Vol.2016-UBI-50 No.12
2016/5/29

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の提案

久賀稜平[1]　前川卓也[1]　松下康之[1]

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．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，ユーザによる事前学習を必要としない本手法が良好な認識精度を示すことを確認した．また，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討した．

1. はじめに

近年，GoPro や Google Glass 等のウェアラブルカメラの普及により，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．

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 [11]．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 [7] 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 [3]．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

[1] 大阪大学大学院情報科学研究科

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 [6] 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．

また，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討する．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 [4], [11]，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 [10] 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．

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．Pirsiavash ら [7] は，part-based model [2] 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，例えば人の場合には，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時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 [5] 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 [9] 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．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．提案手法の概要を図 1 に示す．

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

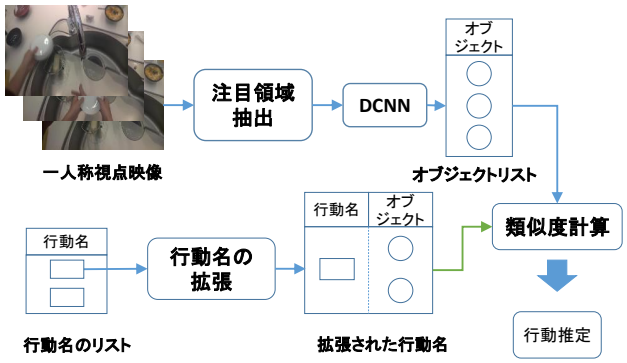

図 1　提案手法の概要

れた行動名を補完する．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．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 [8] 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．本研究では，作成された顕著性マップからもっとも顕著性が高い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 [3] 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．

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 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．

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

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 [6] 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，$V(X)$ はオブジェクト X 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 のリスト同士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，語同士の類似度を計算する手法について述べる．

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．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 節と同じように定義し，$S_{se}(\mathcal{O}_{act_i},\mathcal{O}_{img_t})$ とする．ここで，s_t は時刻 t におけるセンサデータとし，i 番目の行動の GMM との尤度（類似度）は $S_{sd}(M_i,s_t)$ とするとき，時刻 t でのある行動との類似度 S_{re} を以下の式で定義する．

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

ここで，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．上式で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 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．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 30Hz である．また，表 1 の行動名は既存の行動認識研究論文 [5], [7] において利用されているものを基本的に用いた．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．各セッションの平均時間は約 15 分である．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．

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

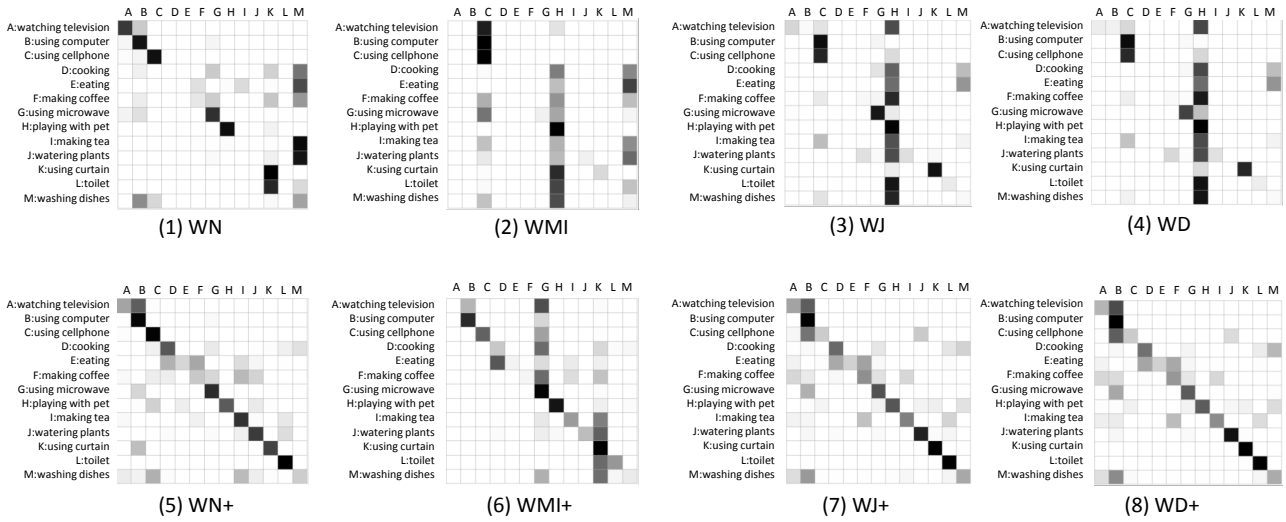

図 2 それぞれの手法の認識結果の混同行列

表 1 実験で行った 13 クラスの行動

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

表 2 それぞれの手法の認識精度

	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

(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．

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．

表 3 加速度と画像を再利用した場合の認識精度

	precision [%]	recall [%]	F-measure [%]
WN+	72.3	69.7	69.0
WMI+	73.5	49.6	54.9
WJ+	67.0	62.8	59.4
WD+	67.7	61.6	59.2

4.3.1 提案手法の認識精度

表 2 にそれぞれの手法の認識精度を示す．また，図 2 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かったことがわかる．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同士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，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，例えば 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．

表 4　画像のみを再利用した場合の認識精度

	precision [%]	recall [%]	F-measure [%]
WN+	72.6	71.0	69.8
WMI+	74.9	49.5	54.8
WJ+	67.3	62.9	59.7
WD+	68.1	62.0	59.7

表 5　加速度のみを再利用した場合の認識精度

	precision [%]	recall [%]	F-measure [%]
WN+	75.9	76.8	73.8
WMI+	73.5	74.4	71.2
WJ+	68.4	64.8	61.3
WD+	69.4	64.8	62.3

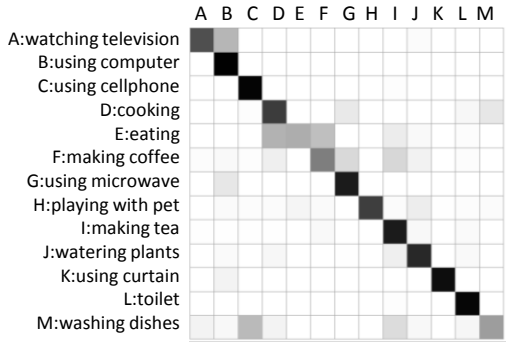

図 3　加速度特徴のみを再利用した場合の認識精度

4.3.2　他環境データを再利用した場合の認識精度

表 3 に画像と加速度の特徴を再利用した場合，表 4 に画像の特徴のみを再利用した場合，表 5 に加速度の特徴のみを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，提案手法よりも良い精度を示すことがわかった．また，これらの表に示されるように，加速度の特徴のみを再利用した手法で最も良い結果となっている．図 3 に加速度の特徴のみを再利用した場合の認識結果の混同行列を示す．特に提案手法では，Watching tv 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，頭の位置や動きには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提案手法よりも精度が低下している．

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，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．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 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, *Proceedings of Pervasive computing*, Springer, pp. 1–17 (2004).
[2] 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).
[3] 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).
[4] Lowe, D. G.: Distinctive image features from scale-invariant keypoints, *International journal of computer vision*, Vol. 60, No. 2, pp. 91–110 (2004).
[5] 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).
[6] 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).
[7] Pirsiavash, H. and Ramanan, D.: Detecting activities of daily living in first-person camera views, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2847–2854 (2012).
[8] Vig, E., Dorr, M. and Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2798–2805 (2014).
[9] Wang, H. and Schmid, C.: Action recognition with improved trajectories, *Proceedings of IEEE Conference on Computer Vision (ICCV)*, pp. 3551–3558 (2013).
[10] 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).
[11] Wu, J., Osuntogun, A., Choudhury, T., Philipose, M. and Rehg, J. M.: A scalable approach to activity recognition based on object use, *Proceedings of IEEE 11th International Conference on Computer Vision*, pp. 1–8 (2007).