大規模映像検索のための潜在的概念抽出技術

研究代表者　　植　木　一　也　　明星大学　情報学部　情報学科　准教授

1 概要

本研究では，インターネット上に日々アップロードされる多種多様な大規模映像の中から，検索クエリに複数の概念を含んだ文（以下，**クエリ文**）を用いて詳細に映像を検索するための技術的課題に取り組んだ．ディープラーニング等の機械学習技術の発展により，学習済みの概念については高精度に検出できるが，新しいトレンドや新しい手口の犯罪等，未知の概念については正確に検出することができないという問題点があった．そのため，本研究では，映像検索のクエリ文中には具体的に明示されていないが，人が知識として持っている潜在的な概念を自動的に抽出することにより，新しい概念を検出可能とする仕組みを構築することを試みた．

2 従来研究

総務省・平成 28 年度版の情報通信白書によると，インターネットに接続する機器の増加，様々なサービスやアプリケーションの登場により,ネットワークを流通するデータトラヒックの量は飛躍的に増大している．米シスコによると，モバイルデバイスからのトラヒックが大きく伸び，中でも映像によるトラヒックが増大し，2020 年までにトラヒックの４分の３を占めると予想している．このような環境下で，インターネット上にアップロードされる映像に不適切な内容を含んでいないかといった安全性の確保のため，多様化する大量の映像を瞬時に自動解析して検索する技術が求められている．

映像検索技術は，YouTube 等のインターネット上の大規模な映像データから見たい映像を検索する，監視カメラの映像から異常を早期に発見するなど，応用範囲が広いことから，多くの研究者が長年にわたり研究を続けている．そのような背景から，映像検索に関する共通の評価基盤の必要性が議論され，2001 年から TRECVID ベンチマーク[1]が始まった．当初は，テキスト情報検索のベンチマークである TREC の中の 1 つのサブタスクとして始まり，2003 年からは独立して開催されるようになった．

TRECVID における大規模映像からの映像検索については,2010 年から 2015 年の 6 年間に実施された映像の意味索引付け（Semantic Indexing：以下，SIN）タスク[2]がある．SIN タスクは，物体・人・シーン・動作等の検出対象（以下，**コンセプト**）を含んだ学習映像が与えられているという条件下で，大量のテスト映像データベースの中から該当するコンセプトを含む映像を高精度に検出することが目的であった．近年，ディープラーニングを用いた手法により，飛躍的に精度が向上した[3][4]．

一方，本稿で取り組む AVS タスクは，上述の SIN タスクを発展させ，2016 年から新しく開始した，さらに難易度の高い課題である．AVS タスクでは，「Find shots of a person in front of a blackboard talking or writing in a classroom」といったクエリ文が与えられ，これに該当する映像を大規模映像データベースから検索することが求められる．このタスクの難しい点は，与えられたクエリ文に合致した学習用映像が与えられていない環境下で映像を検索するという，ゼロショット学習の技術が必要な点である．また，クエリ文中から，物体・人・シーン・動作等のコンセプトを「person」,「blackboard」,「talking」,「writing」,「classroom」のように抽出し，それら複数のコンセプトのすべてを含んだ映像を検索する必要があるという難しさもある．

筆者は，以下の 2 つの特徴を持つ映像検索システムを構築することで，TRECVID AVS タスクにおいて，2016 年と 2017 年に 2 年連続世界一の検索精度を達成することができた[5][6]：

- クエリ文中のキーワードに対応するクラスのカバー率を高めるため，様々な画像・映像データセットで学習された物体・人・シーン・動作等のコンセプトを検出可能なコンセプト識別器を大量に含む**コンセプトバンク**を構築．
- キーワードに対応するコンセプト識別器を数多く選択できるようにするため，自然言語処理の手法を導入．

これらのシステムには，ディープラーニングを活用していることから，学習済みの概念については高精度

に検出可能である．しかしながら，平均適合率の平均（mean average precision）という評価基準で，20%前後とまだまだ低く，ある特定のクエリ文に対しては，合致する映像を全く検索できないケースも見受けられた．これらの原因を分析すると，クエリ文の中に未知の概念が含まれていることが大きな問題であることがわかってきた．これでは，新しいトレンドや新しい手口の犯罪等，未知の概念については正確に検出することができない．そのため，本研究では，映像検索のクエリ文中には具体的に明示されていないが，人が知識として持っている潜在的な概念を自動的に抽出することにより，新しい概念を検出可能とする仕組みを検討した．具体的には，精度が著しく悪いケースに対し，

(1) 既存のコンセプト識別器から潜在的なコンセプトを見つけること，

(2) 視覚的特徴を用いて獲得した複数のコンセプト識別器を組み合わせること，

の 2 つの方法により解決可能かを調査した．

3 提案手法

映像検索システム構築にあたり，評価のための映像データには TRECVID ベンチマークの AVS タスクで使用する大規模映像データベースを使用し，同タスクにおいて 2016 年と 2017 年に 2 年連続世界一の検索精度を達成したシステム[5][6]をベースとする．この章ではまず，評価に使用する TRECVID ベンチマーク AVS タスクの概要（データベースと評価基準）について述べ，その後，ベースラインとなるシステムの概要について述べる．

3-1 TRECVID ベンチマーク AVS タスクの概要

2016，2017 年の TRECVID AVS タスクで出題されたクエリ文を表 1 に示す．各年それぞれ 30 種類（2016 年のクエリ ID は 501～530，2017 年のクエリ ID は 531～560）のクエリ文が与えられた．

表 1. TRECVID 2016，2017 の AVS タスクで出題された 60 種類のクエリ文

クエリ ID	クエリ文
501	Find shots of a person playing guitar outdoors
502	Find shots of a man indoors looking at camera where a bookcase is behind him
503	Find shots of a person playing drums indoors
504	Find shots of a diver wearing diving suit and swimming under water
505	Find shots of a person holding a poster on the street at daytime
506	Find shots of the 43rd president George W. Bush sitting down talking with people indoors
507	Find shots of a choir or orchestra and conductor performing on stage
508	Find shots of one or more people walking or bicycling on a bridge during daytime
509	Find shots of a crowd demonstrating in a city street at night
510	Find shots of a sewing machine
511	Find shots of destroyed buildings
512	Find shots of palm trees
513	Find shots of military personnel interacting with protesters
514	Find shots of soldiers performing training or other military maneuvers
515	Find shots of a person jumping
516	Find shots of a man shake hands with a woman
517	Find shots of a policeman where a police car is visible
518	Find shots of one or more people at train station platform
519	Find shots of two or more men at a beach scene
520	Find shots of any type of fountains outdoors
521	Find shots of a man with beard talking or singing into a microphone
522	Find shots of a person sitting down with a laptop visible
523	Find shots of one or more people opening a door and exiting through it
524	Find shots of a man with beard and wearing white robe speaking and gesturing to camera
525	Find shots of a person holding a knife

526	Find shots of a woman wearing glasses
527	Find shots of a person drinking from a cup, mug, bottle, or other container
528	Find shots of a person wearing a helmet
529	Find shots of a person lightening a candle
530	Find shots of people shopping
531	Find shots of one or more people eating food at a table indoors
532	Find shots of one or more people driving snowmobiles in the snow
533	Find shots of a man sitting down on a couch in a room
534	Find shots of a person talking behind a podium wearing a suit outdoors during daytime
535	Find shots of a person standing in front of a brick building or wall
536	Find shots of children playing in a playground
537	Find shots of one or more people swimming in a swimming pool
538	Find shots of a crowd of people attending a football game in a stadium
539	Find shots of an adult person running in a city street
540	Find shots of vegetables and/or fruits
541	Find shots of a newspaper
542	Find shots of at least two planes both visible
543	Find shots of a person communicating using sign language
544	Find shots of a child or group of children dancing
545	Find shots of people marching in a parade
546	Find shots of a male person falling down
547	Find shots of a person with a gun visible
548	Find shots of a chef or cook in a kitchen
549	Find shots of a blond female indoors
550	Find shots of a map indoors
551	Find shots of a person riding a horse including horse-drawn carts
552	Find shots of a person wearing any kind of hat
553	Find shots of a person talking on a cell phone
554	Find shots of a person holding or operating a tv or movie camera
555	Find shots of a person holding or opening a briefcase
556	Find shots of a person wearing a blue shirt
557	Find shots of person holding, throwing or playing with a balloon
558	Find shots of a person wearing a scarf
559	Find shots of a man and woman inside a car
560	Find shots of a person holding, opening, closing or handing over a box

クエリ文の中には，クエリ ID 544「Find shots of a child or group of children dancing」のように人が一つの動作をしている比較的単純なものから，クエリ ID 534「Find shots of a person talking behind a podium wearing a suit outdoors during daytime」のように，人，物体，動作，シーンを同時に含むものまで多岐にわたる．評価対象の映像は 335944 本あり，この中から与えられたクエリ文に合致している映像を検索する技術が求められる．世界各国の研究グループは，作成したシステムを利用し，それぞれのクエリ文に対してすべてのテスト映像を評価し，上位 1000 位までのランキングの結果を提出する必要がある．各チームの精度は，クエリ文ごとの適合率の平均を取った平均適合率（mean average precision：以下，mAP）という指標で比較される．

3-2 ベースラインシステムの概要

ベースラインとなる映像検索システムの検索パイプラインは以下の通りである．

(1) クエリ文から（複数の）キーワードを抽出する．
(2) キーワードに関連する（複数の）コンセプト識別器を選択する．
(3) 各テスト映像に対し，コンセプト識別器からのスコアを統合することにより，クエリ文に対するスコアを計算する．

この映像検索システムが映像検索の精度向上に寄与した要因は，主に以下の 2 つと考えている．1 つ目は，ImageNet 画像データベース[7]に代表される大規模画像・映像データベースを用いて事前に学習されたコンセプト識別器を用いて，クエリ文中に出現する単語のカバー率を向上させた点である．具体的には，表 2 に示す，50,000（重複を含む）を超えるコンセプト識別器を用いることで，多くのクエリ文は複数のコンセプト識別器の組み合わせで表現することができるようになった．

表 2. 映像検索システムで使用するコンセプトバンク

使用したデータベース	コンセプト数	コンセプトの種類	使用した認識モデル
TRECVID	346	人，物体，シーン，動作	GoogLeNet + SVM
FCVID	239	人，物体，シーン，動作	GoogLeNet + SVM
UCF	101	動作	AlexNet + SVM
Places	205	シーン	AlexNet
Places	365	シーン	GoogLeNet
Hybrid	1183	人，物体，シーン	AlexNet
ImageNet	1000	人，物体	GoogLeNet
ImageNet	4000	人，物体	GoogLeNet
ImageNet	4437	人，物体	GoogLeNet
ImageNet	8201	人，物体	GoogLeNet
ImageNet	12988	人，物体	GoogLeNet
ImageNet	21841	人，物体	GoogLeNet
Pascal VOC	20	人，物体	YOLOv2

2 つ目は，自然言語処理の技術を導入することにより，キーワードに対応するコンセプト識別器が存在しない場合でも，代用できるコンセプト識別器を探し出し，単語のカバー率をさらに向上させた点である．ここでは，word2vec モデルを用いて使われ方が類似している単語を代わりに利用する，単語を概念辞書 WordNet の Synset（同義語）まで拡張してコンセプト識別器を探し出すといった施策を行った．

以上の施策を行っても，ある特定のクエリ文については，該当するコンセプト識別器を見つけることができず，クエリ文に合致した映像が全く検索できない（平均適合率がほぼ 0%）というケースもいくつか見受けられた．この問題に対し，コンセプト識別器のバリエーションを増やし続けることは，一つの解決策ではあるが，すべてのキーワードをカバーするのは不可能である．そこで，クエリ文に合致している少量の画像を収集して評価を行うことで，

- 未知のキーワードに対応する潜在的なコンセプト識別器を見つけることはできないか？
- 保有しているコンセプト識別器の組み合わせでクエリ文を表現できないか？

といった疑問を明らかにしていく．

3-3 コンセプト識別器を用いたスコア算出

TRECVID，FCVID[8]，UCF[9]については，各映像データベースから，学習済みの CNN を用いて特徴を抽出し，各コンセプトに SVM を用いてコンセプト識別器を作成した．映像中のどの場面に対象のコンセプトが存在するかわからないことから，1 つの映像から最大 10 枚の画像を等間隔に切り出し，その中からコンセプトを含んでいる重要な特徴を抽出することとした．具体的には，各画像から得られる特徴ベクトルの要素の値ごとに最大値を得る最大値プーリングにより，複数の特徴ベクトルを束ねる処理を行った．ここで，各画像からの特徴は，画像を ImageNet データベースで学習された Caffe 開発チーム[10]が提供しているモデルや GoogLeNet モデル[11]に入力して，中間層から得られる特徴ベクトルを利用した．映像から特徴ベクトルを抽出した後は，各データベースのアノテーションを用いてサポートベクターマシン（SVM）の学習を行う．TRECVID データベースは，協調的映像アノテーション[12][13]で与えられている 346 カテゴリに対する正例／負例のラベルを用いた．FCVID データセットでは人・物体・シーン・行動のラベルが付与されている 239 カテゴリのデータを正例に，TRECVID からランダムに選択したデータを負例として SVM の学習を行った．UCF データセットも FCVID と同様に，UCF101 データベースで行動のラベルが付与されているデータを正例，TRECVID からランダムに選択したデータを負例として SVM の学習を行った．任意の映像に対するコンセプトのスコアは SVM の境界面からの距離により求めた．

Places（コンセプト数：205, 365）[14]と Hybrid については，Places205-AlexNet モデル（205 カテゴリの 250 万枚の画像を学習したモデル），Places365-GoogLeNet モデル（365 カテゴリの 180 万枚の画像を学習したモデル），Hybrid-AlexNet モデル（205 のシーンカテゴリと 978 の物体カテゴリを含む 1183 カテゴリの 360 万枚の画像を学習したモデル）を利用した．CNN の出力層の各ユニットがシーンや物体のコンセプトに対応しているため，コンセプトのスコアは CNN の出力層（ソフトマックス関数の前）の値を直接用いた．任意のテスト映像に対する各コンセプトのスコアは，その映像から切り出された最大 10 フレームの画像をそれぞれ CNN に入力して得られる各コンセプトのスコアの最大値とした．

コンセプト数を大幅に増やすため，複数の学習済みの ImageNet モデルも用いた．ImageNet（コンセプト数：1000）は Caffe 開発チーム[10]が提供しているモデル，ImageNet（コンセプト数：4000, 4437, 8201, 12988）については，アムステルダム大学が作成した CNN[15]，ImageNet（コンセプト数：21841）は，ImageNet データベースのすべて画像を学習した CNN[16]を用いた．ここでも，1 つの映像から最大 10 フレームを利用し，コンセプトのスコアは出力層の値を直接用いた．

「two or more men」や，「at least two planes」など，2 つ（2 人）以上の物体（人）を含んだクエリ文に対応するため，Pascal VOC データセットの 20 種類の物体を検出するモデル YOLOv2[17]を用いた．物体（人）の数が多い映像に高いスコアを与えるようなコンセプト識別器を作成した．具体的には，検出された各 Bounding Box に対するコンセプトの確率を，すべての Bounding Box で足し合わせたものをスコアとした．

これらの事前に準備されたコンセプト識別器を用いて，335944 本のテスト映像すべてを評価し，各コンセプトのスコアを算出した．コンセプトバンクから得られるスコアの範囲は，コンセプトによって異なるため，テストデータすべてを利用して，各コンセプトのスコアを 1.0（当該コンセプトに最も合致している映像のスコア）から 0.0（当該コンセプトに最もそぐわない映像のスコア）に変更する min-max 正規化を行った．

3-4 検証方法

クエリ文から新たな潜在的なコンセプトを獲得できるか，クエリ文を複数のコンセプトの組み合わせで表現できるか，さらにそれらのコンセプトが映像検索に有効であることを調べるため，

- 手法 1：クエリ文から自動でコンセプト識別器を獲得する方法
- 手法 2：クエリ文に合致した少量の画像からコンセプト識別器を獲得する方法

の比較を行った．

手法 1 では，クエリ文中のキーワードを独立に扱い，キーワードと同じ名前のコンセプト名を持つコンセプト識別器を利用した．同じ名前のコンセプト識別器がない場合は，word2vec を用いて，言語の使われ方が類似しているコンセプト名を持つコンセプト識別器で代用した．1 つのキーワードに対して複数のコンセプト識別器が存在する場合は，その平均スコアを利用した．クエリ文に対する最終スコアは，各キーワードに対するスコアの乗算により求めた．

手法 2 では，まず，Yahoo 画像検索でクエリ文を入力して検索された画像から，明らかに誤っているものを目視で排除し，クエリ文に合致した画像を 100 枚前後収集した．次に，各画像をコンセプト識別器で評価して，TRECVID のテストデータと同様にスコアを算出したのち，コンセプト識別器のスコアの平均が 0.5 以上となるものを選択した．クエリ文は選択されたコンセプト識別器の組み合わせと考えられるため，クエリ文に対するテスト映像の最終スコアは，選択されたコンセプト識別器の加重平均により求めた．ここでの重みは，算出されたコンセプト識別器のスコアを用いた．

その後，スコアの高いコンセプト識別器を見ることにより，今まで見つからなかった潜在的なコンセプトの存在の有無を確認した．また，手法 1 と手法 2 の平均適合率の比較や，検索された映像の目視による確認により，コンセプト識別器の組み合わせることの有効性も検証した．

3-5 実験結果

2016, 2017 年の TRECVID で出題されたクエリ文のうち，精度が著しく悪かったものについて，いくつかのパターン選び，原因を調査した．本稿では，物体のみのもの：「destroyed buildings」，人がある特定の動作をしているもの：「a person communicating using sign language」，人・物体・シーン・動作を組み合わせたもの：「one or more people walking or bicycling on a bridge during daytime」の 3 種類を例に，評価結果と詳細分析結果を述べる．

（1）クエリ文「destroyed buildings」の結果

クエリ文「destroyed buildings」は，人の行動やシーンを含まず，複雑なコンセプトの組み合わせもなく，単に物体が画像中に含まれていれば良いので，一見すると単純なクエリ文のように見える．しかしながら，「destroyed buildings」を直接的に表現可能なコンセプト識別器は存在しないという問題がある．実際，手法 1 で使用できる学習済みのコンセプト識別器は「*building*」のみであり，「destroyed」という単語が完全に無視されてしまうことから，平均適合率は 0.08%であった．検索された映像の上位 15 本を見てみると，図 1 の上部に示すように「destroyed」ではない，通常の「building」のみであった．

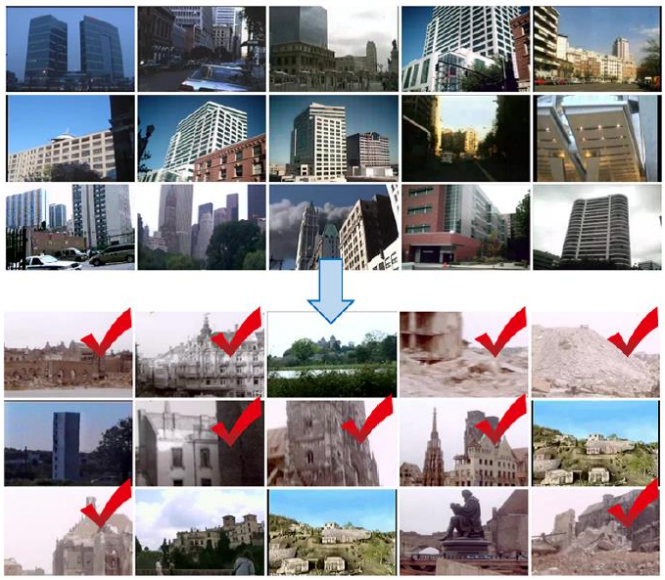

図 1. クエリ文「destroyed building」の映像検索結果．赤いチェックマークが付いているものは正例を表す．上：手法 1，下：手法 2

一方，手法 2 でコンセプト識別器の組み合わせを決定したところ，平均適合率 5.44%と向上し，多少ではあるがクエリ文に合致した映像が検索できていることが確認できた．実際の映像検索結果を見てみると，図 1 の下部に示すように，上位 15 本の映像中の 9 本は正例であり，その他の負例でも「destroyed building」に視覚的に近い映像が得られていることが確認できた．

また，「destroyed buildings」を表現するためには，どのようなコンセプトを選択すべきかを調査するため，選択されたコンセプト識別器をスコアの高い順に並べて出力した．上位 10 個のコンセプトを表 3 に示す．

表 3. 手法 2 において，クエリ文「destroyed buildings」から選択された上位 10 個のコンセプト識別器．データベース名の中の( )はコンセプト数を表す．

スコア	データベース名	コンセプト名
0.7519	ImageNet(12988)	*ruin*
0.7512	TRECVID	*Man_Made_Thing*
0.7024	ImageNet(21841)	*structure*
0.6968	ImageNet(8201)	*ruin*
0.6912	ImageNet(4437)	*ruin*
0.6896	ImageNet(21841)	*lobster_pot*
0.6837	ImageNet(12988)	*lobster_pot*
0.6797	ImageNet(4437)	*lobster_pot*
0.6744	ImageNet(21841)	*recycling_plan*
0.6679	ImageNet(4000)	*ruin*

ここで，「destroyed buildings」に近い意味（言い換え）を表す「*ruin*（[建物などの]廃墟，遺跡）」が選択されている点が興味深い．さらに上位に選択されているコンセプトを確認してみると，「*garbage_heap*（ゴミの山）」「*dump*（ゴミの山）」等，「destroyed buildings」に視覚的にも言語的にも近い意味を持つコンセプトが存在することがわかった．一方，表 3 の中に，図 2 のような画像特徴を持つ，「*lobster_pot*（[ロブスタ

ーを獲るための]わなかご)」といった全く関係ないコンセプトが含まれている.

図 2. ImageNet データベース中の「*lobster_pot*」の画像例.

これは,「destroyed buildings」の視覚的特徴, 特に細かいテクスチャが含まれている特徴が,「*lobster_pot*」の特徴に似ていることを意味していると考えられる. つまり,「destroyed buildings」は,「*building*」というコンセプトに, 細かいテクスチャを多く含む特徴である「*lobster_pot*」のコンセプトを加えることにより表現可能と考えることもできる.

（２）クエリ文「a person communicating using sign language」の結果

手法 1 では, クエリ文中のキーワードから該当するコンセプト識別器を見つける際, 2 つの問題があった. 1 つ目は,「communicating」と「sign language」に対応するコンセプト識別器が存在しないこと, 2 つ目は,「sign language」という複合語から, 誤って「*sign*（標識)」というコンセプト識別器が選ばれてしまうことであった. これらの問題により, 平均適合率は 0.03%と低い精度になってしまった.

一方, 手法 2 を用いて, クエリ文を画像特徴から得られたコンセプトの組み合わせで表現することにより, 平均適合率は 27.06%と大きな改善が見られた. しかしながら, 選択された上位のコンセプトのほとんどが人に関連するということを除き,「communicating」や「sign language」を導き出せる特徴的なコンセプトを見つけることはできなかった.

（３）クエリ文「one or more people walking or bicycling on a bridge during daytime」の結果

このクエリ文は, 人, 物体, 動作, シーンのコンセプトが含まれているため, それらを同時に含んだ映像を見つける必要がある. 手法 1 では,「*people*」,「*walking*」,「*bicycling*」,「*bridge*」,「*Daytime_Outdoor*」といったコンセプト識別器が利用できるため, クエリ中のキーワードをすべてカバーできているように見える. しかしながら, 平均適合率は 1.71%と低い検索精度であった. その理由を分析するため, 上位の映像検索結果を見てみたところ,「bridge」全体を遠くから写しているものが多く,「on a bridge」を表現するものではなかった. 図 3 に示す ImageNet にあるコンセプト「*bridge*」の画像を見てみても,「橋」全体を写したようなものがほとんどであることがわかる.

図 3. ImageNet データベース中の「*bridge*」の画像例

表 4. 手法 2 において，クエリ文「one or more people walking or bicycling on a bridge during daytime」から選択された上位 10 個のコンセプト識別器．データベース名の中の( )はコンセプト数を表す．

スコア	データベース名	コンセプト名
0.7474	ImageNet(21841)	*footbridge*
0.7474	Hybrid	*suspension_bridge*
0.7432	ImageNet(4000)	*footbridge*
0.7362	ImageNet(4437)	*footbridge*
0.7252	ImageNet(12988)	*footbridge*
0.7214	TRECVID	*Man_Made_Thing*
0.7193	ImageNet(21841)	*Tourist*
0.7028	Hybrid	*shopping_cart*
0.6962	ImageNet(21841)	*gangway*
0.6961	ImageNet(21841)	*suspension_bridge*

図 4. ImageNet データベース中の「*footbridge*」の画像例

図 5. ImageNet データベース中の「*shopping_cart*」の画像例

一方，手法 2 では，クエリ文に合致した画像から得られたコンセプトを利用することで，平均適合率は 4.05%と，若干の精度向上が見られた．上位に選択されたコンセプト名を表 4 に示す．ここでは，「*footbridge*」というコンセプトがいくつかのデータベースから選ばれていることが確認できる．このコンセプトの画像を ImageNet データベースで確認すると，図 4 に示すような，橋の近くや橋の上で撮影されている画像が多いことがわかる．また，ここで図 5 に示す「*shopping_cart*」という言語的に全く関係ないコンセプトが選択されているのは橋の一部にある線状の視覚的特徴が「*shopping_cart*」の特徴に近いからである．このような言語的には全く関係ないコンセプト識別器も，クエリ文に対応する視覚的特徴を表現できることから，映像検索の精度を向上させることが可能である．

4 まとめと今後の課題

本稿では，クエリ文からの詳細映像検索のタスクにおいて，対応するコンセプトが見つからなかった場合の対策を検討した．視覚的特徴を用いることにより，潜在的なコンセプトを見つける，クエリ文を既存のコ

ンセプトの組み合わせで表現するといった施策を試みた．

調査の結果，「destroyed building」から「*ruin*」，「on a bridge」から「*footbridge*」のように，映像検索に有益な新たな潜在的なコンセプトを得ることができた．これらは言語処理における言い換え（パラフレーズ）の問題を解くことにより，解決が見込まれる．

また，「destroyed building」から「*lobster_pot*」，「on a bridge」から「*shopping_cart*」といった，画像の視覚的特徴から得られる潜在的なコンセプトも見つかった．これらは一見，誤ったコンセプト選択のようにも受け取れるが，クエリ文に対応する視覚的特徴を持つ，意味のある情報である．実際，実験により，これらのコンセプト識別器を組み合わせることで，映像検索の精度を向上できることを確認できた．

しかしながら，画像特徴から得られるコンセプト識別器を選択するのみでは，解決できない問題も多く存在する．今回はクエリ文中の単語を独立に扱ったが，今後は，クエリ文の構文を解析した上で，複数のコンセプトの関係性，コンセプトの属性等を考慮したコンセプト識別器の選択方法を検討していく予定である．

一方，比較的理想的なコンセプト識別器を選べる手動の映像検索システムであっても，mAP が 20%程度という低い精度しか得られないことから，コンセプトベースの手法だけでは限界も見えている．TRECVID の 2017 年のシステムでは，コンセプトベースの手法と別の手法として，アムステルダム大学のチーム（チーム名：MediaMill）が採用した，映像とそれを記述した文を共通の空間に埋め込む手法[18]や，香港城市大学のチーム（チーム名：VIREO）の TRECVID Video-to-text（VTT）タスクのために学習されたシステムを用いた手法[19]があった．近年，flickr 8k[20]，flickr 30k[21]，MS COCO[22]に加え，Conceptual Captions[23]など，大規模な画像と説明文がセットになったデータベースの整備が進んでいることから，直接的に映像とクエリ文を扱って類似度を計算できる手法を改良していく必要があると考えている．TRECVID ベンチマークにおける AVS タスクも，今後継続予定のため，今後もベンチマークに参加し，他の研究機関と精度を競い合いながら，映像検索技術の向上に貢献していく予定である．

【参考文献】

[1] G. Awad, A. Butt, K. Curtis, J. Fiscus, A. Godil, A. F. Smeaton, Y. Graham, W. Kraaij, G. Quénot, J. Magalhaes, D. Semedo, and S. Blasi, ``TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search,'' In Proc. of TRECVID 2018, 2018.

[2] G. Awad, C. G. M. Snoek, A. F. Smeaton, and G. Quénot, ``TRECVid Semantic Indexing of Video: A 6-Year Retrospective,'' ITE Trans. on MTA vol.4, no.3, pp.187--208, 2016.

[3] C. G. M. Snoek, S. Cappallo, D. Fontijne, D. Julian, D. C. Koelma, P. Mettes, K. E. A van de Sande, A. Sarah, H. Stokman, and R. B. Towal, ``Qualcomm Research and University of Amsterdam at TRECVID 2015: Recognizing concepts, objects, and events in video,`` In Proc. of TRECVID 2015, 2015.

[4] K. Ueki, and T. Kobayashi, ``Waseda at TRECVID 2015: Semantic indexing,'' In Proc. of TRECVID 2015, 2015.

[5] K. Ueki, K. Kikuchi, S. Saito, and T. Kobayashi, ``Waseda at TRECVID 2016: Ad-hoc Video Search,'' In Proc. of TRECVID 2016, 2016.

[6] K. Ueki, K. Hirakawa, K. Kikuchi, T. Ogawa, and T. Kobayashi, ``Waseda Meisei at TRECVID 2017: Ad-hoc Video Search, '' In Proc. of TRECVID 2017, 2017.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ``ImageNet: A Large-Scale Hierarchical Image Database,'' In Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

[8] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, S.-F. Chang, ``Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks,'' arXiv:1502.07209, 2015.

[9] K. Soomro, A. R. Zamir, M. Shah, ``UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild,'' arXiv:1212.0402, 2012.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, ``Caffe: Convolutional architecture for fast feature embedding'', arXiv:1408.5093, 2014.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. ``Going deeper with convolutions,'' In Proc. of Computer Vision and Pattern Recognition (CVPR), 2015.
[12] S. Ayache and G. Quénot, ``Video corpus annotation using active learning,'' In 30h European Conference on Information Retrieval (ECIR'08), pp.187--198, 2008.
[13] J. Blanc-Talon, W. Philips, D. C. Popescu, P. Scheunders, and P. Zemcik, ``Advanced concepts for intelligent vision systems,'' In Proc. of 14th International Conference, 2012.
[14] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, ``Learning deep features for scene recognition using places database,'' Advances in Neural Information Processing Systems (NIPS), 2014.
[15] P. Mettes, D. C. Koelma, and C. G. Snoek, ``The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection,'' In Proc. of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR), pp.175--182, 2016.
[16] S. Ioffe and C. Szegedy, ``Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,'' arXiv:1502.03167, 2015.
[17] J. Redmon and A. Farhadi, ``YOLO9000: Better, Faster, Stronger,'' arXiv:1612.08242, 2016.
[18] C. G. M. Snoek, X. Li, C. Xu, and D. C. Koelma, ``University of Amsterdam and Renmin University at TRECVID 2017: Searching Video, Detecting Events and Describing Video,'' In Proc. of TRECVID 2017, 2017.
[19] P. A. Nguyen, Q. Li, Z. Cheng, Y. Lu, H. Zhang, and C. Ngo, ``VIREO@TRECVID 2017: Video-to-Text, Ad-hoc Video Search, and Video hyperlinking,'' In Proc. of TRECVID 2017, 2017.
[20] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, ``Collecting image annotations using amazon's mechanical turk,'' In Proc. of the NAACL-HLT workshop 2010, pp.139--147, 2010.
[21] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, ``From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,'' Transactions of the Association for Computational Linguistics, vol.2, pp.67--78, 2014.
[22] 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 European Conference on Computer Vision (ECCV), 2014.
[23] P. Sharma, N. Ding, S. Goodman, and R. Soricut, ``Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning,'' In Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, ACL2018, vol.1, pp.2556--2565, 2018.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
クエリ文からのコンセプトの選択に基づくアドホック動画検索	画像の認識・理解シンポジウム (MIRU2018)	2018 年 8 月
ゼロショット映像検索のための潜在的なコンセプトの抽出	画像の認識・理解シンポジウム (MIRU2018)	2018 年 8 月
Fine-grained Video Retrieval using Query Phrases - Waseda_Meisei TRECVID 2017 AVS System -	Proceedings of the 24th International Conference on Pattern Recognition (ICPR2018)	2018 年 8 月
Video Recognition and Retrieval at the TRECVID Benchmark	European Conference on Computer Vision (ECCV2018) Tutorial	2018 年 9 月
Waseda_Meisei at TRECVID 2018: Fully-automatic Ad-hoc Video Search	TRECVID 2018 Workshop	2018 年 11 月

Waseda_Meisei at TRECVID 2018: Ad-hoc Video Search	Notebook paper of the TRECVID 2018 Workshop	2018 年 11 月
Latent Concept Extraction for Zero-shot Video Retrieval	Proceedings of the Image and Vision Computing New Zealand (IVCNZ2018)	2018 年 11 月
複雑のコンセプトを含むクエリ文からのゼロショット映像検索 - TRECVID AVS タスクにおける成果と課題 -	精密工学会誌	2018 年 12 月
クエリ文によるゼロショット映像検索 - TRECVID 2018 AVS タスクの成果報告 -	動的画像処理実用化ワークショップ(DIA2019)	2019 年 3 月