CH-001

行動認識のための人物存在確率を考慮した深層学習
Action Recognition Utilizing Deep Learning based on a Probability Map of an Image for a Person Class

本田 光† 榎田 修一†
Hikaru Honda Shuichi Enokida

1. はじめに

1.1 研究背景

近年，人と共生するロボットや，無人店舗が普及しており，その数は年々増加している．これらの自律システムのセンサとしては，広範囲を詳細に知覚可能でかつ，安価なカメラが利用されることが一般的である．ここで収集される動画像は，システム内で自動認識され適切なサービスを推定するために利用される．加えて，蓄積された動画像は，次世代の自律システムの性能改善に向け，自動解析，整理されるべきである．そこで深層学習を用いて，人の行動を自動的に判別する行動認識が注目されている．行動認識は，カメラやセンサなどが取得した情報に基づき，対象の行動を推定するタスクで，共生ロボットや防犯システム，映像識別など様々な分野での応用が可能である．行動認識では，テクスチャ情報や深度情報，三次元点群情報を使うモデルなど，様々な入力特徴量を用いた研究が行われているが，広く普及している単眼カメラから得られるテクスチャ情報を用いることにより，より幅広い分野での応用が可能になると考える．

1.2 研究目的

行動認識における入力として，カメラから得られた テクスチャ情報 や，隣接フレーム間の動きを計算したオプティカルフローがよく用いられてきた．オプティカルフロー は 3D CNN との組み合わせにより，高い精度が得られることが知られている[6]．これは，生データである RGB の動画像をそのまま入力することに比べ，オプティカルフローを用いて動きに着目すると，精度が上昇することを示している．しかし，例えばカメラ位置が変化するロボット搭載型のシステムでは，カメラの動き情報がオプティカルフローに大きく影響し，行動認識の精度が低下する問題がある．一方，人の位置に着目した入力特徴量は，カメラ運動に頑健な行動認識の実現が期待できる．そこで本研究では，人の位置に着目した人物存在確率マップを入力特徴量として用いる行動認識手法を提案する．

2. 深層学習による行動認識

本章では代表的な行動認識手法である，3D Convolutional Neural Network と，Two-Stream Convolutional Neural Network，それら 2 つの組み合わせである Two-Stream Inflated 3D Convolutional Neural Network について説明していく．

2.1 3D Convolutional Neural Network [1]

3D Convolutional Neural Network [1] （以後 3D CNN ）は代表的な行動認識手法の 1 つである．行動認識は，物体検出などの他画像処理タスクと異なり，一連の動作の時間情報を考慮した処理が必要である．しかし，従来の畳み込み層は，縦横の二次元であり，畳み込み層では時間情報を得ることができない．そこで，3D CNN では畳み込み層を図 2.1 の様に，縦横の二次元から，縦横+奥行きの三次元に拡張し，時間情報も含めて畳み込む．しかし，3D CNN は二次元の CNN と比較して，パラメータ数が増加してしまう問題があり，ネットワークの構造を深くすることが困難である．そのため，使用するデータセットの規模を拡大する，もしくは，ResNet [2] や Inception-V1 [3]などのアーキテクチャを適応させることにより精度向上が図られている．

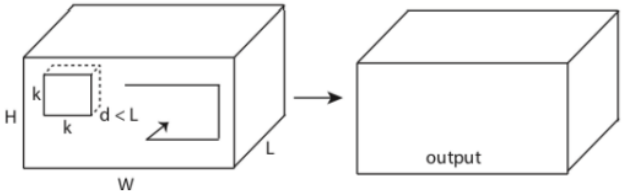


図 2.1 三次元の畳み込み層 ([4]より抜粋)

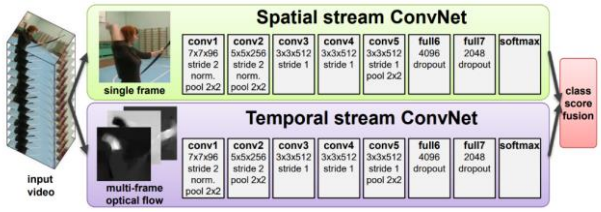


図 2.2 Two-Stream の構造 ([5]より抜粋)

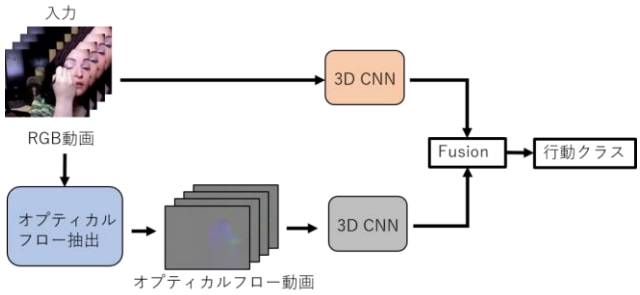


図 2.3 I3D の構造

2.2 Two-Stream Convolutional Neural Network [5]

Two-Stream Convolutional Neural Network [5]（以後 Two-Stream）は，二次元の畳み込み層に RGB 画像と，オプティカルフローの 2 種類の特徴量を入力する（図 2.2 参照）．RGB 画像では空間情報を，オプティカルフローでは時間情報をそれぞれ独立して処理し，最後にそれぞれの推論結果

† 九州工業大学 Kyusyu Institute of Technology

図 2.4 Biking (RGB) ([9]データセットより抜粋)

図 2.5 Biking （オプティカルフロー）

を統合する．オプティカルフローは，画像内に写る隣接フレーム間の物体の動きを計算したベクトルで，Two-Stream ではオプティカルフローによって時間情報を得ることにより精度が向上する．

2.3 Two-Stream Inflated 3D Convolutional Neural Network [6]

Two-Stream Inflated 3D Convolutional Neural Network [6]（以後 I3D）は 2.1 節，2.2 節で紹介した，3D CNN と Two-Stream を組み合わせた手法である．Two-Stream では，二次元の畳み込み層に対して，RGB フレームとオプティカルフローフレームの 2 種類を入力としていたが，I3D では図 2.3 のように，Two-Stream の畳み込み層を三次元に拡張し，RGB 動画と，そこから計算されたオプティカルフロー動画を入力とする．その後 Two-Stream と同様に，それぞれ単独で推論し，結果を統合する．また，バックボーンとして Inception-V1 [3]を使用し，3D CNN によるパラメータ数の増加を抑え，ネットワークを深くしている．

I3D では，Two-Stream の畳み込み層を三次元に拡張することで精度が上昇した．しかし，カメラ位置が大きく変化する撮影状況では，カメラの動きがオプティカルフローに大きく影響し，行動認識の精度が低下する問題がある．図 2.4 と図 2.5 はカメラ位置が変化する動画の RGB フレームとそこから求められた，同時刻のオプティカルフローフレームを抜き出した物である．カメラ位置が変化する際に，画面全体にオプティカルフローが検出されていることが確認できる．これは，オプティカルフローが画面の「動き」に着目した特徴量であるため，カメラ位置の変化により画面全体に，動作と関係の無いオプティカルフローベクトルが発生するためである．一方、人の位置に着目した入力特徴量は，カメラ運動に頑健な行動認識の実現が期待できる．そこで本研究では，オプティカルフローにより推定された「人の動き」ではなく，Semantic Segmentation によって生成した，人物存在確率マップによる「人の位置や概形」に着目した新しい入力特徴量を提案する．

3. 人物存在確率を考慮した行動認識

本章では，提案手法である人物存在確率を考慮した行動認識について説明する．3.1 節で全体の概要について説明し，3.2 節で具体的な生成方法について詳述する．

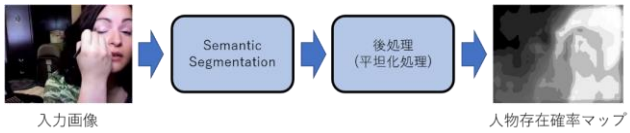
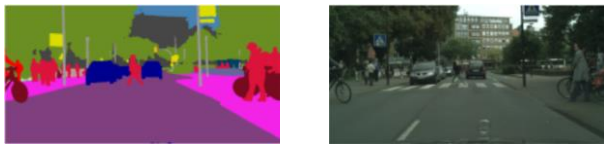


図 3.1 人物存在確率マップ生成



(a) 入力画像　　(b)出力画像

図 3.2 Semantic Segmentation（[7] から抜粋）

3.1 人物存在確率を考慮した行動認識アーキテクチャ

本研究では，Semantic Segmentation を用いて，RGB 画像からクラスごとの存在確率マップを生成する．その後，生成したクラスごとの存在確率マップから，行動認識の対象である人の存在確率マップのみを抽出し，ヒストグラムを平坦化処理する（図 3.1 参照）．I3D への入力は，平坦化後の存在確率マップを用いる

3.2 人物存在確率マップ生成

3.2.1 Semantic Segmentation

Semantic Segmentation は，画像のピクセルごとにクラスを付与するタスクである（図 3.2 参照）．本研究では，Semantic Segmentation に，ResNet-101[2]をバックボーンとして持つ DeepLab V3 [7] を使用する．DeepLab V3 は入力された RGB 画像に対して，21 種類のクラスごとに存在確率マップを生成し，各ピクセルで最も高い確率を持つクラスを最終的なクラスとして決定する．

3.2.2 人物存在確率マップ生成

DeepLab V3 によって生成されたクラスごとの存在確率マップは，入力画像と同じサイズで，各画素に存在確率を輝度として持つ 1 チャネルの画像である．本研究では「人の位置や概形」に着目するため，生成されたクラスごとの確率マップから，人の存在確率マップのみを抽出する．しかし，DeepLab V3 によって生成された確率マップは，輝度が低く，輝度の分布も狭いため，全体が黒く塗りつぶされたような画像となってしまう．そのため，一般的な画像認識に用いられる Deep Neural Network で学習すると，精度が大幅に下がる問題がある．そこで，存在確率マップの輝度が 0~255 の間で均等に分布するよう，式（3.1）および（3.2）[8] を用いて輝度の平坦化処理をする．ただし，cdf は累積分布関数で，確率変数 X と確率 P を用いて式（3.2）によって求められる．また，$\mathrm{cdf}_{\mathrm{min}}$は累積分布関数の 0 以外の最小値を表す．$(M \times N)$ は入力画像のサイズを表し，式（3.1）で求められた輝度は round 関数により四捨五入され，整数値で出力される．以後，存在確率マップは平坦化処理した存在確率マップを指す．元の RGB 画像と，平坦化後の存在確率マップの例を図 3.3，3.4，3.5 に示す


(a) RGB([9]データセットから抜粋)

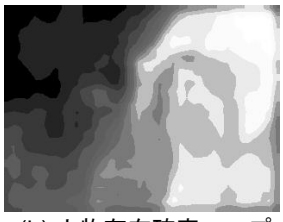

(b) 人物存在確率マップ

図 3.3 Apply Eye Makeup

(a) RGB([9]データセットから抜粋)

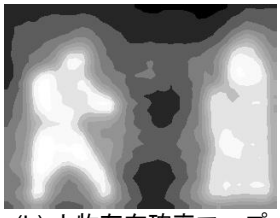

(b) 人物存在確率マップ

図 3.4 Punch


(a) RGB([9]データセットから抜粋)

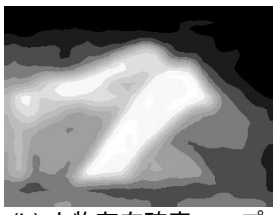

(b) 人物存在確率マップ

図 3.5 例 3 Typing

$$f(x) = \mathrm{round}\left(\frac{\mathrm{cdf}(x) - \mathrm{cdf}_{\min}}{(M \times N) - \mathrm{cdf}_{\min}}\right) \quad (3.1)$$

$$\mathrm{cdf}(x) = P(X \leq x) = \sum_{X \leq x} P(X) \quad (3.2)$$

4. 人物存在確率マップの有効性に関する評価実験

4.1 行動認識モデルの学習詳細

本研究での評価実験には，UCF-101 データセット（split1）[9]を使用する．UCF-101 データセットには 101 種類の行動クラスが含まれており，訓練データとして 9,537 個，評価データとして 3,783 個の動画を使用する．学習時の設定として，入力動画のフレームレートは 25[FPS] とし，入力画像のサイズは224×224 にリサイズする．学習時は 64 フレーム長の RGB 動画に基づき学習する．評価時は，まず図 4.1 のように，入力された動画を 64 フレームごとに分割して，それぞれ I3D Network へ入力し，行動クラスの確率分布を得る．最後に，ビデオ全体における行動クラスの確率分布の平均を求め，確率が最大となる行動クラスを出力する．ここで，本研究では I3D のバックボーンとして，Inception-v1 を用い，オプティカルフローの計算には TV-L1 アルゴリズム [10] を用いる．

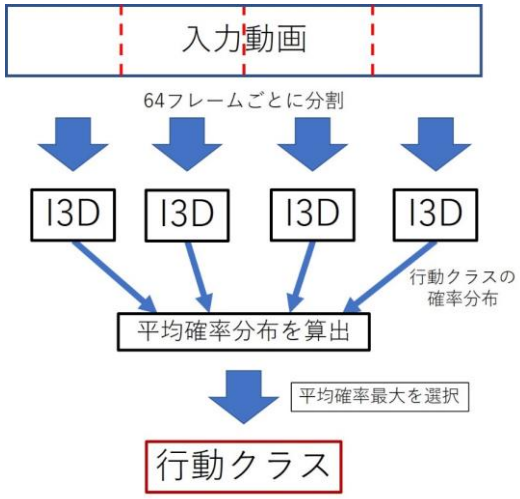


図 4.1 ビデオクラスの推論

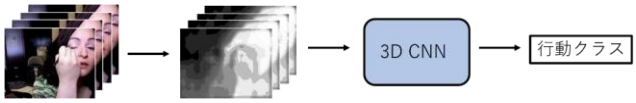


図 4.2 I3D（単独入力特徴量）

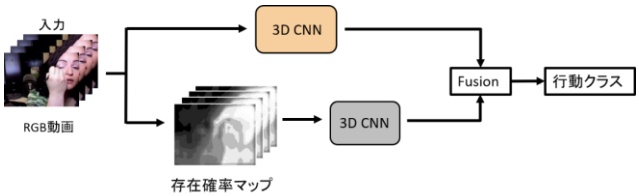


図 4.3 I3D（Two-Stream 形式）

4.2 行動認識精度に基づく評価指標

本研究での評価指標には，Top1 Accuracy と Top5 Accuracy を用いる．Top1 Accuracy は，モデルが推論した最上位クラスが正解である割合 (式 4.1)で，Top5 Accuracy は，モデルが推論した上位 5 クラスの中に正解が存在する割合 (式 4.2)である．

$$\text{Top1 Acc} = \frac{(\text{推論結果の最上位クラス=正解ラベル})\text{の数}}{\text{評価データ数}} \quad (4.1)$$

$$\text{Top5 Acc} = \frac{(\text{推論結果の上位 5 クラス}\supset\text{正解ラベル})\text{の数}}{\text{評価データ数}} \quad (4.2)$$

4.3 行動認識モデルの定量評価に関する基礎実験

本節では，提案手法である人物存在確率マップと，従来手法の RGB，オプティカルフローの 3 種類を入力特徴量として用い，精度を比較する．精度比較の際に，単独入力特徴量と Two-stream の二つの形式に分けて計測する．単独入力特徴量での精度比較では，図 4.2 のように，比較する入力特徴量のうち，いずれか 1 種類のみを用いて推論する．Two-Stream 形式での精度比較では，図 4.3 のように，比較する 3 種類の入力特徴量から，2 種類を組み合わせて用いる．本実験における Two-Stream 形式の結合部は，二つの確率分布の平均値を算出することとした．

表 4.1 単独入力特徴量での精度比較

入力	Top1 Acc	Top5 Acc
人物存在確率マップ	62.8	85.9
RGB	60.6	83.9
オプティカルフロー	77.8	94.0

表 4.2 Two-Stream 形式での精度比較

入力	Top1 Acc	Top5 Acc
人物存在確率マップ ＋オプティカルフロー	79.9	93.6
人物存在確率マップ ＋RGB	72.6	91.1
RGB ＋オプティカルフロー	81.5	94.7

表 4.3 体の動きのみのクラス（Top1 Acc）

入力	スクワット	腕立て伏せ
人物存在確率マップ	83.3	83.3
RGB	40.0	40.0
オプティカルフロー	73.3	80.0

4.3.1 単独入力特徴量における精度比較

単独入力特徴量での精度比較結果を表 4.1 に示す．単独入力特徴量での精度は，Top1 Accuracy，Top5 Accuracy の両方でオプティカルフローが最も精度が高いことを確認した．次いで提案手法である人物存在確率マップの精度が高いことが確認された．このことから，行動認識において，人物存在確率マップが元データである RGB よりも有効であることが確認された．

4.3.2 Two-stream 形式における精度比較

Two-stream 形式での精度比較結果を表 4.2 に示す．単独特徴量での精度比較では，RGB よりも，提案手法である人物存在確率マップで高い精度が得られたが，Two-stream 形式での精度比較では，人物存在確率マップ+オプティカルフローに比べて，RGB+オプティカルフローの組み合わせで，より高い精度が得られた．このことから Two-Stream 形式では，単独特徴量の際の精度が高い二つの入力特徴量を単純に選択するよりも，適切な組み合わせを選択することにより精度が上昇することが確認された．

4.4 体の動きのみの行動に関する実験

「スクワット」や，「腕立て伏せ」などの体の動きのみの行動クラスに対して，同様に実験した．提案した人物存在確率マップを入力特徴量とすると，表 4.3 のように RGB やオプティカルフローと比べて精度が高い傾向にある．精度が上昇した体の動きのみのクラスの例を図 4.4~図 4.8 に示す．

表 4.4 Body-Motion Only クラスでの精度比較

入力	Top1 Acc
人物存在確率マップ	61.9
RGB	51.7
オプティカルフロー	88.5

表 4.5 新たな Body-Motion Only クラスでの精度比較

入力	Top1 Acc
人物存在確率マップ	68.6
RGB	57.8
オプティカルフロー	82.9

表 4.6 道具を使用するクラス（Top1 Acc）

入力	やり投げ	ラケットを振る
人物存在確率マップ	35.4	34.9
RGB	54.8	44.8
オプティカルフロー	62.3	71.4

4.5 UCF-101 Body Motion Only Class について

4.5.1 UCF-101 データセットのクラス分類

UCF-101 データセットは，

1. Human-Object Interaction
2. Body-Motion Only
3. Human-Human Interaction
4. Playing Musical Instruments
5. Sports

の 5 つの行動タイプに分類されている[11]．Body-Motion Only に属する行動クラスは

- Baby Crawling
- Blowing Candles
- Body Weight Squats
- Handstand Pushups
- Handstand Walking
- Jumping Jack
- Lunges
- Pull Ups
- Push Ups
- Rock Climbing Indoor
- Rope Climbing
- Swing
- Tai Chi
- Trampoline Jumping
- Walking with a dog
- Wall Pushups

の 16 クラスである．

4.5.2 Body-Motion Only に属する行動クラスでの精度比較

4.5.1 章で示した Body-Motion Only に属する行動クラスでの精度比較結果を表 4.4 に示す．結果，予想に反し，提案手法である人物存在確率マップにおいて認識精度の低下が確認された．これは Body-Motion Only の行動認識クラスには「Rope Climbing」や，「Walking With Dog」のように体の動きのみではなく，道具や動物などを使用するクラスが含まれていることが原因と考察した．よって本研究では，人の体の動きのみに限定した新たな Body-Motion Only タイプを設定する．

4.5.3 新たな Body-Motion Only クラス分類

4.5.1 章で示した Body-Motion Only に含まれる 16 クラスの行動以外に Human-Human Interaction タイプや，Sports タイプには，「Salsa Spin」や，「Punch」などの体の動きのみのクラスが含まれている．以上を考慮し，新たに作成した Body-Motion Only に属する行動クラスは以下に示す 11 クラスである．以降本論文ではこれら 11 クラスの行動からなる行動タイプを Body-Motion Only V2（BMO V2）と呼ぶ．

- Baby Crawling
- Body Weight Squats
- Handstand Pushups
- Handstand Walking
- Jumping Jack
- Lunges
- Punch
- Push Ups
- Salsa Spin
- Sumo Wrestling
- Tai Chi

4.5.4 BMO V2 に属する行動クラスでの精度比較

BMO V2 タイプでの精度比較結果を表 4.5 に示す．BMO V2 タイプでの精度比較では，従来の Body-Motion Only クラスでの精度比較結果と比べ，提案手法である人物存在確率マップの精度が上昇した．結果より，提案手法である人物存在確率マップは，体の動きのみのクラスで精度が上昇する傾向が表れていることが確認された．

4.6 道具を用いる行動に関する実験

表 4.6 の精度比較結果より，「やり投げ」や「テニスラケットを振る」といった，道具を使用するクラスでは精度が低下する傾向にあることが確認された．精度が低下した道具を使用するクラスの例を図 4.9~図 4.13 に示す．道具を使用するクラスでは，RGB と人物存在確率マップで図の赤く囲まれた道具が消えていることがわかる．これは人物存在確率マップの生成過程で，人以外の背景や，動作に関係する道具の情報が失われたことが原因だと考えられる．

5. まとめ

近年，人と共生するロボットなどの普及により，人の行動を自動的に判別する行動認識が注目されている．行動認識における入力として「テクスチャ情報」に着目した RGB 画像情報や，「人の動き」に着目したオプティカルフローが用いられるが，本研究では，「人の位置や概形」に着目した人物存在確率マップを入力として用いる行動認識手法を提案した．実験結果より，人の位置や概形に着目することで，他の入力特徴量と比べ，体の動きのみのクラスでの精度の上昇率の高さを確認した．一方で，道具を使用した行動クラスにおいては精度が低下した．今後の展望として，動作に関係する道具の情報を加えることにより，さらなる精度向上を目指す．また，Two-Stream における，適切な入力特徴量の組み合わせについても検討する．

参考文献

[1] Shuiwang Ji，*et al.*, “3D Convolutional Neural Networks for Human Action Recognition”, Pattern Analysis and Machine Intelligence, pp. 221-231, (2013).
[2] Kaiming He, *et al.*, ”Deep Residual Learning for Image Recognition”, Computer Vision and Pattern Recognition, (2016).
[3] Christian Szegedy, *et al.*, “Going deeper with convolutions”, Computer Vision and Pattern Recognition, (2015).
[4] Du Tran, *et al.*, “Learning Spatiotemporal Features With 3D Convolutional Networks”, IEEE International Conference on Computer Vision, pp. 4489-4497, (2015).
[5] Karen Simonyan, *et al.*, “Two-Stream Convolutional Networks for Action Recognition in Videos”, Neural Information Processing Systems, (2014).
[6] João Carreira, *et al*,. “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”, Computer Vision and Pattern Recognition, pp. 6299-6308, (2017).
[7] Liang-Chieh Chen, *et al.*, “Rethinking Atrous Convolution for Semantic Image Segmentation”, Computer Vision and Pattern Recognition, (2017).
[8] opencv, “ヒストグラムその 2:ヒストグラム平坦化”, http://whitewell.sakura.ne.jp/OpenCV/py_tutorials/py_imgproc/py_histograms/py_histogram_equalization/py_histogram_equalization.html（閲覧日：2023 年 6 月 5 日）
[9] Khurram Soomro, *et al.*, “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild”, CRCV-TR-12-01, (2012).
[10] C.Zach, *et al.*, “A duality based approach for realtime TV-L1 optical flow”, Pattern Recognition, pp. 214-223, (2007).
[11] UNIVERSITY OF CENTRAL FLORIDA, “UCF101 - Action Recognition Data Set”, https://www.crcv.ucf.edu/data/UCF101.php（閲覧日：2023 年 6 月 5 日）

(a) RGB([9]データセットから抜粋　(b) 人物存在確率マップ

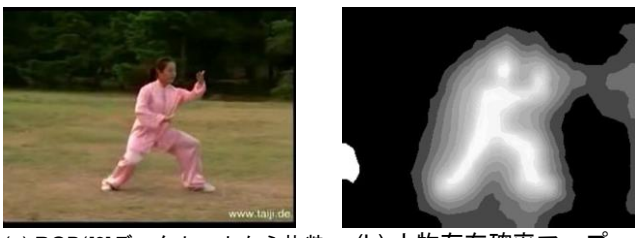
図 4.4 体の動きのみのクラス (Body Weight Squats)

(a) RGB([9]データセットから抜粋　(b) 人物存在確率マップ

図 4.5 体の動きのみのクラス (Push Ups)

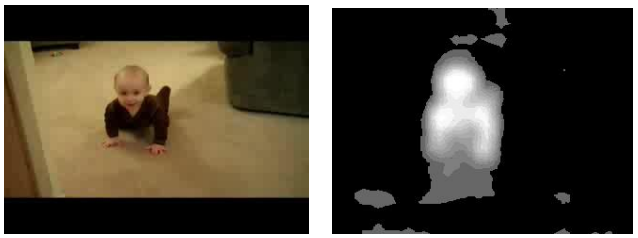
(a) RGB([9]データセットから抜粋　(b) 人物存在確率マップ

図 4.6 体の動きのみのクラス (Salsa Spin)



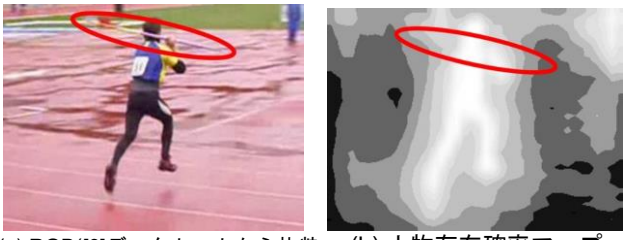
(a) RGB([9]データセットから抜粋　(b) 人物存在確率マップ

図 4.7 体の動きのみのクラス (Tai Chi)



(a) RGB([9]データセットから抜粋　(b) 人物存在確率マップ

図 4.8 体の動きのみのクラス (Baby Crawling)

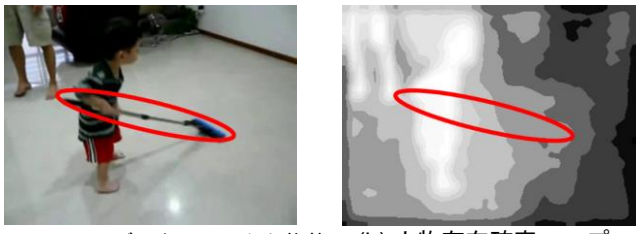


(a) RGB([9]データセットから抜粋　(b) 人物存在確率マップ

図 4.9 道具を使うクラス (Javelin Throw)

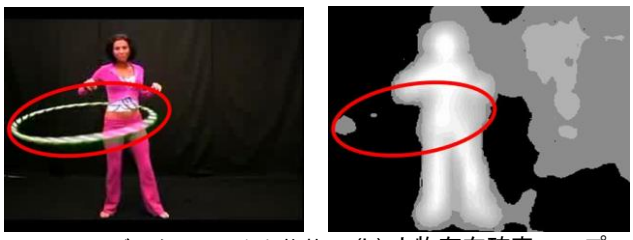
(a) RGB([9]データセットから抜粋　(b) 人物存在確率マップ

図 4.10 道具を使うクラス (Tennis Swing)



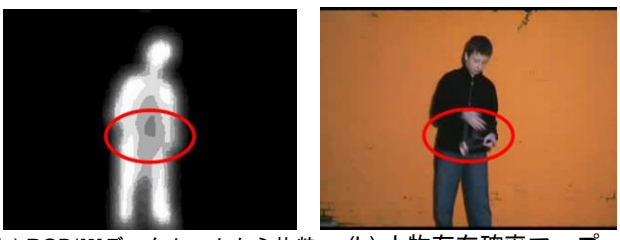
(a) RGB([9]データセットから抜粋　(b) 人物存在確率マップ

図 4.11 道具を使うクラス (Mopping Floor)



(a) RGB([9]データセットから抜粋　(b) 人物存在確率マップ

図 4.12 道具を使うクラス (Hula Hoop)



(a) RGB([9]データセットから抜粋　(b) 人物存在確率マップ

図 4.13 道具を使うクラス (Yoyo)