時空間情報と距離情報を用いたJoint Boostingによる動作識別

池村 翔*，藤吉 弘亘（中部大学）

Action Classification by Joint Boosting Using Spatiotemporal and Depth Information

Sho Ikemura*, Hironobu Fujiyoshi (Chubu University)

Abstract

This paper presents a method for action classification by using Joint Boosting with depth information obtained by TOF camera. Our goal is to classify action of a customer who takes the goods from each of the upper, middle and lower shelf in the supermarkets and convenience stores. Our method detects of human region by using Pixel State Analysis (PSA) from the depth image stream obtained by TOF camera, and extracts the PSA features captured from human-motion and the depth features (peak value of depth) captured from the information of human-height. We employ Joint Boosting, which is a multi-class classification of boosting method, to perform the action classification. Since the proposed method employs spatiotemporal and depth feature, it is possible to perform the detection of action for taking the goods and the classification of the height of the shelf simultaneously. Experimental results show that our method using PSA feature and peak value of depth achieved a classification rate of 93.2%. It also had a 3.1% higher performance than that of the CHLAC feature, and 2.8% higher performance than that of the ST-patch feature.

キーワード：動作識別，時空間情報，距離情報，Joint Boosting
(Action Classification，Spatiotemporal Information，Depth Information，Joint Boosting)

1. はじめに

コンビニやスーパーマーケット等の販売店において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，現在レジの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，POSシステム等では店舗内で来店客が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，商品の配置などのマーケティングにおいて重要であり，自動で商品に注目する行動(商品注目行動)を検出し，どのような動作であるかを識別することが期待されている．そこで，本研究では，来店客が商品を取る動作の検出と，どの高さの商品棚に手を伸ばしたかの動作識別を目的とする．

従来の動作識別手法に用いられる特徴量には，立体高次局所自己相関(Cubic Higher-order Local Auto-Correlation: CHLAC)特徴とST-patch特徴が挙げられる．小林らは，フレーム間差分により得られる時系列2値画像に対して変位パターンを当てはめ，時間方向に積算したCHLAC特徴を用いて判別分析法により動作検出を行う手法を提案している[(1)]．南里らは，このCHLAC特徴を用いて異常行動検知を実現している[(2)]．Sukthankarらは，物体の局所的な「アピアランス」と「モーション」の時間的変化を捉えるSpace-Time Patch(ST-patch)特徴[(3)]を用いた動作識別の手法を提案している[(4)]．村井らは，ST-patch特徴を用いてエスカレータシーンにおける人の異常行動検知を実現している[(5)]．これらの動作識別に用いられているCHLAC特徴やST-patch特徴は，通常の単眼カメラから得られる画像のみを用いているため，商品を取る動作を検出することは可能であるが，高さ情報を得ることができないため，どの高さの棚に手をのばしたかの識別が困難であると考えられる．

そこで本稿では，TOF(Time of Flight)カメラから得られる距離情報を用いて，商品を取る動作の検出と棚の高さの識別を行う手法を提案する．提案手法は，TOFカメラから得られる距離情報を用いてピクセル状態分析により人領域の検出を行う．検出された人領域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，マルチクラス識別のためのBoosting手法であるJoint Boostingにより棚の上段，中段，下段から商品を取る動作を識別する．提案手法は，商品を取る動作の検出と棚の高さの識別を同時に行う．

2. 従来の動作識別のための特徴量

従来の動作識別のための特徴量であるCHLAC特徴はイベント検出[(1)]や異常行動検知[(2)]等に用いられている．また，ST-patch特徴は動作識別[(3)]，イベント検出[(4)]，異常行動検知[(5)]等に用いられている．

〈2･1〉 CHLAC特徴　立体高次局所自己相関特徴（Cubic Higher oder Local Auto-Correlation : CHLAC）は，顔検出等の物体検出に用いられている2次元画像を対象とする高次局所自己相関特徴（Higher oder Local Auto-Correlation : HLAC）[(6)]に時間軸を加えることで3次元に拡張したものであり，動画像中に出現する動物体の「動き」の「形」を表現することができる特徴量である．文献[(2)]ではこの特徴量を異常行動検知に用いており，文献[(1)]ではイベント検出に用いている．CHLAC特徴のN次自己相関関

数の定式化の形は以下の式のようになる．

$$x_f^N(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_N)=\int f(\boldsymbol{r})f(\boldsymbol{r}+\boldsymbol{\alpha}_1),\ldots,f(\boldsymbol{r}+\boldsymbol{\alpha}_N)d\boldsymbol{r}\cdots\cdots(1)$$

ここで，f は時系列 2 値化画像，$\boldsymbol{r}$ は画像内の 2 次元画像座標，N 個の変位 $\boldsymbol{\alpha}_i(i=1,\ldots,N)$ は時間成分として持つ 3 次元のベクトルである．また，時間方向の積分範囲は，どの程度の時間方向の相関を取るのかのパラメータである．HLAC 特徴の場合，変位ベクトルの組み合わせは図 1(a) に示すように，0 次は 1 個，1 次は 4 個，2 次は 20 個の計 25 次元のベクトルとなる．そして，CHLAC 特徴の場合，変位ベクトルの組み合わせは図 1(b) に示すように，0 次は 1 個，1 次は 13 個，2 次は 237 個の計 251 次元のベクトルとなる．CHLAC 特徴は図 1 に示すように，変位パターンに適合するピクセル数をカウントしたものである．

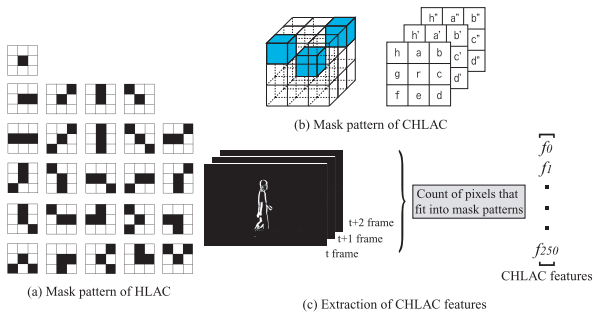


図 1　CHLAC 特徴
Fig. 1.　CHLAC features.

〈2•2〉 Space-Time Patch 特徴　Space-Time Patch (ST-patch) 特徴 [3] は，Shechtman らにより提案されている物体の局所的な「アピアランス」と「モーション」の時間的変化を捉えた特徴である．この ST-patch 特徴はイベント検出 [4] [7] や移動方向を考慮した物体検出とセグメンテーション手法 [8] など，様々な手法に応用されている．図 2 に概要を示す．ここで，x, y は画像の座標軸，t は時間軸，3 本の線は個々の画像の動き，$[u\ \ v\ \ w]^T$ は ST-patch 内の動きベクトル，∇P_i は個々の画素の勾配方向ベクトルを表している．この ST-patch 特徴を得るために，時空間画像においての x 軸，y 軸の勾配を求める．画像中の動きが一定の場合，各軸に対するある画素 i の勾配 $\nabla P_i=(P_{xi},P_{yi},P_{ti})$ は，画素の動き方向ベクトル $[u\ \ v\ \ w]^T$ に対して垂直となる．よって式 (2) の関係が成り立つ．

$$\nabla P_i\begin{bmatrix}u\\v\\w\end{bmatrix}=0\cdots\cdots(2)$$

画素数が n の場合，式 (2) は式 (3) となる．

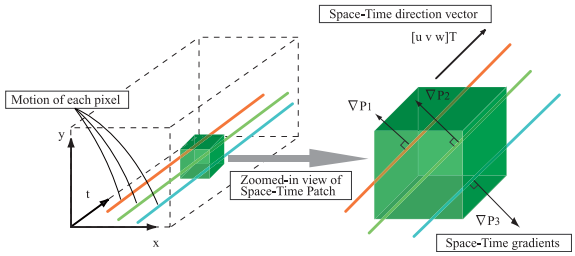


図 2　ST-patch 特徴
Fig. 2.　ST-patch features.

$$\begin{bmatrix}P_{x1}&P_{y1}&P_{t1}\\P_{x2}&P_{y2}&P_{t2}\\\vdots&\vdots&\vdots\\P_{xn}&P_{yn}&P_{tn}\end{bmatrix}\begin{bmatrix}u\\v\\w\end{bmatrix}=\begin{bmatrix}0\\0\\\vdots\\0\end{bmatrix}\cdots\cdots(3)$$

ST-patch 中の n 画素の ∇P_i からなる $n\times 3$ の行列を $\mathbf{G}$ とし，行列 $\mathbf{G}^{\mathrm{T}}$ を掛けると式 (4) となる．

$$\mathbf{G}^{\mathrm{T}}\mathbf{G}\begin{bmatrix}u\\v\\w\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}\cdots\cdots(4)$$

この様に，行列 $\mathbf{G}^{\mathrm{T}}\mathbf{G}$ は 3×3 の行列となる．ここで，行列 $\mathbf{G}^{\mathrm{T}}\mathbf{G}$ を $\mathbf{M}$ とすると式 (5) のように表すことができる．

$$\mathbf{M}=\mathbf{G}^{\mathrm{T}}\mathbf{G}=\begin{bmatrix}\sum P_x^2&\sum P_xP_y&\sum P_xP_t\\\sum P_yP_x&\sum P_y^2&\sum P_yP_t\\\sum P_tP_x&\sum P_tP_y&\sum P_t^2\end{bmatrix}\cdots(5)$$

式 (5) より求められる行列 $\mathbf{M}$ が 1 つの ST-patch から得られる特徴となる．行列 $\mathbf{M}(3\times 3)$ には，左上の 2×2 の行列にアピアランスの情報，3 行目と 3 列目にはモーションの情報を持つ．

〈2•3〉 従来の特徴量の問題点　本研究では，図 3 に示すように，来店客が商品棚から商品を取る動作において，上段，中段，下段のどの棚から商品を取るかの識別を行う．人の手や体の高さを捉えるために，TOF カメラを 2.8m の高さに設置し，人を上部から撮影する．

人を上部から撮影し，商品を取る棚の高さを識別する場合，従来の特徴量である CHLAC 特徴では，画像全体から変異パターンのヒストグラムを作成するため，局所的な動きの情報や，位置の情報が失われてしまうという問題がある．一方，ST-patch は，局所領域の動きを捉えることは可能であるが，勾配情報から得られる特徴量であるため，テクスチャの少ない距離画像からは有効な動き情報を抽出することが困難であると考えられる．

3.　提案手法

商品を取る動作の識別には，商品棚に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を識別する必要がある．商品を取る動作の検出には人の体や手の動きを捉えることが有効であり，高さの識別には人の体や手の高さを捉える必要がある．そのため手の検出や追跡が必要に

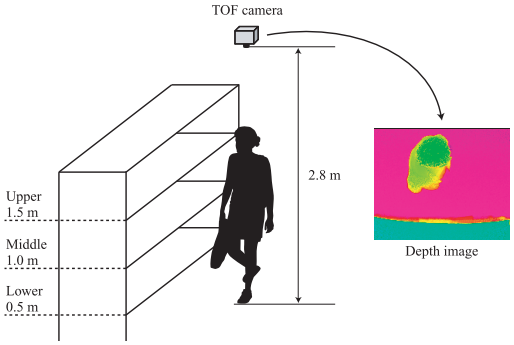


図 3　実験環境
Fig. 3.　Experimental circumstance.

なるが，商品を持つことによる手の形状変化などにより困難となることが考えられる．

そこで，提案手法では図 4 に示すように，TOF カメラにより得られるデプス動画像から，動きを捉えるための時空間特徴である PSA 特徴量と，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．以下に，提案手法における各処理の詳細について述べる．

〈3・1〉　TOF カメラから得られるデプス動画像　本研究での距離情報の取得には TOF(Time of Flight) カメラを用いる．TOF カメラは，カメラの周囲に付いた LED より照射される赤外光が対象物に反射し，カメラで観測されるまでの時間を計測することにより，物体までの距離を計測するカメラである．本研究では，TOF カメラとして MESA 社の SR-4000 を用いる．SR-4000 は，0.8m〜5.0m(絶対距離精度：±1%) までの距離情報をリアルタイムで取得することができる．

〈3・2〉　デプス動画像の時空間情報　提案手法では，デプス動画像の時空間情報としてピクセル状態分析 (Pixel State Analysis：PSA)[9] を用いる．ピクセル状態分析とは，図 5 に示すように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．

提案手法では，人領域から特徴量を抽出するために，ピクセル状態分析を用いてデプス動画像からの人領域検出を行う．ピクセル状態分析は，ピクセルの状態を判別する手

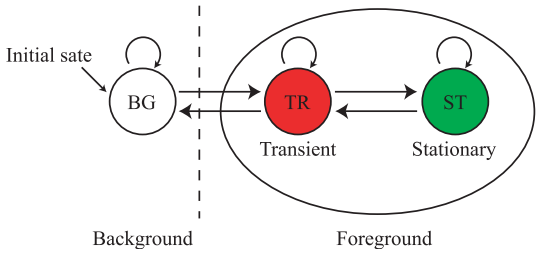


図 5　ピクセル状態の遷移図
Fig. 5.　Diagram of state transition for a pixel.

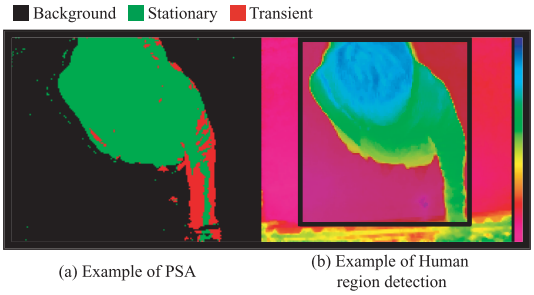


図 6　ピクセル状態分析による人領域検出
Fig. 6.　Human region extraction using PSA.

法であるため，前景と背景を抽出することができる．そこで，ピクセル状態分析により前景（動状態，静状態）と判別された領域を人領域として検出する．図 6 にピクセル状態分析による人領域の検出例を示す．

次に，検出した人領域から PSA 特徴量を抽出する．図 7 に示すように，ピクセル状態分析画像の人領域を 64 × 64[pixel] にリサイズし，8 × 8[pixel] のセル領域に分割する．分割された各セル領域から，ピクセル状態分析結果より PSA ヒストグラムを算出する．これをすべてのセル領域から算出することにより PSA 特徴量とする．PSA 特徴量は，局所的なピクセルの状態をヒストグラム化した特徴量であり，動作識別において人の部分的な動きを表現する特徴量となる．

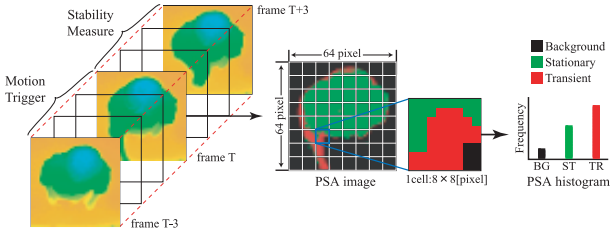


図 7　時空間特徴量 (PSA 特徴)
Fig. 7.　Spatiotemporal feature (PSA features).

〈3・3〉　距離情報　本研究では，図 3 に示すように，TOF カメラを用いて，人を上部から撮影するため，高さを捉えることが可能となる．高さを捉える特徴量として，距離ヒストグラムのピーク値を用いる．図 8 に示すように，検出した人領域を 64 × 64[pixel] のサイズにリサイズする．リサイズした画像を 8 × 8[pixel] のセル領域に分割し，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．提案手法では 0m〜2.8m までを 0.2m 間隔で量子化するため，14 個のビンからなる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．距離ヒストグラムのピーク値は，矩形領域内の出現頻度の高い距離を特徴量としているため，カメラからの絶対的な距離を捉える特徴量である．

〈3・4〉　マルチクラス識別器の構築　提案手法では，上段，中段，下段，その他の複数の動作識別を行うために Joint

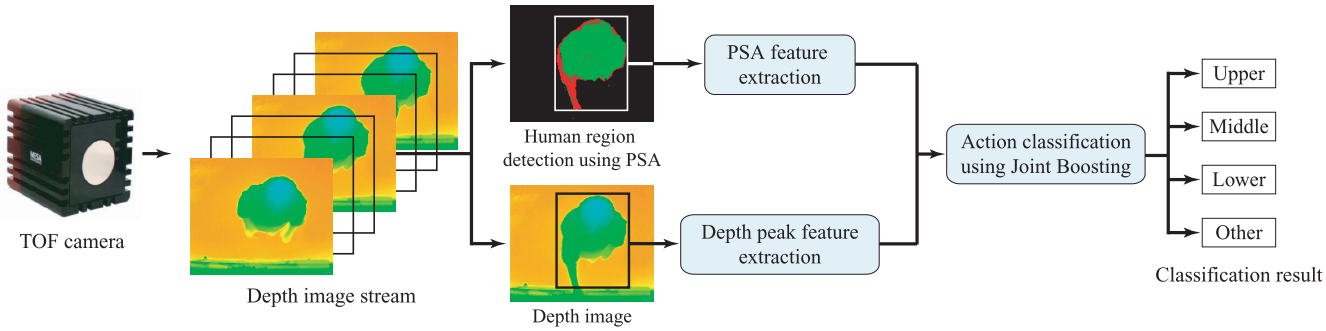


図 4　時空間情報と距離情報を用いた動作識別の流れ

Fig. 4.　Flow of action classification using Spatiotemporal and depth information.

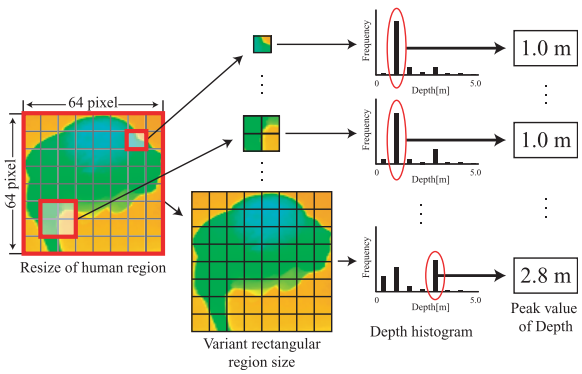


図 8　距離のピーク値による特徴量

Fig. 8.　Depth peak features.

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 [10] 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 $H(v,c)$ は次式となる．

$$H(v,c) = \sum_{m=1}^{M} h_m(v,c) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (6)$$

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集合を $G^{S(n)}(v)$ とすると，以下のように表される．

$$G^{S(n)}(v) = \sum_{m=1}^{M} h_m^n(v) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (7)$$

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．

$$\begin{aligned} H(v,1) &= G^{1,2,3}(v) + G^{1,2}(v) + G^{1,3}(v) + G^{1}(v) \\ H(v,2) &= G^{1,2,3}(v) + G^{1,2}(v) + G^{2,3}(v) + G^{2}(v) \quad (8) \end{aligned}$$

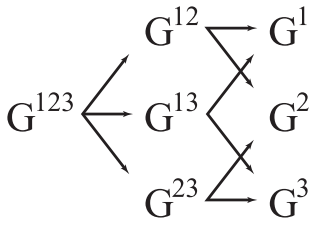


図 9　弱識別器集合の共有

Fig. 9.　Share of weak classifires.

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

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 は図 9 に示すように，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．しかし，識別の際は強識別器間で弱識別器集合を共有していることにより，7 つの弱識別器集合を算出するだけで良いため，高速な識別が可能となる．

4.　識別実験による提案手法の評価

提案手法の有効性を確認するため，評価実験を行う．

〈4・1〉　データベース　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，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．学習サンプルにおいて，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におい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．

〈4・2〉　商品を取る動作の識別実験　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，PSA 特徴，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

表 1　商品を取る動作の識別率 [%]
Table 1.　Classification rate of action classification [%].

	Upper	Middle	Lower	Other	All
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

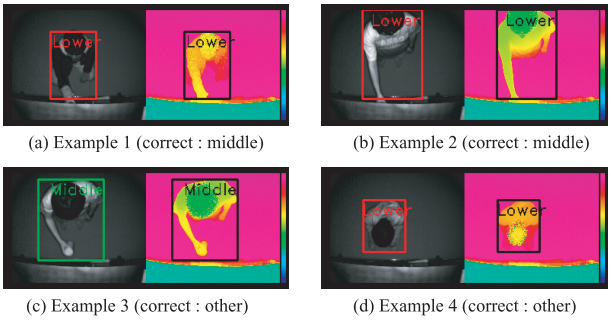
特徴＋距離情報についても比較する．

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，各特徴量に距離情報を付加することで，識別精度を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，従来法である 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 特徴量が 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．

〈4･3〉 商品を取る動作の識別例　図 10 に動作の識別例を示す．識別例では，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．

図 11 に誤識別例，表 2 に PSA 特徴＋距離情報によるコンフュージョンマトリクスを示す．表 2 より「中段」を「下段」と誤識別する場合が多いことがわかる．図 11(a)，(b) は「中段」を「下段」と誤識別した例である．(a) では，しゃがんでいる状態で中段の商品を取るときに「下段」と誤識別している．これは，下段の識別に人の体の高さを捉えているためと考えられる．(b) では，中段から商品を取った際,「下段」と誤識別している．これは，下段から商品を取る動作と類似しているためと考えられる．

これ以外の誤識別例として，(c) では，下段から商品を取った後の動作で「中段」と誤識別されている．これは，下段から商品を取ったあとに，中段から商品を取った場合と類似した高さ，形状になったためと考えられる．(d) では，しゃがんだ際に商品棚に手を伸ばしていないが「下段」と識別されている．これは，しゃがんだ際に前かがみになっているため，手を出した状態と類似したと考えられる．このような誤識別は動きと高さの共起を捉えることで減少させることが可能であると考えられる．



(a) Example 1 (correct : middle)　(b) Example 2 (correct : middle)
(c) Example 3 (correct : other)　(d) Example 4 (correct : other)

図 11　誤識別例
Fig. 11.　Examples of missed classification.

表 2　PSA 特徴と距離情報によるコンフュージョンマトリクス
Table 2.　Confusion matrix by PSA Feature and Depth information.

		out			
		Upper	Middle	Lower	Other
in	Upper	712	5	30	58
	Middle	0	595	46	64
	Lower	0	1	616	125
	Other	34	48	119	4331

〈4･4〉 学習により選択された特徴量　図 12 に，Joint Boosting により選択された特徴量を示す．G^{123}，G^{12}，G^{13}，G^{23} は複数の動作間で共有される弱識別器集合であり，商品を取る動作が共通するため，PSA 特徴は手の動きをよく捉え，体が静状態であることを捉えている．また，距離情報は人の体の高さを捉えている．それに対し，G^1，G^2，G^3 は上段，中段，下段について個々の識別を行う弱識別器集合であるため，PSA 特徴は人の体の動きを捉え，距離情報は人の手の領域の高さをよく捉えている．G^3 においては，下段から商品を取る動作の識別を行うため，しゃがむ動作に対する体の動きを捉えている．

5.　ま と め

本稿では TOF カメラから得られる距離情報を用いた Joint Boosting による動作識別手法を提案した．本研究では，スーパーやコンビニにおいて，来店客が商品棚の上段，中段，下段から商品を取る動作の識別を対象とした．評価実験の結果，PSA 特徴量＋距離情報による識別が 93.2%であり，従来法である CHLAC 特徴＋距離情報と比較し 3.1%，ST-patch 特徴＋距離情報と比較し 2.8%識別率を向上させることができた．今後は，PSA 特徴と距離情報の共起による識別の高精度化を行う予定である．

文　献

（1） T. Kobayashi and N. Otsu. Action and Simultaneous Multiple-Person identification Using Cubic Higher-Order Local Auto-Correlation. *In Proc. International Conference on Pattern Recognition*, pp. 741–744, 2004.
（2） 南里卓也, 大津展之．複数人動画像からの異常動作検出.

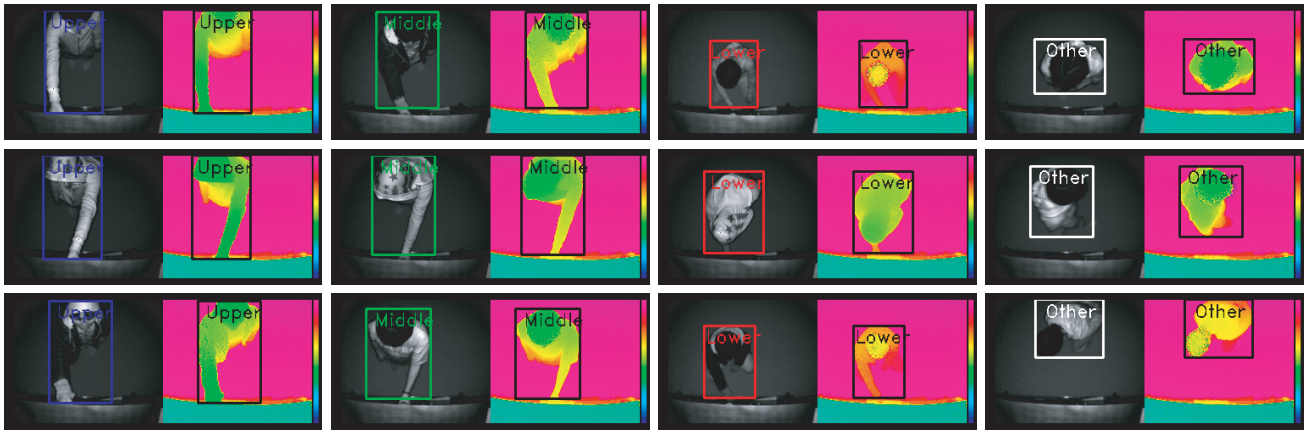


図 10　動作の識別例

Fig. 10.　Examples of action classification.

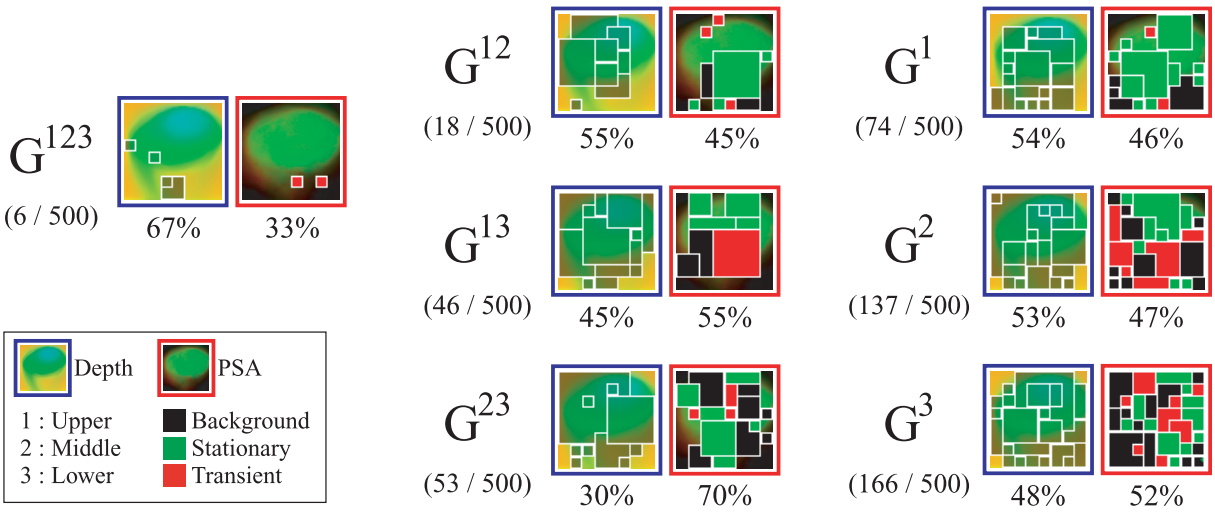


図 12　学習により選択された特徴量

Fig. 12.　Selected features by learning.

情報処理学会論文誌. コンピュータビジョンとイメージメディア, Vol. 45, No. SIG_15, pp. 43–50, 2005.

(3) E. Shechtman and M. Irani. Space-Time Behavior Based Correlation. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 405–412, 2005.

(4) Y. Ke, R. Sukthankar and M. Hebert. Event Detection in Crowded Videos. *In Proc. IEEE International Conference on Computer Vision*, pp. 8–15, 2007.

(5) Y. Murai and H. Fujiyoshi and M. Kazui. Incident Detection based on Dynamic Background Modeling and Statistical Learning using Spatio-temporal Features. *Machine Vision Applications*, pp. 156–161, 2009.

(6) T. Kurita, N. Otsu and T. Sato. A Face Recognition Method Using Higher Order Local Auto-correlation and Multivariate Analysis. *In Proc. International Conference on Pattern Recognition*, pp. 213–216, 1992.

(7) M. Kazui, M. Miyoshi, S. Muramatsu and H. Fujiyoshi. Incoherent Motion Detection using a Time-series Gram Matrix Feature. *In Proc. International Conference on Pattern Recognition*, pp. 1–5, 2008.

(8) Y. Murai, H. Fujiyoshi and T. Kanade. Combined Object Detection and Segmentation by Using Space-Time Patches. *In Proc. 8th Asian Conference on Computer Vision*, Vol. Part I, No. LNCS 4843, pp. 915–924, 2007.

(9) H. Fujiyoshi and T. Kanade. Layered Detection for Multiple Overlapping Objects. *IEICE Transactions on Information and systems*, Vol. E87-D, pp. 2821–2827, 2004.

(10) A. Torralba, K. P. Murphy and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 762–769, 2004.