ろぐろぐ動画: 発話に基づく体験動画の自動要約

長徳 将希[1,a)]　小泉 直也[1,b)]　苗村 健[1,c)]

概要：個人が撮影した動画は，そのままでは冗長で見返しづらいという問題がある．そのため撮影された動画のダイジェストを自動で作成する研究が多くなされている．個人的な体験という観点からすると，自身が発話した内容やその場の雰囲気が重要であり，その場の出来事の全体像を把握できる編集が求められる．そこで本稿では撮影者の発話に着目し，撮影された動画の中から撮影者の発話シーンをピックアップしたダイジェスト動画を，自動で作成するシステムを提案する．また撮影者自身が，発話がピックアップされることを意識しながらシステムを利用することで，撮影時に自然な発話が促され，その場の体験に積極的になれるという効果も期待される．撮影用のカメラは，発話を検出しやすく撮影時の負担の少ないウェアラブルデバイスを用いた．以上を踏まえ，東京大学制作展というメディアアート作品の展示会においてユーザーテストを行った．被験者に対して実施したアンケート結果から，自身の体験を見返す動画としての有効性が確認でき，また自然な発話を促す効果が示唆された．

LogLogVideo: Automatic Summarization of First-Person Video Based on User's Voice

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

Abstract: It is hard to look back raw videos. Therefore many studies have been made to create a digest video automatically. From the purpose of creating a digest video of personal experience, it is important to know the content of own speech, atmosphere of the place, and the overall picture of experience. So we propose a system for creating a digest video automatically by focusing on speech of the cameraman, picked up the cameraman speech scenes from the taken video. Additionally, being aware that the speech is picked up while utilizing the system, natural speech is promoted at the time of recording, the effect is also expected that become actively involved in experience of the place. Based on the above, we conducted a user test in exhibition of media art work called "iii-exhibition 2014" in the University of Tokyo Exhibition. According to the results of the questionnaire performed on the subject, effectiveness as videos for looking back their experience was confirmed, and the effect of promoting the natural speech was suggested.

1. はじめに

人は発話や身振り手振りなどの手段で自身の考えを伝える．しかしそれらは刹那的なものであり，そのままでは確かな形として残らない．そのような中で，ビデオカメラやスマートフォン，ウェアラブルカメラが個人の所有物として一般に流通し，発話などを含めたその場の状況をそのままの形で保存することが容易に行えるようになった．とりわけウェアラブルカメラは，一人称視点の映像を記録でき，個人の思い出を残す際に有用である．

ところが個人で撮影した動画は，その場の状況をありのまま残すためしばしば冗長で，見返す気が起きづらいという問題が指摘されている [1], [2], [3]．見返しやすい動画にするには，どのような場面が重要なのか見極め，適切に編集することが求められる．

そこで本稿では，一人称視点で撮影された動画から発話箇所を抽出し，ダイジェスト動画に自動編集する仕組みを提案する．発話箇所を抽出する理由は主に二つある．第一に，個人的な体験という観点からすると，自身の発話した

[1] 東京大学
Uniersity of Tokyo
[a)] chotoku@nae-lab.org
[b)] koizumi@nae-lab.org
[c)] naemura@nae-lab.org

場面は大きな意味を持っており，重要と言えるため．そして第二に，撮影者が発話に対して，ダイジェスト抽出のトリガーとしての機能を意識することで，その場の体験に対して積極的になれるのではないかという狙いがある．そこで，発話が許容され，むしろ積極的な発話が推奨される，ワークショップや体験型展示のある博物館などを利用シーンに位置づけた．そして，そのような場面で本システムを利用した際に，ユーザーにどの様な変化が起こるのかをユーザーテストを通して明らかにした．

2. 関連研究

2.1 ウェアラブルデバイスを用いたダイジェスト作成

自動でダイジェスト動画を作成する試みは，ライフログやグループワークの分野などで多くなされている．その中でも，本システムのようにウェアラブルデバイスを用いて撮影された動画を，自動編集する研究が過去にいくつかある．

石島ら [1] は，ウェアラブルカメラで撮影した個人体験記録を，脳波を利用して自動編集する仕組みを提案している．また，上田ら [2] は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を提案している．角ら [4] は，赤外線 ID タグシステムを用いて体験を解釈し，ビデオサマリを自動で作成するシステムを提案した．Blum ら [5] が提案したが提案した Insense では，マイクによる音量情報及び加速度センサからの移動情報を用いてユーザーが「楽しい」と思った瞬間を抽出する．

2.2 一人称映像の利用

本システムでは，動画の中でもウェアラブルデバイスを用いて撮影された一人称視点の映像を取り扱う．

一人称視点映像を利用した最近の研究として，笠原ら [6] が提案した JackIn が挙げられる．彼らは，実環境にいる利用者の一人称視点の映像を，別の利用者が観測し状況を共有するシステムを提案した．また，一人称視点映像から空間をモデリングして，疑似的に視点外からの映像を提供することで映像酔いの問題を解決している．また，同様に映像酔いを解消する手法として，Foote ら [3] が提案した Hyperlapse がある．彼らの手法は，ウェアラブルカメラで撮影した映像を 3D 座標上に構成しなおし，その空間を滑らかに移動する仮想カメラのからの映像を合成することで酔いを軽減するものである．

2.3 本研究の位置づけ

[1], [2], [5] では，何気ない日常の中に潜むコミュニケーションや出来事に焦点を当て，ログとして残すことを目指している．一方本システムは，発話というユーザーが意識的に制御可能なものを利用することで，動画撮影時にインデントをつけられ，それを基準にダイジェストを作成する仕組みである．

[4] は，ユーザーがシステムを利用してビデオを撮影している際に，その場でユーザーの振る舞いがどのように変化するかについては言及していないが，本稿ではその点を明らかにする．

JackIn[6] 及び Hyperlapse[3] は，ウェアラブルデバイスで撮影した映像を，視聴する側がストレスなく見られるよう，見やすく編集する手法として本システムと共通している．彼らの手法では疑似的に撮影視点外からの映像を提供したり，仮想カメラからの映像を合成したりすることで映像酔いを解消し見やすい動画作りを実現していた．これらは，動画を空間的に編集し見やすさを向上させている．一方本システムでは，ダイジェスト動画の作成という時間的編集のアプローチで，見やすさを実現した．

3. 提案システム

本システムは，自身の体験動画を，手軽に見やすい動画に編集する仕組みである．その際撮影者の発話に着目し，発話シーンがピックアップされたダイジェスト動画を作成する．要点をまとめると，以下のようになる．

対象：動き，発話のある一人称視点の体験動画

撮影者や被写体に動きのある場面で，動画の有用性が発揮される．また，発話が推奨される場面を利用シーンとして位置付けている．例を挙げると，ワークショップや体験型展示のある博物館，動物園，遊園地，スポーツ観戦などである．

要件 1 ：撮影及び編集の手軽さ

第一に，撮影時の手軽さが必要である．本システムでは撮影に際して，可能な限り体験を邪魔しないウェアラブルデバイスを一つのみ使用することで手軽さを実現している．更に撮影中，ダイジェストにピックアップしてほしい場面をユーザーが指示する際，追加のデバイスを必要とせずハンズフリーで行える，声を出すという行為をトリガーとして用いることで，手軽さを実現している．

第二に，撮影したデータを編集する段階での手軽さが必要である．世間一般で動画の編集技術を持っている人は少ない．したがって，手軽に編集出来るように自動編集をするシステムを作成した．

要件 2：見やすい動画編集

撮影された映像をダイジェスト編集することで動画の冗長性を減らし，見やすさを実現する．

また，本システムではユーザーが発話しているシーンをダイジェストでピックアップされることを目標とした．その理由は以下の 4 点である．

1. 発話されたという重要性

発話をするということは，その瞬間に何か発話をしたくなるような出来事があったということであり，後に

思い出す際に重要なシーンである場合が多いこと．

2. 発話内容の重要性
   発話したシーンは，発話内容という情報が付加されているため情報量が多く，そのシーンを見ることで視聴している側が，撮影者がその時何を考えていたかが分かったり，その場の雰囲気を知ることが出来たりすること．
3. 発話のトリガーとしての機能
   本システムの「発話したシーンをピックアップする」という仕組みを撮影者が知っていることで，体験時に残したいシーンで積極的に発言し，そのシーンにインデントをつけられること．
4. 自然な発話を促す効果
   発話したシーンがピックアップされるということを撮影者が知っていることで，その場の感想や自身の考えを積極的に発言するようになり，その場の体験がより楽しくなることが期待される．

以上の理由から，撮影者の発話箇所を推定し，その場面がピックアップされたダイジェスト動画を自動編集するシステムを作成した．なお，動画撮影のデバイスは，比較的軽量で着脱が容易な Google Glass[7] を用いた．理由は二つあり，まず第一に撮影者の負担が小さくその場の体験に集中しやすいため．そして第二に，マイク位置が撮影者の口に近く，音量を解析することで容易に撮影者の発話区間が推定可能であるためである．

3.1 ダイジェスト動画の構成

ダイジェスト動画の長さは，[8] を参考に 3 分 (180 秒) と規定した．180 秒の構成は，

- 10 秒：動画の OP
- 140 秒：ピックアップされた場面の等倍シーン
- 30 秒：それ以外の場面の早送りシーン

となっている．ダイジェスト動画の模式図を図 1 に示す．時間のパラメータは，予備実験をもとに決定した．

編集の方針として，単に綺麗な絵が撮れた瞬間だけをつなげるのではなく，その場その時その人だけに起きた体験を，個人の体験の全体像が分かるように編集することを目指した．そのため，ピックアップされたシーン以外もカットすることなく，早送り編集をして時系列順に並べた．これにより実際に撮影者が体験した出来事の全体像を，ダイジェストを通してうかがい知ることができる．

3.2 ピックアップシーン選択の方針

処理の流れを説明する．

全体の流れ

(1) ピックアップ候補箇所の抽出：開始位置
(2) ピックアップ候補箇所の抽出：終了位置
(3) ピックアップ候補箇所を前後に 2 秒ずつ拡張

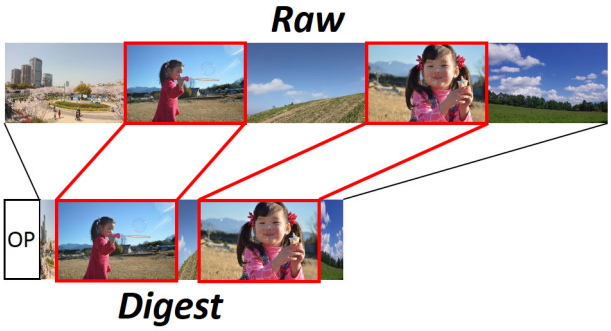

図 1 ダイジェスト動画の模式図

(4) ピックアップ候補箇所の取捨選択：最短時間の設定
(5) ピックアップ候補箇所の取捨選択：尺の調整
(6) 動画全体の尺調整
(7) 繋ぎ合わせ

個々の処理について説明する．

(1) ピックアップ候補箇所の抽出：開始位置

本システムの目標は撮影者が発話したシーンをピックアップすることである．ウェアラブルデバイスで撮影した動画では，装着者が発話を開始した箇所で動画の音量が大きく変化する．そこで，動画全体から音量変化の大きい瞬間の上位を選び出し，それらのシーンをピックアップシーンの開始位置の候補とする．なお，これらは候補にすぎず，この後の処理でそれらを選定するため，最初に選び出す候補数は十分大きければよい．後に示すユーザーテストでは，上位 100 箇所とした．また音量変化は，44.1Hz で音量データを区切り，各区間での平均音量 (dB) の変化量とした．

(2) ピックアップ候補箇所の抽出：終了位置

装着者が発話中の場面は，音量及びその変化が大きい．音量そのものをトリガーとして用いると，周囲の環境音に左右されやすくなってしまうため，音量の変化量を利用する方針を取った．ただし音量の変化量もそのままの値ではノイズが大きすぎるため，平滑化をするために音量の変化値の三角平均を求め，この三角平均の値が動画全体の音量変化の平均値を下回った瞬間に発話が終了したと判断し，終了位置と定めた．

(3) ピックアップ候補箇所を前後に 2 秒ずつ拡張

それぞれのピックアップ箇所の発話に対して，何に対して発話したのかや，発話後どの様な反応が周囲からあったのかが分かるように，発話箇所の前後 2 秒を各候補に追加した．

(4) ピックアップ箇所の取捨選択：最短時間の設定

シーンが頻繁に切り替わることからくる見づらさを回避するため，等倍シーンに最低必要な長さを定義し，それ以下の区間は等倍候補から除外した．具体的なパラメータは，各ピックアップシーンが最低 6 秒の尺を持つことを設計方針とした．

(5) ピックアップ箇所の取捨選択：尺の調整

ピックアップ候補箇所全体の音量変化値の合計を，その箇所の優先度と定義し，優先度の高い順にピックアップを確定させる．ピックアップ箇所の合計時間が 140 秒丁度になるよう，最後にピックアップを確定させる箇所の長さを調整する．

(6) 動画全体の尺調整

動画は OP が 10 秒，ピックアップされた等倍シーンが 140 秒，早送りのシーンが 30 秒である．早送りシーンが 30 秒丁度になるように早送りの速度を調整し，動画の尺が 3 分丁度になるようにする．

(7) 繋ぎ合わせ

最後に，編集されたシーン群を時系列順につなぎ合わせ，冒頭部分に OP を追加し動画を完成させる．

4. ユーザーテスト

2014 年 11 月 13 日から 17 日までの 5 日間開催されたメディアアートの展示会である東京大学制作展 2014[9] でユーザーテストを行った．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ーテストの場として適していた．5 日間で，計 66 名が本システムを利用した．

4.1 目的

本システムを利用することにより，自身の体験動画を，手軽に見やすい動画に編集するということが実現できたのかを検証するためにユーザーテストを行った．またそれと同時に，本システムを利用することで体験時の様子がどのように変化するのかを調査した．

4.2 実験環境

東京大学工学部二号館の 9 階 92B 教室を使用した展示室で実験を行った．本システムも展示会の作品の一部という位置づけであり，自身の体験の動画が自動でダイジェスト化され，それをお土産として持ち帰ることのできるシステムとして展示した．フロアには本システムを除いて 7 作品が展示してあり，被験者には可能な限りすべての作品を体験して貰った．

4.3 手順

4.3.1 実験開始時

本システムは展示会場の入り口に位置しており，システムの説明をした上で実験に協力して貰える人を募った．実験の前段階で説明した内容は以下である．

- 体験をしている間に撮影した動画を自動で 3 分のダイジェストに編集すること
- 作成したダイジェストはお土産としてもらえること
- 喋っているシーンをダイジェストで抽出すること

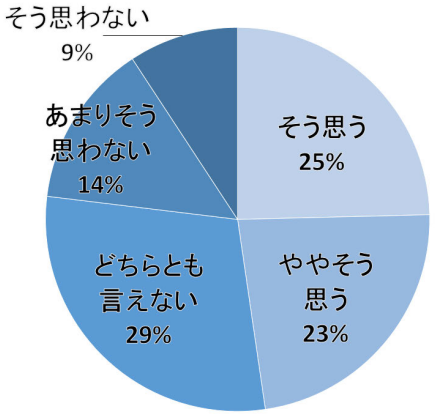

図 2 このシステムを利用することによって，普段より積極的に発言出来たと思いますか？

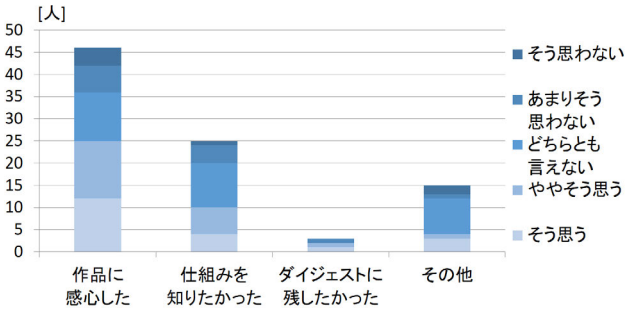

図 3 なぜ発言をしたのか．該当するものをすべて選んでください

- 採取したデータを研究に利用すること

説明の後に Google Glass を録画を開始した状態で被験者に渡し，その状態で展示を見て貰った．

4.3.2 実験中

制作展という展示会は，展示員が展示のそばにいて説明をする形を取っており，また展示自体も体験型の作品ばかりであったため，自然に発話をしやすい環境であった．

4.3.3 実験後

全ての展示を見終わり，録画を停止した後にアンケートを実施した．ダイジェスト動画の作成は，エンコード時間の都合などから 30 分程度の時間を要するため，動画の受け渡しはウェブベースで行った．また動画の公開に同意した被験者については，ダイジェスト動画を Youtube にアップロードすると同時に，制作展の Facebook，Twitter にそのリンクを貼り，誰でも閲覧可能な状態にした（66 名中 55 名が公開に同意した）．

加えて，完成したダイジェスト動画に関するアンケートを後日被験者全員にメールで送信し，31 名から回答を得た．

4.4 結果・考察

4.4.1 実験終了直後のアンケート

実験終了直後に行ったアンケートの結果を図 2,3 に示す．図 2 より，本システムを利用することで普段より発言が積極的になる効果がうかがえる．また図 3 から，発言する

	1位	2位	3位	スコア	人数
a	○	○	○	7	38名
b	○	○	×	6	8名
c	○	×	○	5	8名
d	○	×	×	4	0名
e	×	○	○	3	1名
f	×	○	×	2	0名
g	×	×	○	1	2名
h	×	×	×	0	0名
合計	54/57 (94.7%)	48/57 (84.2%)	49/57 (86.0%)	6.30(平均)	57名

図 4　特に発言していたと被験者が答えた展示がダイジェストでピックアップされていたか否か

動機は，単にダイジェストに残すために発言したのものではなく，作品に感心したり仕組みを知りたいと思ったりといった，その場の体験に付随した自然な反応であることが分かる．本システムを利用することで自然な発話が促され，体験そのものをより楽しむことのできる仕組みであることが示唆された．また，図 3 はそれぞれの回答に対して図 2 の解答をリンクさせて色分けされている．これを見ると，特に「作品に感心した」と答えた人は普段より積極的に発言出来たと答えた傾向にあることが分かる．

図 4 は，会場にあった 7 つの展示のうち，特にどの作品で発言したかというアンケートの結果と編集されたダイジェストとの対応を表したものである．被験者には自身が特に発言したと思う 1 位から 3 位まで，3 つの展示を選択してもらった．その結果に対して，個々のダイジェストを見返し被験者が選んだ展示がピックアップされていれば「○」，されていなければ「×」として，結果をまとめた．スコアは，1 位の作品が「○」ならば 4 点，2 位が「○」なら 2 点，3 位が「○」なら 1 点とし，その合計値を表したものである．なお結果は，全 66 名の体験者のうち，機器の不調等により最後まで録画が出来なかった 2 名，及びアンケートに未回答箇所があった 7 名を除いた 57 名のデータとなっている．

結果を見ると，1 位から 3 位までの全てがピックアップされていた a 群が最も多く 38 名であった．上手くピックアップされていなかった e 群や g 群に関しては，被験者が展示以外の場所で知人と雑談をして，そこが大量にピックアップされてきてしまっていたり，一つの展示に関してずっと話し込んでおりその作品ばかりがピックアップされたダイジェストになっていたりということが見受けられた．順位別に見ると，1 位と回答されたもののうち 95.2%がダイジェストでピックアップされてきていた．また 2 位，3 位もそれぞれ 85.5%，86.0%となっている．更にスコアの

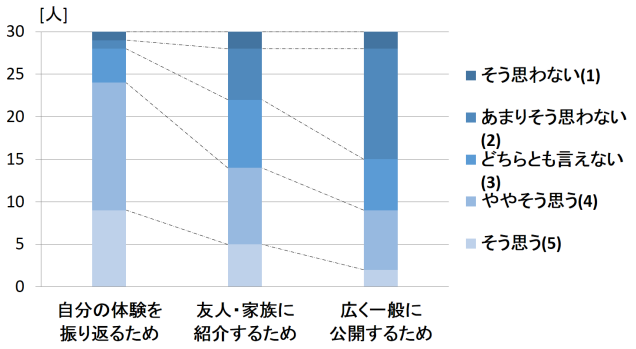

図 5　（それぞれの用途に対して）ダイジェスト動画として過不足なく編集されていると思いますか？

スコア (7点満点)	体験者全体人数	後日アンケート回答人数	後日アンケート回答者の用途別動画評価の平均		
			(1)自分	(2)友人・家族	(3)一般
7	38	19	4.2	3.3	2.8
6	8	3	3.0	3.0	2.7
5	8	3	3.7	3.7	2.7
3	1	0	—	—	—
1	2	1	5.0	5.0	5.0

図 6　スコア別人数と後日アンケートの評価

平均値も 6.30 と高い値となっており，概ね撮影者自身が発話したと思ったシーンのピックアップされたダイジェストとなっていたことが分かる．

一方，66 名全ての動画のピックアップシーンを見返し検証した結果，撮影者が発話していたシーンが 973 個，撮影者の傍にいた第三者が発話していたシーンが 23 個，展示の効果音によりピックアップされてきていたシーンが 2 個，トータルシーン数 998 個であった．つまりピックアップシーンのうち，97.5%が撮影者の発話していたシーンであった．3.2 章 (1),(2) で述べた発話区間の推定が上手く機能していたと言える．

4.4.2　後日実施したウェブベースのアンケート

後日，完成したダイジェスト動画に関するアンケートを実施し，全 66 名中 31 名から回答を得た．結果を図 5 に示す． この結果から，自身の体験を振り返る目的において本システムは有効であることが示唆される．しかしながら，広く一般に公開するための動画としては不十分であるという結果もうかがえる．これは，完成した動画が網羅性を欠いていることに起因していると考えられる．例えば，被験者が第三者と会話している場面で，被験者が発話している場面のみがピックアップされ，相手の発話シーンはピックアップされず，一方的な会話しかダイジェストに残らないといった問題である．解決案として，会話している相手もウェアラブルデバイスを装着し撮影者と同時に録画をすることが考えられる．これにより複数人の視点から発話箇所を切り出すことができ，それを統合することで会話シーンが網羅される．

図 6 は図 4 のスコアに対して，各スコア毎に人数と図 5 の評価の平均をまとめたものである．

スコアが最高の 7 点であった被験者は，自身の体験を振り返る動画としての評価が高い傾向にあることが分かる．逆に友人，家族，広く一般に公開するための動画としての評価は低い．これらは，先に述べた網羅性を向上させることで解決が見込まれると予想する．

4.4.3 自由記述や実験時の発言からの知見

ダイジェスト視聴後のアンケートでは，「完成したダイジェスト動画を見ることで，自身の体験が再構築されたような不思議な感覚があった」という意見が複数あった．自身の体験を振り返る際に，重要だと思われるシーンを動画というリッチな情報で提供することによる効果であると考えられる．

更に，「デバイスを装着していることを忘れるくらい体験に集中できた」という意見が複数あった，可能な限り体験の邪魔をしない設計が実現出来ていたと言える．しかしながら，眼鏡型の Google Glass というデバイスを用いたために，既に自身の眼鏡をかけている被験者はデバイスが気になり，デバイスの位置を調整するためにタッチパッドに触れてしまい体験の途中で録画が止まってしまうなどの問題も発生した．

またそれ以外には，「自身で見返す動画と他者に紹介する動画で異なる編集をした方が良い」という意見や，「動画に映っている作品の情報を提示してほしい」という意見が多かった．

5. まとめと今後の課題

ウェアラブルデバイスで撮影した動画を，撮影者が発話したシーンがピックアップされたダイジェスト動画に自動編集する仕組みを提案した．このシステムを用いてユーザーテストを行い，自身の体験を振り返るためのダイジェスト動画としての有効性が示唆された．更に，本システムを利用することにより自然な発話が促される効果も見受けられた．

一方，他者に紹介したり SNS などで広く一般に公開する目的においては，評価があまり高くなかった．その原因として，ダイジェストにする際に切り落としてしまう情報が多すぎることが考えられる．第三者に見せる動画にする際の課題と解決案をまとめた．

課題 1：会話の網羅性

現在のダイジェストでは，撮影者の発話のみがピックアップされてきているため，複数人で会話していた際に会話の一部しかピックアップされない．

会話の対象は多くの場合，展示員や被験者の同行者であった．そのため解決案としては，展示員と被験者グループの全員がデバイスを装着し撮影者となり，各々で撮影したデータを統合して編集することにより会話全体を網羅するという解決のアプローチが考えられる．

課題 2：撮影対象の説明

ピックアップシーンで見ている展示の説明がないため，何について会話しているのか分からないという問題がある．

解決案としては，動画撮影と同時に撮影者の位置情報を計測し，撮影場所の地図情報と照らし合わせて撮影者の見ていたものを推定することにより，各シーンに字幕を入れるなどの方法が考えられる．

また，これ以外にもダイジェスト動画の適切な長さについては議論の余地がある．

今後，システムを用いた際のユーザーの振る舞いの変化に着目しながら，様々な用途に合わせた適切なダイジェスト動画の形について検討していきたい．

6. 謝辞

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一脳波を利用した映像の自動編集の試み一”，PRMU， pp．85–92(2001)．

[2] 上田隆正, 天笠俊之, 吉川正俊, 植村俊亮: “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”, 情報処理学会第 125 回データベースシステム研究会報告， No．70(2001)．

[3] Johannes Kopf, Michael F. Cohen, and Richard Szeliski: “First-person hyper-lapse videos”, ACM Trans. Graph. 33, pp. 78:1–78:10(2014).

[4] 角康之 , 伊藤禎宣 , 松口哲也 , フェルス シドニー , 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).

[5] Mark Blum, Alex(Sandy) Pentland, and Gehrard Troster: “Insense : Interest-Based Life Logging”，IEEE MultiMedia, Vol.13, No.4, pp.40–48(2006).

[6] Shunichi Kasahara, Jun Rekimoto: “JackIn: integrating first-person view with out-of-body vision generation for human-human augmentation”, In Proceedings of the 5th Augmented Human International Conference (AH ’14). ACM, pp.46:1–46:8(2014).

[7] Google Glass, https://www.google.com/glass/

[8] Jonathan Foote, Matthew Cooper, and Andreas Girgensohn: “Creating Music Videos using Automatic Media Analysis”, Proceedings of the Tenth ACM International Conference on Multimedia(MULTIMEDIA ’02), pp. 553–560 (2002).

[9] 東京大学制作展 2014， http://www.iiiexhibition.com/