Vol.2015-CVIM-198 No.19
2015/9/15

リストマウントカメラによる環境変化に頑健な日常生活動作認識

大西 克典[1,a)]　原田 達也[1,b)]

概要：本研究では，ウェアラブルカメラによる一人称視点からの日常生活動作認識の新規手法を提案する．ウェアラブルカメラによる日常生活動作認識においては，従来ヘッドマウントカメラやチェストマウントカメラといった人の視点に近い位置からの映像が用いられてきた．そのため，背景映像を事前に取得した環境でしか使えなかったり，学習データの用意に多大な労力を要するという問題があった．そこで我々は，このような問題を解決するために操作物体をより大きく捉えられる手首に装着したカメラ，すなわちリストマウントカメラから撮影した日常生活動作を提案する．加えて，動画像全体の特徴を表現するための新たな手法を提案し，それが従来の動画像表現より日常生活動作認識精度を大きく向上させることを示した．

1. 序章

近年，カメラ機能のついた様々なウェアラブルデバイスの開発に伴い，一人称視点からの動作認識に関する研究が盛んに行われるようになった．そうした一人称視点からの動作認識において，認識の対象として日常生活動作 (Activities of Daily Living (ADL)) があげられる．日常生活動作とは，”手を洗う”，”水を飲む”といった基本的な日常生活における動作のことである．日常生活動作が把握できることで，介護・リハビリへの活用や生活習慣の改善といった応用が可能となる．

日常生活動作を認識するためには，それぞれの動作に関連した操作物体に注目することが大切である．例えば，”コーヒーを作る”という動作であれば，マグカップとコーヒーパックに注目していれば可能であり，”掃除機をかける”という動作であれば掃除機に注目していれば可能である．Pirsiavash ら [1] は，チェストマウントカメラから撮影された日常生活動作データセットを作成し，検出された物体及び人の腕の位置から操作物体を特定する手法を用いて，多様な環境における日常生活動作認識を可能とし，実世界における日常生活動作認識においては動作に関わる物体の認識こそが重要であると示した．

しかし，こうした従来の多様な環境におけるヘッドマウントカメラ (HMC) やチェストマウントカメラ (CMC) といった一人称視点カメラによる日常生活動作認識は，物体検出を必要とするため，以下の問題点がある．(1) 物体の位置の特定は完璧に行えるわけではない．(2) 手が常に映像内に映っているとは限らずカメラの視野外で物体操作が行われることもある．(3) 物体の検出を行う上では，学習データ作成時に，大量の画像にバウンディングボックスを付与する必要があり，学習データの準備に多大な労力を要する．

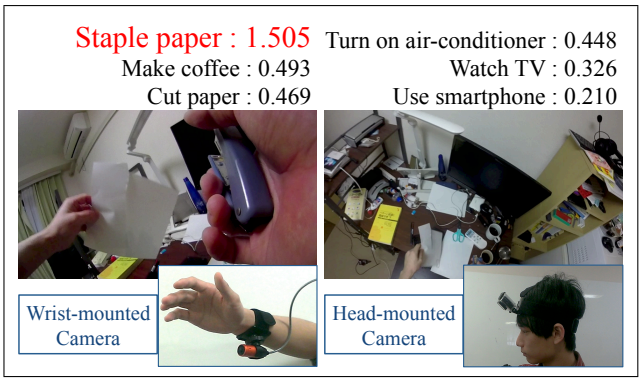

図 1　リストマウントカメラ (左) 及びヘッドマウントカメラ (右) による日常生活動作認識

そういった問題点を解決するために，本研究では，日常生活動作において主に手を利用して物体操作を行うことに着目し，手首にカメラを装着する．手首に装着したカメラ，すなわちリストマウントカメラ (WMC) から撮影された映像であれば，以下の利点がある．(1) 操作物体以外の物体が大きく映り込むことが少ない．(2) 常に操作物体が捉え続けられる．そのため (3) 物体の検出を行う必要がなくなり，物体のバウンディングボックスを付与する必要がなく

[1] 東京大学
The University of Tokyo
[a)] ohnishi@mi.t.u-tokyo.ac.jp
[b)] harada@mi.t.u-toyko.ac.jp

ⓒ 2015 Information Processing Society of Japan

なる．つまり，動作が行われている区間のみを指定するだけでよく，訓練動画作成の労力を大幅に削減できる．

そこで本研究では，リストマウントカメラによる映像を用い，学習データの用意が容易で，より正確かつ環境変化に頑健な日常生活動作認識システムを提案する．

本研究の貢献は以下の通りである．

- リストマウントカメラ及びヘッドマウントカメラから撮影した日常生活動作データセットを作成し，リストマウントカメラとヘッドマウントカメラの公平な比較を可能にした．
- 作成したデータセットを用いて動作認識の実験を行い，リストマウントカメラの有用性を示した．
- 新たな動画像表現手法を開発した．
- 日常生活動作認識において，動画像の新たな認識システムを提案し，実験によりその有用性を示した．

2. 関連研究

2.1 日常生活動作認識

従来，日常生活動作認識は様々な形で行われてきた．手首に加速度センサを操作物体に RFID タグを付けての認識 [2] や操作物体に付けた RFID タグと固定カメラによる三人称視点映像を組み合わせての認識 [3] といった研究が行われてきたが，これらでは事前の準備に多大な労力を要するだけでなく限定的な環境でしか使えない．しかしウェアラブルデバイスであれば，装置の準備が簡単で活動範囲に制限もない．そうした背景と Gopro や GoogleGlass のようなカメラ機能の付いたウェアラブルデバイスの開発に伴い，ヘッドマウントカメラからの映像を用いての日常生活動作認識が行われるようなった．

2.2 Egocentric Vision

一人称視点からの日常生活動作のパイオニアとして Ren ら [4] の研究がある．Ren らは日常生活動作を推定する上で操作物体を認識することが重要であることに着目し，42 種類の日用品の画像データセットと，実生活環境で撮影された 10 本の一人称視点からの日常生活動作動画データセットを作成し，それらの定量的評価を行った [4]．そしてそのデータセットに対し，オプティカルフローを活用して前景と背景を分離し，前景と判定された領域に対して HOG を利用した物体検出を用いることで一人称視点からのより精度の高い操作物体の認識を行った [5]．Faith ら [6] は操作物体と腕を背景から切り出す方法を開発し，これを用いて弱教師付き学習で室内における目線カメラからの動作認識を行った．しかしこれは，限定的な環境にしか対応できない．より自然な日常生活上の環境で動作認識を行う試みとして，腕の位置を検出した後に操作物体を特定する研究 [7], [8] があるが，そもそも腕がカメラの視界からしばしば消えたり操作物体が腕に隠れ見えなくなってしまうという問題もある．また腕が完全に検出されたとしても操作物体を完全に検出することは難しい．Pirsiavash ら [1] は，チェストマウントカメラからの映像に対し，物体と腕の検出を行うことで操作物体の推定を行った．操作物体が完全に検出された状態であれば，多様な環境でも高い精度で動作を認識できることを示したが，完璧な検出そのものが困難である上に，これを行うためには学習時に，映像内の全ての物体に位置の情報を与える必要がある．Pirsiavash らは作成したデータセットに，動作の種類とそれが行われた時間，物体の種類と位置，人の腕の位置といった情報を付与したが，これには 10 人が行って 1 ヶ月要した．これは新たな動作を加える度に多大な労力を要するため実用的ではない．実用的な条件下ではウェアラブルデバイスによる，より環境変化にロバストで，どの区間にどの動作が行われているのかといった簡単な情報だけで学習できる日常生活動作認識が必要である．

2.3 動画像の特徴表現

動画像の特徴表現は動作認識分野において盛んに研究されている．訓練用動画像から局所特徴を得て，局所特徴をコーディングするためのモデルを構築するのが，動作認識に用いる動画像特徴を得る一般的なパイプラインである．HOF [9] や MBH [10] などといった，動き特徴を表現する局所特徴量がよく用いられる．局所特徴のコーディング手法として，Bag of Visual Words (BoVW) [11], [12] が用いられていたが，Wu らは，動作認識において VLAD [13]，Fisher Vector (FV) [14] を BoVW の代わりに用いる手法を提案した [15]．局所特徴量とカメラの動きを取り除ける improved Dense Trajectory (iDT) [16] を組み合わせることで動作認識における最高水準の動作認識結果が得られた [17].

物体認識分野では，近年 Deep Convolutional Neural Networks (DCNN) が優れた性能を示しており，物体認識において標準的なパイプラインとなっている [18]．THUMOS2014 [19] の優勝チームは，DCNN の大規模動作認識への応用を行った．Jain ら [20] は，動画の各フレームから DCNN 特徴を抽出し，その全特徴量を平均化することで動画像全体を表現する特徴を開発した．この簡潔な手法は各フレームの全体的な特徴を取ってきているだけであるが，iDT のような動き特徴と組み合わせることで非常に高い性能を発揮する．我々の取り組む問題は動作認識と物体認識の中間に位置する．そのため我々は前述した研究を活用して，リストマウントカメラからの日常生活動作認識に適した動画特徴を開発する．

2.4 リストマウントカメラ

前川ら [21] は，カメラ，マイク，加速度計，照度計，デジタルコンパスを搭載したリストマウント型デバイスによるマルチモーダルな日常生活動作認識を行った．しかしこれ

ⓒ 2015 Information Processing Society of Japan

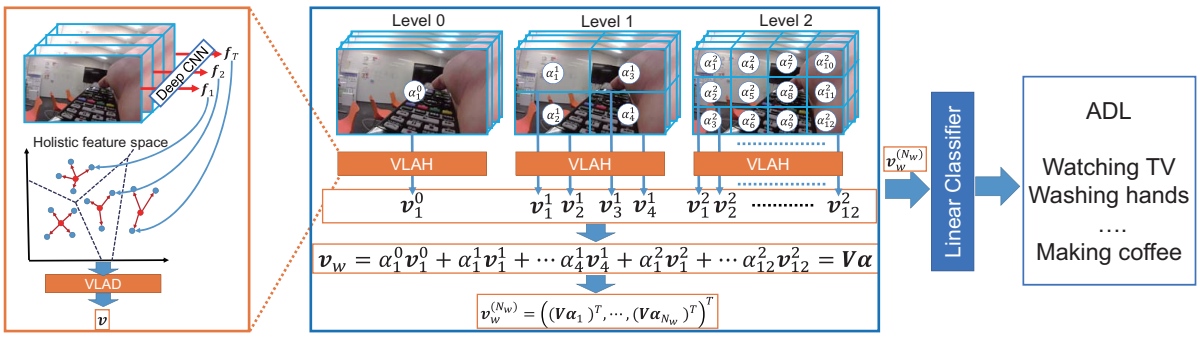

図 2 システムパイプライン

はカメラからはカラーヒストグラムを抽出するだけで，操作物体を認識するには不十分である．そこで，我々の研究ではリストマウントカメラの特性を十分に分析し，その有効性を生かした，動画像のみからの日常動作認識を行う．.

3. 提案手法

3.1 システムパイプライン

図 2 は，リストマウントカメラを用いた我々の日常生活動作認識のパイプラインを示したものである．まず，動画の各フレームをピラミッド領域に分割する．そして各領域から動画特徴を抽出する．最終的な動画特徴は，各領域の動画特徴に識別的な重みをかけた物を足し合わせることで得られる．日常生活動作のクラスを予測するには，最終的な動画特徴に線形識別器を用いることで得られる．

3.2 Vector of Locally Aggregated Holistic Features

まず，提案する動画表現と基本的なパイプラインについて説明する．1 章で述べたように，1 つの動作はいくつかの操作物体によって決定されるので，動画表現はそういった日常生活動作特有の問題を上手く扱えるようなものである必要がある．

T 枚のフレームから成る動画 $\mathcal{V} = \{I_1, I_2, \cdots, I_T\}$ がある時，各フレームに特徴抽出器 (e.g., DCNN) を用いて各フレームから大域特徴量を得る．こうして以下の様な大域特徴量の集合 $\mathcal{F}$ を得る．

$$\mathcal{F} = \{\boldsymbol{f}_t \in \mathbb{R}^d\}_{t=1}^{T}. \tag{1}$$

$\mathcal{F}$ から動画特徴を抽出する上で，各フレームから得た大域特徴の集合を局所特徴のようにみなし，VLAD をフレームの大域特徴の集合に以下のように適用する．

$$\boldsymbol{v}_i = \sum_{f \in \mathcal{R}_i} (\boldsymbol{f} - \boldsymbol{c}_i), \tag{2}$$

ただし，$\boldsymbol{v}_i$ は i 番目の VLAD の構成要素で，$\{\boldsymbol{c}_i\}_{i=1}^{M}$ は訓練動画のフレームの大域特徴からクラスタリング手法 (e.g., kmeans) にって得たコードワードである．そして $\mathcal{R}_i$ は i 番目のコードワードの領域である．$\boldsymbol{v}_i$ を連結して得る，$\boldsymbol{v} = [\boldsymbol{v}_1 \dots \boldsymbol{v}_M]$ は $D = M \times d$ 次元のベクトルになる．この動画特徴 $\boldsymbol{v}$ を Vector of Locally Aggregated Holistic Features (VLAH) と呼ぶことにする．

最後に，VLAH に線形識別器を以下のように用いる．

$$y = \boldsymbol{w}^\top \boldsymbol{v}, \tag{3}$$

ただし，$\boldsymbol{w}$ は線形識別器の重みベクトルである．

VLAD を大域特徴の集合をコーディングするのに用いた理由は，(1) VLAD は物体認識にて良い精度が得られることが知られており，(2) VLAD はコードワードベクトルと特徴量の差を取るだけなので，小規模の高次元特徴量に対して安定した性能を発揮できるからである．(3) また，DCNN などをの高次元ベクトルを用いた時，コードワードは代表的な物体の情報と考えられるので，提案する動画特徴は物体の情報を多く含んだものと考えられる．(4) そして，異なった順番で同じ物体を操作しても，同じ動画特徴が得られる．(5) 加えて，提案手法は動作の長さの違いに影響を受けにくい．

2 章で述べたように，THUMOS2014 [20] の優勝チームは，DCNN 特徴を平均化して動画特徴を得ていた．この特徴では，大域特徴の集合の一次の統計量しか得られない．一方で，本研究の提案手法はより物体情報を含んだ識別的な動画特徴を得ることができるので，より動作認識に対して高い性能を発揮することが期待される．

もちろん，VLAD だけでなく FV，Global Gaussian (GG) [22]，Generalized Local Correlation (GLC) [23] といったコーディング手法も用いる．また，時系列情報を捉える Vector Auto Regression (VAR) も用いた．しかし，これらの手法は小規模の高次元特徴量から分散行列や転移行列を計算するものである．そのため，これらの手法が物体認識に対し発揮するような性能を，日常生活動作認識にも期待することはできない．実験の章にてこれらの手法の比較を行う．

単一のコーディング手法で良い認識性能が得られなくとも，複数の互いに補い合うような手法であれば，組み合わせることでより良い性能を得られるということが知られている．本研究でも，様々なコーディング手法の組み合わせを提案する．m 番目のコーディング手法によって得られ

ⓒ 2015 Information Processing Society of Japan

た動画特徴を $\boldsymbol{v}_m$ とし，それから得た線形識別器の重みを $\boldsymbol{w}_m$ とすると，最終的な識別器の出力は以下のように得られる．

$$y = \sum_{m=1}^{M} \boldsymbol{w}_m^\top \boldsymbol{v}_m. \tag{4}$$

3.3 Discriminative Spatial Pyramid

カメラを右手首に搭載した時，画面の右側にはユーザーの右手が常に映る．また操作物体が，右手から遠い位置にある左下や右下に現れるといったことはほとんどない (図 5)．そのため，リストマウントカメラを用いて操作物体を認識するのであれば，おおまかにではあるものの識別的な領域が固定されている．そこで本研究では，位置情報を活用するために， 3.1 節で述べたように，Discriminative Spatial Pyramid Representation (D-SPR) [24] を動画特徴に組み込むことにする．

動画の各フレームを各ピラミッドレベルにおいて複数の矩形領域に分割する．この時動画特徴は，全ての領域から抽出された特徴量を連結することで得られる．$\boldsymbol{v}_k^l$ をレベル l の k 番目の矩形領域から得た特徴量だとしたとき，Spatial Pyramid Representation (SPR) [25] と呼ばれる次のベクトル $\boldsymbol{v}_s$ が得られる．

$$\boldsymbol{v}_s{}^\top = \left(\boldsymbol{v}_1^0{}^\top, \boldsymbol{v}_1^1{}^\top, \ldots, \boldsymbol{v}_{c(1)}^1{}^\top, \ldots, \boldsymbol{v}_1^{L-1}{}^\top, \ldots, \boldsymbol{v}_{c(L-1)}^{L-1}{}^\top\right),$$

ただし，$c(l)$ はピラミッドレベル l における矩形領域の数であり，L はピラミッドのレベルの数である．

SPR は動画特徴の空間的特徴を捉えるが，各領域の識別的な情報は含んでいない．そこで，レベル l の k 番目の領域が，その領域の識別的な重み α_k^l を持つと考える．重み付けされた SPR は各レベルにおける各領域から得た動画特徴の和で定義される．その式は以下のように書ける．

$$\boldsymbol{v}_w = \sum_{l=0}^{L-1} \sum_{k=1}^{c(l)} \alpha_k^l \boldsymbol{v}_k^l = \boldsymbol{V}\boldsymbol{\alpha}. \tag{5}$$

ただし $\boldsymbol{\alpha} \in \mathbb{R}^{d_w}$，$d_w = \sum_{l=0}^{L-1} c(l)$，$\boldsymbol{V} \in \mathbb{R}^{d \times d_w}$ は下式で定義される．

$$\boldsymbol{\alpha} = \left(\alpha_1^0, \alpha_1^1, \ldots, \alpha_{c(1)}^1, \ldots, \alpha_1^{L-1}, \ldots, \alpha_{c(L-1)}^{L-1}\right)^\top, \tag{6}$$

$$\boldsymbol{V} = \left(\boldsymbol{v}_1^0, \boldsymbol{v}_1^1, \ldots, \boldsymbol{v}_{c(1)}^1, \ldots, \boldsymbol{v}_1^{L-1}, \ldots, \boldsymbol{v}_{c(L-1)}^{L-1}\right). \tag{7}$$

重みベクトル $\boldsymbol{\alpha}$ は，PLS や Fisher SPR によって得られる固有値ベクトルである．α の詳しい導出方法は紙面の都合上省略する．本研究では，PLS を用いた．固有値ベクトルの集合 $\mathcal{A} = \{\boldsymbol{\alpha}_i \in \mathbb{R}^{d_w}\}_{i=1}^{N_w}$ を用いて，$\boldsymbol{v}_W^{(N_w)} \in \mathbb{R}^{N_w d}$ を以下のように定義し直す．

$$\boldsymbol{v}_W^{(N_w)} = \left((\boldsymbol{V}\boldsymbol{\alpha}_1)^\top, \ldots, (\boldsymbol{V}\boldsymbol{\alpha}_{N_w})^\top\right)^\top. \tag{8}$$

3.1 節で述べたように，本研究では最終的なカテゴリを得るために線形識別器を用いた．

$$y = \boldsymbol{w}^\top \boldsymbol{v}_W^{(N_w)}. \tag{9}$$

4. データセット

リストマウントカメラから撮影された日常生活動作認識の公開データセットが存在しないため，本研究では新たにリストマウントカメラ及びヘッドマウントカメラから撮影した日常生活動作データセットを作成した．本章では，作成したデータセットの詳細について説明する．

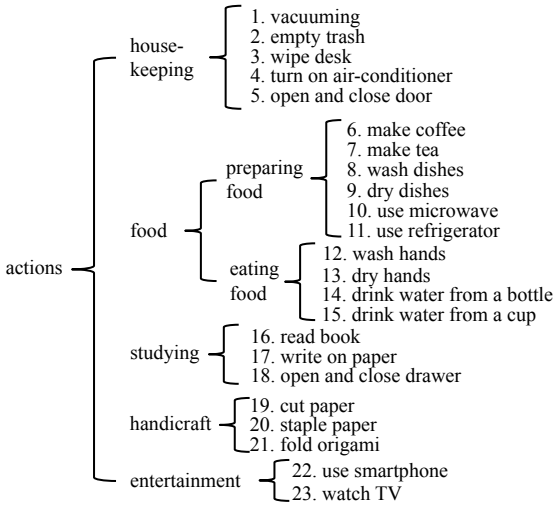

図 3 本研究で作成した日常生活動作動作カテゴリとクラス

4.1 動作クラス

動作リストは，日常生活動作認識に関する研究 [1], [26], [27] を参考に決定した．まず，一人称視点からの日常生活動作認識研究 [1] の 18 種類の動作のうち，歯磨きや洗濯といった多くの被験者が撮影に対し拒否感を示すようなクラスを除いた．そしてより多様なデータセットにするために，他の日常生活動作認識の研究 [26] やアルツハイマー病の治療の評価 [27] に使われた日常生活動作の動作種類を参考に，既存のカテゴリに新たな動作を追加し，また新たなカテゴリとそれに属する動作も増やした．図 3 に示すように，合計で 5 カテゴリ 23 種類の日常生活動作を本研究では認識対象とする．

4.2 データセット収集とアノテーション

データセット収集にあたり，目線カメラとして GoPro 社製の GoPro HERO3+[*1] を，リストマウントカメラとして Panasonic 社製の HX-A100[*2] を図 4 のように装着し，2 種類の動画を同時に撮影した．既存の一人称視点からの日常

[*1] http://jp.shop.gopro.com/cameras
[*2] http://panasonic.jp/wearable/a100

ⓒ 2015 Information Processing Society of Japan

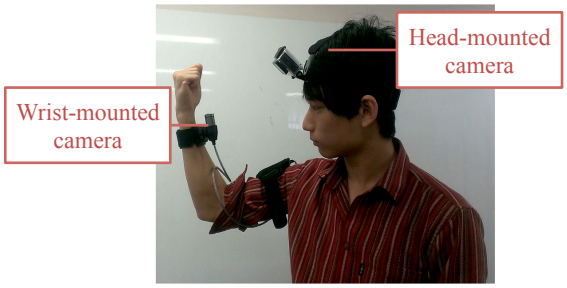

図 4 ヘッドマウントカメラ及びリストマウントカメラ装着図

生活動作データセット [1] 同様に，被験者には，より自然な動作が撮影できるよう細かい動作手順は指示せず，図 3 の動作リストの中から複数選択し，撮影を行ってもらった．動画内における動作のラベル付にあたって，動作はその動作に必要な前後の動作も含めて 1 つの動作と定義している．例えば”手を洗う”という動作であれば，手を洗うために蛇口を捻り手を水で洗って蛇口を閉めるところまでが”手を洗う”と定義される．被験者数は 20 で，撮影された映像は目線カメラとリストマウントカメラそれぞれ約 6.5 時間，約 69 万枚フレームずつとなった．

<environment A>

<environment B>

図 5 ”vacuuming”のデータセットの例．左がヘッドマウントカメラから，右がリストマウントカメラから撮影された映像で，上が環境 A で，下が環境 B で撮影されたもの．

4.3 データセットの特性

日常生活における操作物体は環境によって様々である．そのため，日常生活動作認識を行う上では，多様な環境に対応出来る必要がある．そこで本研究では，20 人の被験者にそれぞれの生活環境で動画を撮影してもらった．20 人の被験者が 23 種類の日常生活動作を撮影した映像は，ヘッドマウントカメラとリストマウントカメラそれぞれ約 6.5 時間，約 69 万枚フレームずつとなった．データセットの例である，図 5 に示すように，同じ動作であってもカメラに映っている環境は，被験者によって全く異なるものとなっている．

5. 実験結果

本章では，提案手法とベースライン，及びリストマウントカメラとヘッドマウントカメラの比較実験の結果について述べる．

実験設定

動画は 1fps に落として使用した．leave-one-out corss-validation 法を用いて評価を行う．フレームの各領域の特徴量を求める際には，一般物体認識において高い性能を示すことで知られている DCNN を利用する．フレームの各領域を ImageNet2012 データセット [28] で学習済みの 8 層 Convolutional Neural Network [29] に入力し，その中間出力である 7 層目の状態を得る．得られた 4096 次元の特徴量をフレームの各領域の特徴量 $\boldsymbol{f}$ とする．本研究ではベースラインとなる手法として，第 2 章で述べた，Jain らの手法のうち，見た目の特徴量を求める手法を参考に，各フレームの画像全体から得られた特徴量を Averaging にてコーディングする手法を採用する．次元圧縮には PCA を用い，識別器には SVM を用いる．

5.1 リストマウントカメラとヘッドマウントカメラの比較

VLAD，FV，GG，GCL 等のコーディング手法において，クラスタリングを行う際に，次元数が大きすぎるとクラスタリングが上手く行かず識別性能を大きく下げてしまう可能性があり，次元圧縮を必要とする場合がある．しかし，同時に次元数を圧縮しすぎることによって性能が下がってしまうことがある．また，VLAD は正規化の手法によって精度が変化する [30] ことが知られているため，本実験に最適な正規化手法を検討する必要がある．そして，VLAD や FV を使用する際に事前に決める必要があるコードブック数について検討する必要がある．そこで，各コーディング手法毎に様々なパラメータを試し，精度と計算量を考慮して決定した．表 2 は，各コーディング手法毎に最も良い性能を示した際の認識率を示したものである．まず VAR を除いた全ての手法においてリストマウントカメラがヘッドマウントカメラより 10%以上高い性能を示している．ここからリストマウントカメラのヘッドマウントカメラに対する優位性が見て取れる．

表 2 各コーディング手法による認識率

	Avg.	VLAD	FV	VAR	GG	GLC
WMC	44.7	48.1	42.3	8.4	41.7	39.0
HMC	34.1	36.3	30.5	8.3	32.9	32.2

表 3 は，各コーディング手法を組み合わせた場合の正答率のうち代表的なものを比較したものである．リストマウントカメラでは，全ての手法を組み合わせた時が 52.8%と最も高い性能を示し，Averaging のみの 44.7%と比べ 7.9%向上させた．本研究でコーディング手法は，全ての手法を組み合わせたものを使用することとする．

ⓒ 2015 Information Processing Society of Japan

表 1 Spatial Pyramid Representation (SPR) 及び Discriminative Spatial Pyramid Representation (D-SPR) の比較

	Level 0-2 1×1, 2×2, 4×4		Level 0-2 1×1, 3×1, 4×3		without SPR (baseline)	
Algorithms	WMC	HMC	WMC	HMC	WMC	HMC
Averaging + SPR	43.9	32.8	47.6	35.5	44.7	34.1
All codings + SPR	52.1	33.7	52.1	38.2	52.8	40.0
All codings + D-SPR(proposed)	54.4	37.3	**56.2**	43.6	-	-

5.2 位置情報

ここまでの実験設定を用いて，提案手法の性能を測り，コーディング手法に Averaging のみを用いた SPR 及び Averaging と VLAD と FV を用いた SPR との比較実験を行う．D-SPR においては，重みベクトルの計算方法は PLS Spatial Pyramid を使用し，パラメータ N_w は，もっともよい性能を示した $N_w = 3$ をヘッドマウントカメラに，$N_w = 7$ をリストマウントカメラに用いた．また，ピラミッドの切り方は，一般的に使用されている 1× 1, 2×2, 4×4 だけでなく，1×1+3×1+4×3 を用いた．これはリストマウントカメラにおいて，画面を縦に 3 分割した時右端に手，中央に操作物体，左端に背景が来る傾向があることに着目したためである．表 1 は，その実験結果である．表に示すように，ピラミッドの切り方が 1×1, 2×2, 4×4 の時に比べ，1×1+3×1+4×3 の時は，平均で 3.15%高い識別を示している．また，全てのコーディング手法を用いた時は，Averaging のみに比べ平均で 4.08%高い性能を示し，SPR ではなく D-SPR を用いることで平均で 3.85%高い性能を示している．

図 6 は，リストマウントカメラのデータセットに提案手法を用いた時の DSPR の各ピラミッドの重みを可視化したものである．Level 1，2 においていずれも右側があまり重要な領域でなく，中心に近い領域が重要な領域となっていることがわかる．これは，右側には撮影者の手だけが映っていることがほとんどであり，識別に重要な情報がほとんどないこと，そして操作物体の現れることの多い中心に近い領域が重要な情報が集まっていることがわかる．

5.3 動作検出

実用的な環境下においてはどの時間に動作を行ったのかといった情報は与えられない．そこで，本節では，時間方向にスライディングウィンドウを用いての提案データセットの認識を行う．表 4 は，その Average Precision を表したものである．この表に示す通り，テストデータに動作区間の情報を与えず動作を検出する場合でもリストマウントカメラに提案手法を用いるものが最もよく検出できている．

表 3 コーディング手法の組み合わせ

Avg.	VLAD	FV	VAR	GG	GLC	WMC	HMC
✓	✓	✓				50.7	37.6
✓	✓	✓		✓	✓	52.3	40.1
	✓	✓	✓	✓	✓	50.6	**40.7**
✓	✓	✓	✓	✓	✓	**52.8**	40.0

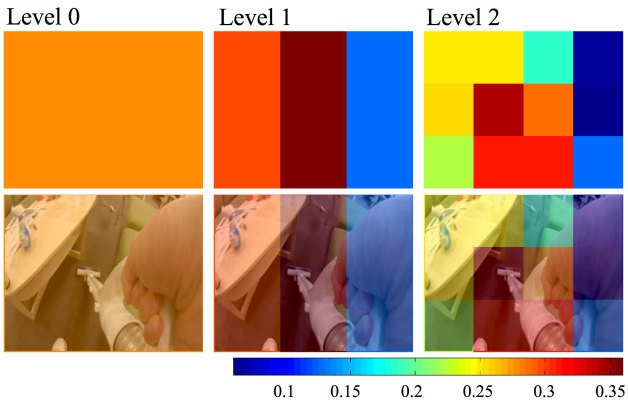

図 6 リストマウントカメラに提案手法を適用した時の D-SPR の重みを可視化した図

図 7 は，我々のデータセットで動作検出を行った時の例である．x 軸が時間軸で，各色がそれぞれの動作を表す．一番上がヘッドマウントカメラにベースラインの手法を用いたもの，中央がリストマウントカメラに提案手法を用いたもの，一番下が正解データである．予測された動作のスコアが閾値以下の区間では，なにも動作が行われなかった区間とした．この閾値は，損益分岐点の値を用いた．

図 7 に示す通り，いくつか検出できていないクラスや違ったクラスが検出されていることがあるものの，リストマウントカメラに提案手法を用いた時の方がヘッドマウントカメラにベースラインの手法を用いた時より良い性能を示している．例えば，Video A では，リストマウントカメラに提案手法を用いた時は”use refrigerator”(正解データの左から 11 番目の茶色のクラス) をほとんど完璧に検出できている一方で，ヘッドマウントカメラにベースラインの手法を用いた時は全く違ったクラスが検出されている．他

表 4 Average precision

	baseline		proposed	
	HMC	WMC	HMC	WMC
Average Precision	0.304	0.319	0.365	0.510

ⓒ 2015 Information Processing Society of Japan

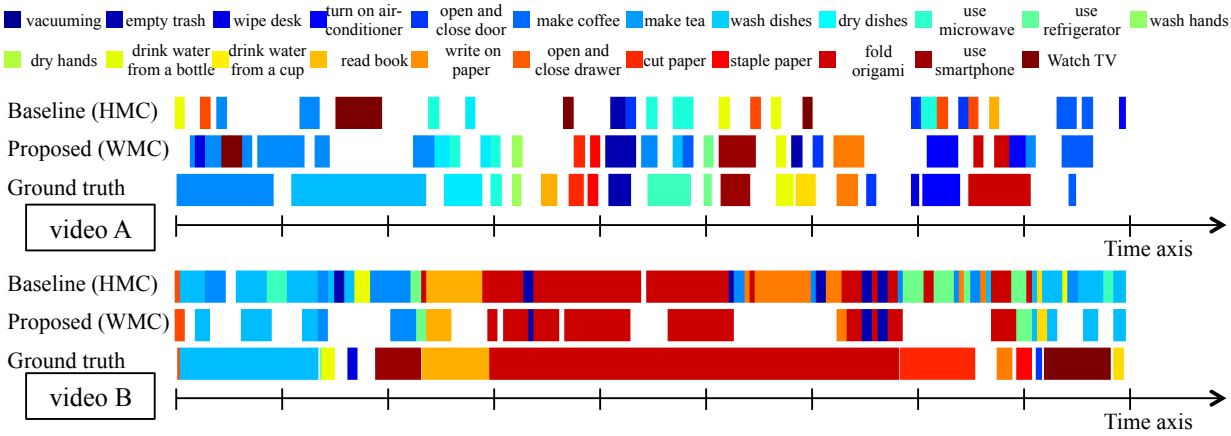

図 7 動作検出の例

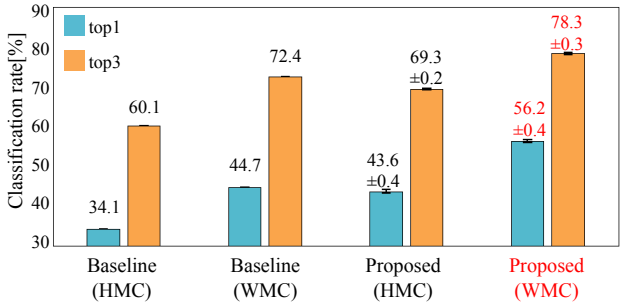

図 8 ヘッドマウントカメラ及びリストマウントカメラに対するベースラインと提案手法の性能

の例として，リストマウントカメラに提案手法を用いた時は，”drink water from a cup”(正解データの左から 14 番目のオレンジ色のクラス) をほとんど完璧に検出できている一方で，ヘッドマウントカメラにベースラインの手法を用いた時はそもそも動作が検出されていない．Video B では，動作時間の長いものが上手く検出できていないように見える．しかし，”make tea”(正解データの左から 2 番目の水色のクラス) は，いくつか”make tea”，ユーザーが紅茶の葉を浸して待っている時の動作等，関係のない動作を含んでいる．そのため，断続的に予測されたクラスというのは，そのクラスの重要な部分のみが予測されていると考えられる．

5.4 考察

図 8 は，提案手法とベースラインの手法から得られた結果を示している．Top1 のスコアでは，リストマウントカメラに提案手法を用いた時が 56.2%と，ヘッドマウントカメラにベースラインの手法を用いた時の 34.1%に比べ 22.1%高い性能を示した．Top3 のスコアでは，リストマウントカメラに提案手法を用いた時が 78.3%と，ヘッドマウントカメラにベースラインの手法を用いた時の 60.1%に比べ 18.2%高い性能を示した．

各手法におけるクラス別の識別率のグラフを図 9 に示す．図 9 から，”wash hand” と ”dry hand”は明らかに目線カメラ＋ベースラインの方がより正しく識別されていることが伺える．これは，リストマウントカメラにおいては手が常に映っているため，手に対する動作を識別するのは難しいためと考えられる．一方で，”wipe desk” や ”open and close drawer” といった動作（グループ 1），”drink water from a bottle” や ”drink water from a cup” といった動作（グループ 2），そして ”cut paper” や ”staple paper” や ”open and close door” といった動作（グループ 3）などにおいて，リストマウントカメラが大きく目線カメラを上回った．グループ 1 に関しては，目線カメラでは机の上の物の動作に関係のない物などが映ってしまい動作を識別しにくいが，リストマウントカメラであれば布巾や引き出しといった必要な操作物体のみ映るためにより正確に識別できたと考えられる．グループ 2 に関しては，目線カメラでは口元が映らず，必要な操作物体であるペットボトルやコップが視界から外れてしまう一方で，リストマウントカメラならそれらを捉え続けられるためにより正確に識別できたと考えられる．そしてグループ 3 に関しては，ハサミやホチキスやドアノブといった認識にとって重要な操作物体は目線カメラでも捉えられはするものの，小さく映ってしまうために認識しづらい．一方でリストマウントカメラであれば，それらの物体が大きく映るためにより正確に識別できたと考えられる．

6. 結論

本研究では，リストマウントデバイスとコンピュータービジョンを組み合わせることを提案し，ヘッドマウントカメラ及びリストマウントカメラによる日常生活動作データセットを作成した．そしてそれらを用いて，より簡単な情報だけで学習を行い，より多様な環境に対しロバストな操作物体認識が行うためには，ヘッドマウントカメラよりもリストマウントカメラが有効であることを示した．また，より精度の高い認識を行えるよう，新たな日常生活動作認識のための動画認識手法を提案し，それが有効であることを実験にて示した．

ⓒ 2015 Information Processing Society of Japan

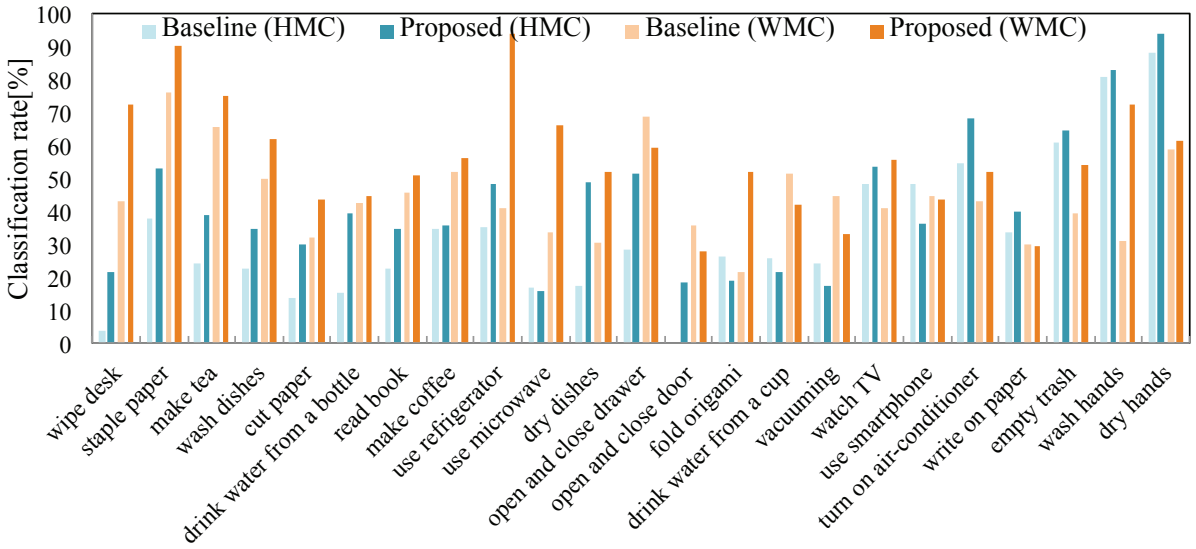

図 9 各クラスごとの認識率

参考文献

[1] Pirsiavash, H. and Ramanan, D.: Detecting activities of daily living in first-person camera views, *CVPR* (2012).
[2] Stikic, M., Huynh, T., Van Laerhoven, K. and Schiele, B.: ADL recognition based on the combination of RFID ,and accelerometer sensing, *Pervasive Computing Technologies for Healthcare* (2008).
[3] Wu, J., Osuntogun, A., Choudhury, T., Philipose, M. and Rehg, J. M.: A scalable approach to activity recognition based on object use, *ICCV* (2007).
[4] Ren, X. and Philipose, M.: Egocentric recognition of handled objects: Benchmark and analysis, *CVPR workshop on Egocentric Vision* (2009).
[5] Ren, X. and Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video, *CVPR* (2010).
[6] Fathi, A., Ren, X. and Rehg, J. M.: Learning to recognize objects in egocentric activities, *CVPR* (2011).
[7] Li, C. and Kitani, K. M.: Model recommendation with virtual probes for egocentric hand detection, *ICCV* (2013).
[8] Li, C. and Kitani, K. M.: Pixel-level hand detection in egocentric videos, *CVPR* (2013).
[9] Lucas, B. D. and Kanade, T.: An iterative image registration technique with an application to stereo vision., *IJCAI*, Vol. 81, pp. 674–679 (1981).
[10] Dalal, N., Triggs, B. and Schmid, C.: Human detection using oriented histograms of flow and appearance, *ECCV* (2006).
[11] Sivic, J. and Zisserman, A.: Video Google: A text retrieval approach to object matching in videos, *ICCV* (2003).
[12] Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C.: Visual categorization with bags of keypoints, *ECCV workshop on Statistical Learning in Computer Vision* (2004).
[13] Jégou, H., Douze, M., Schmid, C. and Pérez, P.: Aggregating local descriptors into a compact image representation, *CVPR* (2010).
[14] Perronnin, F. and Dance, C.: Fisher kernels on visual vocabularies for image categorization, *CVPR* (2007).
[15] Wu, J., Zhang, Y. and Lin, W.: Towards good practices for action video encoding, *CVPR* (2014).
[16] Wang, H. and Schmid, C.: Action recognition with improved trajectories, *ICCV* (2013).
[17] Wang, H., Kläser, A., Schmid, C. and Liu, C.-L.: Dense trajectories and motion boundary descriptors for action recognition, *IJCV*, Vol. 103, No. 1, pp. 60–79 (2013).
[18] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, *NIPS* (2012).
[19] Jiang, Y.-G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M. and Sukthankar, R.: THUMOS Challenge: Action Recognition with a Large Number of Classes (2014).
[20] Jain, M., van Gemert, J. C. and Snoek, C. G. M.: University of Amsterdam at THUMOS Challenge 2014, *ECCV workshop on THUMOS Challenge*, (online), available from ⟨https://ivi.fnwi.uva.nl/isis/publications/2014/JainTHUMOS2014⟩ (2014).
[21] Maekawa, T., Yanagisawa, Y., Kishino, Y., Ishiguro, K., Kamei, K., Sakurai, Y. and Okadome, T.: Object-based activity recognition with heterogeneous sensors on wrist, *ICPC* (2010).
[22] Nakayama, H., Harada, T. and Kuniyoshi, Y.: Global gaussian approach for scene categorization using information geometry, *CVPR* (2010).
[23] Harada, T., Nakayama, H. and Kuniyoshi, Y.: Improving local descriptors by embedding global and local spatial information, *ECCV* (2010).
[24] Harada, T., Ushiku, Y., Yamashita, Y. and Kuniyoshi, Y.: Discriminative spatial pyramid, *CVPR* (2011).
[25] Lazebnik, S., Schmid, C. and Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *CVPR* (2006).
[26] Tapia, E. M., Intille, S. S. and Larson, K.: Portable Wireless Sensors for Object Usage Sensing in the Home: Challenges and Practicalities., *AmI*, Vol. 4794, pp. 19–37 (2007).
[27] Galasko, D., Bennett, D., Sano, M., Ernesto, C., Thomas, R., Grundman, M. and Ferris, S.: An inventory to assess activities of daily living for clinical trials in Alzheimer's disease., *Alzheimer Disease & Associated Disorders*, Vol. 11, pp. 33–39 (1997).
[28] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database, *CVPR* (2009).
[29] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition, *arXiv preprint arXiv:1310.1531* (2013).
[30] Arandjelovic, R. and Zisserman, A.: All about VLAD, *CVPR* (2013).

ⓒ 2015 Information Processing Society of Japan