Vol.2016-CVIM-200 No.1
2016/1/21

Web上の知識を用いた1人称視点映像におけるオブジェクトの利用を伴う行動の認識に関する検討

久賀稜平[1]　前川卓也[2]　松下康之[2]

概要：センサを用いた行動認識技術は，独居高齢者見守りやホームオートメーションなどの基盤的技術であり，近年活発に研究がされている．本研究ではウェアラブルカメラにより撮影された一人称視点映像に着目し，ユーザによる事前学習を必要としない行動認識手法を提案する．従来の研究において，一人称視点映像や日常物に添付したセンサノードを用いて行動認識を行うものは多数存在するが，SVM などの機械学習のアプローチを用いることが多いため，ユーザによるトレーニングデータの収集が必要となる．一方本研究では，Web 上に存在する知識を用いることによって教師なしの行動認識を実現する．提案手法では，事前学習された一般物体認識用ディープニューラルネットワークを用いて，映像に含まれる日常行動で利用されたオブジェクトを認識したあと，そのオブジェクトの名前（例えば「フライパン」など）と，あらかじめ定義した日常行動の名前（例えば「料理をする」など）との類似度を計算することで，学習を必要としない行動認識を実現する．このとき，オブジェクトと行動の名前間の類似度を，Web における語の共起性や概念辞書などを用いて計算する．評価実験では一人称視点映像のデータセットを用いて評価を行い，トレーニングデータを一切利用しない本手法が良好な認識精度を示すことを確認した．

1. はじめに

近年，GoPro や Google Glass 等のウェアラブルカメラの進展により，一人称視点映像を用いた行動認識の研究が盛んに行われるようになっている．一人称視点映像を用いた行動認識研究は，特にライフログやヘルスケアへの応用が期待されており，ユーザのライフスタイルや健康状態の管理に重要な役割を果たすものと考えられる．また，アルツハイマー病や認知症患者の記憶を補助するための研究も行われている [7].

行動認識手法のアプローチには，大まかに分けてユビキタスセンシングとウェアラブルセンシングの 2 つがある．ユビキタスセンシングはユーザの身の回りの環境にセンサを添付し，そのセンサから得られたデータを用いて行動認識を行うものである．特に，ユーザが行動において利用したオブジェクトをセンシングし，その情報を用いてユーザの行動認識を行う方法がユビキタスコンピューティングの分野で盛んに研究されている [19]．このアプローチは，ユーザが使用しているオブジェクトはユーザが行っている行動に強く関連するという考えを基にしており，例えば，包丁やまな板などの利用が検知された場合，その情報から料理をするという行動が推定される．しかし，これらの手法は行動において利用されるあらゆる物にセンサを添付する必要があるため，導入・管理コストが大きくなってしまう．

ウェアラブルセンシングは，ユーザが身に着ける加速度センサやカメラなどのウェアラブルセンサを用いるアプローチである．加速度センサを用いた手法では，身体部位に添付した加速度センサを用いて身体部位の動きを捉えることで，ユーザの歩行や走行などの行動を認識する．しかしながら，身体の動きの情報のみを用いるため，オブジェクトの利用を伴う複雑な行動の認識は難しい．

本研究では，ウェアラブルカメラのみを用いて，オブジェクトの利用を伴う行動の認識を行う．すなわち，ユーザが行動の中で使用しているオブジェクトを一人称視点映像から抽出し，その情報から行動認識を行う．ここで，従来の一般的な行動認識手法 [13] では，教師あり学習のアプローチが取られているが，一般的な環境においてユーザがトレーニングデータを用意することは負担が大きい．特に，一人称視点カメラによって撮影される映像は人や環境によって異なるため，それぞれのユーザごとにトレーニングデータが必要である．このような問題を解決するため本研究では，一人称視点映像を用いた教師なし行動認識を提案する．

近年，一般オブジェクト認識向けのの事前学習された

[1] 大阪大学工学部
[2] 大阪大学大学院情報科学研究科

ディープニューラルネットワークが手軽に利用できるようになりつつある [8]．本研究では，それを用いて，まずユーザが利用しているオブジェクトを認識する．具体的には，時間窓内に含まれる一人称視点画像群から，「テレビ」，「リモコン」など，オブジェクトの名前のセットを抽出する．そして，抽出された名前のセットと，任意につけられた行動の名前との意味的な類似度を計算することで，ユーザによる学習データの収集を必要としない行動認識を行う．例えば，一人称視点映像から「テレビ」と「リモコン」というオブジェクトの名前からなるリストが得られたとする．このリストと，「料理をする」，「テレビを見る」などの行動の名前との意味的な類似度をそれぞれ計算し，最も類似度の高い行動を認識結果とする．このとき，オブジェクトのリストと行動の名前間の類似度の計算に Web 上の情報を利用する．例えば，「料理をする」と「鍋」の語は多くの Web ページにおいて共起率が高くなると考えられ，その共起情報を用いて類似度計算を行う．また，Web 上の概念辞書における語同士の距離を用いた類似度計算方法も提案する．ここで，行動名は一般的に動詞であることが多く，オブジェクトの名前は名詞である．概念辞書では動詞と名詞の距離計算は不可能であり，動詞の名詞形に変換したとしても，その名詞形とオブジェクトとの概念辞書における距離は大きい場合が多い．例えば，「cook」を「cooking」に変換したとしても．概念辞書である WordNet [12] における「pot」との距離は 17 ホップもある．そこで，本研究では行動において利用されると期待されるオブジェクトの名前をあらかじめ Web 上から抽出し，それらを行動の定義として拡張して用いる「セット拡張」を行うことで，行動名とオブジェクト名との距離計算を実現する．

2. 関連研究

2.1 ユビキタスセンシングによる行動認識

ユビキタスセンシングを用いた行動認識では，多くの研究において，環境の物体に添付したセンサを用いている．例えば，Radio Frequency Identification（RFID）タグを用いて，身の回りのオブジェクトの利用をセンシングし，ユーザの利用情報から行動認識を行う手法が研究されている．この手法は，人が行動を行っている際に使用するオブジェクトは，その行動と強く関連しているという考えに基づいている．Wu ら [19] は，ユーザが行動中に手で利用しているオブジェクトは，ユーザの行動を推測する重要な手がかりとなるという考えを基に，手首に装着した RFID リーダとオブジェクトに添付した RFID タグにより得られたオブジェクトの使用情報と，ビデオ映像から得られる手に持っているオブジェクトの SIFT 特徴を組み合わせて，ベイジアンネットによる認識手法を提案している．上記の研究では，周辺のオブジェクトにタグやセンサノードを添付する必要があるため，メンテナンス・導入コストが大きくなってしまう．

2.2 ウェアラブルセンシングによる行動認識

ウェアラブルセンシングの中でも特に一般的な手法であるユーザが装着するセンサを用いて行動認識を行う研究を紹介する．

2.2.1 身体部位の動きを捉えるセンサを用いた行動認識

RFID リーダ，および複数の RFID タグとアンテナをユーザの身体部位に添付し，行動認識を行う研究が行われている [17]．行動中にユーザの身体部位が動き，添付されたリーダとタグの距離が変化することで，行動ごとに異なるアンテナと RFID タグの組み合わせの電波情報を RFID リーダにより読み取ることができる．得られた電波情報からユーザの動きを捉える時空間的特徴情報を抽出し，SVM を用いてスマートフォン上で認識することで，低コストなリアルタイム行動認識を実現している．しかし，ユーザの体に複数のセンサを添付する必要があるため，ユーザへの負担が大きい．

ユーザの身体部位に添付した加速度センサを用いて部位の動きを捉えることで，ユーザの行動を推定する研究が数多く行われている [2], [14]．これらの研究では，「歩く」や「走る」などの行動における身体部位の特徴的な動きを捉えている．また，近年はウェアラブルセンサの 1 つとして，スマートフォンが注目されている．スマートフォンは多くの人が所有しており計算能力も高いことから，データマイニングツールとしても注目を集めており，行動認識の分野にも応用されている．ポケットに入った状態のスマートフォンのセンサからユーザの歩行や走行を認識する研究 [10] や，スマートフォン内蔵の加速度センサおよびジャイロセンサを用いることで，「調理」や「掃除」などのユーザの複雑な行動を認識する試み [5] が行われている．これらの手法は比較的低コストで実現でき，「歩行」や「走行」などの単純な行動は精度良く認識できるものの，オブジェクトの利用を伴う複雑な行動に関しては，高い精度での認識は困難である．

2.2.2 一人称視点映像を用いた行動認識

一人称視点映像から行動認識を行う手法がこれまでに数多く提案されている．ユビキタスセンシングと同様に，まずユーザがどのようなオブジェクトを使用したかを検出し，その情報を用いてユーザが行っている行動を推定する研究が多くなされている．Pirsiavash ら [13] は，オブジェクトを複数のパーツに分割するモデルである part-based model [6] を用いてあらかじめ学習させておいたオブジェクトを，一人称視点映像から認識し，行動認識を行っている．また近年では，Region based Convolutional Neural Network (R-CNN) を用いて行動に利用されるオブジェクトを画像内から切り出して認識する手法が提案されている．具体的には，入力画像からオブジェクト領域を Selective

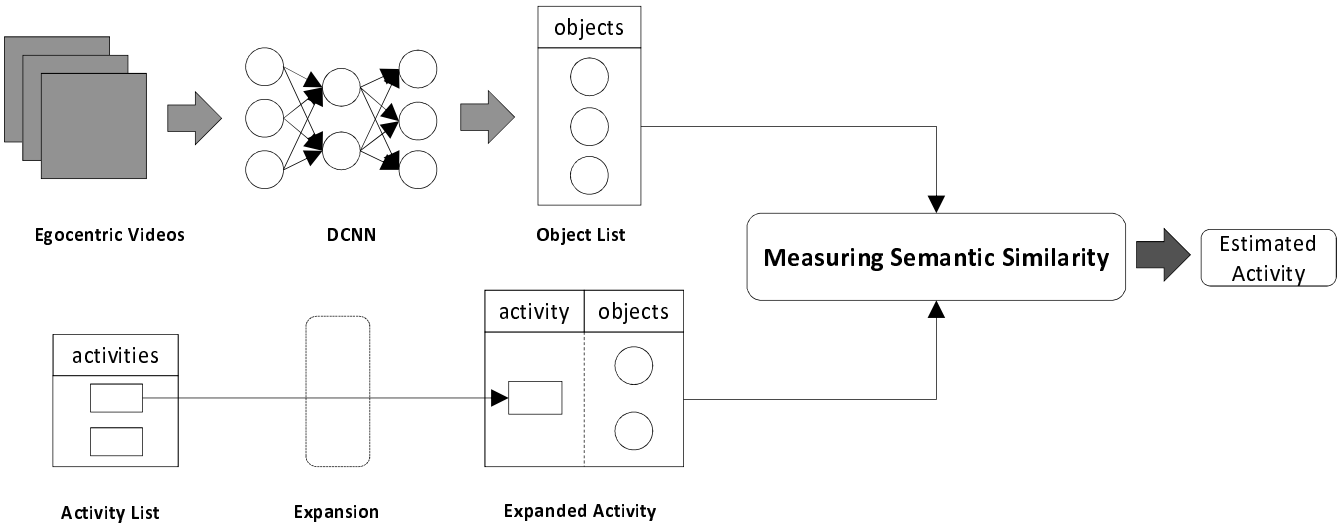

図 1 提案手法の概要

Search [16] や saliency map [1] を用いて検出し，その領域のみから CNN を用いて，オブジェクトを認識している．また，Ren ら [15] は，オプティカルフローを用いて一人称視点映像を背景領域と手の周辺領域とに分割し，手の周辺領域から手で利用しているオブジェクト領域を検出している．さらに，Luo ら [11] は，手に持っているオブジェクトの情報に加え，映像に現れるオブジェクトの動きの特徴なども用いて，行動認識を行っている．一人称視点映像はユーザや環境によって大きく異なるため，上記のような既存研究は，ユーザ・環境ごとにトレーニングデータが必要になるというデメリットが存在する．

本研究でも，CNN を用いてオブジェクトの使用を認識して行動認識を行うが，一般物体認識用の事前学習された DCNN を用いるため，ユーザによって収集されたトレーニングデータを必要としない．

3. 提案手法

3.1 概要

提案手法の概要を図 1 に示す．

まず，あらかじめ設定しておいた行動名を，その行動において使われるであろうオブジェクトのリストにより拡張を行うことで，行動ごとの定義を決定する．次に，認識対象となる一人称視点映像が得られたとき，スライディングウィンドウを設定し，そのウィンドウごとに行動を推定する．まずウィンドウ内に含まれる画像に対して，事前学習された Deep Convolutional Neural Network（DCNN）を用いてその窓内の画像に含まれるオブジェクトのリストを得る．次に，得られたオブジェクトリストに含まれるノイズ，すなわち誤って検出されたオブジェクトを除去したあと，あらかじめ作成した行動の定義ごとに，オブジェクトリストとの類似度を計算することで行動認識を行う．以下では，行動において利用されると期待されるであろうオブジェクトのリストを生成する手法について 3.2 章で述べ，3.3 章で一人称視点映像からの DCNN を用いたオブジェクト認識について述べる．DCNN により，一人称視点映像からユーザが使用したオブジェクトのリストが得られる．得られたリストからノイズを除去する手法を 3.4 章で述べ，得られたオブジェクトリストとあらかじめ作成しておいた行動の定義との類似度の計算方法を 3.5 章で述べる．

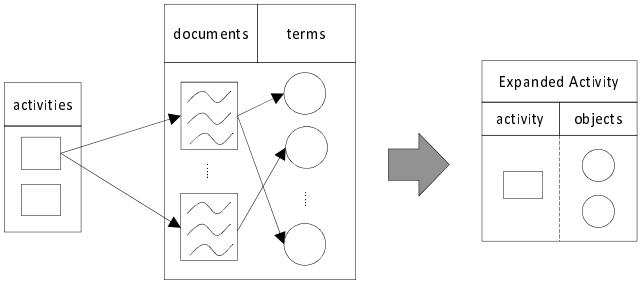

図 2 Web 文書を用いた行動名の拡張

3.2 行動名の拡張

提案手法では設定された行動名を用いて類似度計算を行うが，行動の名前は短いものが多く，類似度計算の際に正しい結果が得られない可能性がある．そこで，あらかじめ設定された行動の名前を，その行動で使用されると期待されるオブジェクトのリストで拡張する．これにより，設定された行動名を補完する．情報検索の研究分野では，ユーザによって入力された短いクエリを，web 上の文書を用いて補完する研究が行われている．例えば，Cui ら [4] や Wen ら [18] は，検索エンジンのクエリログとユーザが閲覧した文書からクエリと共起する語を抽出し，その語のリストをクエリの拡張に用いている．本研究では，このクエリ拡張技術を一般的に短い行動の名前を補完するために用いる．ある行動において使用されるオブジェクトは，web 上の文書においても行動名との共起率が高いと考えられため，行

動名をクエリとする web 検索結果に含まれる文書から，行動名に共起するオブジェクトリストを抽出する．検索結果に含まれる文書内には，行動において使用されるオブジェクト名が頻出し，それらの文書内における重要度は高いと考えられる．そこで，Term Frequency Inverse Document Frequency (tf-idf)[9] を用いてオブジェクト名の重要度を計算し，重要度が大きいものを行動に共起するオブジェクトとする．tf-idf は，term frequency（単語の出現頻度）と inverse document frequency（逆文書頻度）の積から計算される．オブジェクトの出現頻度が高ければ，そのオブジェクトが行動名に大きく関連していると言え，逆文書頻度が高ければ，そのオブジェクトはその行動に固有のオブジェクトであると言える．提案手法では，より日常生活における共起性の高い単語を抽出できるようにするため，how-to サイト [*1] の文書のみを検索対象とし，さらに ImageNet に列挙されているオブジェクト名のみに関して重要度計算を行った．あらかじめ用意したそれぞれの行動名に対してクエリ拡張を行い，得られた tf-idf 値の高い単語を，行動名に対応するオブジェクトリストとする（図 2）．行動名と上記のようにして作成されたオブジェクトのリストを行動の定義とする．

3.3 DCNN を用いた物体認識

一人称視点映像からオブジェクトを認識するために，DCNN を用いる．DCNN は既存の機械学習手法と異なり，画像の特徴量を自動で獲得して学習を行うことができるため，近年注目を集めている．本研究では，オープンソースの DCNN フレームワークである Caffe [8] を利用する．Caffe では，オブジェクトの画像から構成される ILSVRC2012 データセット [*2] を用いてあらかじめ学習されたモデルが用意されており，このモデルを利用することで，トレーニングデータを利用者が用意することなく画像に含まれるオブジェクトを認識することができる．DCNN の出力は，それぞれのオブジェクトクラスのクラス分類確率であるが，この値は出力層のそれぞれのクラスに対応するニューロンの活性化関数から計算されている．本研究で用いた DNN は 1 つのオブジェクトが含まれる画像から学習されているが，入力画像に複数のオブジェクトが含まれていた場合でも，それぞれのオブジェクトに対応する出力層のニューロンが活性化される．すなわち，入力画像に複数のオブジェクトが映っていた場合でも，それぞれのオブジェクトの分類確率が高くなることが期待される．一般的な行動は複数のオブジェクトの利用が伴うため，DCNN を用いることで複数のオブジェクトの利用を同時に認識できる．さらに，ILSVRC2012 データセットを用いた Caffe の学習モデルでは，各画像カテゴリは WordNet の概念の ID となっている．以上まとめると，本研究で用いる DCNN は，1 枚の入力画像に対して，その画像に含まれると推定されるオブジェクト（WordNet の ID）とそのスコアのリストを出力する．

[*1] http://www.wikihow.com
[*2] http://www.image-net.org/challenges/LSVRC/2012/

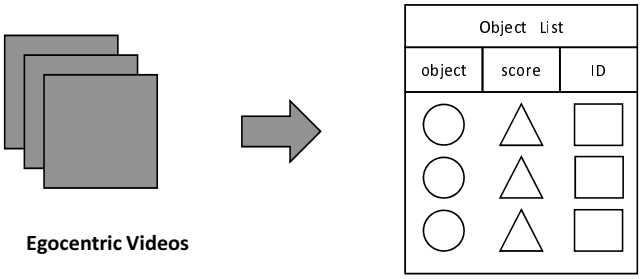

図 3　映像から得られるオブジェクトリスト

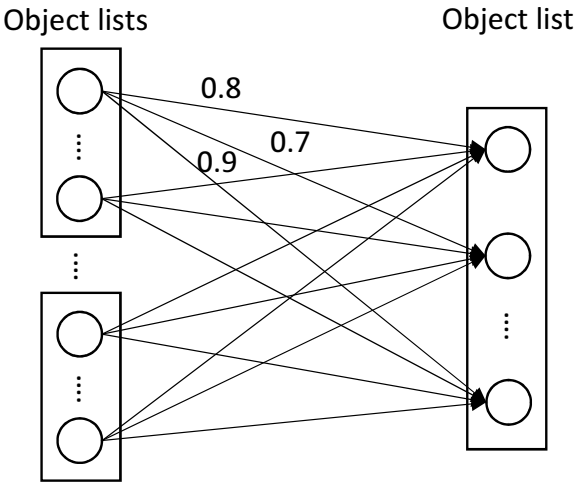

図 4　2 つのオブジェクトリスト間の類似度計算

3.4 オブジェクトリストのフィルタリング

提案手法では時間窓ごとに，窓に含まれる一人称視点画像からオブジェクトリストを抽出し，行動を推定する．このとき，DCNN の認識エラーにより実際に画像に含まれていないオブジェクトが抽出されることがあるが．これらを行動の定義との類似度計算に使用することで認識性能が低下すると考えられる．誤って認識されたオブジェクトはそのクラス分類確率（スコア）が低く，ウインドウ内の画像に含まれる頻度も低いと考えられる．そこで，それぞれのオブジェクトごとにウインドウ内の画像から抽出されたオブジェクトリスト内の対応するスコアの総和を計算し，その総和をウインドウにおけるそのオブジェクトのスコアとする．そして，閾値より大きいスコアをもつオブジェクトのみを保持する．すなわち，ウインドウ内の画像から，オブジェクト（Wordnet ID）とスコアの 2 つの項目からなるリストを得る．

3.5 類似度計算

オブジェクトリストにより拡張された行動の定義と，窓

ごとの一人称視点映像から得られたオブジェクトリストとの類似度を計算し，最も類似度の高い行動名を認識結果とする．本研究では，以下の 2 つの類似度計算方法を考案し，評価実験において比較する．

3.5.1 WordNet を利用した距離計算

1 つ目は WordNet[12] を用いた方法である．WordNet はオンライン上の概念辞書であり，約 11 万 7 千の synset と呼ばれる同義語集合間の関係が木構造で記述されている．そのため，2 つの synset 間の関係を距離として得ることができる．そこで，WordNet を用いて行動名とオブジェクトリストとの距離を計算する手法を提案する．

まず，拡張したオブジェクトリストを用いずに，行動名のみ用いて，類似度を計算する方法を述べる．この場合，あらかじめ設定された行動名から名詞を抽出し，それに対応する WordNet 内の synset を検索する．例えば，“watching television”からは “television”が属する synset が得られる．名詞が行動名に含まれない場合は，動詞を名詞形に変換して用いる．そして，窓内の映像から得られたオブジェクトリスト $\mathcal{O}_{img}$ との類似度を

$$S_{wn}(n, \mathcal{O}_{img}) = \sum_{y_j \in \mathcal{O}_{img}} V(y_j) W(n, y_j)$$

で定義する．n は行動名から抽出された名詞，$V(X)$ はオブジェクト X のスコア，$W(X,Y)$ はオブジェクト X と Y の WordNet 上での類似度であり，$W(X,Y) = 1/D(X,Y)$ で定義される．D は WordNet 上での 2 つの synsetX, Y 間のエッジの数である．

拡張したオブジェクトリストを用いて類似度を計算する場合は，WordNet の synset のリスト同士の類似度計算となる（図 4）．行動名から拡張したオブジェクトのリストを $\mathcal{O}_{act}$ として，2 つのリスト間の類似度を次のように定義する．

$$S_{wn}(\mathcal{O}_{act}, \mathcal{O}_{img}) = \sum_{x_i \in \mathcal{O}_{act}} \sum_{y_j \in \mathcal{O}_{img}} V(y_j) W(x_i, y_j)$$

3.5.2 Web 検索エンジンを用いた類似度計算

この手法では，検索エンジンにより得られる語のヒットカウントの情報を用いて語同士の類似度を計算する．まず，クエリ拡張をする前に行動名のみを用いて距離計算をする場合について説明する．

$h(q)$ を “q”，$h(q_1, q_2)$ を “q_1 q_2” をクエリとした場合の検索エンジンから得られる web ページのヒットカウント数とした場合，ある行動（a）においてあるオブジェクト（o）が使用される確率は条件付き確率 $P(o|a) = h(a,o)/h(a)$ で求められる．例えば，$h(cooking) = 62,800,000$，$h(cooking, stove) = 4,440,000$ であるとき，cooking という行動において stove が使用される確率は 444000/62800000 = 0.07 と計算される．これを行動名とオブジェクトとの類似度と定義し，行動名とウィンドウ内の画像から得られたオブジェクトリストとの類似度を次式で定義する．

$$S_{se}(a, \mathcal{O}_{img}) = P(\mathcal{O}_{img}|a)$$

$$= \sum_{y_j \in \mathcal{O}_{img}} V(y_j) \frac{h(a, y_j)}{h(a)}$$

拡張したオブジェクトリストを用いて類似度を計算する場合，オブジェクトリスト間の距離計算となる．文献 [3] では，コーパスから得られる相互情報量を用いることで，単語間の意味的な類似度を計算している．相互情報量は 2 つの確率変数がどの程度情報量を共有しているかを示す指数であり，

$$I(X = x, Y = y) = \log \frac{P(X = x, Y = y)}{P(X = x)P(Y = y)}$$

で定義される．本手法では，相互情報量を用いてオブジェクト間の距離を計算し，さらにそれを用いてオブジェクトリスト間の距離を計算する．Web 上では，ある語 w の事前確率は検索エンジンがインデックスするページ数である W を用いて，$P(w) = h(w)/W$ のように表されるため，2 つのリスト間の類似度は相互情報量を用いて次のように定義できる．

$$S_{se}(\mathcal{O}_{act}, \mathcal{O}_{img}) = \sum_{x_i \in \mathcal{O}_{act}} \sum_{y_j \in \mathcal{O}_{img}} V(y_j) \log \frac{h(x_i, y_j) W}{h(x_i) h(y_j)}$$

4. 評価実験

4.1 データセット

本研究では，Google Glass を装着したユーザが表 1 に示す 13 種類の行動を行い，Glass のカメラで一人称視点映像を撮影した．Glass のカメラは 1280×720 ピクセルの JPEG 画像を 30fps で撮影する，また，表 1 の行動名は既存の行動認識研究論文において利用されているものを基本的に用いた．1 名の被験者が自身の住居で 13 種類の行動が含まれるセッションを 5 回行った．各セッションの平均時間は約 13 分である．データの取得方法には semi-naturalistic collection protocol [2] と呼ばれる方法を用いた．この手法では，被験者に対して実行して欲しい行動の一覧をランダムな順で提示をするが，具体的にどのように振る舞って欲しいかは伝えない．したがって，より日常生活における自然な状況を想定したデータを収集することができる．図 6 に実験において得られた一人称視点画像の例を示す．

4.2 評価手法

評価実験では以下の 4 つの手法を比較・評価する．

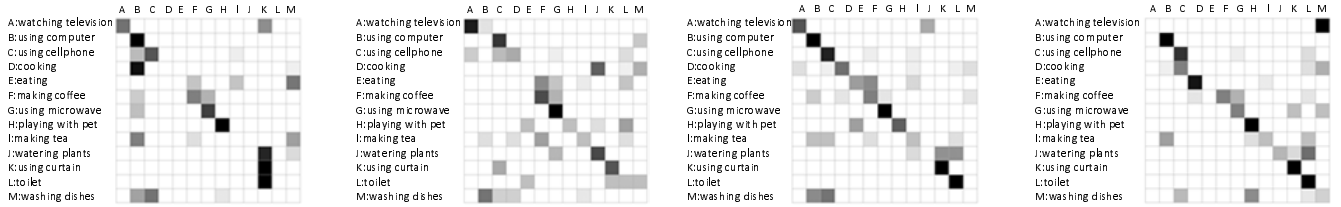

図 5 認識結果の混同行列

図 6 実験時に使用した入力映像の例

表 1 実験で行った 13 クラスの行動

making tea	making coffee	cooking
watching television	using cellphone	using computer
eating	toilet	washing dishes
using microwave	using curtain	playng with pet
watering plants		

(1) WordNet: WordNet を用いた類似度計算
(2) Search Engine: Web 検索エンジンを用いた類似度計算
(3) WordNet (Expansion): 行動名の拡張+WordNet を用いた類似度計算
(4) Search Engine (Expansion): 行動名の拡張+Web 検索エンジンを用いた類似度計算

評価指標: ウィンドウ内の映像に対して，3 章で説明した手法を用いて行動を推定し，手動でラベリングされた正解と比較する．そして，正しく認識されたウィンドウの数を基に，認識率を平均 F 値により評価する．

実験パラメータ: 行動名の拡張において，検索エンジンから得られる検索結果からは多数の文書を得ることができるが，クエリによく適合する文書は高々上位 20 件までと考え，20 件までの文書を用いてオブジェクト名の tf-idf 値を計算した．また，得られた文書から tf-idf 値を算出し，行動名に対してオブジェクトリストを拡張する際，tf-idf 値の高い上位 3 個のオブジェクトのみを用いた．行動の推定の際は，時間窓のサイズは 5 秒とした．一つの窓には約 50 枚の画像が含まれる．オブジェクトリストのフィルタリングの閾値は 0.1 に設定した．さらに，0.1 以上のスコアを持つオブジェクトが 4 つ以上ある場合は，上位の 3 つのみを用いた．

表 2 それぞれの手法の認識精度

	-		+expansion	
	WordNet	Search Engine	WordNet	Search Engine
precision[%]	31.3	32.8	53.0	44.9
recall[%]	42.1	38.4	60.4	50.1
F-measure[%]	32.4	30.9	53.7	41.5

4.3 認識

表 2 にそれぞれの手法の認識精度を示す．行動名の拡張を行わず WordNet を用いて距離計算する手法では 32.4%，Web 検索エンジンを用いる手法では 30.9% の F 値となった．行動名を拡張した場合，WordNet を用いる手法では 53.7%，Web 検索エンジンを用いる手法では 41.5% の F 値となった．行動名を拡張せずに行動認識を行った場合，WordNet を用いる手法が Web 検索エンジンを用いる手法を 1.5 ポイント上回った．図 5 にそれぞれの手法の混同行列を示す．Making coffee, Making tea, Watering plants などの複数のオブジェクトを使用する行動は検索エンジンを用いる手法が精度よく認識できている．一方で，Using computer, Using cellphone, Using curtain などの基本的に 1 つのオブジェクトしか利用されない行動に関しては，行動名に含まれるオブジェクトと同じオブジェクトが映像にも現れたため，WordNet を用いる手法の精度が上回ったと思われる．

表 2 に示すように，WordNet および Web 検索エンジンを利用する両方の手法について，拡張したオブジェクトリストを用いて類似度を計算することで精度の向上が確認された．特に，WordNet を用いた場合の Cooking と Eating において顕著に効果が見られた．例えば，Cooking からは {pot, wok, stove} がオブジェクトとして拡張されており，これらは実際の行動で使用すると期待されるものと一致している．実際に，映像からも同じオブジェクトが認識され，より正確に類似度を計算することができた．一方で，行動名の拡張を行っても精度が向上しなかった場合もあった．Watering plants からは {pot, patio, wine bottle} が拡張されていたが，実際の映像から「patio」（中庭）が認識されることはなく，「wine bottle」は Watering plants とは関係のないオブジェクトである．さらに WordNet の結果では，行動名の拡張を行うことで，Playing with pet の精度が低下してしまっている．Eating に対して「hen」が拡張され

ており，一人称視点画像から得られた「labrador retriever」と WordNet 上での距離が近く，誤認識が起こっていた．一方，Web 上では鶏と犬が共起することは少なく，検索エンジンを用いる手法では精度が向上した．

また，全ての行動において，行動中に常に映像内にオブジェクトが映っているとは限らず，例えばオブジェクトがユーザの手で遮蔽されたり，ユーザがよそ見をしたりすることにより，オブジェクト認識が正しく行えなかった場合もあった．さらに，Washing dishes に関しては，どの手法を用いても正しく認識できないことが多かった．これはオブジェクト認識において，行動で使われていたオブジェクトが認識されていなかったことが主な原因である．実際に食器を洗っている映像を DCNN でオブジェクト認識した場合，シンクや食器などが認識されていなかった．一般的な家庭において，シンク周りには多数の物が置かれている場合が多く，それらがシンクの認識に影響を及ぼしたと考えられる．

5. おわりに

本研究では，Web 上に存在する膨大な情報に着目した一人称視点映像における行動認識手法を提案した．提案手法では，Web 上の知識を用いて行動名と実際に使用されたオブジェクトとの類似度を計算することで，ユーザによるトレーニングデータを必要としない行動認識を行った．評価実験では，Google Glass を用いて撮影した映像を用いて提案手法の評価を行い，トレーニングデータを一切用いずに良好な認識精度を示すことを確認した．今後の課題として，まずオブジェクト認識の改良が考えられる．ILSVRC2012 データセットには 1000 カテゴリの画像が含まれているが，これらの中には日常生活において使用されないであろうカテゴリが含まれている．今回は全てのカテゴリから学習されたモデルを用いたため，日常生活オブジェクトに特化した認識は行えなかった．ImageNet から日常生活に使用されるカテゴリのみを選出して DCNN を訓練することでオブジェクト認識の精度を向上させられると考える．また，本研究では得られた画像をそのまま用いてオブジェクト認識を行ったが，行動に利用されていると考えられるオブジェクトの画像領域のみを切り出して認識することで精度の向上が期待できる．画像内に多数のオブジェクトが映っている場合でも，ユーザが現在使用しているオブジェクトを判別することができれば，より正確な行動認識が行える．また，行動名の拡張に関して，全ての行動名において適切なオブジェクトが拡張されたとは言えなかった．より行動名と共起度の高いオブジェクトを拡張することができれば類似度計算がより正確に行えると思われる．

謝辞 本研究の一部は，JST CREST および JSPS 科研費 26730047 の助成を受けて行われたものです．

参考文献

[1] Achanta, R., Hemami, S., Estrada, F. and Susstrunk, S.: Frequency-tuned salient region detection, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604 (2009).
[2] Bao, L. and Intille, S. S.: Activity recognition from user-annotated acceleration data, *Pervasive computing*, Springer, pp. 1–17 (2004).
[3] Church, K. W. and Hanks, P.: Word association norms, mutual information, and lexicography, *Computational linguistics*, Vol. 16, No. 1, pp. 22–29 (1990).
[4] Cui, H., Wen, J.-R., Nie, J.-Y. and Ma, W.-Y.: Probabilistic query expansion using query logs, *Proceedings of the 11th international conference on World Wide Web*, Association for Computing Machinery, pp. 325–332 (2002).
[5] Dernbach, S., Das, B., Krishnan, N. C., Thomas, B. L. and Cook, D. J.: Simple and complex activity recognition through smart phones, *IEEE 8th International Conference on Intelligent Environments*, pp. 214–221 (2012).
[6] Felzenszwalb, P. F., Girshick, R. B., McAllester, D. and Ramanan, D.: Object detection with discriminatively trained part-based models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 9, pp. 1627–1645 (2010).
[7] González Díaz, I., Buso, V., Benois-Pineau, J., Bourmaud, G. and Megret, R.: Modeling instrumental activities of daily living in egocentric vision as sequences of active objects and context for alzheimer disease research, *Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare*, pp. 11–14 (2013).
[8] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. and Darrell, T.: Caffe: Convolutional architecture for fast feature embedding, *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678 (2014).
[9] Joachims, T.: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization., Technical report, DTIC Document (1996).
[10] Kwapisz, J. R., Weiss, G. M. and Moore, S. A.: Activity recognition using cell phone accelerometers, *ACM SigKDD Explorations Newsletter*, Vol. 12, No. 2, pp. 74–82 (2011).
[11] Luo, C., Ni, B., Wang, J., Yan, S. and Wang, M.: Manipulated Object Proposal: A Discriminative Object Extraction and Feature Fusion Framework for First-Person Daily Activity Recognition, *arXiv preprint arXiv:1509.00651* (2015).
[12] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J.: Introduction to wordnet: An on-line lexical database*, *International journal of lexicography*, Vol. 3, No. 4, pp. 235–244 (1990).
[13] Pirsiavash, H. and Ramanan, D.: Detecting activities of daily living in first-person camera views, *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 2847–2854 (2012).
[14] Ravi, N., Dandekar, N., Mysore, P. and Littman, M. L.: Activity recognition from accelerometer data, *Association for the Advancement of Artificial Intelligence*, Vol. 5, pp. 1541–1546 (2005).
[15] Ren, X. and Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video, *IEEE Conference on Computer Vision and Pattern*

Recognition, IEEE, pp. 3137–3144 (2010).

[16] Uijlings, J. R., van de Sande, K. E., Gevers, T. and Smeulders, A. W.: Selective search for object recognition, *International journal of computer vision*, Vol. 104, No. 2, pp. 154–171 (2013).

[17] Wang, L., Gu, T., Xie, H., Tao, X., Lu, J. and Huang, Y.: A wearable RFID system for real-time activity recognition using radio patterns, *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, Springer, pp. 370–383 (2014).

[18] Wen, J.-R., Nie, J.-Y. and Zhang, H.-J.: Clustering user queries of a search engine, *Proceedings of the 10th international conference on World Wide Web*, ACM, pp. 162–168 (2001).

[19] Wu, J., Osuntogun, A., Choudhury, T., Philipose, M. and Rehg, J. M.: A scalable approach to activity recognition based on object use, *IEEE 11th International Conference on Computer Vision*, pp. 1–8 (2007).