複数フレームの統合による混雑環境での車椅子利用者検出に関する検討

谷川　右京† 　川西　康友†† 　出口　大輔††† 井手　一郎†† 　村瀬　　洋††

秋山　達勇††††

† 名古屋大学 大学院情報科学研究科　〒 464–8601 愛知県名古屋市千種区不老町
†† 名古屋大学 大学院情報学研究科　〒 464–8601 愛知県名古屋市千種区不老町
††† 名古屋大学 情報戦略室　〒 464–8601 愛知県名古屋市千種区不老町
†††† 日本電気株式会社　〒 211–8666 神奈川県川崎市中原区下沼部 1753
E-mail: †tanikawau@murase.m.is.nagoya-u.ac.jp, ††{kawanishi,ide,murase}@i.nagoya-u.ac.jp,
†††ddeguchi@nagoya-u.jp, ††††t-akiyama@df.jp.nec.com

あらまし　近年，車椅子利用者を支援するため，監視カメラ映像から車椅子利用者を検出するシステムが求められている．しかし，車椅子利用者の周囲に多数の歩行者が存在する可能性が高く，歩行者に遮蔽された車椅子利用者を，各フレームから独立に検出することは困難であり，未検出が生じやすい．また，車椅子利用者の見えは歩行者と類似しているため誤検出も生じやすい．本報告では，混雑環境において映像中から高精度に車椅子利用者を検出する手法について検討した結果を報告する．提案手法では，複数フレームにおける車椅子利用者候補の特徴を統合することで，単一フレームのみから車椅子利用者のみを検出することが困難な場合でも正確な検出を目指す．評価実験を行った結果，提案手法は混雑環境において比較手法より高精度に車椅子利用者を検出できることを確認した．
キーワード　車椅子利用者検出，物体追跡，CNN

A study on wheelchair-users detection in a crowded scene by integrating multiple frames

Ukyo TANIKAWA†, Yasutomo KAWANISHI††, Daisuke DEGUCHI†††,

Ichiro IDE††, Hiroshi MURASE††, and Tatsuo AKIYAMA††††

† Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
†† Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
††† Information Strategy O ce, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
†††† NEC Corporation
1753 Shimonumabe, Nakahara-ku, Kawasaki-shi, Kanagawa, 211–8666 Japan
E-mail: †tanikawau@murase.m.is.nagoya-u.ac.jp, ††{kawanishi,ide,murase}@i.nagoya-u.ac.jp,
†††ddeguchi@nagoya-u.jp, ††††t-akiyama@df.jp.nec.com

Abstract　In recent years, to support wheelchair users, there has been an increasing demand for a system to detect wheelchair users from a surveillance video. However, since many pedestrians are often walking around wheelchair users in a crowded scene, miss-detections easily occur when they are occluded in a frame. False detections also easily occur since wheelchair users appear similar to pedestrians. We report on a method to detect wheelchair users accurately from a video observing a crowded scene. The proposed method integrates multi-frames features of wheelchair-user candidates, which enables to detect wheelchair users accurately that cannot be detected without false detections from a single frame. As a result of an experiment, the proposed method achieved the highest accuracy compared with other methods.
Key words　Wheelchair-user detection, object tracking, CNN

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2018　by　IEICE

1. まえがき

近年，多くの公共施設がバリアフリー化され，車椅子利用者が単独でも公共施設を利用可能な環境が整備されつつある．そのような取り組みは，身体の不自由な人々が積極的に参加・貢献することができる共生社会を実現するために重要となる．しかし，車椅子利用者が他者の援助を必要とする場面は依然として多い．そのような場面で必要に応じて適切に支援するためには，事前に車椅子利用者の位置を把握しておくことが重要となる．そのため，既に多くの施設に設置されている監視カメラの映像を用いて，自動的に車椅子利用者を検出するシステムへの重要が高まりつつある．

しかし，鉄道駅などの混雑しやすい環境では，しばしば多くの歩行者が車椅子利用者の周囲に存在する．混雑した環境中を移動する車椅子利用者の例を図 1 に示す．一般的に，車椅子利用者の全高は歩行者と比較して低い．そのため，正面や斜め上に監視カメラが設置されている場合，混雑環境下では周囲の歩行者による遮蔽が原因となり，車椅子利用者の全身を観測できない場面が多い．そのような場面では検出が困難なため，未検出数が増加するという問題がある．また，車椅子利用者と歩行者の上半身の画像における見えは類似している．同一映像に含まれる車椅子利用者と歩行者の例を図 2 に示す．そのため，歩行者に対する誤検出数が増加しやすい．本報告では，混雑環境における高精度な車椅子利用者の検出を目的とし，未検出と誤検出が少ない検出手法について検討した結果を報告する．

事前に検出対象について学習した検出器を用いる手法の多くは，1 枚の画像から検出を行なう．そのような手法を混雑環境における車椅子利用者の検出に適用した場合，遮蔽があるフレームにおいて検出が困難となり，未検出数が増加する．この問題への対策として，車椅子利用者か否かの識別をするしきい値を緩めることで未検出を低減することが考えられる．しかし，そのトレードオフとして誤検出数が増加する．特に混雑環境における車椅子利用者の検出の場合，周囲に歩行者が同時に多数存在するため，誤検出数が増加する．そのため，混雑環境では単一フレームから車椅子利用者を正確に検出するのは困難であるといえる．本研究では，単一フレームから車椅子利用者を識別するのは困難でも，複数フレームの情報を用いることで情報量が増え，識別しやすくなる点に着目する．提案手法では，検出する候補の画像系列を抽出し，それらの特徴を統合して車椅子利用者か否かを識別することで，未検出と誤検出を低減する．

複数フレームの情報を用いて検出を行なうために，提案手法は（1）検出のために車椅子利用者候補を追跡し，（2）追跡して得られた画像系列を用いて車椅子利用者か否か識別する．（1）では，検出器により得られた初期検出結果を時間方向に対応付けることで追跡する．混雑環境では遮蔽により追跡系列が途切れやすいため，提案手法では，遮蔽による検出スコアの低下に対応するため，それらを補う大域的最適化に基づく追跡を行なう．（2）では，Convolutional Neural Network（CNN）を用いて追跡系列の識別を行なう．ここで，追跡系列には遮蔽や位置ずれが含まれている可能性があり，高精度に識別を行なうには

図 1　混雑した環境中を移動する車椅子利用者の例

図 2　同一映像に含まれる車椅子利用者と歩行者の例

系列全体から有効な特徴のみを抽出する必要がある．提案手法では，フレームごとに抽出した特徴を，時空間の情報を用いて算出した重みを用いて統合する．統合のための重みを計算するため，フレームごとに抽出した特徴に対して 3D Convolution を適用する．

以降，2 節では本研究の関連研究について述べる．3 節では，提案手法である複数フレームの統合による混雑環境における車椅子利用者検出の検出手法について説明する．4 節では，提案手法の有効性について調査した評価実験とその結果について述べ，考察を加える．最後に 5 節で，まとめと今後の課題について述べる．

2. 関連研究

既存の物体検出手法の多くは，1 枚の画像に対して検出処理を行うものである．Felzenszwalb らは，物体のモデルを部位の集合として表現する Deformable Part Models（DPM）を用いた物体検出手法を提案している [1]．DPM は，各部位の詳細な形状や位置を考慮するため姿勢変動に頑健である．Girshick らは，画像から物体候補領域を抽出した後，その領域から CNN 特徴を抽出し識別を行なう R-CNN [2] を提案している．Girshick らはさらに R-CNN を改良し，画像全体から CNN 特徴を抽出した後，ROI Pooling により候補領域ごとの特徴を抽出することで計算コストを削減した Fast R-CNN [3] を提案している．Ren らは Fast R-CNN をさらに改良し，候補領域抽出を CNN により行なう Faster R-CNN [4] を提案している．Faster R-CNN は高速かつ高精度な物体検出が可能であるが，対象物体が遮蔽されると候補領域抽出に失敗しやすく，検出精度が低下するという問題がある．

Myles らは，車椅子利用者に特化した検出手法を提案している [5]．この手法では，Hough 変換を用いて車椅子の車輪を，色特徴を用いて車椅子利用者の顔をそれぞれ検出し，車椅子利用者の 3 次元姿勢情報を構築することで検出を行なう．しかし，事前にカメラキャリブレーションを正確に行なう必要があるため，利用できる環境が限られる．Huang らは，単一の固定カメラにより撮影された映像から車椅子利用者を検出する手法を提案している [6]．この手法では，HOG 特徴量などの局所特徴量とカスケード化した AdaBoost による識別器を用いて検出を行

なう．しかし，この手法は車椅子利用者の遮蔽を考慮していないため，混雑環境では高精度に検出することができない．

3. 複数フレームの統合による車椅子利用者の検出手法

本手法は 3 つの段階に分けられる．はじめに，事前に学習した車椅子利用者の検出器を用いて入力映像の各フレームについて初期検出を行なう．次に，初期検出により得られた検出候補の組み合わせについて大域的最適化問題を解くことで追跡を行い，検出候補の画像系列を抽出する．最後に，得られた追跡系列を CNN により車椅子利用者か否か識別する．

3.1 車椅子利用者候補の初期検出

入力映像の各フレームから，ベースライン検出器を用いて車椅子利用者候補を検出する．ベースライン検出器には Faster R-CNN [4] を用いる．一般に，車椅子利用者が遮蔽されたフレームでは，検出スコアが小さくなり，検出に失敗する場合が多い．そのような場合でも候補として検出するために，検出スコアのしきい値を緩めに設定する．これにより，誤検出を含め多くの車椅子利用者候補が得られる．

Faster R-CNN を学習する際，検出対象のクラスに歩行者を別クラスとして追加する．これにより，歩行者に対する誤検出を低減する．検出結果として，車椅子利用者に対応する出力結果のみを用いる．

3.2 検出候補の追跡

3.2.1 最適化問題としての定式化

初期検出により得られた車椅子利用者の検出候補に対して追跡を行い，検出候補の画像系列を求める．提案手法では，追跡を大域的最適化問題として定式化し，最小費用流問題として解く枠組み [7] をもとに追跡を行なう．この枠組みでは，1 つの検出候補 x_i につき 2 つの頂点 u_i，v_i を作成しネットワークを構築する．辺は以下の 4 箇所に作成され，各辺についてコストを算出する．

- 始点 s から頂点 u_i（コスト C_{si}）
- 頂点 u_i から頂点 v_i（コスト C_i）
- 頂点 v_i から頂点 u_j（コスト C_{ij}）
- 頂点 v_i から終点 t（コスト C_{it}）

算出されたコストの合計が最小となるフローを計算し，それらに対応する検出候補の系列を追跡系列として出力する．C_{si} および C_{it} は，検出対象がそのフレームで最初に現れる，もしくは最後に現れる確率に基づくコストである．C_i はその検出候補の信頼度に基づくコストであり，C_{ij} は 2 つの検出候補の類似度に基づくコストである．最小費用流 T は，以下の目的関数を最小化することにより求められる．

$$T = \arg\min_{\mathcal{T}} \Big(\sum_i C_{si} e_{si} + \sum_{i,j} C_{ij} e_{ij} + \sum_i C_{it} e_{it} + \sum_i C_i e_i \Big) \tag{1}$$

ここで e_{ab} は，あるフロー $\mathcal{T}_k \in \mathcal{T}$ が頂点 (a, b) の間に流れる場合に 1，流れない場合に 0 となる．最小費用流問題を解くには流量を与える必要があるが，流量に相当する，入力映像中の検出対象数は未知である．そのため，フローの量が 0 から上限値までの場合の最小費用流をそれぞれ求め，合計コストが最小である場合の流量を選択する．

3.2.2 提案手法におけるネットワークの構築方法

提案手法では，式（1）の各コストを以下のように算出する．

$$C_{si} = C_{it} = -\log \frac{1}{a \cdot r} \tag{2}$$

$$C_i = \log \frac{1-\beta_i}{\beta_i} \tag{3}$$

$$C_{ij} = -\log \left(\max \left(0, \mathrm{IoU}_{ij} - \frac{f_{ij}-1}{F} \right) \right) \tag{4}$$

ここで r は入力映像のフレームレート，a は定数であり，それぞれが大きいほど C_{si} は大きくなる．β_i は検出候補 x_i の信頼度であり，β_i が小さいほど C_i は大きくなる．β_i は，ベースライン検出器による検出スコア S_i をシグモイド関数により $[0, 1]$ に正規化した値である．

$$\beta_i = \frac{1}{1 + \exp(-bS_i - c))} \tag{5}$$

ここで b はシグモイド関数のゲイン，c は β_i の大きさを制御する定数である．遮蔽された検出候補の検出スコアは小さくなるため，c を小さくすることで β_i が大きくなるように調整する．IoU_{ij} は検出候補 x_i と x_j の Intersection over Union（IoU）である．矩形 b_i と矩形 b_j の IoU は以下の式で算出される．

$$\mathrm{IoU}_{i,j} = \frac{|b_i \cap b_j|}{|b_i \cup b_j|} \tag{6}$$

ここで $|\cdot|$ は領域を構成する画素数である．IoU が小さいほど，位置の連続性が小さいため C_{ij} が大きくなる．f_{ij} は x_i と x_j の間のフレーム数，F は辺を作成する最大のフレーム間隔であり，f_{ij} が大きいほど時間の隔たりが大きいため C_{ij} が大きくなる．

また，隣接フレーム間でのみ頂点 v_i から u_j への辺を作成する場合，遮蔽により初期検出に失敗したフレームが存在すると追跡系列が途切れてしまう．そこで，$1 \leqq f \leqq F$ フレーム離れた検出候補に対応する各頂点の組み合わせについて辺を作成する．これにより，フレームを越えた追跡を可能にし，追跡の中断を抑制する．

3.3 追跡系列の識別

CNN を用いて，得られた追跡系列を車椅子利用者か否か識別する．識別する際，系列全体から有効な特徴のみを抽出するため，フレームごとに抽出した特徴から時空間の重みを計算し，これを用いて特徴を統合する．CNN の構造を図 3 に示す．入力は 128×128 の RGB 画像系列，出力は車椅子利用者とそれ以外のクラスの尤度である．入力画像の系列長は 8 とし，追跡系列を 8 フレームずつ分割して入力する．

CNN の図 3（ア）の部分では，入力映像のフレームごとの特徴抽出を行なう．この部分は 5 つの Convolution 層と 4 つの Max-pooling から構成され，フレームごとに特徴マップを計算する．図 3（イ）の部分では，複数フレーム情報を活用して識別を行なうために，各フレームの特徴マップを時間方向で

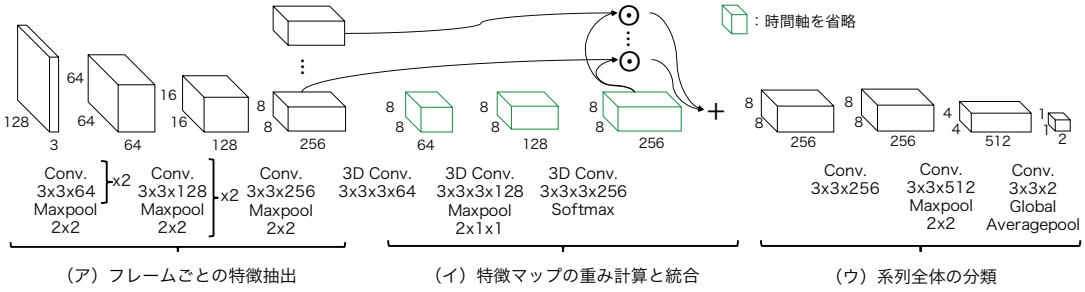


図 3　CNN の構造

統合する．統合のための重みを計算するため，特徴マップの系列に対して 3 つの 3D Convolution 層と時間軸に対する 1 つの Max-pooling を計算する．その出力の各次元に対し時間軸で Softmax 関数を適用し，時間軸での総和が 1 であり $[0, 1]$ に正規化された値を重みとする．算出した重みと特徴マップ系列の要素ごとの積を計算し，時間軸で総和をとることで複数フレームから得た特徴マップを統合する．

CNN の図 3（ウ）の部分では，統合した特徴マップを用いて系列全体の分類を行なう．この部分は，3 つの Convolution 層と 1 つの Max-pooling，1 つの Global Average-pooling から構成される．パラメータ数を削減し学習データが少なくても学習できるようにするため，Fully-connected 層の代わりに Convolution 層と Global Average-pooling を用いる．最後の Convolution 層の出力に対して Global Average-pooling を適用し，1×1 の 2 チャネルにする．各チャネルはそれぞれ車椅子利用者とそれ以外のクラスに対応しており，最後に Softmax 関数を用いて尤度を計算する．

CNN を学習する際，損失関数として交差エントロピーを用い，誤差逆伝播法により学習を行う．各 Convolution 層の活性化関数には Leaky ReLU [8] を用いる．学習を安定化させるため，各層の出力には Batch Normalization [9] を適用する．パラメータの初期値は乱数で決定し，過学習を抑制するために L2 正則化を加える．

4. 評 価 実 験

4.1 データセット

提案手法の有効性を確認するため，混雑環境で撮影された映像を用いた車椅子利用者の検出実験を行なった．本実験では，学習データセットとして車椅子利用者を含む映像 15 本を使用した．また，評価用データセットとして映像 23 本を使用した．各データセットの詳細を表 1 に示す．評価用データセットの各映像には歩行者が多数含まれており，それらによる車椅子利用者の遮蔽がある．

4.2 評 価 方 法

検出精度を評価する際，各手法で出力された検出枠と正解枠との IoU（式 (6)）が 0.5 以上である場合を正検出とし，そうでない場合を誤検出とした．評価指標には Free-response Receiver Operating Characteristic（FROC）曲線を用いた．FROC 曲線は横軸に画像 1 枚あたりの誤検出数（False Positives Per Image，FPPI），縦軸に検出率をとる曲線である．FROC 曲線が図の左上に位置するほど高精度であることを表す．FROC 曲線は検出結果のスコアのしきい値を変化させることで描画した．また，F 値の最大値とその時の再現率および適合率についても評価した．

本実験では，以下の手法の検出精度を比較した．

- 比較：Faster R-CNN（FRCNN）
- 比較：FRCNN ＋追跡＋ CNN（単一フレーム）
- 比較：FRCNN ＋追跡＋ CNN（特徴マップを単純平均）
- 提案：FRCNN ＋追跡＋ CNN（3D Conv. による加重平均）

「CNN（単一フレーム）」は，提案手法において追跡系列の識別に使用する CNN から，特徴マップの統合部分を除いた手法である．すなわち，1 フレームずつ検出を行う．「CNN（特徴マップを単純平均）」は，提案手法において追跡系列の識別に使用する CNN で，特徴マップを単純平均して統合する手法である．

4.3 学 習 方 法

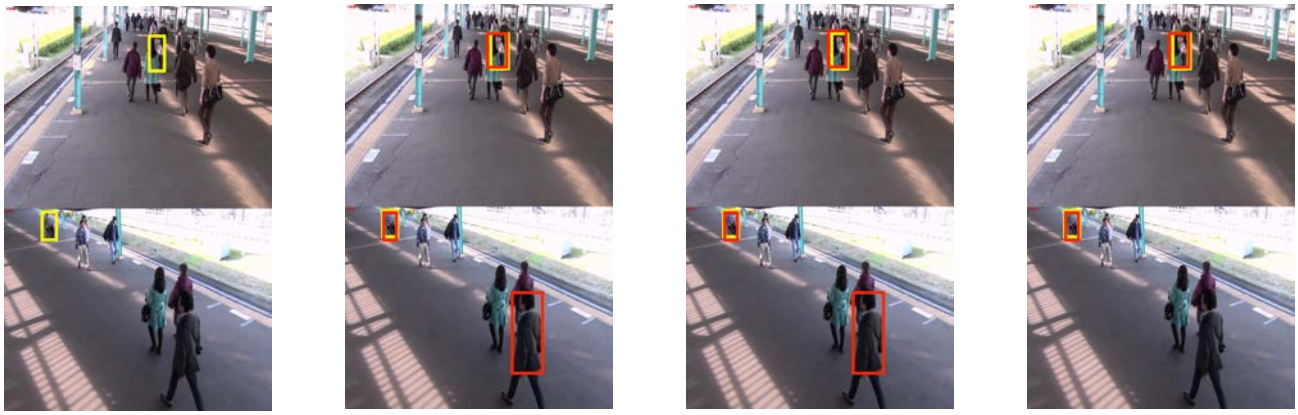
4.3.1 ベースライン検出器

ベースライン検出器として用いる Faster R-CNN の学習では，歩行者の学習データとして，VOC2007 [10] に含まれる歩行者画像 1,025 枚を使用した．Faster R-CNN の実装には，Girshick らによって公開されている実装 [4] を利用した．

4.3.2 CNN

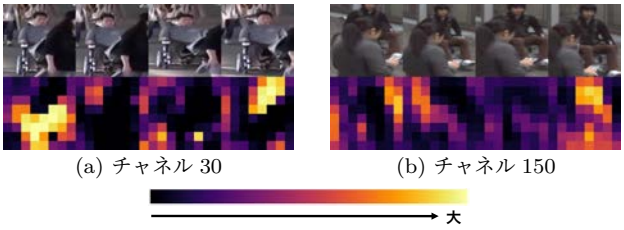
CNN を学習するために，学習用映像から学習用画像系列を生成した．エポックごとに，映像の各フレームを開始位置としてランダムに画像系列を生成した．抽出するフレーム数は CNN の入力系列長と同様に 8 とし，フレーム間隔は乱数で決定した．ポジティブサンプルの系列については，車椅子利用者の正解領域に 15 %のパディングを加えて切り出した．ネガティブサンプルの系列については，最初のフレームの背景領域からランダムに小領域を切り出し，その後のフレームで見えに基づき追跡することで系列を生成した．追跡には Henriques らの手法 [11] を用いた．

生成する際，学習データのパターンを増やすため表 2 に示す Data augmentation を行った．遮蔽の追加に用いる歩行者の画像とそのマスクとして，Daimler Pedestrian Segmentation Benchmark Dataset [12] に含まれる画像 785 枚を用いた．これらの画像から歩行者領域のみを切り出し，画像系列に対してランダムな位置・大きさ・向きに重ねることで遮蔽を加えた．



(a) FRCNN　(b) FRCNN ＋追跡＋ CNN（単一フレーム）　(c) FRCNN ＋追跡＋ CNN（特徴マップを単純平均）　(d) FRCNN ＋追跡＋ CNN（3D Conv. による加重平均）

図 6　車椅子利用者検出の結果例（黄枠：真値，赤枠：各手法における F 値最大時の検出結果）



(a) チャネル 30　(b) チャネル 150

(c) 重みの値のカラーバー

図 7　3D Convolution により算出した重みの可視化例

は，歩行者により遮蔽されている領域の重みが小さい一方，車椅子利用者が観測できる領域の重みが大きくなっている．これらの結果から，提案手法の CNN は識別に有効な次元および時間に注目していることがわかる．しかし，歩行者や背景領域に対して大きな重みが算出される場合も多い．CNN への新たな層の導入や学習における損失関数の工夫など，ネットワーク構造や学習方法について検討が必要である．

5. むすび

本報告では，混雑環境における車椅子利用者の手法について検討した結果をまとめた．提案手法は，混雑環境において単一フレームからの検出では未検出と誤検出が増加しやすいという問題に対し，複数フレーム情報を活用するというアプローチをとった．複数フレーム情報を活用するために，初期検出器により得られた検出候補を追跡し，CNN により複数フレームの特徴を統合して，その系列に含まれる人物が車椅子利用者か否かを識別した．評価実験により，混雑環境において比較手法より高精度な検出が可能であることを確認した．

今後の課題として，追跡におけるネットワークの構築方法の改善や，CNN の構造や学習方法の工夫が挙げられる．

謝辞　本研究の一部は，科学研究費補助金による．

文　献

[1] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.32, no.9, pp.1627–1645, Sept. 2010.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” Proceedings of the 27th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.580–587, June 2014.

[3] R. Girshick, “Fast R-CNN,” Proceedings of the 2015 IEEE International Conference on Computer Vision, pp.1440–1448, Dec. 2015.

[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” Proceedings of the 28th International Conference on Neural Information Processing Systems, pp.91–99, Dec. 2015.

[5] A. Myles, N.D.V. Lobo, and M. Shah, “Wheelchair detection in a calibrated environment,” Proceedings of the 5th Asian Conference on Computer Vision, pp.706–712, Jan. 2002.

[6] C.-R. Huang, P.-C. Chung, K.-W. Lin, and S.-C. Tseng, “Wheelchair detection using cascaded decision tree,” IEEE Transactions on Information Technology in Biomedicine, vol.14, no.2, pp.292–300, Mar. 2010.

[7] L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-object tracking using network flows,” Proceedings of the 21st IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.1–8, Aug. 2008.

[8] A.L. Maas, A.Y. Hannun, and A.Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” Proceedings of the ICML Workshop on Deep Learning for Audio, Speech, and Language Processing, pp.1–6, June 2013.

[9] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” Proceedings of the 32nd International Conference on Machine Learning, pp.448–456, July 2015.

[10] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results,” `http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html`.

[11] J.F. Henriques, R. Caseiro, P. Martins, and J. Batista, “Exploiting the circulant structure of tracking-by-detection with kernels,” Proceedings of the 12th European Conference on Computer Vision, pp.702–715, Oct. 2012.

[12] F. Flohr and D. Gavrila, “Pedcut: An iterative framework for pedestrian segmentation combining shape models and multiple data cues,” Proceedings of the 24th British Machine Vision Conference, pp.66.1–66.11, Jan. 2013.