様式 C-19

科学研究費補助金研究成果報告書

平成２３年３月３１日現在

機関番号：３３９１０
研究種目：若手研究（B）
研究期間：2009～2010
課題番号：２１７００２２８
研究課題名（和文）
TOF カメラを用いた特徴量間の共起による人検出と動作認識技術に関する研究
研究課題名（英文）
Human Detection and Action Recognition by Feature Co-occurrence using TOF Camera
研究代表者
藤吉　弘亘（FUJIYOSHI HIRONOBU）
中部大学・工学部・教授
研究者番号：２０３３３１７２

研究成果の概要（和文）：本研究では TOF カメラから得られる距離情報を用いて人の重なりや複雑なシーンに頑健なリアルタイム人検出手法を提案する．提案手法では，距離画像から 2 つの局所領域の距離関係を捉えることができる距離ヒストグラム特徴量を抽出する．抽出された特徴量を用いて Real AdaBoost 識別器を構築し，人の識別を行う．評価実験の結果，誤検出率 5.0%において検出率 98.9%となり，HOG 特徴量を用いた従来法と比較して 4.9%検出率を向上させることができた．また，提案手法は約 10fps でリアルタイムに人検出が可能であることを確認した．

研究成果の概要（英文）： We proposed a method for detecting humans that uses depth information obtained from a TOF camera. Our method calculates features derived from a depth histogram that represents the relationship between two local regions. Our method achieved a detection rate of 98.9% with a false positive rate of 5.0%. It also had a 4.9% higher performance than the conventional method, and our detection system can run in real-time (10 fps).

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	3,000,000	900,000	3,900,000
２０１０年度	500,000	150,000	650,000
年度			
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：画像処理、パターン認識
科研費の分科・細目：情報学・知覚情報処理・知能ロボティクス
キーワード：人検出、動作認識、距離情報、Joint Boosting、Haar-like フィルタ

１．研究開始当初の背景

1990 年代後半より，ロンドンの大量監視カメラ(1997 年)，日本の歌舞伎町における監視カメラ(2002 年)など，大規模な監視カメラ導入とともに，物体検出や追跡などの動画像理解技術の要求が高まりつつある．さらに，知的な映像監視に関する研究は 1997 年の DARPA の自動ビデオ監視システムの研究プロジェクトであった VSAM(Video Surveillance and Monitoring)プロジェクトをきっかけに，以来これらの動画像理解技術はいっそう進展した．

最近では，これらの技術を基にした製品の実用化が盛んに取り組まれるようになり，監視目的だけではなく，オフィスや家，公共施設などの空間において，その空間内の人の意図を理解し行動を支援する技術への展開が期待されている．ジョージア工科大学の Aware Home とういプロジェクトでは，生活空間である家にカメラなどのセンサ群を埋め込み，

24 時間を通して生活空間における人の動きをセンシングする研究が取り組まれている．Microsoft Research の Easy Living では、リビングルームを対象としたセンシングにより得られた情報を基に，ユーザである人に対して快適な空間をアシストする研究が盛んである．このような活力生活支援技術は QoLT(Quality of Life Technology)と呼ばれる技術の一環である．また，カーネギーメロン大学のロボット工学研究所では、2005 年に PIA(People Image Analysis)プロジェクトを立ち上げ，人に関するあらゆる画像(顔，上半身, 全身など)を解析するための研究を横断的に行っている．
これらの技術には，刻々と変化する人の状態を認識する必要がある．特に，人の姿勢や動きをセンシングする 3 次元物体認識技術が必要不可欠な要素であり，重要な研究課題の一つとなっている．

２．研究の目的
本研究では，リアルタイムに距離情報を取得するとこができる TOF (Time of Flight)カメラを用いた人検出と動作認識の実現を目的とする．

(1)TOF カメラを用いた人検出
可視光カメラを用いた人検出では，局所特徴量と統計的学習からなる手法が多く用いられている．局所特徴量には，勾配ベースの特徴量である HOG, Edgelet 等が利用されていが，複雑な背景や人の重なりが多い場合，人の形状を捉えることが困難となり，検出精度が低下することがある．これに対して，距離情報から局所特徴量を抽出すると，テクスチャや色に依存しない距離の情報を用いるため，人の重なりや複雑なシーンの影響を軽減することが期待できる．そこで，本研究では，複雑な背景や人の重なりに影響されない，高精度な人検出の実現を目的としている．

(2)時空間情報と距離情報を用いた動作認識
コンビニやスーパーマーケット等の店舗では，購買者の年齢や性別，購入商品の種類や個数といった情報は重要な情報であり，レジの POS システムや売り上げの記録等から取得されている．しかしながら，来店客が店舗内でどのような商品に注目し，手に取り，興味を持ったかという情報を取得することはできない．このような情報は，商品の配置や POP 設置などのマーケティングにおいて重要であり，来店客が商品に注目する行動の自動検出の実現が期待されている．そこで，本研究では，来店客が商品を取る動作の検出と，商品棚の上段，中段，下段のどの高さの棚に手を伸ばしたかの動作識別の実現を目的としている．

(3)Haar-like フィルタリングによる人検出
統計的学習を用いた人検出手法[1]は、汎化能力の高い識別が可能であるが、事前に大量のデータが必要となる問題がある．また、作成した学習データの撮影環境と異なるシーンにおける人検出は困難となるため、人検出技術の実用化にはこれらの問題を解決する必要がある．このような問題に対して、本研究では、統計的学習手法を用いない人検出の実現を目的としている．

３．研究の方法
TOF カメラは，LED より強度変調して発光された赤外線が対象物体に反射してカメラで観測されるまでの時間を計測し，その飛行時間から物体までの距離を取得するカメラである．本研究では，TOF カメラとして MESA 社の SR-3100，SR-4000 を用いる．

(1)TOF カメラを用いた人検出
本手法では，統計的学習手法である Real AdaBoost により，事前に識別器を学習する．Real AdaBoost による学習では，人と人以外の距離画像を多数用意し（約 12000 枚），それぞれの距離画像から特徴量を抽出し，人と人以外を最も分離できる特徴量を選択することで学習を行う．学習により構築した識別器を用いて距離画像をラスタスキャンすることで人を検出する．
まず，TOF カメラ（SR-3100）により得られた距離画像上において，その距離に適したサイズの検出ウィンドウを用いてラスタスキャンする．ある決定された検出ウィンドウを距離画像から切り出し，距離情報に基づく局所特徴量を抽出する．本手法で用いる局所特徴量は，検出ウィンドウをセル分割し，セルを最小とする矩形領域から 2 つの領域を選択し，それぞれの領域から距離ヒストグラムを抽出する．Bhattacharyya 距離により 2 つの距離ヒストグラム間の類似度を算出することで，距離に基づく局所特徴量とする．次に算出された特徴量のオクルージョン判定を行い，学習により構築された Real AdaBoost 識別器を用いて各検出ウィンドウが人か人以外かの識別を行う．ラスタスキャンによる人識別後，3 次元実空間における Mean-Shift クラスタリングにより人と識別された検出ウィンドウを統合し，人領域を決定することで人検出を実現する．

(2)時空間情報と距離情報を用いた動作認識
本手法では，TOF カメラ（SR-4000）の距離動画像から，まず，ピクセル状態分析(PSA)により人領域を検出し，動きを捉えるための時空間特徴である PSA 特徴と，高さを捉えるための距離特徴である距離ヒストグラム

のピーク値を抽出する．PSA 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り，各ピクセル毎の動きを捉えることができる特徴である．抽出された特徴量を用いて，マルチクラスの Boosting 手法である Joint Boosting により学習を行う．Joint Boosting は，商品棚の上段，中段，下段から商品を手に取る動作をそれぞれポジティブクラス，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することで識別器を構築する．Joint Boosting は，棚に手を伸ばす動きと，そのときの手や体の高さを捉えるように学習を行うため，手や体の動きと高さを同時に捉える識別器を構築し，動作認識を実現する．

(3)Haar-like フィルタリングによる人検出
本手法では，人を上部から撮影した距離情報に対し背景差分を行うことで物体領域を抽出する．抽出された物体領域が人かどうかを判別するために，人の肩，頭，肩の凸形状を抽出する Haar-like によるフィルタリング処理を行う．Haar-like フィルタとは，黒と白の領域の平均距離の差を応答値とするフィルタであり，黒の領域が高い場合には正の応答値を出力し，白の領域が高い場合には負の応答値を出力する．本手法では，人の肩，頭，肩の高さの違いを捉えるために白，黒，白と領域を配置した Haar-like フィルタを用いる．この Haar-like フィルタの応答値が正の値の場合，黒い領域が白い領域に比べ高いことがわかるため，凸形状を抽出するフィルタとなる．本手法では，人の向きの違いに対応するために，0°，45°，90°，135° の 4 方向のフィルタリング処理を行うことで凸形状の判別を行う．Haar-like フィルタリングにより，凸形状と抽出された点を Mean-Shift クラスタリングにより統合することで，統計的学習手法を用いない人検出を実現する．

４．研究成果
(1)TOF カメラを用いた人検出
本手法の有効性を確認するため，特徴量の評価実験およびオクルージョン対応の評価実験を行った結果を以下に示す．
実験に用いるデータベースには TOF カメラで撮影したシーケンスを用いる．屋内で TOF カメラを約 2.5m の高さに設置し，人の歩行シーンと複数の人が重なり合うシーンを対象とした．撮影した屋内のシーケンスから切り出した学習用ポジティブサンプル 1346 枚，学習用ネガティブサンプル 10000 枚を用いる．また，評価には学習用とは別に作成した評価用ポジティブサンプル 2206 枚，評価用ネガティブサンプル 8100 枚を用いる．TOF カメラは屋内において最長 7.5m までの撮影となるため，複数の人の全身を撮影することが困難である．そのため本実験では人の上半身（全身の上部 60%）を検出対象とした．
実験結果の比較には，Receiver Operating Characteristic (ROC)カーブを用いる．ROC カーブとは，横軸に誤検出率，縦軸に検出率を表したものである．識別器の閾値を変化させることによって，誤検出率に対する検出率の比較を行うことが可能である．グラフ左上に近いほど検出性能が良いことを表す．
図 1 に評価実験結果を示す．まず，特徴量の違いについての識別精度について述べる．距離ヒストグラム特徴量(矩形領域サイズの可変あり，オクルージョン対応なし)は，誤検出率 0.5%において検出率 94.0%であり，距離画像の HOG 特徴量より 13.0%識別率を向上した．これは，矩形領域サイズを可変にすることにより，領域の距離関係を有効に捉えることが可能となるためである．また，距離ヒストグラム特徴量(矩形領域サイズの可変あり，オクルージョン対応なし)のみと，HOG 特徴量と距離ヒストグラム特徴量(矩形領域サイズの可変あり，オクルージョン対応なし)の両者を用いた特徴量を比較すると，検出精度は同精度であった．これは，弱識別器の学習時に距離ヒストグラム特徴量が多く(99%)選択されており，HOG 特徴量が識別に貢献していないといえる．
次にオクルージョン対応について述べる．距離ヒストグラム特徴量(矩形領域サイズの可変あり，オクルージョン対応あり)は，誤検出率 0.5%において検出率 96.9%であり，距離ヒストグラム特徴量(矩形領域サイズの可変あり，オクルージョン対応なし)と比較し 2.9%識別率を向上した．さらに，HOG+距離ヒストグラム特徴量(矩形領域サイズの可変あり，オクルージョン対応あり)においても検出率が向上している．これは，オクルージョン率を用いて識別に有効な弱識別器に重みづけし，最終識別器の出力を求めることにより，オクルージョン領域の影響を抑えることができたといえる．

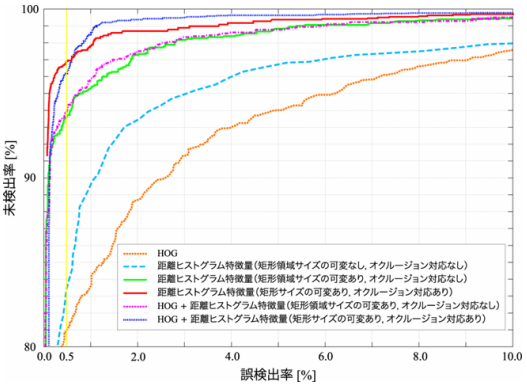

図 1　統計的学習手法による人検出精度

図 2 に距離画像からの人検出例を示す．(a)では，距離情報から得られる人の形状を学習により捉えるため，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(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 3 次元位置を正確に検出できていることがわかる．本手法は，Intel Xeon CPU 3.00GHz を用いた際に 1 フレームの処理時間が約 100ms であるため，約 10fps でのリアルタイム検出が可能である．
距離画像による人検出は，可視光カメラを用いた人検出[1]と比較し，約 17%識別率を向上することができたため，人検出において距離情報が有効であることが確認できた．

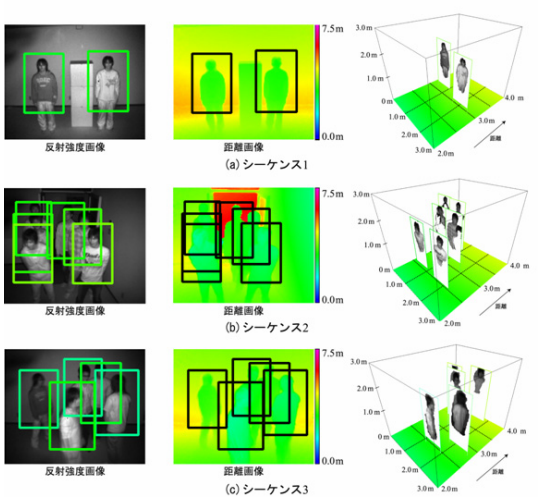

図 2　統計的学習手法による人検出例

(2)時空間情報と距離情報を用いた動作認識
本手法の有効性を確認するため，評価実験を行った結果を以下に示す．
本実験では TOF カメラを用いて人を上部から撮影したシーケンスを用いる．データベースは，棚の高さを上段（1.5m），中段（1.0m），下段（0.5m）とし，実験室において 8 人の学生が各動作を行っている距離画像から人領域を切り出したものである．学習サンプルとして，上段から商品を取る動作(927 枚)，中段から商品を取る動作(1122 枚)，下段から商品を取る動作(1101 枚)，立ち止まる，通過等のその他(6333 枚)を用いる．また，評価には学習サンプルとは別に撮影したシーケンスとして，上段から商品を取る動作(805 枚)，中段から商品を取る動作(705 枚)，下段から商品を取る動作(742 枚)，その他(4532 枚)を用いる．
評価用データベースを用いて，商品を取る動作の識別実験を行い，特徴量の識別精度による比較を行う．比較には，本手法である PSA 特徴と従来の動作認識手法で用いられている CHLAC 特徴，ST-patch 特徴を用いる．さらに，各時空間特徴と距離情報である距離のピーク値の両者を用いた特徴量として，PSA 特徴＋距離情報，CHLAC 特徴＋距離情報，ST-patch 特徴＋距離情報についても

表 1　商品を取る動作の識別率[%]

	上段	中段	下段	その他	全て
PSA	52.9	63.8	65.4	92.9	82.1
CHLAC	13.5	26.5	29.2	93.2	69.8
ST-patch	49.6	56.3	64.6	92.9	80.9
PSA + Depth	88.4	84.4	83.0	96.6	93.2
CHLAC + Depth	83.5	79.1	88.3	93.9	90.1
ST-patch + Depth	88.3	73.6	78.4	95.9	90.4

表 1 に各動作の識別率を示す．PSA 特徴，CHLAC 特徴，ST-patch 特徴のみでは，高さが捉えられないため識別率が低いことがわかる．これに対し，各時空間特徴量に距離情報を付加した特徴量は，識別精度が向上していることがわかる．特に PSA 特徴においては，距離情報を付加することで 93.2%で識別が可能となり，CHLAC 特徴＋距離情報による識別と比較して 3.1%，ST-patch 特徴＋距離情報による識別と比較すると 2.8%識別率を向上した．これは，PSA 特徴が CHLAC 特徴や ST-patch 特徴と比較して，人の部分的な，テクスチャに依存しない動き情報を捉えることができるためである．
図 3 に動作の識別例を示す．識別例から，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」に識別されていることが確認できた．

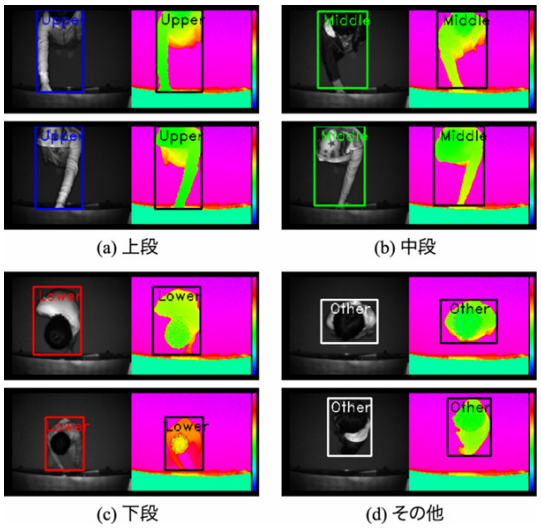

図 3　動作識別例

(3)Haar-like フィルタリングによる人検出

本手法の有効性を確認するため，評価実験を行った結果を以下に示す．

評価用のデータベースには異なる環境で撮影された 3 種類のデータベースを用いる．シーケンス 1 は，大人を対象としたシーケンスとなっており，大勢の人を撮影した混雑しているシーンとなっている．シーケンス 2 は，大人と子供が同時に存在するシーンである．また背景には人以外の物体も存在する．シーケンス 3 は，階段において人が上り下りするシーンであり，階段の高さが変化することにより人の高さも変化する．

表 2，3，4 に各シーケンスに対する検出実験結果を示す．実験結果より，本手法である Haar-like + Mean-Shift による 3 種類のシーケンスに対する検出率の平均は，Mean-Shift のみの検出と比較し 6.6%検出率を向上させることができた．これは図 4 に示すように，人同士が近接した場合に Mean-Shift クラスタリングのみを用いて距離情報を統合すると，近接している人同士の距離情報を統合してしまうため，2 人の場合でも 1 人と検出し，未検出となる．一方，本手法は Haar-like フィルタリングを用いることにより人の凸形状を検出することにより，人の中心部分の距離情報を用いて Mean-Shift クラスタリングを行うため，近接し合う人の距離情報を分離できるため未検出を抑制することができた．

表 2　シーケンス 1 に対する人検出精度

	真値 [人]	検出数 [人]	検出率 [%]
Mean-Shift	477	428	89.7
Haar-like + Mean-Shift		471	98.7

表 3　シーケンス 2 に対する人検出精度

	真値 [人]	検出数 [人]	検出率 [%]
Mean-Shift	283	249	88.0
Haar-like + Mean-Shift		271	95.8

表 4　シーケンス 3 に対する人検出精度

	真値 [人]	検出数 [人]	検出率 [%]
Mean-Shift	291	278	95.5
Haar-like + Mean-Shift		286	98.3

図 4 に本手法による人検出例を示す．本手法は，図 4 (a)のような混雑したシーン，(b)のような大人と子供が同時に存在するシーン，人以外の物体が存在するシーンにおいて高精度な人検出が可能であることがわかる．また，(c) のような階段のシーンにおいても人検出が可能である．本手法は，Haar-like フィルタリングの計算に Integral Image を用いているため，高速なフィルタリング処理が可能であり，約 17fps でのリアルタイム人検出が可能であることを確認できた．

［文献］

[1] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection”, In Proc. Computer Vision and Pattern Recognition, pp. 886-893, 2005.

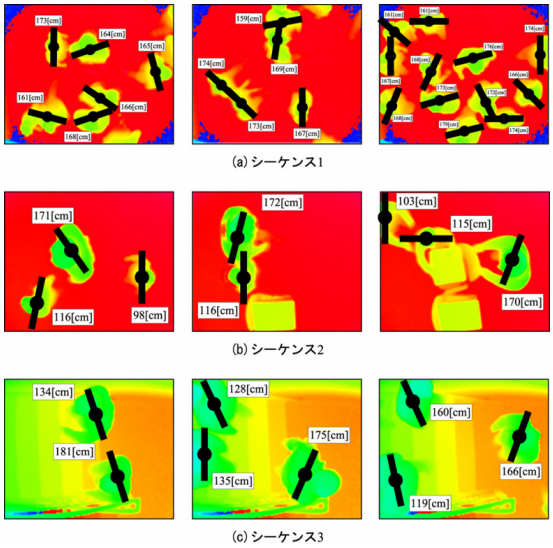

図 4　Haar-like フィルタリングによる人検出例（数値は検出した人の身長を表す）

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

① 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、査読有、Vol. 130、No. 9、2010、pp. 1554-1560

② 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、査読有、Vol. J93-D、2010、pp. 355-364

〔学会発表〕（計 6 件）

① Hironobu Fujiyoshi、Real-Time Human Detection using Relational Depth Similarity Features、The Tenth Asian Conference on Computer Vision、2010 年 11 月 11 日、Queenstown, NZ

② 藤吉弘亘、距離情報を用いた Haar-like フィルタリングによる人検出、第 16 回画像センシングシンポジウム、2010 年 6 月 11 日、パシフィコ横浜

③ 藤吉弘亘、時空間情報と距離情報を用い

たマルチクラス Boosting による動作識別、第 16 回画像センシングシンポジウム、2010 年 6 月 11 日、パシフィコ横浜
④ 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、電気学会一般産業研究会、2009 年 12 月 12 日、徳島大学
⑤ 藤吉弘亘、時空間情報と距離情報を用いたマーケティングのための動作識別、平成 21 年度電気関係学会東海支部連合大会、2009 年 9 月 11 日、愛知工業大学
⑥ 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、第 15 回画像センシングシンポジウム、2009 年 6 月 12 日、パシフィコ横浜

○出願状況（計 1 件）

名称：物体検出装置
発明者：藤吉弘亘
権利者：中部大学
種類：特許
番号：特願 2009-133698
出願年月日：2009 年 6 月 3 日
国内外の別：国内

６．研究組織
(1)研究代表者
藤吉　弘亘（FUJIYOSHI　HIRONOBU）
中部大学・工学部・教授
研究者番号： 20333172