一人称視点映像の高速閲覧に有効なキューの自動生成手法

粥川 青汰 * 樋口 啓太 † 中村 優文 * 米谷 竜 † 佐藤 洋一 † 森島 繁生 ‡

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んでおり，効率的な閲覧手法が不可欠となる．本研究では，入力の一人称視点映像中に現れるさまざまな物体を，閲覧のための手がかり（キュー）として利用し，ユーザの意図に応じて同映像を適応的に高速再生するインタフェースを提案する．具体的には，ユーザが各物体について重要度を，キューを用いて指定することにより，提案インタフェースは重要とされた物体の現れるシーンは通常速度で，それ以外のシーンは高速で再生する．評価実験の結果，既存インタフェースと比較して提案インタフェースが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 (頭部に装着されたカメラによって撮影した映像) が撮影される機会が増加している．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可能となる．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小さいため，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，今後普及して行くことが予想される．さらに，カメラ付きのスマートフォンの普及に伴い，誰でも気軽に写真が撮影可能となったことで，SNS において大量の写真が投稿，共有される時代が到来したように，今後，次なる撮影手段としてウェアラブルカメラが普及することで，一人称視点映像を共有し，閲覧する機会が増えていくことが予想される．

しかしながら，一人称視点映像では常時撮影が基本であるため，撮影者の視覚体験を記録可能である反面，長時間かつ冗長なシーンを含む場合が多くなってしまい，映像を閲覧するのに時間を要する問題点がある．こ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適応的な高速閲覧手法 [3], [11], [2] が研究されている．これらの手法では重要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．

特に本研究では Higuchi ら [2] の手法—EgoScanning に注目する．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．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る．ユーザはそれらのキューの重要度を設定することで，キューに対応したシーンを重要シーンとして注目しつつ，高速再生することが可能となる．しかしながら，キューが上記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

* 早稲田大学
† 東京大学
‡ 早稲田大学理工学術院総合研究所

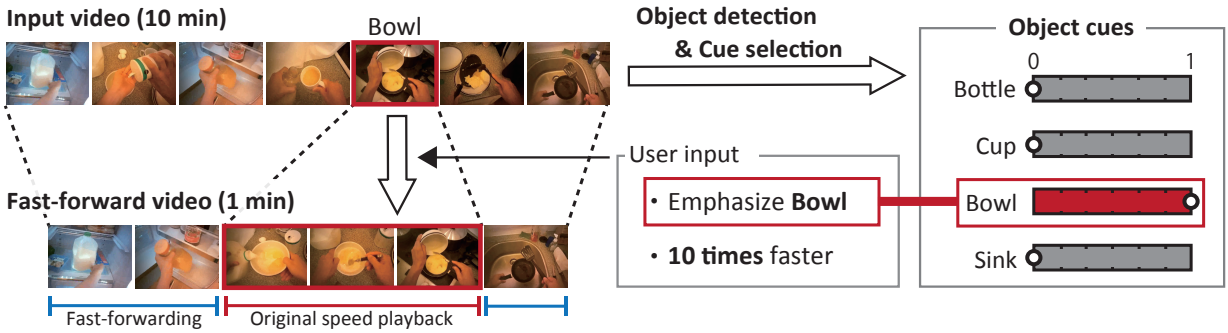

図 1. 提案システムの概要図

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．入力映像の内容を反映したオブジェクトキューを用いることで，Higuchi らが採用した固定のキュー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 の例では，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 を用いて，様々なシーンを撮影し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．実験から (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，(2) ユーザは映像の一部区間を強調するキューに有用性を感じる，(3) オブジェクトキューは映像内容の推定を容易にする，という 3 つの知見が得られた．

2 関連研究

Higuchi ら [2] の手法以外にも一人称視点映像を短時間で閲覧するための手法が研究されており，その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]，注視点 [12]，ストーリーライン [8] 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渡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，シーンの重要度に合わせて再生速度を変更する適応的な高速再生システムの研究がある．Silva ら [11] は映像中の人物や顔に着目し，それらが映ったシーンを重要シーンとする高速再生システムを提案した．しかしながら，このシステムでは高速再生時に注目するシーンが人物や顔など事前に固定され，強調して再生するシーンに対してユーザの意図を反映させることができない．

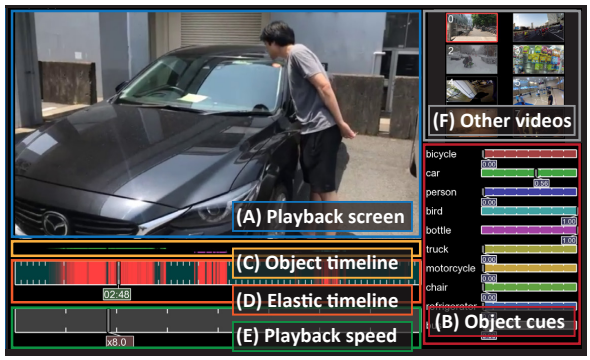

図 2. 提案インタフェース (DO-Scanning).

3 提案手法

本研究では，一人称視点映像の高速閲覧に有効なキューを，入力映像の情報を元に自動生成する手法を提案する．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)．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オブジェクトキューの重要度を設定し，さらに映像全体を何倍の速度で閲覧するかを設定する．これらの入力を元に，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速度で再生され，その他のシーンは高速再生された動画がユーザに提示される．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．

図 2 に提案インタフェース (DO-Scanning) を示す．図 2 内の (A) は再生画面領域，(F) は他ビデオへのリンクとなる．オブジェクトキューは図 2 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オブジェクト名のキューを操作する．操作したキューによって指定されたシーンが映像のオブジェクトタイムライン (C) 上で，それぞれのキューに対応した色でハイライトされる．これにより，映

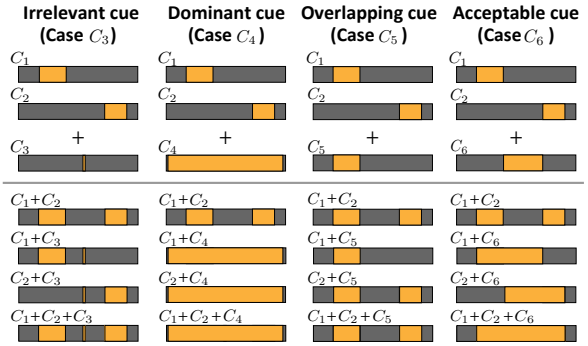

図 3. 異なるキューを追加した時の強調可能箇所のバリエーションの違い. オレンジの領域はキューを個別もしくは複数設定した時に強調される箇所を示す.

像全体のどのタイミングで関心を持ったオブジェクトが登場されるかが一目で確認可能となる．さらに映像全体を何倍の速度で再生するかを (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を通常速度で再生し，その他の箇所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．

3.1 オブジェクト検出

入力映像の内容を考慮したキューを生成するために，まずオブジェクト検出を行う．過去の映像要約 [4] やシーン推定 [5] などの研究において，オブジェクト検出は重要な役割を果たしている．そのため，映像で撮影されたオブジェクト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[10] を用いて，毎フレームごとにオブジェクト検出を行った．今回は COCO dataset[7] を用いて学習した計 80 種類のオブジェクトが検出可能なネットワークを利用した．ここで検出されたオブジェクトを提示するキューの候補とする．

3.2 有効なキューの選択

単純に映像で検出された全ての物体をキューとして利用した場合，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るうる問題がある．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設計する．以下では図 3 を用いて提案アルゴリズムを説明する．ここでは，すでに 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 3.3 項で説明する），ここに新しいキューを 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．

Irrelevant cue：ごく一部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区間を適度に強調するがその強調箇所がすでに選択済みの C_1 で強調される箇所と被っている場合，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

上記のような問題には該当せず，映像の一部区間を適切に強調できる 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 を新たなキューとして追加する．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，その中から上記のアルゴリズムに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．

3.3 キュー選択アルゴリズムの詳細

入力映像全体から検出された N 個の物体を $C_{all} = \{C_1, \ldots, C_N\}$ とし，そこからキューとして選択されたものを $C \subset C_{all}$ 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．

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

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

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

$A(C)$ 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の総数で，映像全体のカバー率を表す．この項を導入することで第 3.2 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．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で，1 つのキューの最大占有率を表す．こ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．第 3.2 項で説明したように有効なキューを追加する際は，$C_{all} \setminus C$ の候補の中から $F(C \cup \{c\})$ を最大にする $c \in C_{all} \setminus C$ を選択する．また，第 3.2 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．

30fps で撮影された 10 分の動画に対して，Intel Xeon E5-1650 v4, NVIDIA TitanX という環境でオブジェクトキューの生成を行なったところ，オブジェクト検出に約 15 分 (NVIDIA TitanX を使用)，有効なキューの選択に約 1.8 秒を要した．

4 評価実験

DO-Scanning が有効に働く映像の種類を調査するために，EgoScanning との比較実験を行なった．一人称視点映像は見回り（監視）や伝統技能保存，ライフログなどの幅広いアプリケーションがあり，それぞれの映像において重要なシーンが異なる．そこで今回は様々なシーンを撮影した一人称視点映像を用意し，その中から特定のシーンを DO-Scanning，又は EgoScanning を用いて発見するタスクを与えることで，DO-Scanning がどのような映像に対して有効であるか検証した．実験では図 4 に示した 8 種類のシーンをウェアラブルカメラを用いて撮影した映像を 2 本ずつ計 16 本用意した．映像の一部（映像 2[9], 映像 7[6]，映像 8[1]) 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．実験参加者は，YouTube などの一般的な映像閲覧システムを使用した経験のある大学生 16 名である．16 本の映像を映像の種類ごとにグループ A と B に，実験参加者をグループ X と Y に分け，グループ X はグループ A を DO-Scanning，グループ B を EgoScanning を用いて閲覧し，グループ Y は X とは逆のインタフェースを用いて閲覧した．

4.1 タスク完了時間の評価

それぞれの一人称視点映像から，特定のシーンを 2 秒程度の映像で抜き出し，実験参加者に提示した．そして各インタフェースを用いて，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．実験では，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，測定されたタスクの完了時間から “平均閲覧速度” を計算した．平均閲覧速度は，Higuchi らがインタフェースの高速閲覧性能を測る尺度として導入したもので，特定のシーンが映像内で位置している時間をそのシーンを発見するタスクの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，効率的に目的のシーンを発見可能なシステムとなる．8 種類の映像それぞれに対するタスクとそれらを合計したタスク全体における平均閲覧速度から 95%信頼区間を計算し，二つのインタフェースの結果を比較した．

4.2 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 目的のシーンに対してキューの選択はどちらが容易だったか
- 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体験だったか

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 評価関数 $I(C_n)$ によるキューの選定結果

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 4 である．映像の種類に合わせてそれぞれ別の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’ はどの種類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．

図 4. 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）．

表 1. 各インタフェースの平均閲覧速度及び 95%信頼区間 (∗ 95%信頼区間で有為差が得られた結果)

	DO-Scanning		EgoScanning[2]	
映像	平均	95%信頼区間	平均	95%信頼区間
映像 1	**29.8**	23.6〜36.0*	11.7	10.3〜13.2
映像 2	**18.1**	13.9〜22.2*	8.31	5.94〜10.7
映像 3	**39.6**	28.2〜51.1*	11.8	7.01〜16.5
映像 4	**18.6**	13.3〜23.9*	9.22	6.95〜11.5
映像 5	**22.8**	10.8〜34.7*	6.92	4.01〜9.83
映像 6	2.03	1.48〜2.57	**3.26**	2.77〜3.75*
映像 7	**14.3**	12.4〜16.3*	8.78	7.33〜10.2
映像 8	**71.6**	45.2〜98.0*	23.8	19.7〜28.0
Total	**13.7**	10.9〜16.5*	7.84	7.01〜8.68

5.2 タスク完了時間の評価結果

各タスクにおける平均閲覧速度の実験参加者平均と 95%信頼区間を表 1 に示した．映像 6 に対するタスクのみ EgoScanning の方が，他の全てのタスクは DO-Scanning の方が優位な有意差が得られた．

5.3 主観評価結果

主観評価の結果を図 5 に示す．各質問に対して，DO-Scanning 優位の回答を暖色で，EgoScanning 優位の回答を寒色で表した．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：「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，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く，適したキューをすぐ選択できた」，「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」「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」

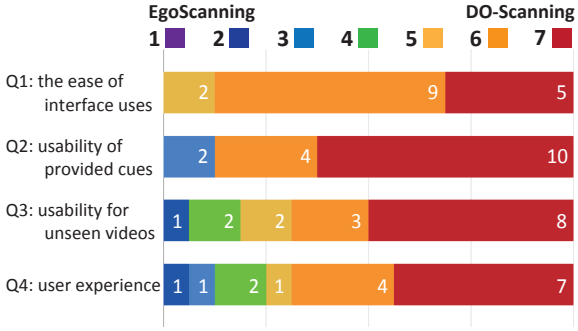

図 5. 主観評価結果．暖色：DO-Scanning 優位の回答，寒色：EgoScanning 優位の回答

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：「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」，

一方，映像 6 に関しては DO-Scanning に対して否定的な意見が得られた：「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」．

6 議論

実験結果からは以下の 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．多くの映像において，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，EgoScanning よりも多様なシーンにおいて，目的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

ンタビューからも目的シーンを探す際，オブジェクトキューが有用であるという意見が多く得られた．

(2) ユーザは映像の一部区間を強調するキューに有用性を感じる．ユーザから DO-Scanning は EgoScanning と異なり，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が得られた．このことから，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．第 5.1 項で述べたように，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入力映像の持つ意味的な情報を反映した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．

一方，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．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．

7 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．特定のシーンを発見するタスクの完了時間と主観評価の結果から，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，オブジェクトに限らないその他の検出結果も利用できる．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場所や行動などさらに広範囲の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] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social Interactions: A First-Person Perspective. In *In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'12)*, pp. 1226–1233, 2012.

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

[3] K. Kurihara. CinemaGazer: a system for watching videos at very high speed. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 108–115, 2012.

[4] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'12)*, pp. 1346–1353, 2012.

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

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

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

[8] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'13)*, pp. 2714–2721, 2013.

[9] Y. Poleg, A. Ephrat, C. Arora, and S. Peleg. Temporal Segmentation of Egocentric Videos. In *In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'14)*, pp. 2537–2544, 2014.

[10] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'17)*, pp. 7263–7271, 2017.

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

[12] J. Xu, L. Mukherjee, Y. Lo, J. Warner, J. M. Rehg, and V. Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'15)*, pp. 2235–2244, 2015.