Vol.2017-CG-168 No.4
Vol.2017-DCC-17 No.4
Vol.2017-CVIM-209 No.4
2017/11/8

物体検出とユーザ入力に基づく一人称視点映像の高速閲覧手法

粥川 青汰[1,a)]　樋口 啓太[2]　米谷 竜[2]　中村 優文[1]　佐藤 洋一[2]　森島 繁生[3]

概要：本研究では長時間の一人称視点映像の効率的な早回し再生を目的として，ユーザが選択可能な手がかり（以下：キュー）の自動生成手法を提案する．一人称視点映像はウェアラブルカメラにより撮影される映像のことであり，両手が空いた状態で少ない負担で撮影可能であるため，今後記録を残す手段として普及していくことが予想される．しかしながら，常時撮影される一人称視点映像では長時間かつ冗長なシーンを多く含むため，その全てを閲覧することは困難である．そこで本研究では，映像を効率よく閲覧するためのキューを，コンピュータビジョン技術により映像から検出された物体名を用いて自動生成する手法を提案する．ユーザは提示されたキューを選択することで，意図したシーンを通常速度で再生して強調しつつ，他のシーンを高速再生することで，映像全体を高速に閲覧することが可能となる．既存手法が採用したキューと，提案手法が生成したキューをそれぞれ搭載した映像再生インタフェースを比較した評価実験の結果から，本研究が生成したキューを用いることで，より効率的に一人称視点映像から特定のシーンを発見可能であることを確認した．

1. はじめに

ウェアラブルカメラの小型化及び普及に伴い，一人称視点映像が撮影される機会が増加している．撮影者の頭部に装着して撮影される一人称視点映像を閲覧することで，撮影者がどこへ行き，何をしていたかなどの詳細な記録を，撮影者の目線を通して共有することが可能となる．さらに，両手が空いた状態で撮影可能な一人称視点映像は撮影時の負担が非常に少ないため，日常生活，レジャー，スポーツ，個人技能の解析など，様々な対象の記録を残す手段として，今後普及していくことが予想される．しかしながら，ウェアラブルカメラは常時撮影が基本であるため，長時間かつ冗長なシーンを多く含み，映像の閲覧に時間がかかるという問題点がある．そこで本研究では，一人称視点映像の高速閲覧を支援するインタフェースを提案する．

一人称視点映像を高速に閲覧するための既存研究は，(1) **自動要約システム** [4], [8], [15]：自動で重要なショットを選択し，それらをつなぎ合わせて短い映像を出力，(2) **高速再生システム** [3], [10]：ビデオ全体を倍速再生，という２つの手法に大別される．しかしながら，それぞれの手法には (1) 出力される映像にユーザーの意図を反映させることができず，ユーザが見たいシーンが出力映像から排除されてしまうリスクがある，(2) 映像全体が高速で再生されるため，映像の内容把握が困難であるといった問題点がある．

これらの問題点を解決する手法として，シーンごとに再生速度を変化させる高速閲覧手法が研究されている．これらの手法では重要なシーンを通常速度で再生し，その他のシーンを高速で再生することで，映像の重要なシーンに注目しつつ映像全体を高速で閲覧することが可能となる．その中でも，本研究では Higuchi ら [2] の伸縮タイムラインを用いた手法（EgoScanning) に注目した．Higuchi らは一人称視点映像を閲覧する手がかりとして Egocentric キューを導入した．Egocentric キューは Movement（移動），Stop（静止），Hand（手の動作），Person（人物との対話）という撮影者の基本的な行動に対応した 4 つのキューで構成されている．ユーザがそれらのキューの重要度を入力することにより，映像の中で注目部分と非注目部分が設定され，注目部分は通常速度で，非注目部分は高速に再生された短時間の映像が出力される．これにより，一人称視点映像から個人の関心の高いシーンを効率的かつ高速に発見し，閲覧することを可能となる．しかしながら，キューは上記の一定のものに固定されており，入力映像の内容を一切考慮していないため，システムが有効に働く入力映像が限定さ

[1] 早稲田大学
Waseda University
[2] 東京大学
The University of Tokyo
[3] 早稲田大学理工学術院総合研究所
Waseda Research Institute for Science and Engineering
[a)] k940805k.lab@gmail.com

Vol.2017-CG-168 No.4
Vol.2017-DCC-17 No.4
Vol.2017-CVIM-209 No.4
2017/11/8

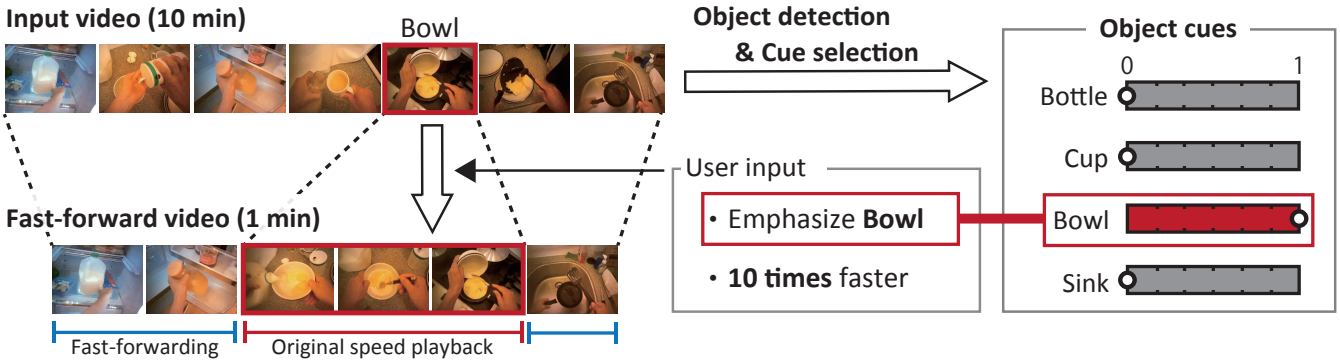

図 1 提案システムの概要図

れるという問題点がある．具体的な例として，ユーザが調理器具の使い方を学ぶために，プロの料理人が撮影した料理工程の一人称視点映像を閲覧する状況を考える．通常，料理工程を撮影した映像では図 1 にあるように，多くのフレームで撮影者の手が写り込む．そのため，Hand キューでは映像の大部分が強調され，特定のシーンに注目すること（ある特定の調理器具が映ったシーンに注目するなど）が難しい．また，閲覧する料理映像に人物との対話のシーンが含まれない場合，person キューは映像内に強調箇所が存在しないため，キューそのものが機能しない．

そこで本研究では，入力映像ごとにそれぞれ映像の内容を反映したキューを搭載するインタフェース（*Dynamic Object Scanning*(以下:DO-Scanning)）を提案する．提案手法では，入力映像の持つ意味的な情報を考慮する一つの手段として，コンピュータビジョン技術により映像から検出された物体名をキューの候補とする．ただし，単純に映像全体にわたって物体検出を行った場合，映像中に数フレームしか現れない物体や，逆に常時現れ続ける物体など，適応的な高速閲覧に必ずしも適さない物体がキューとして利用されうる問題があるそこで提案手法ではキューの有効度を評価する関数を導入することで入力映像に対して有効なキューを絞り込み，ユーザに “オブジェクトキュー” として提示する．オブジェクトキューは入力映像の内容を反映したキューであるため，これらのキューを操作することでユーザは，Higuchi らが採用した固定のキュー (撮影者の行動を指定する 4 つのキュー) だけでは強調できないような様々なシーンが強調可能となる．例えば，図 1 にあるようにユーザが Bowl キューを選択すると，ボウルが映ったシーンを強調した高速再生映像が出力される．

本研究では，DO-Scanning と EgoScanning[2] を用いて，様々なシーンで撮影された映像から特定のシーンを発見するタスクを与え，提案手法の有用性を検証した．実験を通して以下の 3 つの知見が得られた．

(1) **オブジェクトキューを提示することで，様々なシーンにアクセスすることが容易となる．** タスクの完了時間が短縮されたことと，参加者へのインタビューから，入力映像の内容を考慮して生成されたオブジェクトキューは，ユーザが映像の中から特定のシーンを発見する手助けになることがわかった．

(2) **ユーザは映像の一部分を強調するようなキューに有用性を感じる．** 提案手法で導入したキューの有効度評価関数では，映像の一部を指定するキューが採用され，映像の大部分を指定するようなキューは採用されない．ユーザからのフィードバックを通して，そのようなキューは長時間の映像から特定のシーンを発見する際に非常に役に立つということがわかった．

(3) **オブジェクトキューは映像内容の推定を容易にする．** ユーザは映像を閲覧する際，撮影者の動きよりも撮影された物体に注目しているため，映像内で撮影された物体名をキューとして提示することで，映像全体の内容やそれぞれのキューによって強調されるシーンの種類の推定を容易にすることがわかった．

2. 提案手法

本研究では Higuchi らの伸縮タイムラインの考え方に基づき，一人称視点を高速に閲覧するためのインタフェース（DO-Scanning）を提案する．ユーザは以下のように提案インタフェースを利用できる．（図 1 も参照）．まず初めに入力映像に対して，物体検出とキューの絞り込みを行うことで，入力映像の内容を反映したキューの組み合わせ（オブジェクトキュー）を生成する．ユーザは自分の関心のあるシーンに関連付けてそれぞれのキューの重要度を設定し，さらに映像全体を何倍のスピードで閲覧するかを設定する．これらの入力を元に各フレームの再生速度を計算し，ユーザに早回し映像を提示する．出力された映像では，ユーザが重要度を大きく設定したキューに関連したシーンが元のスピードで再生され，そのほかのシーンは高速に再生される．これにより，ユーザはキューを操作することで高速再生時に特に注目したいシーンを設定することが可能となる．

2.1 キューの設計方針

本研究では，一人称視点映像の高速閲覧に有効なキュー

Vol.2017-CG-168 No.4
Vol.2017-DCC-17 No.4
Vol.2017-CVIM-209 No.4
2017/11/8

を生成するために，以下の“キューの設計方針”を導入する．

(1) **セマンティックなキューを動的に生成：**様々な種類の映像に対して有効なキューをユーザに提示するため，全映像に対して同一の固定のキューを用意するのではなく，各映像に対して固有のものを用意するべきである．さらに，その際には映像の内容を考慮したセマンティックなキューを生成するべきである．例えば料理工程を撮影した映像と，散歩をする映像では撮影される場所（屋内か屋外か）や撮影される物体（調理器具か信号か）などが異なるため，それぞれに対応したキューを生成する．

(2) **少数かつ有効なキューの組み合わせを選択：**インタフェース上に大量のキューがあると，その中からユーザが好みのキューを選択する際の負担が大きくなるため，少数のキューを選択して提示する必要がある．さらに，キューを選択する際，(a) 映像中でごくわずかに登場する映像の要旨に無関係なキュー，(b) 映像内の大部分のシーンで登場するキュー（例：撮影場所など映像全体を通して変化しない情報），(c) ほかの物体と全く同じタイミングで登場するキュー（例：食事のシーンにおけるお皿とグラス）などはキューとして有効度が低いものであるため，これらを除外しつつ，有効な少数のキューを選択する．

以下では，この設計方針を元に提案インタフェースにおけるキューの自動生成手順について説明していく．

2.2 セマンティックなキューの動的生成

映像ごとに映像の内容を反映したセマンティックなキューを生成するために，我々は入力映像で撮影された物体をキューとして採用する．過去の映像要約 [4] やシーン推定 [5] などの研究において，物体検出は重要な役割を果たしている．そのため，映像で撮影された物体の一部をキューとして採用することにより，ユーザは提示されたキューから映像全体の内容（撮影した場所や撮影者の行動など）を推定し，かつキューを用いて映像中の特定のシーン（撮影者がある特定の物体を見ているシーンなど）に容易にアクセスすることが可能となる．

提案手法では，一般物体検出手法である YOLOv2[13] を用いて，毎フレームごとに物体検出を行なった．今回は COCO dataset[7] を用いて学習した計 80 種類の物体が検出可能なネットワークを利用した．ここで検出された物体名を提示するキューの候補とする．

2.3 少数かつ有効なキューの組み合わせの選択

一本の映像からは大量の物体が検出されるため，これらを全てをキューとしてユーザに提示すると先述の設計方針 (2) に反するため，少数のキューを選択する必要がある．また，単純に検出回数の多い順に複数のキューをユーザに提示した場合，それらのキューは物体は入力映像の情報を反映したセマンティックなキューであるが，先述したように高速閲覧に有効なキューの組み合わせとは限らない．そこで本研究ではそれぞれの物体の検出回数，映像全体における占有率，そして他のキューとのオーバーラップを考慮に入れ，有効なキューを選択するアルゴリズムを考案した．以下では図 2 を用いて提案アルゴリズムを説明する．ここでは，2 つのキュー（C_1，C_2）が与えられた時（この 2 つのキューの選択方法については第 2.4 項で説明する），新しいキューを C_3, C_4, C_5, C_6 の中から 1 つ選んで追加し，3 つの有効なキューの組み合わせを選択する例を考える．

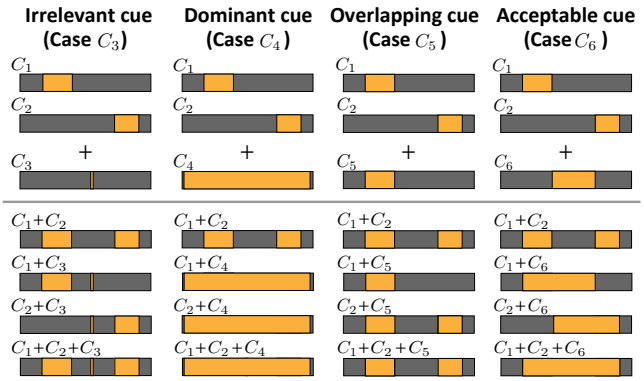

図 2 異なるキューを追加した時の強調可能箇所のバリエーションの違い．オレンジの領域はそれぞれのキューを個別もしくは複数設定した時に強調される箇所を示す．

Irrelevant cue：ごくわずかなシーンのみ強調するキュー
C_3 のように映像内のごくわずかなシーンのみ強調可能なキューは，映像の要旨に無関係なノイズである場合が多く，また C_3 を追加した 3 つのキューの組み合わせに対してどのキューを用いても強調できないシーンが映像内に多く存在してしまう．

Dominant cue：映像全体を強調する冗長なキュー 反対に C_4 のように映像の大部分のシーンを指定するキューは入力映像の内容を反映したキューである反面，ユーザが C_4 を選択した際に大部分が一様に強調された冗長な映像が出力されてしまう．

Overlapping cue：同じ箇所を強調するキュー また，C_5 のように映像の一部分を適度に強調するが，すでに選択済みの C_1 による強調箇所と同一の場所を強調する場合，ユーザが C_1 を選択した場合と C_5 を選択した場合で同一箇所を強調した映像が出力されてしまう．

Acceptable cue：最適なキュー これらの条件を満たさない C_6 のようなキューを選択した場合，得られた 3 つのキューを組み合わせることで，様々なパターンで映像の一部分を強調可能となる．

そのため，今回の例では C_6 が新たなキューとして追加される．提案手法では映像内で検出された全ての物体を追加するキューの候補とし，その中から上記にアルゴリズムに

Vol.2017-CG-168 No.4
Vol.2017-DCC-17 No.4
Vol.2017-CVIM-209 No.4
2017/11/8

従って最適なキューを 1 つ追加する作業を事前に設定したキューの個数まで繰り返すことで，最終的にユーザに提示する最適なキューの組み合わせを決定する．

2.4 キュー選択アルゴリズムの詳細

映像全体から検出された N 個の物体を $C_{\text{all}} = \{C_1, \ldots, C_N\}$ とし，そこからキューとして選択されたものを $C \subset C_{\text{all}}$ とする．さらにフレーム t において物体 C_n が検出された場合は 1，それ以外は 0 となるバイナリデータを $a_{n,t} \in \{0,1\}$ とする．この時，キューの組み合わせ C に対し，以下の式を用いてキューの有効度の評価関数を導入する．

$$F(C) = A(C) - B(C), \tag{1}$$

$$A(C) = \sum_t \left(1 - \prod_{\{m|C_m \in C\}} (1 - a_{m,t})\right), \tag{2}$$

$$B(C) = \max_{\{m|C_m \in C\}} \sum_t a_{m,t}. \tag{3}$$

$A(C)$ は C に含まれる物体の内，どれか 1 つでも検出されたフレーム数を計算したもので，映像全キュー体のカバー率を表す．反対に $B(C)$ は C に含まれる物体の内，最も検出回数の多い物体の検出フレーム数を計算したもので，1 つのキューの最大占有率を表す．第 2.3 項で説明したように有効なキューを追加する際は，$C_{\text{all}} \setminus C$ の候補の中から $F(C \cup \{c\})$ を最大にする $c \in C_{\text{all}} \setminus C$ を選択する．評価関数 $F(C \cup \{c\})$ を最大にする際，$A(C)$ の項を導入することで第 2.3 項の C_3 のような映像内でわずかしか登場しないキューや，C_5 のようにすでに選択済みの組み合わせ C と強調箇所が被るようなキューが排除される．一方で，$B(C)$ の項を導入することで C_4 のように映像の大部分を強調してしまうキューを排除することが可能となる．また，第 2.3 項の説明で最初に選択される 2 つのキュー (C_1 と C_2) は $F(C)$ を最大とするような 2 つのキューの組み合わせを C_{all} から全探索を用いて決定する．

3. 提案インタフェース（DO-Scanning）

本研究では伸縮タイムラインを用いた最新のインタフェースである EgoScanning[2] を下地にインタフェースを設計する．図 3 に提案インタフェースを示す．図 3 内の (A) は再生画面領域，(F) は他ビデオへのリンクとなる．提案アルゴリズムで選択された 10 個のオブジェクトキューが図 3 の (B) のエリアに配置され，ユーザはその中から自分が関心を持った物体名のキューを操作する．さらに映像全体を何倍のスピードで再生するかを (E) 再生速度設定スライダ用いて設定する．これらのユーザからの入力を元に各フレームの再生速度が計算され，(D) 伸縮タイムライン上に反映される．再生時には伸縮タイムライン上で赤くハイライトされた箇所でのみ通常速度で再生し，他のシーンを高速再生することで，オブジェクトキューで指定したシーンを強調しつつ映像全体を高速に俯瞰することが可能となる．また，操作したオブジェクトキューによって指定されたシーンが (C) オブジェクトタイムライン上でオブジェクトキューに対応した色でハイライトされる．

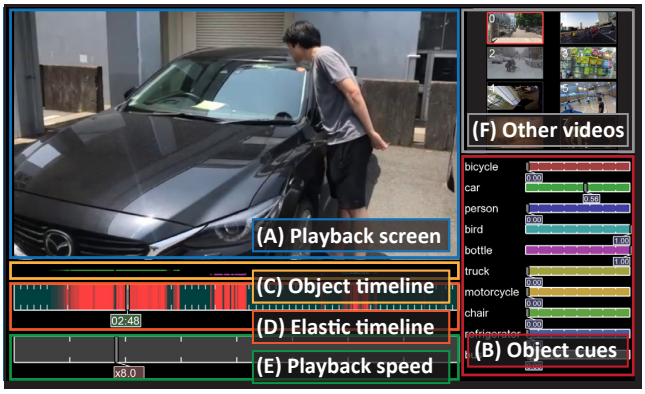

図 3 提案インタフェース（DO-Scanning）

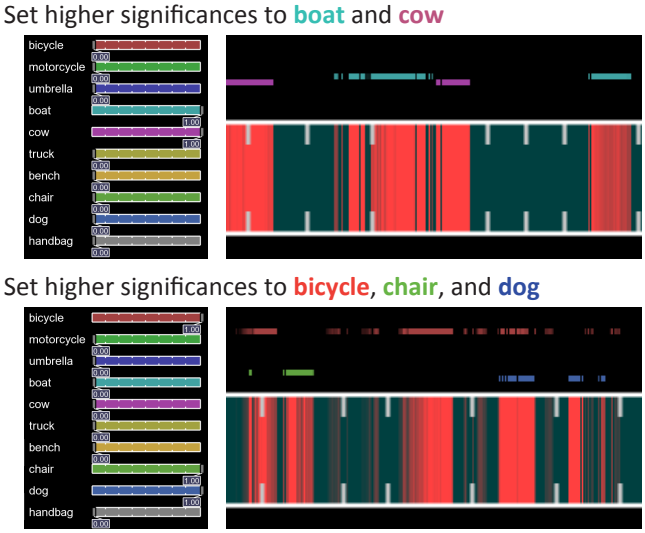

図 4 異なる (B) オブジェクトキューを設定した時の (C) オブジェクトタイムラインと (D) 伸縮タイムラインの例．

異なるオブジェクトキューを操作した時のオブジェクトタイムラインと伸縮タイムラインの結果例を図 4 に示す．オブジェクトキューの入力を変えることで映像内の異なる場所が伸縮タイムライン上でハイライトされ，さらにオブジェクトタイムラインを参照することで，どのタイミングで指定したどの物体が登場するかが一目で確認可能となる．

4. 評価実験

DO-Scanning の有用性を検証するために EgoScanning と比較実験を行なった．実験参加者は一般的な映像閲覧システム (YouTube など) の使用経験のある大学生 16 名である．

4.1 実験に用いたデータベース

実験では一人称視点映像が撮影される様々なシーン（公園の散歩，自転車レース [11]，市街地散策，買い物，犬に

	Scenario	Group	Length	Target scene	Target time	Source
Task 1	Strolling in the park	A	16:09	Recording birds	11:08	YouTube
		B	18:10	Walking near a car	14:05	
Task 2	Road race	A	26:05	Waiting at the traffic light	23:17	[11]
		B	26:05	Waiting at the traffic light	9:48	
Task 3	Strolling in the street	A	15:01	A bike cutting in front of the recorder	12:32	YouTube
		B	18:40	Arriving at the river	26:23	
Task 4	Shopping at a store	A	18:40	Taking a cup	13:05	YouTube
		B	17:58	Taking a bottle	9:36	
Task 5	Dog-centric videos (Videos recorded by a camera mounted on dogs)	A	9:08	Passing a car on the road	15:05	YouTube
		B	21:05	Taking a rest	4:50	
Task 6	Playing volleyball	A	9:08	Blocking a ball	6:10	YouTube
		B	14:21	Setting a ball	7:55	
Task 7	Cooking at home	A	13:27	Taking a bottle out from the refrigerator	10:48	[6]
		B	13:58	Returning eggs to the refrigerator	13:08	
Task 8	Playing in an amusement park	A	34:25	Buying a beverage	31:30	[1]
		B	34:45	Operating a cell phone	32:24	

表 1 評価実験に用いた一人称視点映像のデータセット．

図 5 それぞれの映像に対する，映像中のフレームの例，検出回数の多い物体名（Detected object），提案アルゴリズムで選択されたキュー（Seleted cues）．

装着した映像，バレーボール，料理 [6]，テーマパーク観光 [1] の計 8 種類）に合わせて映像を 2 本ずつ計 16 本用意した．映像の一部は既存データセットを利用し，残りの映像は YouTube 上から取得したもので，実験参加者が未閲覧のものを用意した．16 本の映像を表 1 のようにグループ A と B に分け，実験参加者は片方のグループを DO-Scanning で，もう片方のグループを EgoScanning を用いて閲覧した．

8 種類の映像それぞれに対して，映像中のフレームの例，検出回数が多かった物体名，提案アルゴリズムで選択されたキューをまとめたものが図 5 である．入力映像に合わせて異なる種類のキューがオブジェクトキューとして動的に生成され，提案アルゴリズムを用いることで，単純な検出回数の多い順とは異なるキューが選択された．例えば，‘person’ はどの映像に対しても検出されているが，映像 3（strolling in the street）と映像 8（playing in an amusement park）では映像の大部分で歩行者が撮影されるため，提案アルゴリズムでは ‘person’ は有効でないキューと判断され，選択されなかった．このように，提案アルゴリズムを用いることで映像全体を指定するような冗長なキューを含めずに有効なキューの組み合わせを選択することが可能となる．

4.2 タスク完了時間の評価

それぞれの一人称視点映像から顕著性の高いイベントシーンを 2 秒程度の映像で抜き出し，実験参加者に提示した．そして DO-Scanning と EgoScanning それぞれを用いて，提示したシーンを見つけるタスクを与え，その完了時間を測定した．今回，実験でユーザに提示した目的シーンを表 1 に示した．ある特定の物体に関係したシーン（ある物体を手にとったシーンなど）だけでなく，撮影者が特定の状況や場所にいるシーン（休息をとるシーンや川辺に到着したシーンなど）も目的シーンとして選定した．

さらに，タスク完了時間から平均閲覧速度を計算した．平均閲覧速度はインタフェースの高速閲覧性能を図る尺度として Higuchi ら [2] が導入したもので，目的シーンが映像内で位置している時間（表 1 の Target scene）をタスク完了時間で割ったものである．平均閲覧速度が大きいシステムほど，効率的に目的シーンを発見可能なシステムとなる．

２つのインタフェースを比較するために，DO-Scanning は EgoScanning よりも平均閲覧速度が高いという仮説をたて，平均閲覧速度に関する 95%信頼区間とマン・ホイットニーの U 検定を元にその仮説を検証した．

Vol.2017-CG-168 No.4
Vol.2017-DCC-17 No.4
Vol.2017-CVIM-209 No.4
2017/11/8

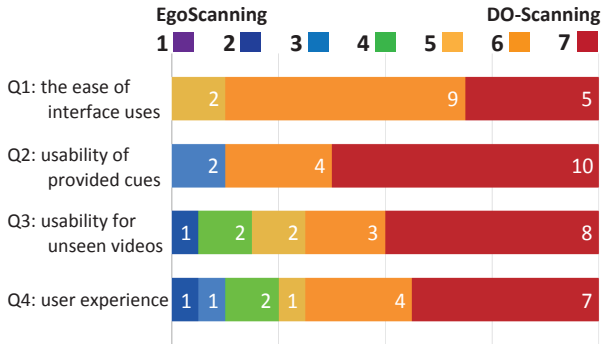

図 6 主観評価結果．暖色：Do-Scanning 優位の回答，寒色：EgoScaning 優位の回答．

4.3 インタフェースの主観評価

タスク完了後，実験参加者に 2 つのインタフェースの主観評価アンケートを行った．質問事項は以下の 4 つである．

Q1: どちらのインタフェースが使いやすかったか
Q2: 目的シーンに対してキュー選択はどちらが容易か
Q3: 初見の映像に対してどちらのインタフェースのキューが提示されると嬉しいか
Q4: どちらのインタフェースを使うのが楽しいか

両端をそれぞれのインタフェース（DO-Scanning を 7，EgoScanning を 1）とした 7 段階の評価軸を用意し，各質問がどちらのインタフェースの方に当てはまるか回答する形式で集計した．また，参加者に対して 10 分程度のインタビューを行い，ユーザがどのようにそれぞれのインタフェースをどのように使用したかを調査した．

5. 結果

5.1 タスク完了時間

2 つのインタフェースを用いて行った実験のタスク完了時間とそこから計算した平均閲覧速度の平均及び標準偏差，平均閲覧速度の 95%信頼区間，マン・ホイットニーの U 検定の p 値を表 2 に示した．95%信頼区間とマン・ホイットニーの U 検定を用いて検定を行ったところ Task6 を以外の全ての映像で DO-Scanning 優位の結果が得られ，Task6 では EgoScanning 優位の結果が得られた．

5.2 主観評価結果

主観評価の結果を図 6 に示す．各質問に対して，Do-Scanning 有意の結果を暖色で，EgoScanning 有意の結果を寒色で示した．全ての質問に対して過半数の参加者が DO-Scanning の方を高く評価した．

また，インタビューでは以下に示すように目的シーンを探す際，EgoScanning よりも DO-Scanning の方が使いやすかったという意見が多く得られた：**A1:**「目的シーンを探す際は撮影者の動きよりも撮影された物体に注目するため，適したキューをすぐ選択できた」，**A2:**「物体名がキューとして採用されている方が，どのキューがどのようなシーンを強調するかイメージしやすかった」，**A3:**「EgoScanning はキューが映像の大部分を強調してしまうことが多かったが，DO-Scanning の方がキューによって強調される範囲が限定されていたため使いやすかった」

また，オブジェクトキューに関して以下のような好意的な意見が得られた：**A4:**「自分の興味があるような物体名がキューとして提示されるとその物体が登場するシーンに注目したいと思う．一方，EgoScanning で提示されたキューは一般的なものであるため，キューが指定するシーンに興味を持たないと思う」，**A5:**「自分で撮影した映像であっても，自分が予想しない物体がキューとして提示されるとそれに注目して映像を見返したくなる．一方，EgoScanning のキューはありふれたシーンしか強調しないため，それらのキューに魅力を感じない」，**A6:**「町中で撮影した映像から “cow” キューが提示されるなど，普段撮影されないような物体がキューとして提示されると，インタフェースの魅力が増すと思う」．

一方，提示されたキューの個数に関しては賛否両方の意見が得られた：**A7:**「DO-Scanning は EgoScanning より提示された cue の数が多く，選択肢が多くてよかった」，**A8:**「DO-Scanning の方がキューの数が多く，適切なキューを探す際に手間取ったが，適切なキューを発見できればキュー自体の効果は大きかった」，**A9:**「キューの数が多く，種類も映像ごとに変化するため，欲しいキューを探す際に苦労した」．

また，Task6 に関しては DO-Scanning に対して否定的な意見が得られた：**A10:**「スポーツ映像のように撮影される物体が映像全体を通して変化せず，シーンが物体ではなく撮影者の動きで特徴付けられる場合はオブジェクトキューは有効ではなかった」，**A11:**「目的シーンに関係した物体が提示されたキューに含まれていないと，シーンを探す際に苦労した．バレーボール映像を視聴する際は “hand” キューが提示されると嬉しい」．

比較対象とした EgoSnanning については以下のような意見が得られた：**A12:**「EgoScanning はキューが指定するシーンの具体性に欠け，シーン特定の役に立たなかった」，**A13:**「キューが映像の大部分を強調する場合が多く，その場合出力される映像の再生速度にも変化がないため，使いづらかった」，**A14:**「あるシーンに物体があるか否かは一意に決まるが，そのシーンで撮影者がどのように動いていたかは一意に決まらないため，自分が意図したシーンを強調することが難しかった」，**A15:**「ある人物の一日を撮影した映像では，動いているか否かで一日の行動を大別できるため便利だと思う」．

6. 考察

評価実験を通じて DO-Scanning は最新の映像の高速閲覧インタフェースである EgoSncanning よりも有用であると

Task	DO-Scanning					EgoScanning [2]					ASS-p
	TCT (A)	TCT (B)	ASS	Lower	Upper	TCT (A)	TCT (B)	ASS	Lower	Upper	
Task 1	17.8 ± 4.1	45.8 ± 11.8	**29.8 ± 12.6***	23.6	36.0	57.8 ± 14.8	78.5 ± 9.6	11.7 ± 2.98	10.3	13.2	0.00000513**
Task 2	64.4 ± 17.4	56.8 ± 22.3	**18.1 ± 8.47***	13.9	22.2	154.5 ± 70.2	139.0 ± 76.2	8.31 ± 4.83	5.94	10.7	0.00301*
Task 3	35.1 ± 22.7	23.5 ± 8.26	**39.6 ± 23.3***	28.2	51.1	248.8 ± 139.5	57.5 ± 17.7	11.8 ± 9.71	7.01	16.5	0.0000409**
Task 4	32.9 ± 10.9	61.4 ± 20.4	**18.6 ± 10.8***	13.3	23.9	81.6 ± 35.5	104.3 ± 53.2	9.22 ± 4.62	6.95	11.5	0.00221*
Task 5	30.8 ± 11.8	57.4 ± 31.0	**22.8 ± 24.3***	10.8	34.7	91.6 ± 35.5	169.9 ± 74.7	6.92 ± 5.94	4.01	9.83	0.00280*
Task 6	218.4 ± 110.0	329.4 ± 149.9	2.03 ± 1.11†	1.48	2.57	128.0 ± 29.6	158.4 ± 64.7	**3.26 ± 1.00†**	2.77	3.75	0.00121†
Task 7	53.6 ± 9.58	53.4 ± 17.9	**14.3 ± 3.95***	12.4	16.3	76.3 ± 25.4	105.0 ± 24.8	8.78 ± 2.94	7.33	10.2	0.000212**
Task 8	48.1 ± 41.8	38.9 ± 26.2	**71.6 ± 53.9***	45.2	98.0	111.8 ± 51.9	76.4 ± 18.9	23.8 ± 8.46	19.7	28.0	0.000119**
Total	501 ± 176.0	666.4 ± 199.8	**13.7 ± 5.64***	10.9	16.5	949.9 ± 191.0	888.5 ± 7.84	7.84 ± 1.70	7.01	8.68	0.000138**

表 2 定量評価結果. TCT: データセット A と B に対するタスク完了時間 (task completion time). ASS: 平均閲覧速度（average scanning speed）の平均及び標準偏差 (∗ 95%信頼区間で有為差が得られた結果). Lower, Upper : 平均閲覧速度の 95%信頼区間の下限と上限. ASS−p: 平均閲覧速度に対するマン・ホイットニー U 検定の p 値 (∗ と ∗∗ はそれぞれ有意水準が 0.01 と 0.001 の時に DO-Scanning 優位の有意差が得られた結果, † は有意水準が 0.01 の時に EgoScanning 優位の有意差が得られた結果).

いう結果が得られた．実験結果から得られた 3 つの知見，問題点，そして今後の発展について以下で議論する．

6.1 得られた知見

(1) オブジェクトキューを提示することで，様々なシーンにアクセスすることが容易となる．DO-Scanning を用いることで EgoScanning よりも特定のシーンを発見するタスクの完了時間が大幅に短縮された．このことから，映像内容を考慮して生成されたオブジェクトキューはシーンにアクセスする際有効であることがわかった．また，インタビューからもオブジェクトキューが有用であるという意見が得られた（**A1**, **A3**, **A4**, **A5**).

(2) ユーザは映像の一部分を強調するキューに有用性を感じる．ユーザから DO-Scanning に関してキューが映像の大部分を強調しないため使いやすかったという意見 (**A3**) が得られ，EgoScanning に関して大部分を指定するキューは使いづらかったという意見（**A13**) が得られた．また，もし映像において撮影者の動きが変化していく場合は，EgoScanning のキューでも映像の一部分を強調可能となるため使いやすいという意見（**A15**）が得られた．これらのフィードバックから，キューの種類に加え，キューが指定する範囲もキューの有用性に影響することがわかった．町中を散歩する映像では “person” キューが選択されなかったように，提案アルゴリズムを用いることで映像の大部分を指定するような効果の小さいキューを取り除くことに成功した．

(3) オブジェクトキューは映像内容の推定を容易にする．映像を閲覧する際には撮影者の動きよりも撮影されたオブジェクトに注目するため，オブジェクトキューは強調されるシーンと結びつきが強く，キュー選択が容易だったという意見（**A1**, **A2**) が得られた．一方で EgoScanning で用意されたキューに関して，キューが指定するシーンに具体性に欠け，シーン特定の役に立たないという意見（**A12**, **A14**) が得られた．これにより，映像内で撮影されたオブジェクトから絞り込まれたオブジェクトキューは映像の内容を反映したセマンティックなキューとして機能し，映像全体の内容やそれぞれのキューによって強調されるシーンの種類の推定を容易にすることがわかった．

6.2 問題点と今後の発展

目的シーンにおいて，映像に特徴的な物体が登場せず，トスのシーンやブロックのシーンといった “撮影者の動き” で特徴づけられるシーンでは DO-Scanning は有効に働かなかった（**A10**, **A11**)．DO-Scanning では物体のみに注目し，動作の検出を行っていないため，動作に顕著性の現れる映像についてはあまり有効な結果は得られないことがわかった．提案アルゴリズムではキューの内容を考慮せず，指定するシーンの頻度やタイミングのみを考慮しているため，今後は検出物体と検出動作を合わせたキューの候補に提案アルゴリズムを適用し，最適なキューを提示するインタフェースへと発展させたい．

ユーザに提示するキューの個数もユーザの使いやすさに影響し，今回提示したキューの個数（10 個）に関しても賛否両方の意見（**A7**, **A8**, **A9**) が得られた．提案アルゴリズムでは決められた個数に対して最適なキューを決定するため，ユーザが指定した個数に合わせて最適なキューを提示するシステムに発展させることも可能である．

7. 関連研究

一人称視点映像を短時間で閲覧する一手法として自動要約システムがある．自動要約システムでは映像の中からシステム固有のルールに従って重要なショットを自動で検出し，要約映像を作成し，ユーザに提示する．ショットの重要度を判断する要因として，それぞれ人物 [4]，ストーリーライン [8]，注視点 [15] に注目した手法がある．これらは映像の概要を短時間で把握することが可能であるが，適用可能なシーンが各システムの定義した重要なシーンに限定

される．そのため，長時間かつ撮影されるシーンが多岐にわたる一人称視点映像において，ユーザが関心を持つシーンが排除されてしまう可能性がある．

映像を高速に閲覧するための別の手法として高速再生手法 [3], [10] も研究されている．これらを用いることで，映像全体を短時間で閲覧することが可能となるが，ユーザにとって重要なシーンも高速に再生されるため，ユーザが関心のあるシーンを見逃してしまう可能性がある．

これらの問題点を解決するために，再生速度をシーンごとに変化させる高速再生手法 [12], [14] が研究されてきた．これらを用いることで，映像の一部分に注目しつつ，映像全体を短時間で閲覧することが可能となるが，シーンごとの重要度の設定方法はシステムごとに決まっており，ユーザの意図を反映させられないという問題点がある．

一方でユーザの意図を反映可能な早回しシステムとして Higuchi ら [2] の手法がある．Higuchi らは一人称視点映像を閲覧する手がかりとして Egocentric キューを導入した．ユーザの設定したキューに応じてフレームごとの再生速度変化させることで，ユーザの意図を反映した早回し映像が出力可能となる．しかし，用意されたキューが固定されており，入力映像の内容を考慮していないため，システムが有効に働く入力映像が限定されるという問題点がある．

8. まとめ

本研究では物体検出結果とユーザ入力に基づいて再生速度を動的に変化させる高速閲覧インタフェースを提案し，その有用性を検証した．本手法の主なコントリビューションは，入力映像ごとに映像の内容を反映したキューを自動で生成するという点である．特定のシーンの発見のタスク完了時間と主観評価の比較結果から，入力映像の内容を考慮したキューを提示する提案インタフェースが，様々な種類の映像を高速に閲覧する際に有効であることを確認した．

DO-Scanning では撮影された物体に，EgoScanning では撮影者の動き，手，人物にのみ注目しているが，他にも既存のコンピュータビジョン技術を用いることで，撮影場所 [16], 注視点 [6], 動作 [9] などを検出することが可能である．今後は，それらに対して提案アルゴリズムを適用することで，より様々な種類の一人称視点映像から，特定の人物，場所，行動などのあらゆるシーンに注目可能なインタフェースに発展させていきたい．

謝辞 本研究は JST ACCEL (課題番号 JPMJAC1602) 及び，JST CREST (課題番号 JPMJCR14E1) の支援を受けた．

参考文献

[1] Fathi, A., Hodgins, J. K. and Rehg, J. M.: Social Interactions: A First-Person Perspective, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1226–1233 (2012).

[2] Higuchi, K., Yonetani, R. and Sato, Y.: EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines, *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*, pp. 6536–6546 (2017).

[3] Joshi, N., Kienzle, W., Toelle, M., Uyttendaele, M. and Cohen, M. F.: Real-Time Hyperlapse Creation via Optimal Frame Selection, *ACM Transaction on Graphics (TOG)*, Vol. 34, No. 4, pp. 63:1–63:9 (2015).

[4] Lee, Y. J., Ghosh, J. and Grauman, K.: Discovering Important People and Objects for Egocentric Video Summarization, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1346–1353 (2012).

[5] Li, L.-J., Su, H., Lim, Y. and Fei-Fei, L.: Objects As Attributes for Scene Classification, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 57–69 (2012).

[6] Li, Y., Fathi, A. and Rehg, J. M.: Learning to Predict Gaze in Egocentric Video, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3216–3223 (2013).

[7] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755 (2014).

[8] Lu, Z. and Grauman, K.: Story-Driven Summarization for Egocentric Video, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2714–2721 (2013).

[9] Ma, M., Fan, H. and Kitani, K. M.: Going Deeper into First-Person Activity Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1894–1903 (2016).

[10] Poleg, Y., Halperin, T., Arora, C. and Peleg, S.: EgoSampling: Fast-forward and stereo for egocentric videos, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4768–4776 (2015).

[11] Poleg, Y., Ephrat, A., Arora, C. and Peleg, S.: Temporal Segmentation of Egocentric Videos, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2537–2544 (2014).

[12] Ramos, W. L. S., Silva, M. M., Campos, M. F. M. and Nascimento, E. R.: Fast-forward Video based on Semantic Extraction, *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 3334–3338 (2016).

[13] Redmon, J. and Farhadi, A.: YOLO9000: Better, Faster, Stronger, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7263–7271 (2017).

[14] Silva, M. M., Ramos, W. L. S., Ferreira, J. P. K., Campos, M. F. M. and Nascimento, E. R.: Towards Semantic Fast-Forward and Stabilized Egocentric Videos, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 557–571 (2017).

[15] Xu, J., Mukherjee, L., Lo, Y., Warner, J., Rehg, J. M. and Singh, V.: Gaze-Enabled Egocentric Video Summarization via Constrained Submodular Maximization, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2235–2244 (2015).

[16] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A. and Oliva, A.: Learning Deep Features for Scene Recognition using Places Database, *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pp. 487–495 (2014).