「マルチメディア，分散，協調とモバイル
(DICOMO2021) シンポジウム」 令和3年6月-7月

深層距離学習を用いたARデバイス向けの人物識別手法

高橋 直也[1]　天野 辰哉[1]　山口 弘純[1]　東野 輝夫[1]

概要：本研究では，AR デバイス視野内の人物追跡用途を想定し，深層距離学習に基づく人物再識別手法を組み合わせた新しい人物追跡手法を提案する．提案手法では Deep Sort など連続フレーム間での人物追跡を前提とした従来手法が，AR デバイスにおいては視野移動による頻繁なフレームアウトで追跡に失敗する点に着目し，検出された人物が過去に追跡した人物と同一か否かを，深層距離学習により識別するモデルを構築する．深層距離学習ではデータの特徴量空間における距離を学習できるため，同一人物の画像間距離が小さく，異なる人物画像間距離が大きくなるよう学習することで，検出された人物画像と過去の追跡人物画像間の特徴距離を計算できる．提案手法では AR グラスを通したスポーツトレーニングや観戦を題材とし，シーン切り替えが頻繁に発生するラグビーの試合映像から得られた 27 万枚以上のデータに対してラベリングを行い，通常のクラス分類学習を通した距離学習が可能な ArcFace を用いた人物再識別モデルを構築した．提案手法を Person Re-identification に利用される既存のデータセット Market-1501 [1] を用い実装し，Market-1501 [1] と独自のデータセットに対する精度を評価した結果，前者では F 値 74.15%，後者では F 値 66.88%を達成した．

1. はじめに

Augmented Reality（AR）技術の発達に伴い，AR デバイスが備える RGB カメラや深度カメラを用いて現実空間の人やモノを認識し，それらに関するサイバー情報を現実空間映像やグラスに重畳して表示する様々なアプリケーションが提案されている．例えばスポーツ分野では，カメラによって試合やトレーニング中の選手を検出・識別し，AR デバイスを装着する観客やコーチに対し，それらのデバイスを通して選手情報を提供するシステムも提案されている [2]．このような AR アプリケーションの実現のためには，カメラが捉えた映像中の人の追跡が不可欠である．映像中の人物追跡手法として，最近では高度化する物体認識アルゴリズムを活用したものが増えてきている．例えば Deep Sort [3] では，よく知られた物体認識アルゴリズムである Yolo を用いて人物検出を行い，連続フレーム間の物体移動を想定したカルマンフィルタで追跡を行う．しかし Deep Sort を含め，既存の映像中の人物追跡手法は定点観測映像を想定しているものが多く，カメラの位置や向きが変化する状況では追跡中の人物がフレームアウトすることにより追跡が頻繁に途切れるといった課題がある．特にモバイル AR やヘッドセット AR では，装着者の挙動により AR デバイスの方向が高頻度で変化するため，非連続フレーム間での人物同定，すなわち，新たに追跡を開始した人物が過去に追跡した人物か否かを判定し，そうであればどの人物かを識別する必要がある．

これまでに，互いに異なる地点に設置された定点カメラから得られる映像に対し，それらのカメラの撮影領域間を移動する同一人物や車両を特定する，Multi-Target Multi-Camera Tracking とよばれる問題に対し，手法が提案されてきた．MTMCT は本質的に，異なる画像に映り込んだ同一オブジェクトの再識別（Re-Identification, Re-ID）を基本手法として用いる．コンピュータビジョン分野における Re-ID とは，与えられた判定対象のオブジェクト画像 i と，これまでに蓄積したオブジェクト画像集合 I が与えられた場合に，i に類似する順で画像集合 I をソートする問題として扱うことができる．本論文で扱う，AR デバイス向けの人物トラッキングにおいては，装着者が視点方向を変更し，新しい人物が登場するような状況を扱うため，複数視点から別の地点を捉える MTMCT と本質的に類似する．その一方で，あるカメラの視野角内に人物が存在する限りは安定的にトラッキング可能な MTMCT とは異なり，AR による人物トラッキングでは装着者の視点変化が予測できず，常に画像間で Re-ID による比較が求められる．トラッキングは人物の特徴量を蓄積していくため，時間経過とともに比較すべき対象データ集合 I が増大し，負荷増大につながる問題がある．

そこで本研究では，Deep Sort [3] など単一カメラの連続フレーム間で複数人物を追跡する手法（Multi-Object

[1] 大阪大学大学院情報科学研究科

Tracking, MOT）を活用し，MOT が異なる人物と判定したタイミングでフレームシーケンスをセグメント化するとともに，深層距離学習に基づく新しい人物再識別手法を組み合わせた追跡手法を提案する．提案手法ではカメラ視点の移動により MOT が人物追跡に失敗した場合，新しい人物と判定することを利用し，連続して捉えた同一人物の連続フレームをセグメント化するとともに，そのセグメントから代表画像を用いて過去の追跡人物の画像と同一か否かを識別する深層距離学習モデルを構築する．深層距離学習ではデータの特徴量空間における距離を学習することができるため同一人物の画像同士の距離は小さく，異なる人物画像同士の距離は大きくなるように学習しておくことにより，検出された人物と過去に追跡されていた人物同士の特徴距離を計算する．提案手法では AR グラスを通したスポーツトレーニングや観戦を題材とし，シーン切り替えが頻繁に発生するラグビーの試合映像から得られた 27 万枚以上のデータに対してラベリングを行い，通常のクラス分類学習を通した距離学習が可能な ArcFace を用いた人物再識別モデルを構築した．

本稿は以下のように構成する．2 章では既存研究について述べ，本研究の位置付けを明確にする．3 章では，提案手法の概要と Deep Metric Learning を用いた人物再識別手法について述べる．4 章では，提案手法の性能評価を AR アプリケーションを想定した独自のデータセットをテストデータとして実証する．5 章では，本研究のまとめと今後の課題について述べる．

2. 関連研究

2.1 Multi-Object Tracking (MOT)

単一カメラ内で人物や車両を追跡する手法は Multi-Object Tracking (MOT) として以前より多く研究されている．MOT 手法は通常, 検出ベーストラッキング (Detection-Based Tracking, DBT) および検出フリートラッキング (Detection-Free Tracking DFT) に分類される．DBT では物体を事前学習した検出器により検出し，それらを繋ぎ合わせた軌跡を導出する [4] のに対し，DFT では，初期フレームにおいて手動によるオブジェクト指定を行ったあと，その移動を追跡する [5]． オブジェクトが消滅，出現を繰り返すより一般的な環境に対応可能である点と，近年の物体認識の高度化により，現在では DBT がより一般的となっている．MOT では大規模なベンチマークとデータセットは MOT Challange として公開されており [6]，任意のオブジェクトのトラッキングや，混雑時におけるトラッキングなど様々なシーンにおけるトラッキング手法の比較が可能となっている．また，Multi-Object Tracking and Segmentation [7] では，トラッキングとセグメンテーションを同時に行う方法を提案しており，KITTI データセットを用いた評価を行っている．MOT に関しては最新のサーベイ論文 [8] を参照されたい．

2.2 Multi-Target Multi-Camera Tracking (MTMCT)

異なる地点に設置されたカメラによる複数の定点観測映像において，複数オブジェクトが出現する中で各オブジェクトをトラッキングする問題（Multi-Target Multi-Camera Tracking）手法が研究されてきている．MTMCT 問題はカメラの設置位置により光量や人物までの距離，撮影方向が大きく異なる中で同一人物を発見できる必要があるため，その高精度化は非常に挑戦的な問題である．

MTMCT は R-CNN や SSD といった深層学習ベースの高精度なオブジェクト検出器や OpenPose [9] などの姿勢検出器の登場により，それらを用いた高精度化に向けて一層注目を集めている研究分野である []．MTMCT 手法の多くは，カメラ同志の FOV（Field of View）の重畳を仮定できないため，単一カメラ内ではオブジェクト追跡（MOT）を用い，複数カメラ間ではオブジェクト再識別 (Re-idenfication, Re-ID）をもとに人物同定を行う．Re-ID 問題は与えられたターゲット（人物）画像と画像集合に対し，その類似度でランク付けを行う問題であり，MTMCT はそれを用い，2 つの画像が与えられたときにそれが同じ人物か否かを判定する 2 値分類器を構成する問題である [10]．[10] では MTMCT 問題と Re-ID 問題に共通する特徴量を発見し，損失関数として動的重み付き Triplet Loss を用いる手法を提案している．

車両の MTMCT 手法に関しては複数の路側監視カメラを用いた CityFlow [11] などが知られている．CityFlow では, 最大 2.5km 離れた 10 の交差点に設置された 40 台のカメラから得られた 3 時間以上の交通 HD 映像に 20 万以上のラベル付けを行ったデータを提供しており，車両再識別のためのベンチマークを提供している．車両検出，トラッキング，車両再識別において，様々な追跡手法やそれらの組み合わせの性能評価を行っており，2019 AI City Challenge (https://www.aicitychallenge.org/) に向けたサーバーを提供した実績がある．

なお, 人物再識別用データセットとして, Market-1501 [12] や Motion Analysis and Re-identification Set (MARS) [13] が知られている．Market-1501 は Deformable Part Model (DPM) による人物検出器を用いた 3 万 2 千以上のバウンディングボックスがアノテーションされており，50 万以上の不正解データも含まれている．また MARS はビデオデータを対象としており，Market-1501 同様，DPM を用いて，1,261 の人物による約 2 万の Tracklet 特徴量（追跡点の軌跡）を提供している．

2.3 提案手法の位置づけ

前述したような MOT または MTMCT では，固定設置

された単一または複数カメラを使用することが前提となっている．しかし，AR グラスを通したトラッキングシステムのように，カメラ視点が固定されておらずシーン切り替えが頻繁に発生する状況を想定したトラッキング手法ではない．これに対して提案手法では，Deep Sort といった連続フレームを想定した MOT 向けトラッキング手法を活用し，そのトラッキングが新しい ID を生成するタイミングを検出したうえで，深層距離学習による識別器を訓練することで，新しい ID 付与が必要か否かを判断する方法論を導入している点で従来手法とは異なる．また AR グラスでのトラッキングを想定した大量のデータセットを生成し，それを用いたテストを行っている点も従来手法とは大きく異なる．

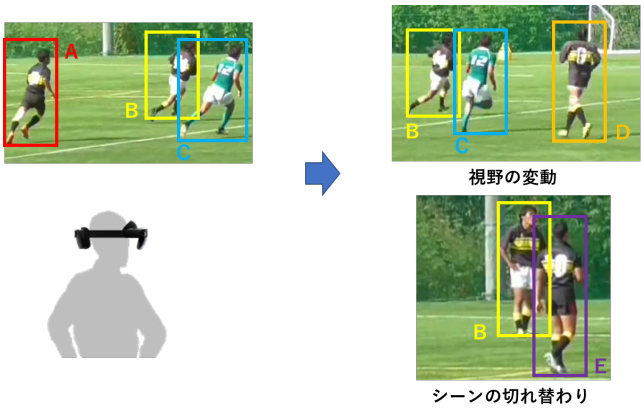

図 1　提案手法の利用シーン

3. 提案する人物識別手法

3.1 提案手法の概要

本システムは，AR グラスを装着した監督やコーチが選手の位置把握に利用したり，観客が選手の情報の動きを追跡するような用途を想定している．本システムが想定する利用シナリオを図 1 に示す．

AR デバイス装着者の視野が変動することによりシーンが切り替わる場合も継続して同一選手をトラッキングすることで，AR デバイス上で選手の位置を把握することを目指している．本研究では，Deep Sort [3] によるトラッキングと，ArcFace の人物再識別モデルを組み合わせた手法を提案する．

本研究で提案するシステムを図 2 に示す．まず AR デバイス装着者が取得する映像データに対し，連続フレームに対するトラッキング手法である Deep Sort [3] により連続してフレームに映り込む選手のフレーム間同定（フレーム間トラッキング）を行う．システムの時刻を t で表し，実行開始時は $t = 0$，i 枚目の画像フレームを得た時刻を $t = i$ とする．また，検出した人物（選手）の画像セグメントを保持する画像セグメントデータベース（以下，単にデータベースとよぶ）の時刻 i における処理後の内容を DB_i で表すとし，$DB_0 = \{\}$ とする．

以下，時刻 k に得られた画像フレーム内で検出された選手の画像セグメント集合を P_k，$p \in P_k$ に対し，Deep Sort が出力する ID を $ID_{deep}(p)$, 提案システムが最終的に決定した ID を $ID_{final}(p)$ とする．また，一般的なマップ関数 $map(f, L) = \{\forall x \in L | f(x)\}$ を用いる．

AR デバイス装着者の視点は常に変化するため，選手が視界からフレームアウトしたり，図 3 で示すような選手同士の重なり（オクルージョン）が発生したりした場合，Deep Sort は類似オブジェクトを発見できない可能性が高い．これを利用し，$map(ID_{deep}, P_k) \neq map(ID_{final}, P_{k-1})$ であれば，シーンの切り替わりにより新しい ID が与えられたとみなす．各画像セグメント $p \in P_k$ とデータベース DB_k に対し，ArcFace による人物再識別器 $(af : S \times 2^S \rightarrow S \cup \epsilon)$ を適用する．ここで，af はある画像 s および画像集合 S が与えられた場合，s に最も類似度が高く，かつ一定の閾値以上の類似度がある S に含まれる画像もしくは不一致シンボル（ϵ）を返す関数であり，提案手法はこれを深層距離学習 ArcFace を事前に（あるデータセットで）訓練した関数を用いる．

$p' = af(p, DB_{k-1})$ において，$p' \neq \epsilon$ であれば，$ID_{final}(p') = ID_{final}(p)$，そうでなければ $ID_{final}(p') = ID_{deep}(p)$（Deep Sort による新しい ID）とし，$DB_k = DB_{k-1} \cup P_k$ とする．

ここで，トラッキングしたい人物が事前に把握できるのであれば，人物検出モデルおよび分類モデルを既存の深層学習等で事前に構築し，各映像フレームごとに各選手を識別することも可能である．つまり，試合に出場する選手全員分のデータを事前に準備できれば，クラス分類問題としてトラッキングを実現することもできる．しかし本研究で想定する利用シーンでは，事前に特定の人物を追跡するためのデータを取得することは困難であることが多い．そこで提案手法では，与えられた 2 つの画像（検出した選手の画像とデータベースにある選手の画像）が同一人物であるかどうかを判定するように学習させた Verification モデルを利用してデータベースとの照合を行う．本研究では，深層距離学習の一種である ArcFace を利用し，Verification モデルを構築する．深層距離学習では，データの特徴量空間における距離を学習することができるため同一人物の画像同士の距離は小さく，異なる人物画像同士の距離は大きくなるように学習しておくことによって，検出された人物と過去に追跡されていた人物同士の特徴距離を計算できる．また，通常のクラス分類モデルに ArcFace 独自のレイヤを追加するだけで容易に構築できる．

3.2 深層距離学習を用いた人物再識別手法

Additive Angular Margin Loss (ArcFace) 関数は顔認識

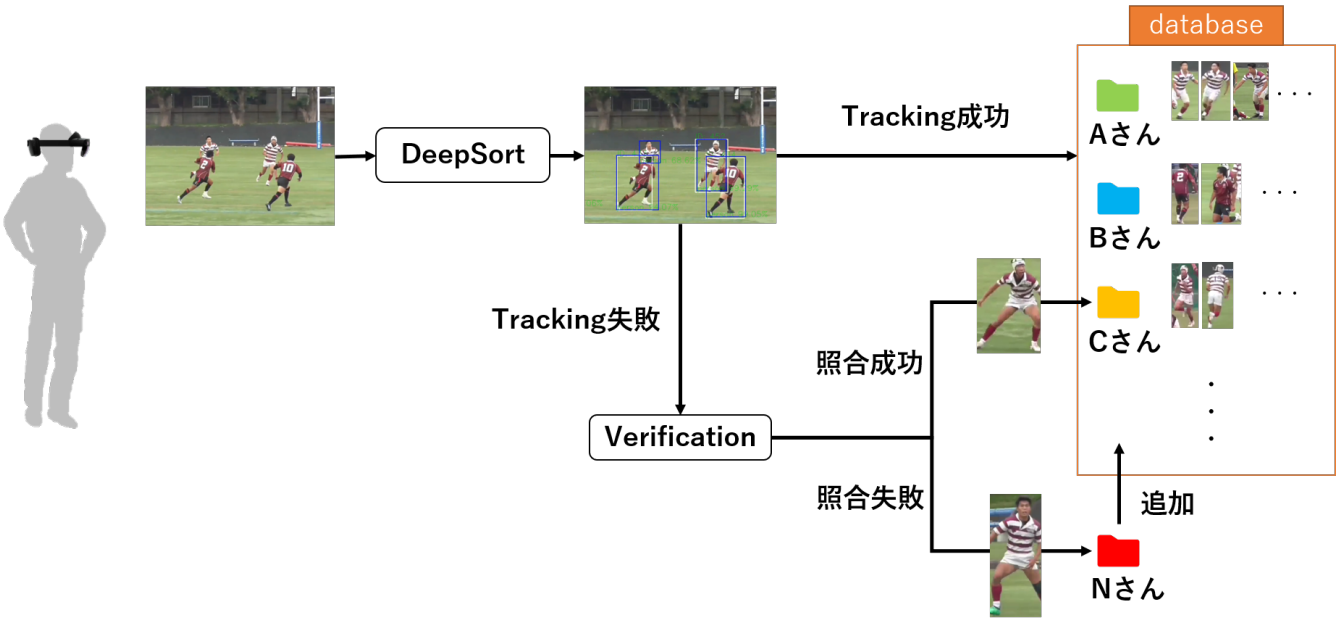

図 2 システム概要

図 3 オクルージョンの例

において高い分類性能を発揮することが示された損失関数である [14]. 距離学習に用いられる損失関数であり，類似度を角度で表現する．提案手法では図 4 に示すように，ResNet50 と ArcFace を組み合わせた深層距離学習で人物再識別を行うための Verfication モデルを構築する．

分類問題でよく利用される損失関数であるソフトマックス関数は次式で表される．

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{1}$$

ここで $x_i \in \mathbb{R}^d$ は i 番目の層から出力される特徴量を示し，図 4 の ResNet50 からの出力に相当する．本稿では ResNet50 から得られる特徴量の次元を 2048 とした．また，$W_j \in \mathbb{R}^d$ は $W \in \mathbb{R}^{d\times n}$ の j 列目の重み，$b_j \in \mathbb{R}^n$ はバイアス項を示す．N はバッチサイズ，n はクラス数を示す．しかし，ソフトマックス関数はクラス内分散を小さくしてクラス間分散を多様化することには適していない．そこで x_i と行列 W を列ごとに L2 正規化し，それぞれのベクトルの積を取ると次のように表せる．

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{2}$$

この式に $(\theta_{y_i} + m)(m > 0)$ のペナルティを加えると次のように表せる．

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos(\theta_j+m)}} \tag{3}$$

ペナルティを与えることで正解クラスに対する logit は小さくなるが，これによってクラス内分散が小さく，クラス間分散が大きくなるように学習を進めることができる [14].

提案手法における人物の再識別方法の概要を図 5 に示す．まず ArcFace を用いて事前学習した ResNet50 に対し，時刻 k で取得した再識別対象の人物画像セグメント p およびデータベース（DB_k）内の全人物画像セグメント（その総数を N とする）を入力とし，特徴量を得る．次に DB_k の N の各特徴量に対し，p の特徴量とのコサイン類似度を算出する．そのうち最大のコサイン類似度が予め定めた閾値を超えた場合，それに対応する画像セグメントと同一人物と判定しトラッキングを継続する．そうでなければ，新たに検出した人物とし新しい ID を付与する．

4. 評価実験

4.1 データセットの生成

本研究の利用シナリオとして挙げた，AR デバイスのスポーツにおける選手トラッキングへの活用においては，同チームの選手全員がほぼ同じユニフォームを着用した状況で人物識別を行う必要があるため，非常に困難なタスクである．しかし，本論文では事前に利用シーンがわからない前提であるため，Person Re-identification の研究分野の評価で利用される一般的なデータセット Market-1501 [1] や CUHK03 [15] で学習したデータを用いて訓練

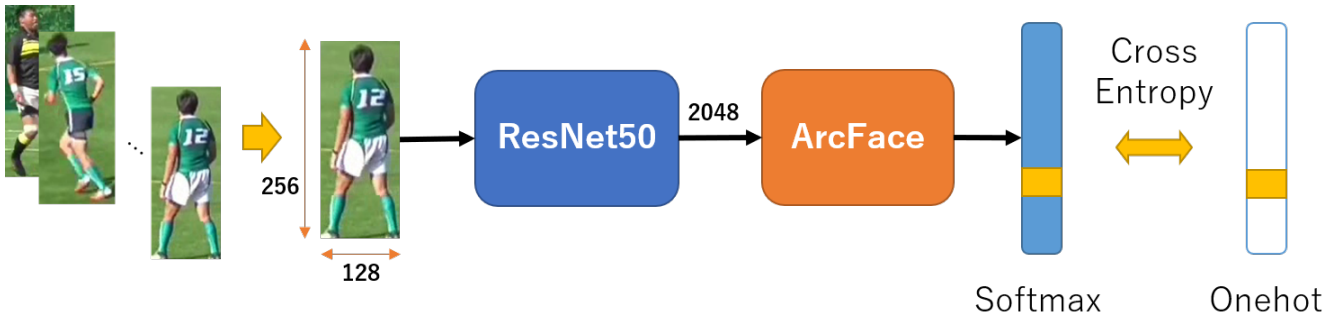

図 4　Verification モデル概要

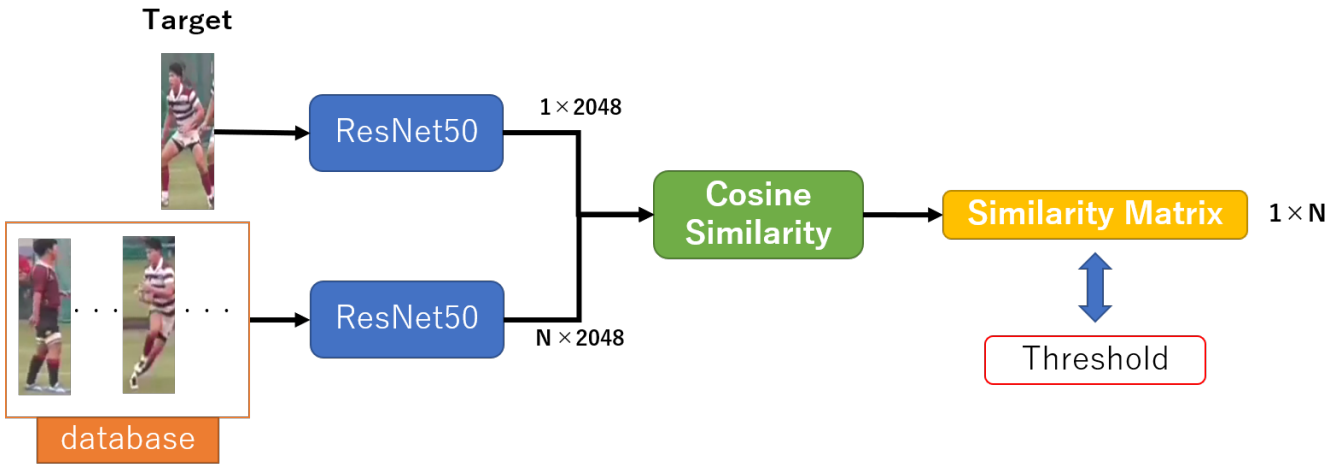

図 5　人物再識別方法概要

したResNet50+ArcFaceを用いた識別実験を実施する．一方で，テストデータとしてはそのシナリオを想定し，ユニフォームを着用したラグビー選手から成る独自のデータセットを作成した．

データセット生成には，ARを想定して撮影した，1試合あたり40分程度のラグビー試合映像を3試合分用いた．これらの動画からDeep Sortを利用して選手画像セグメントを切り出し，Deep Sortでトラッキングが連続して成功している間の選手の画像を1グループ（フォルダ）にまとめて作成した．しかし，単純にDeep Sortを適用すると，トラッキング中に選手間でオクルージョンが発生した場合にIDが入れ替わることがあり，結果として1フォルダに複数選手の画像が混入する．例えば図3はオクルージョン発生時にDeep Sortで切り出された選手画像セグメントであるが，1フレームごとに少しずつ検出対象が替わっていることがみてとれる．そこで作成したフォルダに対し，1フォルダにつき1人の選手データとなるよう，画像のヒストグラムを比較することによる誤分類発見とデータクレンジングを行った．Deep Sortのトラッキング特性上，時間的に近接するフレーム間の画像の差異は微小であるが，離れたフレーム間の画像の差異はその間に入れ替わりが発生していれば大きくなる．そこで，同一フォルダでフレーム間距離が5である2画像のヒストグラムを比較し，一定以上の変化があれば入れ替わり発生とみなし，特定したタイミングの前後でフォルダ分割を行う．これにより同一選手のみを含む画像フォルダ集合を生成した．データクレンジングが完了した総数17532のフォルダに対し，出場している選手とレフリーのラベリングを行った．画像総数は271,606枚である．本研究ではこのデータセット（以下，Rugby3データセットとよぶ）を用いて評価を行った．

4.2　評価指標

評価指標としてF値とRankを利用した．提案システムにおけるRank-kは，k番目以内に正しい照合結果が候補として抽出された率である．本研究ではRank-1，Rank-5，Rank-10を求め，ArcFaceによって同一クラスタに属するデータはResNet50から抽出されるコサイン類似度が高くなることを示す．

4.3　テストデータ

提案手法の性能評価のため，まずRe-identificationの研究分野で利用されるMarket-1501 [1]を利用し検証を行った．学習に751ラベルの12,936データを利用し，テストには750ラベルの3,368データを利用した．これによっ

表 1　評価結果

テストデータ	F 値 (%)	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)
Market-1501	74.15	82.19	94.39	96.79
Rugby3	66.88	61.62	81.31	88.31

て予め用意した Verification モデルを利用して未知の人物の再識別が実現できることを示す．また，スポーツ選手識別のように類似の服装の人物の識別能力の評価を行うため，Market-1501 [1] を用いて学習したモデルを，前述の Rugby3 データセットのうちの 1 試合に相当する 5,789 データを利用して評価を行った．

4.4　評価結果と考察

評価結果を表 1 に示す．F 値はテストデータが Market-1501 のとき 74.15%となった．また，Rank-1，Rank-5，Rank-10 についても Market-1501 での精度が高く，服装が異なる人物に対し，提案するモデルが適用できることが確認できた．一方，Rugby3 では F 値が 66.88%となり，Rank-1，Rank-5，Rank-10 についても Market-1501 に比べると低い値となった．この原因としては以下の 2 点が考えられる．まず，データ量が膨大であり，完全なデータの正当性検証がができていない点が挙げられる．図 6 に示すようにフォルダ内のデータクレンジングを半自動化しているため，検出されない誤分類データも存在し，結果として Market-1501 よりも誤差が大きくなったと考えられる．次に，Market-1501 [1] データセットでは主に服装の色の違いを特徴として学習したため，Rugby データセットのように同じ服装の選手を見分けるタスクには適していない可能性が挙げられる．そのため，訓練にも用いることができるよう，Rugby3 データセットの検証とデータ追加が必要になると考えられるが，アノテーションタスクには負荷が高いため，現在も順次データ整備を行っている．

図 6　データセットの複雑さ

5.　まとめと今後の課題

本研究では，スポーツトレーニング又はスポーツ観戦向け AR アプリケーションのトラックングシステムとして，Deep Sort によるトラッキングと，ResNet50 および ArcFace 損失関数を用いた人物識別のための Verification モデルを組み合わせた手法を提案した．AR デバイス装着者による取得映像に対して Deep Sort を直接適用することは，装着者の視点変更に伴うフレームアウトやオクルージョンにより，同一人物を安定的に追跡することが難しく，そういった状況でのトラッキングは非常に挑戦的なタスクである．既存のデータセットである Market-1501 と 27 万枚以上の画像からなる独自生成した Rugby3 データセットを用いた評価を行った結果，前者では F 値 74.15%，後者では F 値 66.88%を達成した．

今後の課題として，5 章で述べたように，本研究の精度は Market-1501 データセットで学習を行ったモデルによる Rugby3 データセットのテスト結果であり，それらのデータセットの画像特性の乖離度が大きいことから，なるべく近いデータセットで学習したモデルを用いることによる精度向上の余地が大きく残されている．今後，Rugby データセットの検証を進め，充実させることを行っていく予定である．

謝辞

本研究成果は国立研究開発法人情報通信研究機構（NICT）の委託研究「ウイルス等感染症対策に資する情報通信技術の研究開発（課題番号 222）」により得られたものです．

参考文献

[1] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. and Tian, Q.: Scalable person re-identification: A benchmark, *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124 (2015).

[2] Soltani, P. and Morice, A. H.: Augmented reality tools for sports education and training, *Computers Education*, Vol. 155, p. 103923 (2020).

[3] Wojke, N., Bewley, A. and Paulus, D.: Simple Online and Realtime Tracking with a Deep Association Metric, *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp. 3645–3649 (2017).

[4] Bose, B., Wang, X. and Grimson, E.: Multi-class object tracking algorithm that handles fragmentation and grouping, *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2007).

[5] Zhang, L. and van der Maaten, L.: Preserving Structure in Model-Free Tracking, *IEEE Transactions on Pattern Analysis Machine Intelligence*, No. 04, pp. 756–769 (2014).

[6] Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K. and Leal-Taixé, L.: MOT20: A benchmark for multi object tracking in crowded scenes, *arXiv:2003.09003[cs]* (2020). arXiv: 2003.09003.

[7] Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A. and Leibe, B.: MOTS: Multi-Object Tracking and Segmentation, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7934–7943 (2019).

[8] Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W. and Kim, T.-K.: Multiple object tracking: A literature

review, *Artificial Intelligence*, Vol. 293, p. 103448 (2021).

[9] Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y.: Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1302–1310 (2017).

[10] Ristani, E. and Tomasi, C.: Features for Multi-target Multi-camera Tracking and Re-identification, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6036–6046 (2018).

[11] Tang, Z., Naphade, M., Liu, M.-Y., Yang, X., Birchfield, S., Wang, S., Kumar, R., Anastasiu, D. and Hwang, J.-N.: CityFlow: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-Identification, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8789–8798 (2019).

[12] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. and Tian, Q.: Scalable Person Re-identification: A Benchmark, *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124 (2015).

[13] Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S. and Tian, Q.: MARS: A Video Benchmark for Large-Scale Person Re-Identification, *Computer Vision – ECCV 2016* (Leibe, B., Matas, J., Sebe, N. and Welling, M., eds.), Cham, Springer International Publishing, pp. 868–884 (2016).

[14] Deng, J., Guo, J., Xue, N. and Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699 (2019).

[15] Li, W., Zhao, R., Xiao, T. and Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 152–159 (2014).