Vol.2013-CVIM-186 No.26
2013/3/15

統計的学習手法による物体検出の高精度化と効率化 -人検出の実用化に向けて-

藤吉 弘亘[1,a)]　山内 悠嗣[1,b)]　土屋 成光[1,c)]　山下 隆義[2,d)]

概要：本稿では，統計的学習手法を用いた人検出の高精度化と学習における効率化について，実用化の観点から述べる．人検出性能の高精度化として，画像局所特徴量を組み合わせて物体検出に有効な Joint 特徴量を自動生成する学習法と人検出への適用例について紹介する．統計的学習法に基づく人検出では，学習サンプル収集に伴う人的コストと特定シーンに合わせた再学習のための時間的コストが大きな問題である．そこで，学習時の効率化として，特定シーンにおける学習サンプルの収集コストを低減する人体シルエットの生成と MILBoost による学習法と，学習時間の短縮を目的としたハイブリッド型転移学習法について紹介する．

1. はじめに

人々の生活の利便性向上や安心・安全な社会の実現には，人を観る画像認識技術 [1], [2] が重要である．特に，人検出は映像中から人の位置を特定する技術であり，追跡や動作認識を実現するための前処理として必要不可欠である．人検出は画像局所特徴量と統計的学習手法の組み合わせ [3] により実現されている[*1]．画像局所特徴量には，局所領域における勾配強度を方向毎に累積した Histograms of Oriented Gradients(HOG) 特徴量 [4] が用いられ，統計的学習手法には Support Vector Machine(SVM) や AdaBoost などが用いられている．事前に統計的学習手法により多くの学習サンプルを用いて識別器を構築し，検出時は未知の入力画像をラスタスキャンしながらウィンドウを識別することで，人検出を実現している．

このような統計的学習手法を用いた場合，公開されている画像データベースによる評価が一般的であるが，必ずしも評価用データベースで高い性能を獲得した手法が実用化に適しているとは限らない．これは，統計的学習手法を用いた手法の検出性能は学習サンプルに強い依存性があり，学習サンプルと異なる環境において再学習を必要とする場合，どのように大量の学習サンプルを集めるのかという問題があるからである．従って，実用化という観点から最適な手法は，以下の 3 つの条件を満たすことが重要である．

(1) 検出失敗の理由を明確に把握することが可能
(2) 少ない学習サンプルでシステムをチューニング可能
(3) 省メモリで高速な計算アルゴリズム

そこで，本稿では実用化の観点から，統計的学習手法を用いた人検出の高精度化と学習における効率化について述べる．人検出性能の高精度化として，2 章では画像局所特徴量を組み合わせて物体検出に有効な Joint 特徴量を自動生成する学習法と人検出への適用例について紹介する．3 章では，Joint 特徴量を用いた人検出法の FPGA によるハードウェア化について紹介する．統計的学習法に基づく人検出では，学習サンプル収集に伴う人的コストと特定シーンに合わせた再学習のための時間的コストが大きな問題である．そこで，学習時の効率化として，4 章では特定シーンにおける学習サンプルの収集コストを低減する人体シルエットの生成と MILBoost による学習法を，5 章では学習時間の短縮を目的としたハイブリッド型転移学習法について紹介する．

2. Joint 特徴量による人検出の高精度化

画像局所特徴量として HOG 特徴量，統計的学習手法として AdaBoost を用いて構築した識別器は，図 1(a) に示すように，局所領域における人の勾配特徴を捉えた複数の弱

[1] 中部大学工学部情報工学科
Chubu University, 1200 Matsumoto-cho, Kasugai, Aichi 487-8501, JAPAN.
[2] オムロン株式会社
OMRON Corporation, 2-2-1 Nishikusatsu, Kusatsu, Shiga 525-0035, JAPAN.
a) hf@cs.chubu.ac.jp
b) yuu@vision.cs.chubu.ac.jp
c) mtdoll@vision.cs.chubu.ac.jp
d) takayosi@omm.ncl.omron.co.jp

[*1] 統計的学習手法を人検出のサーベイについては，文献 [3] を参照されたい．

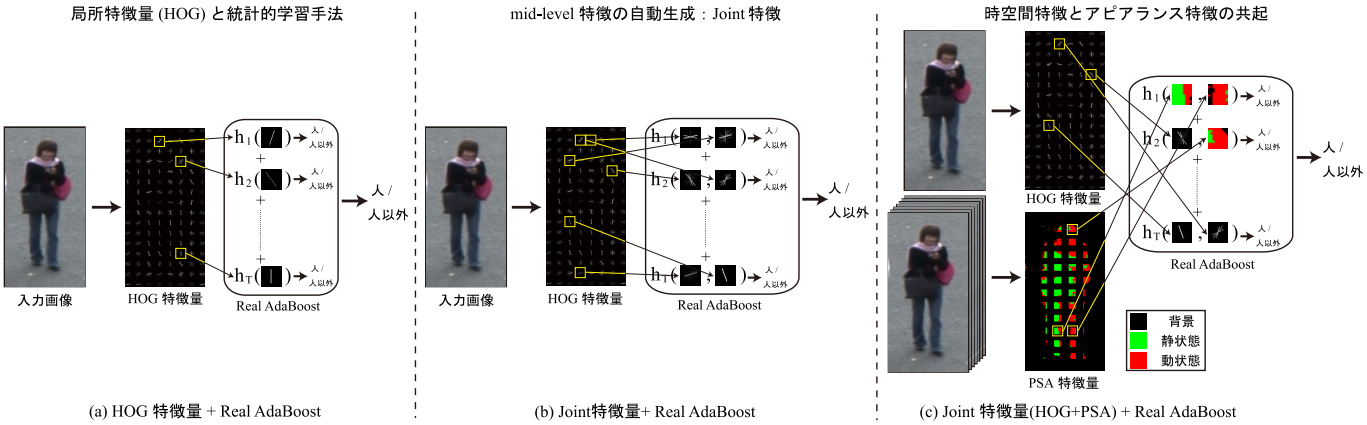

図 1 人検出における特徴量の捉え方.

識別器から構成される．強識別器は，複数の弱識別器の結果を組み合わせて最終的な識別結果を出力する．

人の形状には，大きく分けて下記に示す 2 つの特徴があると考えられ，より高精度な人検出を実現するためには，これらを捉えるような特徴量を統計的学習手法により設計する必要がある．

(1) 頭から肩にかけての Ω に似た形状や上半身から下半身にかけての連続的な形状

(2) 頭や肩，胴，足などの左右対称的な形状

(1) に対しては，局所領域内の 4 方向のエッジ特徴を AdaBoost により組み合わせる Shapelet 特徴量 [5] が提案されている．(2) に対しては，AdaBoost の弱識別器が複数の特徴量を同時に観測することにより，共起性を表現する Joint Haar-like 特徴量 [6] が提案されている．両手法は，複数の low-level な特徴量をブースティングにより組み合わせることで特徴量間の関連性を捉えることができ，高精度な検出を実現している．

我々は，2 段階の Real AdaBoost を用いて図 1(b) に示すような物体形状の対称性や連続性を自動的に捉える Joint 特徴量による物体検出法を提案している．本章では，Joint 特徴量 [7] による人検出法とその効果について述べる．

2.1 Joint 特徴量と 2 段階ブースティング

Joint 特徴量の生成と最終識別器の構築の流れを図 2 に示す．Joint 特徴量は，2 段階の Real AdaBoost による学習により生成される．

2.1.1 1 段階目の Real AdaBoost による Joint 特徴量の生成

Joint 特徴量を生成するために，2 つの異なる局所領域 (セル) から Low-level な特徴量として HOG 特徴量を求め，共起表現法 [8] により異なるセルの HOG 特徴量間の共起を表現する．まず，入力画像から HOG 特徴量 $\mathbf{v}$ を算出する．以降では，low-level な特徴量として HOG 特徴量を用いることを前提として述べるが，より高い表現能力を持つ

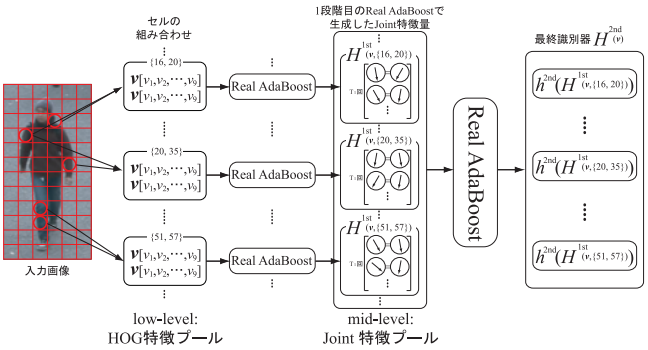

図 2 Joint 特徴量による 2 段階 Real AdaBoost.

多重解像度の HOG 特徴量も利用できる [9]．次に，2 つのセル {m,n} の HOG 特徴量から共起確率特徴量を生成する．その際に，共起確率特徴量は HOG 特徴量の全ての組み合わせに対して求め，Real AdaBoost により最も良い組み合わせを弱識別器として自動的に選択する．この処理を T 回繰り返し，1 段階目の Real AdaBoost により次式で表される 2 つのセル {m,n} の Joint 特徴量である強識別器 $H^{1\mathrm{st}}(\mathbf{v},\{m,n\})$ を学習する．

$$H^{1\mathrm{st}}(\mathbf{v},\{m,n\}) = \sum_{t=1}^{T} h_t^{1\mathrm{st}}(\mathbf{v},\{m,n\}) \qquad (1)$$

上記の処理を全てのセルの組み合わせに対して行い，組み合わせ数と同数の Joint 特徴量を生成する．例えば，入力画像が 30 × 60 ピクセル，セルサイズを 5 × 5 ピクセルとした場合，72 個のセルに分割され，組み合わせ数は ${}_{72}C_2 = 2,556$ となるため，2,556 個の Joint 特徴量 $H^{1\mathrm{st}}()$ を生成する．生成した全ての Joint 特徴量を特徴プール $\mathcal{F}$ とし，後述する 2 段階目の Real AdaBoost の入力とする．

2.1.2 2 段階目の Real AdaBoost による最終識別器の構築

2 段階目の Real AdaBoost では，1 段階目の Real AdaBoost により生成した Joint 特徴量 $H^{1\mathrm{st}}()$ のプール $\mathcal{F}$ を入力として次式で示す強識別器 $H^{2\mathrm{nd}}()$ を構築する．

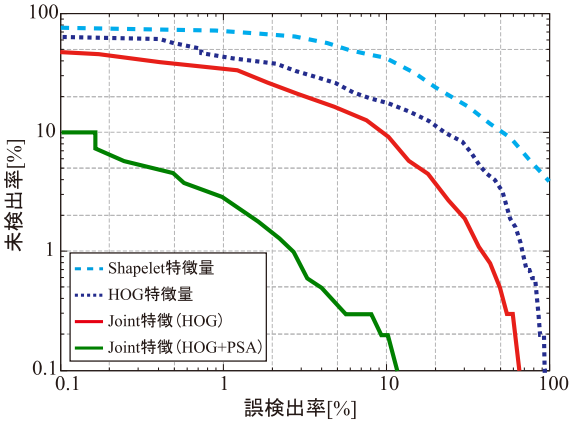

図 3 DET カーブ.

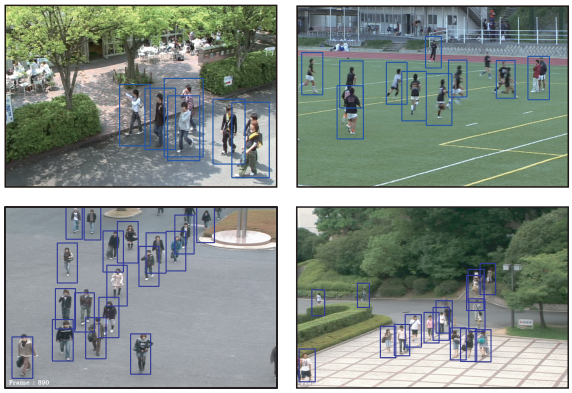

図 4 Joint 特徴量による物体検出例.

$$H^{2\mathrm{nd}}(\mathbf{v}) = \sum_{\{m,n\}\in\mathcal{F}} h^{2\mathrm{nd}}\left(H_t^{1\mathrm{st}}(\mathbf{v},\{m,n\})\right) \quad (2)$$

$$= \sum_{\{m,n\}\in\mathcal{F}} h^{2\mathrm{nd}}\left(\sum_{t=1}^{T} h_t^{1\mathrm{st}}(\mathbf{v},\{m,n\})\right) \quad (3)$$

2 段階目の弱識別器 $h^{2\mathrm{nd}}()$ は，2 つのセルの関係を捉えた 1 段階目の $H^{1\mathrm{st}}()$ の出力となっていることから，異なるセルの low-level の特徴量から mid-level の特徴量を生成していることになる．これにより，識別に有効な Joint 特徴量を自動的に選択することが可能となる．

2.2 Joint 特徴量の効果

2.2.1 検出性能

Joint 特徴量による評価実験結果を図 3 に示す．提案する Joint 特徴量は，従来法である HOG 特徴量 [4] や Shapelet 特徴量 [5] と比較して高い検出性能であることから，位置の異なる 2 つのセル内の HOG 特徴量を組み合わせることの有効性を確認した．図 4 に Joint 特徴量による人検出例を示す．部分的なオクルージョンに対して頑健な検出が可能であることがわかる．

2.2.2 Joint 特徴量の効果

図 5(a) に 1 段階目の Real AdaBoost，図 5(b) に 2 段階目の Real AdaBoost により選択された HOG 特徴量の可視化結果を示す．また，図 5(c) に 2 段階目の Real AdaBoost により選択された学習ラウンド毎の 2 つのセルと Joint 特徴量を示す．また，HOG 特徴量の勾配方向を人は 9 方向で表現しており，輝度が高いほど Real AdaBoost における弱識別器の評価値が高く，識別に有効な特徴量であることを表す．

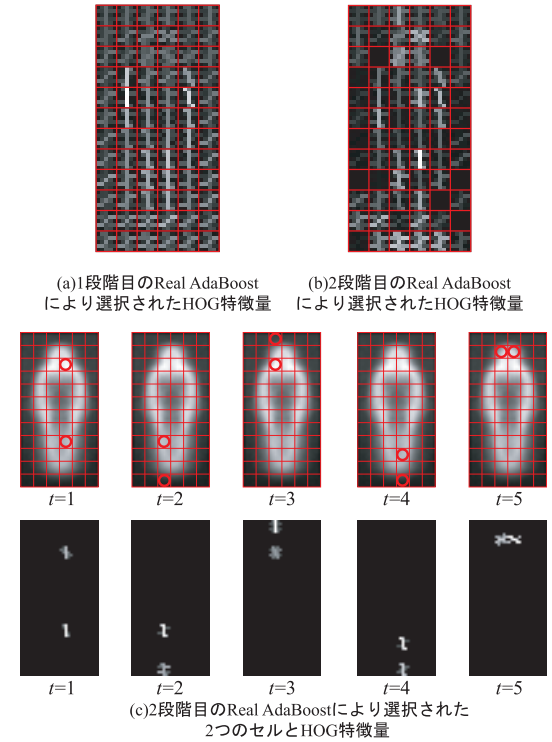

図 5 選択された Joint 特徴量の可視化.

図 5(b) では，図 5(a) で選択された HOG 特徴量であっても人の輪郭以外は選択されにくい傾向がある．これは，2 段階目の Real AdaBoost の特徴選択において，識別に有効ではないと判断されたためである．次に図 5(c) に注目する．2 段階目の Real AdaBoost により選択された Joint 特徴量は，人の輪郭に沿ったセルが選択されていることがわかる．

HOG 特徴量と Real AdaBoost では，図 1(a) に示すように 1 個の弱識別器が 1 個の HOG 特徴量を用いて識別するのに対し，Joint 特徴では，図 1(b) に示すように 1 個の弱識別器が位置の異なる 2 つの領域内に含まれる複数の HOG 特徴量を用いて識別を行う．これにより，従来の単一の HOG 特徴量のみでは捉えることができない物体形状の対称性や連続的なエッジを自動的に捉えることができるため，高精度な人検出が可能となる．

2.2.3 解析の容易さ

図 6 は，未検出画像と誤検出画像に対する各弱識別器の応答を示したものである．図 6(a) の未検出例では，人の特徴は頭部や体の右側面のエッジが重要であるが，入力画像では抽出されなかったため，その弱識別器の出力はマイナス方向に大きく出力され未検出と判定されたことがわか

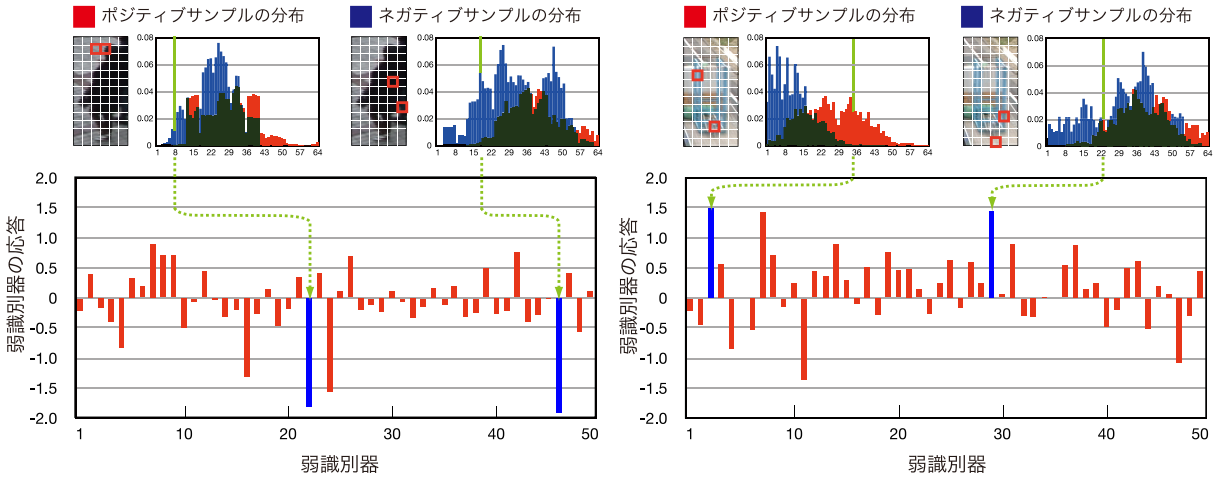

図 6 未検出と誤識別したサンプルの弱識別器の出力と確率密度関数.

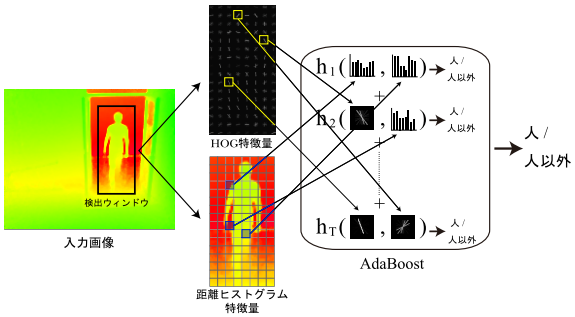

図 7 HOG 特徴量と距離ヒストグラム特徴量の共起.

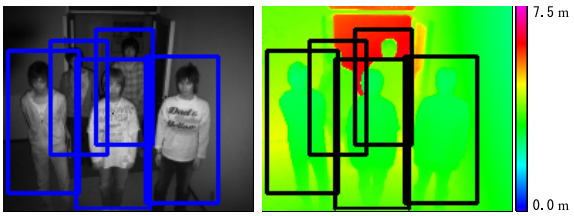

図 8 TOF カメラによる人検出例.

る．実際に該当する弱識別器における学習サンプルの分布を見ると，ネガティブサンプルの頻度が高い bin であることがわかる．また，図 6(b) の誤検出例では，逆に人らしいエッジが観測されたために誤検出となったことがわかる．以上より，Joint 特徴量はどのような局所領域が原因でどのように検出失敗するかという理由の解析が容易であり，学習サンプル追加，局所特徴量の追加などを検討する際に有用であるといえる．

2.3 Joint 特徴量の応用

Joint 特徴量のフレームワークでは，人のアピアランスを表す HOG 特徴量に他の特徴量を追加することが可能である．ここでは，HOG 特徴量に時空間特徴とデプス情報を用いた Joint 特徴量による高精度化について述べる．

2.3.1 時空間特徴量との共起

動体検出に用いられてきた時空間特徴に基づく特徴量として，ピクセル状態分析 (PSA) の結果を加えることにより，さらに高精度な人検出を実現した [10]．ピクセル状態分析とは，ピクセルの状態の輝度の時間変化から背景，動状態，静状態に判定する手法である．これにより，人独特な歩行動作を捉えることができる [11]．アピアランス特徴のみでは人に似た物体を誤検出しているが，時空間特徴量を加えることにより，誤検出を抑制することができる．図 3 の DET カーブの結果から，HOG 特徴のみを用いた Joint 特徴量と比べて，HOG と PSA を用いた Joint 特徴量は誤検出率 1.0%において約 30%検出率を向上することができた．

2.3.2 デプス情報との共起

可視光カメラにより取得した画像から人の検出を行う場合，背景のテクスチャの複雑さによって，人検出に有効なアピアランス情報を取得することが困難となる場合がある．そこで我々は，カメラから物体までのデプス情報を取得できる Time of Flight カメラ (TOF カメラ) を用いた人検出 [12] を提案している．TOF カメラから得られるデプス情報と，アピアランス特徴を同時に捉えることにより高精度な人検出を実現する．

デプス情報から人検出に有効な特徴量を抽出するために，2 つの局所領域間の距離分布の類似度から得られる距離ヒストグラム特徴量を抽出する．類似度には Bhattacharyya 係数を用いる．

図 7 に，アピアランス情報である HOG 特徴量と距離ヒストグラムから得られる特徴量の共起による人検出の流れを示す．本手法は，図 8 に示すように，人の重なりがある場合でも高精度な検出が可能となる．デプス情報を加えることにより，弱識別器は人体と背景の距離関係を捉えることが可能となり，オクルージョンや背景の複雑さの影響を

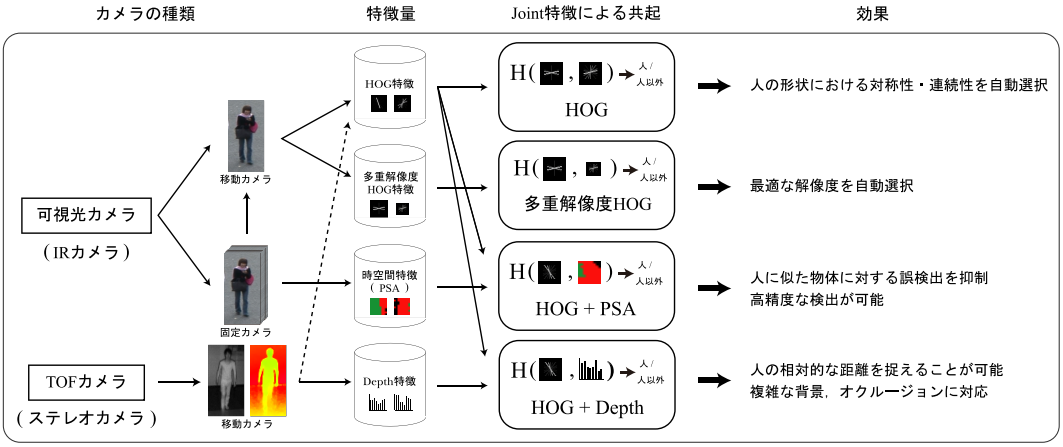

図 9 Joint 特徴量による共起特徴表現.

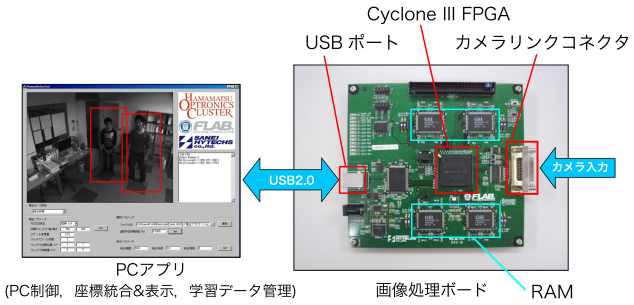

図 10 画像処理 FPGA ボード.

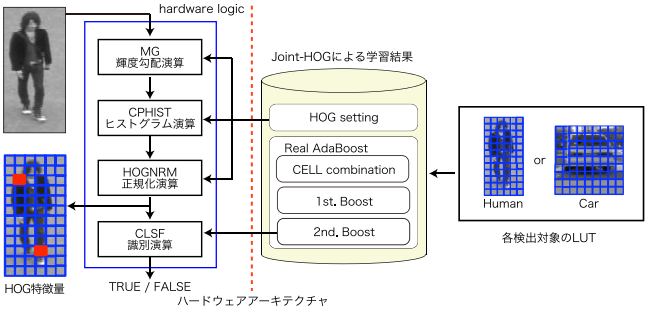

図 11 ハードウェアアーキテクチャ.

抑制することができる.

2.4 Joint 特徴量の課題

本章では，図 9 に示すような複数の特徴量間の共起を用いた Joint 特徴量による物体検出法について述べた．単一の HOG 特徴量では識別困難なパターンに対して，Joint 特徴量は位置の異なる 2 つのセル内の HOG 特徴量を組み合わせることにより，識別困難なパターンを正しく識別することができる．さらに高精度な検出を実現するために，時空間特徴や距離ヒストグラム特徴量との共起の効果について述べた．Joint 特徴量は，高精度な人検出が可能である一方，実用化という観点から考えた時に省メモリ化が課題となる．そのためには，実数で表現される HOG 特徴量を 2 値化することで大幅にメモリ使用量を削減する特徴量のバイナリコード化 [13], [14] が有効である.

3. FPGA による人検出器のハードウェア化

人検出技術のアプリケーションの一つとして，自動車の安全運転支援を目的とした車載カメラでの利用が挙げられる．車載での利用では人検出技術をハードウェア化する必要がある．2011 年には，東芝から Co-HOG[15] による人検出技術を搭載した車載用画像認識プロセッサ LSI が発売されている [16]．本章では Joint 特徴量を用いた人検出器の FPGA(Field Programmable Gate Array) によるハードウェア化 [17] について述べる．我々は人検出器をハードウェア化する上で重要となる (1) Joint 特徴量の高速化，(2) 検出対象の柔軟性の 2 点を考慮して，ハードウェアアーキテクチャを設計した.

3.1 検出対象の柔軟性

利用する環境下での検出対象を想定して学習した結果をハードウェア化することになるが，利用環境が異なる場合や検出対象が変更となると，再度ハードウェアを設計する必要がある．そこで，我々は図 10 に示すように，学習ソフトウェア API と連携して同一ハードウェア上で検出対象を変更可能な Joint-HOG による FPGA システム (Joint-HOG FPGA) を実現した．Joint-HOG FPGA は，人を対象とした場合は縦横 12 × 6 セル，車両を対象とした場合は縦横 9 × 9 セルと検出ウィンドウのサイズや縦横比が異なっても，セルを同一サイズにしておくことで，検出対象を変更可能となるハードウェアアーキテクチャとなっている．事前に学習ソフトウェア API で学習した結果 (各弱識別器の入力に対する応答) を Look Up Table(LUT) で識別器を構成する．これにより，LUT の内容を書き換えることで，同一 FPGA 上で瞬時に検出対象を柔軟に変更することが可能となる.

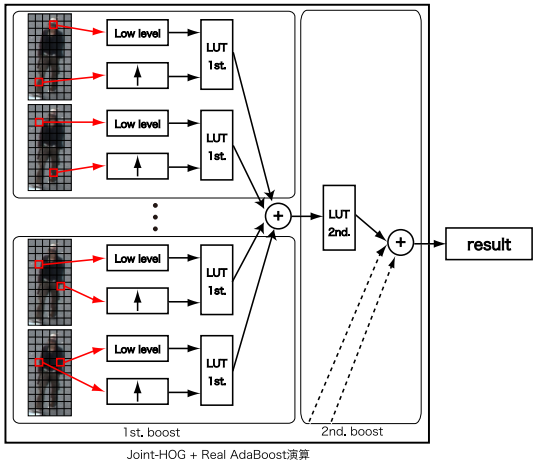

図 12 Joint-HOG 識別器.

3.2 ハードウェアアーキテクチャ

図 11 に Joint-HOG FPGA のハードウェアアーキテクチャを示す．MG/CPHIST/HOGNRM/CLSF の 4 段階の演算ステージによって入力画像から Joint HOG 特徴量を算出し識別演算を行う．各ブロックでは専用の小規模メモリを複数活用することで，演算サイクルの最適化を図った．これにより，演算待ち時間を極力減らすことができた．

3.2.1 HOG 特徴量の演算

グレースケール画像を入力として，MG/CPHIST/HOGNRM の各モジュールを経ることで HOG 特徴量を演算する．MG モジュールでは，画像ウィンドウの輝度情報から一次微分を行い，輝度勾配ベクトルを算出する．1 ピクセルあたりの輝度勾配ベクトルを算出するために，上下左右の 4 ピクセルを使用するが，1 ピクセル毎にデータを扱うことをせずに，上下の 3 ライン分を FIFO にて管理し，シフトレジスタと組み合わせてデータ処理する．これにより，ほぼ 1 ウィンドウ分のデータ読み出しサイクル数で輝度勾配ベクトルを算出できる．

次に，CPHIST モジュールにて，算出した勾配強度と勾配方向からセル毎に輝度の勾配方向ヒストグラムを作成する．勾配方向毎の配列メモリに勾配強度を累積することで算出する．これもほぼ 1 ウィンドウ分のデータ読み出しサイクル数にて，セル数 × 方向数の次元量のデータとして出力する．最後に，HOGNRM モジュールでブロック領域毎に正規化を行う．二乗演算，ブロック毎の累積演算，平方根演算，除算の各処理を経て正規化された HOG 特徴量を算出する．

3.2.2 Joint-HOG による識別演算

図 12 に CLSF モジュールでの識別の演算構成を示す．HOGNRM モジュールで算出した HOG 特徴量から検出に最適なセルの組合わせを使用し，2 段階 Real AdaBoost 処理から TRUE/FALSE を判定する．事前学習結果は LUT 化されており，サイズは約 360kbit である．これにより，事前学習にて選択した検出対象毎に最適な特徴量の組み合わせでの識別演算が可能となる．また，LUT 化することで，学習結果の差し替えが容易となり，検出対象を人/車両と変更可能となる．図 13 に識別結果出力までの流れを示す．MG-CLSF のそれぞれのモジュールは，ほぼ同じ演算時間で処理するように設計できるため，パイプライン状に構成することで後段モジュールの出力を待たずに演算可能である．

MG	q0	q1	q2	q3	q4	···
CPHIST	↳	q0	q1	q2	q3	···
HOGNRM		↳	q0	q1	q2	···
CLSF			↳	q0	q1	···

図 13 パイプライン処理.

表 1 実装結果 (Cyclone III EP3C120).

総 LE 数	17,419(15%)
総レジスタ数	11,306(9%)
内部メモリ bit 数	1,046,647bit(26%)
動作周波数	70MHz
処理時間 (1 画像分)	93.95ms(約 10fps)
処理時間 (1 画像分)	46.98ms(約 20fps)

3.3 Joint-HOG FPGA システム

Joint-HOG FPGA システムでは，図 10 に示すように，FPGA ボードとして Altera 社製 CycloneIII FPGA を使用し，カメラリンク入力された画像に対して検出する．このとき，1 枚の画像に対して検出ウィンドウ毎に HOG 特徴量の計算及び識別を行い，その結果を画像データと共に USB 経由で PC に転送する．PC で受信した識別結果から，Mean Shift クラスタリングにより検出ウィンドウの統合処理を行い，最終的な検出結果を表示する．また，学習データの LUT は USB 経由にて PC から書き換え可能である．

また，Cyclone III FPGA ボードをターゲットとして合成した結果 (検出結果 1 個分) と 1 枚の画像 (2,940 検出ウィンドウ) の処理時間を表 1 に示す．1 検出ウィンドウ分の検出器を比較的小規模な回路として実装でき，約 10fps と実用的な時間内で実行できることを確認した．さらに，2 並列とした場合には約 20fps での検出が可能であることを確認した．

4. CG による学習サンプルの生成と MILBoost による学習の効率化

本章では，学習における効率化として，学習サンプルの重要性について述べた後，3 次元人体モデルを用いた学習サンプルの自動生成と MILBoost による生成型学習法 [18] について述べる．

4.1 学習サンプルの重要性

統計的学習手法を用いて検出性能の高い識別器を構築す

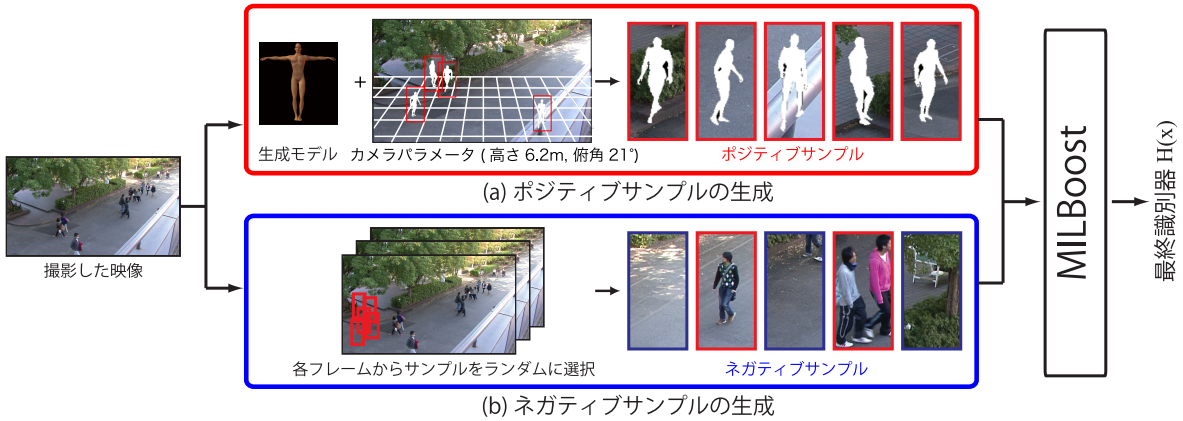

図 14 MILBoost による学習の流れ.

るには，人らしさを捉える特徴量をどのように設計するかという観点だけでなく，学習サンプルの質も重要である．図 15 は，INRIA Person Dataset[4] の全人画像から平均勾配画像を作成したものである．図 15 より，頭部から肩，下半身へのシルエットが人画像に共通した特徴であることがわかる．これは INRIA Person Dataset には人の位置ずれがないことと，ポジティブサンプルの背景領域には共通性がないことを示している．また，ラベルの誤った学習サンプルは含まれてない．もし誤ラベルを持つ画像が学習サンプルに含まれていると，AdaBoost の学習では，このラベルに適応するように学習されてしまい，性能低下を招くことになる．INRIA Person Dataset のように大量の良質な学習サンプルの収集は大きな手間がかかるとともに，人検出器の性能を左右する重要な要因である．

これらの問題を解決するアプローチとして，少数の学習サンプルからスケール変化や回転，ノイズの付加などの実環境で測定されうる変動を含むように変形させた学習サンプルを生成し，生成したサンプルを用いて識別器を学習する生成型学習法 [19] が提案されている．我々は 3 次元人体モデルを用いた学習サンプルの自動生成と MILBoost による生成型学習法を提案しており，以下にその手法について述べる．

4.2 学習サンプルの自動生成

ポジティブサンプル

提案手法で使用する人体モデルには，形状モデルやモデルの各パーツの階層構造，動作データなどが含まれている．人体の形状モデルは，19 のパーツが存在し，これらのパーツは階層的な構造で表現される．そのため，例えば右肩を動かした場合，右腕や右手が連動して動く．本手法では，19 のパーツに歩行動作のパラメータを与えることで，歩行姿勢として人体モデルを表現する．特定シーンに特化した人体シルエット画像を得るために，実環境に設置したカメラのパラメータ*2を 3 次元人体モデルに入力し，CG により生成した人体シルエット画像を学習用ポジティブサンプルとしてい用いる．CG で生成するため，位置ずれのない大量の人体シルエット画像を生成することができる (図 14(a)).

図 15 INRIA Person Dataset と平均勾配画像.

ネガティブサンプル

ネガティブサンプルは，撮影した映像中からランダムで切り出す (図 14(b)). しかし，ランダムにサンプルを収集した場合，ネガティブサンプルとして人画像が収集される問題がある．この問題を解決するために，誤って付与されたラベルを持つサンプルの混在を考慮した MILBoost により識別器を学習する．

4.3 学習サンプルの混在を考慮した MILBoost による学習

ネガティブサンプルに人画像が混在してしまう問題に対して，MILBoost[21] を用いることにより解決する．MILBoost は Boosting に Multiple Instance Learning (MIL) を導入した学習法である．MIL は，複数のサンプルで構成された Bag に対してラベル付けを行い学習する．これにより，ネガティブサンプルの Bag に人画像が混在する様な場合でも悪影響を受けない学習が可能となる．

4.3.1 Bag と MILBoost

図 16 に示すように，ポジティブクラスの Bag はサンプル一つを Bag として作成する．ネガティブクラスの Bag は，映像中からランダムで画像を切り出すことでサンプ

*2 人検出結果からカメラパラメータを自動推定する手法については文献 [20] を参照されたい.

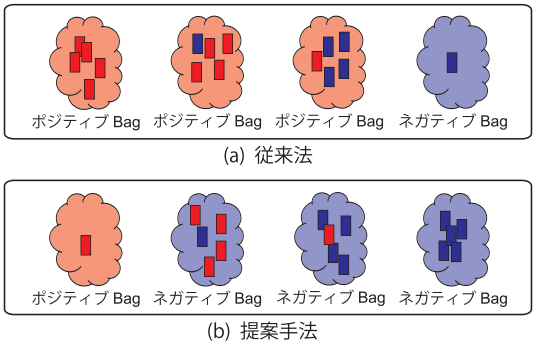

図 16 MIL における従来法と提案手法による Bag の構成.

ルを生成する．そして，切り出したサンプルの集合をネガティブクラスの Bag とする．ネガティブクラスの Bag に，ポジティブサンプルが含まれる場合においても，悪影響を及ぼさないようにする．MILBoost の学習の流れは Real AdaBoost と同様であるが，サンプルの重み w_{ij} の算出方法が異なる．以下に学習の流れを示す．

Step1 ポジティブクラスの Bag に人画像 1 枚を割り当てる．

Step2 ネガティブクラスの Bag に背景からランダムで切り出した画像集合を割り当てる．

Step3 t 個目の弱識別器を学習する．

Step4 式 (4) により j 個目の学習サンプルのクラス尤度 p_{ij} を算出する．

$$p_{ij} = \frac{1}{1+\exp(-H_t(x))} \tag{4}$$

Step5 式 (5) により i 番目の Bag のクラス尤度 p_i を算出する．

$$p_i = \prod_{j \in \mathrm{Bag}_i} p_{ij} \tag{5}$$

Step6 式 (6) により学習サンプルの重み w_{ij} を更新する．

$$w_{ij} = \begin{cases} -p_{ij} & \text{if} \quad y_i = 1 \\ \frac{p_{ij} \times p_i}{1-p_i} & \text{otherwise} \quad y_i = 0 \end{cases} \tag{6}$$

Step3〜**Step6** を繰り返すことにより，最終識別器 $H(x)$ を得る．ネガティブ Bag に含まれているサンプルは，サンプルのクラス尤度 p_{ij} と Bag のクラス尤度 p_i により重みを更新する．誤ラベルの学習サンプルが含まれていた場合，ネガティブ Bag のクラス尤度 p_i が十分に低ければ，そのサンプルはノイズであると捉え，学習サンプルの重み w_{ij} は低下する．このように，Bag 単位での尤度を用いることでノイズの影響を低減することができる．

4.4 自動生成の効果

特定シーンに特化した学習サンプルの自動生成による有効性を評価する．図 17 に示す 4 つのデータベースにより学習した際の識別性能を比較する．DET カーブを図 18 に示す．まず，ネガティブサンプルが同一の Database 1,

図 17 学習用データセットの例.

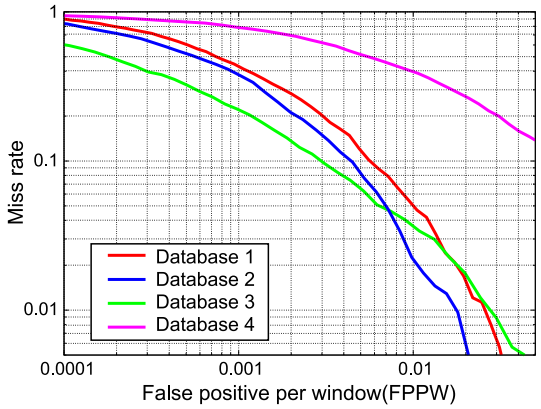

図 18 各学習データベースの実験結果.

Database 2，Database 3 を比較すると，検出性能が最も高いのは人体モデルから生成したサンプルを用いた Database 3 であった．これは，実環境下で撮影した映像に対応した人の見えを CG により生成できたからといえる．実環境下の映像から人手で切り出したサンプルを用いた Database 2 は，自動生成よりも低い結果となった．これは，人画像を人手で大量に切り出す際には，切り出し基準が曖昧になることがあり，これが識別器に悪影響を及ぼしたと考えられる．汎用性のあるデータベースを用いた Database 1 の結果が最も低い検出率となった．これは，学習用データベースの INRIA Person Dataset は実験環境とカメラ位置が異なることが要因である．

次に，Database 3 と Database 4 を比較すると，実環境下で撮影した映像の背景を用いた Database 3 の方が良い結果が得られた．これは，Database 3 では実環境から生成した学習用ネガティブサンプルを用いているため，実環境のシーンに特化した識別器となり検出性能が大きく向上したといえる．

以上より，特定シーンにおいて，3 次元人体モデルから

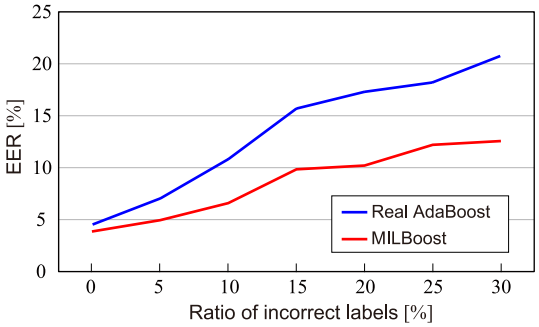

図 19 誤ラベルの割合を変化させた際の性能比較.

生成した学習サンプルを用いることにより，学習サンプルを人手で作成することなく実環境に特化した識別器を構築することができた．さらに，MILBoost を用いることにより，誤ラベルを付与されたサンプルに対して悪影響を受けない学習を実現した．

4.5 誤サンプルの影響

ネガティブの誤サンプルに対応した MILBoost による学習法の有効性を評価する．提案手法と Real AdaBoost を比較する．MILBoost の有効性を確認するために，学習用のネガティブサンプルへ故意に人画像を混在させて識別器を学習する．その際の人画像の割合を 0%〜30% まで変動させ，その際の識別結果を比較する．

実験結果を図 19 に示す．実験結果より，ネガティブサンプル中に人画像の含有率が高くなるに従い，通常の学習手法では Equal Error Rate(EER) が高くなるが，提案手法 (MILBoost) では EER の増大を抑制していることがわかる．人画像の含有率が 15% の場合を比較すると，提案手法は従来法よりも EER が 6.1% 低い．以上より，提案手法はネガティブサンプル中に人画像が含まれていても，識別器の学習に及ぼす悪影響を低減することができた．また，混入率 0%時点においても提案手法の EER がわずかに低いのは，ノイズ低減効果による差であると考えられる．

5. ハイブリッド転移学習による学習の効率化

本章では，学習の効率化として，学習時間の短縮を目的としたハイブリッド型転移学習 [22] について述べる．転移学習を用いることで少量の目標学習サンプルで高精度な識別器の学習が可能となるが，事前学習と目標学習でのシーンが大きく異なると転移が不可能となる．そのため，目標シーンにあわせた再学習が必要となるが，大量のサンプルを用いて学習するには多くの時間を要するという問題がある．そこで我々は，図 20 に示すように転移により得られる特徴量と，再学習により得られる特徴量をそれぞれ特徴空間として用意し，学習効率に基づいて転移特徴空間と全探索空間を選択的に切り替えるハイブリッド型転移学習を提案している．

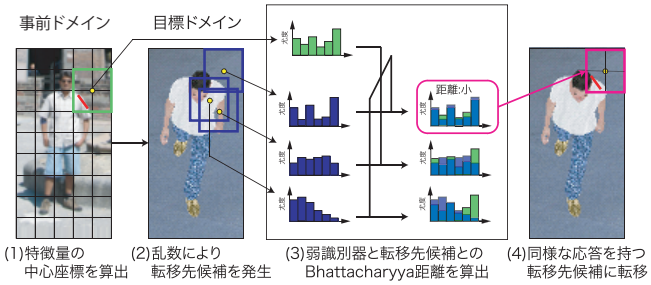

図 21 HOG 特徴量の転移.

5.1 特徴の転移

転移学習の前処理として，事前学習で選択された弱識別器と同様の特徴量を採用する弱識別器を，再学習するシーンのサンプル群 (目標ドメイン) を用いて特徴を転移する．まず，図 21(1) のように，事前学習で選択された弱識別器が捉える特徴量の中心座標を求める．次に，図 21(2) のようにこの座標を中心に，正規分布に従い L 個の候補領域を発生させる．候補領域から局所特徴量のヒストグラムを求め，図 21(3) のように事前学習で選択された弱識別器の局所特徴量のヒストグラムの Bhattacharyya 係数を求め転移尤度とする．

$$Bhattacharyya = \sum_{i=1}^{n} \sqrt{p_i q_i} \tag{7}$$

ここで p_i と q_i はそれぞれ異なるドメインの確率密度関数である．最後に，事前学習で選択された弱識別器を最も高い転移尤度を持つ転移候補へ転移させ，その集合を転移特徴空間 F_{Tr} と定義する．これに対し，再学習と同様に画像から全特徴量を抽出したものを全探索特徴空間 F_{Re} と定義する．

5.2 ハイブリッド型転移学習

ハイブリッド型転移学習では，事前ドメイン T_a と目標ドメイン T_t の目標シーンから切り出した学習サンプルを用いる．これらのサンプルは全てクラスラベルを持ち，ポジティブサンプルには $+1$, ネガティブサンプルには -1 を設定する．次に学習サンプルの重みを初期化する．学習サンプルの重みは目標ドメインと事前ドメインのそれぞれで正規化したものを初期値とし，それぞれの重みを $D_t(x_i)$ と $D_a(x_j)$ と表現する．弱識別器 $h(x)$ は次式より求める．

$$h_m = \mathrm{argmin}_{h_t}\Big(\sum_{(x_i,y_i)\in T_t} e^{-2y_i D_t(x_i)} y_i h_t(x_i) + \sum_{(x_j,y_j)\in T_a} \lambda_j e^{-2y_j D_a(x_j)} y_j h_t(x_j)\Big) \tag{8}$$

ここで，λ は共変量を表し，次式から求める．

$$\lambda = \frac{1+e^{-yH_a(x)}}{1+e^{-yH_t(x)}} \tag{9}$$

共変量 λ は，事前ドメインのサンプルが目標ドメインのサンプルにどれだけ適合しているかを表わし，目標ドメイン

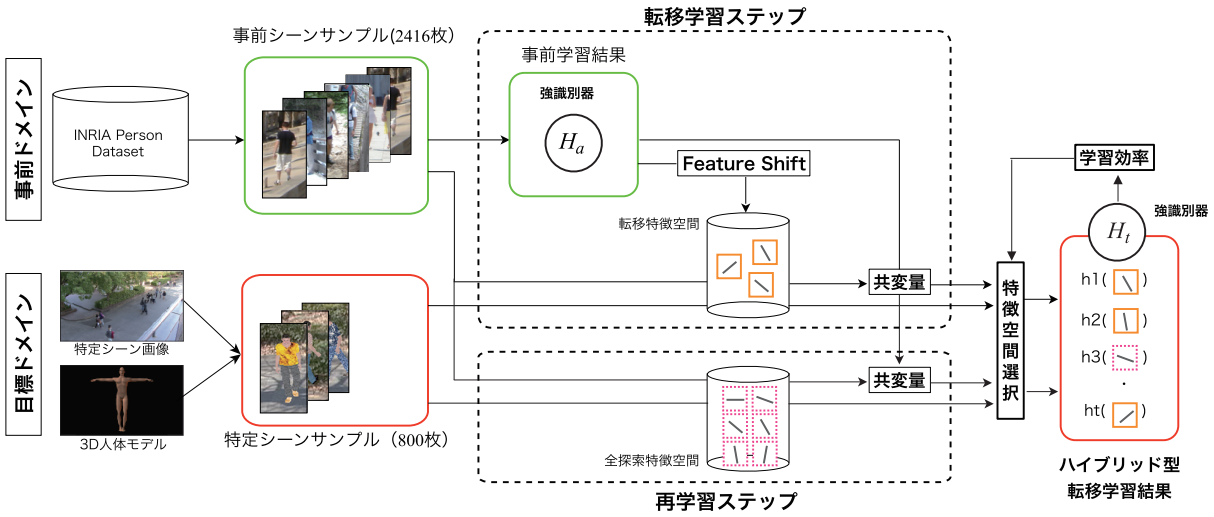

図 20 ハイブリッド型転移学習.

に適合しているほど大きな値となる．各 $h()$ は転移特徴空間 F_{Tr} を探索して求める．次に，式 (10) でエラー率 ϵ_m を算出する．

$$\epsilon_m = \frac{\sum_{h(x_i)\neq y_i} e^{-2y_i D_t(x_i)} + \sum_{h(x_j)\neq y_j} \lambda_j e^{-2y_j D_a(x_j)}}{\sum_i e^{-2y_i D_t(x_i)} + \sum_j \lambda_j e^{-2y_j D_a(x_j)}} \quad (10)$$

ここで，学習効率 ζ を算出し，その値が閾値以下のとき，全探索特徴空間 F_{Re} において弱識別器の再選択が行う．次に選択した弱識別器に対する重み α_m を式 (11) で算出する．

$$\alpha_m = \frac{1}{4} \ln \frac{1-\epsilon_m}{\epsilon_m} \quad (11)$$

次に，学習サンプルの重みを更新する．

$$D_t(x_i) = D_t(x_i) e^{-2y_i \alpha_t h_m(x_i)} \quad (12)$$

$$D_a(x_j) = D_a(x_j) e^{-2y_j \alpha_t h_m(x_j)} \quad (13)$$

以上の処理を事前学習の学習回数と同数繰り返す．最終的に，全ての弱識別器に重みを付けて多数決を取ることにより識別を行う強識別器を構築する．

5.3 学習効率に基づく特徴空間選択

通常の転移学習では，転移尤度の高い特徴量の転移特徴空間 F_{Tr} を対象に学習するため，これにより学習時間の削減（探索コストの低下）が可能である．しかし，事前学習ドメインと目標学習ドメインに大きな変化がある場合，転移特徴のみで補うことは不可能である．そこで，転移特徴空間と再学習同様の全探索特徴空間を選択的に切り替え，転移が有効な場合には尤度に基づく高速な転移学習を，転移が困難な場合には全探索特徴空間を用いて学習を行う．それぞれの特徴空間は以下のように定義する．

転移特徴空間

- 特徴次元：100（事前学習により選択）
- 弱識別器の閾値探索：必要なし
- 学習にかかる計算コスト：低
- ドメイン間の差が大きいと性能低下

全探索特徴空間

- 特徴次元：3,780
- 弱識別器の閾値探索：各次元ごとに 100 段階
- 計算コスト：高
- 目標ドメインに最適化

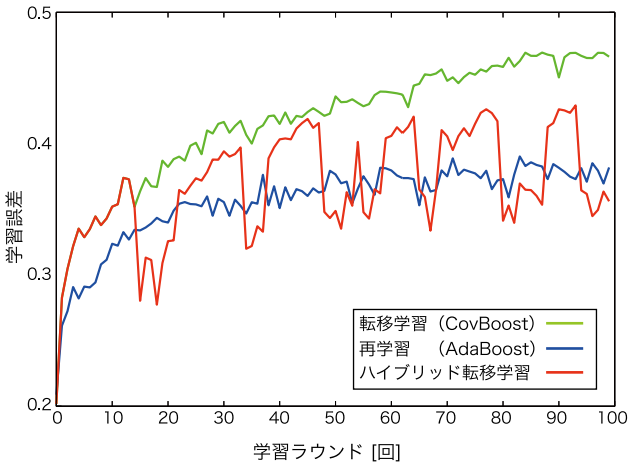

図 22 弱識別器のエラー率 ϵ

図 22 に，転移学習，再学習，ハイブリッド型転移学習の弱識別器のエラー率 ϵ の推移を示す．転移学習では目標シーンに大きな差があるとエラー率が上昇しやすい．ハイブリッド型転移学習では，転移学習が進行して傾きが緩やかになり，その絶対値が閾値を下回る際に全探索に切り替わる．全探索により有効な特徴を発見できれば大幅に ϵ が下がるため勾配が拡大し，再度転移学習へと移行する．このようにハイブリッド型転移学習では，転移学習と全探索を適応的にスイッチングしながら学習が進む．

図 23 俯角の違いによる学習サンプルの見えの変化.

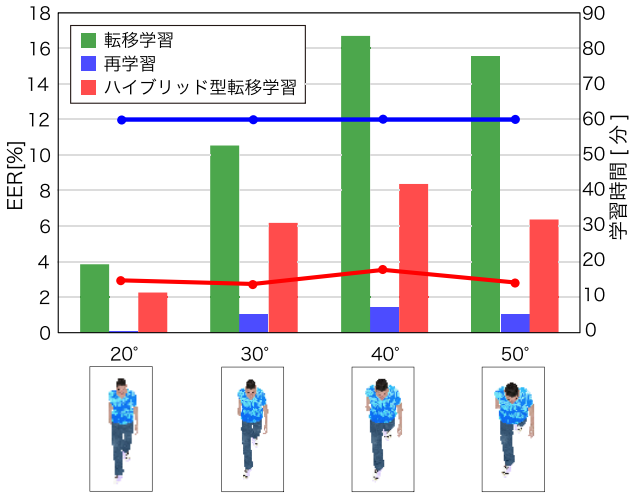

図 24 各手法の EER とハイブリッド型転移学習の学習時間．棒グラフは EER，青線はハイブリッド型転移学習の学習時間，赤線は再学習の学習時間を表わす.

5.4 ハイブリッド型転移学習の効果

ハイブリッド型転移学習の有効性を示すために，識別精度と学習に要する時間の観点から検証を行う．事前学習には，図 23(a) に示す INRIA person dataset をポジティブサンプルとして 2,416 枚使用し，ネガティブサンプルとして 12,180 枚使用する．目標ドメインのポジティブサンプルには，図 23(b) のように事前学習とは異なる俯角 20〜50 ° の CG を用いて生成した人画像を 800 枚用い，ネガティブサンプルとして背景画像 12,180 枚を用いる．ただし，比較手法である再学習については CG を用いて生成した人画像を 2,416 枚使用し，事前学習を行わない．評価用サンプルはそれぞれ特定シーンの俯角に設定した 3D 人体モデルを設置した CG 生成画像 10,000 枚をポジティブサンプルとして，ネガティブサンプルも背景画像を 10,000 枚使用する．

図 24 に，再学習，転移学習，ハイブリッド型転移学習の EER と学習時間を示す．図 24 より，従来の転移学習では，シーンの変化とともに性能が大きく低下することがわかる．一方，大きなシーンの変化を受けても，ハイブリッド型転移学習は再学習には及ばないが，転移学習に比べ 1.59%〜8.35%と性能が向上している．再学習は事前ドメインに頼らず大量のサンプルで目標ドメインに適応できたため，最も性能が高い．また，再学習に必要な学習時間に対して，ハイブリッド型転移学習では学習時間を約 1/3 〜 1/4 に短縮することができた.

5.5 ハイブリッド型学習により選択された特徴

ハイブリッド型転移学習は転移学習で対応しきれない大きな変化に対して，全探索で特徴量を補うことで高い識別能力を持つ識別器を構築することができる．図 25 はハイブリッド型転移学習で選択された特徴量のうち，転移学習ステップで選ばれたものを (a)，再学習ステップで選ばれたものを (b) として可視化したものである．図 25(a) より，転移特徴では標準的な肩のエッジや脚部の縦方向のエッジなどが転移されていることがわかる．一方，図 25(b) の全探索では横エッジが目立ち，俯角の変化により発生した上体部分の見えの変化に適応した新たな特徴が選択されている．図 25(c) にハイブリッド型転移学習全体として (a) と (b) を重ねたものを，(d) に再学習で選択された特徴量を表示する．両者を比較すると，特徴の位置関係や勾配方向が類似していることから，(c) のハイブリッド型転移学習は転移特徴と全探索の組み合わせにより (d) の再学習に近い特徴の構成を獲得している.

6. まとめ

本稿では，人検出の実用化に向けて，著者等の研究グループでこれまでに取り組んできた統計的学習手法による物体検出の高精度化と効率化について述べた．1 章のはじめにで述べた条件 (1) 検出失敗の理由を明確に把握することが可能については，弱識別器の応答とその弱識別器が選択した Joint 特徴量を観測することで可能であることを示した．条件 (2) 少ない学習サンプルでシステムをチューニング可能については，CG により自動生成した小サンプルとハイブリッド型転移学習を用いることで，目標シーンに適応する識別器を短時間で構築できることを示した．条件 (3) 省メモリで高速な計算アルゴリズムについては，検出対象が変更可能な Joint-HOG FPGA システムの実装を紹

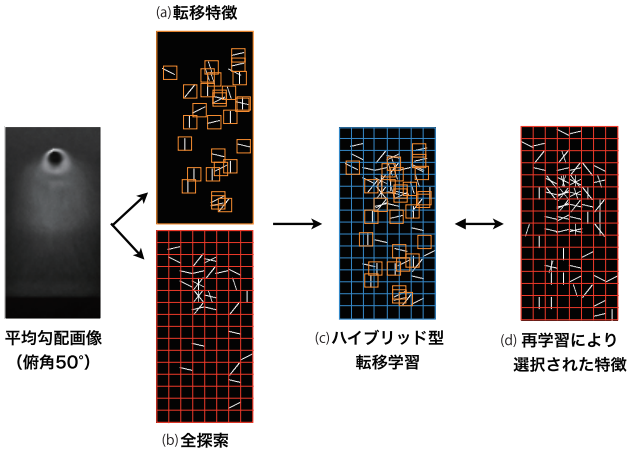

図 25 選択された HOG 特徴量：(a) 転移特徴量，(b) 全探索，(c)(a)+(b)，(d) 再学習．

介し，比較的小規模な FPGA でも約 20FPS で動作することを示した．

本稿で述べた統計的学習手法を用いた物体検出技術は，自動車やエアコン等の想定される環境下での利用が始まっている．統計的学習手法は，事前に作成した学習データへの依存性が高いため，今後は，検出器を稼働しながら入手した新しいデータを用いて識別器をより良くしていくオンラインチューニングが課題であり，全てを自動化するのではなく，能動学習 [23] を導入したアプローチによる実現が好ましいと考える．また，特徴量も同様に，大量のサンプルから Deep Learning[24] 等を用いて自動的に獲得した結果を基に，オンラインで再設計していくような手法が期待されている．

参考文献

[1] 藤吉弘亘，金出武雄：人を観る技術 PIA:People Image Analysis，映像情報メディア学会誌 映像情報メディア，Vol. 60, No. 10, pp. 1542–1546 (2006).
[2] 鷲見和彦：画像認識技術の実用化への取り組み 7. 人を見る画像認識技術，情報処理， Vol. 51, No. 12, pp. 1575–1582 (2010).
[3] 山内悠嗣，山下隆義，藤吉弘亘：[サーベイ論文] 統計的学習手法による人検出，電子情報通信学会パターン認識・メディア研究会 (PRMU) 技術報告, pp. pp. 113–126 (2012).
[4] Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 886–893 (2005).
[5] Sabzmeydani, P. and Mori, G.: Detecting Pedestrians by Learning Shapelet Features, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2007).
[6] Mita, T., Kaneko, T., Stenger, B. and Hori, O.: Discriminative Feature Co-Occurrence Selection for Object Detection, *PAMI*, Vol. 30, No. 7, pp. 1257–1269 (2008).
[7] 三井相和，山内悠嗣，藤吉弘亘：Joint 特徴量を用いた 2 段階 Boosting による物体検出，電子情報通信学会論文誌，Vol. J92-D, No. 9, pp. 1591–1601 (2009).
[8] 山内悠嗣，山下隆義，藤吉弘亘：Boosting に基づく特徴量の共起表現による人検出，電子情報通信学会論文誌，Vol. J92-D, No. 8, pp. 1125–1134 (2009).
[9] 山内悠嗣，藤吉弘亘，山下隆義：Boosting に基づく共起表現による人検出，画像の認識・理解シンポジウム (MIRU2008)，pp. 180–187 (2008).
[10] Yamauchi, Y., Fujiyoshi, H., Iwahori, Y. and Kanade, T.: People Detection Based on Co-occurrence of Appearance and Spatio-temporal Features, *National Institute of Informatics Transactions on Progress in Informatics*, No. 7, pp. 33–42 (2012).
[11] Fujiyoshi, H. and Kanade, T.: Layered Detection for Multiple Overlapping Objects, *IEICE transactions on information and systems*, Vol. 87, No. 12, pp. 2821–2827 (2004).
[12] 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. 93-D, No. 3, pp. 355–364 (2010).
[13] 松島千佳，山内悠嗣，山下隆義，藤吉弘亘：物体検出のための Relational HOG 特徴量とワイルドカードを用いたバイナリーのマスキング，電子情報通信学会論文誌 D, Vol. J94-D, No. 8, pp. 1172–1182 (2011).
[14] 山内悠嗣，金出武雄，山下隆義，藤吉弘亘：量子化残差に基づく遷移ゆう度モデルを導入した識別器の提案，電子情報通信学会論文誌， Vol. J95-D, No. 3, pp. 666–674 (2012).
[15] Watanabe, T., Ito, S. and Yokoi, K.: Co-occurrence Histograms of Oriented Gradients for Human Detection, *Information Processing Society of Japan Transactions on Computer Vision and Applications*, Vol. 2, pp. 39–47 (2010).
[16] 株式会社東芝：画像認識プロセッサ Visconti シリーズ, http://www.semicon.toshiba.co.jp/product/assp/selection/automotive/infotain/visconti/index.html.
[17] 矢澤芳文，吉見 勤，都筑輝泰，土肥知美，藤吉弘亘：検出対象をリコンフィグ可能な Joint-HOG による FPGA ハードウェア検出器, 画像センシングシンポジウム (2011).
[18] 土屋成光，山内悠嗣，藤吉弘亘：人検出のための生成型学習と Negative-Bag MILBoost による学習の効率化，画像の認識・理解シンポジウム (MIRU2012) (2012).
[19] 村瀬 洋：画像認識のための生成型学習，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. 46, No. 15, pp. 35–42 (2005).
[20] 安藤寛哲，藤吉弘亘：人検出結果に基づく自己カメラキャリブレーションと 3 次元位置推定，電気学会論文誌. D, 産業応用部門誌， Vol. 131, No. 4, pp. 482–489 (2011).
[21] Viola, P., Platt, J. C. and Zhang, C.: Multiple instance boosting for object detection, *Neural Information Processing Systems*, pp. 1419–1426 (2006).
[22] 土屋成光，山内悠嗣，山下隆義，藤吉弘亘：ハイブリッド型転移学習による物体検出における学習の効率化，電子情報通信学会 パターン認識・メディア理解研究会 (2013).
[23] Settles, B.: *Active Learning*, Morgan & Claypool Publishers (2012).
[24] 岡谷貴之，斎藤真樹：ディープラーニング，情報処理学会研究報告 CVIM 185 (2013).