示す．

まず，「A. 自身の体験を思い出すために個人的に見る動画として適切か」に関して，分散分析の結果，「a. 作品を見つめているシーン」と，「b. 作品に関して自身が発言しているシーン」の主効果（$p < .001$），および a × b × c（作品を見つめているシーン × 作品に関して自身が発言しているシーン × 作品に関して他者が発言しているシーン）の交互作用効果（$p < .05$）が示された．「a 有」「b 有」の方が評価が高いことから，自身の体験を思い出すために個人的にみる動画は，作品を見つめているシーン，作品に関して自身が発話しているシーンが評価が高い傾向にあることが明らかになった．

表 2 ユーザが回答したシーンラベルに対する，A の観点での順位づけと B の観点での順位づけの差分の分散分析（A：自身の体験を思い出すために個人的に見る動画として適切か，B：展示会の様子を他者に紹介するために一般公開する動画として適切か，a：作品を見つめているシーン，b：作品に関して自身が発言しているシーン，c：作品に関して他者が発言しているシーン）

Table 2 ANOVA of the Difference of Rank Order in Application A and Application B to Participants' Scene Labeling.

			順位の差		各種の*n*
a	*b*	*c*	*M*	*SD*	
有	有	有	1.45	5.56	14
		無	1.21	6.32	96
	無	有	-3.40	5.75	35
		無	-0.61	6.00	55
無	有	有	1.52	5.92	78
		無	4.34	6.61	75
	無	有	-2.67	5.81	81
		無	-0.29	5.84	66
検定結果※					
*a*の主効果			*n.s.*		
*b*の主効果			42.73***		
*c*の主効果			10.14**		
a×*b*			*n.s.*		
a×*c*			*n.s.*		
b×*c*			*n.s.*		
a×*b*×*c*			*n.s.*		

※数値は*F*値を示す。
***：p<.001，**：p<.01，*：p<.05を示す。

次に，「B. 展示会の様子を他者に紹介するために一般公開する動画として適切か」に関して，「a. 作品を見つめているシーン」と，「c. 作品に関して他者が発言しているシーン」の主効（$p < .001$）が示された．「a 有」「c 有」の方が評価が高いことから，展示会の様子を他者に紹介するために一般公開する動画は，作品を見つめているシーン，作品に関して他者が発話しているシーンが評価が高い傾向にあることが明らかになった．

なお，A と B を比較すると，全体的に B の方が評価が低い値となっている．この点に関しては，一般的に他者のための動画を共有するという行為があまりなされておらず，自身が撮影した動画を他者と共有したいという意欲が，自身の思い出を動画としてとっておきたいという意欲に比べてあまり高くなかったためではないかと考えている．

さらに，A，B それぞれの観点での違いを，同一シーンの評価の差から詳細に分析するため，Q4 での順位づけの値を間隔尺度と見なし，A 観点での順位を基準とした B 観点での順位（B 観点順位—A 観点順位）を従属変数，先述の検定と同様に abc の有無を独立変数とした，3 要因分散分析を実施した．結果を**表 2** に示す．順位は値が小さいほど評価が高いため，「順位の差」の値は A 観点での評価が高ければ大きくなり，B 観点での評価が高ければ小さくなる．つまり順位の差が正の場合は個人的に見る動画としての評価が高く，負の場合は一般公開用としての評価が高い．

分散分析の結果，「b. 作品に関して自身が発言している

© 2016 Information Processing Society of Japan

シーン」の主効果（$p < .001$）および「c. 作品に関して他者が発言しているシーン」の主効果（$p < .01$）が示された．「b 有」と「b 無」を比較すると，「b 有」の方が値が大きく，また，「c 有」と「c 無」を比較すると，「c 有」の方が値が小さい．つまり，A と B を比較すると，b は，A の方が点数が高く，c は，B の方が点数が高い傾向にあることが分かる．

4.3 結論

実証実験の結果，以下のことが明らかになった．

(1)「A. 自身の体験を思い出すために個人的に見る動画」，「B. 展示会の様子を他者に紹介するために一般公開する動画」という 2 つの動画使用用途に対して，別々の編集を行うことが望ましい．
(2) シーンカットのアルゴリズムは正しく機能していた．
(3) 動画の使用用途をプライベート・パブリックに分けて両者を比較した場合，自身の発言が含まれているシーンはプライベートな動画としての評価が高く，説明員の発言が含まれているシーンはパブリックな動画としての評価が高くなる傾向にある．

4.4 提案システムの限界

本研究で提案したシステムの限界として以下の点をあげる．

(1)「見る」「聞く」「話す」の抽出精度は 100%ではなく，アルゴリズムの改善などでさらに改善することが望ましい．
(2) 自動要約ができていない．
(3) 動画の意味や内容になどを加味した抽出ができていない．
(4) 撮影対象によっては抽出が困難な場合がある．

(1) に関しては，精度のさらなる向上が望ましいと考えているが，本研究においても発話カットは 92%，静視カットは 76%と高い精度を出すことができていた．しかし現状では抽出のみであり，本研究の最終的な目標である自動要約自体にはまだ到達していない．今後は，抽出した場面をどういった順番で並べるべきかなどの自動要約手法の開発を進めたい．また，その際に (3) の意味内容の解析も重要になると考えており，そういった文脈の解析技術も取り込んで，最終的な一人称鑑賞体験の要約を実現させたいと考えている．(4) に関しては，展示スペースが暗い作品の場合は静視カットで抽出できないなど，撮影対象の条件によって抽出が困難な場合があった．静視カットについては，加速度センサ内蔵のウェアラブルカメラを用い，加速度の値を用いてカットを行うなどの改善策があると考えている．

4.5 考察

「A. 自身の体験を思い出すために個人的に見る動画」の評価と，「B. 展示会の様子を他者に紹介するために一般公開する動画」の評価を比較した際に，A の評価は高いが B の評価が低いシーンがどういったものかを Q3 の回答に基づき考察した．結果として，それらは主に 2 つの種類に分類できることが分かった．

(1) 他者に共有することに抵抗のある，自分の声や行動が含まれている．
(2) 他者に紹介するには情報が不足している．

A 評価が 5 であるにもかかわらず，B 評価が 1 または 2 のものが，全 500 シーン中 42 シーンあった．そのうち (1) に該当するものが 24 シーン，(2) に該当するものが 8 シーンであった．

(1) に関しては，率直な感想が思わず漏れてしまった場面や，作品に対して否定的な意見を述べた場面，自身が大笑いしている場面，作品の操作に失敗している場面など，撮影者個人の振舞いが記録内容の中心となっているものであった．ユーザによる「シーンの悪い点」の記述には，「自分が話しているのがはずかしい」「ネガティブなコメントだけで，楽しさが伝わってこない」「（自分の）声が耳ざわり」「（展示作品の）タブレットの操作が下手」などがあった．

これらは，展示会のパンフレットや Web ページなどの公式情報からは得られない参加者の実際の姿であり，個人の体験の記録という側面だけでなく，展示会そのものの記録としても興味深いものになると考えている．一方で A も B も評価が 5 のシーンは，作品の様子や説明が過不足なく残っているものが多かった．筆者が確認したところ，客観的な作品の様子が中心で，主観的な意見がないシーンが多く，無難だが個性のない場面が多かった．

(2) に関しては，静かに作品を注視している映像が多かった．作品の動く様子が端的に分かる映像である一方で，作品に関する解説情報などが含まれていなかった．このため，実験参加者による「シーンの悪い点」の記述に，「作品の説明がないところ」や「説明の音声がないので他者には伝わりにくい」などの記載があった．これは，4.2 節 (3) で述べた，他者に共有する動画としては，説明員が発言しているシーンの方が評価が高いという結果と一致する．

今後は，Julius などの音声認識システムによって発話内容や発話者を認識し，発話内容のポジティブさや，笑い検出，発話の種類（意見・質問・感嘆詞など）の分類などの発話の意味内容に踏み込んだ場面抽出手法についても検討したい．

さらに，一人称視点動画は自身の顔が映らないこと，映像より音声の方に撮影者の個人性を高く感じていることが，今回のアンケートにおける自由記述から示唆された．今後は．この点をふまえて定量的な評価を検討したい．

© 2016 Information Processing Society of Japan

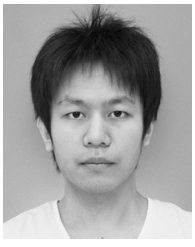
5. おわりに

体験型展示の鑑賞で説明員との対話がなされるような，インタラクティブな要素の多い作品展示における鑑賞体験を一人称視点で記録し，自身での振返りと他者との共有という2つの使用目的ごとに場面抽出する仕組みを提案した．ユーザスタディより，自身での振返りには，作品を見つめているシーンと自身が発言しているシーン，他者との共有には，作品を見つめているシーンと説明員が説明しているシーンが適切だとユーザが感じることが明らかになった．今後の課題としては，音声認識システムによって発話内容や発話者を認識し，発話の意味内容に踏み込んだ自動要約手法について検討したい．本稿では，動画の共有という観点で，動画を撮影した本人が他者と共有したいと感じる動画について調査した．今後は，動画の被共有者の視点で，どのような動画が評価されるかについても検討したい．また開発技術を応用し，個人の体験動画をから適切な抽出を行うことで，映画の予告編やレストランの口コミ情報のような他者との共有の価値をミュージアム体験にも広げ，鑑賞体験共有アプリケーションを実現したい．

謝辞 本研究の一部はJST CREST「共生社会に向けた人間調和型情報技術の構築」領域「局所性・指向性制御に基づく多人数調和型情報提示技術の構築と実践」による助成を受けた．

参考文献

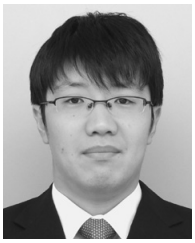
[1] 石島健一郎，相澤清晴：ウェアラブルによる長時間個人体験記録の編集：脳波を利用した映像の自動編集の試み，電子情報通信学会技術研究報告，パターン認識・メディア理解，Vol.100, No.565, pp.85–92 (2001).
[2] 上田隆正，天笠俊之，植村俊亮，吉川正俊：位置情報と地理情報を用いたウェアラブルカメラ映像のダイジェスト作成，情報処理学会研究報告データベースシステム（DBS），Vol.2001, No.70, pp.177–184 (2001).
[3] 長徳将希，小泉直也，苗村 健：ろぐろぐ動画：発話に基づく体験動画の自動要約，インタラクション2015 (2015).
[4] 河村俊哉，福里 司，平井辰典，森島繁生：ラリーシーンに着目した映像自動要約によるラケットスポーツ動画鑑賞システム，情報処理学会論文誌，Vol.56, No.3, pp.1028–1038 (2015).
[5] 栗原一貴：CinemaGazer：動画の極限的な高速鑑賞のためのシステムの開発と評価，コンピュータソフトウェア，Vol.29, No.4, pp.293–304 (2012).
[6] 三浦宏一，浜田玲子，井手一郎，坂井修一，田中英彦：動きに基づく料理映像の自動要約，情報処理学会論文誌コンピュータビジョンとイメージメディア（CVIM），Vol.44, No.9, pp.21–29 (2003).
[7] 角 康之，伊藤禎宣，松口哲也，シドニーフェルス，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).
[8] Blum, M., Pentland, A. and Troster, G.: InSense: Interest-Based Life Logging, *IEEE MultiMedia*, Vol.13, No.4, pp.40–48 (2006).
[9] Buschek, D., Spitzer, M. and Alt, F.: Video-Recording Your Life: User Perception and Experiences, *CHI EA '15*, pp.2223–2228 (2015).
[10] ソンヨンア，橋田朋子，筧 康明，苗村 健：Peaflet：ミュージアムにおける鑑賞体験を反映させた個人別リーフレット，情報処理学会論文誌，Vol.53, No.4, pp.1298–1306 (2012).
[11] Durrant, A., Rowland, D., Kirk, D.S., Benford, S., Fischer, J.E. and McAuley, D.: Automics: Souvenir Generating Photoware for Theme Parks, *CHI '11*, pp.1767–1776 (2011).
[12] 小関 悠，角 康之，西田豊明，間瀬健二：ぱらぱらアニメによる体験データの要約・編集支援システム，コンピュータソフトウェア，Vol.24, No.3, pp.41–50 (2007).
[13] Weilenmann, A., Hillman, T. and Jungselius, B.: Instagram at the Museum: Communicating the Museum Experience Through Social Photo Sharing, *CHI '13*, pp.1843–1852 (2013).
[14] 東京国立近代美術館スクール・プログラム，入手先 ⟨http://www.momat.go.jp/am/wp-content/uploads/sites/3/2015/01/schoolprogram_guide.pdf⟩.
[15] Yenawine, P.，京都造形芸術大学アートコミュニケーション研究センター：学力をのばす美術鑑賞ヴィジュアル・シンキング・ストラテジーズ：どこからそう思う?，淡交社 (2015).
[16] Julius, available from ⟨http://julius.osdn.jp/⟩.
[17] adintool, available from ⟨https://julius.osdn.jp/juliusbook/ja/adintool.html⟩.
[18] Instagram, available from ⟨https://www.instagram.com/⟩.
[19] Liu, C., Yuen, J., Torralba, A., Sivic, J. and Freeman, W.T.: SIFT Flow: Dense Correspondence Across Different Scenes, *ECCV '08*, pp.28–42 (2008).
[20] 宮田章裕，林 剛史，福井健太郎，重野 寛，岡田謙一：思考状態と発話停止点を利用した会議の動画ダイジェスト生成支援，情報処理学会論文誌，Vol.47, No.3, pp.906–914 (2006).
[21] Rabiner, L.R. and Sambur, M.R.: An algorithm for determining the endpoints of isolated utterances, *Bell System Technical Journal*, Vol.54, No.2, pp.297–315 (1975).
[22] 第17回東京大学制作展 (2015), 入手先 ⟨http://www.iiiexhibition.com/⟩.
[23] HX-A500 (Panasonic), available from ⟨http://panasonic.jp/wearable/a500/⟩.



長徳 将希

2014年東京大学工学部電子情報工学科卒業．同年より同大学大学院修士課程に在学．2014年よりミュージアムにおける体験拡張に関する研究に従事．

© 2016 Information Processing Society of Japan



小泉 直也 （正会員）

2012 年慶應義塾大学大学院メディアデザイン研究科後期博士課程修了．博士（メディアデザイン学）．日本学術振興会特別研究員 PD を経て，現在，東京大学情報学環研究員．知覚作用インタフェースやクロミック作用を利用したディスプレイの研究に従事．

苗村 健 （正会員）

1997 年東京大学大学院工学系研究科電子工学専攻博士課程修了．博士（工学）．米国スタンフォード大学客員助教授（日本学術振興会海外特別研究員）を経て，2013 年東京大学大学院情報学環教授．文部科学大臣表彰若手科学者賞，日本バーチャルリアリティ学会論文賞，グッドデザイン賞等受賞多数．

© 2016 Information Processing Society of Japan