Vol.2020-GN-109 No.34
Vol.2020-CDS-27 No.34
Vol.2020-DCC-24 No.34
2020/1/24

自然言語を用いた映像ライフログ検索方式の検討

森本麻代[1]　見上紗和子[1]　本橋洋介[1]

概要：近年，データの収集や分析が盛んになっている．データの中でも映像データは多くの情報が含まれており，今後更に長時間かつ数多く活用されることが期待される．そこで，映像データの新しい活用先拡大の一助として，映像データを自然言語で検索する方式を検討した．検討した方式では，曖昧な記憶や一部の情報を頼りに目的の物体が映っている映像箇所を検索することができる．本稿では検討した方式および試作システムについて述べる．

キーワード：映像ライフログ，一人称視点

A Study of Video Lifelog Retrieval System by using Natural Language

MAYO MORIMOTO[†1]　SAWAKO MIKAMI[†1]
YOSUKE MOTOHASHI[†1]

Abstract: Recently, data acquisition and data analysis are being very popular. Especially, video data are contained many kind of information, therefore, longer and many amount of video are expected to utilize widely in the future. Accordingly, for one of the new utilizing method to help increasing, we studied the method of video retrieval by using natural language. This method can search the position that the target object project rely on vague memory or one part of the information. This paper explain about the method we examined and the trial system.

Keywords: Video Lifelog, First-Person Point of View Video

1. はじめに

近年，AI ブームに伴い，様々な企業や組織があらゆるデータを収集したり，それらを活用して分析したりすることで，社内の効率化を図ったり，新しいサービスを検討したりしている．

データでも数値データ，言語データ，画像(静止画)データ，音声データなど様々な種類があり，用途に応じたデータが収集・利活用され，其々の特色を生かした分析方法が用いられている．その中でも最も情報量が多く含まれるデータとして，映像データがある．

映像データには，他とは異なる特徴がある．まず映像データには他よりも多くの情報量が含まれている．例えば，時間，場所情報，映っている物体や人などがあり，それらの軸で切り出して活用することができる．更に，映像データは一部を切り出したり，処理したりすることで，他の種類のデータとして利用することもできる．例えば，映像を切り取ることで画像データとして利用することや，音声を抜き出すことで音声データとして利用することが可能である．この音声データは音声認識することで言語データとして利用することも可能である．

ただし映像データにはメリットだけではなく，デメリットもあると考えられる．近年では映像データを収集するためのカメラには様々な種類があるが，高性能かつ高解像度になるものも多く，データ容量や通信速度が懸念される．故に収集し蓄積するには大容量の映像データを保存できる環境や，送信できる環境を整えておく必要がある．全てを蓄積しなくても良い場合は，防犯カメラや車載カメラのように，ある一定容量を超えると上書きする仕組みを活用することも可能である．また今後 5G の実用化が見込まれるため，今以上の容量のデータが通信されることが想定され，容量の大きい映像データが更に収集・活用されることが期待される．

映像データを収集するにはカメラを利用することが一般的であるが，カメラにも様々な種類が存在し，大きくは 2 種類，ある場所に固定して撮影するカメラと持ち運んで撮影するカメラがある．固定カメラには，防犯カメラや車載カメラや web カメラなどがあり，客観的な視点から広い範囲を撮影される時に使用される．また長期間撮影されることが多いのも特徴である．一方，持ち運び撮影用のカメラには，ビデオカメラやアクションカメラ，ウェアラブルカメラなどが存在し，一人称視点から，狭い範囲をピンポイントで短期間を撮影することが多い．

現在，人が記憶に残しておきたい印象的な場面に遭遇する場合，画像データとして残したり，断片的な映像データを残したりすることが多い．しかし本来であればその記録したい一瞬までの過程や，その後の経過など全てを記録し

1 NEC
NEC Corporation.

Vol.2020-GN-109 No.34
Vol.2020-CDS-27 No.34
Vol.2020-DCC-24 No.34
2020/1/24

ておくことで,一瞬を逃さずに記録することが期待できる．また何気ない場面でも記録すれば良かったと思う機会が増えることも考えられる．これら記録をする際の映像データは,固定カメラで客観的視点から撮影されたものではなく，自分が見たありのままの光景を記録する方が，自分の行動や記憶と結び付けやすいことが考えられる．つまり一人称視点を撮影できるカメラで長期間撮影することが望ましいと考える．将来，小型のカメラやカメラ付きのスマートグラスなどを装着し，日常生活を長時間撮影することで，あらゆる行動を記録できる可能性が期待できる．ただしこれらの撮影された映像データは膨大な量になるため，映像データを利用する際に的確なものを見つけ出すことが必要である．人間は断片的な情報に頼って記憶していることも多いため，曖昧な情報でも的確に検索できる仕組みがあることが望ましい．

そこで，本研究では一人称視点のカメラで撮影した映像データをあいまいな記憶情報でも検索可能な方式を検討し,システムを試作した．人は全ての行動や状況を明確に記憶することは難しく，断片的にかつ曖昧に記憶していることが多い．本研究で作成した方式では，一部情報が欠如していても，大量の映像データの中から目的の映像データを検索し，目的の物体が映っている場面から再生することができる．本稿は，大量の長時間映像データから目的の映像データを検索する方式と,その試作システムについて述べる．

2. 関連研究

人間の生活をデジタルデータとして記録するライフログに関する研究や，目的の物体が映っているかどうかを判断し，その映像データを検索する研究など，映像ライフログを記録・管理する方式に関する研究について述べる．

2.1 映像ライフログ

ライフログとは，人間の行動や生活などを記録したデジタルデータのことであり，映像，音声，位置情報などのデータや，記録する技術のことを指す．映像ライフログについて 2003 年より研究している相澤らは,「ライフログビデオ」という小型カメラにより個人の日常生活や体験を記録するシステムを開発している．[1]ライフログビデオでは，カメラやマイクの他に脳波計，GPS，ジャイロ，加速度計の各種センサからもデータを収集しており，多種多様なデータを利活用している．しかし複数のセンサを装着すると利用者に負担が掛かるため，日常生活での利用や実用化させるには難しいと考えられる．また目的の映像データを検索する際に，正しい情報を入力しなければヒットさせることが難しい．つまり，曖昧な記憶や断片的な情報からは検索しにくい構造になっている．しかし，人が記憶を思い出す時には，一部の情報しか思い出せないことが多いため，正しい検索結果が得られず，目的の映像データに辿り着けない可能性が高い．

スマートグラスのカメラを利用して一人称視点の映像データを収集し，ある特定の屋内の環境下での行動を判別する研究も進められている．[2]物体認識には DCNN を用いており，加速度を特徴量として用いることで認識精度を改善させている．更にユーザによるトレーニングデータを必要としないため，導入が容易である一方，認識範囲が制限される．またこの研究では行動を認識することを目的としているため，それらの映像を検索することは出来ない．よって本稿の検討している方式とは目的が異なる．

2.2 映像メタデータによる映像コンテンツ検索

映像データに付随するメタデータを活用して映像コンテンツを検索する研究も既に行われている．特に放送業界では活発に研究が進められている．

映像データに対してメタデータを付与する方法は大きく二つに分けることができる．まず一つ目は，映像データを物体認識することによって付与する方法である．既に学習されている物体を映像データから認識し，タグを付与する方法などが用いられる．二つ目の方法は，人手で付与する方法である．対象となる映像のシーンやカテゴリーなどのラベル付けなど，物体認識から容易に判断できないメタデータは人手で行う必要がある．

物体認識でのメタデータを付与する方法としては，河合らが提案している映像検索システムがある．[3]このシステムでは放送局の映像データを対象にしており番組表や字幕データなどの外部情報も学習データの一部として利用している．物体認識では特定の物体を判別できる判別器をそれぞれ作成することで，一定種類の物体を認識している．本システムでは対象としているデータが放送局の映像データであるため，カメラの切り替え点が非常に明確であり，映像切り出しが容易である．

人手でメタデータを付与する手法は，様々な企業により製品として販売されている．例えば「映像検索配信システム StreamGallery」[4]は映像やテキストスライドなどのコンテンツを作成・配信するサービスであるが，映像に関する音声テキストやシーン，使用しているスライドなどからキーワード検索することで目的の映像を見つける機能が備わっている．

このように，映像データに関するメタデータを利用した検索方法について研究が進められているが，物体認識する手法は認識するもののデータを事前に学習しておく必要があり，学習したものと条件が異なれば低い精度を出すことも多い．また人手で付与する方法はコストや時間が掛かるためデータ作成できる量に限界がある．よってどちらの方法にもメリット・デメリットがある．

本稿では，物体認識で一部のメタデータを付与し，物体認識で収集できなかったメタデータを人手で付与することとした．人手で一部を補うことで，データの質が高まり，

Vol.2020-GN-109 No.34
Vol.2020-CDS-27 No.34
Vol.2020-DCC-24 No.34
2020/1/24

より正確な検索が出来ることが期待できる．

3. 提案システムの全体構成

本稿で提案する自然言語を用いた映像ライフログ検索方式とは，過去に撮影した映像ライフログを，自然言語の入力文で検索する方式である．本方式では，映像ライフログに関する記憶が曖昧であっても，目的の物体が映っている映像ライフログの箇所を検索することができる．

3.1 システムの全体イメージ

本稿で提案する方式の概要図を図 1 に示す．

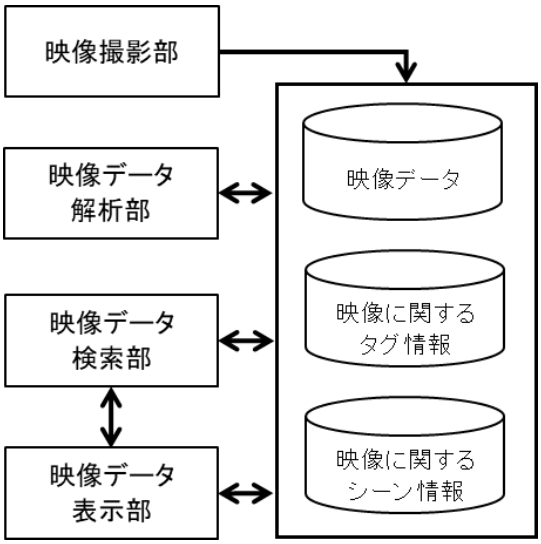

図 1　提案方式の概要

Figure 1　Summary of proposal method.

各部の構成と機能について 3.2 節以降に述べる．

3.2 映像撮影部

映像撮影部では映像データを撮影し，保存する．ライフログを収集することを想定するため，人に装着するカメラや，車載カメラなどの移動することができるカメラを想定する．

3.3 映像データ解析部

映像データ解析部は，蓄積された映像に対してタグを付与することや，シーンの判定を行い，結果を保存する．あらかじめ用意された学習データによって作成した学習済みのモデルによりタグやシーンを判別する．

3.4 映像データ検索部

映像データ検索部は入力された検索クエリに対して映像データを検索する．検索クエリに入力された文章と，保存されたタグ情報やシーン情報との一致度や類似度を評価し，クエリに関係する映像情報を出力する．

3.5 映像データ表示部

映像データ表示部は，映像データ検索部が出力した映像情報を基に，映像データに関するタグ情報や映像データの再生などの表示を行う．目的の映像データや関連する情報を表示まで完了させることで，自然言語による映像データの検索を実現する．

4. システムの試作と動作検証

3 章で述べたシステムの実現上の課題を明らかにするため，システムの試作を実施し，実際の映像データを用いた動作検証を実施した．本章では試作したシステムの構成・実装方法と，試作システムの動作検証において保存・解析に用いたデータについて述べる．試作したシステムの構成図を図 2 に示す．

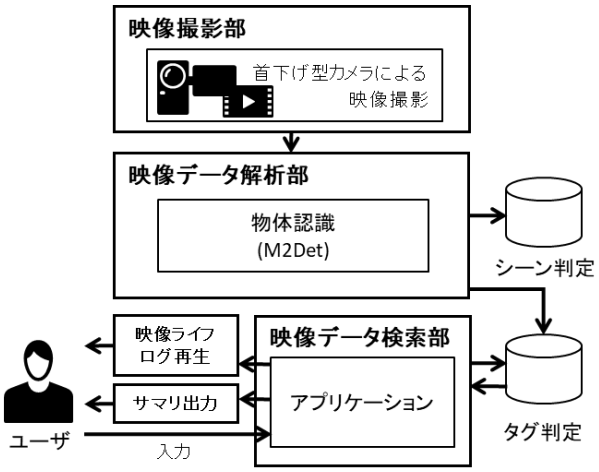

図 2　システムの構成図

Figure 2　Configuration of system.

試作システムの詳細な構成や実装方法，動作検証において用いたデータについて 4.1 節以降に述べる．

4.1 映像撮影部

映像撮影部では，ウェアラブルカメラ（GoPro HERO7 BLACK）[5]にアタッチメントを装着し，首から掛けられる状態にしたものを用いた．前記カメラを用い，被験者（3 名）が日常生活や業務中に装着し映像データを撮影した．撮影したデータは主に日常生活(例：食事，移動，買い物，スマートフォン操作)や業務中(例：国内・海外出張，対話，パソコン操作，会議)に撮影したものである．映像データの内訳を表 1 に示す．

表 1　収集した映像データの内訳

Table 1　Detail of collected movie data .

総映像データ数	総映像データ時間	検索対象映像データ数	検索対象データ時間	撮影者数
174	29h 27m 36s	70	10h 58m 59s	3

今回収集した総映像データ数は 174 ファイル，時間は 29 時間 27 分 36 秒である．またそのうち検索対象とした映像データ数は 70 ファイル，時間は 10 時間 58 分 59 秒である．

Vol.2020-GN-109 No.34
Vol.2020-CDS-27 No.34
Vol.2020-DCC-24 No.34
2020/1/24

残りの 104 ファイル，18 時間 38 分 37 秒分は，シーン判定のための学習データとして用いた．

4.2 映像データ解析部

映像データ解析部で行ったタグ情報付与処理やシーン情報付与処理について述べる．

4.2.1 タグ情報付与処理

本システムの学習アルゴリズムには，処理時間や精度の観点から M2Det[6][7]を用いた．なお M2Det を使用する際に，M2Det 用に公開されている Pretrained Model を学習モデルとして使用している．なおこの Pretrained Model は 80 種のタグ情報が付与されており，今回シーン判定や映像データの検索に利用する．Pretrained Model は Microsoft COCO (Microsoft Common Objects in Context) [8]で提供されている 32 万 8000 枚の静止画のデータセットを利用している．

4.2.2 シーン情報付与処理

シーン情報は，検索対象外の 104 ファイルの映像データのシーンラベルを学習することでシーン判定モデルを作成し，検索対象の映像データに対しシーン判定を実施した．学習用のシーンラベルは人手で正解データを作成した．作成においては，予め日常生活において網羅性のあるようにシーン情報の候補を作成した．作成したシーンラベルのリストを表 2 に示す．

表 2　シーンラベルリスト
Table 2　List of the Scene Label

シーンラベル名	
移動(徒歩)	観光(人工物)
移動(バス)	観光(自然)
移動(自転車)	エンターテイメント
移動(車)	レジャー
移動(電車・駅)	買い物
移動(飛行機・空港)	スマホ操作
食事	テレビ視聴
仕事(パソコン)	勉強
仕事	スポーツ
会議	通院(病院・歯科医院)
発表・講演・講義	美容院・ネイルサロン
会話	冠婚葬祭
ホテル・旅館	学校
不明	

4.3 映像データ検索部

映像データ検索部では，図 3 の処理によって映像データを検索する手順とした．以下にステップを記す．

- ステップ 1. Wikipedia をベースにした word2vec モデル内の単語をクラスタリングして各単語をクラスタに割り当てる．クラスタリングアルゴリズムには k-means を用い，50 クラスタに単語を分類した．
- ステップ 2. 動画に付与されたタグのそれぞれについて，そのタグに関する記事内の単語を抜きだす．その後，ステップ 1 で作成したクラスタ割り当て基準に基づき，各タグに関する記事内の単語の全てについてそれぞれクラスタを割り当てる．ここまでで，動画のあるタイミングに付与されたタグに関する単語のクラスタが複数割り当てられた状態になる．

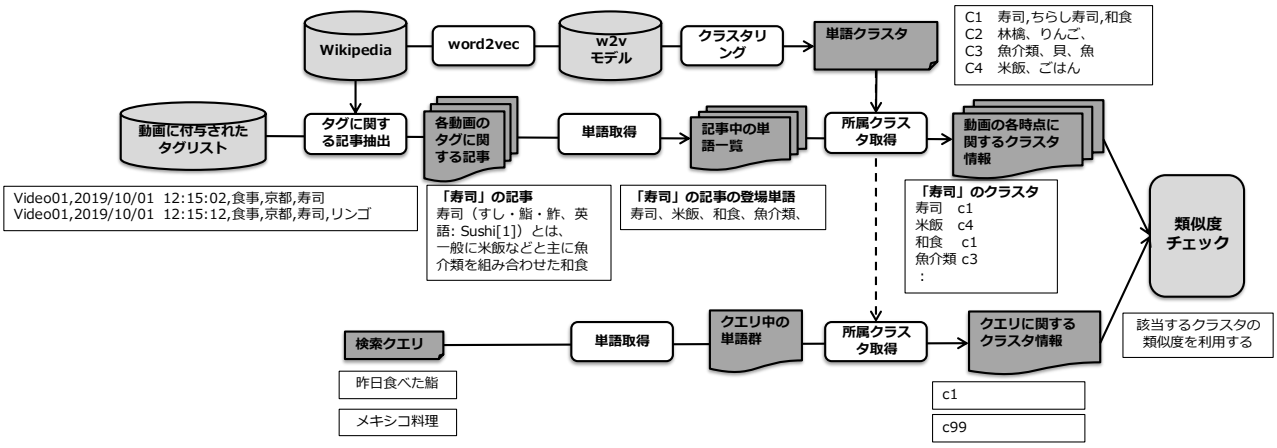

図 3　映像データ検索部の構成図
Figure 3　Configuration of searching video data part.

Vol.2020-GN-109 No.34
Vol.2020-CDS-27 No.34
Vol.2020-DCC-24 No.34
2020/1/24

これを，動画クラスタ情報とする．

- ステップ 3. 検索クエリの各単語に対して，クエリ内に含まれる単語を，ステップ 1 で作成したクラスタ割り当て基準に基づき，各タグに関する記事内の単語の全てについてそれぞれクラスタを割り当てる．これをクエリクラスタ情報とする．
- ステップ 4. 作成したクエリクラスタ情報と，動画の各時点の動画クラスタ情報の類似度を算出し，類似度が高いものを出力する．類似度判定においては，動画クラスタ情報とクエリクラスタ情報をそれぞれ５０次元のベクトルと見立て，当該２ベクトルの正規化後のコサイン類似度を類似度として用いた．

クエリ内の単語および動画のタグを word2vec による単語のエンベディング状態を用いたクラスタ情報に割り当てることで，クエリの表記と動画のタグが一致していない場合も，近しい意味の時に検索結果として出力される効果を期待して上記のような手順とした．

なお単語のエンベディングにおいては，web 上に公開されている日本語 Wikipedia エンティティベクトル[9] [10]の 20190520 版[11]を使用して word2vec で学習した 200 次元のモデルを用いた．

4.4 映像データ表示部

映像データ表示部は，Web アプリケーションとして実装した．表示部の画面例を図 4，図 5，図 6，図 7 に示す．

図 4　ホーム画面
Figure 4　Home screen.

図 5　検索クエリ入力後の画面
Figure 5　Screen after typing search query.

図 6　日にち選択後の画面
Figure 6　Screen after selecting the date.

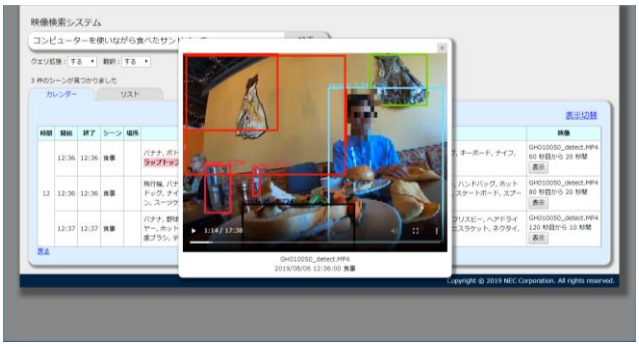

図 7　映像データ再生画面
Figure 7　Screen of playing video data.

図 4 が本システムのホーム画面である．上部には検索欄があり，検索したい映像に関する情報を入力し，右端にある「検索」ボタンを押すことで，図 5 のような結果が表示される．検索結果は該当する映像データが存在すると，画面下部のカレンダーの該当する日にちに件数として表示される．また検索結果の表示方法はタブで切り替えることができ，リストタブを選択すると該当する映像データが再生可能な形のリストで表示される．

カレンダータブ内の「○件」を選択すると，図 6 が表示される．選択された日にちの映像が時間順で表示されており，各映像データの特徴が中央列に羅列される．羅列された特徴は，一定値以上の特徴量が表れた 80 種のいずれかのタグである．またピンク色でハイライトされているタグが，検索時に入力した情報と関連性が高いタグである．左の映像欄の「表示」を選択すると，検索に該当する箇所から映像が再生される．映像が再生されている様子を図 7 に示す．認識されているタグは四角の枠で囲われて表示されている．

5. 動作検証結果の考察

今回試作したシステムについて動作検証を実施した．その動作検証の結果と考察を本章で述べる．

Vol.2020-GN-109 No.34
Vol.2020-CDS-27 No.34
Vol.2020-DCC-24 No.34
2020/1/24

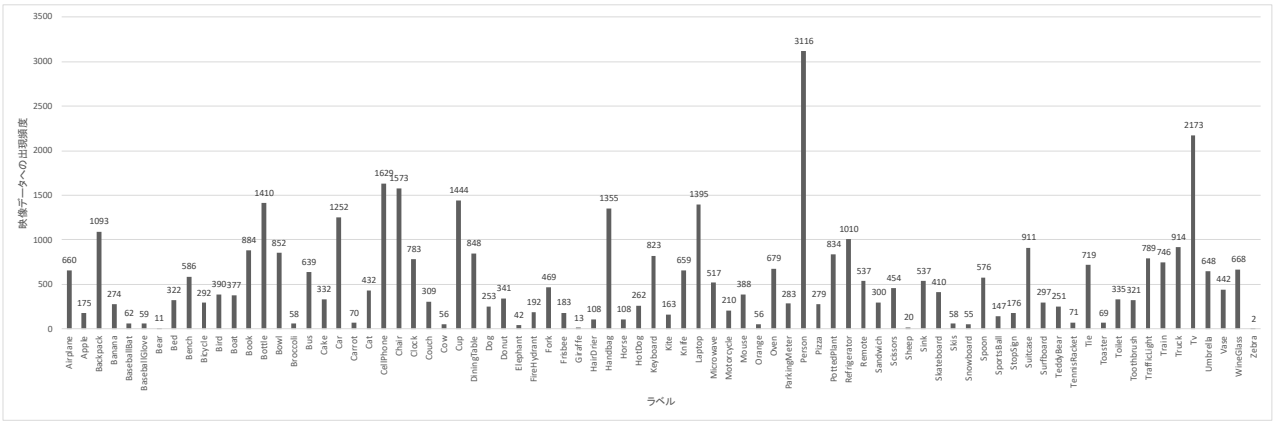

図 8　タグ情報の出現頻度

Figure 8　Appearance frequency of tag

5.1 動作検証結果

5.1.1 映像データ解析部の動作検証

今回検索対象とした 70 ファイルの映像データの中で，シーン判定時や検索時に使用した 80 種のタグが出現した回数の結果を図 8 に示す．タグには偏りがあり，Person が突出して多く，他には Tv，CellPhone，Chair が多く検出されていることが分かった．

また表 2 で作成したシーンラベルを検索対象の映像データに付与した内訳を図 9 に示す．

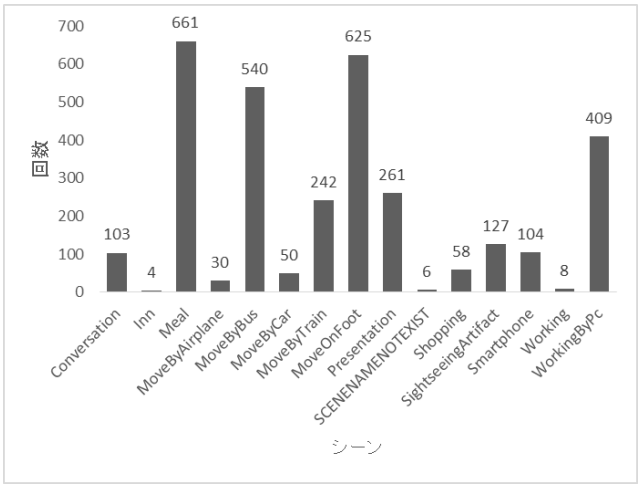

図 9　映像データのシーンの内訳

Figure 9　Detail of the scene in video data.

図 9 の結果から，Meal(食事)，MoveOnFoot(移動(徒歩))，MoveByBus(移動(バス))が多く，Inn(ホテル・旅館)，Working(仕事)が非常に少ないことが分かる．また，表 2 のシーンによっては 0 回のものも多くあり，撮影した映像データのシーンにも偏りがあることが分かった．

5.1.2 映像データ検索部の動作検証

動作検証ではいくつかの検索クエリを準備し，どのような結果が得られるか確認した．動作検証用に作成した検索クエリと，その際に確認したポイントを表 3 に示す．なおヒット件数とは検索時に目的の映像データの候補に選ばれた数のことを指す．

表 3　動作検証用検索クエリ

Table 3　Search query for operation verification

検索クエリ	確認のポイント	ヒット件数	対象の映像に辿り着けたか
コンピューターを使いながら食べたサンドイッチ	・タグ以外の類義語でもヒットするか(コンピューター)	51	○
見学した蒸留窯	・タグ以外の類義語でもヒットするか(蒸留窯)	265	○
北海道で食べた昼食	・場所情報 ・時間情報	1	×
空港で売っていたぬいぐるみ	・シーン情報 ・タグ以外の類義語でもヒットするか(ぬいぐるみ)	79	○

5.2 考察

動作検証の結果，4 件の検索クエリのうち 3 件から目的の映像データに辿り着くことができた．よって目的の映像データをヒットさせることはできることが確認できたが，ヒット件数が多いためヒットした候補の中から目的の映像データを探し出すのに時間掛かることが分かった．つまり，ヒットさせる映像データの候補を更に絞り込まなければ便利とは言えない．よって今後ヒットさせる映像データ数を減らしつつ，目的の映像データはヒット内に確実に入れて検索精度を高めていく必要がある．

またタグ以外の類似する単語を検索クエリに入力することで，目的の映像データをヒットさせ検索できることを確認した．しかし，必ずしも類似する単語とタグの関連度を高く算出できないため，改善する余地があると考える．

動作検証により提案システムの各箇所において課題があることが分かった．課題点について整理して 5.2.1 節以

Vol.2020-GN-109 No.34
Vol.2020-CDS-27 No.34
Vol.2020-DCC-24 No.34
2020/1/24

降に述べる．

5.2.1 検索アルゴリズムの課題

検索アルゴリズムに対する課題として以下が挙げられる．

- 検索バーに「場所」「時間」「食べ物」「購入品」「人物」など 5W1H を入力しても同じクエリ内を検索しているためにヒットしにくく，クエリの構成を再検討必要がある．

今回は同じクエリ内で複数種類の情報を検索しているため，例えば「昨日」や「今年の夏」などと入力しても正しい時制の映像データを検索することができない．曖昧な情報でも検索出来るようにするためには，これらの情報を同じクエリとして処理するのではなく，それぞれ別のクエリとして処理できるように構成を再検討する必要がある．

5.2.2 映像データ表示部の課題

映像データ表示部に対する課題として以下が挙げられる．

- 多くのタグが検知されているため，どの映像に何が映っているのか即座に判断できない．
- 複数人が撮影した映像データを対象とする場合，誰が撮影したものか判断できない．
- 検索後に日にちを選択した際の表示において，シーン判定結果が活用できておらず，何をしている映像か推測することが難しい．

提案システムの検索結果では，映像データを探している利用者に最低限必要な結果が出力されているため，利便性が欠けることが分かった．特に，各映像データのタグやシーン判定結果が有効活用されていないため，より的確に目的の映像データを検索するためには表示方法に改善の余地があると考える．

5.2.3 映像撮影部の課題

本システムで利用するために映像データの撮影する際には以下の課題がある．

- 今回利用したウェアラブルカメラでは GPS 情報や時間情報が正確に収集出来なかった．正確なデータを複数収集する場合は，別のセンサ等も併用することを検討する必要がある．
- 現在のウェアラブルカメラにはバッテリーや撮影容量に限界があるため，長時間撮影は向いていない．
- 日常生活やプライベートを撮影する際に抵抗感があり撮影しにくい．
- 全てのライフログ映像を残すことは難しい．特に公共の場などの撮影に抵抗感がある．
- 人の顔，パソコンの画面，スマホの画面などプライバシーが懸念される．
- カメラを首から下げているため目線の位置で撮影できず，想定される角度でものが映っていない．

課題は大きく分けると 2 つに分類できる．一つ目は機材(ハードウエア))の課題，二つ目は撮影環境の課題である．今回使用した機材は事前に調査し選定したが，使用してみると正しいデータが収集出来ないことが多く，1 つの機材だけを利用することに限界を感じた．また撮影環境では，プライバシーや公共でのマナーなどの影響から場所やシーンが限定され，活動時間全ての映像ライフログを収集するのは難しいことが分かった．

これらの課題を解決するには，別の機材もウェアラブルカメラとの併用の検討や，撮影範囲を限定するなどプライバシーに配慮する方法で映像ライフログ収集するなどの改善案が考えられる．

6. おわりに

本稿では，曖昧な記憶や情報から容易に映像ライフログデータを検索できる方式を提案し，試作システムを構築した. 映像データの撮影部, 物体認識や検索のアルゴリズム，映像データの表示部などそれぞれに課題があるが，それらを解決することで，より利便性と精度の高いシステムが期待できる．

今回はデータ収集時のカメラ性能，容易性，抵抗感の低さなどから，ウェアラブルカメラを使用したが，将来的にはスマートグラスなど新たなハードウエアを用いることで, 新しい用途や業界への展開が期待できると考えている．

将来はカメラで映像を長時間撮影し記録することがより容易にかつ一般的に出来るようになると考えられる．そのような環境が実現した場合，自分の映像ライフログから過去を思い出し，活用するシーンが拡大すると想定している．本稿で分かった課題を中心に，今後も機能改善していき，より良い映像ライフログ検索システムを開発していきたい．

参考文献

[1] 堀鉄郎, 相澤清晴.ライフログビデオのためのコンテキスト推定. 映像情報メディア学会技術報告. 2003, vol. 27, no. 72, p. 67-72.
[2] 久賀稜平, 前川卓也, 松下康之. 一人称視点映像を用いた Web 上の知識に基づく環境非依存な行動認識手法. 情報処理学会論文誌. 2017, vol. 58, no. 10 p. 1664-1673.
[3] 河合古彦, 望月貴裕, 住吉英樹, 藤原忍. 物体認識を利用した映像検索システム. 2014 年度映像情報メディア学会年次大会, 2014, 2-5.
[4] “映像検索配信システム Stream Gallery”. https://www.nec-solutioninnovators.co.jp/sl/stg/, (参照 2019-12-16)
[5] “GoPro HERO7 Black”.

Vol.2020-GN-109 No.34
Vol.2020-CDS-27 No.34
Vol.2020-DCC-24 No.34
2020/1/24

https://gopro.com/ja/jp/shop/cameras/hero7-black/CHDHX-701-master.html, (参照 2019-12-16)

[6] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, Haibin Ling. M2Det: A Single-Shot Object Detector Based on Multi-Level Feature Pyramid Network. The Thirty-Third AAAI Conference on Artificial Intelligence. 2019, vol. 33, no. 01, p. 9259-9266

[7] “M2Det”. https://github.com/qijiezhao/M2Det, (参照 2019-12-17)

[8] “Microsoft COCO”. http://cocodataset.org/, (参照 2019-12-17)

[9] 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. 言語処理学会第 22 回年次大会(NLP2016). 2016, p. 797-800.

[10] “日本語 Wikipedia エンティティベクトル”. http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/, (参照 2019-12-17)

[11] “Wikipedia Entity Vectors”. https://github.com/singletongue/WikiEntVec/releases, (参照 2019-12-17)