畳み込みニューラルネットワークによる 距離学習を用いた動画像人物再同定*

加藤直樹** 箱崎浩平** 田磨雅基† 古山純子† 里 雄二† 青木義満**

Deep Metric Learning for Video-Based Person Re-Identification

Naoki KATO, Kohei HAKOZAKI, Masamoto TANABIKI, Junko FURUYAMA, Yuji SATO and Yoshimitsu AOKI

This paper proposes a novel approach for video-based person re-identification that exploits deep convolutional neural networks to learn the similarity of persons observed from video camera. By Convolutional Neural Networks (CNN), each video sequence of a person is mapped to a Euclidean space where distances between feature embeddings directly correspond to measures of person similarity. By improved parameter learning method called Entire Triplet Loss, all possible triplets in the mini-batch are taken into account to update network parameters at once. This simple change of parameter updating method significantly improves network training, enabling the embeddings to be further discriminative. Experimental results show that proposed model achieves new state of the art identification rate on iLIDS-VID dataset and PRID-2011 dataset with 78.3%, 83.9% at rank 1, respectively.

Key words: computer vision, pattern recognition, machine learning, convolutional neural networks, person re-identification, video recognition, metric learning

1. 序 論

視野を共有しない複数台のカメラで観測された人物の対応付けを行う課題は人物再同定と呼ばれる。人物再同定はカメラ間人物追跡や映像からの人物検索システムに応用可能なコンピュータビジョンにおける重要課題である。しかし、異なるカメラで得られた人物画像はカメラ視点、人物姿勢、照度、背景環境、遮蔽物などに大きな変動を持つため、同一人物同士を正しく対応づけることは容易ではない。また、学習データに含まれるクラスをテストの際にも用いる分類問題とは異なり、再同定では学習データに含まれない人物同士の類似度を定義する必要がある。人物を区別するための適切な類似度指標を得るためには人物画像が持つ細かな特徴の違いに着目する必要がある。非常に難度の高いタスクとなっている。既存の人物再同定手法の多くは静止画像を用いたものであるが、本稿では動画像データに基づいた人物再同定手法を提案する。これを動画像人物再同定と呼称する。実用上のシステムでは監視カメラなどによって人物情報が得られるため、動画像データを用いるのは自然な発想であるといえる。動画像データに基づいた人物再同定には静止画像の場合と比較し優位性が存在する。動画像データは複数枚の静止画像によって構成されているため、単純に含まれている情報量が多く、様々なカメラ視点、人物の体向きや姿勢、背景環境や遮蔽物などを含んだ画像の集合となっている。人物画像から特徴を抽出する際に、このような複数フレームから得られる様々な見えの情報を統合することによって、同一人物と異なる人物を区別し易くなるはずである。

従来手法として、手動設計特徴を用いて人物間の距離関数の

Fig.1 Output distances of our network between each pair of sequence. Vertically aligned pairs are the same person and horizontally aligned pairs are different persons. Larger distance value indicate that the pair is dissimilar. A threshold of 1.3 classifies all the pairs properly

学習を行う手法がある¹⁾。このような手法では特徴量の設計が重要であるが、あるデータに対して有効であった特徴量が他のデータには上手く作用しないことがあり、特徴量設計のデータセット依存が大きという欠点がある。

本稿では畳み込みニューラルネットワークを識別モデルとして利用する。ニューラルネットワークは生物の神経細胞を工学的に模倣した識別モデルである。深層学習とも呼ばれる階層的な構造を持ったニューラルネットワークは複雑な非線形関数を近似することが可能であり、その汎化性能の高さから近年大いに注目されている。ニューラルネットワークに生物の視覚野中の受容野をモデル化した構造である畳み込み層を導入したものは畳み込みニューラルネットワーク (Convolutional Neural Networks; CNN) と呼ばれ、コンピュータビジョンの分野で専ら利用されている。CNN が利用されるタスクとして、物体認識²⁾、物体検出³⁾、領域分割⁴⁾などが挙げられ、いずれのタスクにおいても高精度な物体認識を可能としている。CNN は入力からの特徴抽出の方法を学習によって最適化するため、手

* 原稿受付 平成 29 年 5 月 8 日

掲載決定 平成 29 年 8 月 31 日

** 慶應義塾大学 (神奈川県横浜市港北区日吉 3-14-1)

† パナソニック株式会社 (神奈川県横浜市都築区佐江戸町 600 番)

動設計特徴を用いた識別手法と比べ、より汎用的な特徴を用いた識別が可能となり、それが高精度な認識に寄与していると考えられる。CNN は物体同士の類似度をモデル化するタスクに対しても用いられ⁵⁾⁶⁾、人物再同定もそのタスクの一種であると見做すことができる。CNN を人物再同定に対して適用した研究事例は数多く存在するものの^{7)~10)}、その大半は静止画像に基づいた手法である。動画画像データを用いた人物再同定に対して深層学習のアプローチを用いた手法は McLaughlin ら¹¹⁾、Zheng ら¹²⁾、Wu ら¹³⁾ によるもの等少数の事例に限られる。それらの中でも McLaughlin らの手法が特に高精度な再同定を達成している、彼らは再帰型ニューラルネットワーク (Recurrent Neural Networks; RNN) を用いた人物画像シーケンスからの時系列特徴の抽出、Contrastive Loss¹⁴⁾ という類似度学習のための誤差関数と学習データ中の人物を一意に分類するクロスエントロピーロスを組み合わせたネットワークのパラメータ学習による動画画像人物再同定手法を提案している。この手法において RNN を用いたことによる精度向上はあまり大きくなく、ネットワークの入力として RGB 画像だけでなく連続するフレーム間のオプティカルフロー画像を組み合わせた場合に RNN の寄与が確認されている。また、物体の類似度学習のタスクにおいては Triplet Loss⁶⁾¹⁵⁾ や Lifted Structured Loss⁵⁾ などの Contrastive Loss よりも優れた誤差関数が提案されており、ネットワークのパラメータ学習の方法についても改善の余地が見られる。

本稿では CNN による距離学習を用いた動画画像人物再同定手法を提案する。本手法では CNN 中の各層で入力画像を非線形変換していく際に得られる特徴マップにおいて、複数フレームから得られる情報を統合する。この方法では人物シーケンスの時系列的な依存関係を学習することはできないが、それでも十分に動画画像データの恩恵を受けられることを実証する。また、ネットワークのパラメータ学習のための誤差関数として、Triplet Loss をベースとしパラメータ更新の仕方に工夫を入れた Entrie Triplet Loss を提案する。この誤差関数は、ミニバッチと呼ばれる学習サンプル内の全てのポジティブペア、ネガティブペアを考慮してパラメータ更新を行う Lifted Structured Loss より洞察を得ており、これを Triplet を利用したシンプルなパラメータ更新の枠組みへと変更を加えている。具体的には、ネットワークの一度のパラメータ更新の際に、ミニバッチ内の全ての組み合わせの Triplet と呼ばれる 3 つ組を考慮してパラメータ更新を行う。これにより、効率的かつ高精度なネットワーク学習が可能となる。ネットワークの学習の結果、入力である人物シーケンスから特徴抽出がなされ、最終的に低次元な埋め込みベクトルが出力される。埋め込みベクトルが写像される特徴空間をユークリッド空間により定義し、人物間の類似度を埋め込みベクトル同士のユークリッド距離によって計算する。距離学習によるネットワークの最適化により、図 1 のように、同一人物、異なる人物を距離によって区別できるような埋め込みベクトルが学習される。人物シーケンスの埋め込みベクトルが一度得られれば、人物同士の類似度は低次元な埋め込みベクトル同士のユークリッド距離により算出されるため、非常に高速な類似度計算が可能である。

提案手法の有効性を検証するために、動画画像人物再同定において広く用いられる公開データセットである iLIDS-VID データセット、PRID-2011 データセットを用いた評価実験を行った。既存の再同定手法と提案手法の精度比較を行い、我々の手

法により最先端手法を凌ぐ精度での再同定が可能となることを示す。また、静止画像を用いた場合と動画画像を用いた場合での精度比較を行い、人物再同定を動画画像データに基づいて行うことの有効性を検証する。さらに、ネットワークのパラメータ学習を複数の誤差関数を用いて行い、我々の提案する Entire Triplet Loss により効果的な埋め込みベクトルの学習がなされることを示す。

以降の本稿の構成は以下の通りである。第 2 章では提案手法の詳細について解説する。続く第 3 章では評価実験による提案手法の有効性の検証と可視化実験を行う。最後に第 4 章で本研究により得られた知見と課題点をまとめる。

2. 提案手法

2.1 システム構成

本節では提案手法の全体構成について説明する。本稿では畳み込みニューラルネットワークを利用し、人物間の類似度を学習することによる動画画像人物再同定手法を提案する。ネットワークの入力の動画画像データである人物画像シーケンスは $X = [x_1, \dots, x_T]$ と表される。このとき、 x_t はフレーム t における人物画像を表す。シーケンスの各フレームはネットワークに入力され、特徴抽出がなされる (2.2 節で詳細を述べる)。CNN は畳み込み層やプーリング層、全結合層などの層を持っており、入力は各層で処理され、最終的にネットワークの出力として人物シーケンスの埋め込みベクトル $f(X)$ が出力される。この特徴抽出の処理は入力ベクトルを特徴空間へと写像する非線形関数 $f(\cdot)$ の適用であると見做すことができる。我々はこの特徴空間をユークリッド空間によって定義する。これにより、2 つの人物シーケンス X_i, X_j の距離 $D_{i,j}$ は次式で表される。

$$D_{i,j} = \|f(X_i) - f(X_j)\|_2^2 \quad (1)$$

$\|f(X_i) - f(X_j)\|_2^2$ は人物シーケンス同士のユークリッド距離である。埋め込みベクトル同士の距離 $D_{i,j}$ が大きいほど 2 人の人物の類似度が低く、距離が小さいほど人物間の類似度が高いものとする。カメラによって観測される複数の人物シーケンスが存在するとき、同一人物のシーケンスから得られる埋め込みベクトル間の距離が小さく、異なる人物同士のベクトル間の距離が大きくなっていけば、それらをユークリッド距離を用いて区別することができる。我々の目的は、そのような人物の埋め込みベクトルを得るための写像関数 $f(\cdot)$ を取得することである。CNN は畳み込み層や全結合層に重み W やバイアス b と呼ばれるパラメータを持っており、これらの値を教師データを用いた学習により最適化する。写像関数 $f(\cdot)$ の取得はネットワークのパラメータ学習によって実行される。一度同一人物と異なる人物をユークリッド距離によって区別できる埋め込みベクトルが得られれば、人物の分類問題をユークリッド空間における最近傍探索問題に置き換えることができる。本手法で用いるような距離学習の大きな利点は、学習データに含まれていないクラスに関する分類問題を類似度指標の利用により解くことができるようになることである。ネットワークのパラメータ学習のためには適切な誤差関数を用いる必要がある。距離学習のためのいくつかの誤差関数が既存研究により提案されているが、本研究では Entrie Triplet Loss と呼ばれる新たな誤差関数を利用する。パラメータ学習に関する詳細は 2.3 節で説明する。

Fig.2 Network architecture of our convolutional neural networks. Note that pooling layers and activation functions in the first and second layers are omitted in the figure

2.2 特徴抽出

図2に我々の用いるCNNのネットワーク構成の概略、表1にネットワークの詳細パラメータを示す。このネットワークは4層の畳み込み層、全結合層から成り、近年主流となっているVGG¹⁶⁾やResidual Networks¹⁷⁾などのモデルと比べ浅い構造となっている。このような設計としたのは、実験において用いることのできる学習データ量が少なく、これ以上ネットワークの規模を大きくするとパラメータ増加による過学習により精度の低下が確認されたためである。汎化性能の高いモデルを設計するためには、パラメータ数の増加を抑えつつ表現力の高いネットワーク構成にすることが学習データが少ない場合とりわけ重要である。

各層の詳細について説明する。Conv1, Conv2, Conv3層は畳み込み層である。畳み込み層では、入力に対してフィルタの畳み込みを行い、特徴マップを出力する。フィルタは重みをパラメータとして持ち、入力から特徴的なパターンを抽出する効果がある。Conv1層、Conv2層の後には活性化関数としてRectified Linear Unit (ReLU) 関数を配置する。ReLU関数は $\sigma(x) = \max(0, x)$ と表現され、入力を非線形に変換する役割を持つ。MaxPool1, MaxPool2, AvgPool3層はそれぞれマックスプーリング層、平均プーリング層である。これらの層は入力画像の局所領域内で最大値、平均値を出力する。プーリング層により、畳み込み層で出力された特徴の位置感度を低下させ、入力の位置ずれに頑健な特徴抽出が可能となる。FC4層は全結合層である。この層は、入力ユニットと出力ユニットが全結合しており、その全ての結合に重みパラメータが存在する。L2層では、入力に対して $n(x) = x/\|x\|_2$ で表されるL2正規化を行う。これによって出力である埋め込みベクトルのL2ノルムが1に正規化される。

人物シーケンスの複数フレームの特徴を用いることにより、再同定の精度向上を図る。入力シーケンスの全フレームに対し、AvgPool3層により出力される特徴マップの平均をとりFC4層に入力することにより、シーケンス全体の情報を持った埋め込みベクトルを取得する。

2.3 パラメータ学習

2.2節で説明したネットワークは、畳み込み層のフィルタと全結合層のユニット同士の結合に重み W 、バイアス b のパラメータを持ち、これを学習によって決定する必要がある。このパラメータ学習は、誤差逆伝播法を用いて行われる。誤差逆伝

Table1 Network parameter detail of proposed model. The input and output sizes are denoted as *height* × *width* × *channels*. The kernel and the stride sizes are described by *height* × *width*

layer	size-in	size-out	kernel	stride
Conv1	112 × 48 × 3	112 × 48 × 24	3 × 3	1 × 1
MaxPool1	112 × 48 × 24	56 × 24 × 24	2 × 2	2 × 2
Conv2	56 × 24 × 24	56 × 24 × 24	3 × 3	1 × 1
MaxPool2	56 × 24 × 24	28 × 12 × 24	2 × 2	2 × 2
Conv3	28 × 12 × 24	28 × 12 × 24	3 × 3	1 × 1
AvgPool3	28 × 12 × 24	14 × 6 × 24	2 × 2	2 × 2
FC4	1 × 1 × 2016	1 × 1 × 128		
L2	1 × 1 × 128	1 × 1 × 128		

播法では、学習データの教師信号との誤差を各層に逆伝播させ、誤差が小さくなる方向へと各パラメータを更新する。学習にどのような誤差関数を用いるかが、ネットワークの構造の設計と並んで重要である。

2.3.1 パラメータ学習手法の概観

本節では、人物の埋め込みベクトルをユークリッド距離により区別できるようなネットワークのパラメータを得る、距離学習のための誤差関数について説明する。

Contrastive Loss: Contrastive Lossは人物の埋め込みベクトルのペア $(f(X_i), f(X_j), y_{ij})$ を用いる誤差関数である。 $y_{ij} \in \{0, 1\}$ は、 i, j が同一人物なら1を、異なる人物なら0を返すラベルである。この誤差関数は2つの埋め込みベクトル同士のユークリッド距離の2乗を、それらが同一人物から得られた埋め込みベクトルであれば0に、異なる人物のものであればマージンのパラメータよりも遠ざけようとする設計となっている。ミニバッチを用いた学習において、この誤差関数は次のように定義される。

$$L = \frac{1}{2N} \sum_{(i,j)} y_{ij} D_{i,j} + (1 - y_{ij}) \max(\alpha - D_{i,j}, 0) \quad (2)$$

ここで、 N はミニバッチサイズ、 $D_{i,j} = \|f(X_i) - f(X_j)\|_2^2$ 、 α は負例に対して適用されるマージンである。学習の際は、全てのポジティブペアと、それと同数のランダムに選択したネガティブペアを交互に用いてパラメータ更新を行い、それを1エポックとする。この誤差関数は対象人物の見え方や背景環境、遮蔽などにかかわらず、同一人物から得られる埋め込みベクトルが埋め込み空間における一点へと写像されるまで誤差を与え続ける。埋め込みベクトルの写像点に自由度を与えていないため、学習データに対して過学習しやすくなってしまふ。また、入力の組がポジティブペアかネガティブペアかによって目的関数変動するため学習が収束しにくくなってしまふ。

Triplet Loss: Triplet LossはTripletと呼ばれる3つ組 $(f(X_a), f(X_p), f(X_n))$ を用いて誤差計算を行う。ここで、 $f(X_a)$ は基準となる人物（アンカー）の埋め込みベクトル、 $f(X_p)$ はアンカーと同一人物（ポジティブ）の埋め込みベクトル、 $f(X_n)$ はアンカーと異なる人物（ネガティブ）の埋め込みベクトルである。この誤差関数はTriplet内の負例同士のユークリッド距離の2乗が正例同士のそれよりもマージンパラメータよりも大きくなるよう誤差を与える。ミニバッチ法の各イテレーションにおける誤差は次式で表される。

$$L = \frac{1}{N} \sum_{(a,p,n)} \max(D_{a,p} - D_{a,n} + \beta, 0) \quad (3)$$

Fig.3 Examples of sequence pairs in the datasets investigated in this study. Each pair of sequences denotes the same person observed from different cameras

ここで、 N はミニバッチサイズ、 $D_{a,p} = \|f(X_a) - f(X_p)\|_2^2$ 、 $D_{a,n} = \|f(X_a) - f(X_n)\|_2^2$ 、そして β はマージンである。学習データ中の 1 人の人物に対して複数のカメラから得られたサンプルが存在するが、1 エポックの学習において、それぞれの人物におけるアンカーサンプルをランダムに決定し、ポジティブサンプルはアンカーと同一人物の他カメラから取得されたサンプルをランダムに 1 つ選択する。ネガティブサンプルはアンカーと異なる人物の全てのカメラで取得されたサンプルの中からランダムに 1 つのサンプルを選択する。ミニバッチを用いた学習において、このようにして形成された Triplet をミニバッチサイズの数だけランダムな順番で選択を行い、個々のミニバッチを形成し、パラメータ更新を行う。この誤差関数は正例、負例の相対的な距離に対して誤差を与えるため、同一人物の埋め込みベクトルは他の人物と距離を保った状態において多様な点へと写像されることが許容されている。このような関数設計により、同一人物の埋め込みベクトルであっても人物の見え方の違いや背景環境の違いによって写像点変動し、埋め込みベクトルがこれらの特徴も含んだものとなることが期待される。結果的に、画像内の細かな特徴に対応した類似度指標が獲得できると考えられる。

2.3.2 Entire Triplet Loss によるパラメータ学習

我々の提案する誤差関数は、Triplet Loss と本質的に同一の目的関数となっているが、パラメータ更新の仕方に変更を加えている。通常の Triplet Loss では、ミニバッチサイズを N とするとき、 N 個の Triplet を用いて誤差、勾配を計算し、パラメータの更新を行う。一方、提案手法ではミニバッチサイズを誤差を計算する際に利用する人物の数と定義する。そのため、この方法では一回のパラメータ更新の際に、対象とする人物の全てのカメラで得られた埋め込みベクトルを用いる。各ミニバッチに対する誤差関数は、次式のようにミニバッチ内の考えられる全ての組み合わせの Triplet から求められる Triplet Loss の平均により定義される。

$$L = \frac{1}{|\mathcal{T}|} \sum_{(a,p,n) \in \mathcal{T}} \max(D_{a,p} - D_{a,n} + \beta, 0) \quad (4)$$

ここで、 $\mathcal{T}$ はミニバッチ内の全通りの Triplet である。ミニバッチは人物数に基づき形成され、1 エポックの学習において学習データ中の全ての人物が一度ずつ選択される。我々はこの誤差関数を Entire Triplet Loss と呼称する。この誤差関数を用いることにより、一回毎のパラメータ更新の際に考慮される Triplet の数が著しく増加する。人物サンプルを取得するカメラの台数を K とするとき、Entire Triplet Loss の一度のパラメータ更新の際に用いられる Triplet の数は $K P_2 K N (N - 1)$ となる。こ

のように、考慮される Triplet の数は N^2 のオーダーで増加するが、比較的計算コストの大きい CNN による特徴抽出が必要なサンプルの数は $K N$ であり、カメラ台数、ミニバッチサイズの増加に対して線形となっている。それに対し、Triplet Loss の場合、一度のパラメータ更新の際に考慮される Triplet の数はミニバッチサイズの N で、特徴抽出が必要なサンプル数は $3N$ である。例えば、本稿の実験におけるパラメータ設定であるミニバッチサイズ $N = 64$ 、人物を観測するカメラの台数が $K = 2$ の場合、Entire Triplet Loss の誤差計算の際に用いられる Triplet の数は 16,128 に上る。このとき、CNN による特徴抽出が必要なサンプルの数は 128 であり、通常の Triplet Loss の際の 192 よりも少なくなっている。Entire Triplet Loss において誤差計算は特徴抽出後の低次元な埋め込みベクトルに対して行われるため、組み合わせの数が増大しても計算コスト、メモリ消費共にあまり大きくはならず、効率的に行うことができる。データセット中の Triplet の組み合わせの数は一般的に非常に多くなるため、通常の Triplet Loss では全ての Triplet を用いて学習を行うためにはたくさんのイテレーションを必要とする。それに対し、我々の Entire Triplet Loss は大量の Triplet を考慮して誤差の計算を行うため、少ないイテレーションの数で効率的に学習を行うことができる。また、一つの Triplet が勾配に対して与える影響が小さくなるため、少数のサンプルを用いた学習の際にはノイズとなっていた成分が打ち消され、全体として正しい勾配を得やすくなる。この誤差関数により、明示的なハードネガティブマイニングを行わなくても効果的なパラメータ学習を行うことができる。本稿の実験ではカメラの台数 $K = 2$ の場合のみを扱っているが、カメラ台数が増加すると Triplet の組み合わせ数が膨大になってしまうことが考えられる。このような場合、サンプルの取得に用いるカメラの数を限定することにより Triplet の組み合わせ爆発を防ぐ必要があると考えられる。

3. 実 験

本章では提案手法の有効性を検証するため、2 つのデータセットを用いて複数の比較項目について定量的な精度評価を行う。まず、本節では実験に用いるデータセットの詳細、評価手法、実装に関する詳細設定について説明する。3.1 節、3.2 節、3.3 節では各比較項目の実験結果の提示とそれに関する考察を行う。また、3.4 節では本手法により得られる人物の埋め込みベクトルの可視化実験を行う。

データセット：図 3 に本実験で用いる 2 つのデータセットに含まれる人物シーケンス例を示す。いずれのデータセットも動画人物再同定の精度評価に広く利用される公開データセッ

Fig.4 CMC curves of our models evaluated on the iLIDS-VID and PRID-2011 datasets

Fig.5 Learning curves of rank 1 identification rate, evaluated on iLIDS-VID dataset

Fig.6 CMC curves of our models trained with Entire Triplet Loss as a function of sequence lengths, evaluated on the iLIDS-VID dataset

Table2 The effect of changing mini-batch size in Triplet Loss and Entire Triplet Loss

Loss function	Triplet Loss				Entire Triplet Loss			
	8	16	32	64	8	16	32	64
Mini-batch size								
Memory consumption MiB	274	328	446	682	256	293	377	552
Time of single iteration ms	25	31	48	79	71	132	262	518
Identification rate @Rank 1 iLIDS VID %	54.0	59.5	61.1	61.7	75.5	75.7	76.6	78.3
Identification rate @Rank 1 PRID-2011 %	71.7	74.6	75.8	75.8	82.1	82.4	83.2	83.9

トである。iLIDS-VID データセット¹⁸⁾¹⁹⁾は空港において視野を共有しない2台のカメラを用いて追跡された人物シーケンスのデータセットとなっている。このデータセットは300人物の画像シーケンスから成り、それぞれの人物に対して2つの異なるカメラ視点から得られたシーケンスが存在する。各人物画像シーケンスは異なる長さとなっており、フレーム数は23フレームから192フレームであり、平均73フレームとなっている。このデータセットは異なる人物の服装の類似や、カメラの視点変化、画像の照度変化、背景変化、オクルージョンを含んでおり、とても難度の高いデータセットとなっている。PRID-2011 データセット²⁰⁾は2つのカメラで観測された200人物のシーケンスを含んでいる。シーケンスの長さは5フレームから675フレームであり、平均100フレームである。このデータセットは混雑が少なく比較的明瞭な背景環境であり、遮蔽物も少ないため、iLIDS-VID データセットと比較すると難度は低くなっている。

評価手法: 人物再同定で広く利用される Cumulative Matching Characteristics (CMC) カーブを用いて各モデルの定量的な精度評価を行う。まずデータセットの50%の人物を学習データ、残りの50%の人物をテストデータにランダムに分割する。カ

メラ1から得られるシーケンスをクエリ、カメラ2のシーケンスをギャラリーと設定する。各クエリシーケンスに対して、全てのギャラリーシーケンスとの距離を埋め込みベクトル間のユークリッド距離を用いて計算する。クエリシーケンスとの距離が近い順にギャラリーシーケンスを並び替えたとき、クエリシーケンスと同一の人物が全体の中で k 番目以内に位置する割合をランク k における再同定率と定義する。再同定率が高いほど、人物間の類似度指標が適切に得られていることを表す。各モデルの学習、テストは10回行い、その結果の平均を評価値として報告する。

実装詳細: 深層学習では一般的に、汎化性能の高い識別モデルを得るために大量の学習データが必要である。学習データの量が少ない場合、ネットワークは少ない学習サンプルに最適化しよう学習されテストデータに対する精度が出ない、過学習と呼ばれる現象を引き起こしてしまう。そのため学習データが持つ意味情報が失われない程度に線形変換などを施すことにより学習データの数を増大させる、データ拡張と呼ばれる手法が一般的に利用される。我々は画像の水平方向の反転、上下左右3%以内のクロップを均等な確率でランダムに行うことでデータ拡張を行った。

Table3 Comparison of our method with the state of the art methods

Dataset	iLIDS-VID				PRID-2011			
	1	5	10	20	1	5	10	20
CMC rank								
DVR ¹⁹⁾	39.5	61.1	71.7	81.0	40.0	71.7	84.5	92.2
STFV3D + KISSME ¹⁾	44.3	71.7	83.7	91.7	64.1	87.3	89.9	92.0
SI ² DL ²²⁾	48.7	81.1	89.2	97.3	76.7	95.6	96.7	98.9
TDL ²³⁾	56.3	87.6	95.6	98.2	56.7	80.0	87.6	93.6
AvgTAPR ²⁴⁾	55.0	87.5	93.8	97.2	68.6	94.6	97.4	98.9
Deep RCN + KISSME ¹³⁾	46.1	76.8	89.7	95.6	69.0	88.4	93.2	96.4
CNN + XQDA ¹²⁾	53.0	81.4	-	95.1	77.3	93.5	-	99.3
RNN + Contrastive ¹¹⁾	58.0	84.0	91.0	96.0	70.0	90.0	95.0	97.0
CNN + Entire Triplet	78.3	92.9	97.6	99.3	83.9	96.3	99.2	100.0

全ての実験において、CNN のネットワーク構造は第 2.2 節で説明したものを採用している。また、各ハイパーパラメータは以下のように設定した。ミニバッチサイズ $N = 64$ 、埋め込みベクトルの次元数 $d = 128$ 、誤差関数におけるマージンパラメータ $\alpha = 2.0$ 、 $\beta = 0.2$ 。

ネットワークの入力であるシーケンス中の各フレームは RGB 画像とし、画像中の画素値の平均が 0、分散が 1 となるよう正規化を行った。また、正則化のためにドロップアウト層をドロップ率 $p = 0.5$ として全結合層の前に配置した。ネットワーク学習の際はそれぞれの人物シーケンス中の 1 フレームのみをランダムに選択してミニバッチを形成し、誤差計算を行った。パラメータ学習における誤差関数を最小化するための最適化手法として Adam²¹⁾ を学習率 0.001 で使用した。誤差逆伝播法を用いて各層における勾配を計算するが、このとき全ての勾配が $[-5, 5]$ の範囲に収まるよう、勾配のクリッピングを行った。全ての比較項目において、各モデルの学習は 100,000 イテレーション行ったが、これには NVIDIA TITAN X を一台用いておよそ 15 時間を要した。本実験には深層学習ライブラリの Chainer を採用している。

3.1 パラメータ学習手法の比較

第 2.3 節では CNN による距離学習を実現するための誤差関数についての解説、そして我々の提案する Entire Triplet Loss と呼ばれる新たな誤差関数の説明を行った。本節では、CNN のパラメータをこれらの誤差関数で学習させたときの再同定率の変化について実験する。2つのデータセットに対する CMC カーブを図 4 に示す。どちらのデータセットにおいても、CMC カーブは Contrastive Loss、クロスエントロピーロス、Triplet Loss、そして Entire Triplet Loss の順で精度が向上している。誤差関数を変えた場合の精度の違いは大きく、これは特に低いランクにおいて顕著である。この結果より誤差関数の選択によってモデルの性能が大きく左右されることが確認できる。特に着目すべき点は、Triplet Loss と Entire Triplet Loss の精度の差の大きさである。Entire Triplet Loss はベースとする目的関数は Triplet Loss と同一であるが、ネットワークの一度のパラメータ更新を、ミニバッチ内の全通りの Triplet から計算される勾配に基づいて行う。一見すると両者の差はあまり大きくないように思えるが、パラメータ更新の方法の単純な変更により、精度が大幅に向上することが確認された。このような精度向上の理由として、一度に大量の Triplet を考慮して誤差を計算することにより、得られる勾配の方向が安定するものと考えられる。

また、iLIDS-VID データセットにおける学習イテレーション数とランク 1 の再同定率の関係を図 5 に示す。クロスエント

ロピーロスは距離学習の誤差関数と比較し過学習が顕著に現れている。Entire Triplet Loss は学習が収束したときの精度が高いだけでなく、Triplet Loss よりも学習の進む効率も改善していることが確認できる。triplet を用いた学習の場合、データセット中の Triplet の組み合わせ数は非常に大きくなる。通常の Triplet Loss の場合、一度のパラメータ更新の際にミニバッチサイズの数の Triplet しか考慮できないため、学習データ中の全組み合わせの triplet を網羅するためには非常に多くのパラメータ更新回数を必要とする。また、各 Triplet が勾配に与える影響が大きいため、最終的な収束性もあまりよくない。それに対し、Entire Triplet Loss では個々の Triplet から求められる勾配の影響が小さく、学習データの組み合わせを効率的に周回することができるため、学習の効率、最終的な精度共に Triplet Loss よりも優れた結果となったと考えられる。

表 2 に Triplet Loss、Entire Triplet Loss においてミニバッチサイズを変化させた時のメモリ消費量、学習時の 1 イテレーションにかかる時間、それぞれのデータセットにおけるランク 1 の再同定率をまとめて示す。Entire Triplet Loss ではミニバッチサイズ N を増加させると考慮される Triplet の数が N^2 のオーダーで増加するが、計算コストの大きな CNN による特徴抽出が必要なサンプル数は N の増加に対し線形な関係となっている。また、個々の triplet を用いた誤差計算は特徴抽出後の低次元ベクトルを用いて行われる。そのため、1 イテレーションにかかる時間は Triplet Loss と比べると比較的大きくはなるもののミニバッチサイズの増加に対しほぼ線形な関係となっていることがわかる。また、本実験でのカメラの台数は $K = 2$ であり、ミニバッチサイズが同じ場合、Triplet Loss よりも Entire Triplet Loss の方が特徴抽出に必要なサンプル数が少なくなっている。これにより Entire Triplet Loss ではメモリ消費量を少なく抑えて学習を行うことが可能となっている。また、いずれのロス関数の場合も、ミニバッチサイズを大きくするほど再同定率が向上することが確認された。しかし、Entire Triplet Loss の方が再同定率が大きく上回っており、Triplet Loss におけるミニバッチサイズを単純に大きくしても、Entire Triplet Loss の再同定率には届かないと思われる。これは、Entire Triplet Loss では限られた人物の組の中での全通りの Triplet を用いて誤差、勾配を計算するため、汎用的な特徴抽出に寄与しない、ノイズとなる勾配の成分同士が効果的に打ち消しあうことにより、各パラメータ更新が有効に機能するためであると考えられる。

3.2 シーケンスのフレーム数による精度比較

ここではシーケンスのフレーム数を変化させたときの再同定率の変化を検証する。この実験では iLIDS-VID データセット

を用い、テストデータに含まれるシーケンスの先頭から必要なフレーム数 l (1, 8, 32, 全フレーム, とする) の分だけのシーケンスを切り出して評価に用いるシーケンスとする。データセットのシーケンスのフレーム数が所望するフレーム数に満たない場合、用いることのできる全てのフレームを用いてテストを行う。直感的には、フレーム数を多くするほど精度は向上すると考えられる。図 6 に示す実験結果より、想定通りフレーム数を増やすに従い再同定率が向上することが確認できる。フレーム数を増加させることによる精度向上は顕著であり、静止画像による再同定と同義である $l=1$ の場合は 33.5% であったランク 1 における再同定率が、テストデータの全フレーム (平均 73 フレーム) を用いることにより 78.2% まで向上している。本手法では人物シーケンスが持つ時系列の情報を、特徴マップの平均によって埋め込みベクトルへと組み込んでいる。この方法では時系列的な人物や背景の動きの特徴を抽出することはできていないが、それでもこのように静止画像ベースの場合と比較し大きく精度が向上することが確認された。このような結果となった理由として、埋め込みベクトルが人物の複数の見えの情報をを持ったベクトルとなったこと、また複数フレームの平均をとることにより埋め込みベクトル中の背景や遮蔽などによるノイズ成分の影響が低減されたことなどが挙げられる。

3.3 最先端手法との精度比較

我々の手法と既存の動画画像人物再同定の最先端手法の精度比較を行う。比較手法の詳細に関する説明は本稿では行わないが、¹⁾¹⁹⁾²²⁾²³⁾²⁴⁾ は手動設計特徴を用いた手法で、^{11)~13)} は深層学習のアプローチを用いた手法である。2つのデータセットでの精度比較結果を表 3 に示す。結果より、両方のデータセットにおいて、我々のモデルが他の全ての手法を大きく上回る精度となっていることがわかる。特に、iLIDS-VID データセットにおける精度の向上が顕著である。既存手法の中で最も iLIDS-VID データセットで高い精度を成している McLaughlin らの手法は、RNN を Contrastive Loss, クロスエントロピーロスを組み合わせて学習させている。我々の手法が彼らの手法を大きく上回る精度となったのは、Entire Triplet Loss の導入による効果的な距離学習による影響が大きいと考えられる。本手法の特徴抽出の部分に改良を施し、時系列特徴を埋め込みベクトルへと適切に組み込むことのできるモデルを構築すれば、更なる精度向上を見込めるはずである。

3.4 埋め込みベクトルの可視化

Entire Triplet Loss を用いた距離学習により、ユークリッド空間において同一人物同士が近く、異なる人物同士が遠くの点へと写像される。本節では CNN による特徴抽出という形で特徴空間へと写像された複数人物の埋め込みベクトルの関係性の可視化を試みる。可視化のため、t-Distributed Stochastic Neighbor Embedding (t-SNE)²⁵⁾ を用いて 128 次元の埋め込みベクトルを 2 次元ベクトルへと次元削減する。t-SNE は高次元データの構造を保ったまま次元を削減できるため、データの可視化によく用いられるアルゴリズムである。

iLIDS-VID データセットのテストデータ中の 100 人物のシーケンスの埋め込みベクトルを t-SNE を用いて 2 次元空間へと写像した結果を図 7 に示す。図中の個々の数字の番号は人物アイデンティティを示しており、数字の色は人物シーケンスを観測したカメラの種類を表している。異なるカメラで撮影された同一人物が概ね近くの点へと写像されており、異なるアイデン

Fig.7 t-SNE²⁵⁾ visualization of our embeddings on the test split of the iLIDS-VID dataset. Numbers and colors represent the identity of an individual and the camera that captured the individual's sequence, respectively

ティティの人物と距離関係によって区別できることが分かる。それぞれの埋め込みベクトルは人物同士を区別するために必要な高次元な意味情報が含まれていると考えられる。この結果から Entire Triplet Loss による距離学習の有効性が視覚的に確認できる。既存手法同士を比較すると、深層学習を用いた手法が比較的高い精度を達成しており、データドリブンな特徴抽出のパラメータ学習が人物再同定に関しても高い汎化性能に寄与していると考えられる。

4. 結 論

本稿では CNN を用いた人物の距離学習による動画画像人物再同定を行った。動画画像データがもつ複数フレームの画像から得られる人物の様々な見え情報を統合した埋め込みベクトルを用いることにより静止画像の場合と比べ非常に高精度な再同定ができることを示した。また、ニューラルネットワークのパラメータ学習のための誤差関数として Entire Triplet Loss を提案し、人物シーケンスの埋め込みベクトルの学習を行った。この誤差関数はネットワークの一回のパラメータ更新の際にミニバッチ内の全ての組み合わせの Triplet を考慮し勾配の計算を行う。これにより、効率的かつより同一人物、異なる人物を見分けやすい埋め込みベクトルが学習されることを実証した。

改善点として、時系列特徴を抽出可能なモデルの導入が挙げられる。RNN に代表される時系列の依存関係を学習できるモデルを用いることにより、対象人物や背景、遮蔽物の動きを捉え、それらを区別したり、対象人物の歩様特徴を捉えた特徴の抽出が可能になると考えられる。

今後の展望として、人物再同定と人物追跡器を統合した人物追跡システムの研究を考えている。今回の実験設定では対象とする人物を囲む矩形が与えられた前提で再同定を行ったが、実用時には人物矩形を検出器や追跡器などを用いて推定する必要がある。再同定を適用する対象、そして手順が新たな課題点となると考えられる。人物追跡器が追跡に失敗した対象同士を再同定を用いて適切に割り当てることにより、カメラ間の移動や遮蔽に対応した人物追跡を行うことが可能となるはずである。

参 考 文 献

- 1) Liu, Kan, et al. "A spatio-temporal appearance representation for video-based pedestrian re-identification." Proceedings of the IEEE International Conference on Computer Vision. 2015.

- 2) Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.
- 3) Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems*. 2015.
- 4) Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- 5) Oh Song, Hyun, et al. "Deep metric learning via lifted structured feature embedding." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- 6) Wang, Jiang, et al. "Learning fine-grained image similarity with deep ranking." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.
- 7) Li, Wei, et al. "Deepreid: Deep filter pairing neural network for person re-identification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.
- 8) Cheng, De, et al. "Person re-identification by multi-channel parts-based cnn with improved triplet loss function." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- 9) Ahmed, Ejaz, Michael Jones, and Tim K. Marks. "An improved deep learning architecture for person re-identification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- 10) McLaughlin, Niall, Jesus Martinez Del Rincon, and Paul Miller. "Data-augmentation for reducing dataset bias in person re-identification." *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*. IEEE, 2015.
- 11) McLaughlin, Niall, Jesus Martinez del Rincon, and Paul Miller. "Recurrent convolutional network for video-based person re-identification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- 12) Zheng, Liang, et al. "Mars: A video benchmark for large-scale person re-identification." *European Conference on Computer Vision*. Springer International Publishing, 2016.
- 13) Wu, Lin, Chunhua Shen, and Anton van den Hengel. "Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach." *arXiv preprint arXiv:1606.01609* (2016).
- 14) Hadsell, Raia, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping." *Computer vision and pattern recognition, 2006 IEEE computer society conference on*. 2. IEEE, 2006.
- 15) Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- 16) Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).
- 17) He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- 18) Wang, Taiqing, et al. "Person re-identification by video ranking." *European Conference on Computer Vision*. Springer International Publishing, 2014.
- 19) Wang, Taiqing, et al. "Person re-identification by discriminative selection in video ranking." *IEEE transactions on pattern analysis and machine intelligence* 38.12 (2016): 2501-2514.
- 20) Hirzer, Martin, et al. "Person re-identification by descriptive and discriminative classification." *Scandinavian conference on Image analysis*. Springer Berlin Heidelberg, 2011.
- 21) Kingma, Diederik, and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).
- 22) Zhu, Xiaoke, et al. "Video-Based Person Re-Identification by Simultaneously Learning Intra-Video and Inter-Video Distance Metrics." *IJ-CAI*, 2016.
- 23) You, Jinjie, et al. "Top-push video-based person re-identification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- 24) Gao, Changxin, et al. "Temporally aligned pooling representation for video-based person re-identification." *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016.
- 25) Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of Machine Learning Research* 9.Nov (2008): 2579.