定点カメラ映像における群衆の挙動予測のための歩行者追跡に関する検証

A Validation of Pedestrian Tracking for Predicting Crowd Behavior in Fixed-point Camera Videos

北海道大学工学部　○学生員　　鴨藤功武 (Isamu Kamoto)
北海道大学　　正会員　髙橋翔 (Sho Takahashi)
北海道大学　　正会員　萩原亨 (Toru Hagiwara)

1. はじめに

近年，交通事故の発生件数や死傷者数は減少傾向にある．交通事故発生件数は平成 16 年の 95 万件をピークに減少し，令和元年には 59 年ぶりに 40 万件を下回った．

しかしながら，減少傾向にあるといっても，未だ重大な社会問題であることに変わりはない．交通事故は，道路構造物や家屋・車両の破損や，重傷，場合によっては命を落とすことになるのはもちろん，加害者・被害者の親族や知人，事故の目撃者らに精神的なダメージを与え，事故によって渋滞が発生することで経済的な損失につながる．平成 24 年 3 月に公表された調査結果 [1]では平成 21 年度の交通事故による損失額は 6 兆円以上とされており，交通事故発生件数を 0 に近づけることは重要な社会的な課題であるといえる．

交通事故の発生要因の一つに死角がある．道路構造物や建造物，自動車の車体が遮蔽物となることで死角ができる．車体によって生じる死角については，自車周辺の死角をカバーするために，車載カメラやミリ波レーダといった車載センサが活用されている．ドライバ視点では未だに死角のままであるが，車載センサによって認知することで，自車周辺の死角はほとんどカバーされる．しかし，車載センサで周辺状況を把握できる範囲は，デバイスの性能による制約を除けば，自車から視認できる範囲のみである．死角の状況を把握し，死角に存在する対象を認識することで事故率の低下につながり，さらにその対象の挙動を予測できれば事故の回避につなげられる．そのため，道路構造物や建造物が遮蔽物となって生じる死角の状況を把握し予測する必要がある．

元データの取得から情報の伝達までの一連の流れは以下の通りである．

① 交通空間利用者の座標データの取得・集積
② 数秒先の予測
③ 各利用者に情報を提示・伝達

ここで①交通空間利用者の座標データの取得・集積については，車載センサや映像データを検出する様々な手法が研究されている．高精度に検出できる手法も確立されているが，検出漏れや検出ミスは課題として残っている．さらに，検出だけでは瞬間的な座標データにしかならないが，同一の物体を追跡することで時系列の座標データとして取得できる．高精度に追跡できる手法が確立されているが，検出が外れたときや物体が遮蔽物などで遮られてしまったときに追跡しきれないなどの課題がある．

②の数秒先の予測は，座標や軌跡がすべて揃っている

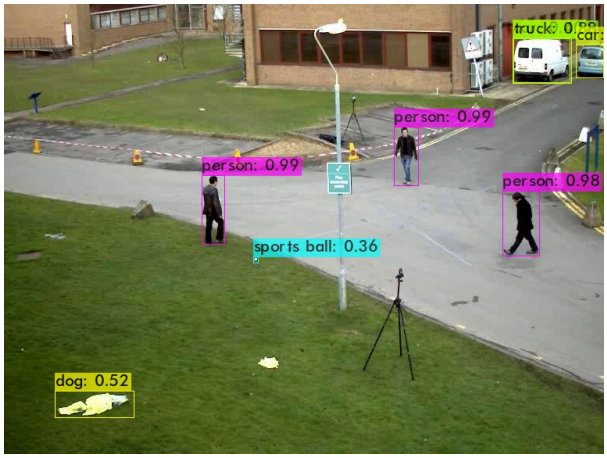


図-1 YOLOv4 による物体検出

前提で，歩行者を対象にした数秒先の予測を行う手法に関する研究 [2),3)]が種々行われている．これらの共通項として，予測対象の個々をそれぞれ別々に扱っていることが挙げられる．

ここで①で挙げられる課題を考慮すると，個々の対象物それぞれに精度良く予測することは困難であり，新たな考え方を導入する必要がある．

②で得られた予測結果をもとに各交通空間利用者に情報を伝達することで，より安全な利用が可能になる．しかしながら，予測結果をもとにした死角の情報をすべて伝達してしまうと，情報を処理しきれない可能性があると同時に，どの情報が真に重要なのかわからない．そのため，重要でない情報を省いて伝達する必要がある．

そこで本研究では，交通空間の状況を把握・予測し，各交通空間利用者にとって重要になりうる情報を伝達することを大きな目標とする．これに向けてまず，本稿では，交通空間利用者の位置を取得する情報源として定点カメラの映像を考える．これを対象に，既存の物体検出・物体追跡の手法を用いて映像データから人物の検出と追跡を行い，その性能を確認し，より正確な予測を行う方法を模索する．

2. YOLOv4 および Deep SORT による歩行者の検出・追跡

本章では，YOLOv4[4)]による物体検出および Deep SORT[5)]による物体追跡について説明する．

動画像を入力として，物体の検出・追跡を行う．物体の検出は高速で高精度な YOLOv4 で行う．学習済みの物体を検出すると，図-1 のように Bounding box を表示す

(a)No.1　(b)No.2　(c)No.3

(d)No.4　(e)No.5　(f)No.6

(g)No.7　(h)No.8

図-2 YOLOv4 による検出結果

る．同時に，Bounding box 内に存在する物体のラベルと存在する確率を出力する．

物体の追跡は Deep SORT を用いて行う．Deep SORT は深層学習を用いた物体追跡の手法で，YOLOv4 などで出力された Bounding bo

x に前後のフレーム間で一貫した

表-1 YOLOv4 による検出数，未検出数，エラー数

No.	検出数	未検出数	誤検出数
1	20	15	0
2	25	5	0
3	24	≥20	0
4	48	≥30	1
5	64	≥20	1
6	56	11	1
7	52	≥20	1
8	25	≥30	1

TrackID を指定する．

3. 実験

本章では，映像データから人物を対象にした検出と追跡の性能を確認するために行った実験について説明する．映像データを入力し，人物を対象に検出と追跡を行い，その結果が可視化された映像を出力する実験を行った．また，本実験では，VIRAT Video Dataset[6](30fps) と MOT20[7](30fps)の 2 つのオープンデータセットを使用した．ただし，MOT20 は映像ではなく連続した画像のデータセットのため，画像から映像に変換する前処理を行った．

また本実験では，同一人物判定を行うための閾値を 0.5 とし，各 ID の特徴ベクトルを保持するフレーム数を 100 とした．さらに，割り当てられなかった ID が削除されるフレーム数を 70，新規の ID が有効化されるフレーム数を 2 とした．

3.1 定点カメラ映像における検出

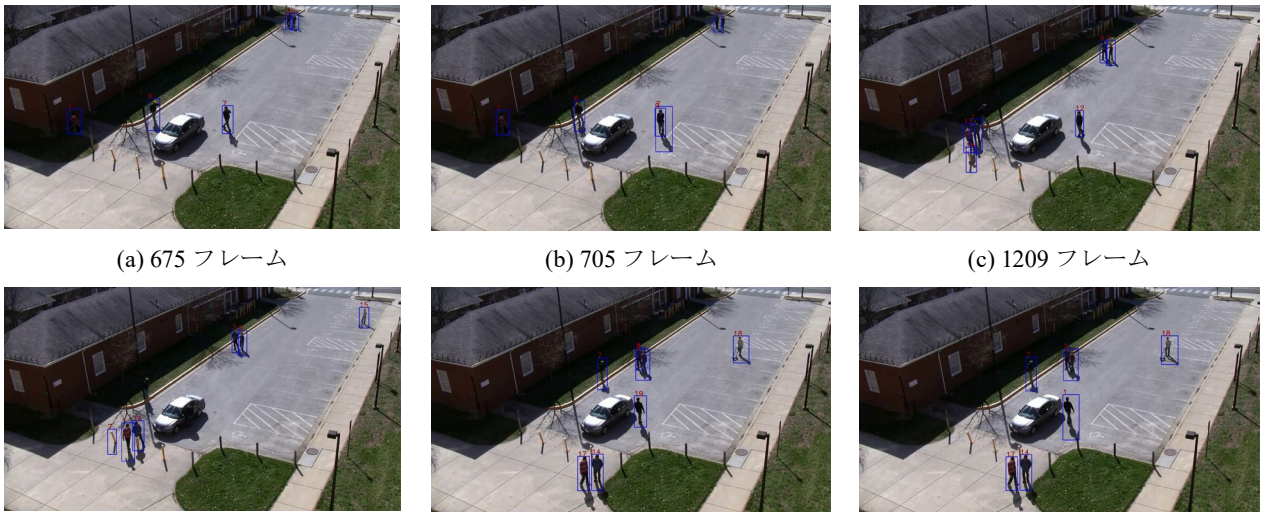
本節では，YOLOv4 による検出性能を確認するための実験と結果について説明する．

本実験では，MOT20 の Test Set および Training Set 計 8 本の映像データからそれぞれ 1 フレームを抽出し，物体検出の対象とした．

YOLOv4 による検出結果を図-2 に，人物として検出した数，未検出数，誤検出数を表-1 に示す．

未検出数は検出されていないが，明らかに人物だと断定できるものを人手によってカウントした．誤検出数は，明らかに人物でない対象を人物として検出している数をカウントした．

図-2 より，No.4-8 には共通して，明らかにサイズの大きい Bounding box があることがわかる．また，人口密度が比較的疎な No.1,2,6 は未検出が比較的出にくい結果となり，密な状態にある No.3-5,7,8 では未検出が多い結果となった．さらに No.4,7 では，人がいるにも関わらず，



(a) 675 フレーム　(b) 705 フレーム　(c) 1209 フレーム

(d) 1323 フレーム　(e) 1533 フレーム　(f) 1536 フレーム

図-3 Deep-SORT-YOLOv4 による追跡結果

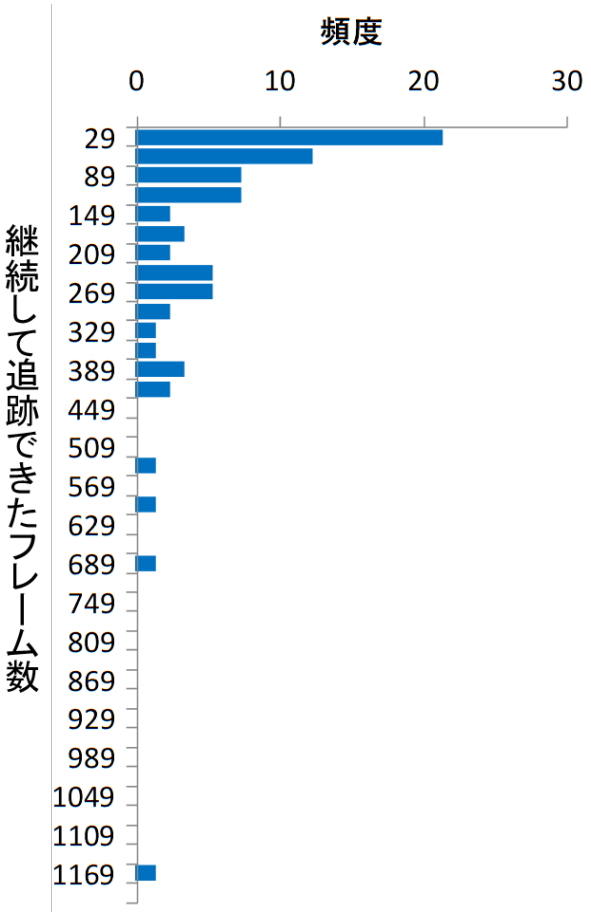


図-4 継続して追跡できたフレーム数のヒストグラム

1 人も検出されていない領域がある．No.4 については，夜の暗さで色の識別が難しいことで検出できないと考えられる．No.3 については，手前の集団に遮られてしまうことで奥の集団を検出できないことが考えられる．

3.2 定点カメラ映像における追跡

本節では，Deep SORT による追跡性能を確認するための実験と結果について説明する．実験には，VIRAT Video Dataset の Sample Dataset 内映像から 1 シーン(1:13-2:36)を切り抜いた映像を入力データとして使用した．本実験では，対象が遮蔽物に隠れたり影に入ったりしたときに，同一人物と判定されやすいように変更した上で追跡を行った．追跡結果を図-3 に示す．また，各フレームにおける ID を出力し，図-3 の追跡結果と照らし合わせ，同一の対象を継続して追跡できたフレーム数をカウントした．ヒストグラムとして図-4 に示す．

図-3(e),(f)のように計 7 名の人物が登場する映像データを入力し，1~36 の ID が出力され，同一の対象を追跡した回数は計 77 回だった．

図-3(a),(b)では，検出ミスと追跡ミスを確認した．図-3(a)には，ID7 の人を確認できる．図-3(b)では，その人の影を含めて一人の人と検出しており，なおかつ影を含んで検出した方を ID7 として追跡し，元々ID7 だった人には ID9 が付与されており，検出・追跡双方に誤りが生じている．

また，図-3(c)では設置物を人と誤検出し新たな ID14 を付与している．図-3(d)では，図-3(c)で誤検出した設置物に対し，ID7 が付与されている．

図 -3(e),(f) では，7 人全員が検出されている．図-3(e)中の ID19 は，図-3(f)には ID1 となっている．ID1 は 50 フレーム前まで別の人に付与されていたため追跡ミスが生じている．

図-4 は，各フレームに存在する ID と追跡結果を照らし合わせて，同一の対象を追跡できているかどうか確認し，追跡が何フレームに渡って行われたのかを表したヒストグラムである．全 77 回の追跡の内，2 秒以上追跡できたのは 44 回．3 秒以上追跡できたのは 37 回だった．また，計 36 の ID に対して 77 回の追跡を行っていることから，ID の付与におけるミスが多いことが分かった．

一方，図-3(a)における ID1 と 7，図-3(e)における ID17 とこの後 ID1 となる ID14 の 4 ケースで，長い時間追跡したケースも確認できた．

4. 考察

本章では，3.で得られた結果を踏まえて，次のステップである予測をより正確にするための手法を模索する．まず検出の性能については，図-2 に示されるように，対象が密集している状態や対象が他の対象や遮蔽物によって遮蔽されている状態での検出漏れが多いことが分かった．次に追跡の性能については，上記の検出性能に左右された上で，違う ID を付与するケースが多くあることを図-3 から確認できた．また，追跡できず新規の ID を付与するケースの多くは，再度，長時間の追跡ができずに新規の ID になることから，追跡時間が 1 秒未満の頻度が最多となることが予想される．図-4 より，予想が正しいことが確認できる．一方で，長時間追跡できた事例には，遮蔽物に遮られたり歩行者同士ですれ違ったりすることが少ない傾向があった．

以上のことから，交通空間利用者を個々に検出し追跡することは非常に困難なことであるため，個々を対象に予測を行うことは困難といえる．そこで，検出・追跡結果がある程度欠けていても予測できるようにするための手法が必要になる．一人ひとりを追跡できないのであれば，複数の個人で形成される“群”を考えるべきではないかと考える．ただし，群をどう定義するかによって，その特性は大きく変わってしまう．今後，この方法を具体化することは，本研究の課題である．

5.まとめ

本稿では，映像データから人物を対象に検出と追跡を行いその性能を確認した．また，検出と追跡の困難な点をカバーしつつ予測を行うために，“群”を形成する手法を模索した．今後は，今回模索した群を形成し群の挙動予測の手法の構築に取り組む．

謝辞

本研究の一部は，JSPS 科研費 JP17K00148,JP19H02254 の助成を受けて行われた．

参考文献

1) 内閣府：交通事故の被害・損失の経済的分析に関する調査結果について、交通事故の被害・損失の経済的分析に関する調査、2012．https://www8.cao.go.jp/koutu/chou-ken/index-c.html (2020/12/15 参照)
2) Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 961-971, 2016
3) Ma, W. C., Huang, D. A., Lee, N., & Kitani, K. M.: Forecasting interactive dynamics of pedestrians with fictitious play. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 774-782. 2017
4) Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M.: YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934. 2020.
5) Wojke N., Alex B., & Dietrich P.: Simple online and realtime tracking with a deep association metric. 2017 IEEE international conference on image processing. IEEE, 2017.
6) Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C. C., Lee, J. T., ... & Swears, E.: A large-scale benchmark dataset for event recognition in surveillance video., conference on computer vision and pattern recognition 2011, pp. 3153-3160, 2011
7) Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., ... & Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003.2020