Vol.2017-CG-168 No.3
Vol.2017-DCC-17 No.3
Vol.2017-CVIM-209 No.3
2017/11/8

大量な映像における高速な動的場面検索*

胡 晟[1,a)]　劉 健全[2,b)]　西村 祥治[2,c)]

概要：本論文では，膨大なカメラ映像から，「動的場面」を高速に検索できる技術を提案する．動的場面とは，人や物の間にインタラクティブな関係を持ち，動的な変化が起こるシーンである．既存技術は，人や物の特徴量を用いて該当の人・物を含んだ静的なシーンを検索することができるが，その関係が変化する動的なシーンを検索できない．この問題を解決するため，本稿は人と物の間にあるインタラクティブな関係を抽象化した述語モデルを提案して，提案モデルにより動的場面の検索を実現する．しかしながら，述語モデルを用いた動的場面の検索を従来のシーケンシャルスキャン法で行う場合は，処理時間が非常に長い問題が生じる．これに対して，本稿はさらに接尾辞木と転置索引を導入し，新たな索引構造を提案することにより，高速な検索を実現する．実験では，大量な実映像データを用いて検索効率および検索結果の精度を評価し，提案手法の有効性を示す．

キーワード：映像検索，動的場面，カメラ映像，述語モデル，インタラクション

1. はじめに

近年，ショッピングモール，ビル，駅，公共施設など至るところに防犯カメラが設置されるようになってきた．従来より，防犯カメラは，警備や捜査といったパブリックセーフティの用途として利用されている．その際，映像の確認や解析は，主に人手によって行なわれている．

しかしながら，昨今，防犯カメラの利用が普及しつつある中で，カメラが撮影した映像データも増大している．2020 年には，防犯カメラが撮影した映像が 5.6 ZB に達し，ビッグデータ全体の 42%を占めると予測される [1]．このため，人手による解析が困難になってきている．例えば，複数箇所に設置されたカメラの映像を対象に，人の目による確認作業では多くの時間が必要あるため，同じ場所に何度も出現する，あるいは複数の場所に出現する人物，さらに人と物の間にある動的なインタラクションなどの検索をおこなうのは，非常に困難である．

このような人的資源の限界を克服するため，映像に写った人，車両，物体をリアルタイムに解析する技術が開発されてきた [2], [3] が，多くの既存技術は典型的な映像解析，例えば，物体の認識，識別，追跡，行動およびイベント分析に留まり，大量なカメラを横断した映像検索技術は今なお欠けている [4]．

一方で，マルチメディアデータの複雑性が高いため，映像データは従来のテキストや数値データと異なり，検索処理は莫大な時間がかかり非常に困難である．さらに，映像に写った物体のセマンティックを分析して検索を可能にする自体はこの十数年の難題になっている．コンピュータビジョンの分野では，主に映像の内容を細かく分析することで，物体の動線データや，物体のモーションまたは外見を抽出して特徴量化する．映像検索をする際は，データベースに蓄積される物体の特徴量を入力データの特徴量と照合して類似した物体を含む映像を結果として返す．一般的に，このような映像検索を内容に基づく映像検索と言う．しかしながら，内容に基づく映像検索では，物体間のセマンティックを捉えることができず，物体間で起こるインタラクションも検索できない．

具体的に，例えば，図 1 に示すように，ある人が部屋に置かれた鞄を拾い去ったというセマンティックを持つシーンを大量の映像から探したい．この際は，従来の内容に基づく映像検索によれば，人と鞄の特徴を入力として検索をおこない，該当条件の結果が大量に返されることが多い．しかし，これらの結果には，必ずしも「鞄を拾い去った」というインタラクションを含むと限らない．逆に，検索目的に合わない結果が混ざっているため，このインタラクションが含まれたとしても見分けられない可能性が高い．特に，サーベイランス分野では，このような検索結果を見分けるには，膨大な時間を費やしてしまい，犯罪捜査活動を妨げる問題が生じる．

[1] 名古屋大学情報科学研究科
[2] 日本電気株式会社システムプラットフォーム研究所
[a)] hu@db.ss.is.nagoya-u.ac.jp
[b)] j-liu@ct.jp.nec.com
[c)] s-nishimura@bk.jp.nec.com

*本稿は NEC におけるインターン期間中の研究成果の一部．

Vol.2017-CG-168 No.3
Vol.2017-DCC-17 No.3
Vol.2017-CVIM-209 No.3
2017/11/8

図 1 動的場面の例： ある人が部屋に置かれた鞄を拾い去った[*1].

本研究では，前記の例示で述べた問題に対して，独自の映像検索コンセプトを提案し，高度なセマンティックを持つインタラクションの検索を実現した．本稿では，このような検索を新たなコンセプトとして「動的場面検索」を提案する．動的場面とは，映像に写った人や物の間にインタラクティブな関係を持ち，動的な変化が起こるシーンである．本稿は，動的場面を適切に表現するデータモデルと，動的場面の検索を高速に実現するアルゴリズムを中心に述べる．動的場面を表現するには，物体間のインタラクティブな関係を捉える必要がある．例えば，図 1 で示した人と鞄の空間関係が動的に変化している．最初は，人が鞄に向かって歩いて行く．人が鞄を拾ったあとに，鞄は人に付いて一緒に移動する状態に変わる．このような空間関係について，実は多様な映像の中にはさまざまな細かい変化が発生しうる．本稿では，Hall 氏が提唱した近接学の理論 [5] を参考し，細かな空間上の変化を大まかに 4 種類に抽象化した．接近，分離，同行，および静止という 4 つの空間関係を用いて物体間のインタラクションを表現する．図 1 で示した空間関係を抽象化すれば，人と鞄は最初に接近の空間関係から，次に同行の空間関係に変わった．そして，このような空間関係の変化を用いて物体間のインタラクションが抽象化され，適切に表現できる．さらに，物体間のインタラクションを用いて時間順に連続した変化が起こる動的場面を表すことができる．図 1 で示したような動的場面を表現すれば，人と鞄の間に発生したインタラクションの遷移を図 3 のように抽象化すれば表すことができる．

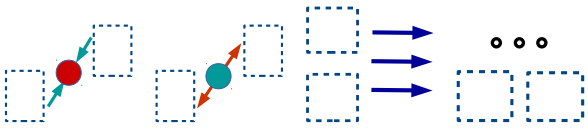

図 2 近接学の理論により抽象化した 4 種類の空間関係：接近，分離，同行，静止．

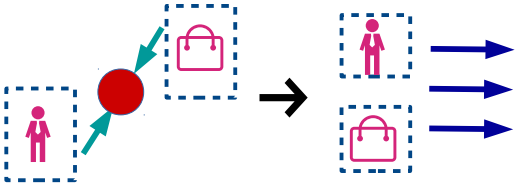

図 3 図 1 で示した動的場面の例を表す空間関係の連続した変化

このような人と物が連続なインタラクションが起こるシーンが，日常生活に非常に重要な行動意味が含まれている．人間の行動パタンの識別・分析の研究にも非常に役立てると思っている．

本論文では，上記の問題を解決するために，連続な動的場面を表現できるハイレベルなインタラクション表現を定義し，新たな述語モデルを提案した．その上に，接尾辞木と転置索引も導入し，高速な検索を実現した．また，手書きの線画・色により動的場面を入力できるため，より精度の高い検索を実現した．

本論文の構成は以下の通りである．2 章で問題設定を紹介する．3 章で関連研究を紹介する．4 章で解決手法を提案し，5 章でその提案手法を評価するために実験で提案の有効性を示す．最後に，6 章で今後の課題についてまとめる．

2. 問題設定

2.1 インタラクション

インタラクションについて，我々は空間上の動的関係 [5], [6] で，すべての人や物間のインタラクションを 4 つの述語に分類した．

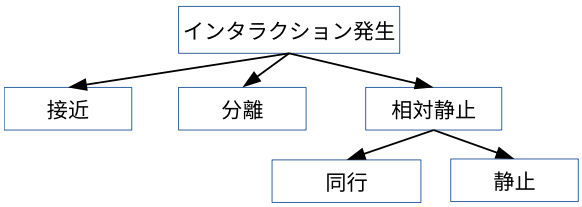

図 4 空間上動的関係による分類

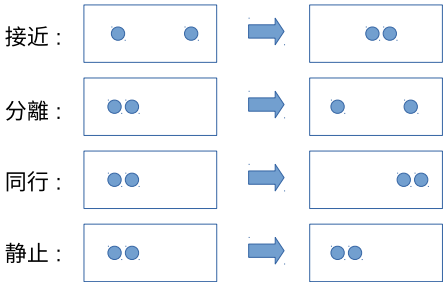

図 5 4 つの動的関係の例；接近：2 つの物体は距離を縮める；分離：2 つの物体は距離を置き続ける；同行：2 つの物体が距離を保ちつつ動く；静止：2 つの物体が距離を保ちつつ静止する．

図 4 と**図 5** はその分類の根拠である．空間上の動的な関係で，2 つの物体は起始状態から次の状態どのように変化するのかによって，図 5 のように 4 つの状況に分類した．一つ目の「接近」はその 2 つの物体が空間距離を縮めて近接関係に至る述語である．2 つ目の「分離」はその 2 つの物体が空間距離を置き続けて遠距離関係に至る述語である．3 番目の「同行」は 2 つの物体がずっとある固定距離を保ちつつ，同じ方向に同行する述語である．4 番目の「静止」は 2 つの物体が固定距離を保ちながらその自身も静止になっている述語である．本論文では，その 4 つの述

[*1] Youtube より映像引用．https://www.youtube.com/watch?v=clnLLF6kqvc

Vol.2017-CG-168 No.3
Vol.2017-DCC-17 No.3
Vol.2017-CVIM-209 No.3
2017/11/8

語をインタラクション述語と呼ぶ.

これらの 4 種類のインタラクション述語はそれぞれスタンプとして，簡便に入力できるようにしている．さらに，スタンプ中の dashbox に物体の外見を sketch として入力できる．これにより，それぞれのインタラクション述語に関わる物体をより詳細にすることができる．**図 2** に 4 種類のインタラクション述語を表すスタンプを示す.

2.2 動的場面

動的場面とは，人や物の間にインタラクティブな関係があることによって，動的な変化が起こるシーンである．1 章で示した「部屋に置かれた鞄を，ある人が拾って，鞄と一緒に去った」という例を 2.1 節のインタラクション述語で表現すると，**図 3** のようにになる．この動的場面は，2 つのインタラクション述語の順列として表現できる．まず、「部屋に置かれた鞄を，ある人が拾う」にあたるシーンが，「人が鞄に近づく」、すなわち、「接近-人-鞄」と表現できる．そして、その後の「鞄と一緒に去った」にあたるシーンが「同行-人-鞄」と表現できる．これら 2 つのインタラクション述語を図 2 のスタンプで表現すると，図 3 となる．ここで，各スタンプの並び順が各インタラクション述語の時間順を表している.

このように，動的場面は必ず時間的に連続して起こるため，一つの動的場面を一連のインタラクション述語の列として簡単に表現できる.

ちなみに，図 3 の示すように，各スタンプのついている dash box の画像がそのインタラクション述語と関連した物体を表している.

2.3 映像検索

本稿で提案する映像検索システムは，そのインタラクション述語モデルを使って問合せを構築し，大量な映像データの中から問合せとマッチした映像のセグメントを検索結果として返す．以下では，検索の入力および検索の出力について説明する.

本稿の検索入力は，図 3 の示すように，動的場面をスタンプの順列として表現する．そして，各スタンプの dashed box にユーザが手書きスケッチを入力し，物体の外見を表現する．その手書きスケッチを使って映像で出た物体の外見と類似度を計算し，ランキングを行う.

検索後に，問合せの動的場面と照合した映像のセグメントを類似度ランキング上位 K 件を結果として出力する.

システムの構成一覧は**図 6** となっている.

3. 関連研究

3.1 軌跡に基づく映像検索

軌跡に基づく映像検索は以下の 3 種類が知られている.

IMOTION

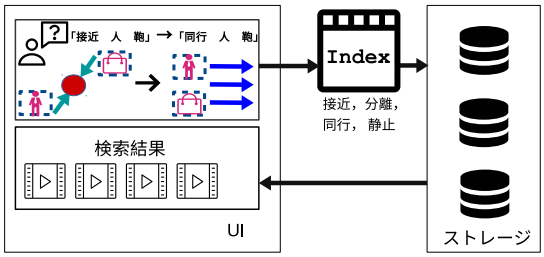

図 6 システム構成図

IMOTION システム [7], [8], [9], [10], [11] は Rossetto らによって開発されたスケッチに基づく映像検索エンジンである．Query-by-sketch, query-by-example, query-by-motion など複数の問合せパラダイムに基づいた入力が可能である． 特に，一つのモーション矢印を使い，その方向に動くものがあったすべての映像セグメントを検索することが可能である．IMOTION システムは映像検索システムである Cineast [12], [13] を元に開発された．IMOTION システムは映像の特徴量とメタデータを独自のデータベース ADAM [14], [15] に格納する．最近，Rossetto らは IMOION 最新版のマルチモーダルインターフェースである vitrivr [16], [17] を発表している.

ストーリボードによるモーションスッケチ映像検索

Collomosee らはストーリボードで手書きスケッチとして物体の簡略図と移動方向を線画と矢印で表現することで検索問合せを構成することを特徴とする，映像検索システム [18], [19] を提案している．確率モデルで映像クリップの特徴量を抽出し，スケッチの特徴量と照合する手法を用いている．彼らの研究は既存の軌跡照合映像検索研究 VideoQ [20] からの発想したものである．その後 R. Hu らはより性能が優れた trellis of space-time token 確率モデル手法 [21], [22], [23] を提案している．最後に, S.James ら [24] が特徴をベクトルに索引する手法を提案し，関連性フィードバック手法も加えることで，さらに効率を改善している.

軌跡映像検索

他の研究としては軌跡データのみを使い、スケッチを使わない映像検索がある．古典的な研究については，文献 [25] が詳しい．近年, K. Ghosal ら [26] は軌跡の定性特徴量を用いた映像検索手法を提案している．Lai ら [27] は 3-D の入力インターフェースを提案してより精確な問合せ軌跡を指定できるようにすることで、検索の精度の改善を図っている．Zhang ら [28] は軌跡をハイレベルな簡略図として表現するグラフモデルを提案している．Wu ら [29] はクラウドの環境でモーションレベルでの軌跡を問合せとする検索手法を提案している.

3.2 異常イベント検出

異常イベント検出分野において，多くの既存研究が提案されている．それらの研究は、主に人間行動検出と人間-オ

ブジェクトインタラクション検出の２つカテゴリに分けることができる．

人間行動検出

L. Kratz ら [30] は混雑な環境でローカルな時空間動線データを使って行動を検出する統計的なモデルを提案している．B. Ni ら [31] は人のグループ行動を３種類の因果関係で６つのカテゴリに分類している．その分類は，walk-in-group, run-in-group, stand-and-talk, gathering, fighting と ignoring である．Y. Benezeth ら [32] は背景差分から抽出した動線ラベルによるイベントモデリング手法を提案した．M. S. Ryoo ら [33] は２つの映像間の類似度を算出するために映像の時空間特徴量のモデリングとマッチング手法を提案した．彼らの実験結果は複雑な行動である push や hug などを検出できることを示した．Chang ら [34] は不審および挑発的なグループ行動を検出する手法を提案している．彼らの研究では凝集クラスタリングと決定クラスタリングなどの技術を用いている．Lan ら [35] はキーワードで問合せを構成し，行動映像のセグメントを検索する．また，SVM 技術によってその結果をランキングする．Cheng ら [36] はより少人数のグループに着目し，動線データと人間の外見データを使って細かい行動を検出する．Choi ら [37] はグループの行動を検出できる手法を提案した．D. Gowsikhaa ら [38] は防犯カメラ映像における人間とオブジェクトの動線データによる追跡や識別に関する技術の調査結果をまとめている．

人間-オブジェクトインタラクション検出

人間-オブジェクト間が関連する行動として，「置き去り」に着目した研究がいくつかある [39], [40]．Y. Yoshimitsu ら [41] はルールに基づく場面検出手法を提案し，人間-オブジェクト間の不審な行動を検出できることを示している．

3.3 手書きスケッチに基づく映像検索

ここでは動線データを使わずに，人や物の外見データだけを用いた映像検索を紹介する．

外見データによる映像検出

Yuan ら [42] はユーザの記憶による映像検索を提案している．彼らの研究ではスケッチ問合せ，テキスト問合せ，関連度フィードバックなどの要素を統合している．Huang ら [43] はテキスト問合せとスケッチ問合せの優劣を比較している．さらに，２つ種類を組み合わせた問合せについても評価しており，ユーザは様々な問い合わせを組み合わせることを好むことを報告している．

3.4 ビジュアル関係検出

Shang ら [44] はビジュアル関係検出の視点で，映像における物同士の間の述語関係で，検出する手法を提案している．我々の高度抽象な述語提案と違い，Shang らの研究は軌跡データを使って，物と物の間の詳細な述語関係の検出を目指している．このため，カメラの向きなど詳細に記述する必要があり，本稿の問題定義とは異なっている．また，Shang らの研究はリアルタイムで述語関係の検出を目指しており，この点に関しても本稿の研究目的である映像検索とは異なっている．

4. 提案手法

4.1 前処理

まずは，映像データに対する前処理の必要性について説明する．本稿が提案した述語モデルは高度に抽象化した特徴量であるため，映像データからより高速に検索できるデータ形式に変換しておく必要がある．つまり，検索はオリジナルの映像データでなく，映像データから抽出した特徴量を用いて実行する．それゆえ，特徴量抽出に関わった前処理も提案手法の重要な一要素である．

大量な映像から独自な特徴量を抽出するために，まずは，既存の state-of-the-art[45], [46] な識別・追跡手法によって，映像データを処理し，映像に物体を示すラベルと動線データをつける．そして，そのラベルと動線データを使って，我々の独自なアルゴリズム 1 でインタラクション述語に解析する．表示の便宜のため，それぞれのインタラクション述語を一つのタプルで表す．例えば，$interaction - sub_1 - sub_2, subimg_1, subimg_2, fbeg, fend, vid$ の場合，$interaction$ は述語の「接近」「分離」「同行」「静止」のいずれかを表し，$-sub_1 - sub_2$ はその行動に参加した物体は sub_1 と sub_2 であることを示す．$subimg_1, subimg_2$ はそれらの物体の画像特徴量を表し，$fbeg, fend, vid$ はその述語が行った映像のフレーム起始時間，終止時間と映像 id を表す．

Algorithm 1: ExtractDynamicSceneRecord ($frames$)

Input: A video clip $frames$
Output: A formatted dynamic scene record DS

```
DS ← ∅                                  /* a dynamic scene set */
for Each interaction ∈ frames do
    DS ← DS ∪ ⟨interaction − sub1 −
      sub2, subimg1, subimg2, fbeg, fend, vid⟩
return DS
```

4.2 索引構築

前処理の後，一つの特徴量レコードセット (以下略称レコードセット) が得られる．そのレコードセットを使って本稿の索引を構築する．本稿では，レコードを効率的に検索するために接尾辞木と転置索引を用いた．本稿の接尾辞木では，一つの節点は一つのインタラクション述語を表す．(表示便宜のため，それぞれのインタラクション述語をコードで表す．０は「接近」，１は「分離」，２は「同行」，３は

Vol.2017-CG-168 No.3
Vol.2017-DCC-17 No.3
Vol.2017-CVIM-209 No.3
2017/11/8

「静止」である.）このように表現することで，上から下にたどる木構造でインタラクション述語の時間順を表現することができる．また，接尾辞木を構造したため，問合せインタラクション述語との部分マッチングも可能となる．

接尾辞木が構築終わると，それぞれのリーフ節点に転置索引をつける．本稿の転置索引を用いることで，問合せに指定された物体が含まれたインタラクション述語の出現列を効率的に抽出することが可能になる．

次ははレコードのインタラクション述語を使って接尾辞木を構築する手順を，アルゴリズム 2 と 3 で示す．

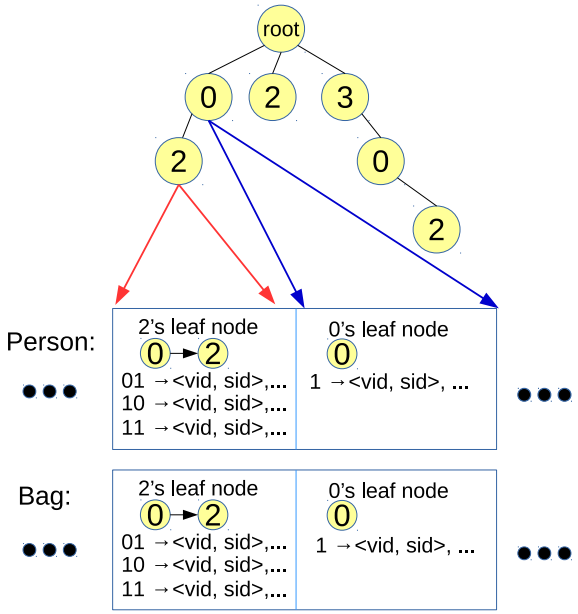

図 7 索引構造

Algorithm 2: BuildIndex (D, T)

Input: A record set D, a suffix tree root T
Output: A suffix tree rooted at T

```
1 for Each r ∈ D do
2     InsertDynamicSceneRecord(T, r)
3 return T
```

アルゴリズム 2 の中で，まずはレコードセットのすべての接尾辞を挿入する (Line 1).

アルゴリズム 3 の中で，一つのインタラクション述語を一つの節点として接尾辞木に挿入する (Line 2 – 7)．その後接尾辞木のそれぞれの節点に転置索引へのポインタを置いておく (Line 8)．その転置索引の構築はアルゴリズム 4 で示す.

アルゴリズム 4 の中で，まずは物体のバイナリコードを生成する (Line 2)．そのバイナリコードを用いることで，その物体はどのインタラクションに出現したかが分かる．例えば，物体「人」が 2 つのインタラクションで出現した

Algorithm 3: InsertDynamicSceneRecord (T, r)

Input: A suffix tree root T, a record r
Output: A suffix tree rooted at T

```
1 n ← T                                   /* the root of T */
2 for Each interaction ∈ r do
3     if n has a child n' through label interaction then
4         n ← n'
5     else
6         Create a child n' through label interaction for n
7         n ← n'
8 BuildInvertedList (n, r)
9 return T
```

Algorithm 4: BuildInvertedList (n, r)

Input: A suffix tree node n, a record r
Output: A suffix tree node n

```
1 IL ← ∅                          /* initilize inverted list */
2 for Each interaction ∈ r do
3     map['subject'] ←' binarycode'      /* binary code
        represents occurence */
4 if n is a new created node then
5     ia ← n's interval array
6     ia[subject] ← IL[subject]'s current length
7 lentry ← IL[subject]'s last entry
8 lentry.Insert(r)
9 return n
```

場合は，「人」を表すバイナリコードは 11 となる．そして節点からの区間ポインタを構築する (Lines 5 – 6)．最後は転置索引を更新する (Lines 7 – 8)．転置索引を更新する時に，posting list を挿入する．**図 7** のように，vacabulary は binary code で，posting list は挿入されるレコードの $\langle vid, sid \rangle$ である．vid は映像の id で，sid はレコードのフレーム起始時間，終止時間である $fbeg$ と $fend$ などの組合せである．つまり，$\langle vid, sid \rangle$ を用いることで検索結果となる映像のセグメントを特定できる.

最後に構築された索引構造は図 7 のようになる.

以下は一つの例を用いて図 7 を説明する.

Example1 **図 1** の例を見て，その映像の動的場面は「静止-人-鞄」，「接近-人-鞄」，「同行-人-鞄」と解析できる．この動的場面をアルゴリズム 1，2 で処理すると，図 7 のような索引になる.

4.3 結果検索

ここで検索アルゴリズムを説明する．まずは問合せにあった物体のバイナリコードを計算する (Line 1 – 2)．次にインタラクション述語の時間順を使って索引をたどる (Lines 3 – 9) ．そして，インタラクション物体の出現順で

Vol.2017-CG-168 No.3
Vol.2017-DCC-17 No.3
Vol.2017-CVIM-209 No.3
2017/11/8

転置索引で結果を絞り込み (Lines 10 – 12)，最後に処理時間が最もかかるスケッチと画像との類似度を計算し，ランキングを行う (Line 13)．スケッチと映像画像の類似度の計算 Sim(sketches,imgs) は state-of-the-art の [47] を用いる．

Algorithm 5: FetchResult (q, T)

Input: A query q, a suffix tree T
Output: A set of video record R

```
1  for Each interaction ∈ q do
2      map['subject'] ←' binarycode'
3  n ← T
4  for Each interaction ∈ q do
5      if n has a child n' through label interaction then
6          n ← n'
7      else
8          n ← null
9          break
10 for Each appeared subject ∈ q do
11     ILrange ← n[subject]
12     R ← R ∩ IL entry located in ILrange with subject's
         binarycode
13 TopkRank(R) with regard to Sim(sketches, imgs)
14 return R
```

以下は一つの例を用い、アルゴリズム 5 を説明する．

Example2 問合せ q は図 3 であるとする．索引は図 7 であるとする．q を解析すると，「0-人-鞄」→「2-人-鞄」になる．まずは図 7 の索引で 0 → 2 のラベルを辿り着く．問合せの「人」と「鞄」は 2 つのインタラクションで全部出たので，バイナリコードは 11 と 11 になる．赤い矢印が指した転置索引に入ると，そのバイナリコードを使って，”11” 番目のリストを取得してインターセクション操作をする．

5. 実験と評価

本章では，実映像データを用いて，提案する映像検索の性能を評価した．評価実験は，メモリ 12GB と Intel i7-6700 3.40GHz の CPU を持つサーバを用い，Ubuntu16.04 の OS 上で実行した．公開映像データセットを用い，映像検索効率と，結果の再現率と精度を評価した．

5.1 データセット

映像データの汎用性を配慮して，本実験では，最も一般的な公開映像データセットを使った．ILSVRC2016-VID[44], [48] から 85 件の映像データを選出した．**図 8** はデータセットの全貌である．85 件の映像データの詳細は **表 1** に記載した．

さらにその 85 件の映像の中で明確な意味を持つ 10 件の映像を問合せとして生成した．**図 9** は問合せの例である．

図 8 映像データ例

表 1 データセットの詳細

映像数	85
合計映像サイズ	1.1GB
最長フレーム	1615
最短フレーム	30
平均フレーム	149
fps	25

図 9 問合せの例

5.2 評価結果

本稿と類似の既存技術がないため，ベースライン手法として，映像セグメントを切り出してそれぞれの物体が揃ってるシーンだけを返すものを用いた．つまり，既存研究には，一個一個の単独の述語しか検出できないので，ベースラインのデータセットは，単独な述語が入ってる映像セグメントを単独に切り出して，単独な検索結果として返す．本稿は精度 (P) と再現率 (R) の評価を行った．精度 (P) と再現率 (R) の定義は以下となる．R は問合せ q による正解セットであり，A は提案手法による検索結果セットである．$R \cap A$ は 2 つセットのインタラクションとなる．$|R|$，$|A|$ と $|R \cap A|$ はそれぞれのセットのサイズを表す．Precision は $|R \cap A|$ と $|A|$ の比率であり，Recall は $|R \cap A|$ と $|R|$ の比率である．

$$Precision = P = \frac{|R \cap A|}{|A|}$$

$$Recall = R = \frac{|R \cap A|}{|R|}$$

再現率 (R) と精度 (P) の評価結果は以下**表 2** となる．すべての問合せの詳細は**表 3** となる．**表 4** は索引の効率を評価した結果を示し，**表 5** は問合せの詳細実行時間を示した．

Vol.2017-CG-168 No.3
Vol.2017-DCC-17 No.3
Vol.2017-CVIM-209 No.3
2017/11/8

表 2 評価結果

手法	P	R
ベースライン	35.3	92.5
提案手法	88.3	92.5

表 3 問合せ評価結果：精度と再現率

問合せ	P	R	$P_{baseline}$	$R_{baseline}$
q_1	2/2	2/2	2/4	2/2
q_2	2/3	3/4	2/5	3/4
q_3	1/1	1/1	1/3	1/1
q_4	1/1	1/1	1/7	1/1
q_5	1/1	1/1	1/5	1/1
q_6	1/1	1/1	1/1	1/1
q_7	1/2	1/1	1/5	1/1
q_8	1/1	1/1	1/4	1/1
q_9	1/1	1/1	1/4	1/1
q_{10}	2/3	2/4	2/5	2/4

表 4 効率評価結果：実行時間 (ms)

問合せ	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}
時間	28	26	20	21	20	15	20	19	26	24

表 5 問合せ効率評価結果：索引の性能

索引作成時間	2582 ms
平均問合せ時間	22.9 ms
索引サイズ	23.8 Mb

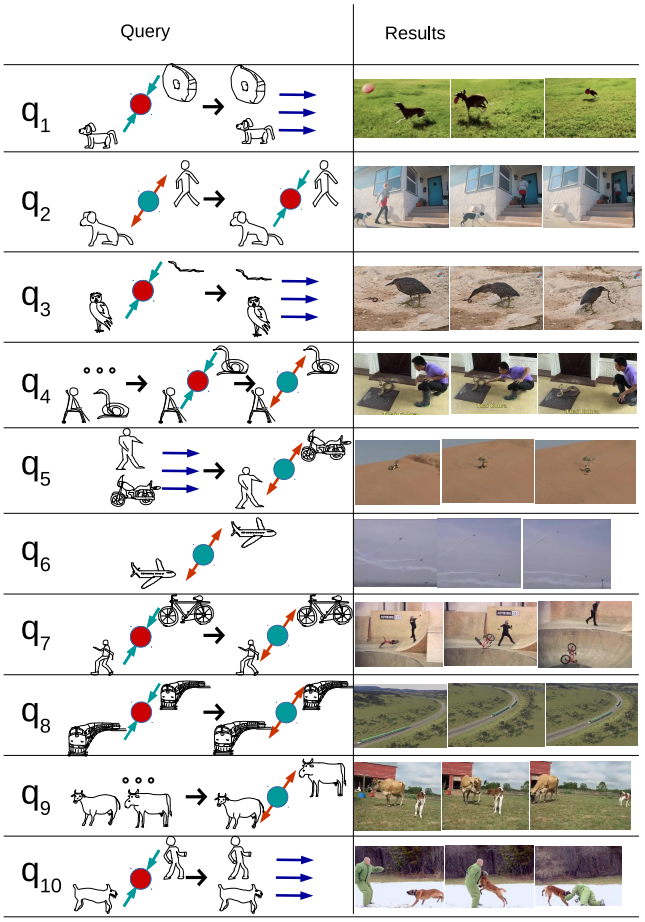

図 10 一部問合せの検索結果

6. まとめ

本稿では，防犯カメラ映像から，抽象化した述語モデルを用いて，犯罪に繋げる可能性の高い「動的場面」の検索を実現した．評価実験によってベースラーン手法より精度40%のアップを検証した．

今後の課題について，もっと大量な映像を用いた実験を実施する予定である．索引構造を改善し，多様な問合せに対する対応も考えられる．

参考文献

[1] Gantz, J. and Reinsel, D.: The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East (2012).

[2] Shah, M., Javed, O. and Shafique, K.: Automated Visual Surveillance in Realistic Scenarios, *IEEE Transactions on Multimedia*, Vol. 14, No. 1, pp. 30–39 (2007).

[3] Räty, T.: Survey on Contemporary Remote Surveillance Systems for Public Safety, *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 40, No. 5, pp. 493–515 (2010).

[4] Hu, W., Xie, D., Fu, Z., Zeng, W. and Maybank, S. J.: Semantic-Based Surveillance Video Retrieval, *IEEE Transactions on Image Processing*, Vol. 16, No. 4, pp. 1168–1181 (2007).

[5] Edward, T.: *Hall, The Hidden Dimension*, Garden City, NY: Doubleday (1966).

[6] Helbing, D. and Molnar, P.: Social force model for pedestrian dynamics, *Physical review E*, Vol. 51, No. 5, p. 4282 (1995).

[7] Rossetto, L., Giangreco, I., Schuldt, H., Dupont, S., Seddati, O., Sezgin, M. and Sahillioğlu, Y.: IMOTION—a content-based video retrieval engine, *MMM 2015*, Springer, pp. 255–260 (2015).

[8] Rossetto, L., Tănase, C. and Schuldt, H.: Dealing with Ambiguous Queries in Multimodal Video Retrieval, *MMM 2016*, Springer, pp. 898–909 (2016).

[9] Tanase, C., Giangreco, I., Rossetto, L., Schuldt, H., Seddati, O., Dupont, S., Altiok, O. C. and Sezgin, M.: Semantic Sketch-Based Video Retrieval with Autocompletion, *ACM IUI 2016*, ACM, pp. 97–101 (2016).

[10] Rossetto, L., Giangreco, I., Tănase, C., Schuldt, H., Dupont, S. and Seddati, O.: Enhanced Retrieval and Browsing in the IMOTION System, *MMM 2017*, Springer, pp. 469–474 (2017).

[11] Rossetto, L., Giangreco, I., Tănase, C. and Schuldt, H.: Multimodal Video Retrieval with the 2017 IMOTION System, *ACM ICMR 2017*, ACM, pp. 457–460 (2017).

[12] Rossetto, L., Giangreco, I. and Schuldt, H.: Cineast: a multi-feature sketch-based video retrieval engine, *ISM 2014*, IEEE, pp. 18–23 (2014).

[13] Rossetto, L., Giangreco, I., Heller, S., Tănase, C. and Schuldt, H.: Searching in Video Collections Using Sketches and Sample Images–The Cineast System, *MMM 2016*, Springer, pp. 336–341 (2016).

[14] Giangreco, I., Al Kabary, I. and Schuldt, H.: Adam-a database and information retrieval system for big multimedia collections, *IEEE Big Data Congress 2014*, IEEE, pp. 406–413 (2014).

[15] Giangreco, I. and Schuldt, H.: ADAMpro: Database

support for big multimedia retrieval, *Datenbank-Spektrum*, Vol. 16, No. 1, pp. 17–26 (2016).
[16] Rossetto, L., Giangreco, I., Tanase, C. and Schuldt, H.: vitrivr: A Flexible Retrieval Stack Supporting Multiple Query Modes for Searching in Multimedia Collections, *ACM MM 2016*, ACM, pp. 1183–1186 (2016).
[17] Goel, P., Giangreco, I., Rossetto, L., Tănase, C. and Schuldt, H.: "Hey, vitrivr!" –A Multimodal UI for Video Retrieval, *ECIR 2017*, Springer, pp. 749–752 (2017).
[18] Collomosse, J. P., McNeill, G. and Watts, L.: Free-hand sketch grouping for video retrieval, *IEEE ICPR 2008*, IEEE, pp. 1–4 (2008).
[19] Collomosse, J. P., McNeill, G. and Qian, Y.: Storyboard sketches for content based video retrieval, *IEEE ICCV 2009*, IEEE, pp. 245–252 (2009).
[20] Chang, S.-F., Chen, W., Meng, H. J., Sundaram, H. and Zhong, D.: VideoQ: an automated content based video search system using visual cues, *ACM MM 1997*, ACM, pp. 313–324 (1997).
[21] Hu, R. and Collomosse, J.: Motion-sketch based video retrieval using a trellis levenshtein distance, *IEEE ICPR 2010*, IEEE, pp. 121–124 (2010).
[22] Hu, R., James, S. and Collomosse, J.: Annotated free-hand sketches for video retrieval using object semantics and motion, *Advances in Multimedia Modeling*, pp. 473–484 (2012).
[23] Hu, R., James, S., Wang, T. and Collomosse, J.: Markov random fields for sketch based video retrieval, *ACM ICMR 2013*, ACM, pp. 279–286 (2013).
[24] James, S. and Collomosse, J.: Interactive video asset retrieval using sketched queries, *Proceedings of the 11th European Conference on Visual Media Production*, ACM, p. 11 (2014).
[25] Hu, W., Xie, N., Li, L., Zeng, X. and Maybank, S.: A survey on visual content-based video indexing and retrieval, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 41, No. 6, pp. 797–819 (2011).
[26] Ghosal, K. and Namboodiri, A.: A Sketch-Based Approach To Video Retrieval Using Qualitative Features, *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, ACM, p. 54 (2014).
[27] Lai, Y.-H. and Yang, C.-K.: Video object retrieval by trajectory and appearance, *IEEE TCSVT*, Vol. 25, No. 6, pp. 1026–1037 (2015).
[28] Zhang, Y., Chen, X., Lin, L., Xia, C. and Zou, D.: High-level representation sketch for video event retrieval, *Science China Information Sciences*, Vol. 59, No. 7, pp. 1–15 (2016).
[29] Wu, S., Yang, H., Zheng, S., Su, H., Zhou, Q. and Lu, X.: Motion sketch based crowd video retrieval, *Multimedia Tools and Applications*, pp. 1–29 (2017).
[30] Kratz, L. and Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models, *IEEE CVPR 2009*, IEEE, pp. 1446–1453 (2009).
[31] Ni, B., Yan, S. and Kassim, A.: Recognizing human group activities with localized causalities, *IEEE CVPR 2009*, IEEE, pp. 1470–1477 (2009).
[32] Benezeth, Y., Jodoin, P.-M., Saligrama, V. and Rosenberger, C.: Abnormal events detection based on spatio-temporal co-occurences, *IEEE CVPR 2009*, IEEE, pp. 2458–2465 (2009).
[33] Ryoo, M. S. and Aggarwal, J. K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities, *IEEE ICCV 2009*, IEEE, pp. 1593–1600 (2009).
[34] Chang, M.-C., Krahnstoever, N., Lim, S. and Yu, T.: Group level activity recognition in crowded environments across multiple cameras, *2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, pp. 56–63 (2010).
[35] Lan, T., Wang, Y., Mori, G. and Robinovitch, S. N.: Retrieving Actions in Group Contexts., *ECCV Workshops (1)*, pp. 181–194 (2010).
[36] Cheng, Z., Qin, L., Huang, Q., Jiang, S., Yan, S. and Tian, Q.: Human group activity analysis with fusion of motion and appearance information, *ACM MM 2011*, ACM, pp. 1401–1404 (2011).
[37] Choi, W. and Savarese, S.: A unified framework for multi-target tracking and collective activity recognition, *ECCV 2012*, pp. 215–230 (2012).
[38] Gowsikhaa, D., Abirami, S. and Baskaran, R.: Automated human behavior analysis from surveillance videos: a survey, *Artificial Intelligence Review*, pp. 1–19 (2014).
[39] Lim, S.-N. and Davis, L. S.: A one-threshold algorithm for detecting abandoned packages under severe occlusions using a single camera, Technical report (2006).
[40] Smith, K. C., Quelhas, P. and Gatica-Perez, D.: Detecting Abandoned Luggage Items in a Public Space, *IEEE Performance Evaluation of Tracking and Surveillance Workshop (PETS)*, No. LIDIAP-CONF-2006-033 (2006).
[41] Yoshimitsu, Y., Naito, T., Fujimura, K. and Kamijo, S.: Behavior understanding at railway station by association of locational semantics and postures, *2010 IEEE International Conference on Systems Man and Cybernetics (SMC)*, IEEE, pp. 3033–3038 (2010).
[42] Yuan, J., Zhao, Y.-L., Luan, H., Wang, M. and Chua, T.-S.: Memory recall based video search: Finding videos you have seen before based on your memory, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, Vol. 10, No. 2, p. 21 (2014).
[43] Huang, Y., Ma, C. and Wang, H.: Exploring the Benefits of Text and Sketch in Video Retrieval of Complex Queries, *Proceedings of the 8th International Symposium on Visual Information Communication and Interaction*, ACM, pp. 115–120 (2015).
[44] Shang, X., Ren, T., Guo, J., Zhang, H. and Chua, T.-S.: Video Visual Relation Detection, *ACM MM 2017*, Mountain View, CA USA (2017).
[45] Kang, K., Ouyang, W., Li, H. and Wang, X.: Object detection from video tubelets with convolutional neural networks, *IEEE ICPR 2016*, pp. 817–825 (2016).
[46] Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X. et al.: T-cnn: Tubelets with convolutional neural networks for object detection from videos, *IEEE TCSVT* (2017).
[47] Jin, C., Wang, Z., Zhang, T., Zhu, Q. and Zhang, Y.: A novel visual-region-descriptor-based approach to sketch-based image retrieval, *ACM ICMR 2015*, ACM, pp. 267–274 (2015).
[48] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al.: Imagenet large scale visual recognition challenge, *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211–252 (2015).