社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

[招待講演] 機械学習を用いた距離画像からの物体認識技術

藤吉　弘亘†

† 中部大学 〒 487-8501 愛知県春日井市松本町 1200
E-mail: †hf@cs.chubu.ac.jp

あらまし　本稿では，機械学習を用いた距離画像からの物体認識技術として，ポイントクラウドを用いた物体認識と距離画像を用いた物体認識手法について解説する．距離画像からの人検出，人流計測，動作認識等の講演者の研究グループで取り組んでいるアプローチから，ゲーム機の入力デバイスに採用された人体姿勢推定のアルゴリズム等の最新動向についても紹介する．事例を紹介しながら，距離画像と機械学習の組み合わせによる認識手法のしくみやメリットについて示す．

キーワード　距離画像，ポイントクラウド，機械学習，物体検出，動作認識，人体姿勢推定

[Invited Talk]Object Recognition with Depth Image by Machine Learning

Hironobu FUJIYOSHI†

† Chubu University，1200 Matsumoto-cho，Kasugai Aichi 487-8501 Japan
E-mail: †hf@cs.chubu.ac.jp

Abstract　This paper presents object recognition with point cloud and depth images, and its trends. There have been many various researches with depth information for detecting human, analyzing human flow, and recognizing action. In this paper, we report that machine learning with depth images obtains better performance than that with gray-scale images. We also present a human pose estimation method used for practical application such as gesture interaction, and discuss trends in depth image processing.

Key words　depth image, point cloud, machine learning, object detection, action recognition, human pose estimation

1. ま え が き

汎用コンピュータの進化に伴い，大量のデータを高速に処理できるようになったことから，画像から高次元の特徴量ベクトルを抽出し，機械学習を用いて識別する手法が実用化された．機械学習では，クラスラベルが付与された大量の学習サンプルを必要とするが，ルールベースの手法のように研究者がいくつかのルールを設計する必要がないため，汎用性の高い識別器を学習できる．画像から対象物の位置と大きさを求める物体検出問題において，顔検出では Haar-like 特徴量 [1] と AdaBoost [2]，人検出では HOG 特徴量 [3] と SVM [4] の組み合わせが広く使用されている．物体検出は対象クラスと非対象クラスに識別する 2 クラスの問題設定であったが，画像分類やセマンティックセグメンテーションのような応用では，多クラスの問題設定が扱われるようになった．マルチクラス識別器の 1 つである Random Forest [5] は，ランダム性を取り入れたアンサンブル学習手法であり，2006 年以降に画像認識の分野で利用され始めている．このような画像局所特徴量と機械学習の組み合わせは，可視光カメラで撮影された通常の画像だけでなく，距離画像へも適用され始めている．

2010 年以降では，リアルタイムに距離情報を出力する TOF 方式の距離画像センサが市販化された．中でも，ゲーム機の入力デバイスである Kinect では，距離画像と Random Forest によりリアルタイムの三次元人体姿勢推定を実現した．我々の研究グループでは，これまでに距離画像からの物体認識として，人検出，人流計測，動作認識の研究に取り組んできた．距離情報を用いることで，従来の画像認識フレームワークを利用しつつ，通常の可視光カメラと比べて認識性能を大幅に向上させることが可能となる．本稿では，機械学習を用いた距離画像からの物体認識技術とその動向について述べる．

2. 距離情報の表現

距離情報の表現方法は，図 1 に示すように，ポイントクラウドと距離画像の 2 種類ある．各表現方法に合わせて，その認識処理過程は異なる．以下では，ポイントクラウドと距離画像の定義と性質について述べる．

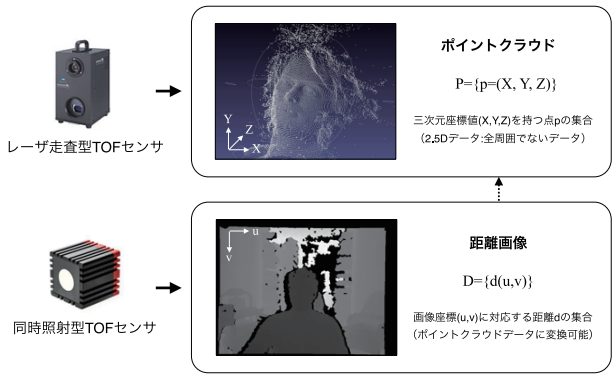


図 1　距離画像の表現方法

2.1　ポイントクラウド

ポイントクラウドは，三次元座標値 (X, Y, Z) を持つ点 p の集合として定義され，レーザー走査型の TOF センサ等から取得された 2.5 次元の点群データを指す．2.5 次元とは，カメラの視点方向から物体や空間の三次元形状を表したものである．また，ポイントクラウドを対象としたライブラリとして，Point Cloud Library(PCL) が幅広く利用されている．PCL は，フィルタリング，物体検出，サーフェス再構成，レジストレーション，モデルマッチング，セグメンテーション等のポイントクラウド処理に必要な関数が用意されている．

2.2　距 離 画 像

距離画像は，画像座標 (u, v) に対応する距離 d の集合として定義され，同時照射型 TOF センサから取得できる画像を指す．距離画像は，テクスチャに影響されることなく，物体の形状や前後関係を知ることができる．一方，データ形式は，RGB 画像と変わらないため従来の画像処理を利用することも可能である．また，距離画像は，カメラパラメータを用いることにより，ポイントクラウドデータに変換することができる．

3.　ポイントクラウドを用いた物体認識

本章では，ポイントクラウドデータを用いた物体認識の手法として，PCL に採用された人検出法とレイヤー毎にモデルを学習する人検出法について述べる．

3.1　PCL の人検出

PCL に採用された人検出法 [6] は，ポイントクラウドデータから床面を推定し，クラスタリングにより人の候補位置を求めるアプローチである．本手法は Point Cloud Library(PCL) のバージョン 1.7 から採用された．本手法の流れを以下に示す．

- 床面の推定と除去
- 三次元クラスタリング
- 画像情報を用いた識別

本手法では，まず，ポイントクラウドデータをボクセル化と，RANSAC アルゴリズムを用いた平面モデルを当てはめにより，平面の推定をする．そして，平面である床面を除去したポイントクラウドデータにクラスタリングを行い，検出対象候補を推定する．次に，床面から頭部までの高さとクラスタリング領域のサイズから候補を絞り込む．最後に，ポイントクラウド上で検出した候補領域に対応する画像領域から HOG 特徴量を抽出し，SVM により識別して最終判定を行う．本手法では，ポイントクラウドを用いて人の候補領域を絞り込み画像情報を用いて人であるかどうかの最終判断を行うことで，高速な人検出を可能とした．図 2 に，本手法の画像上での検出過程と結果を示す．図 2(a)(b)(c) に示すように，高さやサイズによる制約により，検出対象候補を限定していることがわかる．また，図 2(d) は最終判定結果であり，緑はポジティブ，赤はネガティブと識別されたことを示す．

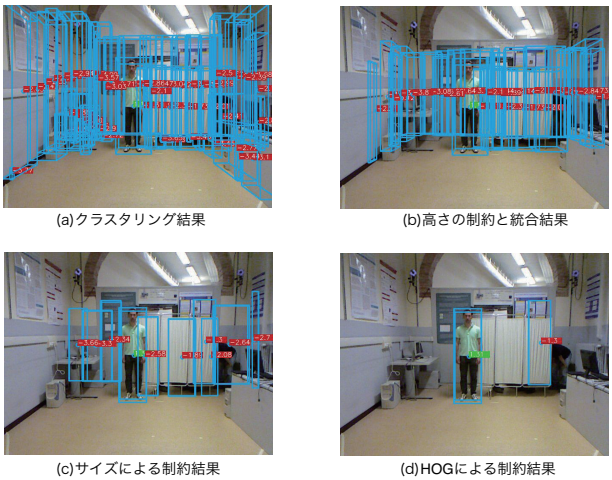


図 2　床面を歩行する人の検出

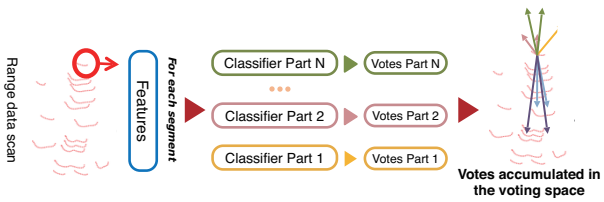


図 3　レイヤー毎にモデルを学習した人検出の概要

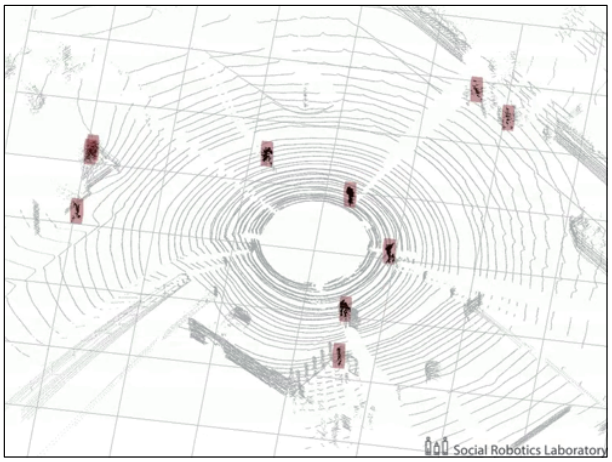


図 4　レイヤー毎にモデルを学習した人検出の検出結果

3.2　レイヤー毎にモデルを学習する人検出

レイヤー毎にモデルを学習する人検出 [7] では，体の断面のレイヤーの集合をパーツとして捉えて識別する． 本手法では，パーツの識別と重心位置の投票による検出の 2 つの処理により人の検出を行う．

パーツの識別では，まず，パーツに含まれるすべての各レイヤーに含まれるポイントクラウドから特徴を抽出する．ポイン

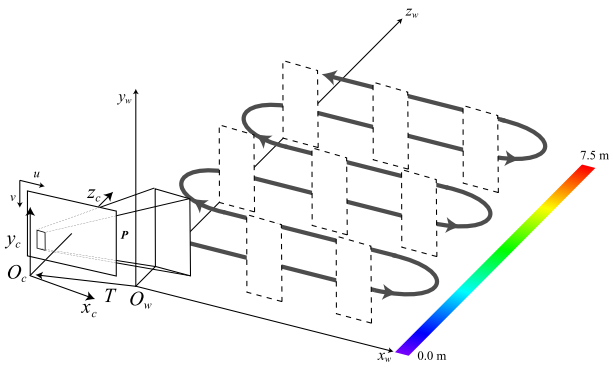


図 5　三次元実空間におけるラスタスキャン

トクラウドからの特徴抽出は，ポイントクラウドデータの点群の幅や量，円形度等の 17 種類の特徴を抽出し，これを特徴ベクトルとする．次に，図 3 に示すように，各パーツ毎に AdaBoost により識別器を学習する．検出時は，AdaBoost を用いてパーツの識別を行う．パーツの識別結果を基に，パーツの重心位置を三次元空間上に投票し，人の重心位置を決定する．本手法によるポイントクラウドからの検出例を図 4 に示す．図 4 の赤い枠で囲われた領域は，ポジティブと判定された領域を示す．

4.　距離画像を用いた物体認識

本章では，距離画像を用いた物体認識における人検出，動作認識，三次元姿勢推定について説明する．距離画像は画素の近傍演算が可能であり，従来の画像処理や認識で用いられた手法で利用できるというメリットがある．

4.1　距離画像からの人検出

距離画像を用いた Real AdaBoost による人検出法 [8] は，距離ヒストグラム特徴量と Real AdaBoost を用いる．

距離ヒストグラム特徴量の抽出方法は，距離画像を 8×8pixel のセルに分割し，セルで構成される 2 つの矩形領域を選択する．選択された 2 つの矩形領域の距離情報から距離ヒストグラムを算出し，ヒストグラムから Bhattacharyya 距離 [9] による類似度を算出する．これを全ての矩形領域の組み合わせに対して行い，各類似度を特徴ベクトルとする．

識別器である Real AdaBoost [10] は，ポジティブクラスの特徴量とネガティブクラスの特徴量の各次元の確率密度関数から分離度を求め，ポジティブクラスとネガティブクラスを最も分離できる特徴量を弱識別器として選択する．学習により選択された弱識別器を $h_t(x)$ とすると，構築される最終識別器 $H(x)$ は弱識別器を $h_t(x)$ の総和となる．オクルージョン領域から抽出される距離情報は，弱識別器の誤った応答を出力する原因となる．従って，このようなオクルージョン領域を捉える弱識別器の出力は，そのまま最終識別器に統合しないようにする．本手法では検出ウィンドウを三次元実空間においてラスタスキャンする．三次元ラスタスキャンとは，三次元空間にて $y_w = 0$ とした地面上において，60 × 180[cm] の検出ウィンドウを z_w を変化させながら x_w 方向へのラスタスキャンを繰り返すことにより，三次元実空間の地面上をラスタスキャンする手法である．この三次元実空間中のラスタスキャンにより得られる検出ウィンドウの三次元座標を画像座標に投影し，投影された座標

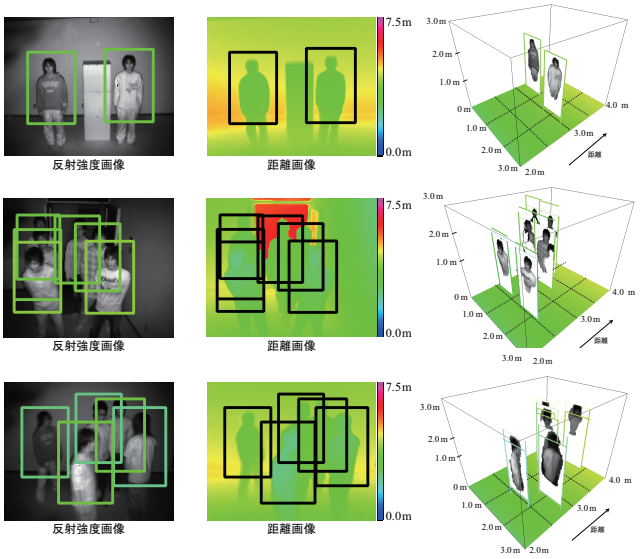


図 7　人検出例

位置に対する領域内に対して検出処理を行うため，ウィンドウの世界座標が既知である．そこで，カメラに対して検出ウィンドウより手前に存在する物体領域をオクルージョンとして判別し，識別に利用する．図 6(a) に示すように，オクルージョン領域を考慮せず最終識別器により識別を行うと，多くの弱識別器の出力がマイナスとなり，その結果，人以外と誤識別される．一方 (b) は，オクルージョン率を考慮して，最終識別器の出力を求めるため，人と正しい識別が可能となる例である．

三次元実空間における検出ウィンドウのラスタスキャンを用いた人検出結果を図 7 に示す．図 7(a) では，人と同様の高さの物体を誤検出しないで，人のみを検出していることがわかる．さらに図 7(b)(c) では，向きの異なる人の重なりが存在しても，それぞれの人とその三次元位置を正確に検出できていることがわかる．

4.2　距離画像からの動作認識

4.2.1　距離動画像からの動作認識

距離動画像からの動作認識法 [11] では，TOF カメラにより得られる距離動画像から，動きを捉えるための時空間特徴と，高さを捉えるための距離特徴を抽出し，統計的学習法により手を伸ばす動作の検出と，どの棚に手を伸ばしたかの手の高さの識別を同時に行う．本手法の流れを図 8 に示す．

a）　時空間特徴と距離特徴

時空間特徴量として，PSA 特徴量を用いる．PSA 特徴量は，ピクセル状態分析 (Pixel State Analysis：PSA) [12] の結果を特徴量とする．ピクセル状態分析とは，ピクセル状態の時間変化をモデル化することにより，各ピクセルを背景 (Background)，静状態 (Stationary)，動状態 (Transient) の三状態に判別する手法である．

算出方法は，まずピクセル状態分析により，前景を人領域として検出する．次に，検出した人領域から PSA 特徴量を抽出する．ヒストグラムを作出することを考慮し，ピクセル状態分析画像の人領域をリサイズし，セル領域に分割する．分割され

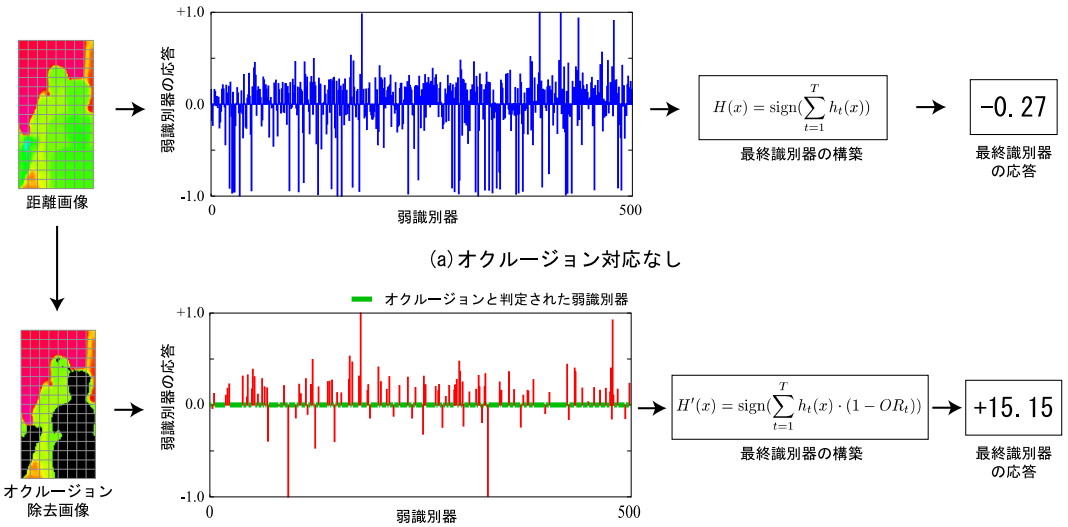


図 6　オクルージョン領域を考慮した識別例

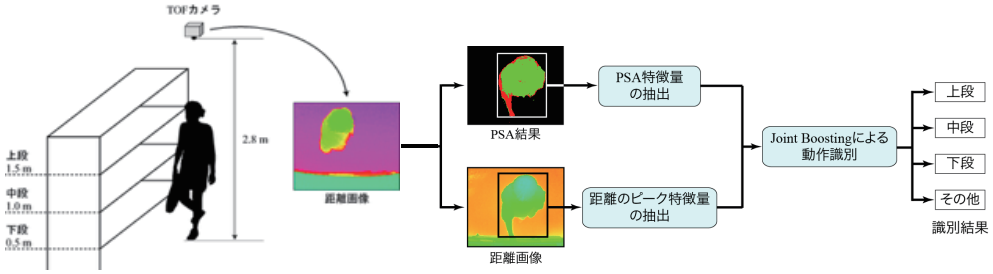


図 8　距離動画像からの動作認識の流れ

たセルを最小とする矩形領域を選択し，各矩形領域からピクセル状態分析結果に基づき PSA ヒストグラムを算出する．これを全てのセル領域から算出することにより PSA 特徴量とする．

距離特徴量として，距離ヒストグラムのピーク値を用いる．算出方法は，検出した人領域をリサイズし，セル領域に分割する．分割したセルを最小とする矩形領域を選択し，各矩形領域から距離ヒストグラムを算出する．算出された各距離ヒストグラムのピーク値を特徴量とする．

b）　マルチクラス識別器の構築

提案手法では，上段，中段，下段，その他の複数クラスの動作識別を行うために Joint Boosting を用いる．Joint Boosting により商品棚の上段，中段，下段から商品を手に取る動作をポジティブクラス，それ以外の動作（立ち止まる，通過等）をネガティブクラスとして学習を行い識別器を構築する．

Joint Boosting [13] は，マルチクラス識別のための Boosting 手法であり，共通する弱識別器集合を共有しながら学習を行う手法である．Joint Boosting における弱識別器を $h_m(v,c)$ とすると，強識別器 $H(v,c)$ は弱識別器 $h_m(v,c)$ の総和である．ここで c はクラスラベルである．Joint Boosting では，クラス集合 $S(n)$ の識別に対して用いられる弱識別器を $h_m^n(v)$ として，強識別器 $G^{S(n)}(v)$ は弱識別器 $h_m^n(v)$ の総和をとる．3 クラスについてのマルチクラス識別器を考える場合，それぞれのクラスを識別する強識別器は $G^{S(n)}(v)$ を用いて以下のように表される．

$$H(v,1) = G^{1,2,3}(v) + G^{1,2}(v) + G^{1,3}(v) + G^{1}(v)$$

$$H(v,2) = G^{1,2,3}(v) + G^{1,2}(v) + G^{2,3}(v) + G^{2}(v) \quad (1)$$

$$H(v,3) = G^{1,2,3}(v) + G^{1,3}(v) + G^{2,3}(v) + G^{3}(v)$$

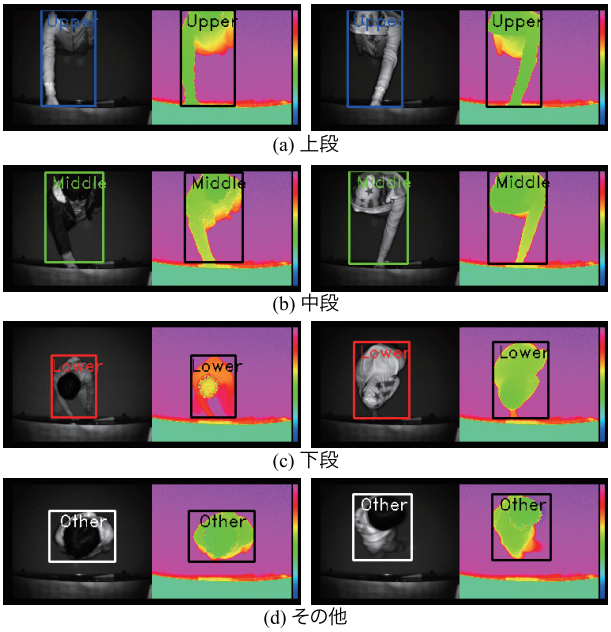
このとき $G^{1,2,3}$ は検出対象 1〜3 すべてと背景を，$G^{1,3}$ は検出対象 1 と 3 すべてと背景を識別するのに有効な弱識別器集合である．Joint Boosting は，強識別器間で共通する弱識別器集合を共有する．3 クラスの識別では，各クラスの強識別器が 4 つの弱識別器集合を必要とするため，合計で 12 個の弱識別器集合が必要となる．

c）　商品を取る動作の識別

図 9 に動作の識別結果を示す．識別結果では，商品棚に手を伸ばしたときの棚の高さの識別が可能であることがわかる．さらに，左右どちらの手で商品を取る場合でも正しい識別が可能である．下段の識別においては，立った状態で商品を取る場合としゃがんだ状態で商品を取る場合どちらでも「下段」の識別が可能である．また，手を伸ばす動作以外の商品を見ている人や通り過ぎる人は「その他」に識別されている．

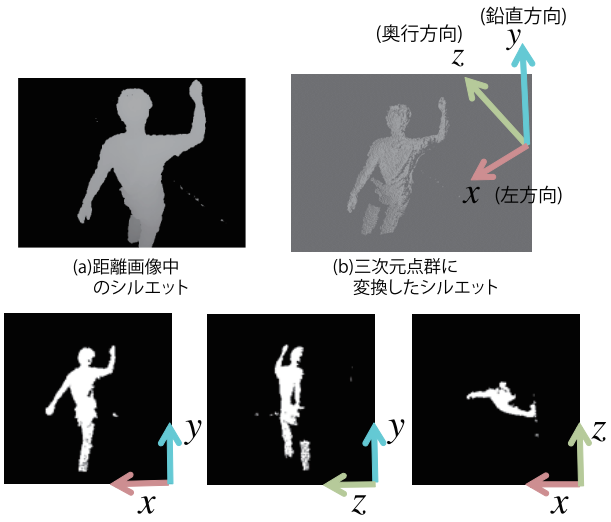
4.2.2　視点に依存しない距離動画像からの動作認識

視点に依存しない距離動画像からの動作認識 [14] は，認識時に動作する人の向きや重心等の人の形状に関わる特徴量を抽出する必要がないため，従来手法の課題である複数の人が接触したり大きく姿勢を変える状況に対応することが可能な手法である．また，認識時のカメラ視点の学習サンプルの数を小さく抑えることで，学習サンプルの収集の手間を小さく抑えることができる．



(a) 上段
(b) 中段
(c) 下段
(d) その他

図 9　商品を取る動作の識別例



(a)距離画像中のシルエット
(b)三次元点群に変換したシルエット
(c)三面図展開後のシルエット

図 10　距離画像の三面図展開

a）　距離画像の三面図展開を用いた生成学習

本手法は，事前ドメインにおける弱識別器の学習と目標ドメインにおける強識別器の学習に分けられる．事前ドメインにおける弱識別器の学習では，事前ドメインの距離画像および動作の教師信号を入力として，距離画像の前処理，カメラ視点を変えたデータの生成，時空間特徴量の抽出，弱識別器の学習を順に行う．本手法では，事前ドメインにおける弱識別器の学習時に，図 10 のように，距離画像の三面図展開することにより，視点の変化への対応を行う．次に，目標ドメインにおける強識別器の学習では，少数の目標ドメインの距離画像と教師信号から，距離画像の前処理，時空間特徴量の抽出を行い，事前ドメインで学習した弱識別器の中から目標ドメインの時空間特徴量に最適な強識別器を構築する．本手法では，以上述べた事前ドメインのデータと少数の目標ドメインのデータを組み合わせた学習により，カメラ視点の変化に追従した強識別器を構築する．

また，本研究では，距離画像中の人のシルエットの見えと動

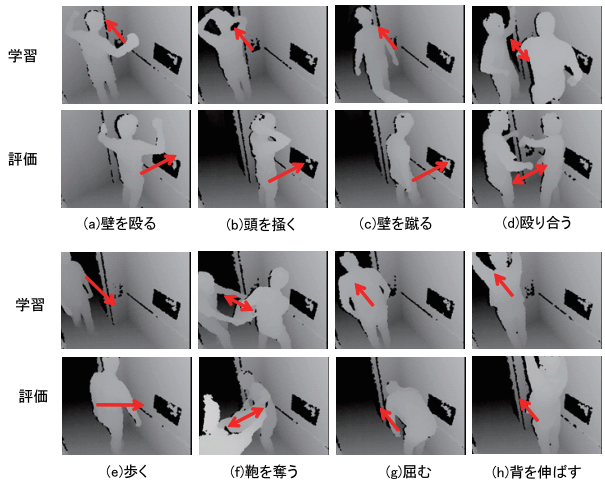


図 11　評価実験のデータ

表 1　F 値の比較

手法	Recall[%]	Precision[%]	F 値
提案手法	90.7	91.3	91.0
目標ドメイン	52.5	54.2	53.4
事前ドメインと目標ドメイン	84.4	88.9	86.6

きを捉える時空間特徴量として，Motion History Image(MHI) を利用する．MHI は，画像上の時間変化を濃淡で記録した特徴量である．MHI 特徴量は，三面図に展開した各平面において抽出する．ビン数が 18 のとき，時空間特徴量は 54 次元となる．これを時間方向に拡張して情報量を増やす．時刻の数が 6 のとき，時空間特徴量は三平面合計で 324 次元である．

b）　動 作 認 識

距離画像センサを室内の天井付近に取り付けて，斜め下を向けて撮影した．この距離画像センサの視野内において，異なる方向を向いて所定の動作する人を撮影して評価データとした．動作のデータは，図 11 に示す 8 種類のカテゴリとした．動作のデータのケース数は, 学習用と評価用の 2 種類，動作の 8 種類, 実験者の 3 名の組み合わせで，48 ケースである．動作のデータのフレーム数は，動作カテゴリと実験者の組み合わせで，学習用に計 1,235 フレーム, 認識用に計 1,242 フレームを用いる．

動作認識の結果を表 1 に示す．比較対象は，目標ドメインのみで Random Forest の学習を行った手法と，事前ドメインと目標ドメインで Random Forest の学習を行った手法である．表 1 より，提案手法は異なる視点においても三面図展開を用いることで，最も高い F 値を得ることができる．これは，距離画像をポイントクラウドに変換し，その後，三面図に展開するために，カメラの視点位置に依存しないシルエット形状を得られたからである．

4.3　人体三次元姿勢推定

人体の三次元姿勢推定技術は，自然な動作を取り入れたユーザーインターフェース (Natural User Interface) として，ゲームなどの操作入力に利用されている．このような人体姿勢推定は，距離画像から検出した人体に対してパーツの識別を行い，各パーツの重心位置を求めることで実現されている．人体パーツ識別とは，検出した人領域の各画素が頭や腕，足といった人

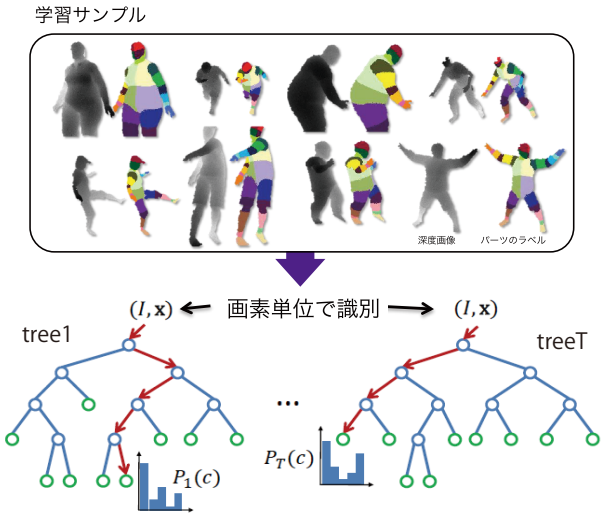


図 12　人体パーツ識別の概要

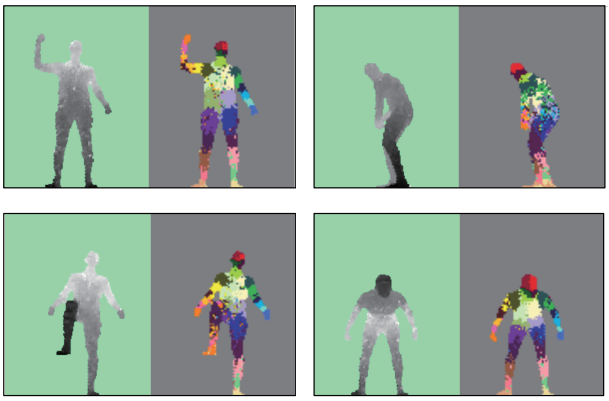


図 13　人体パーツ識別結果

体パーツのどのクラスに属するかを識別することである．人体パーツは複数あるため，多クラス識別器である Random Forest が利用されている [15]．図 12 の上部にあるように，距離画像と各パーツを色で表した正解ラベルの組み合わせを学習サンプルとして，Random Forest を構築する．Random Forest の各ノードでは，距離画像 I の注目画素からオフセット $\boldsymbol{u}$ と $\boldsymbol{v}$ 離れた 2 点の画素が持つ距離値の差を特徴量として，しきい値処理により左右の子ノードに分岐する関数を用いる．パーツ識別時は，距離画像を決定木に入力し，各分岐関数において指定された 2 点の距離値の差を用いて分岐していく．到達した末端ノードのクラス確率を用いて，注目画素のパーツラベルを推定する．人体領域の全画素についてパーツ識別を行い，パーツラベルごとに重心を求めることで，人体の姿勢を求めることができる．本手法におけるパーツの識別結果を図 13 に示す．

5. ま と め

本稿では，距離情報の表現方法についてと，距離情報を用いた物体認識法について述べた．ポイントクラウドデータを用いた人検出と，距離画像を用いた人検出，動作認識，三次元姿勢推定の手法について述べ，距離情報と機械学習との組み合わせにより高精度化が可能であることを示した．図 14 に，距離画像とポイントクラウドにおける物体検出のメリットとデメリットと，その応用先をまとめた．今後は，さらなる人式性能の高

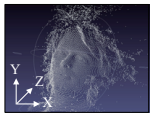
距離画像	ポイントクラウド
 D={d(u,v)}	 P={(X, Y, Z)}
ラスタスキャンベース物体検出 メリット： 従来の画像処理アルゴリズムや、物体認識フレームワークを適用可能 デメリット： ラスタスキャンによる計算コスト	クラスタリングベースの物体検出 メリット： ラスタスキャンを必要としない PCL等で今後さらに充実 デメリット： 前処理が多い（フィルタリングや穴埋め等）
人検出、人体姿勢推定、動作認識等	人検出、平面検出、障害物検知等

図 14　距離画像とポイントクラウドの比較

精度化を目的とし，距離画像により適した特徴量を学習により自動獲得する Convolutional Neural Netwark(CNN) [16] の導入が期待されている．

文　献

[1] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features”, CVPR, vol.1, pp.I-511-I-5i8, 2001.

[2] Y. Freund and R. E. Schapire, “Experiments with a new boosting algorithm”, ICML, vol.96, 1996.

[3] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection”, CVPR, vol.1, pp.886-893, 2005.

[4] C. Cortes and V. Vapnik, “Support-vector network”, Machine Learning, vol.20, no.3, pp.273-293, 1995.

[5] L. Breiman, “Random Forests”, Machine Learning, vol.45, pp.5-32, 2001.

[6] M. Munaro, F. Basso and E. Menegatti, “Tracking people within groups with RGB-D data”, IROS, pp.2101-2107, 2012.

[7] L. Spinello, K. O. Arras, R. Triebel and R. Siegwart, “A layered approach to people detection in 3D range data”, AAAI, 2010.

[8] 池村翔，藤吉弘亘，“距離情報に基づく局所特徴量によるリアルタイム人検出”，電子情報通信学会論文誌，vol.J93-D，no.3，pp.355-364，2010.

[9] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by probability distributions”, Bull. Calcutta Math. Soc., vol.35, pp.99-109, 1943.

[10] R. E. Schapire and Y. Singer, “Inproved boosting algorithms using confidence-rated predictions”, Machine Learning, no.37, pp.297-336, 1999.

[11] 池村翔，藤吉弘亘，“時空間情報と距離情報を用いた Joint Boosting による動作識別”，電気学会論文誌 C (電子・情報・システム部門誌)，vol.130，no.9，pp.1554-1560，2010.

[12] H. Fujiyoshi and T. Kanade, “Layered detection for multiple overlapping objects”, IEICE, vol.E87-D, pp.2821-2827, 2004.

[13] A. Torralba, K. P. Murphy and W. T. Freeman, “Sharing features: efficient boosting procedures for multiclass object detection”, CVPR, pp.762–769, 2004.

[14] R. Yumiba and H. Fujiyoshi, “Viewpoint-independent action recognition method using depth image” , CVIM, vol.197, no.30, pp.1-16, 2015.

[15] J. Shotton, M. Johnson and R. Cipolla, “Semantic texton forests for image categorization and segmentation”, CVPR, 2008

[16] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition”, Neural Computation, vol.1, pp.541-551, 1989.