Vol.2014-CVIM-190 No.12
2014/1/23

個と群の相互作用に基づいた人の行動・属性の認識

浮田 宗伯[1,a)]

概要：個人と個人のインタラクション，および個人と群（人のグループや集合）との間のインタラクションを解析することによって，1）個々人の行動や属性の認識性能を向上させ，2）さらに従来は認識できていなかった群の状態・行動まで認識可能とする技術を紹介する．従来のコンピュータビジョンにおいては，個人の行動はその個人の動きにのみ注目して認識されていたのに対して，周辺の他者を中心としたシーンのコンテクストを考慮することにより，静止画・動画を問わず認識性能を向上できる．本稿では，複数対象追跡，行動認識，性別や年齢などの属性認識，複数人のグルーピングや関係性の認識など，個と群のインタラクションに注目した関連研究を広くサーベイする．

1. はじめに

様々な実世界応用システムを構築する上で，人は最も重要な観測・認識の対象である．コンピュータビジョンにおいては，顔認識，対象追跡，人体姿勢推定，行動認識などが盛んに研究されており，それぞれサーベイも多くなされている（顔認識 [109][108]，対象追跡 [106][107][71] ，姿勢推定 [53][62]，行動認識 [63][96][10] ）．

各研究分野において，初期の研究では，簡単化された問題が取り組まれていた．特に観測対象数（観測人数）については，いずれの分野でも 1 人のみが対象とされていた．例えば，見え方のオンライン学習による一人の姿勢追跡 [64]，行動認識における視点非依存な特徴量 [50][110] や，それらのアイディアを組み合わせた視点非依存な特徴量のオンライン学習 [66]，視点変化に加え遮蔽に対する頑健性も向上させる特徴量 [95][52]，対象形状の変化に頑健な検索・回帰ベースの姿勢追跡 [89][87] や複数行動に渡る大きな動きの変化や微妙な変化も姿勢推定・行動識別加能とする手法 [88][86] などである．人の属性推定の研究においても，盛んに研究されている顔認識だけでなく，性別識別や年齢推定などにおいても，加齢による変化の個体差に対する頑健性向上 [30] や年齢間の学習画像数のアンバランスに対する頑健性向上 [11] などが研究されている．

このように 1 人を対象にした研究から，手法の発展に伴い複数対象が同時に観測されるシーンへと研究が進み，さらに，複数対象がインタラクションし合うシーン，多数対象が同時に観測されるシーンなどが研究対象となっている．

[1] 奈良先端科学技術大学院大学
NAIST, Ikoma, Nara 630–0192, Japan
[a)] ukita@is.naist.jp

図 1 本稿の対象範囲．2 人以上の複数人における属性・姿勢・行動の推定から，群集における対象追跡までを取り扱う．

対象追跡： Data association に基づくアルゴリズムの高速化 [61][74] による多人数・長時間の動画への追跡の大域的最適化や，Data association と mean sift tracker [16] との組み合わせ [97] など．

姿勢推定： 人体動作の事前知識を活用した遮蔽に対する頑健性の向上 [90] や重なり合う複数人の姿勢推定に適した特徴量の重み設定 [24] など．

行動認識： Bag-of-words アプローチによる遮蔽に対して頑健な行動認識 [55].

これらの研究は，主に遮蔽に対する頑健性向上の点で優れた成果を得ている一方，個々の対象の特徴にのみ注目している．このような手法の性能をさらに向上させるため，個々人ではなく，個と個の間のインタラクションによって生じる相互作用や，個とグループや，個と群の間の相互作用，さらには人以外の物体や環境などの周辺コンテクストとの相互作用を考慮に入れるアプローチの有効性が確認されてきている．本稿では，こうした個と群やコンテクストの相互作用に基づいた手法について，図 1 に示すような比較的少人数における詳細な情報である属性・姿勢・行動の認識から，群集における追跡にわたって広く紹介する．

ⓒ 2014 Information Processing Society of Japan

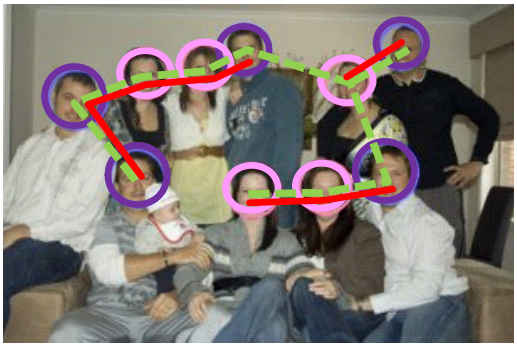


図 2 群における性別推定のための顔配置のモデル化（文献 [27] を参照）．点線が最近傍対応づけによるすべての検出顔を結んだグラフ，実線がすべての顔をノードにした最小全域木（より近づき合うペアになる男女が結びついている）である．それぞれが，複数顔の配置に関する異なる情報を提供している．これらに顔の絶対座標および相対座標（男性の方が女性よりも背が高い，子供よりも大人のほうが背が高い，などの現象を表現）を加えてモデル化することにより，個々人の属性を推定している．図の例は，性別の識別（紫とピンクの丸が，それぞれ男性と女性を示す）の結果である．

2. 属性推定

画像からの人の属性（年齢，性別，職業など）推定では，その人の顔，服装，動きなどがヒントとなる．中でも盛んに研究されている顔認識では，古典的な Strongly-supervised な問題設定下では，既に 100%近い精度が得られている．しかし，大量人数を対象にした際に必要となりうる Weakly-supervised な問題設定では [7][79][29][*1]，認識精度は大きく落ちてしまう．また，属性推定では，その属性と画像情報量との相関性に応じて，問題はさらに難しくなってしまう．

このようなより推定が難しい属性推定のために，個人だけでなく画像中の他者や集団との関係性を利用することが有効である．文献 [7] が画像とキャプションやタグ中の名前の同時生起確率に基づいて顔学習と認識を実現しているのに対し，文献 [79][29][25] では，ある人と人とが同時に観測されている確率まで参照することによって，認識精度を向上させている．

このアプローチを発展させ，年齢や性別まで推定することも可能である [26]．文献 [27] では，図 2 に示すようにさらに多人数のグループにおける人の間の位置関係までモデル化することによって，より多くの人間が写っている，すなわち各人の顔が小さく写ってしまっている画像においても，年齢推定と性別識別を可能にしている．上記文献 [26][27] が各人の顔にラベルが与えられている Strongly-supervised な設定であるのに対し，同様の問題を Weakly-supervised な設定で実現している手法 [91] もある．最新の研究では，

[*1] 例えば，ニュース・新聞やインターネット上における顔画像を対象として，個々人の顔に対してではなく，複数人が写っている画像全体に対するキャプション，記事，タグなどを学習データとする問題設定．

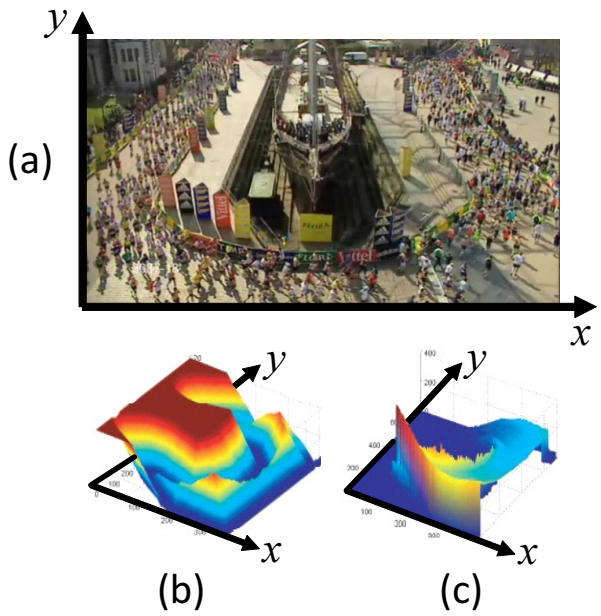


図 3 集団の流れの学習例（文献 [2] より引用）．(a)：サンプル画像．(b)：(a) における人の流れの境界を表現するフロー場（青いほど人が流れる領域で，赤いほど人が侵入しない領域を示す）．(c)：各領域における人の時間的な流れを表現するフロー場（赤い領域から時間にそって青い領域へ人が移動していく）．いずれのフロー場も，確率的にモデル化されている．これらシーン固有のフロー場と各瞬間に計測される局所的なフロー場（各対象の移動を表している）を統合することにより，対象追跡を実現している．

各身体部位の見え特徴量から服装に関する情報に加え，同時に観測されている他者との位置関係に応じて，各人の職業まで推定している [75]．

3. 対象追跡

3.1 多人数追跡

従来は，複数対象追跡というと 10 人から 20 人程度を指していたが，近年では，100 人を越える多人数を視野内で同時に観測・追跡することも可能になってきている．これほどの多人数を視野内で同時追跡する場合，図 3(a) に示す例のように各対象の撮影像は極めて小さく（10〜20 ピクセル程度）遮蔽の程度も大きいため，HOG [17] による対象検出や Mean Shift [15] による領域追跡などの広く用いられている手法をそのまま適用することは困難である．

このような多人数の混雑環境における追跡問題に対するためにも，各追跡対象とその周辺の他者や集団との相互作用を利用できる．このような手法を，2 種に大別する．

集団による（遮蔽による）見えの変化の学習：多人数が同時観測される混雑環境下における遮蔽による見えの変化の学習 [41] や複数人の見え方の変化の学習 [81]，遮蔽の見えの識別に適したモデルや識別空間の学習 [42][92] などが挙げられる．文献 [68] では，混雑程度に応じた遮蔽モデル

ⓒ 2014 Information Processing Society of Japan

(a) Scene (b) Grouping people trajectories

図 4 観測空間中を移動する人々．(a) は観測空間の様子のサンプル画像であり，同一グループの人同士が近づき合い，他者とは距離を維持できる程度の混雑環境が，移動軌跡からのグループ検出に必要な条件である．(b) は人の軌跡の鳥瞰図（床面に平行な 2D 平面座標）．この人々の軌跡を，それらの時空間的な関係性に基づいて，(b) に示すようにグループごとに分割することを目標とする．

を定義し，更なる精度向上を実現している．文献 [69] では，混雑時の見えをローカルパッチで学習し，学習データとクエリデータのマッチングによって手段の動きを解析している．

対象が多人数でなくとも，追跡対象の周辺の見えを学習することによって追跡の頑健性を向上させる研究は従来も存在したが（文献 [100] など），上記手法では混雑による遮蔽状態まで学習しようとする．

集団の流れ・見えの学習： 多くの撮影対象が意図的にほぼ同じ方向に進むようなシーン（例：図 3(a) に示すようなマラソン）や，混雑状態のせいで全員同じ方向にしか進めないようなシーン（例：駅構内）などでは，その人の流れを参照することによって追跡頑健性を向上できる．文献 [1] では，流体力学で用いられている定式を人の流れに適用し（すなわち，個々をパーティクルと見なし，その全体の流れを水などの流体の流れとしてシミュレートして移動を予測する），集団の動きをセグメンテーションしている．この考えを発展させ，シーン固有の人の流れをモデル化し（図 3(b)(c) を参照），人の群の中で各人の追跡精度を向上させることも可能である [2]．文献 [67] では，Correlated Topic Model [9] を利用していくつかのパターンの組み合わせによって，シーン全体の流れが時間によって変化するようなシーンにも適用できるようにしている．

3.2 人のグループ検出

3.1 節では，複数人による遮蔽の局所的な見え，または逆にシーン全体の人の流れをモデル化する研究を紹介した．これに対し，それらの中間的な表現，すなわち，共に移動する複数名で構成されるグループを検出することにより，対象追跡に利用することも可能である．グループに関する情報は，他にもグループ構成に応じたマンナビゲーションなど多くの実世界応用を可能にする有用な情報である．

ほとんどのグループ検出法は，図 4 に示すような対象の移動軌跡の近接性や類似性に基づいている．文献 [12][99] では，文献 [73][60] のように人の動きのシミュレーションで広く利用されている Social Force Model[*2][32] のようなモデルによって，人の間に働く引力や斥力を表現し，それらを均衡させるような最適化によって人のグループを検出している．文献 [56] では，任意の軌跡のペアから計算される時空間的な特徴量を全時刻で計算し，それらの類似性および共起性を評価することによって，複雑な最適化やクラスタリング無しでグループ検出を実現している．手法 [12][99][56] がペアごとのような小さい単位でグループ検出のための力を計算しているのに対し，文献 [28] では，検出対象全体の階層的なクラスタリングによってグループを検出している．上記の手法がいずれも対象の軌跡を抽出後にグループ検出を行っているのに対して，文献 [59] では，各フレームにおける対象の検出座標を入力にした条件付確率場によって，各対象の移動軌跡とグループ検出を同時最適化している．

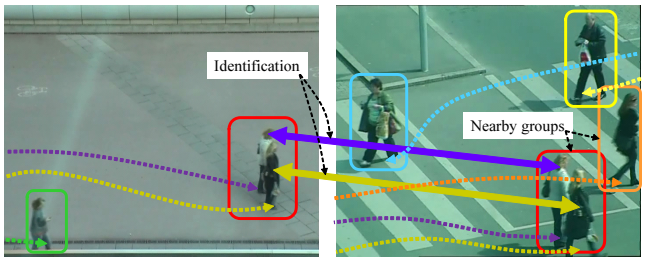


図 5 分散カメラ間の個人同定例の例．広域人追跡の実現には，各カメラの視野内で撮影された人の追跡（破線）と，各カメラで撮影された人の間で同一人物判定を行うカメラ間人同定（実線）が必要．この精度向上のため，個人の見え特徴に加えて，各人が属するグループの特徴量も利用する．図中の各枠が，それぞれグループを示している．著者らの手法では，各人が属するグループの人数を特徴量化する際，グループ外の人が近接している（例：図中の赤枠とオレンジ枠のグループ），すなわち正確な検出が難しいグループほど，そのグループ特徴量の重みを下げる処理が加わっている．

本節の最後に，グループ検出結果の利用例として，図 5 に示した著者らによる分散カメラでの個人同定（文献 [98] など）の頑健化について述べる．この手法では，カメラ間をまたいだ個人同定において，個人の見え特徴に加え，新たに提案するグループ特徴量を併用する．グループを特徴量化して対象同定に利用するためには，手法 [111] のようにグループ構成員の見えをまとめて特徴量化することが可能である．このアイディアは，基本的には視野内対象追跡にも用いられている（例：手法 [100]）．このようなグループの見え特徴量に加え，著者らの手法では，グループの構成人数を特徴量化している．この特徴量化では，グループ検出の失敗に対処するため，各視野内で同時観測されている他グループとのグループ識別性の難しさやなども考慮に入れている．

[*2] 人の親密度に応じた物理的な距離関係をモデル化している．

ⓒ 2014 Information Processing Society of Japan

4. 姿勢推定

人体姿勢[*3]は，人の行動認識のような人の状態推定に重要な情報となるため，その推定手法が広く研究されている．多くの姿勢推定では，人体構造を頭や胴体のような人体パーツをノードとし，パーツ間の依存関係をリンクで表したグラフィカルモデルで表現する．特に，Deformable Part Models (DPM)，中でもグラフを木構造に限定することで学習・探索効率を向上させた Picotial Structure Models (PSM) [23], およびそれらを識別学習する Discriminatively-Trained DPM [22][*4]が広く利用されている．

DPM は，人体の階層構造のモデル化 [94][82] や，PSM の利点を部分的に捨てることで推定の頑健性を向上させるアプローチ（例：ループを含んだグラフによるパーツ間の高次依存関係の表現 [8][80] や，リンクで接続されている親子パーツ間の見え特徴のモデル化 [72][84]）などが研究され続けている．しかし，人体の遮蔽が大きくなるとその推定精度は大きく落ちてしまう．自己遮蔽を陽にモデル化することで対頑健性を向上させる手法 [77][93] も，まだ十分な推定精度を得ることは難しい．この問題への対処にも，推定対象個人のみでなく，周辺の他者との相互作用の参照が有効な手段の一つとなる．

姿勢推定における複数人の相互作用を陽に取り扱った研究の歴史は浅く，文献 [20][21] に始まっている．この手法では，近接した人の間に誤って姿勢推定がなされてしまわないように（例：「ある人の左腕」と「その人の左隣の人の右腕」で一人の姿勢が得られてしまう），1）人の前後関係をパラメータに取り入れたパーツ間の遮蔽表現，2）異なる人間が同じ画像領域にパーツを持つこと抑制，を姿勢最適化のコスト関数に導入している．また，文献 [6] では，人と人のインタラクションの種類に応じて PSM における人体パーツの配置確率を調節している．文献 [54] では，より高次のシナリオ情報を仮定することにより，長時間動画における複雑に遮蔽し合う複数人の姿勢推定を実現している．

人の姿勢や行動などの状態は，人に限定されることなく，その姿勢や行動に関わる多くの周辺物や環境との間に関連を持つ．例えば，テニスでスィングする際には，手にはラケットがあり，その前面にはボールがある．文献 [105][33] などでは，人と物との関連性を同時最適化することにより，人の姿勢推定と物体検出の精度を共に向上させている．

[*3] 本稿で紹介する人体姿勢推定法は，2 次元画像上における人体パーツ配置の推定がほとんどであるが，一部 3 次元姿勢まで推定している手法も含まれている．

[*4] 文献 [22] で提案されている Latent SVM は，人体姿勢モデルの識別的学習だけでなく，後述する姿勢と行動の同時最適化や，姿勢・行動・グループ活動などの階層的なモデル化など，隠れ変数を持つ多くのモデル最適化に非常に有効である．

5. 行動認識

5.1 個々の行動認識

高性能な行動認識の実現のためには，識別器などいくつかの研究要素が重要であるが，本稿では認識対象を含んだ人の集団をはじめとした周辺のコンテクストから得られる特徴量に焦点を当てる．行動認識のための特徴量は，前述の姿勢推定でも用いられている HOG [17] ベースの特徴量 [34] の他にも，SIFT [51] などの特徴点群から得られる軌跡ベースの特徴量 [45] や Bag-of-words 特徴量 [55]，身体のの形状・位置と動画中の動き（オプティカルフローなど）を組み合わせたもの [37] などがある．

こうした特徴量の一つとして人体姿勢を元にした特徴量の有効性も多くの文献で示されている [78][52] ．姿勢ベースの特徴量の最大の欠点は，姿勢推定そのものが難しく，それが失敗していると続く行動認識も失敗しやすいということである．この問題に対処するため，姿勢と行動の同時最適化が提案されている（Latent SVM を用いた最適化 [101][76] など）．また，文献 [102] では，3 次元姿勢を例にして推定姿勢に含まれるノイズに対する行動認識の頑健性が示されており，文献 [36] では，オプティカルフロー，HOG，Bag-of-features などの一般的な特徴量に対する姿勢ベース特徴量の優位性が示されている．

5.2 人と物や環境の相互作用を参照した行動認識

行動認識においても，認識対象個人の特徴だけでなく，周辺コンテクストとの相互作用が有用な情報となる．文献 [104] では，各行動に何か物体（例：フルート演奏とフルート）が必要なシーンにおいて，物体とその操作に関わる人体部位の見えを学習することによって，コンテクストを含めた差異の小さな行動認識（例：フルートを持っているだけ，または，フルートを演奏している）を静止画において実現している．物体に加え，前節で挙げた人体姿勢も参照した行動認識 [31]（動画における研究例）も可能である．

この手法 [31] では，まず姿勢推定し，それを行動認識に利用するため，前節で論じた問題と同様に姿勢推定の誤りが行動認識に伝播する可能性があるが，手法 [18][19] は，静止画における姿勢推定，物体検出，行動認識の同時最適化を実現している．同時最適化に，姿勢と物体配置の事例との相関度を特徴量として組み込み，遮蔽や異なる行動間の類似姿勢に対する認識頑健性を向上させた手法もある [33].

上記手法は，一つの物体と人とのインタラクションを取り扱っているが，その他多くのシーン属性との相互作用を参照することも可能である．複数の行動クラスに関与する物体や背景物体を特徴量化した手法には，要素的な人体の動き（例：平行移動，銅の上下運動）やシーン（例：屋内か屋外か）を属性に持つ特徴ベクトルを Laten SVM に適用し

ⓒ 2014 Information Processing Society of Japan

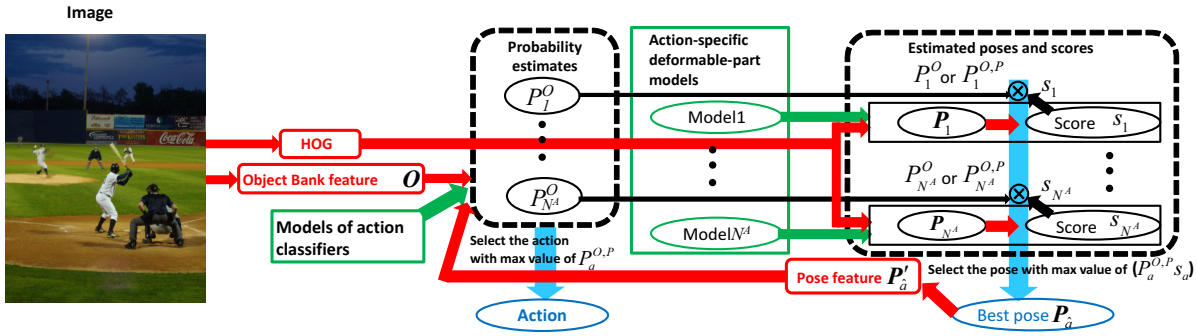


図 6 手法 [85] 概要．シーンの特徴量（“Object bank”）によってシーン識別する．同時に，HOG と各行動に最適化された DPM により姿勢推定する．各 DPM の推定姿勢にシーン識別のスコアを重みにして，ベストな姿勢推定を得る．その姿勢を，シーン特徴量と統合して行動認識に利用する．この姿勢推定と行動認識の反復で，相互に精度向上させる．

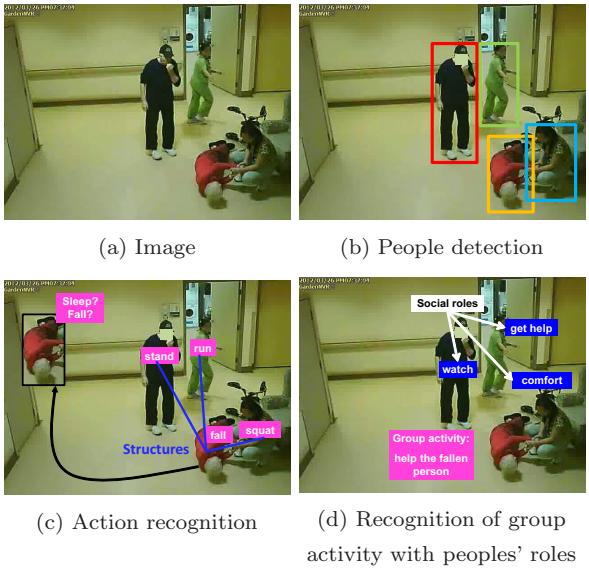
て，各行動クラスの識別に必要な属性を学習する手法 [49]，複数物体の検出結果を特徴ベクトルにした SVM [103]，物体検出結果や背景特徴量（GIST[57]）を入力にした Multi Instance Learning [5] による解法 [35] などがある．

上記手法にも，要素運動や対象検出の誤りが行動認識の失敗を引き起こす問題がある．この問題に対処するため，ここまでに紹介してきた同時最適化ではなく，1）安定に推定可能なシーン特徴量を最初に求め，2）その推定信頼度を重みにして姿勢推定し，3）その結果から得られる姿勢特徴量とシーン特徴量と併用して再度行動認識を行う，という処理の反復によって，誤りを補正しつつ推定精度向上させる手法 [85] が提案されている（図 6）．この手法では，シーン特徴として，シーン識別に広く用いられている GIST [57] のような背景構造を主に表現する特徴量ではなく，画像中における任意の物体の検出スコアの配置から得られる特徴量 [46] を利用した（図 6 中の “Object bank features $\boldsymbol{O}$”）．これにより，理想的には任意の物体の位置関係の表現が可能になる．手法 [85] では，さらに，観測行動ごとに最適化させた姿勢推定モデルを学習することによって，姿勢推定の精度を向上させている．

5.3 人と人の相互作用を参照したグループの活動認識

人と人の相互作用から，個々の行動推定精度の向上も可能である．このためには，2 人以上の行動が組み合わさって意味を成すグループの活動（Group activity）の認識が重要になる．本節では，複数人の行動が一つのグループ活動になる動画を認識対象としている手法について述べる．

文献 [47] では，3.2 節で述べたような対象群の移動軌跡のみを入力として，視野内で起きているグループ活動を認識している．手法 [3] では，グループ活動認識に適した時空間軌跡特徴量の集合を発見している．軌跡推定（時間的な検出対象の対応付け）とグループ活動の同時最適化も提案されている [39][13]．軌跡に加えて人体姿勢もキューとすることにより精度を向上させることも可能である [14]．



(a) Image (b) People detection

(c) Action recognition (d) Recognition of group activity with peoples’ roles

図 7 グループ活動とその中の各人の役割を想定した行動認識（文献 [44] 参照）．入力画像 (a) から，各人をそれぞれ検出することは比較的容易になってきた（図中の (b)）．しかし，床に倒れている女性が寝ているのか倒れているのかは，その女性が写っている局所領域からだけでは判定が難しい（(c) 中の左上）．(d) のように，視野内の 4 人が一つのグループ活動に従事していることを検出できると，床の女性が倒れていて，助けを必要としていることが認識できる．

上記した軌跡以外の特徴量も数多く利用されている．数例を以下に列挙する．文献 [4] では，高次のグループ活動から低位の画像特徴量までを上位層から下位層で表現した分類木を提案している．手法 [38] は，文献 [44] で提案されている Action Context 記述子に加えて，シーンの種類に応じた行動の事前知識を用いることで，記述子が類似する行動間の識別性を向上させている．文献 [58] では，Structured SVM [83] によりインタラクションしている人の相対的な位置と頭部方向，およびそのインタラクションのクラスをモデル化している．文献 [40] では，行動要素の検出の有無を示す 2 値の集合を特徴量とし，その組み合わせから推定できるインタラクティブな行動を隠れ変数とし

ⓒ 2014 Information Processing Society of Japan

た Latent SVM によって，その行動を認識している．多くの手法が SVM による識別ベースであるのに対し，手法 [48] は，比較的混雑している視野内（同様の姿勢の人間が 20 人程度）におけるインタラクティブな行動を，検索ベースのアプローチで検出している．

5.4 グループの活動認識と個の役割認識

前節では比較的単純なグループ活動（例：列に並ぶ，会話する）の認識のみが対象であった．スポーツにおけるチームプレイなど，より複雑なグループ活動においては，グループ活動の認識がより難しくなることに加え，グループ中の各人の役割となる行動認識も必要になってくる．しかし，このようなグループ中の相互作用に関わる役割認識は，ここに注目するだけでは認識困難な行動の識別も可能とする [44]（図 7 参照）．

初期の研究である手法 [70] では，各グループ活動における複数人の動きに関する知識が手動で与えられていた．文献 [44] では，そのような知識を必要としない学習フレームワークが Latent SVM により実現されている．手法 [65] では，個々の役割が学習データとして与えられず，グループ活動のみの Weakly-supervised な問題設定を解いている．また，文献 [43] では，より多い対象数（ホッケーのチームプレイ）におけるグループ活動の時間遷移，および各プレイヤーの役割分担が認識されている．

6. おわりに

人の複雑な属性や行動を認識するため，周辺他者や環境との間の相互作用をモデル化する手法について述べてきた．各種属性推定，群追跡，人体姿勢推定，行動認識，いずれの分野においても基礎的な特徴量やアルゴリズムは共通しているものも多く，関連深い分野であるといえる．

貴重な助言や資料を提供してくれた Greg Mori，Tsuhan Chen に謝意を表する．

参考文献

[1] Ali, S. and Shah, M.: A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis, *CVPR* (2007).
[2] Ali, S. and Shah, M.: Floor Fields for Tracking in High Density Crowd Scenes, *ECCV (2)* (2008).
[3] Amer, M. R. and Todorovic, S.: A chains model for localizing participants of group activities in videos, *ICCV* (2011).
[4] Amer, M. R., Xie, D., Zhao, M., Todorovic, S. and Zhu, S. C.: Cost-Sensitive Top-Down/Bottom-Up Inference for Multiscale Activity Recognition, *ECCV (4)* (2012).
[5] Andrews, S., Tsochantaridis, I. and Hofmann, T.: Support Vector Machines for Multiple-Instance Learning, *NIPS* (2002).
[6] Andriluka, M. and Sigal, L.: Human Context: Modeling Human-Human Interactions for Monocular 3D Pose Estimation, *AMDO*, pp. 260–272 (2012).
[7] Berg, T. L., Berg, A. C., Edwards, J., Maire, M., White, R., Teh, Y. W., Learned-Miller, E. G. and Forsyth, D. A.: Names and Faces in the News, *CVPR (2)* (2004).
[8] Bergtholdt, M., Kappes, J. H., Schmidt, S. and Schnörr, C.: A Study of Parts-Based Object Class Detection Using Complete Graphs, *International Journal of Computer Vision*, Vol. 87, No. 1-2, pp. 93–117 (2010).
[9] Blei, D. M. and Lafferty, J. D.: A Correlated Topic Model of Science, *The Annals of Applied Statistics*, Vol. 1, No. 1, pp. 17–35 (2007).
[10] Chaquet, J. M., Carmona, E. J. and Fernández-Caballero, A.: A survey of video datasets for human action and activity recognition, *Computer Vision and Image Understanding*, Vol. 117, No. 6, pp. 633–659 (2013).
[11] Chen, K., Gong, S., Xiang, T. and Loy, C. C.: Cumulative Attribute Space for Age and Crowd Density Estimation, *CVPR*, pp. 2467–2474 (2013).
[12] Choi, W. and Savarese, S.: Multiple Target Tracking in World Coordinate with Single, Minimally Calibrated Camera, *ECCV (4)* (2010).
[13] Choi, W. and Savarese, S.: A Unified Framework for Multi-target Tracking and Collective Activity Recognition, *ECCV (4)* (2012).
[14] Choi, W., Shahid, K. and Savarese, S.: Learning context for collective activity recognition, *CVPR* (2011).
[15] Comaniciu, D. and Meer, P.: Mean Shift: A Robust Approach Toward Feature Space Analysis, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 24, No. 5, pp. 603–619 (2002).
[16] Comaniciu, D., Ramesh, V. and Meer, P.: The Variable Bandwidth Mean Shift and Data-Driven Scale Selection, *ICCV*, pp. 438–445 (2001).
[17] Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection, *CVPR* (2005).
[18] Delaitre, V., Sivic, J. and Laptev, I.: Learning person-object interactions for action recognition in still images, *NIPS* (2011).
[19] Desai, C. and Ramanan, D.: Detecting Actions, Poses, and Objects with Relational Phraselets, *ECCV* (2012).
[20] Eichner, M. and Ferrari, V.: We Are Family: Joint Pose Estimation of Multiple Persons, *ECCV* (2010).
[21] Eichner, M. and Ferrari, V.: Human Pose Co-Estimation and Applications, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 34, No. 11, pp. 2282–2288 (2012).
[22] Felzenszwalb, P. F., Girshick, R. B., McAllester, D. A. and Ramanan, D.: Object Detection with Discriminatively Trained Part-Based Models, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 32, No. 9, pp. 1627–1645 (2010).
[23] Felzenszwalb, P. F. and Huttenlocher, D. P.: Pictorial Structures for Object Recognition, *International Journal of Computer Vision*, Vol. 61, No. 1, pp. 55–79 (2005).
[24] Fihl, P. and Moeslund, T. B.: Pose Estimation of Interacting People using Pictorial Structures, *AVSS* (2010).
[25] Gallagher, A. C. and Chen, T.: Using Group Prior to Identify People in Consumer Images, *CVPR* (2007).
[26] Gallagher, A. C. and Chen, T.: Estimating age, gender, and identity using first name priors, *CVPR* (2008).
[27] Gallagher, A. C. and Chen, T.: Understanding images of groups of people, *CVPR* (2009).
[28] Ge, W., Collins, R. T. and Ruback, B.: Vision-Based Analysis of Small Groups in Pedestrian Crowds, *IEEE*

ⓒ 2014 Information Processing Society of Japan

Trans. Pattern Anal. Mach. Intell., Vol. 34, No. 5, pp. 1003–1016 (2012).

[29] Guillaumin, M., Mensink, T., Verbeek, J. J. and Schmid, C.: Face Recognition from Caption-Based Supervision, *International Journal of Computer Vision*, Vol. 96, No. 1, pp. 64–82 (2012).

[30] Guo, G., Fu, Y., Dyer, C. R. and Huang, T. S.: Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust Regression, *IEEE Transactions on Image Processing*, Vol. 17, No. 7, pp. 1178–1188 (2008).

[31] Gupta, A., Kembhavi, A. and Davis, L. S.: Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 31, No. 10, pp. 1775–1789 (2009).

[32] Helbing, D. and Molnr, P.: Social force model for pedestrian dynamics, *Physical Review E*, pp. 4282–4286 (1995).

[33] Hu, J.-F., Zheng, W.-S., Lai, J.-H., Gong, S. and Xiang, T.: Recognising Human-Object Interaction via Exemplar based Modelling, *ICCV* (2013).

[34] Ikizler-Cinbis, N., Cinbis, R. G. and Sclaroff, S.: Learning actions from the Web, *ICCV* (2009).

[35] Ikizler-Cinbis, N. and Sclaroff, S.: Object, Scene and Actions: Combining Multiple Features for Human Action Recognition, *ECCV (1)* (2010).

[36] Jhuang, H., Gall, J., Black, M. and Schmid, C.: Towards understanding action recognition, *ICCV* (2013).

[37] Jiang, Z., Lin, Z. and Davis, L. S.: Recognizing Human Actions by Learning and Matching Shape-Motion Prototype Trees, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 34, No. 3, pp. 533–547 (2012).

[38] Khamis, S., Morariu, V. I. and Davis, L. S.: Combining Per-frame and Per-track Cues for Multi-person Action Recognition, *ECCV (1)* (2012).

[39] Khamis, S., Morariu, V. I. and Davis, L. S.: A flow model for joint action recognition and identity maintenance, *CVPR* (2012).

[40] Kong, Y., Jia, Y. and Fu, Y.: Learning Human Interaction by Interactive Phrases, *ECCV (1)*, pp. 300–313 (2012).

[41] Kratz, L. and Nishino, K.: Tracking Pedestrians Using Local Spatio-Temporal Motion Patterns in Extremely Crowded Scenes, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 34, No. 5, pp. 987–1002 (2012).

[42] Kuo, C.-H., Huang, C. and Nevatia, R.: Multi-target tracking by on-line learned discriminative appearance models, *CVPR* (2010).

[43] Lan, T., Sigal, L. and Mori, G.: Social roles in hierarchical models for human activity recognition, *CVPR* (2012).

[44] Lan, T., Wang, Y., Yang, W., Robinovitch, S. N. and Mori, G.: Discriminative Latent Models for Recognizing Contextual Group Activities, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 34, No. 8, pp. 1549–1562 (2012).

[45] Le, Q. V., Zou, W. Y., Yeung, S. Y. and Ng, A. Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, *CVPR* (2011).

[46] Li, L.-J., Su, H., Xing, E. P. and Li, F.-F.: Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification, *NIPS* (2010).

[47] Li, R., Chellappa, R. and Zhou, S. K.: Learning multi-modal densities on Discriminative Temporal Interaction Manifold for group activity recognition, *CVPR* (2009).

[48] Li, R., Porfilio, P. and Zickler, T.: Finding Group Interactions in Social Clutter, *CVPR* (2013).

[49] Liu, J., Kuipers, B. and Savarese, S.: Recognizing human actions by attributes, *CVPR* (2011).

[50] Liu, J., Shah, M., Kuipers, B. and Savarese, S.: Cross-view action recognition via view knowledge transfer, *CVPR* (2011).

[51] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110 (2004).

[52] Maji, S., Bourdev, L. D. and Malik, J.: Action recognition from a distributed representation of pose and appearance, *CVPR* (2011).

[53] Moeslund, T. B., Hilton, A. and Krüger, V.: A survey of advances in vision-based human motion capture and analysis, *Computer Vision and Image Understanding*, Vol. 104, No. 2-3, pp. 90–126 (2006).

[54] Morariu, V. I. and Davis, L. S.: Multi-agent event recognition in structured scenarios, *CVPR* (2011).

[55] Niebles, J. C., Wang, H. and Li, F.-F.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words, *International Journal of Computer Vision*, Vol. 79, No. 3, pp. 299–318 (2008).

[56] Okada, A., Moriguchi, Y., Ukita, N. and Hagita, N.: People Groping by Spatio-Temporal Features of Trajectories, *IAPR MVA* (2013).

[57] Oliva, A. and Torralba, A.: Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, *International Journal of Computer Vision*, Vol. 42, No. 3, pp. 145–175 (2001).

[58] Patron-Perez, A., Marszalek, M., Reid, I. and Zisserman, A.: Structured Learning of Human Interactions in TV Shows, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 34, No. 12, pp. 2441–2453 (2012).

[59] Pellegrini, S., Ess, A. and Gool, L. J. V.: Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings, *ECCV (1)* (2010).

[60] Pellegrini, S., Ess, A., Schindler, K. and Gool, L. J. V.: You'll never walk alone: Modeling social behavior for multi-target tracking, *ICCV* (2009).

[61] Pirsiavash, H., Ramanan, D. and Fowlkes, C. C.: Globally-optimal greedy algorithms for tracking a variable number of objects, *CVPR* (2011).

[62] Poppe, R.: Vision-based human motion analysis: An overview, *Computer Vision and Image Understanding*, Vol. 108, No. 1-2, pp. 4–18 (2007).

[63] Poppe, R.: A survey on vision-based human action recognition, *Image Vision Comput.*, Vol. 28, No. 6, pp. 976–990 (2010).

[64] Ramanan, D., Forsyth, D. A. and Zisserman, A.: Tracking People by Learning Their Appearance, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 29, No. 1, pp. 65–81 (2007).

[65] Ramanathan, V., Yao, B. and Li, F.-F.: Social Role Discovery in Human Events, *CVPR* (2013).

[66] Reddy, K. K., Liu, J. and Shah, M.: Incremental action recognition using feature-tree, *ICCV* (2009).

[67] Rodriguez, M., Ali, S. and Kanade, T.: Tracking in unstructured crowded scenes, *ICCV* (2009).

[68] Rodriguez, M., Laptev, I., Sivic, J. and Audibert, J.-Y.: Density-aware person detection and tracking in crowds, *ICCV* (2011).

ⓒ 2014 Information Processing Society of Japan

[69] Rodriguez, M., Sivic, J., Laptev, I. and Audibert, J.-Y.: Data-driven crowd analysis in videos, *ICCV* (2011).
[70] Ryoo, M. S. and Aggarwal, J. K.: Stochastic Representation and Recognition of High-Level Group Activities, *International Journal of Computer Vision*, Vol. 93, No. 2, pp. 183–200 (2011).
[71] Salti, S., Cavallaro, A. and di Stefano, L.: Adaptive Appearance Modeling for Video Tracking: Survey and Evaluation, *IEEE Transactions on Image Processing*, Vol. 21, No. 10, pp. 4334–4348 (2012).
[72] Sapp, B., Toshev, A. and Taskar, B.: Cascaded Models for Articulated Pose Estimation, *ECCV* (2010).
[73] Scovanner, P. and Tappen, M. F.: Learning pedestrian dynamics from the real world, *ICCV* (2009).
[74] Segal, A. and Reid, I.: Latent Data Association: Bayesian Model Selection for Multitarget Tracking, *ICCV* (2013).
[75] Shao, M., Li, L. and Fu, Y.: What Do You Do? Occupation Recognition in a Photo via Social Context, *ICCV* (2013).
[76] Shapovalova, N., Vahdat, A., Cannons, K., Lan, T. and Mori, G.: Similarity Constrained Latent Support Vector Machine: An Application to Weakly Supervised Action Classification, *ECCV* (2012).
[77] Sigal, L. and Black, M. J.: Measure Locally, Reason Globally: Occlusion-sensitive Articulated Pose Estimation, *CVPR* (2006).
[78] Singh, V. K. and Nevatia, R.: Action recognition in cluttered dynamic scenes using Pose-Specific Part Models, *ICCV* (2011).
[79] Stone, Z., Zickler, T. and Darrell, T.: Toward Large-Scale Face Recognition Using Social Network Context, *Proceedings of the IEEE*, Vol. 98, No. 8, pp. 1408–1415 (2010).
[80] Sun, M., Telaprolu, M., Lee, H. and Savarese, S.: An efficient branch-and-bound algorithm for optimal human pose estimation, *CVPR* (2012).
[81] Tang, S., Andriluka, M., Milan, A., Schindler, K., Roth, S. and Schiele, B.: Learning People Detectors for Tracking in Crowded Scenes, *ICCV* (2013).
[82] Tian, Y., Zitnick, C. L. and Narasimhan, S. G.: Exploring the Spatial Hierarchy of Mixture Models for Human Pose Estimation, *ECCV* (2012).
[83] Tsochantaridis, I., Joachims, T., Hofmann, T. and Altun, Y.: Large Margin Methods for Structured and Interdependent Output Variables, *Journal of Machine Learning Research*, Vol. 6, pp. 1453–1484 (2005).
[84] Ukita, N.: Articulated pose estimation with parts connectivity using discriminative local oriented contours, *CVPR* (2012).
[85] Ukita, N.: Iterative Action and Pose Recognition using Global-and-Pose Features and Action-specific Models, *ICCVWS–HACI* (2013).
[86] Ukita, N.: Simultaneous particle tracking in multi-action motion models with synthesized paths, *Image Vision Comput.*, Vol. 31, No. 6-7, pp. 448–459 (2013).
[87] Ukita, N., Hirai, M. and Kidode, M.: Complex volume and pose tracking with probabilistic dynamical models and visual hull constraints, *ICCV* (2009).
[88] Ukita, N. and Kanade, T.: Gaussian process motion graph models for smooth transitions among multiple actions, *Computer Vision and Image Understanding*, Vol. 116, No. 4, pp. 500–509 (2012).
[89] Ukita, N., Tsuji, R. and Kidode, M.: Real-Time Shape Analysis of a Human Body in Clothing Using Time-Series Part-Labeled Volumes, *ECCV (3)* (2008).
[90] Urtasun, R., Fleet, D. J. and Fua, P.: 3D People Tracking with Gaussian Process Dynamical Models, *CVPR* (2006).
[91] Wang, G., Gallagher, A. C., Luo, J. and Forsyth, D. A.: Seeing People in Social Context: Recognizing People and Social Relationships, *ECCV (5)* (2010).
[92] Wang, X., Hua, G. and Han, T. X.: Discriminative Tracking by Metric Learning, *ECCV (3)* (2010).
[93] Wang, Y. and Mori, G.: Multiple Tree Models for Occlusion and Spatial Constraints in Human Pose Estimation, *ECCV* (2008).
[94] Wang, Y., Tran, D. and Liao, Z.: Learning hierarchical poselets for human parsing, *CVPR* (2011).
[95] Weinland, D., Özuysal, M. and Fua, P.: Making Action Recognition Robust to Occlusions and Viewpoint Changes, *ECCV (3)*, pp. 635–648 (2010).
[96] Weinland, D., Ronfard, R. and Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition, *Computer Vision and Image Understanding*, Vol. 115, No. 2, pp. 224–241 (2011).
[97] Wu, B. and Nevatia, R.: Tracking of Multiple, Partially Occluded Humans based on Static Body Part Detection, *CVPR (1)*, pp. 951–958 (2006).
[98] Xu, Y., Lin, L., Zheng, W.-S. and Liu, X.: Human Re-identification by Matching Compositional Template with Cluster Sampling, *ICCV* (2013).
[99] Yamaguchi, K., Berg, A. C., Ortiz, L. E. and Berg, T. L.: Who are you with and where are you going?, *CVPR* (2011).
[100] Yang, M., Wu, Y. and Lao, S.: Intelligent Collaborative Tracking by Mining Auxiliary Objects, *CVPR (1)* (2006).
[101] Yang, W., Wang, Y. and Mori, G.: Recognizing human actions from still images with latent poses, *CVPR* (2010).
[102] Yao, A., Gall, J., Fanelli, G. and Gool, L. V.: Does Human Action Recognition Benefit from Pose Estimation?, *BMVC* (2011).
[103] Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L. J. and Li, F.-F.: Human action recognition by learning bases of action attributes and parts, *ICCV* (2011).
[104] Yao, B. and Li, F.-F.: Grouplet: A structured image representation for recognizing human and object interactions, *CVPR* (2010).
[105] Yao, B. and Li, F.-F.: Modeling mutual context of object and human pose in human-object interaction activities, *CVPR* (2010).
[106] Yilmaz, A., Javed, O. and Shah, M.: Object tracking: A survey, *ACM Comput. Surv.*, Vol. 38, No. 4 (2006).
[107] Zhan, B., Monekosso, D. N., Remagnino, P., Velastin, S. A. and Xu, L.-Q.: Crowd analysis: a survey, *Mach. Vis. Appl.*, Vol. 19, No. 5-6, pp. 345–357 (2008).
[108] Zhang, X. and Gao, Y.: Face recognition across pose: A review, *Pattern Recognition*, Vol. 42, No. 11, pp. 2876–2896 (2009).
[109] Zhao, W.-Y., Chellappa, R., Phillips, P. J. and Rosenfeld, A.: Face recognition: A literature survey, *ACM Comput. Surv.*, Vol. 35, No. 4, pp. 399–458 (2003).
[110] Zheng, J. and Jiang, Z.: Learning View-invariant Sparse Representations for Cross-view Action Recognition, *ICCV* (2013).
[111] Zheng, W.-S., Gong, S. and Xiang, T.: Associating Groups of People, *BMVC* (2009).

ⓒ 2014 Information Processing Society of Japan