[招待講演] 動画像理解による人を観る技術

藤吉　弘亘†

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

あらまし　動画像からの人を検出は，ビデオ監視において重要な技術である．既に，固定カメラによる背景差分に基づく人等を対象とした動体検出法が数多く提案されているが，このような動体検出をベースとしたアプローチでは，対象とする動体の検出に失敗すると次段の処理である物体識別が不可能となる問題がある．この問題を解決するアプローチとして，近年のコンピュータの高速化に伴い，画像全体を検出ウィンドウによってラスタスキャンし，low-level な特徴量と統計的学習手法の組み合わせによる物体検出法が提案されている．本稿では，従来のアプローチとして VSAM プロジェクトで開発された動画像理解アルゴリズムと，その実用化例について紹介する．また，人画像解析ための人を検出する新しいアプローチとして，検出ウィンドウのラスタスキャンベースによる人検出法について紹介する．
キーワード　Video Surveillance and Monitoring(VSAM)，人画像解析，人検出

People Image Analysis by Video Understanding Technology

Hironobu FUJIYOSHI†

† Dept. of Computer Science, Chubu Univ. 1200 Matsumoto, Kasugai, Aichi, 487–8501 Japan
E-mail: †hf@cs.chubu.ac.jp

Abstract　Automatic people detection is a key enabler for applications in visual surveillance. A number of methods have been developed based on background subtraction for detecting motion from images captured by fixed cameras. Because motion detection-based methods such as background subtraction use a top-down approach, object classification in the next step becomes impossible if the object's region is not segmented correctly. A window-scanning approach, which is a combination of low-level features and statistical learning, has been proposed for solving this problem because of the great improvement in computer speed over the last several years. In this paper, we describe the conventional approach of a video understanding algorithm, developed during the VSAM project, for detecting motion, and we introduce applications that use it. We also introduce an example of how a window-scanning approach for detecting people can be implemented as a novel approach for analyzing the images of people.
Key words　Video Surveillance and Monitoring(VSAM)，People Image Analysis，People Detection

1. はじめに

近年，街中での犯罪率が増加傾向にあり，セキュリティに対するニーズが高まっている．既にイギリスでは 100 万台以上の監視カメラが設置，国内においても新宿歌舞伎町において 50 台の監視カメラが導入され，犯罪防止や事件の検挙に活用されている．それとともに，物体検出や追跡等の動画像理解技術の要求は高まりつつあり，知的な映像監視に関する研究は 1997 年の DARPA(Defence Advanced Research Projects Agency) 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 [1] を一つのきっかけに，以来いっそう進展し，現在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ている．

さらに動画像理解技術は監視目的だけでなく，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 [2] プロジェクトでは，生活空間である家にカメラをはじめとする数多く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．また，リビングルームを対象とした Microsoft Research の Easy Living [3] のように，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技術は QoLT(Quality of Life Technology) と呼ばれる技術の一環で，QoLT のためのセンシング技術は，刻々と変化する人の状態を実時間で認識する必要がある．特に，人画像解析 (People

Image Analysis) として，動画像からの人の検出，追跡，顔の検出，顔の部位の追跡，モーション理解が不可欠な技術要素となる．近年のコンピュータの高速化に伴い，画像全体を検出ウィンドウによってラスタスキャンし，low-level な特徴量と統計的学習手法の組み合わせによる人検出法が提案されている．本稿では，従来のアプローチとして VSAM プロジェクトで開発された動画像理解アルゴリズムと，その実用化について紹介する．また，人検出法の新しいアプローチとして，検出ウィンドウのラスタスキャンベースによる人検出法についても述べる．

2. 動画像理解技術を用いたビデオ監視システム：VSAM

重要施設の入退出管理を目的とした従来のビデオ監視システムは，監視カメラ映像を記録するものや，監視員が複数のカメラ映像を同時にモニタリングするものが多い．監視する範囲が広く 24 時間監視となると監視員への負担が大きくなり，問題とされている．これに対し，米国では 1997 年より 2000 年の 3 年間，DARPA の下，画像理解技術を用いたビデオ監視システムの研究プロジェクト VSAM が行われた．CMU では，キャンパスに 12 台のカメラを配置し，テストシステムを構築した [1]．本システムは，動画像理解技術により検出した侵入物体を複数のカメラが協調してトラッキングし，その状況をリアルタイムで監視員に提示する．これにより，監視員の負担軽減と作業効率化に大きく貢献できる新しいビデオ監視システムとして注目された．

2.1 VSAM における動画像理解技術

図 1 に VSAM における動画像理解の流れを示す．まず，入力画像と背景画像との差分を計算し，動体検出を行う．次にセグメンテーションした結果から特徴量を抽出し，この特徴量を基に識別を行う．ここでは，VSAM システムに用いられている動画像理解技術として，動体検出と物体クラス識別について述べる．

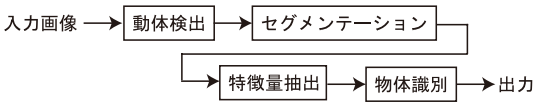
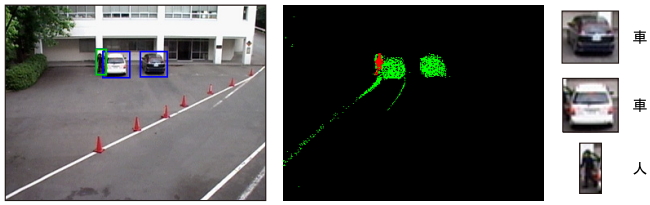


図 1　VSAM システムによる物体識別の流れ

2.1.1 レイヤー型動体検出

侵入物体の検出には，検出すべき物体が存在しない背景画像を予め用意しておき，入力画像と背景画像の差分を計算する背景差分処理が多く用いられている．人と自動車のアクティビティ認識するには，画像上の複数物体の重なりを検出する必要があるが，背景差分処理と領域クラスタリングを組み合わせた手法では隣接する複数物体を 1 つの領域として検出してしまうという問題がある．これに対して，我々はピクセル分析とリージョン分析の 2 つの処理からなるレイヤー型検出法を提案した [4]．ピクセル分析では，各画素の輝度値の時間変化を観測し，その変化軌跡により画素の状態を静もしくは動と判定する．リージョン分析では，動とラベル化された画素領域を移動体と判定する．静とラベル化された領域は静止物体と判定し，背景上のレイヤーとして記憶する．一度停止した物体は，再び動き出すまでレイヤーとして登録されているため，レイヤー上を通過する移動体を区別して検出することが可能となる．図 2 は，停止した車から人が降りる動画に対してのレイヤー型検出例 (図 2(a))，ピクセル状態分析例 (図 2(b))，物体識別例 (図 2(c)) である．停止した車は，動き出すまでレイヤーとして登録されているため，車と人が画像上で重なった場合にも区別して検出することが可能である．



(a) レイヤー型検出例　(b) ピクセル状態分析例　(c) 識別例

図 2　レイヤー型動体検出例

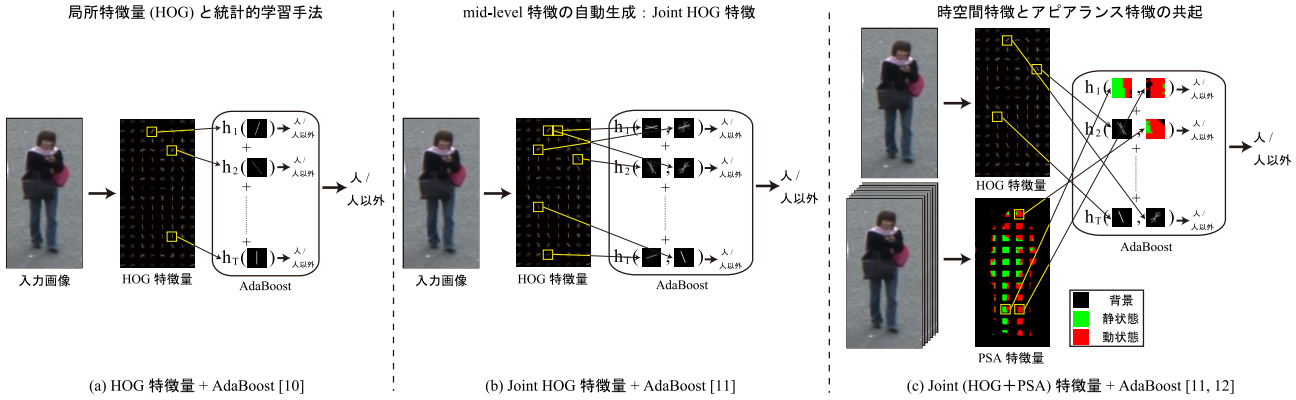
2.1.2 物体クラスの識別

レイヤー型検出により検出した物体は，対象物体の姿勢変化，天候による照明変化，カメラの位置等の要因により，その「見え」は逐次変化するため，人と自動車の識別は容易でない．VSAM システムでは，12 台のカメラを見えの大きく異なるカメラ毎にグループ化し，人工ニューラルネットワーク (ANN) を用いて，それぞれの識別器を作成することで対処している [5]．検出した画像から計算した形状の複雑度 (周囲長^2/面積)，面積，縦横比，カメラのズーム倍率を求め，ANN への入力パラメータとした．出力は，人 (一人), 人のグループ (二人以上), 自動車, その他の 4 クラスとした．小規模な ANN(入力層 4, 中間層 16, 出力層 4) で構成しているため，リアルタイムでの識別処理を実現することができた．

2.2 VSAM 動画像理解技術の実用化

1997 年に始まった VSAM プロジェクトでは，当時 Pentium II 300Mhz 相当の PC を用いて，約 5 ～ 10fps のリアルタイム動画像処理を実現していた．10 年後の現在では，計算処理能力が大幅に向上していることから FPGA(Field Programmable Gate Array) によるハードウェア化や DSP(Digital Signal Processor) を用いた高速処理が可能である．VSAM プロジェクトマネージャが起業した米 ObjectVideo 社 [6] では，VSAM 動画理解技術をベースとした，国境，空港，化学工場，核燃料施設等，重要な施設の警備・防犯用途の製品を提供している．また，ObjectVideo 社のアプリケーションでは，ユーザがセキュリティ規則 (アクティビティ) を作成し，監視の自由度を高めている．ユーザは進入する物体に対して，進入領域や物体の種類などをセキュリティ規則を設定することにより，事前に定義した規則に反する状態のみをユーザへリアルタイムに通知する．近年では，Texas Instruments 社により，ObjectVideo 社の技術を組み合わせたインテリジェンス機能付きビデオ監視カメラも開発されている．

VSAM で開発されたカメラ間のマスタースレーブ技術は，エンターテイメント分野へ応用されている．アメリカのテレビ局 CBS が，映画 Matrix で使用された映像効果をリアルタ



(a) HOG 特徴量 + AdaBoost [10]　(b) Joint HOG 特徴量 + AdaBoost [11]　(c) Joint (HOG+PSA) 特徴量 + AdaBoost [11, 12]

図 4　特徴量の捉え方

(a) 入力画像　(b) 背景差分画像

図 3　背景差分による動体検出の失敗例

イムで提供する EyeVision システム [7] を CMU と開発した．EyeVision システムは 32 台のロボットカメラ [8] を活用し，複数の角度から撮影された映像を連続的に合成した 3 次元的なテレビゲームのような感覚でユーザに映像が提供できる．2001 年の第 35 回スーパーボール (Super Bowl XXXV) にて初めて実用化され，テレビ放送で用いられた．

3.　人を観る技術：PIA

VSAM における動体検出法は，入力画像と背景画像の差分を計算する背景差分ベースの手法であった．このような動体検出をベースとした動画像理解のアプローチは，移動体同士が画像上で重なった場合，セグメンテーションに失敗するため，その後の処理である物体識別が不可能となる問題があった．図 3 に背景差分による動体検出の失敗例を示す．図 3 の A，B，C のように人同士が画像上で重なった場合，背景差分結果を用いると 1 つの動体として検出される．これに対して，Viola と Jones は Haar-like と呼ばれる局所特徴と統計的学習の組み合わせによる高速かつ高精度な顔検出法 [9] を提案した．この手法は，入力画像に対して検出ウィンドウをラスタスキャンし，検出ウィドウを顔/非顔と判別するため，前処理としての動体検出を必要としない．そのため，ラスタスキャン方式による顔検出は，デジタルカメラ等に搭載され実用化されている．近年では，形状変化が大きいため検出が難しいとされている人を対象とした研究が盛んである．本章では，近年のアプローチである局所特徴量と統計的学習を組み合わせた人検出法と，さらに高精度化のための有効な特徴量の捉え方について述べる．

3.1　局所特徴量 (HOG) と統計的学習手法による人検出

人は動きとともに形状が変化する非剛体な物体であるため，顔検出と比較して難しい問題である．また，画像中での人同士の重なりによるオクルージョンの発生や衣服の違い，照明や影の影響も検出を困難とする要因である．このような問題に対して，Dalal 等により局所領域における勾配方向をヒストグラム化した Histograms of Oriented Gradients(HOG) 特徴量と統計的学習手法を組み合わせた人検出法 [10] が提案された．HOG 特徴量は，照明の変動による影響が少なく，局所的な幾何学的変化に頑健であるため高精度な人検出を可能とした．

図 4(a) に AdaBoost による HOG 特徴量の捉え方を示す．AdaBoost の弱識別器により 1 個の HOG 特徴量が選択され，最終的に多数ある弱識別器の重み付き多数決により人と人以外に判別する．

3.2　mid-level 特徴の自動生成：Joint HOG 特徴

人の形状は左右対称性や連続したエッジ等の特徴があり，これらの特徴を捉えることで検出精度を向上させることができると考えられる．三井等は，人独特の形状を捉えるために，複数の HOG 特徴量を組み合わせた Joint HOG 特徴量と，2 段階に構築した AdaBoost による学習法 [11] を提案した．複数の low-level な特徴量である HOG 特徴量を AdaBoost により組み合わせることで mid-level な特徴量である Joint HOG 特徴を自動的に生成し，この Joint HOG 特徴を入力とした 2 段階目の AdaBoost により最終識別器を構築する．図 4(b)AdaBoost による Joint HOG 特徴の捉え方を示す．HOG 特徴量と AdaBoost による識別時には，1 個の弱識別器が 1 個の HOG 特徴量を用いて識別したのに対し，Joint HOG 特徴と AdaBoost では，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可能となり，高精度な人検出法を実現した．図 5(a) に人の平均勾配画像，図 5(b)，(c) に AdaBoost により選択された HOG 特徴量を可視化した結果を示す．HOG 特徴量の勾配方向を 9 方向で表現しており，輝度が高いほど AdaBoost における弱識別器の重みが高いことを表す．1 段階目で選択された HOG 特徴量 (図 5(b)) は全ての局所領域において選択されているが，2 段階目で選択された HOG 特徴量 (図 5(c)) では，人の輪郭に沿った HOG 特徴量が自動的に選択され，識別に有効であることがわかる．

3.3　時空間特徴とアピアランス特徴の共起

Joint HOG 特徴では，アピアランス特徴である HOG 特徴量以外の特徴を追加することが可能であり，文献 [12] にて有効性が確認された時空間特徴に基づく特徴量としてピクセル状態

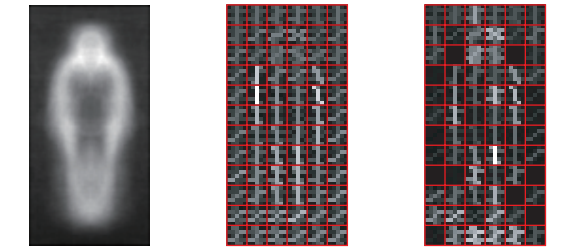


図 5　選択された HOG 特徴量の可視化

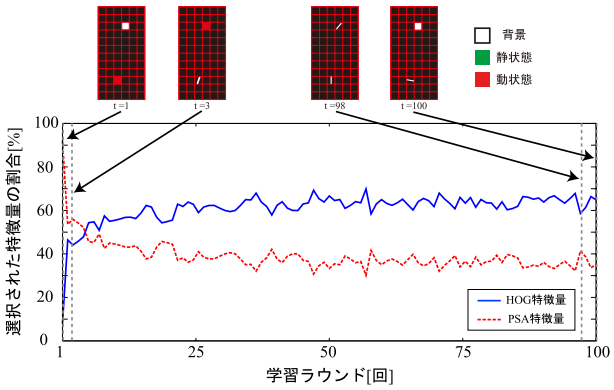


図 6　選択された特徴量の割合

分析 (PSA) による時空間情報を加えることにより，より高精度な人検出を達成した．ピクセル状態分析とは，2. 1. 1 で示したレイヤー型検出に用いられた手法であり，ピクセル状態の時間変化をモデル化し，各ピクセルを背景と動状態，静状態に判別する手法である．ピクセル状態分析の結果である時空間特徴とアピアランス特徴を図 4(c) に示すように同時に捉えることで歩行者の人らしさをより捉えることが可能となる．これにより，アピアランスの情報のみでは誤検出する人に似た物体に対して誤検出を抑制することができる．

図 6 に AdaBoost の各学習ラウンドにおける HOG 特徴量と PSA 特徴量の選択された割合と，その際に選択された特徴量の可視化の例を示す．初期ラウンドにおける弱識別器では PSA 特徴が多く選択され，学習ラウンド数が進むにつれて HOG 特徴量が選択される割合が多い．これは，まず物体の動きを表すことが可能な PSA 特徴により，大まかに人と人以外を判別することにより，動いている物体と背景を判別する．次に，アピアランスを表す HOG 特徴量を用いて，人の輪郭を捉えることにより人の判別を行うと考えられる．

図 7 に本手法による人検出の例を示す．人の大きさの変化や人同士の画像上での重なりによる部分的な隠れに対しても高精度な人検出が可能である．図 8 は，評価実験の結果を示す DET(Detection Error Tradeoff) カーブであり，原点に近いほど識別器の性能が高いことを示す．従来人検出に用いられている HOG 特徴量のみを使用した場合より，Joint HOG 特徴の識別性能が高いことがわかる．また，人のアピアランス特徴と時空間特徴を同時に捉えることで，さらに高精度な人検出が可能となることがわかる．Joint HOG+PSA 特徴を用いた際，誤検出率 5.0%において検出率を 99%まで向上させることができた．

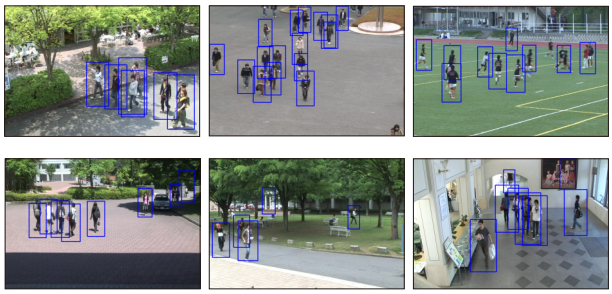


図 7　人検出例

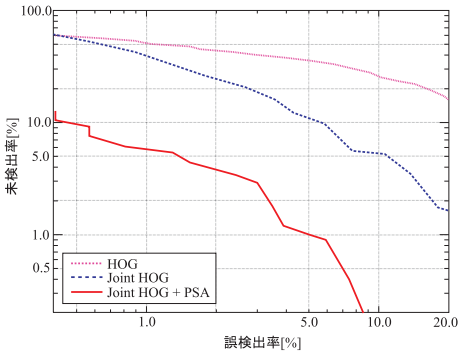


図 8　DET カーブ

4. おわりに

本稿では，動画像理解技術を用いたビデオ監視システムのプロジェクト VSAM と，人を観る技術である人検出法について紹介した．近年のコンピュータの発展とともに動画像理解技術も発展し，監視目的だけでなく，一般社会の中で人の生活を支援する技術への応用とその実現が期待されている．

文　献

[1] R. Collins, A. Lipton, H. Fujiyoshi and T. Kanade: “Algorithms for cooperative multi-sensor surveillance”, Proceedings of the IEEE, Vol. 89, No. 10, pp. 1456 - 1477 (2001).
[2] Aware House, http://www.gatech.edu/innovations/futurehome/
[3] Easy Living, http://research.microsoft.com/easyliving/
[4] H. Fujiyoshi and T. Kanade, “Layered Detection for Multiple Overlapping Objects”, IEICE Transactions on Information and systems, vol. E87-D, pp. 2821-2827 (2004).
[5] 長谷川修, 金出武雄：“一般道路映像中の移動物体の識別・色の推定と特定対象の検出”, 情報処理学会論文誌, Vol.44, No.7, pp.1795-1807 (2003).
[6] ObjectVideo, http://www.objectvideo.com/
[7] Eye Vision, http://www.eyevision.com/
[8] 川内直人, 金澤宏幸, 見持圭一, 宮内礼三, 大西献, 中山淳二, 藤田淳, 大道武生: “新映像システム EYE VISION 向けカメラ制御パンチルトの開発”, ロボティクス・メカトロニクス講演会, p. 59 (2002).
[9] P. Viola and M. Jones, “Robust Real-Time Face Detection”, Int. Journal of Computer Vision, 57(2), pp. 137-154 (2004).
[10] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection”, IEEE Computer Vision and Pattern Recognition, pp. 886-893 (2005).
[11] 三井相和, 山内悠嗣, 藤吉弘亘, “Joint HOG 特徴を用いた 2 段階 AdaBoost による人検出”, 第 14 回画像センシングシンポジウム SSII08, IN1-06 (2008).
[12] 山内悠嗣, 藤吉弘亘, Bon-Woo Hwang, 金出武雄, “アピアランスと時空間特徴の共起に基づく人検出”, 画像の認識・理解シンポジウム (MIRU2007) (2007).