CI-002

StrongSORT と密度推定による密度に依存しないグループ検出
Density-independent group detection with StrongSORT and density estimation

植野直次朗 † 中川博之 ‡ 土屋達弘 †
Naojiro Ueno Hiroyuki Nakagawa Tatsuhiro Tsuchiya

概 要

カメラで撮影された動画から人間の行動を分析する技術が求められている．特に動画中の群衆からグループを検出する技術に関心が高まっている．例えば，公共施設に設置されたカメラによるグループ検出は，混雑状況に合わせた警備員配置の最適化を可能とする．本研究では群衆の映る動画からグループを余すことなく検出することを目標とする．オブジェクトトラッキングツールとして StrongSORT を，密度推定技術として DM-Count を用い，群衆密度に依存しないグループ検出技術を提案する．実世界で撮影された動画を用いた評価実験を通じて，提案手法により様々な密度でグループ検出が可能となることを確認した．

1. はじめに

画像から目的の物体を検出する技術は進展し，動画から目標物体を検出してトラッキングする技術が確立された．現代では，同技術を活かした様々な方面での技術の開発が進められている．例えば，自動運転においては，車両の周囲の状況を認識しながら次の行動を決定する必要がある．自動運転技術が確立されると人が運転する必要性がなくなり，移動の利便性が格段に向上する．一方で車両の場合，数メートルの誤差が事故を招く可能性があり，認識技術の正確性が求められる．また身近な例では，近年ではコロナの影響を受け入店時に自動的に顔認証を行い体温を測ってくれるようなシステムがある．非接触で検温をおこなうことができたり，検温を無人で行えるようになった．またマスクを着用したままでも検温を受けることができる．これらの例からもわかるように物体検出やトラッキング周辺の技術が向上したことによる利点が多くある．

物体検出の手法は，古典的な R-CNN[1] や精度が向上した Faster R-CNN[2] や SSD[3]，精度に加え実行速度が向上した YOLO[4]，そして最新の自動運転技術への導入が期待されている SVNet[5] など多岐にわたる．

オブジェクトトラッキング技術を実現するためのフレームワークとして StrongSORT[6] がある．StrongSORT がトラッキングできるのは，あくまで StrongSORT 内で使用している物体検出アルゴリズムが事前に学習してるクラスのみである．本研究の目的であるグループの検出については扱われていない．またこのフレームワークは活用できる群衆密度に限界がある．

一方で，高密度の群衆を大まかに検出するためのフレームワークとして群衆密度推定の DM-Count[7] がある．このフレームワークを用いることで StrongSORT では人物を検出できないような密度の群衆に対してアプローチをすることができる．しかし大まかな検出や DM-Count に用いられている学習済みモデルの仕様上，群衆の一部の検出漏れやカメラに近接しているグループの検出漏れが生じる．

本研究では，あらゆる群衆の密度に対してグループを検出することを目的としている．まず StrongSORT で中密度以下にいる人物のトラッキングを行い，その際に得られる各パラメータを使用してグループを検出する．しかしその場合，群衆密度が高くなった際に StrongSORT で人物検出のために用いている YOLO の技術上検出できない人物が存在してしまう．そこで StrongSORT と同時に密度推定によるグループ検出技術も使用し，高密度でのグループ検出を実現している．群種密度推定は先述した DM-Count というフレームワークが提供している．実験では動画から目視で確認できるグループを正解データとし，本手法の検出結果と比較することで性能を測っている．

結果として，検出精度が高いことが確認できた．また既存研究に比べ，様々な密度の群衆が存在する動画から，高精度のグループ検出が実現できていることが確認できた．

本論文の構成は以下の通りである．2 節では関連研究及びグループ検出のために使用したフレームワークを示す．3 節では研究目的と，提案手法の実装方法について説明する．4 節では比較実験の方法と結果，出力結果の一部を示す．5 節では本研究に関する考察を行う．最後に 6 節で本研究のまとめと将来的な本論文の総括，今後の課題を示す．

2. 研究背景

2.1. 物体検出

物体を検出する検出技術は古くから存在し，今日まで改善が繰り返し行われより実用性や有用性が高まっている．P.Viola と M.Jones[9] によって提案された物体検出に始まり，最新のものはニューラルネットワークや深層学習を用いた物体検出に進展している．画像での物体検出の技術を基に動画から目標物を検出する方法も確立された．動画を細かく分割し画像に変換して，各画像で物体検出技術を使用する．次に各画像で目標物を検出した後ラベリングを行い，各画像間でそのラベルを共有することで検出物をトラッキングすることも可能になっている．この技術は工場での異常や不良の検査を自動化するために使用されている．自動運転技術の開発や工場での異常や不良の検知にはこれらの技術が必要不可欠である．他にも防犯や監視用としての使用 [10] や，医療方面での利用 [11] もされている．今やこの技術は多方面で必要不可欠なものとなりつつある．

†大阪大学大学院情報科学研究科, Graduate School of Information Science and Technology, Osaka University
‡岡山大学, Okayama University

2.2. 関連研究

人が多く行きかう動画の中からグループを検出することは重要な問題として考えられている．Qi Wang ら [12] や Shao ら [13] は人間に共通する特徴量から人を検出し，人を粒子とみなし粒子の流れを取得している．その粒子の集まりからおおまかなグループを検出する．そこで検出されたグループには流れが同じ方向である他人もまとめて含まれている．すなわちこの検出はマクロ的なグループの検出であり，これは群衆行動解析の補完を行っている [14]．また Julio Cezar ら [15] はより高密度の群衆内からの集団を検出方法を提案している．彼らも個人の検出と識別は行っておらず，人を粒子とみなしその粒子の流れと密度から集団を検出している．こちらもマクロ的なグループの検出である．

またより小さなグループを検出している手法も提案されている．波田ら [16] は人物間距離や速度ベクトルに加えて，ジェスチャーの発生度を特徴量として小グループを検出する手法を提案している．また Weina Ge ら [15] は独自の特徴量で人を検出し，人物間距離と速度ベクトルからグループかどうかを判定するアルゴリズムを提案している．しかしこの手法では 1 フレームあたり高々20 人の密度が小さい群衆の場合でないと精度を高く保つことができない．よって 1 フレームあたり 40 人ほど存在する群衆について同手法を用いると，精度がかなり低くなっている．人が多く集まる駅のホームやショッピングモールのような人が集まる場所での使用が難しいく実用的ではないと考える．

Giovanna Castellano ら [17] は DM-Count を用いて，ドローンを用いて上空から撮影した群衆からプループを大まかに検出する手法を提案している．この手法は先述の小グループの検出手法とは異なり，人が多く集まる街中や広場に対して効力を発揮している．人が集まる場所で効力を発揮するので，本質的にはオクルージョンに対してロバストである．一方で人物がカメラに近接した際や，群衆の密度が小さいときは検出精度が低くなることにも言及している．

2.3. 本研究で用いる技術

本研究では，主に人物検出，トラッキング，群衆密度推定，クラスタリングなどの技術を用いる．これらのために使用したツールを以下で説明する．

2.3.1.YOLO v7

人物検出のために，本研究では YOLO (You Only Look Once) を用いる．YOLO は，推論速度が他の多くのアルゴリズムより高速な物体検出アルゴリズムの一つである．もともと物体検出は検出と識別の 2 段階で行われていたが，YOLO はこの過程を一度の作業にすることで高速化に成功した．画像を一度正方形にリサイズすることでニューラルネットワークを使用した分析を行いやすくしている．リサイズした画像をグリッド・セルと呼ばれる小さな正方形にさらに分割し，このグリッド・セルを基にして物体の検出が行われる．画像内で検出物体の範囲を特定するためにバウンディングボックスと呼ばれる矩形を用いる．

YOLO には様々なバージョンが存在し，今回は使用した他フレームワークとの親和性から YOLO v7[18] を使用した．YOLO v7 は現時点での YOLO シリーズの最新のバージョンの 1 つである．それまでのバージョンと比較して，最も正確性が高く，検出にかかる時間も最短である．

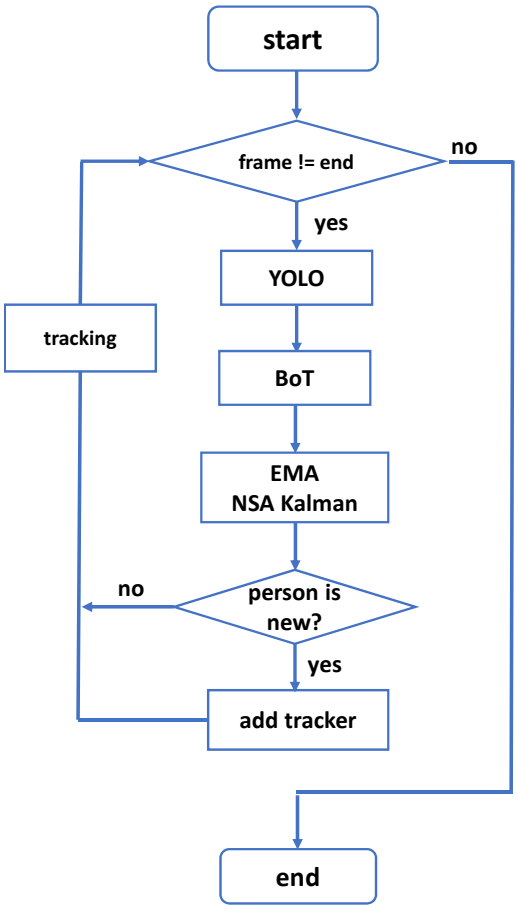

図 1: StrongSORT のフローチャート

2.3.2.StrongSORT

YOLO による検出物体をトラッキングするために StrongSORT を用いる．StrongSORT とは入力動画から人物を検出し，各人物をトラッキングするためのフレームワークである．まず，入力動画をフレームに分割し，各フレームに対して YOLO を用いて物体の検出を行う．次に検出した物体の個体同士の特徴を区別し，よりロバストなトラッキングを行うために外観特徴抽出器である BoT[19] を用いる．BoT により外観の特徴を抽出し，EMA（Exponential Moving Average）[17] を用いて同一の検出物体をトラッキングする．EMA では注目物体の短期間の軌跡の外観特徴を保持している．EMA は過去の特徴を慣性項として保持し，特徴を更新していく．

$$e_i^t = \alpha e_i^{t-1} + (1 - \alpha) f_i^t$$

f はフレーム t で検出され tracklet i に割り当てられた物体の特徴，tracklet は正確な追尾をしているが短期間しか追尾できていない不完全な軌跡，e は $t-1$ フレームまでの tracklet の特徴である．これを α で重み付けることで効率的且つノイズを軽減しながら特徴を更新することができ，BoT と合わせて使用することでロバストなトラッキングを実現する．また，時間変化する移動

物体の観測値を推定するために NSA Karman が必要となる．StrongSORT 以前のトラッキングアルゴリズムでは，単純な線形カルマンフィルタ [23] を使用したが，検出物体が全て同じ観測ノイズを持っていることが前提とされてた．そこで NSA Karman によって，ノイズを検出の信頼度に応じて適用的に変化させている．複雑な動きをした物体では検出器が出力する物体の信頼度が低いことが考えられるので，カルマンフィルタでの補正を強化している．また，よりロバストなトラッキングを実現するために，カメラの視点の変化に対応する必要があり，補正アルゴリズムである ECC が用いられている．図 1 に StrongSORT のフローチャートを示す．

2.3.3.DM-Count

群衆検出のために，本研究では DM-Count（Distribution Matching for crowd COUNTing）を用いる．DM-Count は群衆密度推定を行うためのフレームワークである．群衆密度推定とは，群衆を対象に非常に密集している群衆の中の人物の数をカウントする技術である．群衆密度推定手法は detection-then-count，direct count regression，density map estimation の大きく 3 つに分類される．1 つ目の detection-then-count は，YOLO など対象物体を個別に検出してカウントするアプローチである．このアプローチでは出力のバウンディングボックスごとにアノテーションが必要となる．人数が数百人になる群衆に対して個別にアノテーションすることは非常に高コストであり，検出アルゴリズムの精度も低くなる．対象物が比較的少ない場合に精度が出る手法である．2 つ目の direct count regression は直接対象物を回帰するアプローチである．点によるアノテーションを用いたり，明示的にアノテーションを用いない場合もあり，detection-then-count に比べアノテーションコストが低い．direct count regression に分類される手法の 1 つである CNN ベースの手法では，5 層の畳み込み層と 2 層の全結合層からなるネットワークを二乗誤差で学習させる回帰モデルである．それ以前の画像特徴を手で作っていた手法を深層学習で置き換えた点で優れている．そして 3 つ目の density map estimation は現時点で他の 2 つに比べて最も性能が高いとされている．特に density map estimation に分類されるフレームワークの中で，本手法では DM-Count を使用している．

DM-Count は最適輸送を損失関数に導入した手法である．DM-Count で使用する損失関数は，3 つの損失関数から構成されており以下の式で表される．

$$\mathcal{L}^{DM-Count}(z,\hat{z}) = \mathcal{L}_C(z,\hat{z}) + \lambda_1\mathcal{L}_{OT}(z,\hat{z}) + \lambda_2||z||_1\mathcal{L}_{TV}(z,\hat{z})$$

z は密度アップをベクトル化したものでありアノテーションを示す．また，$\hat{z}$ はニューラルネットワークによる予測密度マップをあらわす．は 3 つの損失関数 $\mathcal{L}_c$，$\mathcal{L}_{OT}$，$\mathcal{L}_{TV}$ について，以下で詳しく説明する．

1 つ目の $\mathcal{L}_C$ は Counting Loss と呼ばれる損失関数である．これは予測密度マップの値の合計を最終的な対象の数とみなすことで，真の対象物の数と一致させるように学習させる損失関数である $\mathcal{L}_C$ の式は以下である．

図 2: DM-Count による出力

$$\mathcal{L}_c(z,\hat{z}) = |||z||_1 - ||\hat{z}||_1|$$

2 つ目の $\mathcal{L}_{OT}$ は Optimal Transport Loss（OT Loss）と呼ばれる損失関数である．これは Monge の最適輸送問題 [24] の定式化に基づき，それぞれ正規化されたアノテーション点による密度マップと予測密度マップの 2 つの分布間の輸送コストが最小になる場合のコストを表す．

3 つ目の $\mathcal{L}_{TV}$ は Total Variation Loss と呼ばれる損失関数である．OT Loss を Sinkhorn アルゴリズム [25] で計算しても，近似が不十分であるという課題があった．これを解決するために導入された損失関数である．$\mathcal{L}_{TV}$ の式は以下である．

$$\mathcal{L}_{TV}(z,\hat{z}) = \frac{1}{2}\left|\left|\frac{z}{|z|_1} - \frac{\hat{z}}{|\hat{z}|_1}\right|\right|$$

これら 3 つの損失関数を用いることにより学習が安定することが知られている．

2.3.4.DBSCAN

クラスタ数が不特定の場面でクラスタリングを実行するために DBSCAN[26] が必要となる．DBSCAN とは密度準拠クラスタリングのアルゴリズムである．半径以内に点がいくつあるかでその領域をクラスタとして判断し，近傍の密度がある閾値を超えている限りクラスタを成長させ続ける．ある点から一定距離以内にある点の数を数え，その数が閾値以上であるとクラスタとみなす作業を繰り返すという特徴がある．これにより形が複雑なクラスタに対しても対応できる．最初にクラスタ数を決める必要がないため，本研究のようにクラスタ数が不明なものを対象とする際に適している．

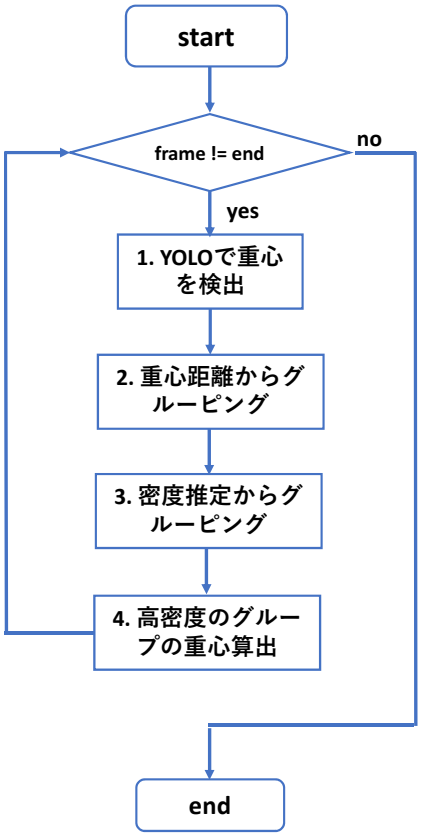

図 3: 提案手法のフローチャート

3. 提案手法

3.1. 研究目的

本研究の目的は，グループ検出を群衆の密度に依存せず検出精度を高く保つことである．中密度の群衆からの各人物に着目したグループ検出や，高密度の群衆から大まかにグループを検出する手法は存在する．本研究では，あらゆる群衆密度からのグループ検出を同時に行うことができる手法を提案する．場面によっては高密度な領域と中密度以下の領域が存在するが，これに対しても対応できる手法を提案している．また高密度の群衆でのグループ検出は大まかであるがために取りこぼしが存在していたが，各検出をクラスタリングし高密度群衆の重心を算出することでより精度の高い検出を実現している．

3.2. 提案手法のアルゴリズム

図 3 に提案手法のアルゴリズムのフローを示す．基本的な構造は StrongSORT の構造を用いており，適宜変化を加えている．処理の流れを以下に示す．

Step 1： YOLO による出力から重心を算出する．StrongSORT 内の YOLO を使用して人物を検出し，YOLO の出力のバウンディングボックスから重心を計算し出力する．

Step 2： 重心距離からグループを検出する．Step1 で出力した重心の座標を元に座標間距離からグループであるかを判定する．判定の方法としては，動画ごとに実験して適切な閾値を設定し，その閾値よりも人物間の重心距離が小さいとグループと判定している．これにより中密度以下の群衆から小グループを検出している．

図 4: DM-Count の出力結果

Step 3： 群衆密度推定を行ってグループを検出する．DM-Count を使用した群衆密度推定を実行している．ここで高密度の群衆がどこに存在しており，何人程度の群衆であるかを大まかに検出している．

Step 4： Step3 で検出したグループから重心を出力する．Step3 で検出した大まかなグループに対してクラスタリングを実行する．クラスタリングを行うことで，Step 3 で検出した大まかなグループで近接しているもの同士をまとめ上げ，グループの分裂や検出漏れを防ぎ精度を高めている．

また，Step3 での出力である密度マップの出力結果から，Step4 でどのようにクラスタリングを行っているかについての詳細を説明する．図 4 はある場面に対して DM-Count を使用し，特徴マップを出力した例である．この結果から高密度の群衆から大まかなグループ検出が行われていることがわかる．しかし DM-Count の性能上，検出グループが分裂してしまっているためクラスタリングで統合する．Step4 でこの出力画像について，クラスタリングを行う前に正規化を行う．正規化を行うことにより特徴の尺度の統一と，外れ値の影響の軽減を実現している．前者に関して，クラスタリングは一般的に各特徴量の尺度が同程度であることが好ましいためである．これにより異なる尺度の特徴量がクラスタリングアルゴリズムに適切に影響を与えることができる．後者に関して，正規化により外れ値がクラスタリングに与える影響を軽減することができる．正規化により，外れ値が通常のデータと同じスケールに置かれ，クラスタリングの結果が安定する．

そしてその後，正規化した特徴マップを，指定した閾

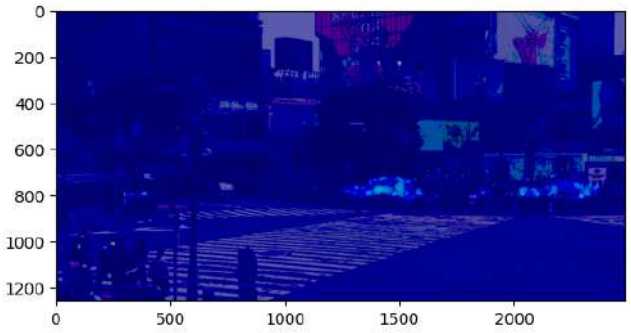

図 5: 処理後の特徴マップ

値を用いて画像のしきい値処理を行う．しきい値処理後の画像からクラスタリングの前処理として改善した特徴マップを作成する．ここでは DM-Count カウントの出力画像は青色成分を主としていることから，特徴抽出をしやすいため青チャンネルをもとに特徴マップを作成している．正規化と青チャンネルによって改めて作成した特徴マップを図 5 に示す．最後に作成した特徴マップの黒色のピクセル，すなわち閾値以下となったピクセル以外を取得しクラスタリングを行う．これらの処理を行い残ったピクセルに対して DBSCAN を用いて，各同一クラスタ内の点から重心を算出し出力する．以上の処理により，大まかな検出グループをまとめあげ，グループ検出としての精度を高めている．

4. 評価実験

本研究では，群衆が行きかう動画からキャプチャした静止画に対して評価実験を実施した．手法の評価には評価基準を設ける必要があるが，歩行者からグループ分けを行っているデータセットがないため直接的な精度比較は困難である．そこで評価対象を目視で確認できるグループに所属している人物の数とし，動画から目視で判断できるグループに所属する人数と本手法で検出された各人物がグループに所属しているかどうかを比較し，これにより定量的な面で本手法を評価する．また本手法で設定している動画や各パラメータについての言及する．

4.1. 実験設定

実験対象は MOT charenge に設定されているベンチマーク動画（動画 1，2）と，Youtube にて LIVE 配信されている渋谷のスクランブル交差点の動画（動画 3）を使用した．定量的な比較ができるように目視で複数のグループを検出できる場面を選択した．また本手法の汎用性を示すため，各動画において群衆が明確に分割されている場面と混雑している場面の 2 場面を静止画で示す．動画 3 に関しては，群衆が分割されているものをシーン 1，混雑しているものをシーン 2 としている．

評価基準には Precision，Recall，F-measure を使用する．本実験では Precision は正しいグループの検出率，Recall は検出結果の正答率，F-measure はこれらの総合的な評価が該当する．これらの定義を以下に示す．

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F\text{-}measure = \frac{2 \times precision \times recall}{precison + recall}$$

TP: 検出されたグループである人物の数

FP: 誤検出されたグループである人物の数

FN: 検出されていないグループである人物の数

TN: 検出されていないグループでない人物の数

状況が大きく異なる場面を用いて評価を行っており，Precision と Recall はトレードオフの関係であることから変動する可能性がある．そこでこれらの調和平均である F-measure を確認し性能を評価している．また本手法と比較するのは 3 つの手法

手法 1： YOLO(StrongSORT) のみでグループ検出を行ったもの

手法 2： DM-Count のみでグループ検出を行ったもの

手法 3： YOLO と DM-Count のグループ検出を単純に OR 結合したもの

である．手法 1 は YOLO のみであるので，中密度の群衆に対してのみ対応しているグループ検出手法である．手法 2 は DM-Count のみであるので，高密度の群衆に対してのみ対応しているグループ検出手法である．手法 3 は YOLO のみのグループ検出と DM-Count のみのグループ検出を単純に OR 結合したものであり，密度に依存はしない．

提案手法では高密度の群衆においては，DM-Count によるグループを検出した後にその出力結果をもとにクラスタリングを行いグループの重心を出力している．また評価実験の動画 3 では，同じ動画内の異なる場面を使用している．シーン 1 は信号待ちの場面であり，群衆が領域ごとにきれいに分断されている場面である．またシーン 2 は信号が青に変わり，人が交差点を移動している最中であり，群衆としてのまとまりはなくなっている場面である．ベンチマーク動画である動画 1 はシーン 1 と似た場面であり，動画 2 はシーン 2 と似た場面となっている．

4.2. 出力結果

YOLO のみでのグループ検出である手法 1 は，YOLO の人物の重心の検出結果を緑色の点，グループの検出には青色の線を用いており人物の重心を接続する形で表示している．DM-Count のみによるグループ検出である手法 2 は青色で出力されている．YOLO と DM-Count のグループ検出を単純に OR 結合した手法 3 は，各出力画像の透過度を調整し重ねた形で出力している．提案手法は YOLO の出力はそのままで，DM-Count によって検出したグループはまた本論文においては，画像として示すのは MOT ベンチマークの動画のみである．

図**??**は MOT ベンチマーク動画 1 での評価画像であり，

図 6: 動画 1 の実験結果（左上：手法 1，右上：手法 2，左下：手法 3，右下：提案手法，赤点は密度推定で検出したグループの重心）

図 7: 動画 2 実験結果

渋谷の動画の場面 1 を想定された群衆の領域が分かれている場面である．図??は MOT ベンチマーク動画 2 での評価画像であり，場面 2 を想定している．各図の左上は YOLO のみによる出力，右上は DM-Count のみによる出力，左下がそれぞれを単純に OR 結合した出力，左下が提案手法の出力である．

4.3. 評価結果

各検出手法について，2 つのベンチマーク動画（動画 1，2）の静止画と，渋谷の映像（動画 3）の 2 つのシーンから観測した TP，FP，FN の値を表 1〜4 に示す．場面 1 の評価結果を表 1 に，場面 2 の評価結果を表 2 に示す．

5. 考察

実験結果から，群衆密度推定と人物検出ベースでの利点を活かしつつ，既存研究より検出漏れを少なくグループを検出できていることが分かる．またその中で精度を高く保ったまま検出できていることが確認できた．

評価実験の結果から精度が場面の密集具合や人物とカメラまでの距離に強い関係があることが確認できる．これは高密度での群衆検出を行っている DM-Count が，遠くからの方が，またある程度密集している場合でないと高い精度が保てないことが原因である．適宜入力動画によってパラメータを調整したが，より適切なパラメータを見つけることができれば精度の向上が期待できる．

また実行速度の改善も見込める．現在はリアルタイム性という観点で実行速度に関して問題がある．提案手法では，入力動画をフレームに分解し，各フレームに関して YOLO による検出と DM-Count による検出を順に実行している．しかし DM-Count で検出した領域には YOLO を適用する必要がないため，DM-Count で検出した領域以外で YOLO を適用することができればより処理速度が向上する．本手法のリアルタイム性を向上させると利用用途の幅も広がると考察する．

YOLO でのグループ検出と DM-Count 実行後のクラスタリングには，実験的に閾値やパラメータを設定している．YOLO の座標間距離に関しては実験的に一番性能が良いと考えたものを採用している．DBSCAN クラスタリングのパラメータである eps と min_sample に関しては，eps に関しては経験的に設定し，min_sample は 5 人以上であればグループであるという考えから値を 5 に設定している．YOLO に関して，カメラと人物の遠近感とグループと判定する閾値には相関があると考える．人物が近い場合はバウンディングボックスの面積が大きく出力され，グループと判定する重心間距離も大きくなる．この関係性を考慮することで，理論的に定式化できると考える．またクラスタリングのパラメータに関しても，DM-Count によって出力されたグループの色から密集具合を判定し，クラスタと判定するパラメータである eps を導出できると考える．また AutoEpsDBSCAN[29] という各パラメータを自動的に決定する技術について研究がされており，これを本研究に適用することが可能であればクラスタリングに関しては自動化が可能である．この技術についても今後調査する必要がある．

6. おわりに

本研究では様々な密度の群衆から，あますことなくグループを検出するために，YOLO による重心間座標距離と DM-Count による群衆密度推定を使用したグループ検出の手法を提案した．また既存研究に比べより幅広い密度の群衆に対してもグループの検出を行える手法を提案した．検出結果のグループの集合と，目視で確認できるグループの集合を定量的な比較を行うことで本手法の性能を評価した．

今回の実装について懸念する点が 3 つ存在する．1 つ目は YOLO での検出や DBSCAN によるクラスタリングにおいて，閾値やパラメータを経験的に設定している点である．本手法の一般化という観点では改善する必要がある．2 つ目は実行速度を上げることである．YOLO や DM-Count を適用する領域を分けることで全体的な計算量を削減し，リアルタイム性の観点で有用性を高めることを目指す．3 つ目はクラスタリングの結果の出力方法である．今回は DM-Count による密度推定結果と比較することで，高密度の群衆からグループの重心を算出できたことを示している．しかし実用性の観点から，重心だけの出力であるとグループの範囲が特定できないという課題がある．今後は，これら 3 つの課題に取り組み，より実用性の高い技術の実現を目指したい．

表 1: 動画 1 における各手法の評価値

	TP	FP	FN	TF	Precision	Recall	F-measure
手法 1	9	1	120	8	0.818	0.070	0.129
手法 2	63	2	66	7	0.964	0.488	0.648
手法 3	74	2	55	7	0.974	0.574	0.722
提案手法	78	3	51	7	0.963	0.605	0.743

表 2: 動画 2 における各手法の評価値

	TP	FP	FN	TF	Precision	Recall	F-measure
手法 1	23	0	206	2	1.000	0.010	0.182
手法 2	73	0	156	2	1.000	0.468	0.638
手法 3	94	0	135	2	1.000	0.696	0.821
提案手法	105	2	124	0	0.981	0.847	0.909

表 3: 動画 3 シーン 1 における各手法の評価値

	TP	FP	FN	TF	Precision	Recall	F-measure
手法 1	24	0	137	11	1.000	0.149	0.259
手法 2	70	0	91	11	1.000	0.435	0.606
手法 3	81	0	80	11	1.000	0.503	0.669
提案手法	143	0	18	11	1.000	0.888	0.941

表 4: 動画 3 シーン 2 における各手法の評価値

	TP	FP	FN	TF	Precision	Recall	F-measure
手法 1	4	2	161	33	0.667	0.024	0.023
手法 2	40	5	125	30	0.889	0.242	0.380
手法 3	44	7	121	30	0.863	0.267	0.408
提案手法	103	5	62	30	0.954	0.624	0.754

参考文献

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hier-archies for accurate object detection and semantic segmentation, 2013.

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot MultiBox detector. In Computer Vision – ECCV 2016, pp. 21–37. Springer International Publishing, 2016.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2015.

[5] Shaoshuai Li and Fuyan Liu. Svnet: A single view network for 3d shape recognition. In 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 1648–1653, 2019.

[6] Yunhao Du and Zhicheng Zhao and Yang Song and Yanyun Zhao and Fei Su and Tao Gong and Hongying Meng. StrongSORT: Make DeepSORT Great Again

[7] Boyu Wang, Huidong Liu, Dimitris Samaras, Minh Hoai Nguyen. Distribution Matching for Crowd Counting. Advances in neural information processing systems 33 (2020): 1595-1607.

[8] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition.

[9] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Vol. 1, pp. I–I, 2001.

[10] 藤吉弘亘, 金出武雄. Vsam. 映像情報メディア学会誌, Vol. 57, No. 9, pp. 1068–1072, 2003.

[11] Hayit Greenspan, Bram van Ginneken, and Ronald M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging, Vol. 35, No. 5, pp. 1153–1159, 2016.

[12] Qi Wang, Mulin Chen, Feiping Nie, and Xuelong Li. Detecting coherent groups in crowd scenes by multiview clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, No. 1, pp. 46–58, 2020.

[13] Jing Shao, Chen Change Loy, and Xiaogang Wang. Scene-independent group profil- ing in crowd. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2227–2234, 2014.

[14] 高柳英明, 長山淳一, 渡辺仁史. 歩行者の最適速度保持行動を考慮した歩行行動モデル : 群衆の小集団形成に見られる追跡-追従相転移現象に基づく解析数理. 日本建築学会計画系論文集, Vol. 71, No. 606, pp. 63–70, 2006.

[15] Julio Cezar Silveira Jacques Junior, Soraia Raupp Musse, and Claudio Rosito Jung. Crowd analysis using computer vision techniques. IEEE Signal Processing Magazine, Vol. 27, No. 5, pp. 66–77, 2010.

[16] 波部斉, 橋本知典, 満上育久, 鷲見和彦, 八木康史. 人物のジェスチャーを加味した歩行者グループ検出. 知能と情報, Vol. 29, No. 3, pp. 605–610, 2017.

[17] Yunhao Du, Junfeng Wan, Yanyun Zhao, Binyu Zhang, Zhihang Tong, Junhao Dong. GIAO-Tracker: A comprehensive framework for MC-MOT with global information and optimizing strategies in VisDrone 2021

[18] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Train- able bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022.

[19] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.

[20] Nicolai Wojke, Alex Bewley, Dietrich Paulus. Simple Online and Realtime Tracking with a Deep Association Metric

[21] Zhang, Hang and Wu, Chongruo and Zhang, Zhongyue and Zhu, Yi and Lin, Haibin and Zhang, Zhi and Sun, Yue and He, Tong and Mueller, Jonas and Manmatha, R. and Li, Mu and Smola, Alexander. ResNeSt: Split-Attention Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2022

[22] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification, 2019.

[23] Rudolf Emil Kalman, A New Approach to Linear Filtering and Prediction Problems., ASME. J. Basic Eng, March 1960; 82(1): 35–45.

[24] G.Monge,M ´ emoiresurlaTh ´ eorie des D ´ eblais et des Remblais, De l ' Imprimerie Royale, 1781.

[25] Sinkhorn Richard, Knopp, Paul. Concerning nonnegative matrices and doubly stochastic matrices. Pacific J. Math. 21, 343–348. 1967

[26] Ester, Martin and Kriegel, Hans-Peter and Sander, Jörg and Xu, Xiaowei and others: A density-based algorithm for discovering clusters in large spatial databases with noise

[27] Weina Ge, Robert T. Collins, and R. Barry Ruback. Vision-based analysis of small groups in pedestrian crowds. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 5, pp. 1003–1016, 2012.

[28] Edward T. Hall. A system for the notation of proxemic behavior. American An- thropologist, Vol. 65, No. 5, pp. 1003–1026, 1963.

[29] Manisha Naik Gaonkar, Kedar Sawant. AutoEps-DBSCAN: DBSCAN with Eps automatic for large dataset. International Journal on Advanced Computer Theory and Engineering, 2013