情報処理学会論文誌 Vol.58 No.10 1664–1673 (Oct. 2017)

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

久賀 稜平[1] 前川 卓也[1,2,a)] 松下 康之[1]

受付日 2016年12月10日, 採録日 2017年7月4日

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．本論文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環境非依存な行動認識手法を提案する．これまでに，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研究は数多くなされているが，その多くがユーザによるトレーニングデータの収集を必要としている．一方本研究では，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，Web 上に存在する知識を用いることによって環境非依存な行動認識を実現する．提案手法では，入力画像から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，ユーザが利用したオブジェクトを認識し，認識したオブジェクトの名前とあらかじめ定義した日常行動の名前との類似度を Web 上の知識を用いて計算することで，環境非依存な行動認識を実現する．

キーワード：行動認識，ウェアラブルセンサ，一人称視点映像

Environment-independent Activity Recognition Based on Web Knowledge Using Egocentric Video

Ryohei Kuga[1] Takuya Maekawa[1,2,a)] Yasuyuki Matsushita[1]

Received: December 10, 2016, Accepted: July 4, 2017

Abstract: In this paper, we recognize daily activities based on a wearable camera without using training data prepared by a user in her environment. Recently, deep learning frameworks have been publicly available, and we can now easily use deep convolutional neural networks (DCNNs) pre-trained on a large image data set. In our method, we first detect objects used in the user's activity from her first-person images using a pre-trained DCNN for object recognition. We then estimate an activity of the user using the object detection result because objects used in an activity strongly relate to the activity. To estimate the activity without using training data, we utilize knowledge on the Web because the Web is a repository of knowledge that reflects real-world events and common sense. Specifically, we compute semantic similarity between a list of the detected object names and a name of each activity class based on the Web knowledge. The activity class with the largest similarity value is the estimated activity of the user.

Keywords: activity recognition, wearable sensor, egocentric video

1. はじめに

近年，GoPro や Google Glass などのウェアラブルカメラの普及により，一人称視点映像を用いた行動認識の研究がさか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

[1] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565–0871, Japan
[2] 国際電気通信基礎技術研究所
Advanced Telecommunications Research Institute International (ATR), Souraku-gun, Kyoto 619–0237, Japan
[a)] maekawa@ist.osaka-u.ac.jp

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．

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さかんに研究されている [16]．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，たとえば，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とらえることで，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ともな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ともな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 [11] では，ユーザが環境ごとにトレーニングデータを収集することを想定しているが，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．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環境非依存な行動認識を提案する．近年，一般オブジェクト認識向けの事前学習されたディープニューラルネットワーク（DNN）が手軽に利用できるようになりつつある [6]．本研究では，DNN 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．たとえば，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．たとえば，「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どうし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．たとえば，「cook」を「cooking」に変換したとしても．概念辞書である WordNet [10] 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．このように，公開されているデータセットで学習されたオブジェクト認識器や Web 上のリソースを用いた環境非依存な行動認識を行う．

また，対象とは異なる実際の環境で得られたセンサデータを用いる環境非依存な手法として，他の環境から得られた画像や加速度データを用いて認識精度を向上させる手法についても検証する．環境が異なっても，ある行動の際に得られる加速度データは類似していると考えられ，行動認識に有用である．また，環境が異なっても，行動に利用されるオブジェクトは類似した画像特徴を持つと期待される．加速度データを用いる場合はその平均や分散を，画像を用いる場合は得られた画像を DNN に入力して中間層から得られる特徴を特徴量とし，Gaussian Mixture Model（GMM）を用いて行動ごとに特徴の分布を学習する．そして，テストデータと各行動ごとの GMM の類似度を計算し，上記の類似度計算に組み込む．

2. 関連研究

ユビキタスセンシングやウェアラブルセンシングを用いた行動認識では，環境の物体に添付したセンサを用いた研究や [8], [16]，ユーザの身体部位に添付した加速度センサを用いた研究 [14] などが多く行われている．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場合にメンテナンス・導入コストが大きくなってしまったり，ユーザの体に複数のセンサを添付する場合にユーザへの負担が大きくなってしまうといった問題がある．また，ウェアラブル加速度センサを用いた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ともな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．

近年は，ウェアラブルカメラが一般的に普及してきており，ウェアラブルカメラから得られる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．

表 1 一人称視点映像を用いた既存研究における行動認識精度
Table 1 Activity recognition accuracies for existing studies that employ ego-centric video.

手法	クラス数	精度 [%]
Luo ら [9]	18	53.0
Pirsiavash ら [11]	18	60.7
Castro ら [3]	19	65.9

Pirsiavash ら [11] は，part-based model [5] 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．Part-based model とはオブジェクトを複数のパーツに分割するモデルであり，たとえば人の場合には，人体を頭，胴体，手，足などのパーツに分割する．このモデルを用い，実際にユーザが行動しているときの一人称視点画像のオブジェクトを学習し，18 種類の行動を認識した．さらに，Luo ら [9] 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．CNN から抽出した特徴をオブジェクトの情報とし，オブジェクトの移動軌跡情報 [13] を動きの特徴として，Pirsiavash らと同様の 18 種類の行動を認識対象としている．また，近年は DNN を用いて行動認識を行う研究も行われている．Castro ら [3] は，一人称視点映像とそれが撮影された時間や曜日などのコンテキスト情報を用いて，DNN により 19 種類の行動を認識した．**表 1** に，一人称視点映像を用いた既存行動認識研究についてまとめた．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「料理をする」,「食器を洗う」などの複雑な行動の認識を行うが，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

提案手法は大きく以下の 4 つのステップに分けることができる．

(1) 行動名の拡張
(2) 注目領域抽出
(3) DCNN を用いた物体認識
(4) 類似度計算

提案手法の概要を**図 1** に示す．図に割り当てられた番号は上記の各ステップの番号と一致している．提案手法ではまず，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

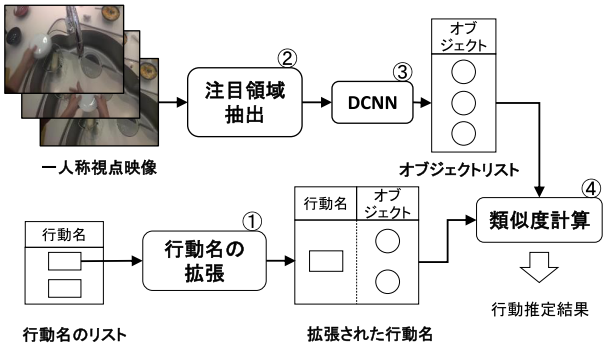

図 1 提案手法の概要
Fig. 1 Overview of proposed method.

対象となる一人称視点映像が得られたとき，スライディング時間窓（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，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．以下の節では，各ステップの詳細について述べる．

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し，これを用いて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．たとえば，Cui ら [4] や Wen ら [15] 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．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るため，行動名をクエリとする web 検索結果に含まれる文書から，単語の重要度を基に行動名に共起するオブジェクトリストを抽出する．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．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) [7] 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

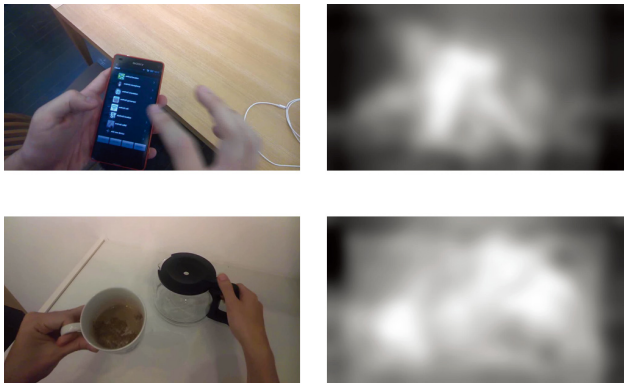

図 2 顕著性マップの例：上段がスマートフォンを操作しているとき，下段がコーヒーを作っているときの画像とその顕著性マップ．右側の画像が顕著性マップを示す．白い領域ほど顕著性が高い

Fig. 2 Examples of saliency map: Upper and lower images correspond to smart phone use and making coffee. Right one is saliency map. Brighter pixel shows higher saliency.

いえ，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いえる．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，得られた重要度の高い単語を，行動名に対応するオブジェクトリストとする．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 注目領域抽出

本研究で得られる入力画像はユーザそれぞれの環境の一人称視点から得られたものであり，環境によってはオブジェクトの周囲に存在するオブジェクトがノイズとなり，DCNN の認識エラーにつながる恐れがある．そこで提案手法では，入力画像に対して Vig らの手法 [12] を用いて人の注目領域を模倣した顕著性マップを作成し，それを基に画像からユーザの注目領域を抽出する．実際に得られた顕著性マップの例を**図 2** に示す．本研究では，作成された顕著性マップから顕著性が閾値より高い点をすべて包含する矩形領域をユーザの注目領域とし，この注目領域を DCNN の入力画像とする．

3.4 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．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 [6] を利用する．Caffe では，約 15 万枚のオブジェクトの画像から構成される ILSVRC2012 データセット[*1]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．

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

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．このオブジェクトリストの抽出に DCNN を用いる．DCNN の出力にはノードがクラスの数だけ存在し，それぞれのノードから出力される値がクラスの分類確率（スコア）となる．このとき，DCNN の認識エラーにより実際には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，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積を計算し，その積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．また，Caffe の学習モデルでは，各画像カテゴリは WordNet [10] の概念の ID となっている．以上まとめると，あるウィンドウに対して，そのウインドウ内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．

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 を利用した類似度計算

1 つ目は WordNet を用いた手法である．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ネットワーク構造で記述されている．そこで，WordNet を用いて行動名とオブジェクトリストとの類似度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．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，$V(Y)$ はオブジェクト Y の DCNN 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，$W(X,Y) = 1/D(X,Y)$ で定義する．D は WordNet 上での 2 つのオブジェクト X，Y 間の最短経路のホップ数である．

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 のリストどうしの類似度計算となる．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．

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

類似度計算にオブジェクトのスコアを用いることで，ウィンドウ内に頻出するオブジェクトほど類似度が大きくなるように重みづけされた計算ができる．

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，語どうしの類似度を計算する手法について述べる．以降の類似度計算指標は検索エンジンを用いた後の類似度計算によく用いられている [2]．

相互情報量を用いた手法

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．

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．

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

$h(q)$ は “q”，$h(q_1, q_2)$ は “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である．また，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上記のように定義できる．

Jaccard 係数を用いた手法

Jaccard 係数とは以下で定義される類似度である．

$$J(x, y) = \frac{h(x, y)}{h(x) + h(y) - h(x, y)}$$

よって，オブジェクトリスト間の距離を以下の式で計算する．

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

Dice 係数を用いた手法

Dice 係数とは以下で定義される類似度である．

$$D(x, y) = \frac{2h(x, y)}{h(x) + h(y)}$$

よって，オブジェクトリスト間の距離を以下の式で計算する．

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

Web 検索におけるヒットカウントを基にしたこれらの類似度は，クエリとなる 2 つの単語がどの程度文書を共有しているかを示すことになる．2 つの単語が同じ文書を共有していればいるほど，これらの類似度は高くなる．

3.6 他環境で得られたデータを用いた類似度計算

他環境で得られたラベリングありデータを再利用して類似度計算をする場合，他環境でそれぞれの行動から得られる画像や加速度の特徴をラベルありデータを用いて GMM のパラメータをあらかじめ学習しておく．GMM を用いることで，特徴量を複数の正規分布の混合分布で表現することが可能である．画像から抽出する特徴には，本研究で用いた DCNN の中間層から得られる 4096 次元の特徴を用い，加速度から抽出する特徴には，3 軸それぞれの平均および分散の計 6 次元の特徴を用いる．以降の評価実験では，GMM を学習する際，(1) 加速度と画像，(2) 画像のみ，(3) 加速度のみを利用する計 3 パターンについて検証を行う．ここで，推定したい行動の集合を $\mathcal{A}$ したとき，$\mathcal{A}_i$ を i 番目の種類の行動と表記する．また，時刻 t において，提案手法により得られる i 番目の行動との類似度を 3.5 節と同様に定義し，$S_{se}(\mathcal{O}_{act_i}, \mathcal{O}_{img_t})$ とする．このとき，時刻 t でのある行動との類似度 S_{re} は以下の式で定義される．

$$S_{re}(\mathcal{O}_{act_i}, \mathcal{O}_{img_t}, M_i, s_t) = \lambda S_{se}(\mathcal{O}_{act_i}, \mathcal{O}_{img_t}) + (1 - \lambda) S_{sd}(M_i, s_t) \quad (1)$$

ここで，s_t は時刻 t におけるセンサデータ，i 番目の行動の GMM との尤度（類似度）を $S_{sd}(M_i, s_t)$ とする．また，M_i は i 番目の行動の特徴から学習されるガウス分布であり，λ は 0 から 1 で定義される重みである．すなわち，上式は図 1 における類似度計算の処理において用いられる式に，GMM との尤度を組み合わせた式となる．このように定義される類似度が最も高い行動を時刻 t における推定結果とする．

3.7 スムージング

提案手法ではウィンドウごとに行動認識を行うが，行動中ユーザのよそ見や DCNN の認識エラーによって行動とは関係のないオブジェクトが映ることでノイズが発生することが考えられる．そこで提案手法では，ウィンドウごとに類似度計算を行った後，その前後のウィンドウの類似度も用いることでこのようなノイズの影響を低減させることを考え，各ウィンドウの類似度をその前後数ウィンドウの類似度との平均値とする．

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 2** 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 1,280 × 720 ピクセルの JPEG 画像を 30 fps で撮影する．さらに，Glass には 3 軸加速度センサが搭載されており，サンプリングレートは 30 Hz である．表 2 の 13 種類の行動名は既存の行動認識研究論文 [9], [11] において利用されているものからオブジェ

表 2 実験で行った 13 クラスの行動とその平均時間
Table 2 Activities performed in experiment and their average durations.

行動名	平均時間（秒）
using cellphone	40.3
making tea	35.6
using computer	51.7
toilet	16.6
watering plants	25.3
watching television	51.7
cooking	66.8
eating	55.1
using microwave	23.9
making coffee	34.1
washing dishes	42.0
playing with pet	36.5
using curtain	11.0

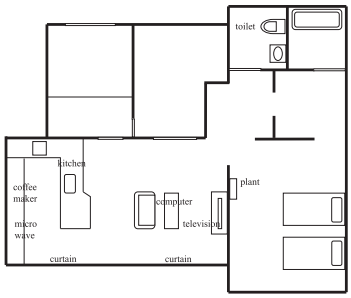

図 3 環境 1 の見取り図
Fig. 3 Floor plan of environment 1.

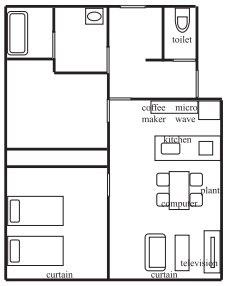

図 4 環境 2 の見取り図
Fig. 4 Floor plan of environment 2.

クトの利用をともなうものを基本的に用いた．また，本研究が一人称視点映像を用いているため，Glass のカメラにオブジェクトが映り込むような行動のみを用いた．実験では，2 名の被験者が 3 つの環境で 13 種類の行動が含まれるセッションを 5 回ずつ行った．このとき，1 名の被験者は全 3 環境で実験を行い．残りの 1 名は 2 環境のみで実験を行った．各環境の見取り図を**図 3**，**図 4**，**図 5** に示す．各セッションの平均時間は約 15 分である．また，被験者が行う各行動と行動の間には平均して約 20 秒の間隔が含まれる．データの取得方法には semi-naturalistic collection protocol [1] と呼ばれる方法を用いた．この手法では，被験

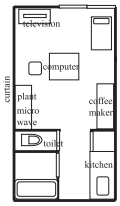

図 5 環境 3 の見取り図
Fig. 5 Floor plan of environment 3.

者に対して行ってほしい行動の一覧をランダムな順で提示をする．具体的にどのように振る舞ってほしいかは伝えない．したがって，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．また提案手法において，ウィンドウ幅を 0.1 秒分に設定し，スムージングに前後合わせて 15 枚のウィンドウを用いた．クエリ拡張の際には，行動名をクエリとして取得した上位 20 ページから抽出した重要度の高い上位 2 つを拡張するオブジェクトとして選択した．

4.2 評価手法

4.2.1 提案手法

評価実験では以下の 8 つの手法を比較・評価する．

(1) WN：WordNet を用いた類似度計算
(2) WMI：相互情報量を用いた類似度計算
(3) WJ：Jaccard 係数を用いた類似度計算
(4) WD：Dice 係数を用いた類似度計算
(5) WN+：行動名の拡張+WordNet を用いた類似度計算
(6) WMI+：行動名の拡張+相互情報量を用いた類似度計算
(7) WJ+：行動名の拡張+Jaccard 係数を用いた類似度計算
(8) WD+：行動名の拡張+Dice 係数を用いた類似度計算

(1)，(2)，(3)，(4) は行動名の拡張を行っていない場合の手法である．

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．なお，本論文における適合率，再現率および F 値は各クラスにおいてその値を計算した平均を記載している．

4.3 結果

提案手法および他環境データを学習する両手法について，その認識精度を示す．

4.3.1 提案手法の認識精度

表 3 にそれぞれの手法の認識精度を示す．また，**図 6** にそれぞれの手法の混同行列を示す．ただし，これらは他環境のセンサデータを再利用していない結果である．まず，クエリ拡張を行わない手法では全体的に認識精度が良くな

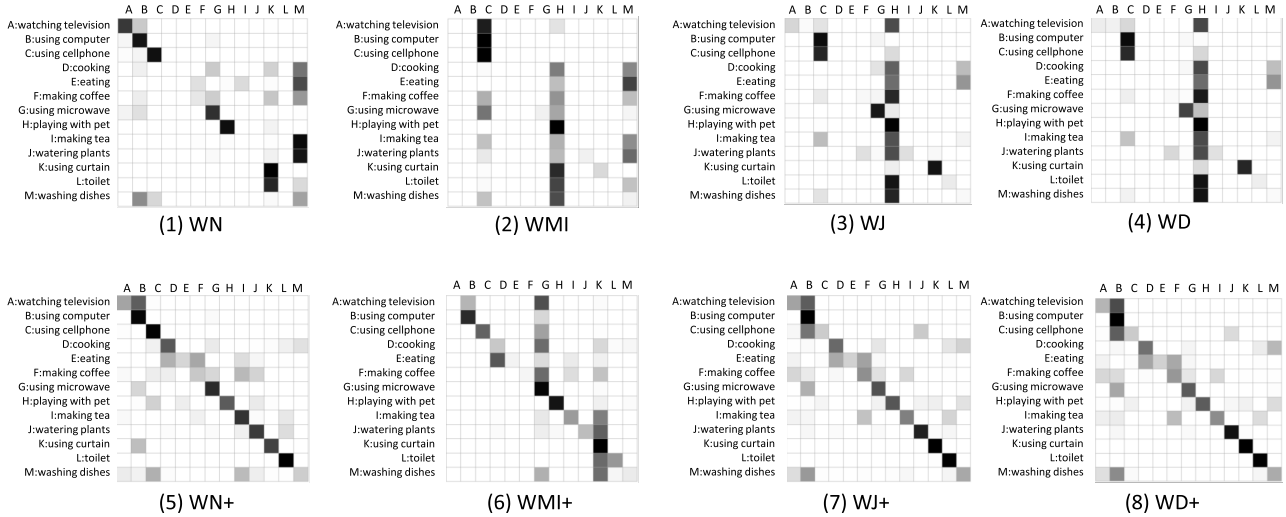

図 6 それぞれの手法の認識結果の混同行列
Fig. 6 Confusion matrices of methods.

表 3 それぞれの手法の認識精度
Table 3 Recognition accuracies for methods.

	precision [%]	recall [%]	F-measure [%]
WN	33.9	45.2	35.9
WMI	26.4	17.8	9.1
WJ	32.9	30.7	22.3
WD	33.4	27.8	20.7
WN+	63.8	64.3	59.2
WMI+	61.6	44.7	38.1
WJ+	64.4	60.3	56.3
WD+	64.3	58.9	55.8

かったことが分かる．Web 検索を用いた手法においては偏ったクラスに認識されており，WordNet を用いた手法においては，より多くのオブジェクトの利用をともなうような複雑な行動ほど誤ったクラスに認識されている．しかし，WordNet，Web 検索を用いた両手法について，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．Cooking，Eating などの行動名と「pot」，「plate」などのオブジェクト名との WordNet における距離は大きかったが，行動名をオブジェクトで拡張することでオブジェクト名どうしの距離計算ができたため，類似度計算の精度が上がった．

さらに，すべての行動において，行動中につねに映像内にオブジェクトが映っているとは限らず，たと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，たとえば Making coffee と Making tea では同じオブジェクト（cup）を使用するように，複数の行動で同じオブジェクトが使用される場合がある．各オブジェクトが 1 つの行動のみと対応しているとは限らず，認識精度の向上が困難であったと思われる．

Web 検索を用いた手法に関して，WMI+の精度が最も低かった．図 6 の WMI+の結果では，多くの行動が Using microwave，もしくは Using curtain に推定されてしまっていることが分かる．相互情報量では分母が乗算となっているため，他の手法よりも 1 つの単語の検索結果の影響を受けやすい．たとえば Watching television では，television が実際に拡張された単語であり，これの検索結果ページ数は約 8,000 万であった．一方で Using microwave では，microwave が拡張されるが，これの検索結果ページ数は約 1,500 万である．この差が影響されやすくなっているため，特定の行動に偏って推定されてしまったと考えられる．また，Jaccard 係数と Dice 係数を用いた手法の精度はほぼ変わらなかった．Dice 係数は Jaccard 係数に比べてその計算結果に積集合のサイズが影響しやすいため，積集合のサイズが小さい場合でも（類似度が低い場合でも）類似度の差異を表現しやすい．しかし本研究のタスクは最も類似度の高い行動を決定するタスクであるため，類似度が低い場合の表現能力に違いがあるこれらの係数の違いの影響が小さかったものと考える．また，WN+は WJ+や WD+に比べて Using cellphone の精度が高かった．WJ+や WD+では cellphone と computer の類似度が高く，Using computer に誤って分類されていた．これらのオブジェクト名が多くのページで共起して現れるため，類似度が高くなったと考える．また，表 3 に示すように，WN+の精度が最も高かった．表 1 に示す既存研究と比較してもほぼ同様の精度を達成していた．本研究で用いたクラス数は既存研究より若干少ないが，本研究では環境非依存の行動認識を行っている．

4.3.2 他環境データを再利用した場合の認識精度

本研究において収集された 3 つの環境のうち，2 つの環境をトレーニングデータとし，1 つをテストデータとすることで，他環境データを再利用する手法の評価実験を行った．この際，他環境で収集されたデータであっても，テストユーザの行ったセッションはトレーニングデータに含

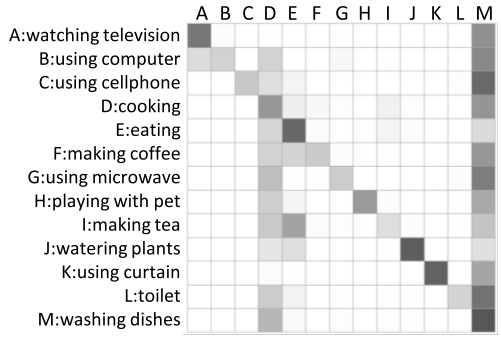

図 7 画像特徴のみを再利用した場合の認識結果の混同行列：$\lambda = 0$

Fig. 7 Confusion matrix when only image features are re-used: $\lambda = 0$.

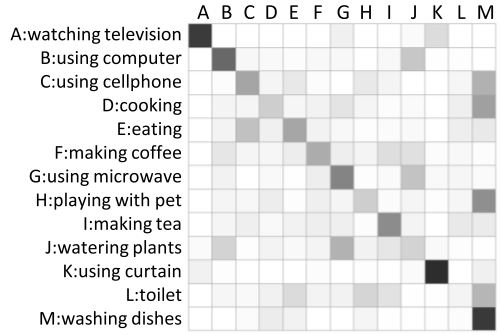

図 8 加速度特徴のみを再利用した場合の認識結果の混同行列：$\lambda = 0$

Fig. 8 Confusion matrix when only acceleration features are re-used: $\lambda = 0$.

まないようにした．なお，類似度計算の手法に関しては，WN+を用いた．

まず，式 (1) における λ の値を 0 にしたとき，すなわち，他環境から得られる各特徴量から GMM を構築し，それのみを用いて行動推定を行った場合の精度を示す．**図 7** が画像特徴，**図 8** が加速度特徴のみを再利用した場合の混同行列である．図に示されるように画像特徴を用いた場合には Eating が，加速度特徴を用いた場合では Washing dishes が提案手法と比べ特に精度が高いことが分かる．しかし，このとき平均 F 値はそれぞれ 42.5%，40.4%とほぼ同様の精度となっており，提案手法と比べて低い値となっている．他環境の画像のみを再利用した手法では，環境ごとにオブジェクトの画像特徴が異なると精度が低下する．他環境の加速度データのみを再利用した手法では，加速度データに違いが少ない行動を識別できない．

次に他環境データを用いる手法と提案手法の組合せについて検証する．ここで，式 (1) の λ の値を変動させることで λ の影響を考察する実験を行った．λ の値と認識精度（F 値）との関係を表すグラフが**図 9** である．この結果から λ の値としては 0.96 が最も良い値であると判断した．ただし，$S_{se}(\mathcal{O}_{act_i}, \mathcal{O}_{img_t})$ と $S_{sd}(M_i, s_t)$ の値のスケールは異なることに注意されたい．実験では，$S_{se}(\mathcal{O}_{act_i}, \mathcal{O}_{img_t})$ と $S_{sd}(M_i, s_t)$ の値の平均は 0.303 と 1.0 であった．**図 10** に

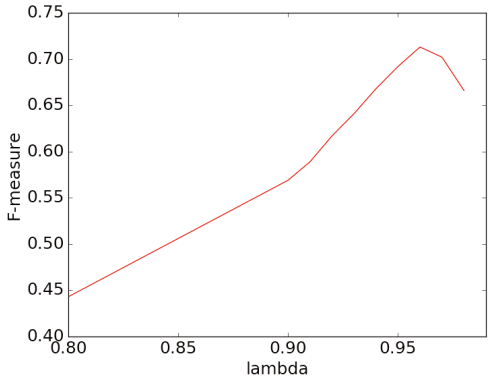

図 9 λ と認識精度の関係：画像と加速度両方の特徴を使用

Fig. 9 Relationship between λ and recognition accuracy: both image and acceleration features are used.

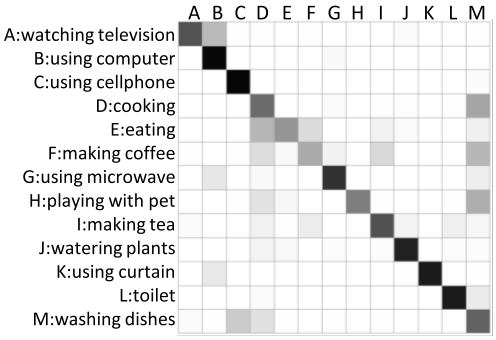

図 10 画像と加速度特徴を再利用した場合の認識結果の混同行列：$\lambda = 0.96$

Fig. 10 Confusion matrix when image and acceleration features are re-used: $\lambda = 0.96$.

表 4 画像と加速度特徴を再利用した場合の認識精度：$\lambda = 0.96$

Table 4 Recognition accuracies when image and acceleration features are re-used: $\lambda = 0.96$.

	precision [%]	recall [%]	F-measure [%]
WN+	75.6	71.2	71.3
WMI+	68.0	41.5	45.8
WJ+	67.4	62.4	59.3
WD+	68.6	61.2	55.4

表 5 画像のみを再利用した場合の認識精度：$\lambda = 0.96$

Table 5 Recognition accuracies when only image features are re-used: $\lambda = 0.96$.

	precision [%]	recall [%]	F-measure [%]
WN+	74.6	70.4	70.3
WMI+	66.1	42.0	46.2
WJ+	67.7	63.3	60.1
WD+	69.0	62.1	60.0

$\lambda = 0.96$ としたときの認識結果の混同行列を示す．また，**表 4**，**表 5**，**表 6** に画像と加速度両方の特徴を再利用した場合と，それぞれ単独で再利用した場合の認識精度を示す．他環境で収集されたラベルありデータを再利用することにより，提案手法よりも高い精度を示すことが分かった．特

表 6 加速度のみを再利用した場合の認識精度：$\lambda = 0.96$
Table 6 Recognition accuracies when only acceleration features are re-used: $\lambda = 0.96$.

	precision [%]	recall [%]	F-measure [%]
WN+	71.4	70.0	67.0
WMI+	54.0	53.3	50.4
WJ+	68.2	64.7	61.4
WD+	69.0	64.4	61.8

に提案手法では，Watching television は Using computer に誤分類されてしまうことが多かったが，加速度データを用いることにより，精度が大きく改善された．ImageNet に登録されている「テレビ」と「コンピュータ」には画像特徴的な違いがあまりなかったが，頭の姿勢に明確な違いがあったため，精度が向上したと思われる．また，画像特徴を用いることで特に Eating の認識精度が向上した．Eating では皿やカップ，フォークといったオブジェクトが検出されるが，そういったオブジェクト情報のみを用いた場合，提案手法では Eating は Cooking に分類されることが多かった．Cooking でも同じオブジェクトが出現することがあり，それらのオブジェクト名の間に意味的な違いはないためである．しかし，画像の特徴量を用いることで，食事中の机や料理中のキッチンなどの要素も考慮することができるようになり，精度が向上したものと考えられる．

5. おわりに

本研究では，Web 上に存在する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評価を行い，トレーニングデータをいっさい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，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 1,000 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．そこで，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．

謝辞 本研究の一部は，JST CREST JPMJCR15E2 の助成を受けて行われたものです．

参考文献

[1] Bao, L. and Intille, S.S.: Activity recognition from user-annotated acceleration data, *Proc. Pervasive Computing*, pp.1–17, Springer (2004).
[2] Bollegala, D., Matsuo, Y. and Ishizuka, M.: Measuring semantic similarity between words using web search engines, *www*, Vol.7, pp.757–766 (2007).
[3] Castro, D., Hickson, S., Bettadapura, V., Thomaz, E., Abowd, G., Christensen, H. and Essa, I.: Predicting Daily Activities from Egocentric Images Using Deep Learning, *Proc. 2015 ACM International Symposium on Wearable Computers*, *ISWC '15*, pp.75–82, ACM (2015).
[4] Cui, H., Wen, J.-R., Nie, J.-Y. and Ma, W.-Y.: Probabilistic query expansion using query logs, *Proc. 11th International Conference on World Wide Web*, pp.325–332 (2002).
[5] Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D.: Object detection with discriminatively trained part-based models, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.32, No.9, pp.1627–1645 (2010).
[6] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. and Darrell, T.: Caffe: Convolutional architecture for fast feature embedding, *Proc. ACM International Conference on Multimedia*, pp.675–678 (2014).
[7] Joachims, T.: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, Technical Report, DTIC Document (1996).
[8] Lowe, D.G.: Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, Vol.60, No.2, pp.91–110 (2004).
[9] 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).
[10] 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).
[11] Pirsiavash, H. and Ramanan, D.: Detecting activities of daily living in first-person camera views, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.2847–2854 (2012).
[12] Vig, E., Dorr, M. and Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.2798–2805 (2014).
[13] Wang, H. and Schmid, C.: Action recognition with improved trajectories, *Proc. IEEE Conference on Computer Vision* (*ICCV*), pp.3551–3558 (2013).
[14] 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*, pp.370–383, Springer (2014).
[15] Wen, J.-R., Nie, J.-Y. and Zhang, H.-J.: Clustering user queries of a search engine, *Proc. 10th International Conference on World Wide Web*, pp.162–168 (2001).
[16] Wu, J., Osuntogun, A., Choudhury, T., Philipose, M. and Rehg, J.M.: A scalable approach to activity recognition based on object use, *Proc. IEEE 11th International Conference on Computer Vision*, pp.1–8 (2007).

久賀 稜平 （学生会員）

2016 年大阪大学工学部電子情報工学科卒業，同大学大学院情報科学研究科マルチメディア工学専攻博士前期課程入学．機械学習による画像認識の研究に従事．

前川 卓也 （正会員）

2003 年大阪大学工学部電子情報エネルギー工学科卒業．2006 年同大学院情報科学研究科博士後期課程修了．同年日本電信電話株式会社入社．2012 年 4 月より大阪大学大学院情報科学研究科准教授．2013 年 8～10 月スイス連邦工科大学ローザンヌ校招聘教授．博士（情報科学）．本会平成 22 年度山下記念研究賞，日本データベース学会平成 25 年度上林奨励賞等受賞．ACM，IEEE，電気学会，日本データベース学会各会員．

松下 康之

1998 年東京大学工学部卒業．2003 年同大学大学院工学系研究科電子情報工学博士後期課程修了．同年 Microsoft Corp. に入社し Microsoft Research Asia の Visual Computing Group に研究員として勤務．2015 年 4 月より大阪大学情報科学研究科教授，現在に至る．コンピュータビジョン・機械学習・最適化の研究に興味を持つ．