博　士　学　位　論　文

真正歩容動画の匿名化と
なりすまし偽歩容動画
に対する防御

2023年1月

大阪大学大学院工学研究科
電気電子情報通信工学専攻

廣瀬　雄基

内容梗概

本論文は，筆者が大阪大学 大学院工学研究科 電気電子情報通信工学専攻 在学中に行った「真正歩容動画の匿名化となりすまし偽歩容動画に対する防御」の成果をまとめたものであり，以下の 6 章より構成される．

第 1 章は序論であり，歩容認証の発達により顕在化した真正歩容動画からのプライバシー情報流出問題，および，マルチメディアデータ生成技術の進展に伴って今後の深刻化が危惧される偽歩容動画を用いたなりすまし攻撃の危険性について述べ，本研究の目的を明らかにする．

第 2 章では，歩容およびその他の生体情報である顔や声などについて，生体認証や匿名化，なりすまし攻撃とその防御法に関する関連研究とそれらの課題を挙げ，本研究の位置づけを明らかにする．

第 3 章では，真正歩容を含む動画が非意図的に Web 上で公開されることによるプライバシー情報流出問題について検討し，それ対する防御手法として，シルエットベース歩容認証手法を想定した歩容動画の匿名化について述べる．歩容により個人を識別する歩容認証に関して，シルエットベースの歩容認証手法が盛んに研究されている．この手法には低解像度で個人を識別できるという特徴があり，これを Web 上の動画に対して適用されれば，動画中の個人が同定され，それに紐づくプライバシー情報が流出する危険性がある．従って，動画中の歩容情報を匿名化することによりこれを保護する必要がある．歩容の匿名化を実現する最も簡便な方法は，動画中の人物に対して黒塗りやモザイク処理を行うことである．しかし，動画としての見た目が不自然になり，視聴されることが前提の Web 動画においては適切ではない．そこで，歩容シルエットの形状の変形と姿勢変化パターンの変更により特徴を変化させ，それに合わせて色を再配置した上で元の動画に埋め戻す，という手法を提案する．提案手法により匿名化した歩容について，その匿名性を歩容認証器の認証精度の観点から，見た目の自然さを動作認識器の認識精度の観点から，それぞれ実験的に評価し，見た目の自然さを保ったまま匿名化が実現可能であることを示す．

第 4 章および第 5 章では，Web 上の動画への偽歩容の埋め込みによる偽情報拡

散問題について検討する．まず，第 4 章では，そのような偽歩容の生成・埋め込みがどの程度現実的なリスクとなり得るのかを議論する．昨今のディープラーニング技術の発達により，生体認証システムの性能は向上したが，一方で生体情報システムに対するなりすまし攻撃のリスクも増大している．特に敵対的生成ネットワークを用いて偽の顔画像や音声データなどを生成することで顔認証や音声認証を突破するといったなりすまし攻撃のリスクについては既存研究で広く議論されている．また，それらの攻撃に対する防御法についても多くの手法が提案されている．しかし，歩容に対するなりすまし攻撃のリスクについて言及した既存研究は少ない．そこで本研究では，ある対象者の一枚の歩容画像からその人物を模倣した偽歩容動画を深層ニューラルネットワークにより生成する手法を考案し，歩容認証におけるなりすまし攻撃の実現可能性について検証する．ここで，このなりすまし攻撃が対象とする歩容認証器では，歩容シルエットを用いて認証を行うことが想定されるので，実際には偽歩容シルエット動画の生成を行うことで上記のなりすまし攻撃の目的は達成される．評価実験において，生成した偽歩容シルエット動画が歩容認証器に高精度で認証されることを確認し，上記のなりすまし攻撃が実際に実現可能であることを示す．

一方，第 5 章では，上述の偽情報拡散に対する防御手法として，真正歩容動画と偽歩容動画の識別手法について述べる．第 4 章にて示すように，偽歩容動画を用いたなりすまし攻撃は実現可能であり，シルエットベースの歩容認証において脅威となり得る．こうしたなりすまし攻撃は，歩容動画の真偽を識別する識別器があれば防御可能であるが，そのような識別器を機械学習により得るためには，真／偽以外の人物，服装，姿勢，視点の 4 つの要素が統制された歩容動画（なりすまし攻撃の場合と同様の理由により，実際には歩容シルエット動画）対からなる学習データセットが必要となる．本研究では，自己符号化器を用いて真正歩容動画から偽歩容動画を作成することにより，そのような学習データセットを用意し，それに基づいて識別器を構築する手法を提案する．複数のデータセットを用いた評価実験を行い，高精度な真偽識別が可能であることを確認することにより，本手法の有効性を示す．

第 6 章は結論であり，本研究により得られた成果を総括するとともに，今後の展望を述べる．

目 次

図 目 次

表 目 次

第1章 序論

SNS 等の普及に伴って，画像・動画・音声・テキストなどのマルチメディアデータが Web 上に大量に蓄積されつつある．それらのマルチメディアデータには個人の顔や声といった生体情報が含まれるものも多い．一方で，深層学習の登場によりマルチメディアデータ生成技術が急速に進歩しており，個人の顔や声を模倣したデータの生成が可能となりつつある [1–5]．極めて写実的な顔画像など，実写・実録のデータと見紛うものも決して珍しくない．以上の状況から，真正な生体情報とそれを模倣した偽の生体情報がともに Web 上で大量に流通する時代が訪れようとしている．

真正な生体情報は，現在でも携帯端末による本人確認などに利用されており，その利活用は日常生活の利便性向上に寄与するものと期待される．しかし，生体情報は典型的な個人情報でもあることから，その流通は，真正か偽かを問わず，大きなリスクを伴う．例えば，Web 上の画像・動画は撮影日時や場所といった情報と紐づいているため，画像・動画中に映っている顔（真正な情報）から個人が同定されれば，その人物がいつどこにいたか，といったプライバシー情報の流出につながる．これは画像・動画中の顔がマルチメディアデータ生成技術により生成された偽情報である場合でも同様である．ただしその場合，実際にはその時間・場所に存在しなかった人物があたかも存在していたかのように偽装されること，すなわち「なりすまし」を意味するため，偽情報の拡散とそれに伴う名誉棄損などが具体的なリスクとなる．上記のリスクに対処するため，マルチメディアデータに含まれる真正な生体情報を保護する手法や，所与の生体情報が真正か偽かを識別することにより偽情報を検出する手法が，主に顔や声を対象に活発に研究されている [6–9]．

生体情報の一種として「歩容」と呼ばれる情報が注目されている．歩容とは「人間の歩行の様子」のことであり，具体的には，歩行者の体型や歩行姿勢，歩行リズムなどを指す．歩容は個人ごとに異なっており，個人を同定し得る情報を含むことが明らかとなってきている．実際，歩容から個人を同定する技術である「歩容認証」が広く研究されており [10–12]，既に犯罪捜査などに利用されている [13]．歩容は，改正個人情報保護法にも保護すべき対象として記載されており，顔や声

と並んで重要な生体情報であると言える．従って，歩容においても，上記の真正／偽情報に纏わるリスクが十分に想定される．しかし，歩容は，顔や声と比較して新しい，近年になって注目され始めた生体情報であることから，現状では，リスク対策手法の検討が進んでいない．

以上の背景を踏まえ，本論文では，真正な歩容情報の流通に伴うプライバシー流出のリスク，ならびに偽歩容情報の流通がもたらす「なりすまし」のリスクに対処するための手法を提案する．そのようなリスク対策手法が実現されれば，歩容を含む生体情報をより安全な形で利活用できるようになり，今後の社会の発展に大きく寄与することが期待される．上述のリスクが現在の社会においてどのように存在するかを以下に述べる．

今日では，Web の発達やスマートフォンなどのカメラ付き携帯端末の普及により，多くの人々が YouTube 等の動画共有サービスを日常的に使用している．動画共有サービスのユーザは，時に投稿者として日々多くの動画コンテンツをアップロードするとともに，時に視聴者としてそうした動画コンテンツにアクセスし視聴する．その結果，動画共有サービス上では膨大な量の動画（Web 動画）が不特定多数の人物に共有される状況となっているが，それらの Web 動画には，動画中に意図せずに映りこんだ人物の真正歩容情報が含まれる．一方で，Web 動画から人物領域を検出することは画像処理技術の発展により容易化しつつあり，切り出した人物領域に対し歩容認証処理を適用すれば，その個人を同定することができる．これは，当該個人の真正歩容情報からそれに紐づくプライバシー情報（当該個人がいつどこにいたか，など）が詐取されることを意味している．

また，歩容認証による個人同定は，プライバシー情報の詐取だけでなく，特定個人の真正歩容情報自体の収集も可能にする．この真正歩容情報から当該個人の偽歩容動画を作成される可能性が考えられる．これは，深層学習を用いたマルチメディアデータ生成技術が急速に発達している昨今では十分に想定されるリスクである．このような偽歩容情報が，社会通念上不適切とされる場所の背景動画と合成され，動画共有サービス上にアップロードされれば，偽情報の拡散につながる恐れがある．例えば，歩容認証器を所有する第三者がこうした動画を悪意なく解析し，結果，動画に映る当該個人が社会通念上不適切とされる場所に存在すると誤判断した場合，このような偽の情報が拡散され，当該個人の名誉が棄損される．

これらを踏まえて，本論文における考察の対象は，具体的には，以下の 2 種類の攻撃を受けるリスクとなる．

(A) 真正歩容動画に対するプライバシー情報詐取攻撃: Web 上の動画に含まれる真正歩容情報から歩容認証技術により個人を同定し，その人物に紐づくプライバシー情報を取得する攻撃

(B) 偽歩容動画を用いたなりすまし攻撃: マルチメディアデータ生成技術により特定個人の偽歩容情報を作成して当該個人になりすまし，それにより偽情報を拡散する攻撃

以上に述べた 2 種類の攻撃に関して，本研究では，主として以下の三つの研究課題に取り組む.

(1) 攻撃 (A) に対する防御手法として，真正歩容動画を匿名化することにより歩容認証技術による個人同定を困難化し，プライバシー情報の流出を防ぐ手法を提案する.

(2) 攻撃 (B) に関して，現実的な攻撃条件の下で偽歩容動画を生成する手法，すなわち攻撃法を検討する．本検討を通じて，偽歩容動画が実際に本人であると誤認識される割合を調査し，本攻撃の実現可能性を検証する.

(3) 攻撃 (B) に対する防御手法として，真正歩容動画と偽歩容動画を識別するための識別器の具体的な設計手法を提案する.

以上の研究課題 (1)〜(3) について，その関連研究と提案手法の具体的な内容や性能評価実験の結果などを，第 2 章以降で詳しく述べる．また，それらの関係性を図 1.1 に示す.

まず，第 2 章では，Web 動画中の歩容に纏わる問題とその対処法について，それらに関する既存研究を歩容以外の生体情報とも比較しながら俯瞰し，本研究の位置づけを明らかにする.

第 3 章では，歩容認証手法として主流であるシルエットベース認証手法を対象として，そのような手法による個人同定を困難化するための歩容匿名化手法について述べる (上記の研究課題 (1)). 歩容の匿名化を実現する最も簡便な方法は，動画中の人物に対して黒塗りやモザイク処理を行うことである．しかし，動画としての見た目が不自然になり，視聴されることが前提の Web 動画においては適切ではない．そこで，歩容シルエットの形状の変形と姿勢変化パターンの変更により歩容に関する特徴を変化させ，それに合わせて色を再配置した上で元の動画に埋め戻すことにより，見た目を保った状態で匿名化を実現する手法を提案する.

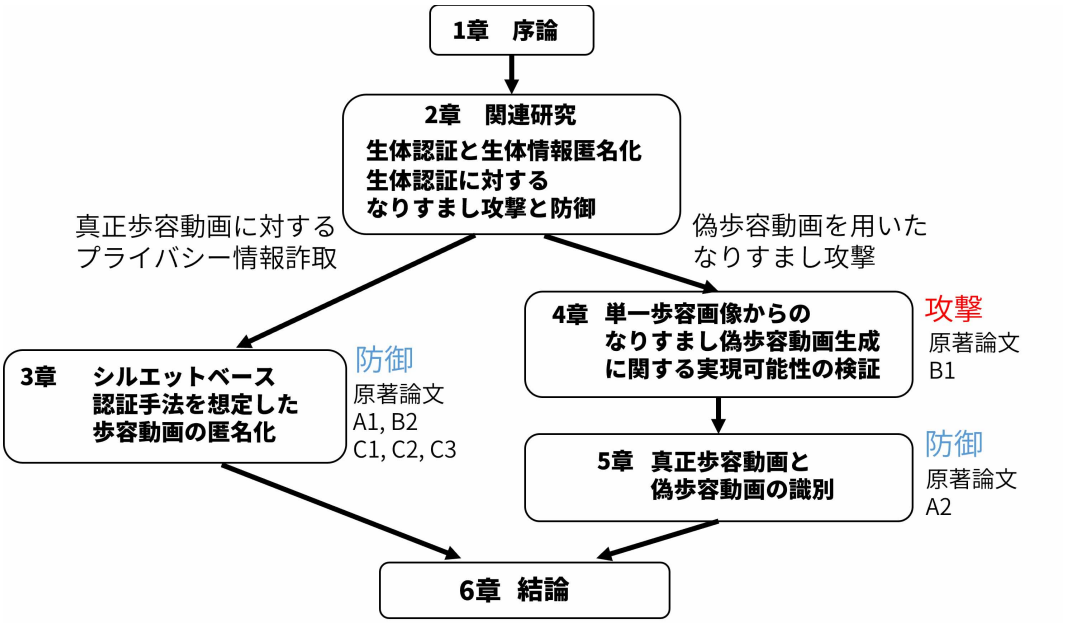


図 1.1: 各章の関連図

第 4 章では，特定個人の一枚の歩容画像からその人物のなりすまし偽歩容動画を生成する手法を検討し，それによるなりすまし攻撃がどの程度現実的であるかを検証する (上記の研究課題 (2))．このなりすまし攻撃が対象とする歩容認証器では，歩容シルエット情報のみを用いて認証を行うと想定される．そこで，歩容シルエット動画の段階でなりすまし対象の個人の特徴を保持していれば，その歩容に付与されている色情報に関わらず，なりすました該当個人の歩容であると歩容認証器に誤認識させることができる．この考えに基づき，第 4 章では偽歩容シルエット動画の生成手法について議論する．この偽歩容シルエット動画の生成は，歩容画像から姿勢に非依存の特徴を抽出するエンコーダと，その特徴量から歩容シルエット動画を生成するデコーダを構築することにより可能となる．しかし，一枚の歩容画像のみから歩行リズムなどの動的な情報を完全に抽出することは容易ではないため，デコーダにより生成される歩容動画は個人性の特徴が一部欠落したものとなる．そこで，個人性特徴を強調することにより欠落した情報を補完し，なりすましの精度を向上させる手法を提案することで，本攻撃の実現可能性を検証する．

第 5 章では，偽歩容動画の問題に対する防御手法として，真正歩容動画と偽歩容動画を識別する手法を述べる (上記の研究課題 (3))．第 4 章で述べたような手法

により偽動画（なりすまし偽歩容動画）が作成され，それを用いたなりすまし攻撃が実現可能であった場合，それを防御する仕組みが必要となる．ここで，偽歩容動画がなりすまし攻撃に用いられる際には，一度シルエット化されたのちに歩容認証器に入力されることが想定される．従って，シルエット化された後の「歩容シルエット動画」の段階で，それが真正であるか偽であるかを識別することができれば，偽歩容動画によるなりすましを防ぐことが可能となる．よって，上記の防御の際には，所与の歩容シルエット動画が真正であるか偽であるかを識別する識別器があれば良いが，そのような識別器を機械学習により得るためには，真／偽以外の人物，服装，姿勢，視点の 4 つの要素が統制された歩容シルエット動画対からなる学習データセットが必要となる．そのようなデータセットを自己符号化器 (Auto Encoder; AE) を用いて作成し，これにより上述の識別器を学習する手法を提案する．

最後に，第 6 章で，Web 動画中に存在する歩容情報に纏わる問題とその対処法に関する検証結果をまとめ，今後の展望を述べる．

第2章 関連研究

2.1 緒言

本章では，歩容による個人認証や歩容の匿名化，歩容に関するなりすまし攻撃とその防御など，本研究に関連する既存研究を概観し，本研究の位置づけを明らかにする．ただし，第 1 章でも述べたように，歩容は比較的新しい生体情報であるため，顔や声と比較して，その匿名化手法やなりすまし防止手法に関する従来研究は多くない．このため，歩容以外の生体情報を対象とした従来研究についても取り上げ，それと比較する形で歩容に関する諸研究の現状を整理する．

まず，2.2 節では，歩容認証およびその他の生体認証に関する既存研究についてまとめ，歩容が重要な生体情報の一つであることを確かめる．次に，2.3 節では，生体情報の匿名化について主に顔と歩容を比較するかたちで論じ，歩容匿名化研究の現状とその課題を明らかにする．その後，2.4 節および 2.5 節で，生体情報の偽造によるなりすまし攻撃とその防御について，やはり顔と歩容を中心に論じ，歩容に対するなりすましの危険性が未だ十分に検討されていない現状を詳らかにする．

2.2 歩容認証およびその他の生体認証

生体認証とは，個人の生理的・行動的な特徴に基づいて個人を自動的に認証する技術のことである．生体認証に用いられる主な生体情報の例を表 2.1 に示す．生体情報のうち最も一般的なのは顔と指紋であり，それゆえに最も古くから研究が進められてきた．顔は，個人を識別するための特徴を多量に含んでおり，認識や認証に用いることができるが，表情や加齢などの別要因に左右されやすいという課題もある．一方で，指紋は指先の表面の凹凸の模様であり，個人ごとに独特のパターンを持つため個人認証に利用される．顔や指紋を用いた認証では，ユーザはカメラやセンサに近づく必要があり，ユーザの協力が必要となる．こうした生体情報のうち Web 動画中に存在する生体情報を以後では考察の対象とする．

生体情報のうち，比較的最近になって発展してきたものの一つが歩容である．歩

Web 動画上	顔，音声，歩容など
それ以外	指紋，掌紋，光彩など

表 2.1: 生体認証に用いられる主な生体情報

容認証は，顔認証や指紋認証と異なり，比較的遠距離から観測された情報に基づいて個人を同定し得るため，距離依存性が小さい点やユーザの協力が必要ない点で優れており，主に犯罪捜査などへの応用 [13] を目的として発展してきた．その手法はモデルに基づく方法とモデルフリーな方法に大別される．

モデルに基づく方法とは，人体を表す 3 次元構造モデル（関節点座標とその接続関係を表現した骨格モデルなど）に基づいて歩行中の人物の姿勢を陽に推定し，その時系列に基づいて個人を認証する手法である [14]．この手法には，画像中に含まれるセンサノイズや背景の多様性に頑健であるという利点がある一方で，計算コストが大きいという欠点がある．また，姿勢推定を正しく行うためには動画中の人物領域に一定程度以上の解像度が求められる．一方，モデルフリーな方法とは，上述のような 3 次元構造モデルを必要としない手法である．中でも，歩行中の人物のシルエット（歩容シルエット）から直接個人を認証する手法が最も一般的であり，近年盛んに研究されている [11,19,20,35] ．歩容シルエットは，解像度の低い動画からでも比較的頑健に抽出可能であり，二値画像であるため計算コストが少ないという特長がある．これらの特長は多量の動画を対象として処理を行う場合に大きな利点となるため，第 1 章で述べたような，歩容認証技術を悪用したプライバシー情報詐取攻撃に用いられる恐れも高いと言える．

歩容シルエットに基づく認証手法は，シルエットの系列を直接扱うものと，シルエット系列を一枚また少数枚の画像に集約した上で認証を試みるものの二種類に大別される．前者の例としては，個人ごとの歩容シルエットの時系列パターンを隠れマルコフモデルによりモデル化する手法が挙げられる [15]．この方法は，比較的扱うシルエット画像の枚数が多いため，モデルに基づく手法にかかる程度のものではないが，計算コストが増加するという問題がある．このため，近年では研究例は多くない．一方，後者は近年非常に活発に研究されている．歩容シルエット系列を画像に集約する手法をまとめたものを図 2.1 に示す．この手法のうち最も基本的なものは Gait Energy Image (GEI) [16] である．GEI は歩容シルエット系列をその周期で平均化した画像であり，初期の研究において高い認証精度が報告されている．

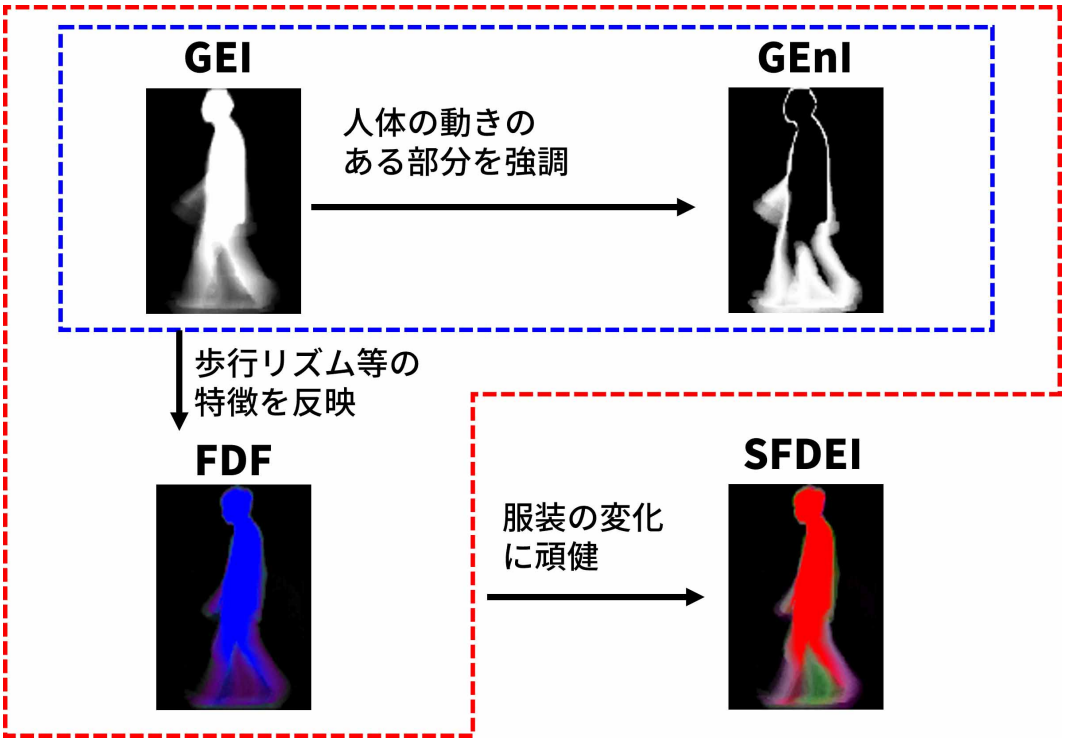


図 2.1: シルエット系列を画像に集約し認証する手法

GEI の問題として，平均化処理に起因して頭部や胴体など人体中の動きのない部分が強調される傾向がある，という点が挙げられる．そこで Bashir らは，手や足などの動きのある部位を強調するため，画素ごとに輝度強度のシャノンエントロピーを計算する Gait Entropy Image (GEnI) [18] へと GEI を拡張した．ただし，GEI も GEnI も，その特性上，歩行リズムなどの動的な特徴が失われるという問題もある．これを解決するものとして，画素ごとに画素値の時系列をフーリエ変換し，それにより得られる低周波成分の強度を画像化した Frequency Domain Feature (FDF) が提案されている [17]．FDF の直流成分は GEI に一致するため，FDF は GEI を周波数方向に拡張した特徴量と位置づけられる．GEI，GEnI および FDF は，歩容シルエットの形状，すなわち，体形や服装に大きく依存する特徴量であるため，服装の変化に脆弱であるという問題がある．これを解決するものとして，Signed Frame Difference Energy Image (SFDEI) [19] が提案されている．SFDEI は歩容シルエット系列中のフレーム間で差分画像を求めることから，動きのある個所に特化した特徴量と言え，服装などの静的な特徴の変化に対する頑健性が期待できる．

シルエットベース歩容認証の研究においては，当初，集約画像そのものの提案に主眼が置かれ，認識アルゴリズム自体に注目した研究は少なかった．しかし，近年では，認識アルゴリズムを工夫することにより認証精度の向上を目指した研究も多い．最も端的な例として，GEI に対し畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) を適用した手法が挙げられる [11]．また，より現実的な環境に対応できるよう認証手法を拡張する研究も試みられつつある．具体的な例として，鈴木らは，荷物を所持している人物の歩容シルエットに対しても頑健に機能する認証手法を提案している [20]．ほかにも，Muramatsu らは，何らかの原因で歩容シルエットが完全に抽出できず，不完全なシルエットしか得られていないような状況下においても，部分空間法に基づいて完全なシルエットに対応する特徴量を復元する手法を提案している [35]．

以上のように，歩容認証手法，とりわけ，歩容シルエットに基づく手法は近年活発に研究され，日々高度化していることから，歩容は顔や指紋と同等の生体情報であることが明らかとなりつつある．従って，その匿名化技術やなりすまし防止技術の確立は重要かつ喫緊の課題と言える．

2.3 生体情報の匿名化

画像・動画中の人物の外見には個人を同定し得る生体情報が多々含まれており，顔や歩容はその最たる例である．それらの生体情報は，画像･動画の撮影場所や時間，あるいは，当該人物の行動など，プライバシー情報と紐づいている．これらのことから，生体認証によって個人が同定され，それに紐づくプライバシー情報を取得される危険性が画像や動画には常に存在していると言える．このため，防犯カメラの動画やWeb動画等に意図せず写りこんだ人物の外見を匿名化することにより個人の同定を防ぎ，それに紐づくプライバシー情報を保護する研究が従来より活発に行われてきた．その最も端的な例として顔情報の匿名化が挙げられる．

画像・動画中の顔領域から個人を同定し得る外見特徴を除去する処理は一般に Face De-identification と呼ばれる．Face De-identification の目的は，主に次の 2 点を満たすように顔領域を匿名化することである．

(a) 人間の観察により個人が同定されないこと

(b) 顔認識器などの自動認識システムにより個人が同定されないこと

初期の研究では，上記のうち主に (a) の要件を満たすような研究が行われ，対象の顔領域にぼかし処理やモザイク処理等の視覚的抽象化 [21, 22] を施す手法が提案されてきた．例えば Boyle らは，ぼかし処理に用いるフィルタのサイズに応じて匿名化効果がどのように変化するかを詳細に評価している [23, 24]．しかし，これらの手法は必ずしも (b) の要件を満たすとは限らないことが後年の研究により明らかとなってきた．顔認識は一般に認識対象の顔画像をデータベース中の顔画像と照合することにより行われるが，Newton らは，認識対象の顔画像に視覚的抽象化が施されている場合でも，同様の視覚的抽象化をデータベース中の顔画像にも施した上で両者を照合すれば，ある程度の精度で顔認識が可能となることを示している [25]．また，視覚的抽象化は見た目の自然さを損なうため，後に視聴される可能性のある画像・動画に対する処理としてはあまり望ましくないという問題も存在する．

上記の理由から，近年の研究では，対象の顔領域を別の顔領域と置換することにより顔情報を変化させ匿名化する，という処理が広く研究されている．例えば Bitouk らは，様々な年齢・性別・向きの顔画像を含むデータセットを予め構築したのち，その中から匿名化対象の顔領域に類似した顔画像を探索し，その画像を元の顔領域と置き換える手法を提案している [6]．この際，置換前の画像と置換後の画像における照明環境等の違いを考慮し，色が滑らかに変化するように両者を合成することにより見た目の不自然さを軽減している．Gross らはこれを発展させた手法として k-Same と呼ばれる手法を提案している [26, 27]．この手法では，匿名化対象の顔領域に類似した顔画像をデータベース中から k 枚探索し，その平均画像で元の顔領域を置き換える．より近年の研究として，Nakashima らは，匿名化対象の顔画像を別の顔画像に置き換える際，パッチ単位で置換処理を行うことにより，元の顔画像における表情は保存したまま個人性のみを除去する手法を提案している [28]．場の状況にそぐわない表情は見る者に不自然さを感じさせるため，表情の保持は見た目の自然さを保つ上で重要な特性と言える．

歩容も顔と同様，個人の同定につながる情報であるが，古くから研究されてきた顔認証技術と異なり，歩容認証技術は近年になって発展した技術であるため，歩容情報の匿名化に取り組んだ従来研究は歩容認証技術と異なり，現状では数少ない．例の一つとして，Agrawal らは動画中の人物領域にぼかし処理を施すことにより歩容情報を匿名化する手法を提案している [29]．また，Mitsugami らは，監視カメラ動画中の各人物の領域を棒状のシンボルと置き換えることにより，人物

	顔	歩容	
視覚的抽象化	Boyle et al. [23] Neustaedter et al. [24]	Agrawal et al. [29] Mitsugami et al. [30]	
置換・変形	Bitouk et al. [6] Gross et al. [26, 27] Nakashima et al. [28]	Tieu et al. [31, 33, 34]	本研究
歩行リズム変化			

表 2.2: 生体情報の匿名化における本研究の位置づけ

のプライバシーを保護することを提案している [30]．しかし，前述の通り，このような処理は見た目の自然さを損なうため，Web 動画のような，視聴されることが前提の動画には適さない手法であると言える．

上記の他に，歩容認証の主流がシルエットベース手法であることを踏まえ，歩容シルエットの微小変形により歩容情報の匿名化を試みた手法も少数ながら提案されている．具体的な例としては Tieu らの手法が挙げられる [31]．この手法は，匿名化対象の歩容シルエットとともに何らかの別シルエット（ノイズシルエット）を入力とする手法であり，基本的には元の歩容シルエットに近いものの，ノイズシルエットにも多少類似するような歩容シルエットを生成・出力することにより，元の歩容シルエットの匿名化を図るものである．しかし，元の歩容シルエットとの差異がある程度大きいノイズシルエットを適切に選択できなければ匿名化の性能が下がるなどの問題もある．これを踏まえて，Tieu らはさらに敵対的生成ネットワーク（Generative Adversarial Networks; GAN）[32] を用いてそのようなノイズを生成，付与することによって匿名化シルエットを生成する手法 [33] を提案している．このようなシルエット変形に基づく匿名化手法では，クロマキー処理など，背景部分から人物領域を切り出す処理が事前に行われることを前提としているが，その過程で歩容シルエットが正しく抽出できない場合もあり得る．そのような場合でもシルエットの匿名化を可能とするため，Tieu らは身体の一部分が欠けた低品質な歩容シルエットを対象に匿名化を施す手法 [34] も提案している．

以上に述べた Tieu らの手法はいずれも，歩容が持つ体形や服装などの静的な特徴の匿名化には非常に有効であるが，同じく歩容が持つ歩行リズム等の動的な特徴については匿名化できない．しかし，より精度の高い匿名化を実現するためには，静的な特徴と動的な特徴をともに匿名化する必要がある．本研究ではこれら

二つの特徴を両方とも匿名化することを目指す．これは第 1 章で述べた研究課題 (1) に相当し，第 3 章にてその詳細を述べる．これらを踏まえて，生体情報の匿名化における本研究の位置づけを表 2.2 に示す．

2.4 生体認証に対するなりすまし攻撃

生体認証を不正に突破することを目的としたなりすまし攻撃の手法は，主に顔認証や音声認証を対象として，古くから研究されてきた．こうした攻撃の中で最も単純な方法は「プレゼンテーション攻撃」と呼ばれるものである．これは，生体認証システムに対し人工物等が提示された場合に誤ってそれが生体情報として認識される，といった脆弱性をついた攻撃である．例えば，Patel らはカメラを搭載した顔認証システムに対して，登録されている個人の写真や動画を提示することで顔認証システムを突破する手法を提案している [36,37]．また，顔の場合と同様，マイクを搭載した音声認証システムに対するプレゼンテーション攻撃として，Cheng は登録されている個人の録音データを再生することで音声認証システムを突破する手法を提案している [38]．

以上のように，プレゼンテーション攻撃を行うためには個人の写真や録音音声などの真正な情報が必要である．これらのうち顔情報については，SNS 上から真正顔画像を取得できる場合がある [39]．一方で，音声情報の場合は，プレゼンテーション攻撃に必要な真正音声の取得は必ずしも容易ではない．これは，真正音声自体は取得できたとしてもそれが認証に必要な音声の言語や発音であるとは限らないからである．よって，音声合成技術を利用して偽のなりすましデータを作成する場合が多い．このような合成データを用いた音声なりすまし攻撃は一般に音声合成攻撃と呼ばれている．このタイプの攻撃に用いられる音声合成技術は音声変換 (Voice Conversion; VC) とテキスト音声合成 (Text-To-Speech; TTS) の二つに大別される．このうち VC は，ソースとなる話者の音声をその言語情報は変えずにターゲットの話者の音声へと変換する技術である．もう一方の TTS は，任意のプレーンテキストを特定のターゲット話者の声質で発声されたかのような音声へと変換するものである．これらの技術を攻撃者自身の音声やテキストに対して適用し，特定個人の音声に変換することにより，なりすましデータを取得することが可能となる [4,5]．なお，このようなマルチメディアデータ生成技術は顔情報を偽装する場合にも悪用され得る．現在では，2 次元の画像・動画だけでなく，顔

の 3 次元ボリュームデータも GAN により生成できることが示されつつある [40]. そのようなボリュームデータから作成された三次元フェイスマスクはなりすまし攻撃に悪用されるリスクがある [41].

上述のように，顔や声などの従来から存在する生体情報に関しては，マルチメディアデータ生成技術を悪用したなりすまし攻撃の可能性が広く研究されている．これに対し，同様に生体情報である歩容に関しては，合成データを用いて歩容認証器の突破を試みるようななりすまし攻撃の可能性は未だ十分に検討されていない．そこで，本研究では，深層ニューラルネットワーク（Deep Neural Networks; DNN）に基づくマルチメディアデータ生成技術を悪用して歩容認証器に対するなりすましを試みる攻撃に着目し，その実現可能性を実験的に検証する．これは第 1 章で述べた研究課題 (2) に相当し，第 4 章にてその詳細を述べる．

2.5 なりすまし攻撃に対する防御

なりすまし攻撃の可能性が検討されている一方で，それを防ぐためのなりすまし防止手法も盛んに研究されている．上述したように，なりすまし攻撃にはプレゼンテーション攻撃とマルチメディアデータ生成処理を悪用した攻撃の二種類が存在するが，なりすまし防止手法としても，それぞれの攻撃に対応したものが存在する．

顔認識システムへのプレゼンテーション攻撃に対して，Usman は入力された顔動画からフレーム間差分を取り出して一つの画像へと集約し，それを特徴量として入力動画がなりすまし顔であるか否かを判別する手法を提案している [42]．紙に印刷された顔やディスプレイに表示された顔を撮影した動画は，実際の顔を映した動画とは異なるフレーム間差分特徴を有するため，それを CNN により解析することはなりすまし防御法として有効である．

一方，マルチメディアデータ生成処理を悪用した攻撃に対する防御法としては，例えば，コンピュータグラフィックスにより生成した顔画像や GAN で生成した顔画像と実際の顔画像との識別を試みた先行研究がある [36,37]．また，近年では CNN を用いたなりすまし対策も提案されている．Chen らは，顔画像の輝度成分が GAN により生成された顔画像の検出に有用であることを見出し，入力顔画像の色情報を YCbCr 色空間へと変換した上で CNN に入力し，それが GAN による生成物か否かを判別する手法を提案している [8].

音声に関しては，ポップノイズに基づくなりすまし防止法が研究されている [9, 43]．人間がマイクに向かって話す際には，話し声とは別に息がマイクに届くことがあり，それが原因でポップノイズが発生する．こうしたノイズはGANのような最新技術を用いたとしても正確に再現することは難しい．そのため，ポップノイズはなりすまし音声の対策に役立つ重要な手がかりとなっている．

以上のように，顔認証や音声認証を対象としたなりすまし防止手法が数多く研究されている．これに対し，歩容認証に対しては，2.4節で述べたように合成データを用いたなりすまし攻撃法自体がまだ十分に検討されていないため，その防止手法に言及した研究はまだ見られない．しかし，DNNを用いたマルチメディアデータ生成技術の発達は目覚ましく，特定個人の歩容特徴を模倣した歩行動作動画の合成も可能となりつつあることから，そのような合成データを悪用したなりすまし攻撃が近い将来実現する可能性は高い．そこで本研究では，所与の歩容動画がマルチメディアデータ生成技術により合成された偽歩容動画であるか実際に撮影された真正歩容動画であるかを識別する手法を考案する．これは第1章で述べた研究課題 (3) に相当し，第5章にてその詳細を述べる．

2.6 結言

本章では，本研究に関連する研究分野の現状について，歩容とその他の生体情報を比較しながら論じた．

研究課題 (1) の歩容匿名化について，歩容のもつ静的な特徴の匿名化を試みた研究は数少ないながらも存在する一方で，動的な特徴の匿名化は試みられていない．より安全性の高い匿名化を実現するため，本研究では静的な特徴と動的な特徴を共に匿名化可能な手法の実現を目指す．

研究課題 (2), (3) のなりすまし攻撃とその防御については，マルチメディアデータ生成技術を歩容に適用した攻撃例は未だ検討されていないものの，適用自体は可能と考えられる．顔や声を対象にマルチメディアデータ生成技術を悪用した攻撃が実行される可能性は広く指摘されていることからも，偽歩容動画によるなりすまし攻撃の実現可能性を探ることは喫緊の課題であると考えられる．同時に，その防御手法の確立も極めて重要な課題である．本研究では，攻撃手法・防御手法の双方をマルチメディアデータ処理の観点から論じ，具体的な手法を提案する．

第3章 シルエットベース認証手法を想定した歩容動画の匿名化

3.1 緒言

本章では，Web 上に存在する真正歩容動画に映る人物がシルエットベース歩容認証によって同定され，それに紐づくプライバシー情報が流出する問題に着目し，その解決に努めるため，シルエットベース歩容認証を想定した歩容動画の匿名化手法について議論する．Web 動画は視聴されることが前提なので，匿名化の際はユーザエクスペリエンスを下げないような方法でなければならない．この考えをもとに，個人の同定は行えない程度に特徴を変更し，一方で動画としての自然な見た目を保持することを目指す．まず，3.2 節で歩容動画の匿名化手法の概要について論じる．次に，3.3 節で歩容シルエット動画の匿名化の具体的な生成手法について述べ，続いて，3.4 節で生成した匿名歩容シルエットに対してテクスチャを転写する方法について詳述する．その後，本手法の有効性を 3.5 節で実験的に評価し，最後に 3.6 節で本章をまとめる．

3.2 歩容動画の匿名化手法の概要

歩行中の人物のシルエットは解像度が低くノイズの多い低品質な動画からでも比較的容易に抽出することが可能である．よって，シルエットベースの歩容認証手法は動画共有サービス上の動画に適している．このため，歩容認証によって真正歩容動画から個人を同定し，その人物に紐づくプライバシー情報を取得するプライバシー情報詐取攻撃に悪用される危険性も高い手法であると考えられる．このことから，シルエットベース認証手法の歩容認証により個人が同定されることを想定し，人物領域の歩容の匿名化を行うことを考える．シルエットベース認証手法では，動画から人物領域を切り出し，それをシルエット化したのち，その歩容シルエットに対し歩容認証を行うことが想定される．よって，動画中の歩容シルエット情報が匿名化されていた場合，個人の同定を行うことが不可能となる．こ

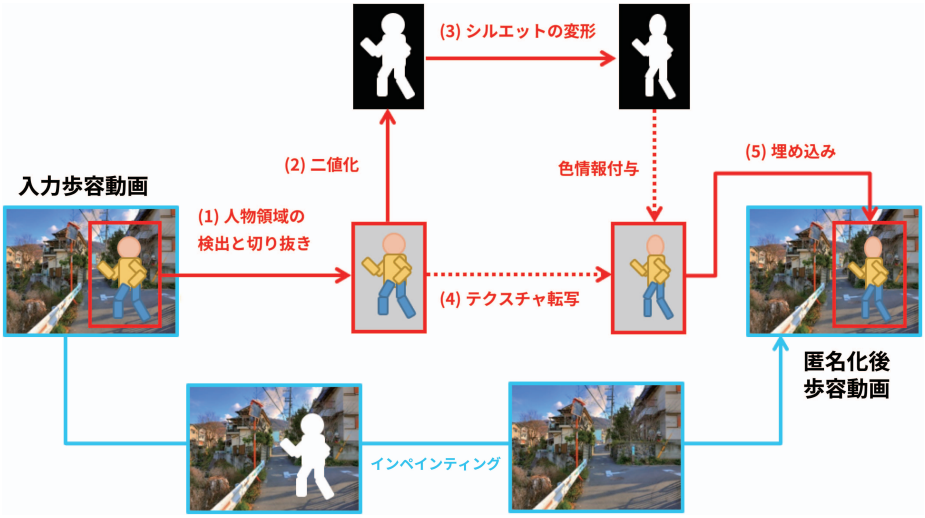


図 3.1: 歩容動画の匿名化手順

れらを踏まえると，本研究の歩容動画匿名化における匿名性はそのシルエットに対する匿名化によって実現されると言える．

第 2 章で述べたように，歩容を匿名化するための簡単な方法としては，塗りつぶしやモザイク処理などの視覚的抽象化が挙げられる．しかし，本研究では，匿名化すべき歩容情報が含まれる動画として YouTube 等の動画共有サービスの Web 動画を想定しているため，人物領域が視覚的に不自然になり動画の品質が劣化する上述のような方法は望ましくない．そこで，顔の匿名化で実施されている手法である領域置換型の手法を採用する．これらを踏まえて，匿名化の具体的な手順は以下の通りとなる．

(1) 入力された歩容動画の各フレームから人物領域を検出し，切り取る．

(2) 人物領域を二値化し，シルエットを取得する．

(3) わずかにシルエットを変形し，歩容認証器が個人を識別できないようにする．

(4) 変形後のシルエットに対し元の人物領域のテクスチャを転写し，匿名化済み人物領域を得る．

(5) 入力動画に対し匿名化された人物領域を埋め戻す．

これらをまとめた匿名化の手順を図 3.1 に示す．

(1) と (2) は，既存の人体検出／セグメンテーション手法を用いることで，容易に実現することができる．また，(5) については，[44] などの画像インペインティングの手法により容易に実現することができる．そこで，歩容動画の匿名化の過程のうち，特に (3) と (4) のステップに着目し，これらを実現するための手法を提案する．以下，この 2 つのステップをそれぞれ「歩容シルエット変形」,「人物領域テクスチャ転写」と呼ぶ．

本研究の目的は，歩容動画を個人の同定が行えないように匿名化し，かつ動画としての自然な見た目を保持することである．このうち匿名性については歩容シルエット変形により実現し，一方で見た目の自然さについては人物領域テクスチャ転写により実現する．しかし，人物領域テクスチャ転写の際に見た目の自然さを実現するためには，歩容シルエット変形で大きくシルエットを変形させないということも重要となる．そこで，次節で述べる歩容シルエット動画の匿名化の際は歩容シルエットを個人の同定は行えないが大きくは変えない程度に変形させることを目標とする．

3.3 歩容シルエット動画の匿名化

3.3.1 歩容シルエット動画の匿名化手法の概要

$\mathcal{S} = (S_1, \cdots, S_N)$ を匿名化すべき歩容シルエット系列とする．ここで，S_i は入力歩容動画の i 番目のフレームを二値化したシルエット画像であり，N はフレームの数である．歩容シルエット動画匿名化の目標は $\mathcal{S}$ を変形し，新たな歩容シルエット系列 $\mathcal{T} = (T_1, \cdots, T_N)$ を取得したのち，$\mathcal{T}$ に含まれる人物を歩容認証器によって正しく識別できないようにすることである．前節で述べたように，本手法ではフレーム単位での変形を実現する．すなわち，S_i から T_i へと各フレームごとに個別に変形し，それらを最終的に連結し一つの動画にする．そこで，本節ではフレーム単位での変形処理に着目する．

歩容シルエットは体型（衣服の形状を含む）と姿勢の 2 つの要素から決まる．前者は 1 つの動画の中で変化しない静的な特徴であり，後者は各人の歩行リズムに応じて周期的に変化する動的な特徴である．これら 2 つの要素は，いずれも歩容認証の重要な手がかりとなる．よって，提案手法ではこの両方を匿名化可能な手法を提案する．

人間の歩行は周期的な動作であるため，歩行中の任意の姿勢に対し 0 から 2π ま

での位相を定義することができる．すなわち，位相 θ は $\theta \in [0, 2\pi)$ と表現できる．これを踏まえて，提案手法では，歩容シルエットを位相成分と形状成分の 2 つに分けて考える．形状成分は人の体型や服装などの静的な特徴に相当し，一つの動画内で常に同じ特徴を持つ．一方，位相成分は，その変化パターンが歩容シルエットの動的な特徴に相当する．提案手法では，入力の歩容シルエット動画をこれらの 2 成分に分解し，各々に摂動を加えたのち，それらの成分から歩容シルエットを再生成する．これにより，静的な特徴と動的な特徴の双方を匿名化することが可能となる．ここで，静的な特徴の匿名化は 2.3 節で述べた Tieu らの手法 [31,33,34] と同様に，シルエットの形状情報を変化させるという考えの元，匿名化を実現する．一方で，それに限らず動的な特徴の匿名化も同時に実施することで，歩容認証の手がかりとなる要素をより多く匿名化させることができるため，提案手法の有効性が向上する．

上記の手法を実現するためには，位相成分および形状成分を表す特徴量を定義した上で，歩容シルエットを特徴量へと変換する写像と，特徴量を歩容シルエットへと変換する写像が必要となる．このような写像を学習する手段として，近年発展が著しい DNN を本研究では用い，上記の手法を実現する．上記の匿名化手法の概要を以下に記す．

人物 a の i フレーム目の位相を θ_i として，このときの歩容シルエット S_i を $S_i = \mathrm{Sil}_a(\theta_i)$ と定義しなおす．提案手法では，各フレーム $\mathrm{Sil}_a(\theta_i)$ にわずかな摂動を加え $\mathrm{Sil}_{a'}(\theta_i')$ へと変形する，という処理を全フレームに対し適用することにより歩容シルエット動画を匿名化する．ここで，$a'(\neq a)$ は仮想的な別人物に相当し，全フレームにおいて a' の形状に関する特徴量が同一となるようにすることにより，匿名化後の動画において人物の服装や体形が不自然に変化することを防ぐ．また，任意の i について $\theta_i' \neq \theta_i$ となるようにする．一枚の歩容シルエット画像を変形するための具体的な手順は次の通りである．但し，$\boldsymbol{p}_\theta$ は位相に関する特徴量（以降では位相コードと呼ぶ），$\boldsymbol{z}_a$ は形状に関する特徴量（以降では形状コードと呼ぶ）である．

(1) 入力画像 $\mathrm{Sil}_a(\theta_i)$ からその位相 θ_i を推定する．

(2) $\mathrm{Sil}_a(\theta_i)$ から形状コード $\boldsymbol{z}_a$ を抽出する．理想的には任意の i について同様の形状コードを得る．

(3) θ_i と $\boldsymbol{z}_a$ に摂動を加える．$\Delta\theta_i$，$\Delta\boldsymbol{z}$ をそれぞれ位相の摂動，形状コードの摂

動とし，これらの摂動を用いて θ_i を $\theta_i' = \theta_i + \Delta\theta_i$ へと変換する．同様に $\boldsymbol{z}_a$ を $\boldsymbol{z}_{a'} = \boldsymbol{z}_a + \Delta\boldsymbol{z}$ へと変換する．但し，$\boldsymbol{z}_{a'}$ は単一の歩容シルエット動画中では常に同一となるようにする．

(4) θ_i' から $\boldsymbol{p}_{\theta_i'}$ を計算し，$\boldsymbol{p}_{\theta_i'}$ と $\boldsymbol{z}_{a'}$ を用いて，入力画像とはわずかに異なる歩容シルエット画像 $\mathrm{Sil}_{a'}(\theta_i')$ を生成する．

以上を概要図にまとめたものを図 3.2 に示す．上記の各ステップのうち，まずステップ (1) については 3.3.2 節で述べる．次に，ステップ (2) については 3.3.3 節で説明する．さらに，ステップ (2)，(4) で使用するエンコーダ・デコーダ等の DNN の具体的なネットワークの設計およびその学習法については 3.3.4 節で詳述する．最後に，ステップ (3) については 3.3.5 節で述べる．

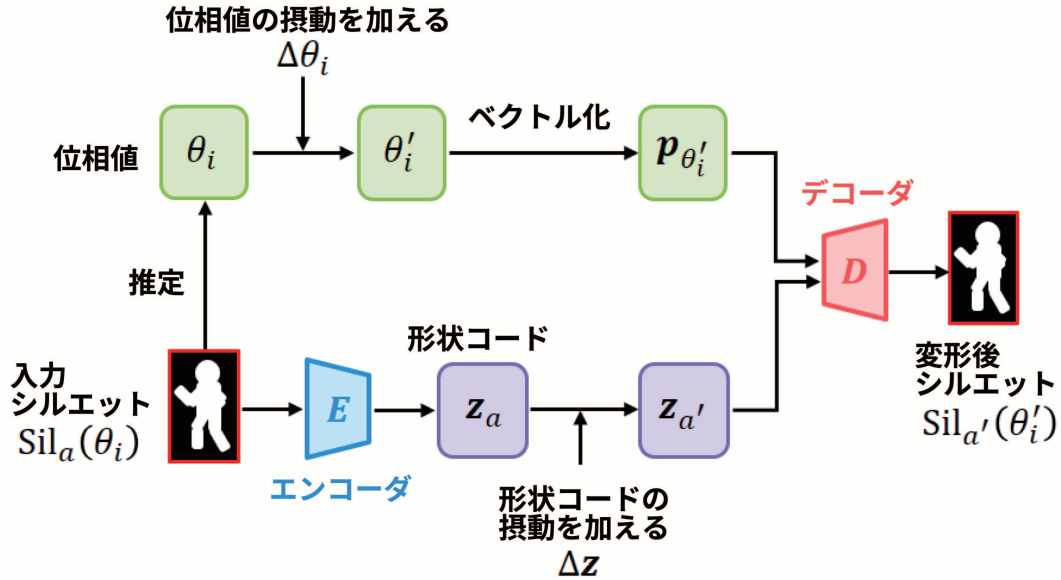


図 3.2: DNN を用いた歩容シルエット変形の概要図

3.3.2 位相値の推定と位相コードの定義

歩容シルエットの位相は，同一の姿勢に対し同一の値が対応づいてさえいれば，どのように定義しても手順 (2) 以降の内容に影響を与えない．例えば，「両足が地面に接している状態」の位相は，個人によらず常に同じ値でさえあれば，その値自体は 0 でも π でも他の値でも支障はない．このことを踏まえ，本手法では以下の手順で任意の歩容シルエットに位相を与える．

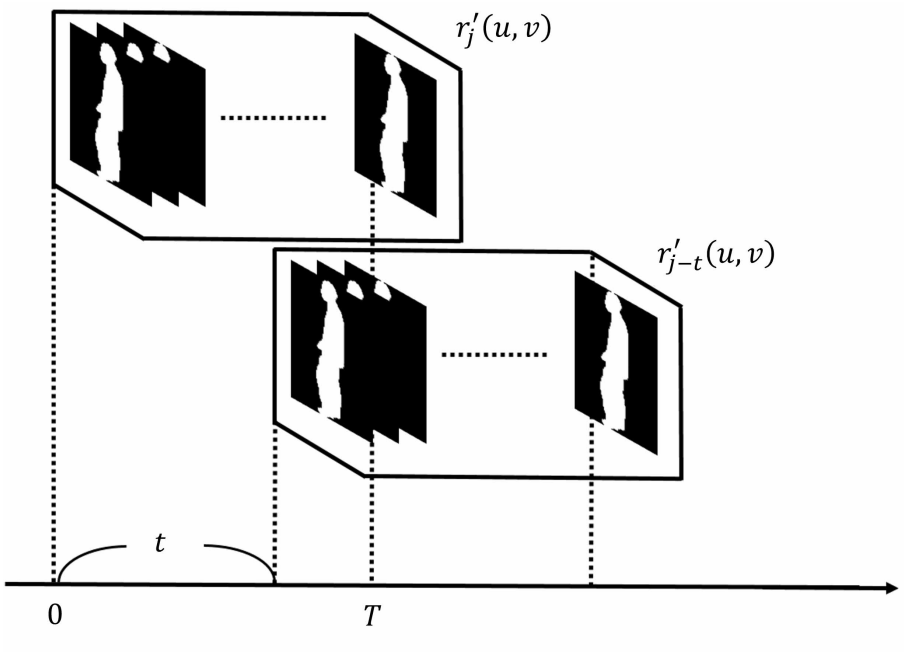


図 3.3: 自己相関

まず，位相を定義するための基準となる歩容シルエット動画 $Q = \{q_1, \cdots, q_{N_{\mathrm{sd}}}\}$ を用意する．以降ではこれを基準系列と呼ぶ．基準系列は一周期分のシルエット系列のみを含み，その長さ（フレーム数）を N_{sd} とする．次に，基準系列中の各歩容シルエット q_i に対し，その位相を $2\pi(i-1)/N_{\mathrm{sd}}$ と定義する．続いて，位相推定の対象となる入力動画からも一周期分のシルエット系列 $R = \{r_1, \cdots, r_M\}$ を抜き出す（M は入力動画における一周期分の長さを表す）．その後，R と Q の間で対応付けを行い，各 $r_j (j = 1, \cdots, M)$ がどの q_i に対応付くかを求める．これにより r_j が q_i に対応付いた場合，r_j に q_i と同じ位相値，すなわち，$2\pi(i-1)/N_{\mathrm{sd}}$ を与える．入力動画のうち R に含まれない箇所の位相については，同じ姿勢の歩容シルエットを含むフレームを R から探索したのち，対応する位相値を与えれば良い．以下，入力動画からその一周期分を抜き出す手法，および R と Q の間で対応付けを行う手法について詳述する．

まず，一周期分の抜き出しは，時間方向に関する自己相関（図 3.3 参照）を用いて行う．入力動画を $R' = \{r'_1, \cdots, r'_T\}$ とし（T は入力動画の長さであり $T > M$ を満たす），r'_j の画素 (u, v) における画素値を $r'_j(u, v) \in \{0, 1\}$（0 は人物領域外部，1 は人物領域内部）と表記すると，R' を t フレーム分ずらして R' 自身と比較

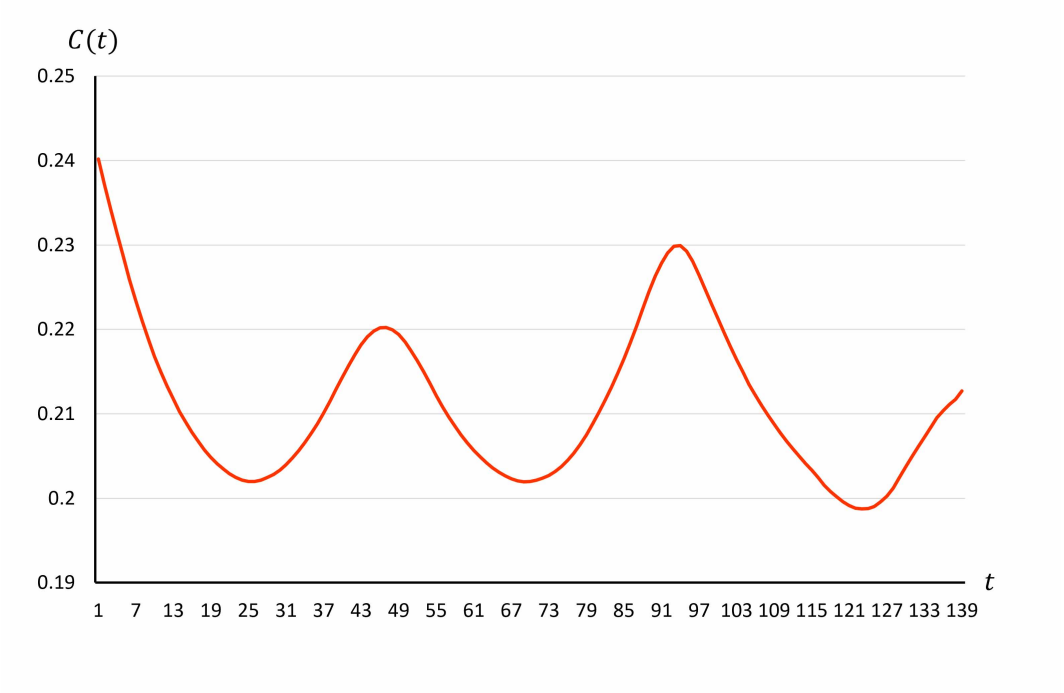


図 3.4: シフトしたフレーム数 t に対する自己相関の大きさ

した際の自己相関は

$$C(t) = \frac{1}{T-t} \sum_{j=t+1}^{T} \sum_{v=0}^{H-1} \sum_{u=0}^{W-1} r'_j(u,v) r'_{j-t}(u,v) \tag{3.1}$$

で与えられる．ここで H および W は入力動画の各フレームの縦幅および横幅である．この $C(t)$ は $t = 0$ において最大値を取ることは自明であるが，それ以外では，t が入力動画の歩行周期に一致するときに極大値をとる．ある入力動画における t と $C(t)$ の関係の一例を図 3.4 に示す．図 3.4 に示されているように $C(t)$ は一定の間隔で極大値をとる．今回の入力動画は歩行中の人物を側面から撮影した場合のシルエットであったため，実際の歩行周期の約半分ごとに類似したシルエットが現れる．このため，$C(t)$ の極大値が一周期分の間に 2 回現れているが，このような場合でも 2 回目の極大値の方が大きな値をとる．これらのことを踏まえ，$1 \leq t < T$ において $C(t)$ が偶数回目に極大となるときの t の集合を求め，これらの中で最も値が小さい t(これを $\hat{t}$ とする) を入力動画の歩行周期とする．その上で，R' から先頭 M フレームを抜き出し，これを R とする．

次に，R と Q の対応付けに際しては，DP マッチングを用いる．ここで，DP マッ

チングでは対応付けの対象となる 2 系列の先頭要素同士および終端要素同士が互いに類似していることが前提となるが，今回の場合は，一周期分の切り出し位置が個々の歩容シルエット動画によって異なるため，r_1 と q_1 の姿勢は必ずしも一致しない．このため，本手法では，図 3.5 のようにして R の先頭要素を m だけずらした（周期的シフトした）系列 $R_m = \{r_m, r_{m+1}, \cdots, r_M, r_1, r_2, \cdots, r_{m-1}\}$ を任意の $2 \leq m \leqq M$ について作成し，R_m と Q の間で DP マッチングを行う．その際のマッチングコスト $\mathrm{cost}(Q, R_m)$ に対し，

$$\hat{m} = \underset{m}{\mathrm{argmin}} \{\mathrm{cost}(Q, R_m)\} \tag{3.2}$$

としてマッチングコストが最小となるときの $\hat{m}$ を求める．$R_{\hat{m}}$ と Q は，その先頭要素同士および終端要素同士が互いに類似していることが期待される．よって，$R_{\hat{m}}$ と Q のマッチング結果を最終的な対応付け結果として採用する．

なお，Q には歩容シルエット動画の中で一周期のフレーム数 N_{sd} が最も長いものを選ぶ．これは，Q の位相値の分解能 (位相値のパターン) を向上させるためである．$R_{\hat{m}}$ の位相は上記の Q とのマッチング結果により付与されるが，Q の位相値の分解能が低い場合は $R_{\hat{m}}$ に付与できる位相値のパターンが少なくなってしまい，姿勢に対して適切な位相値を付与できない場合がある．よって，上記のように Q を選択する必要がある．

位相コード $\boldsymbol{p}_\theta$ は，上記のように推定した位相値 $\theta \in [0, 2\pi)$ に対し

$$\boldsymbol{p}_\theta = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \in \mathbb{R}^2 \tag{3.3}$$

と定義する．この定義に際しては，θ そのものを一次元の位相コードとして用いた場合，位相 0 と位相 2π が別のものとして区別され，本来両者の間に存在している連続性が失われてしまう．この問題を避けるため，本手法では上記のように位相コードを定義する．

3.3.3 形状コードの定義

同一人物の歩容シルエットからは位相の値にかかわらず同一の形状コード値が得られる必要がある．これを踏まえ，本手法では次のようにして形状コードを定義する．まず，画像からの特徴量抽出と特徴量からの画像生成を DNN により行う手法の一種である変分オートエンコーダ (Variational Autoencoder; VAE) [57] に

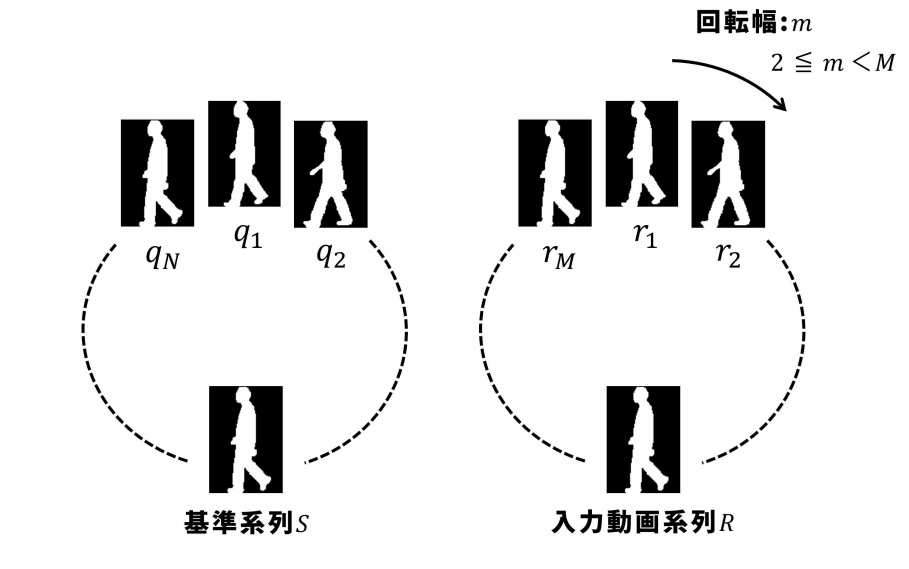


図 3.5: 入力動画系列の位相の決定方法

より，歩容シルエット画像を固定次元の特徴ベクトルに変換するエンコーダ E_{vae}，および，その特徴ベクトルを歩容シルエット画像に変換するデコーダ D_{vae} を学習する．次に，入力動画の各フレーム $\mathrm{Sil}_a(\theta_i)(i = 1, \cdots, N_a)$ に対し E_{vae} から特徴ベクトル $\boldsymbol{\xi}_a(\theta_i) = E_{\mathrm{vae}}[\mathrm{Sil}_a(\theta_i)]$ を抽出する．以上により得られた特徴ベクトルは位相によってその値が異なるため，これらの平均

$$\boldsymbol{z}_a^{\mathrm{gt}} \;=\; \frac{1}{N_a}\sum_{i=1}^{N_a}\boldsymbol{\xi}_a(\theta_i) \tag{3.4}$$

を計算し，これを入力動画中の人物 a の形状コード z_a の正解データとする．後述する実験において実際に使用した VAE の構造は図 3.6 に示す通りである．

3.3.4　歩容シルエット生成ネットワークの設計と学習

提案手法において必要となる写像は，歩容シルエット画像を 3.3.2 節および 3.3.3 節で定義した位相・形状コードに変換するものと，位相・形状コードを歩容シルエット画像に逆変換するものの 2 つである．このうち位相コードは，3.3.2 節の手法により DNN を用いずとも容易に取得可能である．また，位相・形状コードから歩容シルエット画像への逆変換には，3.3.3 節で述べた特徴量 $\boldsymbol{\xi}_a(\theta_i)$ およびデコーダ D_{vae} が再利用できる．従って，実際に学習が必要な写像は，入力の歩容シルエッ

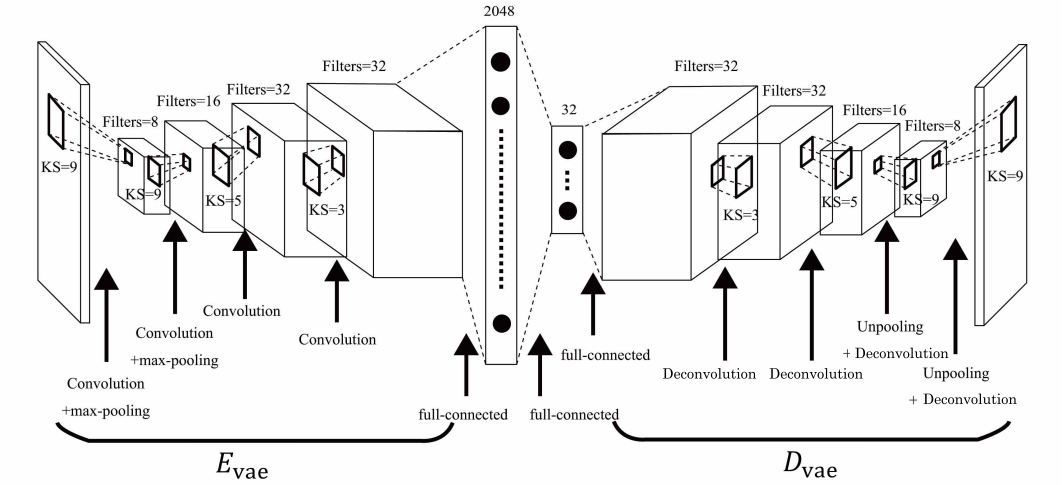


図 3.6: 実装した VAE

ト画像 $\mathrm{Sil}_a(\theta_i)$ を形状コード $\boldsymbol{z}_a$ に変換するエンコーダ E_s，および，$\boldsymbol{p}_\theta$ と $\boldsymbol{z}_a$ を一つの特徴ベクトル（統合コード）$\tilde{\boldsymbol{\xi}}_{a,i}$ へと統合するコード統合器 $\mathscr{F}$ の二つとなる．本手法では，図3.7に示すように，E_s および $\mathscr{F}$ を構成要素として含む単一の DNN を考える．

図 3.7 の DNN の学習に際しては，$\boldsymbol{\xi}_a(\theta_i)$ および $\boldsymbol{z}_a^{\mathrm{gt}}$ が既知であるような歩容シルエット画像 $\mathrm{Sil}_a(\theta_i)$ を訓練データとして用いる．これらは，一周期分のシルエットが完全な形で得られている歩容シルエット動画に対し，3.3.3 節で述べた VAE を適用することにより取得できる．これらを用いて，損失関数

$$
\begin{aligned}
L_1(E_s, \mathscr{F}) = \sum_a \sum_{i=1}^{N_a} \Big\{ & ||\boldsymbol{z}_a - \boldsymbol{z}_a^{\mathrm{gt}}||^2 \\
& + \lambda \left(||\tilde{\boldsymbol{\xi}}_{a,i} - \boldsymbol{\xi}_a(\theta_i)||^2 + ||D_{\mathrm{vae}}[\tilde{\boldsymbol{\xi}}_{a,i}] - \mathrm{Sil}_a(\theta_i)||^2 \right) \Big\},
\end{aligned} \tag{3.5}
$$

が最小化されるように E_s および $\mathscr{F}$ を同時に学習する．この時，$\tilde{\boldsymbol{\xi}}_{a,i} = \mathscr{F}[E_s[\mathrm{Sil}_a(\theta_i)], \boldsymbol{p}_{\theta_i}]$，$\boldsymbol{z}_a = E_s[\mathrm{Sil}_a(\theta_i)]$ となる．上式において，第一項は E_s により形状コードが正しく得られることを保証するための項であり，第二項は元の歩容シルエット $\mathrm{Sil}_a(\theta_i)$ を復元可能な統合コード $\tilde{\boldsymbol{\xi}}_{a,i}$ が，$\mathscr{F}$ により正しく得られることを保証するための項である．さらに，第三項はネットワークの出力画像が元の歩容シルエット $\mathrm{Sil}_a(\theta_i)$

を正しく復元することを保証する項である．なお，学習時には摂動 $\Delta \boldsymbol{z}$，$\Delta\theta_i$ はともに 0 とし，考慮しない．また，E_s および $\mathscr{F}$ に相当する部分の具体的なネットワーク構造は図 3.7 および図 3.8 に示す通りである．

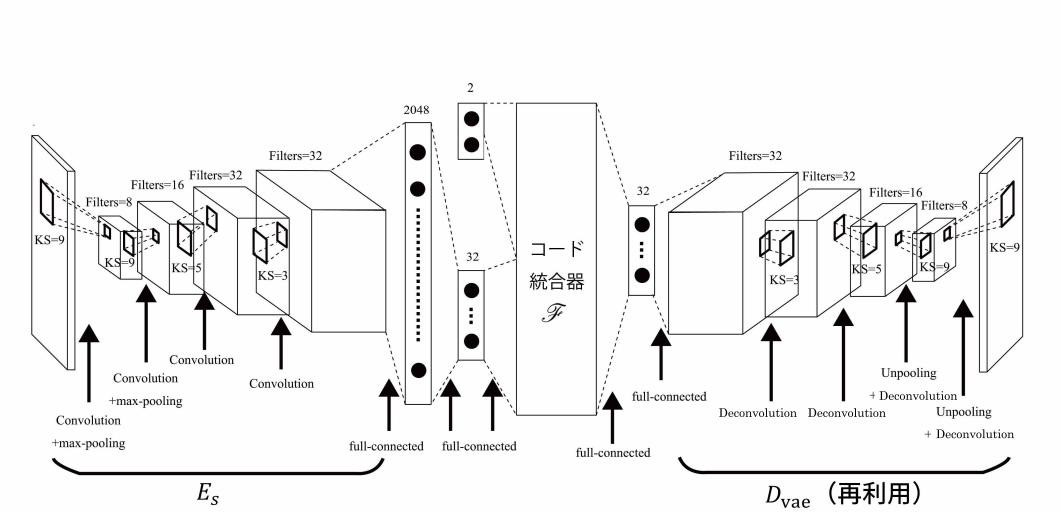


図 3.7: 実装した DNN

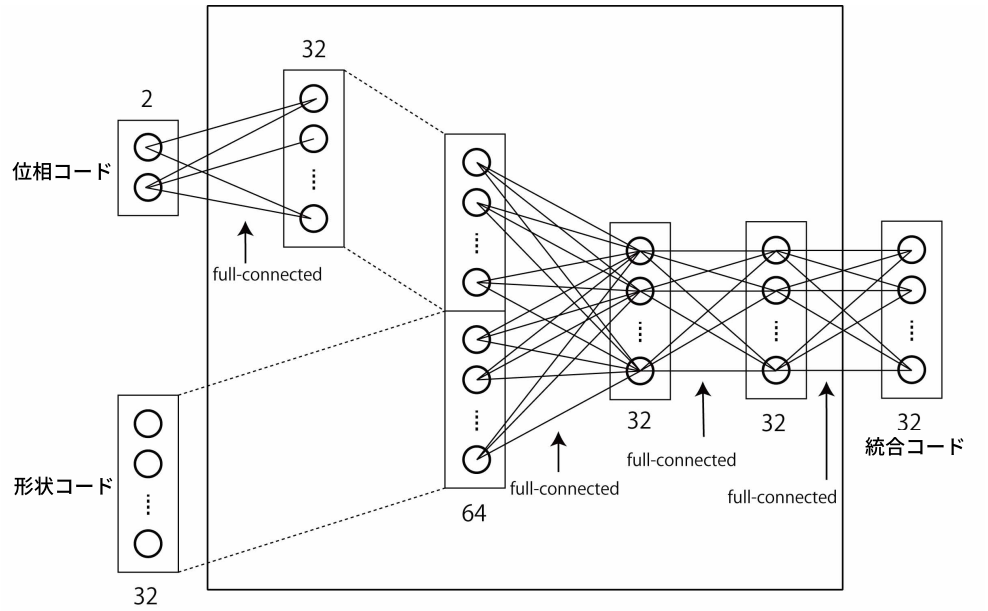


図 3.8: コード統合器 $\mathscr{F}$

3.3.5　位相と形状の摂動による歩容シルエットの匿名化

前節で設計・学習した DNN は，位相成分および形状成分に摂動を与えない場合，入力の歩容シルエット画像と同一の画像を出力するように機能する．従って，匿名化のためには，摂動 $\Delta\theta_i$，$\Delta \boldsymbol{z}$ を与えることが必須となる．本節では，その与え方を述べる．位相コードの摂動の与え方については【位相値の摂動】節で述べ，形状コードの摂動の与え方については【形状コードの摂動】節で説明する．

位相値の摂動

人間の歩行の動的側面は，単一の位相値ではなくその系列で表現される．従って，位相の摂動 $\Delta\theta_i$ はフレームごとに個別に決定されるべきではない．そこで，提案手法ではすべての i $(1 \leq i \leq N)$ に対して同時に $\Delta\theta_i$ を決定する．

一般的に，1 つの動画の中で人の歩行方向は一定であり，人や物の姿勢は連続的に変化する．従って，歩行運動の 1 周期を表す位相値の列 $(\theta_1, \cdots, \theta_N)$ は，以下の式を満たす．

$$\sum_{i=1}^{N} \phi_i = 1 \ \text{ where } \ \phi_i = \frac{1}{2\pi}\{(\theta_i - \theta_{i-1}) \bmod 2\pi\} \tag{3.6}$$

なお，便宜上 $\theta_0 = \theta_N$ とする．全ての i について $\theta_i \geq 0$ を満たすので，$\Phi = (\phi_1, \cdots, \phi_N)$ は実質的に確率分布であると考えることができる（図 3.9 参照）．同様に，摂動後の位相値系列 $(\theta'_1, \cdots, \theta'_N)$ についても，$\Phi' = (\phi'_1, \cdots, \phi'_N)$ は確率分布であると考えることができる．このとき，

$$\phi'_i = \frac{1}{2\pi}\{(\theta'_i - \theta'_{i-1}) \bmod 2\pi\} \tag{3.7}$$

となる．入力系列の動的特徴を匿名化するためには，Φ' を Φ とできるだけ類似しないようにする必要がある．

2 つの確率分布 Φ と Φ' の非類似度を測定するために，Jensen-Shannon (JS) divergence を採用する．これは

$$\mathrm{JS}(\Phi||\Phi') = \frac{1}{2}\sum_{i=1}^{N}\left\{\phi_i \log\frac{2\phi_i}{\phi_i+\phi'_i} + \phi'_i \log\frac{2\phi'_i}{\phi_i+\phi'_i}\right\} \tag{3.8}$$

で計算される値であり，Φ と Φ' が類似しているほど小さくなる．$\mathrm{JS}(\Phi||\Phi')$ を最

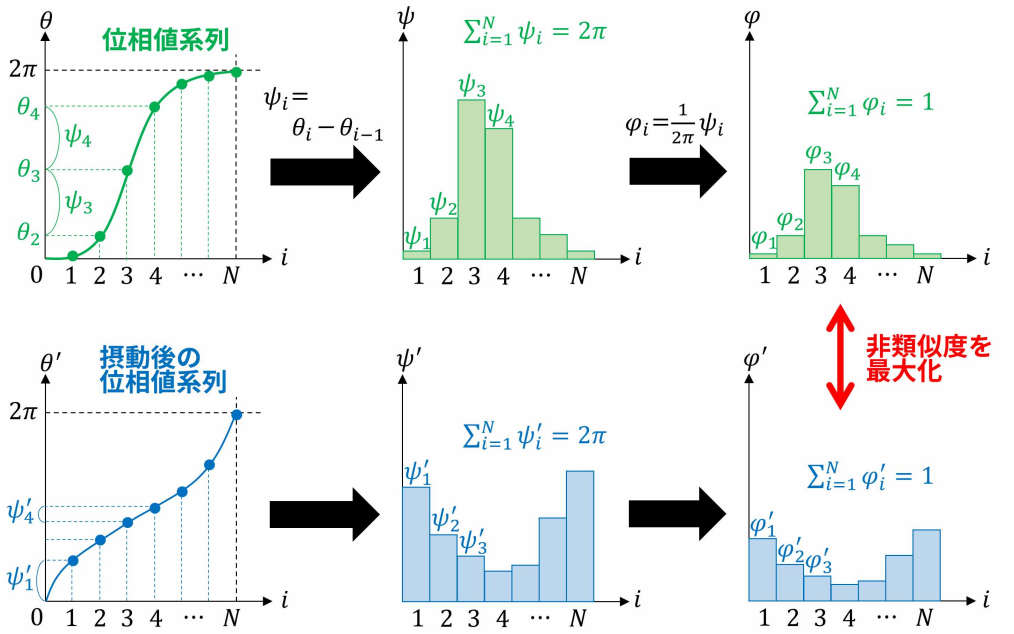


図 3.9: 位相値系列に摂動を与える提案手法

大化する最適な Φ' は

$$\phi'_i = \frac{(\phi_i)^\alpha}{\sum_{j=1}^{N}(\phi_j)^\alpha} \tag{3.9}$$

の $\alpha \to -\infty$ における極限値を全ての i について求めることで得られる．しかし，この解は分布 Φ' の偏りを最大化することにもなり，出力される歩容シルエット系列の姿勢変化が不連続になってしまう．この問題は $\alpha = 0$ とすることで回避可能であるが，その結果，すべての i について $\phi'_i = 1/N$ となり，匿名化能力が低下する．そこで，α をハイパーパラメータとみなし，3.5 節で述べる実験において，経験的にその最適値を探索する．以上の過程により見出された最適な Φ' を用いて，摂動後の位相値を以下のように計算する．

$$\theta'_i \;=\; \theta_0 + 2\pi \sum_{j=1}^{i} \phi'_j \tag{3.10}$$

これは位相の摂動 $\Delta\theta_i$ を $\Delta\theta_i = 2\pi\sum_{j=1}^{i}(\phi'_j - \phi_j)$ として与えることに相当する．摂動後の位相値に対応する位相コードは

$$\boldsymbol{p}_{\theta'_i} = \begin{pmatrix} \cos\theta'_i \\ \sin\theta'_i \end{pmatrix} \tag{3.11}$$

として計算される．

形状コードの摂動

次に，形状コードの摂動 $\Delta \boldsymbol{z}$ をどのように決定するかについて述べる．この $\Delta \boldsymbol{z}$ の大きさ，つまりノルムは重要な要素となる．もし $||\Delta \boldsymbol{z}||$ が大きすぎると，出力されるシルエット画像の形状が入力のシルエット画像と大きく異なってしまい，人体のシルエットとしては不自然な形状になる可能性がある．一方で，$||\Delta \boldsymbol{z}||$ が極端に 0 に近いと，匿名化能力が著しく低下してしまう．よって，適切な大きさとなるように $\Delta \boldsymbol{z}$ を定める

k-Same と呼ばれる顔画像匿名化手法 [26, 27] では，匿名化対象の顔領域に類似した顔画像をデータベース中から k 枚検索し，その平均画像で元の顔領域を置き換える．この方法に基づき入力動画中の人物に類似した形状コードを持つ複数の人物を用意し，それらの人物の形状コードの平均値で元の形状コードを置き換えるとすると，このためには，様々な人物の形状コードを保存したデータベースが必要になる．

本手法では，3.3.3 節の VAE の学習に使用したものと同じデータセットを用いて形状コードのデータベースを用意する．その後，この形状コードのデータベースの中から $\boldsymbol{z}_a$ に近いものを最近傍探索により K 個選択する．これらを $\boldsymbol{z}_{a,k}(k = 1, \cdots, K)$（$\boldsymbol{z}_{a,k}$ は k 番目に $\boldsymbol{z}_a$ に近い形状コード）と置くと，匿名化後の形状コード $\boldsymbol{z}_{a'}$ は

$$\boldsymbol{z}_{a'} \;=\; \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{z}_{a,k} \;=\; \sum_{k=1}^{K} \frac{1}{K}\boldsymbol{z}_{a,k} \tag{3.12}$$

となる．しかし，この手法で計算された $\boldsymbol{z}_{a'}$ は特徴空間上で常に $\boldsymbol{z}_a$ から大きく離れた場所に存在するとは限らない（つまり，常に歩容シルエットが匿名化できるとは限らない）．仮に，$\boldsymbol{z}_{a,k}$ が特徴空間上で均一に配置されていた場合，$\boldsymbol{z}_{a'}$ は $\boldsymbol{z}_a$ と非常に近い場所に存在する．これは匿名化後のシルエットとして匿名化前のシルエットと非常に近しいシルエットが出力されることに相当し，出力されたシルエットが十分に匿名化されない可能性がある．この問題を解決するために，$\boldsymbol{z}_{a,k}$ に対しそれぞれ係数 c_k を付与し，式 (3.12) を以下のように拡張する．

$$\boldsymbol{z}_{a'} \;=\; \sum_{k=1}^{K} c_k \boldsymbol{z}_{a,k} \tag{3.13}$$

ここで，さらに Z を k 番目の列が $\boldsymbol{z}_{a,k}$ である行列 $Z = (\boldsymbol{z}_{a,1} \ \cdots \ \boldsymbol{z}_{a,K})$，同様に，$\boldsymbol{c}$ を k 番目の列が c_k である行列 $\boldsymbol{c} = (c_1 \ \cdots \ c_K)^\top$ とすると，Z と $\boldsymbol{c}$ を用いて，$\boldsymbol{z}_{a'} = Z\boldsymbol{c}$ と書き換えることができる．これを踏まえ，

$$\hat{\boldsymbol{c}} \ = \ \underset{\boldsymbol{c}}{\mathrm{argmax}} \, ||\Delta \boldsymbol{z}||^2 \ = \ \underset{\boldsymbol{c}}{\mathrm{argmax}} \, ||Z\boldsymbol{c} - \boldsymbol{z}_a||^2 \tag{3.14}$$

となる $\hat{\boldsymbol{c}}$ を求めることができれば，十分な匿名化能力を持つ程度に $||\Delta \boldsymbol{z}|| = ||\boldsymbol{z}_{a'} - \boldsymbol{z}_a||$ が大きくなり，特徴空間上で $\boldsymbol{z}_a$ と近い場所に存在しない $\boldsymbol{z}_{a'}$ が得られる．

上記の手法の問題点としては，$\boldsymbol{c}$ に制限がないため，出力結果の歩容シルエット動画の見た目が非常に歪む可能性があることである．これを防ぐために，以下の 2 つの制約を与える．

$$\sum_{k=1}^{K} c_k \ = \ 1 \quad \Longleftrightarrow \quad \boldsymbol{c}^\top \mathbb{I} \ = \ 1 \tag{3.15}$$

$$\forall k \in \{1, \ \cdots, \ K\} \quad 0 \leq c_k \leq 1 \tag{3.16}$$

ここで，$\mathbb{I} = (1 \ \cdots \ 1)^\top$ はすべての要素が 1 の K 次元のベクトルである．これらの制約により，出力される歩容シルエット動画の見た目は自然となることが保証される．計算を簡単にするために，制約式 (3.16) を直接考慮せず，この制約を目的関数に導入する．すなわち，制約式 (3.15) の下で次式の $\hat{\boldsymbol{c}}$ を計算する．

$$\begin{aligned} \hat{\boldsymbol{c}} \ &= \ \underset{\boldsymbol{c}}{\mathrm{argmax}} \left\{ \eta ||Z\boldsymbol{c} - \boldsymbol{z}_a||^2 + \sum_{k=1}^{K} c_k(1 - c_k) \right\} \\ &= \ \underset{\boldsymbol{c}}{\mathrm{argmax}} \left\{ \eta ||Z\boldsymbol{c} - \boldsymbol{z}_a||^2 + \boldsymbol{c}^\top (\mathbb{I} - \boldsymbol{c}) \right\} \end{aligned} \tag{3.17}$$

上式の第二項は制約式 (3.16) に対応し，η は第一項と第二項のバランスを定める正の定数である．これは二次最適化問題であるため，その解はラグランジュの未定乗数法により求めることができる．具体的には，I_K を $K \times K$ の単位行列とし，また，$\boldsymbol{q} = \eta \, G^{-1} Z^\top \boldsymbol{z}_a$，$\boldsymbol{\nu} = G^{-1}\mathbb{I}$，そして $G = \eta \, Z^\top Z - \mathrm{I}_K$ とすると，

$$\hat{\boldsymbol{c}} \ = \ \boldsymbol{q} + \frac{1 - \mathbb{I}^\top \boldsymbol{q}}{\mathbb{I}^\top \boldsymbol{\nu}} \boldsymbol{\nu} \tag{3.18}$$

となる．ここで，$\eta = 0$ の場合には $\boldsymbol{c} = \mathbb{I}/K$ が得られる．このとき，計算結果は式 (3.12) で計算された $\boldsymbol{z}_{a'}$ と同等のものとなる．

次に，パラメータ η が取るべき値について詳述する．上記のように，式 (3.17) は $\boldsymbol{c}$ に関する二次式であり，二次の項のみを取り出すと

$$\eta \boldsymbol{c}^\top Z^\top Z \boldsymbol{c} - \boldsymbol{c}^\top \boldsymbol{c} \ = \ \boldsymbol{c}^\top G \boldsymbol{c} \tag{3.19}$$

となる．このことから，G が半負定値である場合のみ，上記の最適化問題を正しく解くことができる．G が半負定値となるためには，$Z^\top Z$ の最大固有値を τ としたとき，η は $1/\tau$ より小さくなくてはならない．これらのことから，本研究では $0 \leq \omega < 1$ を満たすハイパーパラメータ ω を導入し $\eta = \omega/\tau$ とする．$\omega = 1$ とした場合は G は非正則となり，G^{-1} の計算が不可能になる．

以上を踏まえ，$\eta = \omega/\tau (0 \leq \omega < 1)$ の上で式 (3.18) を計算し，求めた $\hat{\boldsymbol{c}}$ と式 (3.13) により得られる $\boldsymbol{z}_{a'}$ を匿名化後の形状コードとする．これは，形状成分に関する摂動 $\Delta \boldsymbol{z}$ を

$$\Delta \boldsymbol{z} = \boldsymbol{z}_{a'} - \boldsymbol{z}_a \tag{3.20}$$

として与えることに相当する．

3.4　匿名化歩容シルエット動画へのテクスチャ転写

本節では，前節で得られた変形シルエットに，匿名化前の人物領域のテクスチャを転写する方法について説明する．

3.4.1　テクスチャ情報転写の概要

3.3 節では，歩容シルエットの位相特徴と形状特徴に摂動を加え，匿名歩容シルエットを生成した．これを踏まえて，本節以降では，位相特徴と形状特徴の変化に対応したテクスチャ情報の転写手法を提案する．

フォトリアリスティックな人物動画の生成手法 [48, 49] では，動画中の人物領域のテクスチャを任意のターゲット姿勢に合わせて転写するために CNN ベースのエンコーダ・デコーダ構造がよく利用される．これは，理論的には本研究のテクスチャ転写のタスクに適用することができる．しかし，上記の手法で必要となる膨大な学習データを収集することは困難であるため，この手法を本研究のタスクに適用した場合，様々な服装に対応することができない．そこで，本研究では，変位ベクトル場の考え方に基づき，学習の必要がないテクスチャ転写手法を採用する．

まず，変位ベクトルと変位ベクトル場の定義について述べる．変位ベクトル場とは，2 つの固定サイズの画像 $J(x, y)$ と $I(x, y)$ の間の画素単位の変位ベクトルを表す 2 次元のベクトル場である．このとき，J と I は互いにわずかに異なるものとする．一方の画像 J 上の画素 (x, y) が，もう一方の画像 I 上の画素 (u, v) に対応するとき，それらの変位ベクトルは $(u - x,\, v - y)$ と定義される．水平成分 $u - x$

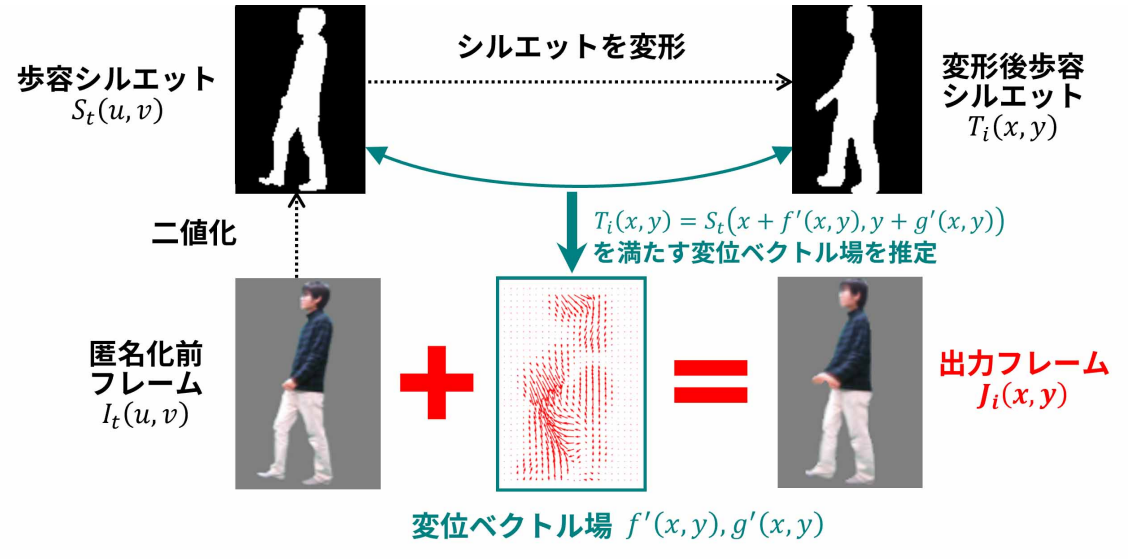


図 3.10: テクスチャ転写の提案手法の概要図

と垂直成分 $v - y$ はともに画素の位置によって変化するので，それぞれ $f'(x,y)$, $g'(x,y)$ と表現する．つまり，画像 J と I は全ての (x,y) について以下の式を満たす．

$$J(x,y) \approx I(u,v) = I(x + f'(x,y),\ y + g'(x,y)) \tag{3.21}$$

ここで，f' と g' のペアを変位ベクトル場とする．変位ベクトル場の典型的な例としては，動画の連続する 2 フレーム間で計算されるオプティカルフローがある．

3.3 節で提案した手法により，匿名化対象のシルエット系列 $\mathcal{S}$ から匿名化後のシルエット系列 $\mathcal{T}$ が得られている．ここで，$T_i(u,v)$ を $\mathcal{T}$ の i 番目のフレーム，$S_t(u,v)$ を $\mathcal{S}$ の t 番目のフレームとし，$S_t(u,v)$ の二値化前画像を $I_t(u,v)$ とする．このとき，T_i と S_t の間で

$$T_i(x,y) \approx S_t(x + f'(x,y),\ y + g'(x,y)) \tag{3.22}$$

を満たす適切な変位ベクトル場 f' と g' が得られれば，それを用いて T_i 上に I_t のテクスチャを転写することができる（図 3.10 参照）．テクスチャ転写後の画像を J_i とすると，J_i はアルゴリズム 1 により得られる．ここで，H と W はそれぞれ画像の高さと幅である．従って，ここでは，T_i と S_t の組から如何に適切な f' と g' を推定するかが重要となるが，S_t の位相値が T_i と大きく異なる場合にはこれらの推定は困難となる．そこで，T_i と S_t として同位相のシルエット画像のペアを選

択することも重要である．

アルゴリズム 1 変位ベクトル場に基づくテクスチャ転写

入力: $I_t,\ S_t,\ T_i$

出力: J_i

1: S_t と T_i の間で変位ベクトル場を計算し，その水平成分と垂直成分をそれぞれ f'，g' とする．
2: **for** $y = 0$ to H **do**
3: **for** $x = 0$ to W **do**
4: $u = x + f'(x, y)$
5: $v = y + g'(x, y)$
6: 画素 $I_t(u, v)$ の色を画素 $J_i(x, y)$ にコピーする（$J(x, y) \leftarrow I(u, v)$）．
7: **end for**
8: **end for**

3.4.2 同位相のシルエット画像の選択

前節で述べたように，変位ベクトル場を求めるシルエット対のそれぞれの位相が大きく異なっている場合，対応する画像同士の形が異なるため，妥当な変位ベクトル場を求めることはできない．よって，各 $T_i \in \mathcal{T} = \{T_1, \cdots, T_N\}$ について，それに最も近い位相を持つフレームを $\mathcal{S} = \{S_1, \cdots, S_N\}$ の中から選択する必要がある．これは，全ての $t \in \{1, \cdots, N\}$ について T_i と S_t の間のシルエット類似度を計算し，最も似ているものを選択することで達成される．この処理を各 $i \in \{1, \cdots, N\}$ について別々に行うのが最も単純な方法である．しかし，この方法では，出力画像列 $\{J_1, \cdots, J_N\}$ にフリッカーが多々発生する．これは，$\mathcal{S}$ において T_i の対応部分と T_{i+1} の対応部分が必ずしも時間的に隣接していないためである．フリッカーを避けるために，3.3.2 節で述べたものと同じシルエットマッチング法を採用する．具体的には，DP マッチングと周期的シフトを用いて，$\mathcal{T}$ と $\mathcal{S}$ の最適な対応関係を求める．その結果に基づいて，各 i について T_i のペアを選択する．

3.4.3 最適な変位ベクトル場の推定

3.4.2 節で選択されたシルエット対を用いて変位ベクトル場を推定する．変位ベクトル場を推定することは，最適な $f'(x, y)$，$g'(x, y)$ を求めることに相当する．以

下これらを求める手順について詳述する．

変位ベクトル場の推定では，2 つのシルエット画像 $S_t(x,y)$ と $T_i(u,v)$ 上の対応する画素同士で可能な限り色が同じになるように，以下のコスト関数を最小化する f' と g' を探索する．

$$Q_1(f',g') = \sum_{(x,y)\in\mathcal{A}} \{T_i(x,y) - S_t(x+f'(x,y), y+g'(x,y))\}^2 \tag{3.23}$$

ここで，$\mathcal{A}$ は画像 $T_i(u,v)$ の全領域である．しかし，$S_t(x,y)$ と $T_i(u,v)$ はともに画素値が 0 か 1 の二値画像であるため，上記のコスト Q_1 は 0 になりやすくなり，不連続かつ不自然で大きな変位ベクトル場となりやすい．これを避けるために，以下の 3 つの制約を導入する．

(1) 平滑性制約

(2) 単調性制約

(3) 境界制約

(1) の平滑性制約とは f'，g' がともに隣接する画素同士で値が極端に変化せず，空間的に滑らかであることを保証する制約である．隣接する画素同士で f'，g' が大きく変わることがあれば，転写後のテクスチャが元のテクスチャに対して大きく歪む．これを避けるために，f'，g' はともに滑らかであることが必要である．これは，f'，g' のラプラシアン

$$\begin{aligned}\mathcal{L}_{f'}(x,y) &= \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) f' \\ &= f'(x-1,y) + f'(x+1,y) + f'(x,y-1) + f'(x,y+1) - 4f'(x,y)\end{aligned} \tag{3.24}$$

$$\begin{aligned}\mathcal{L}_{g'}(x,y) &= \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) g' \\ &= g'(x-1,y) + g'(x+1,y) + g'(x,y-1) + g'(x,y+1) - 4g'(x,y)\end{aligned} \tag{3.25}$$

が小さいことに相当する．つまり，

$$Q_2 = \sum_{(x,y)\in\mathcal{A}} [\{\mathcal{L}_{f'}(x,y)\}^2 + \{\mathcal{L}_{g'}(x,y)\}^2] \tag{3.26}$$

を最小化することを平滑性制約として設定する．

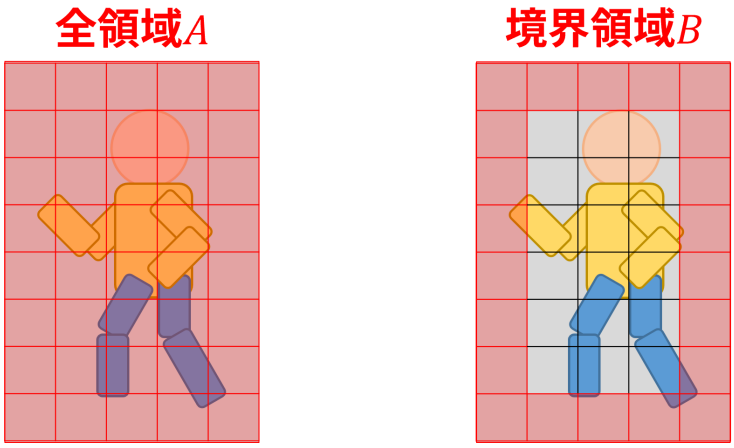


図 3.11: 画像の全領域 $\mathcal{A}$ と境界領域 $\mathcal{B}$

次に，(2) の単調性制約について，これは対応点の上下関係や左右関係が逆転することがないことを保証する制約である．この制約は u が x に対し単調増加，かつ，v が y に対し単調増加であることに相当し，次式のように定式化される．

$$\frac{\partial u(x,y)}{\partial x} = 1 + \frac{\partial f'(x,y)}{\partial x} \approx 1 + f'(x,y) - f'(x-1,y) > 0 \tag{3.27}$$

$$\frac{\partial v(x,y)}{\partial y} = 1 + \frac{\partial g'(x,y)}{\partial y} \approx 1 + g'(x,y) - g'(x-1,y) > 0 \tag{3.28}$$

これらは，$\mathcal{C}_{f'}(x,y) = \max\{0, -1 - f'(x,y) + f'(x-1,y)\}$，$\mathcal{C}_{g'}(x,y) = \max\{0, -1 - g'(x,y) + g'(x,y-1)\}$ とした上で

$$Q_3 = \sum_{(x,y)\in\mathcal{A}} [\{\mathcal{C}_{f'(x,y)}\}^2 + \{\mathcal{C}_{g'(x,y)}\}^2] \tag{3.29}$$

を最小化する問題に置き換えることができる．最後に，(3) の境界制約について，これは外周部（埋め込み対象とその外部の境界）を変形させない制約である．外周部では埋め込み対象との境界部分の不自然さを防ぐために，f'，g' をともに 0 とする．これは，

$$Q_4 = \sum_{(x,y)\in\mathcal{B}} [\{f'(x,y)\}^2 + \{g'(x,y)\}^2] \tag{3.30}$$

を最小化することに相当する．ここで，$\mathcal{B}$ は T_i の境界領域である（図 3.11 参照）．

これらの制約を踏まえ，最終的なコスト関数は以下のようになる．

$$\hat{Q}(f', g') = Q_1 + \beta_2 Q_2 + \beta_3 Q_3 + \beta_4 Q_4 \tag{3.31}$$

ここで，β_2，β_3，β_4 はそれぞれ各制約のバランスを調節する正の定数である．このコスト関数を最小化する際には，最急降下法を用いる．その手順を以下に示す．

(1) $f'(x,y), g'(x,y)$ を適当な値で初期化する．

(2) $\dfrac{\partial \hat{Q}}{\partial f'(x,y)}, \dfrac{\partial \hat{Q}}{\partial g'(x,y)}$ を求める．

(3) $f'(x,y) \leftarrow f'(x,y) - o\dfrac{\partial \hat{Q}}{\partial f'(x,y)}$，$g'(x,y) \leftarrow g'(x,y) - o\dfrac{\partial \hat{Q}}{\partial g'(x,y)}$ として，$f'(x,y), g'(x,y)$ を更新する (o は更新率)．

(4) 手順 (2) と (3) を繰り返す．

このようにして，$T_i(x,y)$ と $S_t(u,v)$ から $f'(x,y)$ と $g'(x,y)$ を求めることができる．よって，これらと $I_t(u,v)$ から 3.4.1 節のアルゴリズム 1 を用いて，$J_i(x,y)$ にテクスチャ情報を転写することができる．

3.5 評価実験

3.5.1 実験設定

使用するデータセット

本実験では，歩容シルエット動画のデータセットと通常の歩容動画のデータセットの 2 種類の歩容動画データセットを用いた．上記のうち一つは，歩容シルエット動画に関するデータセットとして，The OU-ISIR Gait Database [47] の Treadmill-dataset-(A) と Treadmill-dataset-(B) を用いた．Treadmill-dataset-(A) は，34 人の人物が 2km/h から 10km/h までの歩行速度で歩行した際の様子を側面から観測して取得した歩容シルエット動画のデータセットであり，計 612 本存在する．しかし，中には速度が遅すぎるものや速すぎるものなど歩行動作として普通でないものも存在するため，4，5，6[km/h] の速度で歩いている 204 本の動画のみを使用した．以下では，この 204 本の動画の集合を DS_a とする．Treadmill-dataset-(B) は，68 人の人物が 32 種類の服を着用して歩行した際の様子を同じく側面から観測して取得した歩容シルエット動画のデータセットであり，服装の多様性に富んでいるが，歩行速度の多様性には乏しい．動画の総数は $68 \times 32 = 2176$ 本である．

以下では，このうちランダムに抽出した 90000 フレーム分*の歩容シルエット画像を DS_b とする．

一方で，歩容動画に関するデータセットとしては，実際に歩く姿を横から Web カメラで撮影した独自のデータセットを構築した．このデータセットには 14 人の人物の動画が含まれており，1 人あたり 13〜14 パターンの動画が存在し，動画の総数は 190 本である．これは，各フレームの人物領域をクロマキー処理によって抽出し，背景部分をグレーで塗り潰したものとなっている．このように処理することで，歩容動画を簡単に二値化し，シルエットを抽出できる．以下では，このデータセットを DS_c とする．また，DS_c の動画の解像度は，DS_a および DS_b の動画の解像度と同じ 88×128 画素である．

上述のデータセットに含まれる動画は，3.3.2 節で述べた手法により，一周期分の動画となっている．

ネットワークの学習とテスト

3.3.3 節で述べた VAE のうちエンコーダ E_{vae} とデコーダ D_{vae}，3.3.4 節で述べた歩容シルエット生成ネットワークのうちエンコーダ E_s とコード統合器 $\mathscr{F}$ の学習データセットとして，DS_b を用いた．3.3.5 節で述べた，形状コードの摂動の計算に必要な形状コードのデータベースを構築するために，学習済みの E_{vae} と DS_b を用いた．その後，提案手法により DS_c のうち 64 本の歩容動画の匿名化を試み，テストを行った．

また，形状コードの摂動と位相の摂動の匿名化効果を別々に評価するため，位相成分のみを匿名化した場合（*phase-only*），形状成分のみを匿名化した場合（*shape-only*），両方を匿名化した場合（*both*）の 3 種類の匿名化について比較した．

評価の基準

第 1 章で述べた通り，匿名化後の歩容動画は，その歩容特徴から個人が認識されないことに加え，見た目が自然であることが望ましい．このため，本実験では匿名性と見た目の自然さの両面から匿名化後歩容動画を評価した．

---

* 後述するように，DS_b はネットワークの学習データとして用いられるため，基本的にはデータ数が多い方がよい．但し，歩容シルエット動画中の隣接するシルエットは類似した特徴のものも多いため，全てのシルエットを使用する必要はない．従って，ネットワークの学習時間の観点からデータ数を削減している．

これらの評価基準のうち匿名性に関しては，歩容認証精度を評価基準として採用した．匿名化後歩容動画をある歩容認証システムに入力したとき，認識精度が低ければ低いほど，匿名化性能が高いことを意味する．そのため，GEINet [11] をベースとした歩容認証システムを構築し，匿名化の前後で歩容認証精度がどの程度低下するかという観点から評価した．GEINet は歩容認証のためのニューラルネットワークで，畳み込み，プーリング，正規化層の 3 層の組が 2 つとそれに続く 2 つの全結合層から構成されている．本来，GEINet は，歩容シルエット動画の GEI を入力特徴量とすることを想定している．しかし，この GEI は 2.2 節で述べたように，歩行動作の動的情報を失っている．そこで，GEI [16]， FDF [17]，SFDEI [19] をそれぞれ入力特徴量とする 3 つのネットワークを別途構築した．これら 3 つのネットワークは，入力層のチャネルの数を除けば，GEINet と同じ構造となっている．

ネットワークの入力に用いる三つの特徴量についてその詳細を以下に述べる．まず，GEI とは，歩容シルエットの系列をその周期で平均化したものであり，

$$GEI(x,y) = \frac{1}{N_{\mathrm{GEI}}} \sum_{t=1}^{N_{\mathrm{GEI}}} B_t(x,y) \tag{3.32}$$

として計算される．上式において，$GEI(x,y)$ は GEI を表し，$B_t(x,y)$ は歩容シルエット動画中の t フレーム目を表す．また，N_{GEI} は歩容シルエット動画の一周期分の長さである．GEI は一周期分の歩容シルエットを平均化したものであるため，位相の変化パターンに関する情報は残りにくく，形状の特徴のみを反映したものとなる．

一方，FDF とは，各画素の時間軸方向の一次元離散フーリエ変換を計算し，歩行周期で正規化した低周波成分の振幅スペクトルを特徴としたものであり，

$$FDF(x,y,k) = \left| \frac{1}{N_{\mathrm{FDF}}} \sum_{t=1}^{N_{\mathrm{FDF}}} B_t(x,y) \exp\left(-j\frac{2\pi}{N_{\mathrm{FDF}}}kt\right) \right| \tag{3.33}$$

として計算される．上式において，$FDF(x,y,k)$ は FDF を表し，$B_t(x,y)$ は歩容シルエット動画中の t フレーム目を表す．また，k は周波数成分の次数を表し，N_{FDF} は歩容シルエット動画の一周期分の長さである．高周波成分はノイズが多くなるため，0 次，1 次，2 次の成分が用いられる．0 次成分は GEI と同等の特徴であり，1 次成分には左右非対称の動きが，2 次成分には左右対称の動きが現れる．本研究では画像の RGB 成分のうち R チャネルに 0 次成分，G チャネルに 1 次成分，B

チャネルに 2 次成分を使用し，これを FDF 特徴としている．FDF は動きの成分に頑健なので，位相の変化パターンの特徴を反映しやすい．

さらに，SFDEI とは，時刻 t の歩容画像と時刻 $t-\Delta t$ の歩容画像で差分をとった際の正の成分と負の成分からなる差分画像を特徴としたものである．正の成分からなる差分画像 $FD_p(x,y)$ は

$$FD_p(x,y) = \begin{cases} 255 & (I(t,x,y) - I(t-\Delta t,x,y) > 0) \\ 0 & (I(t,x,y) - I(t-\Delta t,x,y) \leq 0) \end{cases} \tag{3.34}$$

と表され，負の成分からなる差分画像 $FD_n(x,y)$ は

$$FD_n(x,y) = \begin{cases} 255 & (I(t,x,y) - I(t-\Delta t,x,y)) \leq 0 \\ 0 & (I(t,x,y) - I(t-\Delta t,x,y)) > 0 \end{cases} \tag{3.35}$$

と表される．3 チャネル画像を用いて GEI を R チャネル，正の差分画像 $FD_p(x,y)$ の平均を G チャネル，負の平均差分画像 $FD_n(x,y)$ の平均を B チャネルに設定したものが SFDEI となる．SDFEI は差分画像を特徴として用いているので，FDF と同様に動きの成分に頑健で，位相の変化パターンの特徴を反映しやすい．また，本研究では Δt を $\Delta t = 1$ として SFDEI を作成した．

これらの 3 つの特徴を入力としたそれぞれのネットワークの具体的な学習方法を以下に述べる．なお，以下で使用される歩容動画はこれらの特徴量に変換されてデータとして保存されているものとする．まず，DS_c に含まれる動画のうち，テストに使用されていない 126 本を二値化し，その結果のシルエットを学習データとして使用した．さらに，ネットワークの性能を向上させるため，DS_a に含まれる全ての動画も学習データとして使用した．また，DS_c には 14 人，DS_a には 34 人の動画が含まれているため，学習されたネットワークは合計で 48 クラス（または 48 人）を含むものとなる．

これら 3 つの GEINet ベースのネットワークを学習した後，上記の DS_c の残りの 64 本の動画を匿名化した．これらの動画に含まれる人数は DS_c の学習データ同様に 14 人となる．ネットワークのクラスは 48 クラスあるが，テスト時にはこれらのうち DS_c に含まれる 14 クラスについて分類を行う．そして，匿名化の結果を以下の 2 つの方法で評価した．一つは，3.4 節で提案した人物領域テクスチャ転写（Human Region Texture Transfer; HRTT）を行う前に，3.3 節で提案した歩容シルエットの匿名化後の結果を直接 3 つのネットワークに入力する方法である．も

う一つは，HRTT を行った後，最終的に匿名化した結果を再び二値化したものを入力する方法である．以下，この 2 つの方法を *before-HRTT*，*after-HRTT* と呼ぶ．

一方で，見た目の自然さに関しては歩容動画を客観的に評価することは容易ではない．こうした中，Tieu らは，匿名化後歩容動画の見た目の自然さの評価に人物動作認識システムを採用している [33]．これは，視覚的に自然な動きをする歩容動画は動作認識システムで「歩行」と正しく認識されやすく，不自然な動きをする動画は誤認識されやすいという考察に基づくものである．このような背景から，3D-ResNet [50] と呼ばれる動作認識用の事前学習済み CNN モデルを見た目の評価に用いた．また，生成された匿名化後歩容動画の人物領域が正しく検出されるかを確かめるために YOLO [51] と呼ばれる物体検出モデルを採用し，その検出精度を測定した．最終的に，提案手法で匿名化した歩容動画を 3D-ResNet に対して動画単位で入力し認識精度（「歩行」であると認識される割合）を測定し，また，YOLO に対して動画のフレーム単位で入力しその検出精度（人物領域であると検出される割合）を測定した．

3.5.2 ハイパーパラメータの設定

本節では，提案手法の匿名化性能と歩容動画の見た目の自然さに大きな影響を与えるハイパーパラメータについて述べる．これらのパラメータのうち，形状コードの摂動に関するものは ω と K，位相値の摂動に関するものは α である．また，このようなハイパーパラメータを調整するために，まずその影響について調べ，その結果を本節で述べる．後述するように，最終的に実験を通して $\omega = 0.99$，$K = 20$，$\alpha = -0.5$ を採用するに至った．これら 3 つのパラメータの影響を調べるにあたっては，様々な組み合わせを考えたが，本研究ではその一部についてのみ述べる．具体的には，3 つのパラメータのうち 2 つのパラメータを固定し，残り 1 つのパラメータのみを変更した場合について述べ，それをもとに，変更したパラメータの影響を考察した．また，式（3.31）における人物領域テクスチャ転写のハイパーパラメータ，すなわち，β_2，β_3，β_4 は経験に基づき，$\beta_2 = 0.3$，$\beta_3 = 10$，$\beta_4 = 5$ とした．

形状コードの摂動に関するハイパーパラメータ

3.3.5 節で述べたように，形状コードの摂動は元の形状コードをその形状コードに最も近い K 人分の形状コードの線形結合で置き換えることで実現される．ま

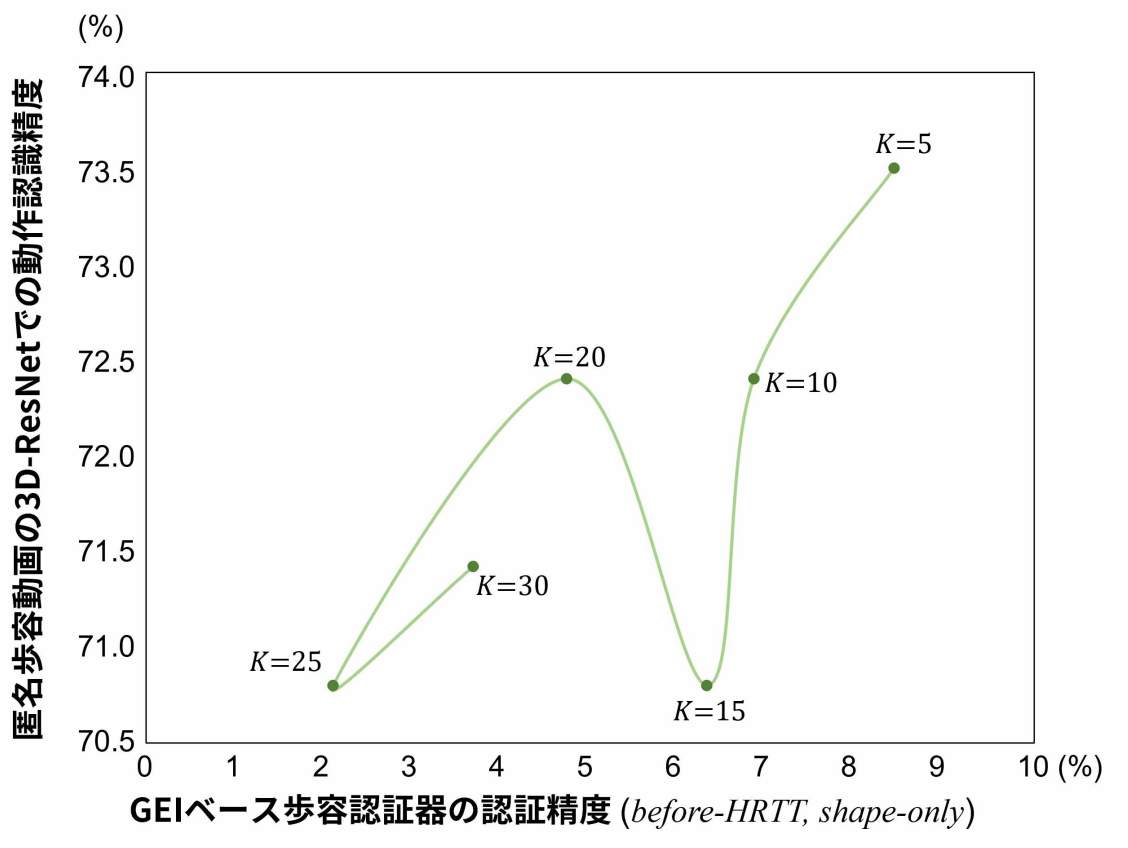


図 3.12: K が匿名化性能と歩容動画の見た目の自然さに与える影響

ず，その影響を調べたところ，図 3.12 と図 3.13 に示すような結果が得られた．これらの図の横軸は，*shape-only* の匿名化を行い，*before-HRTT* のシルエットを用いた場合の GEI による歩容認証精度を示している．また，これらの図の縦軸は *shape-only* 手法によって生成された匿名化後歩容動画を 3D-ResNet に入力した際の動作認識精度を示している．つまり，これらの図では図の中に存在する点が左側にあればあるほど匿名化性能が高く，上側にあればあるほど見た目の自然さが向上していることを示している．よって，より最適なハイパーパラメータを設定するためにはこの図の点ができるだけ左上に近づくように各ハイパーパラメータを設定する必要がある．

図 3.12 において，基本的に K が大きいほど匿名化性能が高く，特に $K \geq 20$ の場合に高い匿名化性能となることがわかる．一方，見た目の自然さについては，K は動作認識精度に大きな影響を与えず，精度が若干変化する程度となっている．

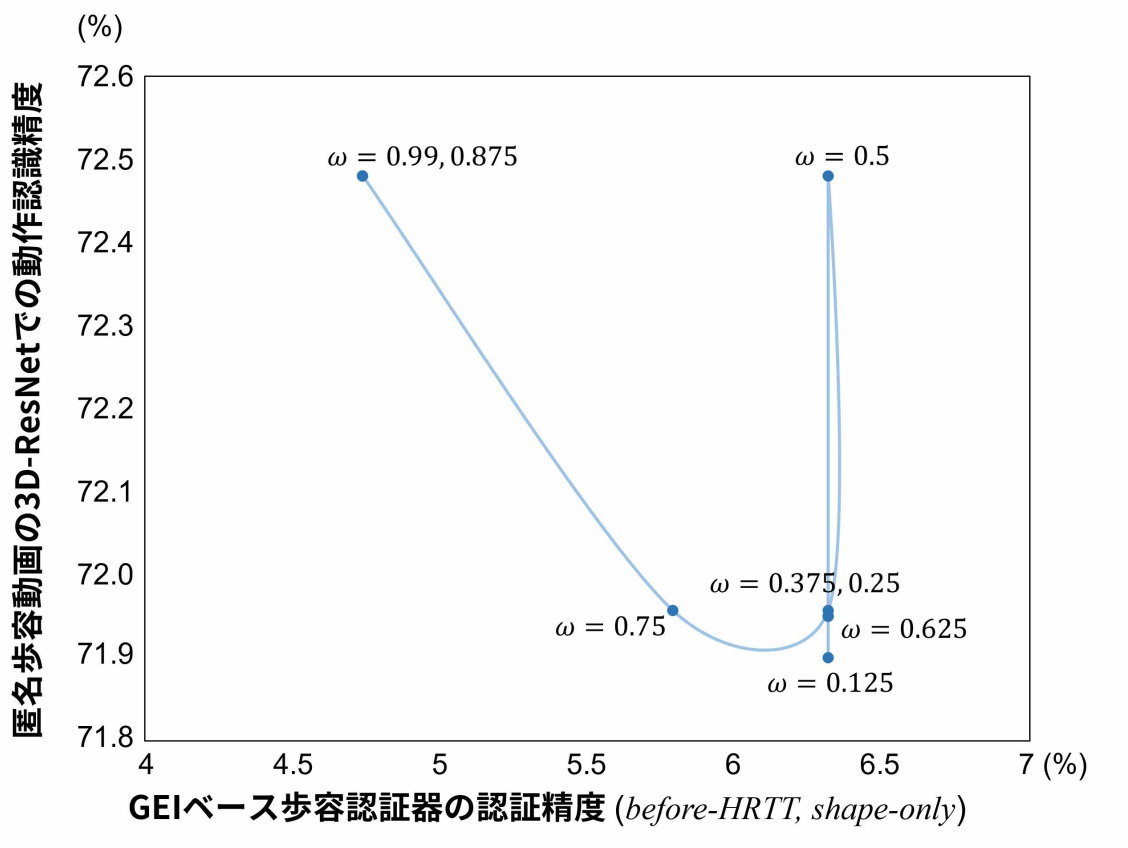


図 3.13: ω が匿名化性能と歩容動画の見た目の自然さに与える影響

ここでは，図 3.12 の左上に最も近いと思われる $K = 20$ を採用した．

また，図 3.13 でも同様の傾向が見られた．つまり，ω が大きいほど匿名化性能が高くなるが，見た目の自然さには大きな影響を与えないという結果が得られた．そこで，図 3.13 の左上隅に最も近い点である，$\omega = 0.99$ を採用した．

位相値の摂動に関するハイパーパラメータ

位相値の摂動を付与する際に必要となる，式 (3.9) のハイパーパラメータ α は，元の位相配列と摂動を受けた位相配列がどれだけ差異を持つかを決定づける．理論上は，$\alpha \to -\infty$ でこの差異が最大となるが，この場合，匿名化後歩容動画の見た目の自然さが損なわれてしまう．上記の理論的背景を実験的に検証した結果を図 3.14 に示す．この図における横軸，縦軸の意味は図 3.12，図 3.13 と基本的には同様であるが，*shape-only* ではなく，*phase-only* を使用しているという点で異な

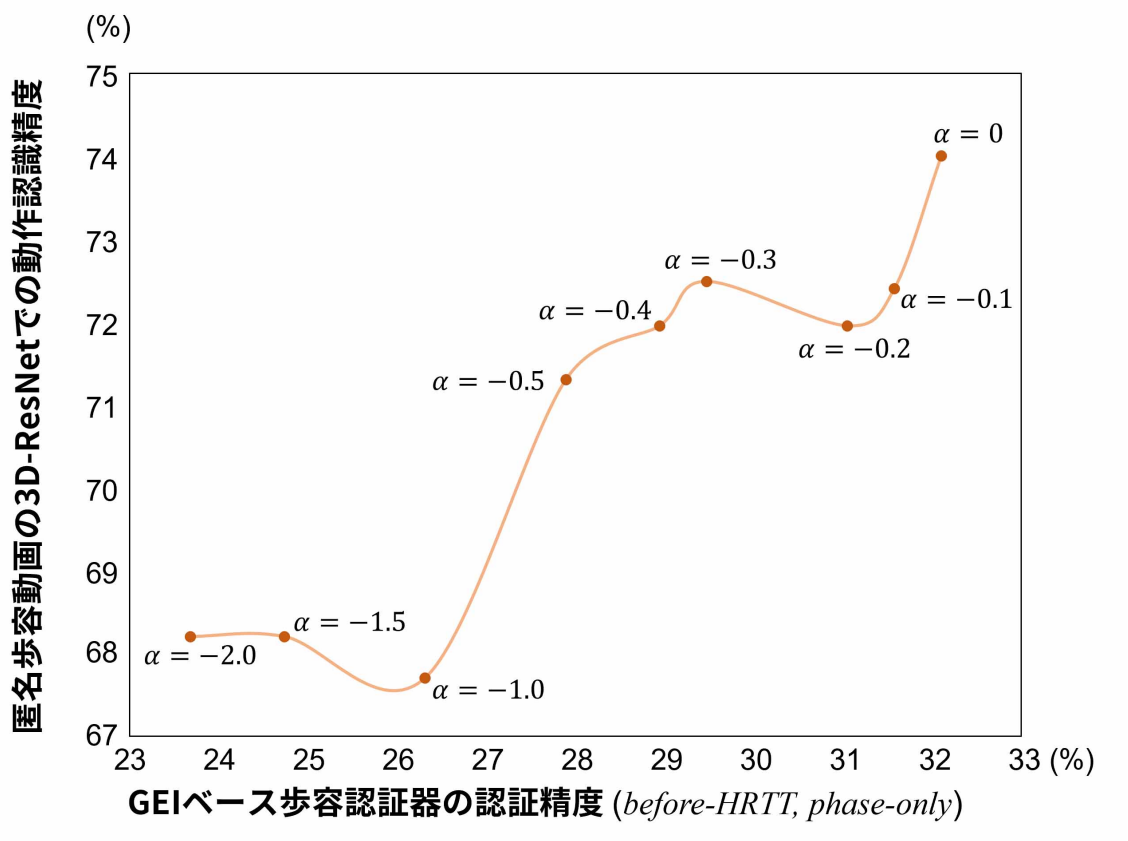
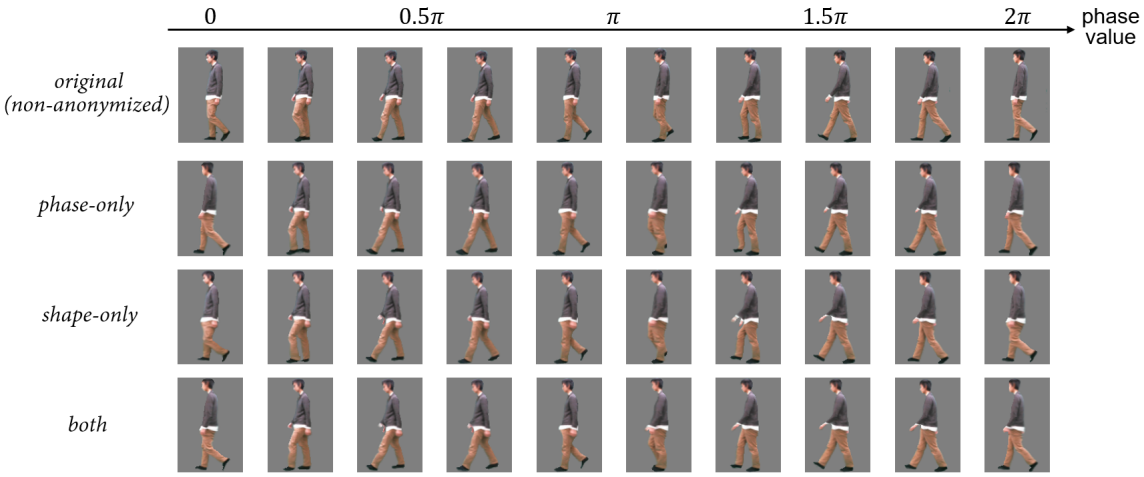


図 3.14: α が匿名化性能と歩容動画の見た目の自然さに与える影響

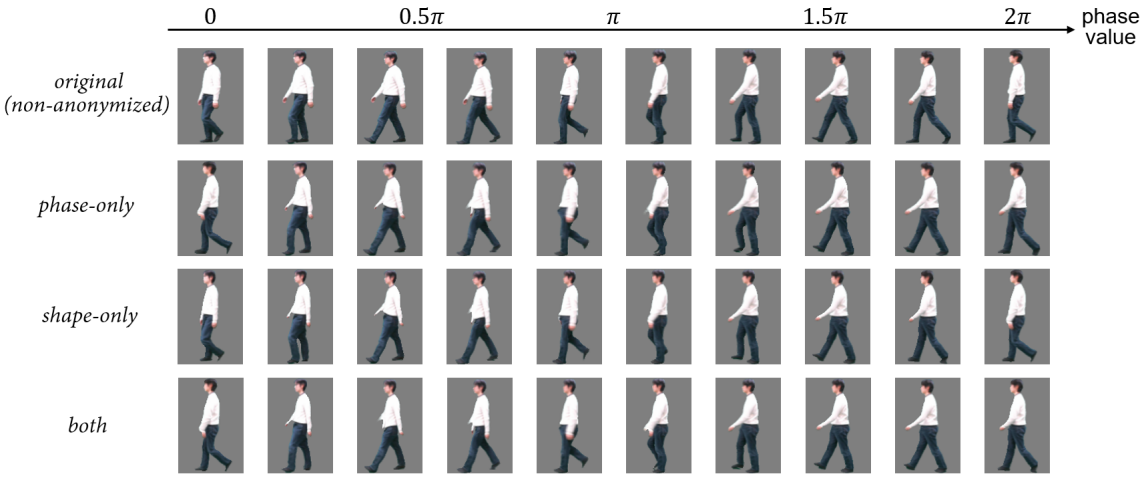
る．この結果は上記の理論的背景に一致する．つまり，α の絶対値が大きいほど匿名化性能が高く，かつ見た目の自然さが低くなるということを示している．特に，$\alpha < -0.5$ の場合，見た目の自然さが急激に低下する．これは，α の絶対値が大きい場合，HRTT を適切に実現することが難しくなるためである．よって，匿名化性能と見た目の自然さのバランスを考慮し，最終的に $\alpha = -0.5$ を採用した．

3.5.3 実験結果

前節で得られたハイパーパラメータを用いて，DS_c のテスト動画に対し匿名化を行い，それらの結果を定性的・定量的に評価した．



(a) ID:1 の人物の匿名化前歩容動画と匿名化後歩容動画



(b) ID:2 の人物の匿名化前歩容動画と匿名化後歩容動画

図 3.15: 匿名化前歩容動画と匿名化後歩容動画の例

定性的評価

図 3.15 に，匿名化前歩容動画（*non-anonymized*）と，提案手法により得られた匿名化後歩容動画の例を示す．この図から，匿名化後歩容動画は匿名化前歩容動画と同程度に自然な外観を保てていることがわかる．また，*both* の場合でも，匿名化前歩容動画と匿名化後歩容動画の間に大きな差は見られない．また，他の多

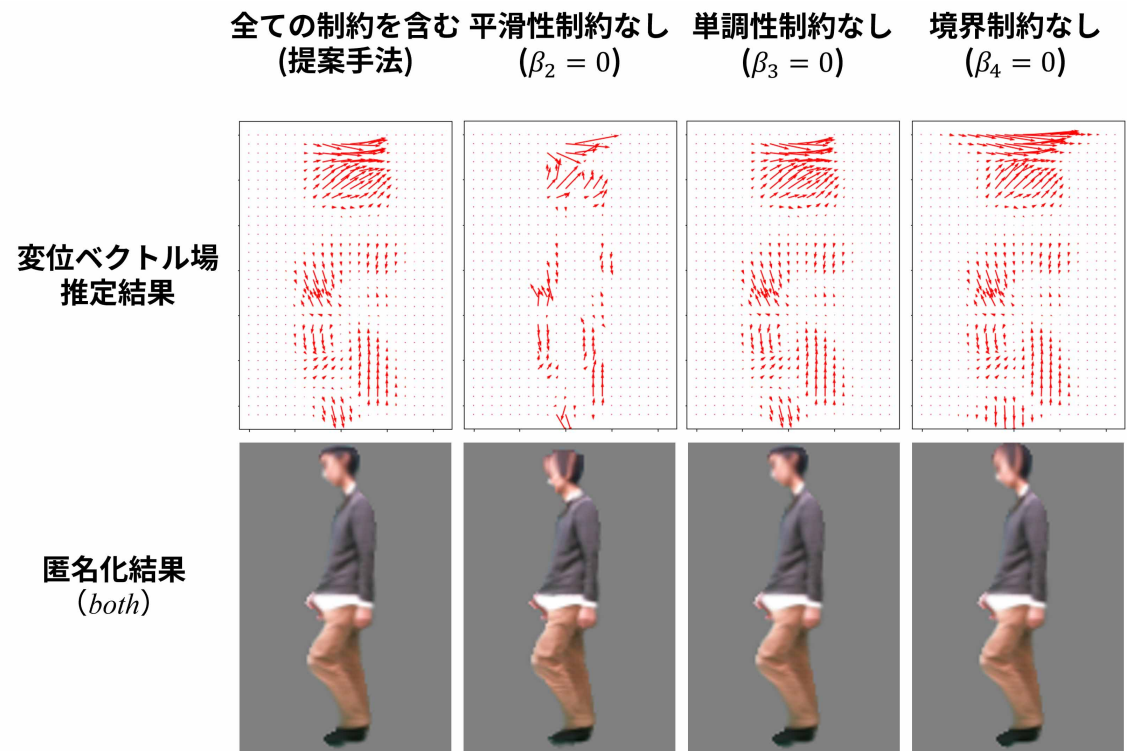


図 3.16: 3 つの制約を全て適用した場合と 3 つのうち 1 つを適用しない場合の変位ベクトル場の推定結果および最終的な匿名化結果

くのテストデータにおいても同様の結果が得られており，提案手法による匿名化では見た目の自然さが損なわれないことが確認できた．

こうした匿名化後歩容動画の見た目の自然さは HRTT の性能に影響され，その性能は 3.4.3 節で述べた 3 つの制約に依存する．本節では，それらの効果を定性的に検証した．図 3.16 に，全ての制約が存在する場合と，3 つの制約のうちそれぞれ 1 つが含まれていない場合について，変位ベクトル場の推定結果と最終的な匿名化結果の一例を示す．図 3.16 から，平滑性制約がない場合は不連続な変位ベクトル場が得られ，結果として，不自然なテクスチャの歪みが生じることが確認できる．この場合，人物領域のうち特に頭部領域が大きく歪んでおり，このことから，この制約の重要性が伺える．一方で，単調性制約がない場合は，深刻なテクスチャの歪みは生じない．これは，推定される変位ベクトル場の変位が小さい場合，平滑性制約が単調性制約を満たすことが多いためである．変位ベクトル場の推定には同位相のシルエット対を用いるため，ほとんどの場合，変位が小さい変

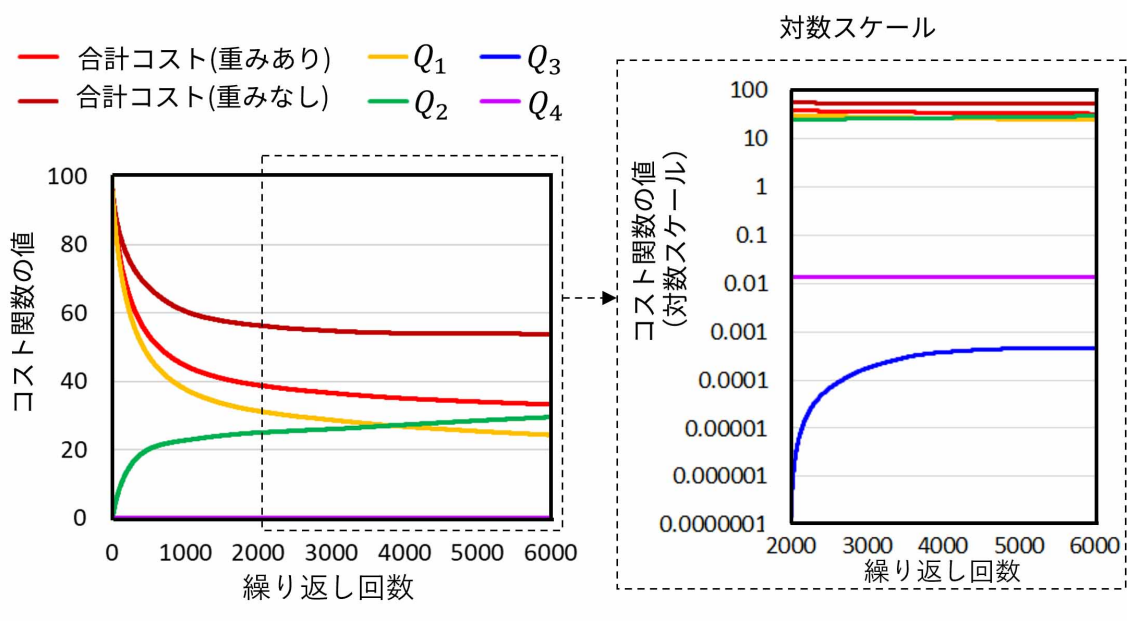


図 3.17: 変位ベクトル場の推定過程における収束傾向

位ベクトル場が得られる．このことから，単調性制約は必ずしも必要ではないことがわかる．最後に，境界制約がない場合，境界付近では比較的大きな変位ベクトルが得られ，テクスチャが不自然に伸張される可能性がある．従って，境界制約も平滑性制約と同様に重要である．図 3.17 は，式 (3.31) の各項の収束傾向を示したもので，Q_3 が他の項に比べて非常に小さくなっている．このことからも，単調性制約に比べ，平滑性制約と境界制約が重要であることがわかる．

匿名化性能に関する定量的評価

表 3.1 は，GEINet ベースの歩容認証器に，匿名化前歩容動画と匿名化後歩容動画を入力したときの歩容認証精度を示している．また，比較のため，提案手法に加えて，モザイク処理 (*pixelization*) およびぼかし処理 (*blurring*) といった視覚的抽象化に基づく 2 つの比較手法も検証した．さらに，GEI を特徴として使用し，かつ，*before-HRTT* の場合の各人物の個別の歩容認証精度を図 3.18 に示す．

Method		歩容認証精度		
		GEI	FDF	SFDEI
non-anonymized		100%	100%	100%
before-HRTT	*phase-only*	27.8%	21.0%	23.7%
	shape-only	4.73%	5.78%	8.94%
	both	**2.10%**	**1.57%**	**3.15%**
after-HRTT	*phase-only*	33.6%	27.3%	25.2%
	shape-only	8.94%	11.5%	12.1%
	both	**8.42%**	**8.42%**	**7.36%**
pixelization		79.4%	81.0%	81.0%
blurring		28.9%	23.1%	27.3%

表 3.1: GEINet ベースの歩容認証器による認証精度

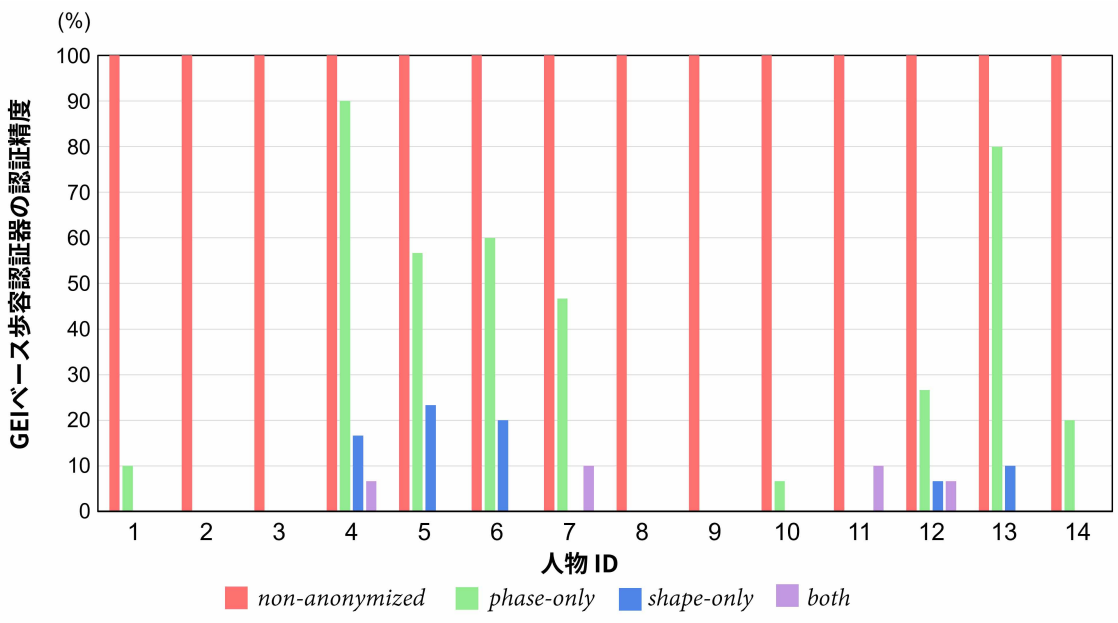


図 3.18: GEI 特徴かつ *before-HRTT* の条件下での各人物ごとの歩容認証精度

まず，匿名化処理を行わない場合についてはすべての歩容特徴量（GEI，FDF，SFDEI）で 100%の認識精度が得られた．よって，歩容認証器によって個人の同定が行えることから，匿名化処理が必要であることが確認できる．これに対し，提案手法を適用した場合には歩容認証精度が大きく低下しており，提案手法には高い匿名化能力があることがわかる．また，*shape-only* や *both* の場合と比較して，

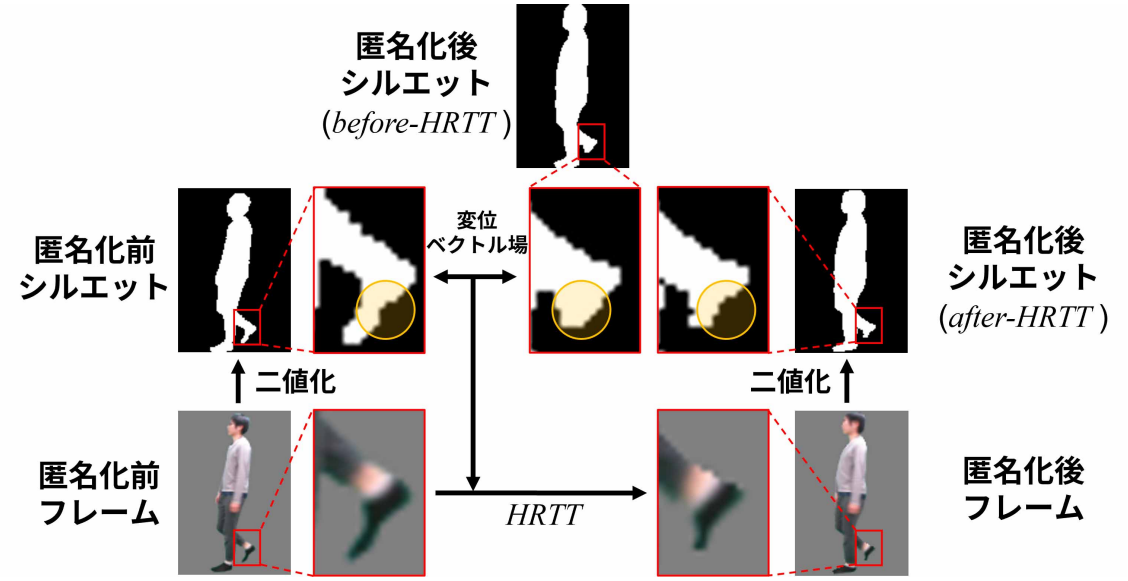


図 3.19: *before-HRTT* と *after-HRTT* の局所形状の比較

phase-only の場合は匿名化性能が限定的である．このことから，人物の歩容を認証する際には，動的特徴よりも静的特徴による影響が大きいということがわかった．上記の理由から，動的特徴を失った GEI でも，匿名化前歩容動画では 100% の精度を達成できている．しかし，ID:1 から ID:14 のほぼ全ての人物，特に FDF と SFDEI の場合，*both* は *shape-only* よりも歩容認証精度を下げることができるので，位相値の摂動も同様に有効であることがわかる．この 2 つの特徴は GEI とは異なり，人物の歩容の動的な側面を扱うことができる．こうした位相値の摂動は，動的な側面を匿名化できるため，FDF や SFDEI のような動的な成分が含まれる特徴量を使用した歩容認証器では特に効果的である．

また，*before-HRTT* と *after-HRTT* のシルエットを比較した場合，後者の方が匿名化性能は低くなる．これは，HRTT が匿名化性能を低下させ得ることを示している．この匿名化性能の低下は 3.4.3 節で述べた平滑性制約が原因となっている可能性がある．平滑性制約により，推定された変位ベクトル場の高周波成分が弱まり，図 3.19 に示すように，匿名化前歩容動画の人物領域の局所的な形状情報が匿名化後歩容動画に反映される可能性がある．このような平滑性制約の影響はあまり大きくはない．しかし，Web 動画からプライバシー情報を詐取しようとする攻撃者は，匿名化した人物の歩容動画を再度二値化したもの（*after-HRTT*）を歩容

Method	YOLO accuracy	3D-ResNet accuracy
non-anonymized	**100%**	**75.6%**
phase-only	**99.9%**	**70.3%**
shape-only	**100%**	**73.0%**
both	**100%**	**73.0%**
pixelization	0.00%	29.6%
blurring	25.7%	23.2%

表 3.2: YOLO による人物検出精度と 3D-ResNet による「歩行」動作認識精度

認証器に入力できるため，*after-HRTT* の場合に匿名化効果が低減するといった傾向は望ましくない．しかし，今後，変位ベクトル場の推定処理において，平滑性制約を適切に設定することができれば，このような問題の解消につながる可能性も十分に考えられる．

また，*pixelization* や *blurring* といった視覚的抽象化は，本手法と比べて歩容認証精度を大きく低下させない．特に，*pixelization* の場合は，80%程度の精度となっており，匿名化効果が低い．これは，GEINet のような CNN ベースの歩容認証器は，内部で *pixelization* に近い処理（平均プーリングなど）を行っているため，*pixelization* を適用した画像に対して頑健であるためと考えられる．

見た目の自然さに関する定量的な評価

3.5.1 節で述べたように，匿名化後歩容動画の見た目の自然さは，YOLO によるフレームごとの人物検出精度と 3D-ResNet による動画ごとの「歩行」に分類される動作認識精度の 2 つの基準で評価した．これらの結果を表 3.2 に示す．また，図 3.20 に各人物ごとの 3D-ResNet の精度を示す．匿名化前動画に対して，YOLO は 100%の検出精度であるのに対し，3D-ResNet は 75.6%の認識精度にとどまった．これは，動画からの動作認識が，画像からの人物検出よりも困難なタスクであるからである．ここでは，動作認識精度が匿名化処理によってどの程度低下するかに着目している．もし，精度が低下しなければ，匿名化後歩容動画は見た目の自然さを保てているということになる．

phase-only，*shape-only*，*both* の 3 つの場合において上述の評価を行った．上記全ての場合で，YOLO による人物領域の検出精度は低下せず，一方で，3D-ResNet の動作認識精度はわずかに劣化する程度であることが確認できた．これらの結果か

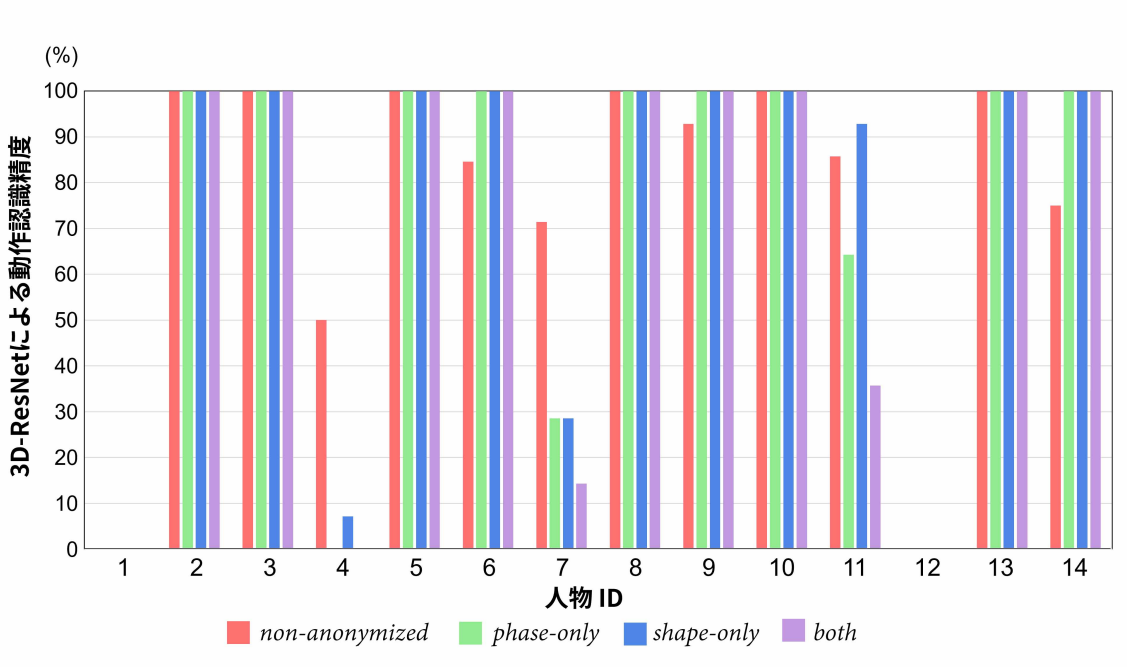


図 3.20: 3D-ResNet による各人物ごとの「歩行」動作の認識精度

ら，提案手法では特に *shape-only* と *both* の場合，見た目の自然さを保つことができたといえる．また，ID:1，ID:4，ID:7，ID:12 の人物の匿名化後歩容動画は，3D-ResNet ではほとんど「歩行」動作として認識されなかったが，これは匿名化前歩容動画でも同様となっている．

また，*pixelization* や *blurring* では YOLO の精度と 3D-ResNet の精度を維持することができない．これは，2.3 節で述べたように，視覚的抽象化では見た目の自然さを全く保てないことを示している．

3.6 結言

本章では，シルエットベース歩容認証により個人が同定され，それに紐づくプライバシー情報を取得されてしまうプライバシー情報詐取攻撃に対する防御法として，歩容動画の匿名化手法を提案した．歩容動画の匿名化は位相成分と形状成分の摂動による歩容シルエットの変形と変位ベクトル場によるテクスチャ情報の転写によって達成された．

実験結果から，匿名化後歩容動画では匿名化前歩容動画と比べて，歩容認証精

度を 100%から 1.57%まで大きく低下させることができた．しかし，平滑性制約の影響で *after-HRTT* の場合には *before-HRTT* と比べて匿名化精度が低くなってしまうため適切な平滑性制約の設定が重要である．また，見た目の自然さについては 3D-ResNet による動作認識精度が匿名化前から匿名化後で 75.6%から 73.0%とほとんど変化しないことから，見た目の自然さが保てていることがわかった．これらのことから，提案した歩容動画の匿名化手法は Web 動画中の歩容情報を匿名化する場合に有用な手段となることが確認できた．

第4章 単一歩容画像からのなりすまし偽歩容動画生成に関する実現可能性の検証

4.1 緒言

マルチメディアデータ生成技術により特定個人の偽歩容情報を作成して当該個人になりすまし，それにより偽情報を拡散する攻撃が行われるといった問題がある．本章では，このなりすまし攻撃の実現可能性について考察する．ここで，このなりすまし攻撃が対象とする歩容認証器では，歩容の持つシルエットの情報のみを用いて認証を行うと想定される．そこで，歩容シルエット動画の段階でなりすまし対象の個人の特徴を保持していれば，その歩容に付与されている色情報に関わらず，なりすました該当個人の歩容であると歩容認証器に誤認識させることができる．この考えに基づき，本章では偽歩容シルエット動画の生成手法について議論する．まず，4.2 節にて，上記のなりすまし攻撃が行われる具体的なシナリオについて述べる．次に，4.3 節で，提案手法に関連するウルフ攻撃についての研究を概観する．続いて，4.4 節で一枚の歩容画像から偽歩容シルエット動画を生成する手法について述べる．その後，4.5 節で本手法の有効性を実験的に評価し，最後に，4.6 節で本章の内容をまとめる．

4.2 なりすまし攻撃が行われる状況

本節では，実際に偽歩容動画を用いてなりすまし攻撃が行われる状況について詳述する．図 4.1 に歩容認証に対するなりすまし攻撃の想定シナリオの例を示す．まず，人物 A と B が対立しているとする．次に，A はこのシナリオにおける攻撃者であり，B の偽の動画を作成し，その評判を落とすことを試みるといった目的がある．このような中で，なりすまし攻撃は「偽歩容動画の生成」と「偽情報の拡散」の二つの過程に分けられる．攻撃者 A が行う，偽歩容動画の生成の具体的な手順は以下の通りとなる．

(1) ある社会通念上不適切とされる場所の動画を自身の端末で撮影する．

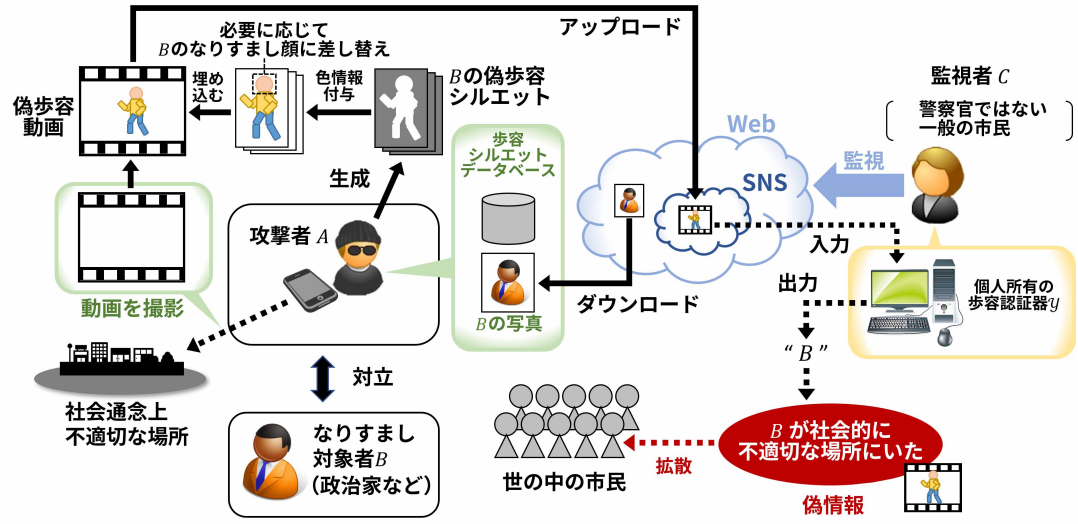


図 4.1: 偽歩容動画を用いたなりすまし攻撃の想定シナリオ

(2) それと並行して，B の歩き方を模倣または偽装した偽歩容シルエット系列を生成する．

(3) 生成したシルエットに対して色情報を付与する．

(4) 上記で撮影した社会通念上不適切とされる場所の動画に色付き偽歩容を埋め込むことで B の偽歩容動画を生成する．

手順 (4) では必要に応じて B の偽の顔を生成し，偽歩容動画に挿入することも可能である．これにより，偽歩容動画のリアリティが高まるが，動画の解像度が低い等，偽の顔の挿入が困難な場合は必ずしも必要ない．なお，人間の目は体の色情報で人を識別できないことや，歩容認証器はシルエット情報のみを利用することから，こうした色付け処理自体はあまり重要ではない．

ところで，昨今，Web 上の著名人などの社会的行動を監視している IT に詳しい人物が多々存在している．こうした人々のうちの 1 人を監視者 C とする．これを踏まえて，偽情報の拡散の具体的な手順は以下の通りとなる．

(1) 攻撃者 A は偽歩容動画を Web 上，その中でも特に SNS にアップロードする．

(2) 監視者 C は SNS をチェックし，自分の所有する歩容認証器 $\mathscr{y}$ に偽歩容動画を入力する．

(3) $\mathscr{Y}$ は偽歩容動画中の人物を B と判定する．

(4) 上記の結果を監視者 C が真正な情報であると考え，悪意なく配信し，その結果，偽情報の拡散が行われる．

このような「偽歩容動画の生成」と「偽情報の拡散」の二つの過程により，なりすまし攻撃が成立する．こうした偽歩容動画が拡散される前に偽物であると検出されることが望ましいが，最近のディープフェイク検出器の多くが顔のみに着目しているため，これを通過してしまう可能性がある．つまり，偽の顔で作られた偽情報は偽の歩容と組み合わせることで，より検出が難しくなる．

ここで，上記のシナリオの中では，攻撃者 A はなりすまし対象者 B の写真 1 枚と，B とも $\mathscr{Y}$ とも無関係な歩容シルエットの大規模データベースを利用できるといった条件を設定する．これは，攻撃のために必要な真正情報を写真一枚とすることによって，政治家やタレントなどの有名人だけでなく，Web や SNS に顔写真が掲載されている一般市民もなりすまし対象者となり，なりすまし攻撃の汎用性がより向上するからである．さらに，上記のなりすまし攻撃のシナリオにおける攻撃者の目的は未知の歩容認証器 $\mathscr{Y}$ がなりすまし対象人物として認識できるような偽歩容シルエット系列を生成することである．よって，本研究の目的も同様に，このような偽歩容シルエット系列，すなわち，偽歩容シルエット動画の生成となる．

4.3 マスター生体情報サンプルによるウルフ攻撃

本章ではなりすまし攻撃についての議論を行うが，このなりすまし攻撃の目的は，一般的にターゲットとなる 1 人の人物に類似し，それ以外の人物には類似しない偽の生体情報サンプル（顔など）を作成することである．しかし，中には 1 人の人物にのみ類似するのではなく，2 人以上の人物に類似した偽のサンプルを 1 つ作るといった場合もあり，生体認証の研究において，これの作成が可能であることが知られている．この偽のサンプルは「ウルフ」と呼ばれ，このウルフを用いた生体認証器への攻撃は「ウルフ攻撃」[54] と呼ばれる．

ここで，このウルフが悪用されてしまう状況について考える．まず，生体認証器を保有する人が 2 人以上いると仮定する．この時，各生体認証器は入力された生体情報サンプルがその所有者のものかどうかを予測する 2 クラス分類器とする．

次に，生体認証器の認証を突破することを目的とした攻撃者は 1 つのウルフを使うが，この時，ウルフの特性により，これらの生体認証器の多くを同時に騙すことができる．このことは，ウルフがマスターキーの役割を持つことを意味している．

このような問題は深刻であるため，現在では，ウルフ攻撃の手法やその対策が研究されている．例えば，Ohki らは音声認証器に対するウルフ攻撃の実行可能性を評価している [55]．また，Nguyen らは顔認証器に対するウルフを GAN により生成する手法を提案している [56]．この手法で生成されたウルフは「マスター顔」と呼ばれている．一方で，歩容認証器に対するウルフ攻撃に着目した研究は未だ行われていないが，ウルフの特徴は偽歩容動画を用いたなりすまし攻撃を行う上で有用であると考えられる．そこで，提案手法では「マスター歩容」という概念を導入し，その特徴について 4.4.3 節で述べる．

4.4 偽歩容シルエット動画の生成

4.4.1 偽歩容シルエット動画の生成手法の概要

第 3 章で述べたように，一枚の歩容シルエット s の形状は衣服の形状を含む体型と姿勢の二つの要素によって決定づけられる．歩行は周期的な動作であるため，人間の歩行一周期の姿勢は位相値 $\sigma \in [0, 2\pi]$ で表現することができる．また，人間の体型は一つの動画の中で変化しない特徴となっており一意に定まる．この特徴は，d' を次元数として形状コード $\boldsymbol{\gamma} \in \mathbb{R}^{d'}$ で表すことができる．これらのことから，歩容シルエット s は $\boldsymbol{\gamma}$ と σ によって決定することができる．一方で，2.2 節で述べたようにこうした歩容シルエットを用いた歩容認証は，一般的に，シルエットを一枚の画像に集約したのちに認証するという手法により行われる．この集約された画像を本章では特徴マップ f と呼ぶ．また，シルエットから特徴マップを抽出するものを特徴マップ抽出器 F とする．特徴マップの典型的な例としては 2.2 節で述べた，GEI，FDF などがあるが，この種類の数は限られている．また，歩容認証の認証結果は，入力した特徴マップが歩容認証器に含まれる人物それぞれに対してどれほどその人物である確率が高いかをスコアにし，それをまとめたスコアベクトルで出力される．

上記を踏まえたうえで，なりすまし偽歩容の生成処理は次のように定式化できる．デコーダ D_{sil} によって生成される偽のシルエット列を $S'(\boldsymbol{\gamma}) = \{D_{\mathrm{sil}}(\boldsymbol{\gamma}, \sigma_i) | i = 1, \cdots, N_s\}$ とし，$S'(\boldsymbol{\gamma})$ の特徴マップを $f(\boldsymbol{\gamma}) = F(S'(\boldsymbol{\gamma}))$ とする．また，$f(\boldsymbol{\gamma})$ を

監視者 C の歩容認証器 $\mathscr{Y}$ に入力した時のスコア（B らしさ）を $\zeta_b(f(\boldsymbol{\gamma}))$ とする．攻撃者の目標は $\zeta_b(f(\boldsymbol{\gamma}))$ を最大化する $\boldsymbol{\gamma}^*$ を求めることであり，求めた $\boldsymbol{\gamma}^*$ から生成される $S(\boldsymbol{\gamma}^*)$ が攻撃結果の偽歩容シルエットである．この時，位相列 $\boldsymbol{\sigma} = \{\sigma_1, \cdots, \sigma_{N_s}\}$ は任意に与えることが可能である．この際，攻撃者 A はネットワークの構造について既知ではないが，上記で述べたように歩容認証に用いられる特徴マップは数種類しかないため，F を推測することが可能である．本研究では，特徴マップ抽出器 F は FDF を抽出できる抽出器とする．

4.2 節で想定したように，攻撃者は $\boldsymbol{\gamma}^*$ を取得する際に，対象人物 B の一枚の写真のみを用いることができる．この写真から抽出された歩容シルエットを p とする．ここで，$\boldsymbol{\gamma}^*$ を最も簡単に見つける方法は，位相に非依存な形状コードを抽出可能なエンコーダ E_{sil}，すなわち $\boldsymbol{\gamma} = E_{\text{sil}}(D_{\text{sil}}(\boldsymbol{\gamma}, \sigma))$ を満たすようなエンコーダ E_{sil} を別途学習し，その E_{sil} により $\boldsymbol{\gamma}^* = E_{\text{sil}}(p)$ として $\boldsymbol{\gamma}^*$ を抽出することである．しかし，一枚の歩容シルエット p では B の歩容特徴を完全には再現できず（例えば動的な特徴を単一の画像から再現することは困難である），一部の情報の欠落した特徴しか得られないため，この方法では実際には最適な $\boldsymbol{\gamma}^*$ を取得することはできない．よって，歩容シルエット p から抽出された $\tilde{\boldsymbol{\gamma}} = E_{\text{sil}}(p)$ と最適な $\boldsymbol{\gamma}^*$ の間には一定の抽出誤差が存在するということがわかる．これは $\tilde{\boldsymbol{\gamma}}$ が $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^* + \Delta\boldsymbol{\gamma}$ となることに相当する．この $\Delta\boldsymbol{\gamma}$ をなりすまし偽歩容動画を生成する際には推定する必要がある．

なりすまし偽歩容を生成するための具体的な手順は以下の通りである．

(1) 攻撃者はまず，自身の歩容シルエットデータベースを使用し，D_{sil} と E_{sil} を学習する．

(2) 学習済みの E_{sil} と対象人物 B の一枚の写真を用いて，$\tilde{\boldsymbol{\gamma}} = E_{\text{sil}}(p)$ を得る．

(3) 次に，$\Delta\boldsymbol{\gamma}$ を推定することにより，$\tilde{\boldsymbol{\gamma}}$ を $\boldsymbol{\gamma}^* = \tilde{\boldsymbol{\gamma}} - \Delta\boldsymbol{\gamma}$ へと最適化する．

(4) 最後に，学習された D_{sil} と任意に与えられた $\boldsymbol{\sigma} = \{\sigma_1, \cdots, \sigma_{N_s}\}$ から，偽歩容シルエット列 $\{D(\boldsymbol{\gamma}^*, \sigma_i) | i = 1, \cdots, N_s\}$ を得る．

これらの手順により，最終的に B になりすました偽歩容シルエット動画を生成することができる．

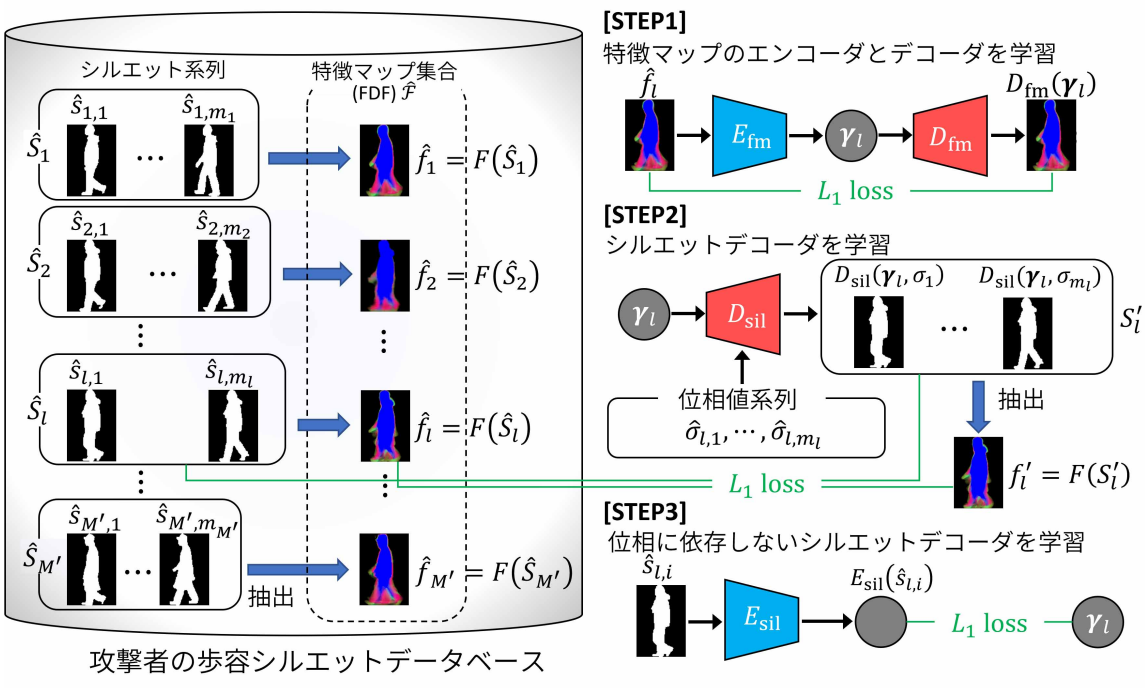


図 4.2: 歩容シルエットデコーダ D_{sil} とエンコーダ E_{sil} の学習過程

4.4.2 偽歩容シルエットエンコーダとデコーダの学習過程

図 4.2 は提案する D_{sil} と E_{sil} の学習手順を示したものであり，3 つのステップから構成されている．ここで，E_{sil} は入力された歩容シルエットの位相の如何に関わらず，同一の特徴（＝形状コード）を出力する必要がある．しかし，位相に依存しないシルエットエンコーダ E_{sil} を直接学習するためには，そのための学習データとして「形状コードの正解値」を有するシルエットが多数必要になる．しかし，形状コードには本来正解値は存在しない（3.3.3 節の $\boldsymbol{z}_a$ は便宜的に「正解」とみなした値に過ぎない）．このため，E_{sil} の直接学習は困難である．そこで，まず，STEP1 として特徴マップレベルのエンコーダ E_{fm} とデコーダ D_{fm} を学習する．そのために，攻撃者の所有するデータベースの各シルエット列 $\hat{S}_l = \{\hat{s}_{l,1}, \cdots, \hat{s}_{l,m_l}\}$（$m_l$ は l 番目の歩容シルエット動画のフレーム数）をそれぞれ F で圧縮し，特徴マップ $\hat{f}_l = F(\hat{S}_l)$ を用意する．こうして，得られた特徴マップの集合を $\hat{\mathcal{F}} = \{\hat{f}_l | l = 1, \cdots, M'\}$（$M'$ は特徴マップの個数）とし，$\hat{\mathcal{F}}$ を用いて AE を学習する．この AE

のエンコーダ部分を E_{fm}，デコーダ部分を D_{fm} とすると，STEP1 の損失関数は下記のように定義される．

$$\mathrm{Loss}_1=\sum_l \left\| \hat{f}_l - D_{\mathrm{fm}}(\boldsymbol{\gamma}_l) \right\| =\sum_l \left\| \hat{f}_l - D_{\mathrm{fm}}(E_{\mathrm{fm}}(\hat{f}_l)) \right\| \tag{4.1}$$

この特徴マップエンコーダ E_{fm} によって $\hat{f}_l$ から抽出された $\boldsymbol{\gamma}_l = E_{\mathrm{fm}}(\hat{f}_l)$ は位相に依存しない．そのため，全ての $i \in \{1, \cdots, m_l\}$ に対して，歩容シルエット $\hat{s}_{l,i}$ の位相に依存しない形状ベクトル，すなわち形状コードとして使用することができる．それらを用いて，STEP2 ではシルエット生成デコーダ D_{sil} を学習する．STEP2 の損失関数は下記のように定義される．

$$\mathrm{Loss}_2 = \sum_l \left\{ \left\| \hat{f}_l - F(S'_l) \right\| + \frac{1}{m_l} \sum_{i=1}^{m_l} \| \hat{s}_{l,i} - D_{\mathrm{sil}}(\boldsymbol{\gamma}_l, \hat{\sigma}_{l,i}) \| \right\} \tag{4.2}$$

ここで，S'_l は $S'_l = \{D(\boldsymbol{\gamma}_l, \hat{\sigma}_{l,i}) | i = 1, \cdots, m_l\}$ となる．それぞれの画像 $\hat{s}_{l,i}$ に対する位相値 $\hat{\sigma}_{l,i}$ は 3.3.2 節で述べた手法により計算する．最後に，STEP3 として位相に依存しないシルエットエンコーダ E_{sil} を学習するが，その損失関数は以下の通りである．

$$\mathrm{Loss}_3 \ = \ \sum_l \left\| E_{\mathrm{sil}}(\hat{s}_{l,i}) - \boldsymbol{\gamma}_l \right\| \tag{4.3}$$

これらの手順により，位相に依存しないシルエットエンコーダ E_{sil} とシルエット生成デコーダ D_{sil} が得られる．

4.4.3 マスター歩容を用いた歩容形状ベクトルの最適化

4.4.1 節で述べたように，エンコーダ E_{sil} によって得られた形状ベクトル $\tilde{\boldsymbol{\gamma}} = E_{\mathrm{sil}}(p)$ には抽出誤差 $\Delta\boldsymbol{\gamma}$ が含まれる．この誤差により，$\tilde{\boldsymbol{\gamma}}$ はなりすまし対象 B の特徴を十分に保てていない．従って，なりすまし攻撃の成功率を向上させるために特徴を強調する必要がある．

下記では，特徴の強調に際して必要となる歩容による本人認証器（以後，本人認証器とする）とマスター歩容について詳述する．まず，本人認証器は 1 人の人物に対して入力シルエットから特徴マップ f を抽出し，入力シルエットから得られた特徴マップ f が本人のものであるかどうかを分類する 2 クラス分類器と定義する．この本人認証器は特徴マップ f を入力した際にスコア $\omega'(f) \in [0, 1]$ を出力する．このスコアが $\omega'(f) \geq 0.5$ のときの入力シルエットは「本人である」と分類

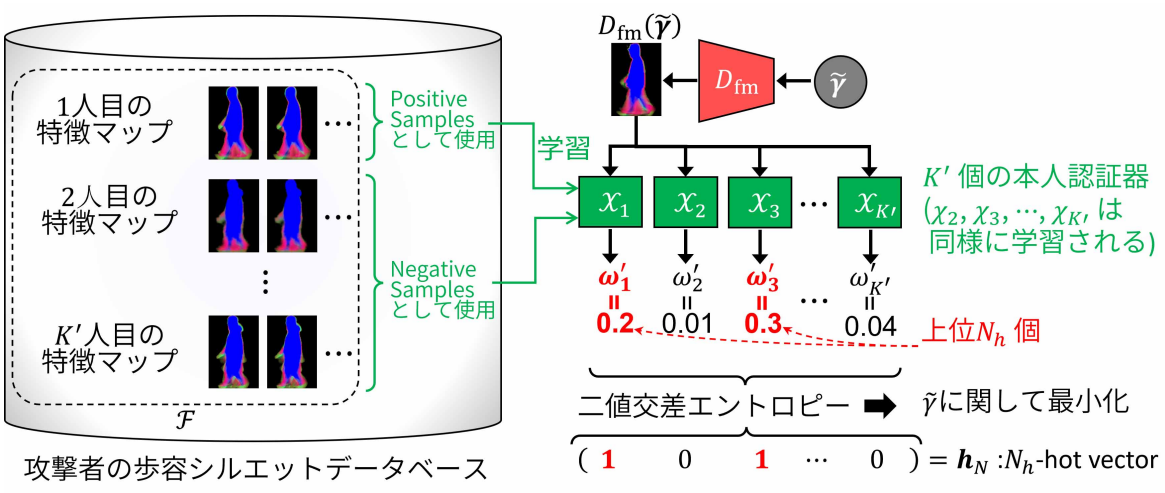


図 4.3: 個人性強調による $\tilde{\gamma}$ の更新手順

される．このような本人認証器を多数の人物について用意した場合に，2 つ以上の本人認証器から 0.5 以上のスコアが得られるような特徴マップを 4.3 節で述べたマスター歩容と定義する．

ここで，攻撃者が所有する自身のデータベースの各個人についてのそれぞれの本人認証器を構築した場合を考える（図4.3参照）．j 番目の個人の認証器を $\mathcal{X}_j$ $(j = 1, \cdots, K')$，$\mathcal{X}_j$ のスコアを ω'_j，K' を攻撃者のデータベースに登録されている人数とする．攻撃者は全ての j について $\mathcal{X}_j$ に特徴マップ $\tilde{f} = D_{\mathrm{fm}}(\tilde{\gamma})$ を入力することで，スコア集合 $\{\omega'_j(D_{\mathrm{fm}}(\tilde{\gamma}))|j = 1, \cdots, K'\}$ を得ることができる．なりすまし対象 B は攻撃者のデータベースに存在しない個人なので，スコアはすべて 0.5 未満になる．しかし，データベースが十分に大きい場合，B とある程度似た特徴を持つ人物がデータベース内に存在するため，スコア集合のいくつかの要素は他の要素に比べて相対的に大きくなる．これは，対象者 B の特徴を表していると考えることができる．よって，提案手法では，相対的に大きい要素がさらに大きくなり，他の要素が小さくなるように $\tilde{\gamma}$ を摂動させることで特徴を強調する．この時，2 つ以上の要素で $\omega'_j(D_{\mathrm{fm}}(\boldsymbol{\gamma}^*)) > 0.5$ を満たすものを $\boldsymbol{\gamma}^*$ として使用する．これは $D_{\mathrm{fm}}(\boldsymbol{\gamma}^*)$ がマスター歩容に相当することを意味する．以降では，上記の処理を $\tilde{\gamma}$ に対する「個人性強調」と呼ぶ．

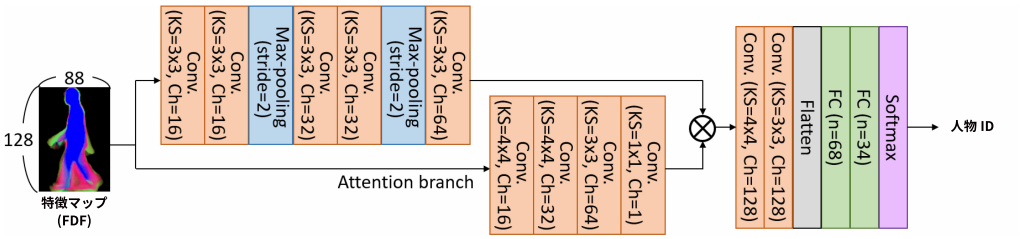


図 4.4: 歩容認証器 $\mathscr{Y}$ のネットワーク構造図（Conv.; 畳み込み層，KS; カーネルサイズ，Ch; チャネル数，FC; 全結合層（n はユニット数），$\otimes$; 画素単位の乗算）

具体的な個人性強調の過程は以下の通りである．まず，攻撃者は自分のデータベースを用いて $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{K'}$ を学習させる．次に，各 $\mathcal{X}$ に対して $\tilde{\boldsymbol{\gamma}} = E_{\mathrm{sil}}(p)$ を入力し，下記のスコアベクト集合 $\boldsymbol{\omega}'(\tilde{\boldsymbol{\gamma}})$ を得る．

$$\boldsymbol{\omega}'(\tilde{\boldsymbol{\gamma}}) = \begin{pmatrix} \omega'_1(D_{\mathrm{fm}}(\tilde{\boldsymbol{\gamma}})) \\ \vdots \\ \omega'_{K'}(D_{\mathrm{fm}}(\tilde{\boldsymbol{\gamma}})) \end{pmatrix} \in [0,1]^{K'} \tag{4.4}$$

そして，$\boldsymbol{\omega}'(\tilde{\boldsymbol{\gamma}})$ の中から上位 N_h 個の大きな要素を見つけ，N_h-hot vector を作成する．これは $\boldsymbol{h}_{N_h} = (h_{N_h,1} \ \cdots \ h_{N_h,K'})^\top \in \{0,1\}^{K'}$ と表現される．ここで，$\boldsymbol{h}_{N_h}$ の各要素は，$\boldsymbol{\omega}'(\tilde{\boldsymbol{\gamma}})$ 中の対応する要素が上位 N_h 個の要素に含まれる場合のみ 1 とし，残りの要素は 0 とする．その後，$\boldsymbol{\omega}'(\tilde{\boldsymbol{\gamma}})$ と $\boldsymbol{h}_{N_h}$ の二値交差エントロピーを下記のように計算し，$\tilde{\boldsymbol{\gamma}}$ に関して最小化することで，最適な $\boldsymbol{\gamma}^*$ を求める．

$$-\sum_{j=1}^{K'} \left[h_{N_h,j} \log\{\omega'_j(D_{\mathrm{fm}}(\tilde{\boldsymbol{\gamma}}))\} + (1 - h_{N_h,j}) \log\{1 - \omega'_j(D_{\mathrm{fm}}(\tilde{\boldsymbol{\gamma}}))\} \right] \tag{4.5}$$

この最小化処理は勾配降下法によって行われる．この処理は $\Delta\boldsymbol{\gamma}$ を $\Delta\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*$ と推定することに相当する．

最終的に，得られた $\boldsymbol{\gamma}^*$ とデコーダ D_{sil} と $\boldsymbol{\sigma} = \{\sigma_1, \cdots, \sigma_{N_s}\}$ から偽歩容シルエット列 $\{D_{\mathrm{sil}}(\boldsymbol{\gamma}^*, \sigma_i) | i = 1, \cdots, N_s\}$ を生成することができる．

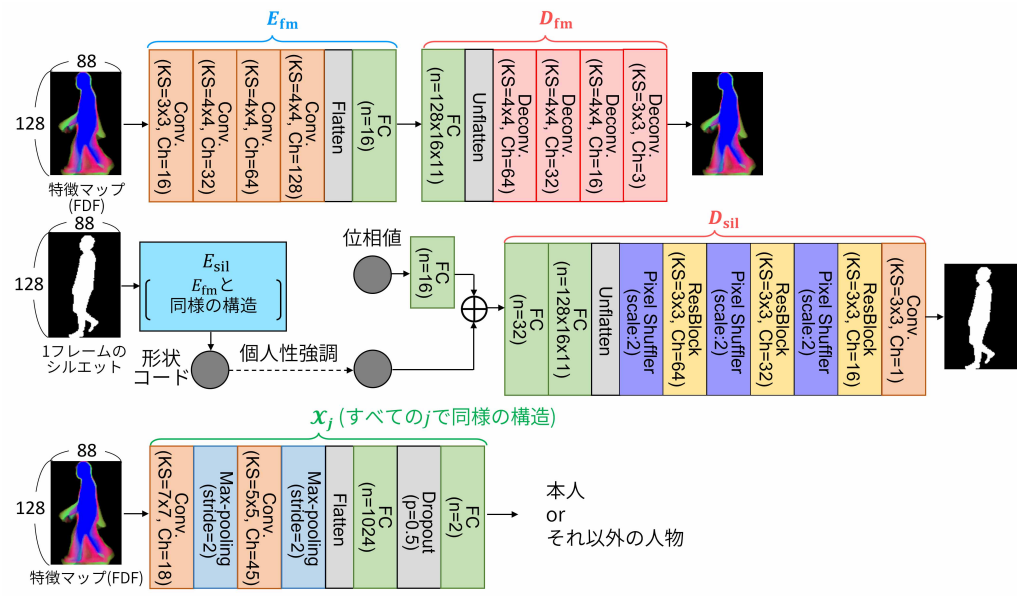


図 4.5: E_{sil}，D_{sil}，E_{fm}，D_{fm}，$\mathcal{X}_j$ のネットワーク構造図（Deconv.; 逆畳み込み層，$\oplus$; 連結演算子）

4.5 評価実験

4.5.1 実験設定

提案手法の性能を検証するために，OU-ISIR Gait Database [47] をデータセットとして用いた．このデータセットには複数のサブセットが有り，そのうちの Treadmill-dataset-(A) と Treadmill-dataset-(B) の 2 つを使用した．Treadmill-dataset-(A) は 34 人の 612 個の歩容シルエット列（一人あたり 18 個），Treadmill-dataset-(B) は 68 人の 2176 個の歩容シルエット列（一人あたり 32 列）から構成される．本実験では，Treadmill-dataset-(A) を用いて調査を行う人物 C の歩容認証器 $\mathscr{Y}$ を構築するとともに，Treadmill-dataset-(B) を攻撃者 A のデータベースとして扱った．$\mathscr{Y}$ は図 4.4 に示すようなネットワーク構造を持つ DNN として学習させた．$\mathscr{Y}$ の学習後，Treadmill-dataset-(A) の各列から 1 フレームを選択する．それをなりすまし対象 B の写真として，その写真から偽の歩容シルエット列を生成して $\mathscr{Y}$ に与え，正しく認識されるかどうかを確認した．この処理を Treadmill-dataset-(A) の全てのフレームについて繰り返し，最終的に認識精度を評価した．認識精度の高さは

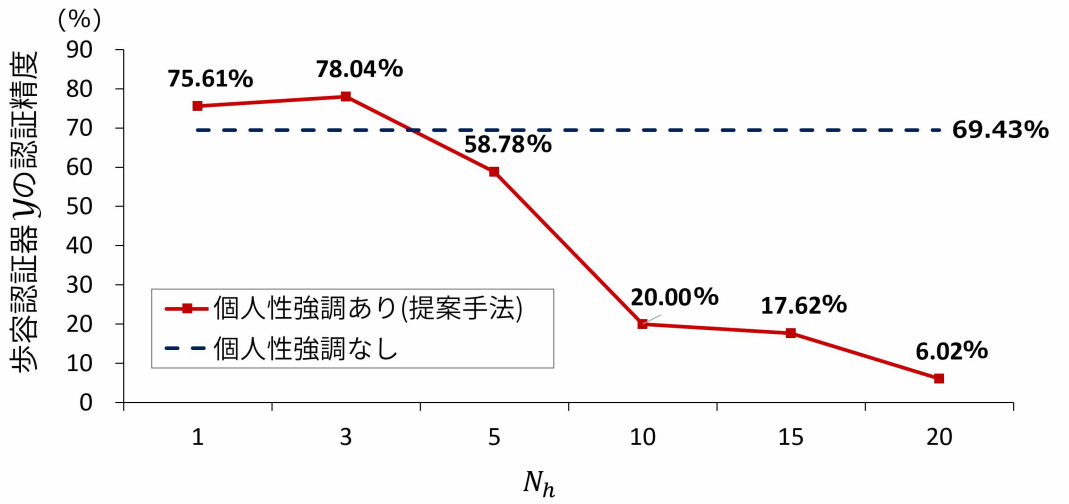


図 4.6: 様々な N_h における歩容認証器 $\mathscr{Y}$ の認証精度

歩容のなりすまし攻撃の成功率の高さに相当し，これは攻撃者にとって望ましい結果となる．シルエットレベルのエンコーダ E_{sil} とデコーダ D_{sil}，特徴マップレベルのエンコーダ E_{fm} とデコーダ D_{fm}，そして本人認証器 $\{\mathscr{X}_j\}$ は攻撃者のデータベースである Treadmill-dataset-(B) を用いて DNN として学習させた．これらの DNN のネットワーク構造を図 4.5 に示す．ここでは，形状ベクトル $\boldsymbol{\gamma} \in \mathbb{R}^{d'}$ の次元を 16，すなわち $d' = 16$ とした．

4.5.2 実験結果

図 4.6 は N_h を変化させながら歩容認証を行った場合の歩容認証器 $\mathscr{Y}$ の認証精度である．赤色の実線は提案手法で生成した偽の歩容シルエット列を与えたときの結果であり，青色の破線は個人性強調を行わない場合の結果である．両者を比較すると，$N_h = 1$ と $N_h = 3$ でより高い認識精度が得られていることがわかる．この結果は，個人性強調が歩容なりすまし攻撃を行う際の有効な手法であることを示している．一方で，$N_h \geq 5$ の場合，個人性強調によって歩容認識精度が著しく低下するという結果も得られた．個人性強調の目的は，スコア集合 $\{\omega'_j(D_{\mathrm{fm}}(\tilde{\boldsymbol{\gamma}}))|j = 1, \cdots, K'\}$ の中で比較的大きな要素を強調することである．しかし，なりすまし対象者は攻撃者のデータベースに存在しない個人であるため，これらのスコアのほとんどは小さなものである．従って，少なくとも本実験では，スコア集合の 4

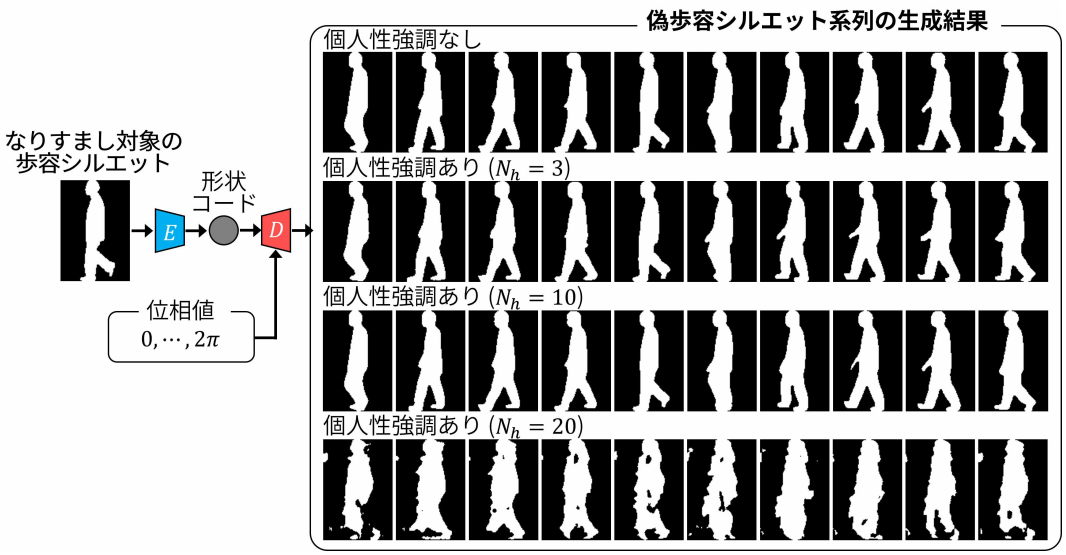


図 4.7: 提案手法で生成した偽の歩容シルエットの例

番目や 5 番目のスコアの大きな値でさえかなり小さい．このような値を大きくしてもなりすましをするうえで有用な効果は得られない．以上の考察から，N_h の最適な設定は攻撃者のデータベースの大きさに依存して異なることがわかった．攻撃者のデータベースに含まれる人物の数に応じて N_h をどの程度に設定すべきか，その依存関係が解明されれば，なりすましのリスクがより高まる可能性も十分に考えられる．

図 4.7 に提案手法で生成した偽の歩容シルエットの例と個人性強調を行わない場合のシルエットの例を示す．$N_h = 20$ の場合，生成されたシルエットの形が崩れていることがわかる．一方，N_h が比較的小さい場合（例えば，$N_h = 3$）には，自然なシルエットを生成できている．これらの結果から，提案手法は N_h を適切に設定することで，偽の歩容シルエットに大きな歪みを与えないことがわかる．また,「個人性強調なし」の場合，シルエットに含まれる腕の領域がうまく生成されない．これは，入力シルエットに腕の形状の情報がないためである．しかし，$N_h = 3$ の提案手法では，より自然に腕の領域を生成することができる．これが，図 4.6 において，提案手法が個人性強調を行わない場合よりも高い精度を実現できている理由である．

4.6 結言

本章では，DNN を用いて，一枚の写真から特定個人の偽歩容シルエット動画を作成して該当個人になりすまし，それにより偽情報を拡散する攻撃が行われるといった問題について，その実現可能性を考察した．本手法の主眼は，一枚の歩容シルエットをシルエット動画へと変換することと，その過程で歩容特徴を個人性強調によって補完することの二つである．実験結果より，個人性強調を行わなかった場合に比べて，個人性強調を行った場合のほうが認識精度が向上し，個人性強調の効果が歩容を用いたなりすまし攻撃に有用であることが確認された．また，その際の認識精度は 78.04%となった．よって，一枚の写真という比較的少ない情報に対してなりすまし攻撃が行われるという無視できないリスクが存在していることが示された．

第5章 真正歩容動画と偽歩容動画の識別

5.1 緒言

前章において，マルチメディアデータ生成処理を悪用した偽歩容動画によるなりすまし攻撃のリスクは無視できないことを示した．従って，それを防御する仕組みが必要となる．ここで，偽歩容動画がなりすまし攻撃に用いられる際には，一度シルエット化されたのちに歩容認証器に入力されることが想定される．従って，シルエット化された後の「歩容シルエット動画」の段階で，それが真正であるか偽であるかを識別することができれば，偽歩容動画によるなりすましを防ぐことが可能となる．すなわち，防御策としては，歩容シルエット動画の真偽を識別できる識別器があればよい．ただし，そのような識別器を機械学習するためには，真／偽以外の人物，服装，姿勢，視点の 4 つの要素が統制された歩容動画対からなる学習データセットが必要となる．本章では，その理由について，まず，5.2 節で論じ，続いて，そのような学習データセットを構築し，識別器を学習するための具体的な手法を 5.3 節で詳述する．その後，本手法の有効性を 5.4 節で実験的に評価し，最後に，5.5 節で本章の内容をまとめる．

なお，本章では，真正な歩容シルエット動画を GGS (Genuine Gait Silhouettes) と呼ぶ．また，偽の歩容シルエット動画については，Babaguchi が提唱した「メディアクローン (Media Clone)」の概念（本物ではないが限りなく本物に近い偽マルチメディアデータ全般）に因み，GSC (Gait Silhouette Clones) と呼称する．ここで，シルエットはカメラにより直接撮影・取得できる情報ではないため，実写であることを真正性の定義とすることはできない．そこで，GSS とは，カメラで実際に撮影された動画から人物領域抽出技術によって抽出されたシルエットのことと定義する．一方，それ以外の歩容シルエット動画，例えば，マルチメディアデータ処理により生成された歩容動画を二値化したものや，DNN の出力として直接得られた歩容シルエット動画などは，全て GSC として扱う．

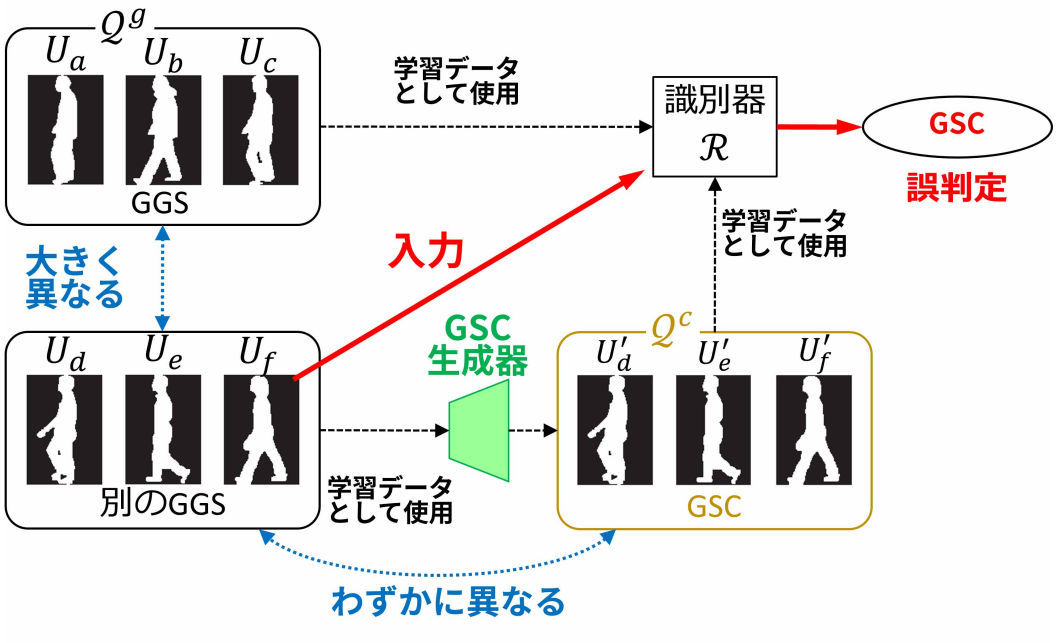


図 5.1: 真／偽以外に統制されていない条件が存在するデータセットの問題点

5.2 歩容動画の真偽識別器とその学習要件

5.2.1 識別器学習の概要

歩容シルエット動画の真偽を識別する識別器を，教師あり機械学習により構築するためには，GGS と GSC の双方を含む学習データセットが必要となる．以下，GGS の集合を $\mathcal{Q}^g = \{V_i^g | i = 1,\ 2,\ \cdots\}$，GSC の集合を $\mathcal{Q}^c = \{V_j^c | j = 1,\ 2,\ \cdots\}$ とし，それらを合わせて学習データセット $\mathcal{Q} = \mathcal{Q}^g \cup \mathcal{Q}^c$ を得るものとする．V_i^g および V_j^c は，それぞれ i 番目の GGS および j 番目の GSC を表す．上記のうち $\mathcal{Q}^g$ は，実際の人物を観測した動画に対して人物領域抽出技術を適用し，抽出結果を二値化することにより得ることができる．一方，$\mathcal{Q}^c$ は，2.4 節で述べたようなマルチメディアデータ生成技術により収集することが可能である．以上のようにして $\mathcal{Q}^g$ と $\mathcal{Q}^c$ を収集できれば，$\mathcal{Q}^g$ および $\mathcal{Q}^c$ に含まれる各動画から識別に効果的な特徴を自動抽出して，実際に識別を実行するニューラルネットワーク $\mathcal{R}$ を学習することは，それほど困難なことではない．従って，識別器の学習に際しては，$\mathcal{Q}^g$ と $\mathcal{Q}^c$ を如何に収集するかに主眼が置かれる．

ここで，5.1 節で言及したように，$\mathcal{Q}^g$ と $\mathcal{Q}^c$ は真偽以外の面で互いに異なっては

ならない．例として，図 5.1 に示すように，実在する三人の人物 U_a, U_b, U_c から収集した GGS により $\mathcal{Q}^g$ が構成されている一方，別の人物 U_d, U_e, U_f の GGS を模倣して生成された GSC（これを U'_d, U'_e, U'_f とおく）により $\mathcal{Q}^c$ が構成されている場合を考える．これは，$\mathcal{Q}^g$ と $\mathcal{Q}^c$ で個人性の条件が統制されていない場合に相当する．この $\mathcal{Q}^g$ と $\mathcal{Q}^c$ を学習データセットとして識別器を学習しても，その結果得られる識別器は期待通りには動作しない．具体的には，GGS である U_f が識別器に入力された場合，GSC であると誤判断される可能性が極めて高い．これは，歩容シルエットの形状は人によって異なる一方で，高度なマルチメディアデータ生成手法を用いて生成された GSC は，その元となった人物の GGS に極めて類似した形状を持つため，結果として，U_f が U_a, U_b, U_c よりも U'_f に類似している，という状況を招くためである．

以上の考察から，$\mathcal{Q}^g$ と $\mathcal{Q}^c$ は真／偽以外の条件が統制されていなければならない．これを満たすように $\mathcal{Q}^g$ と $\mathcal{Q}^c$ を収集するための手法を次のように提案する．

(1) 実在の人物を多数撮影して $\mathcal{Q}^g$ を収集する．

(2) 各 $V_i^g \in \mathcal{Q}^g$ に対し，それを模倣した GSC を生成して V_i^c とする．このとき，個人性や姿勢などの真／偽以外の条件は全て統制されるように V_i^c を生成する．

(3) 手順 (2) で生成した V_i^c を用いて $\mathcal{Q}^c = \{V_i^c | i = 1,\ 2,\ \cdots\}$ を構成する．

(4) $\mathcal{Q} = \mathcal{Q}^g \cup \mathcal{Q}^c$ を用いて，識別器 $\mathcal{R}$ を学習する．

上記のうち手順 (2) について，V_i^g と V_i^c に要求される性質を次節でより詳細に説明する．

5.2.2 学習データセットに求められる要件

V_i^g と V_i^c に要求される性質を考えるにあたり，まず，一枚一枚の歩容シルエットがどのような要素から構成されているかについて議論する．以下，V_i^g の k フレーム目を $s_{i,k}^g$ とする．同様に，V_i^c の k フレーム目を $s_{i,k}^c$ とする．このとき，歩容シルエット $s_{i,k}^g$ および $s_{i,k}^c$ は，主に人物，服装，姿勢，視点の 4 つの要素によって決定される．これは，同一人物が同じ服装のもとで同じ姿勢を取っているとき，その人物を同じ視点から観察すれば，常にほぼ同じシルエット画像が得られるためである．

$s^g_{i,k}$ における人物，服装，姿勢，視点を表現する特徴ベクトルをそれぞれ $\boldsymbol{f}^{\mathrm{pe}}_{i,k}$，$\boldsymbol{f}^{\mathrm{cl}}_{i,k}$，$\boldsymbol{f}^{\mathrm{po}}_{i,k}$，$\boldsymbol{f}^{\mathrm{vi}}_{i,k}$ とする．また，これら 4 つのベクトルを連結したものを

$$\boldsymbol{f}_{i,k} = \left((\boldsymbol{f}^{\mathrm{pe}}_{i,k})^\top,\ (\boldsymbol{f}^{\mathrm{cl}}_{i,k})^\top,\ (\boldsymbol{f}^{\mathrm{po}}_{i,k})^\top,\ (\boldsymbol{f}^{\mathrm{vi}}_{i,k})^\top\right)^\top \ . \tag{5.1}$$

とする．ここで，理想的な特性を持つ画像デコーダ D^{ideal} の存在を仮定すると，$\boldsymbol{f}_{i,k}$ と $s^g_{i,k}$ の関係は

$$s^g_{i,k} \ = \ D^{\mathrm{ideal}}[\boldsymbol{f}_{i,k}] + \epsilon^g_{i,k} \tag{5.2}$$

のように定式化できる．数学的には，D^{ideal} は特徴ベクトルからシルエット画像への写像である．また，$\epsilon^g_{i,k}$ は，人物領域抽出前の画像に含まれる画像ノイズ（画像センサーにおけるノイズ）や人物領域抽出処理の誤差などに起因する誤差項を意味する．ただし，歩容シルエット画像に対する誤差項の影響は，上記 4 つの要素に比べてはるかに小さいと仮定する．

$s^g_{i,k}$ と同様にして，$s^c_{i,k}$ における人物，服装，姿勢，視点を表現する特徴ベクトルを $\boldsymbol{h}_{i,k}$ とおくと，$s^c_{i,k}$ と $\boldsymbol{h}_{i,k}$ の関係は

$$s^c_{i,k} \ = \ D^{\mathrm{ideal}}[\boldsymbol{h}_{i,k}] + \epsilon^c_{i,k} \tag{5.3}$$

のように定式化できる．実際には，$s^c_{i,k}$ の生成には現実の画像デコーダ D が用いられるので，上式はさらに

$$s^c_{i,k} \ = \ D^{\mathrm{ideal}}[\boldsymbol{h}_{i,k}] + \epsilon^c_{i,k} \ = \ D[\boldsymbol{h}_{i,k}] \tag{5.4}$$

のようにおくことができる．このとき，式 (5.3)，式 (5.4) における $\epsilon^c_{i,k}$ は，デコーダ D に起因する誤差項に相当する．

以上 2 つの誤差項 $\epsilon^g_{i,k}$ および $\epsilon^c_{i,k}$ は，それぞれ異なる要因により生じるものであるため，その分布は互いに異なるものであると仮定する．この場合に，識別器 $\mathcal{R}$ はこの違いを捉える必要があり，そのためには，誤差項以外の要件は全て統制されていなければならない．すなわち，全ての i と k について

$$\boldsymbol{f}_{i,k} \ = \ \boldsymbol{h}_{i,k} \tag{5.5}$$

が満たされる必要があり，これが学習データセットに求められる必須要件となる．

5.3 自己符号化器による学習データセット構築に基づく歩容動画の真偽識別

前節で述べた要件を踏まえ，本節では，歩容シルエット動画の真／偽を識別する手法を具体的に述べる．重要となるのは，5.2.1 節で提案した手順のうち (2) および (4) である．そこで，まず，手順 (2) について 5.3.1 節にて詳述する．その後，手順 (4) の詳細を 5.3.2 節にて述べる．

なお，これ以降，式 (5.3)，式 (5.4) 等で用いた表現において i と k を指定する必要がない場合，これらを省略する（例えば，$\epsilon^g_{i,k}$ の代わりに ϵ^g を使う場合もある）．

5.3.1 自己符号化器による学習データセットの構築

5.2 節で述べた必須要件は，以下の方法で満たすことができる．まず，あるエンコーダ E を用いて $s^g_{i,k}$ から $\boldsymbol{f}_{i,k}$ を

$$\boldsymbol{f}_{i,k} \;=\; E\big[s^g_{i,k}\big] \tag{5.6}$$

のように抽出する．数学的には，エンコーダ E はシルエット画像から特徴ベクトルへの写像である．その後，D により抽出された特徴量から $s^c_{i,k}$ を

$$s^c_{i,k} \;=\; D[\boldsymbol{f}_{i,k}] \;=\; D\big[E\big[s^g_{i,k}\big]\big] \tag{5.7}$$

として生成する．ここで，特徴ベクトル $\boldsymbol{f}_{i,k}$ は様々に定義することが可能であるが，どのような定義であれ，すべての i と k について $\boldsymbol{f}_{i,k}$ の正解データを求めることは容易ではない．この問題を避けるため，提案手法では $\boldsymbol{f}_{i,k}$ を明示的に定義しないこととする．

5.2.2 節で述べたように，誤差項 ϵ^g は，人物，服装，姿勢，視点の 4 要素に比べ，歩容シルエット画像に与える影響が非常に小さいと仮定している．つまり，ϵ^g の変化による歩容シルエット画像の分散は，$\boldsymbol{f}$ の変化による分散よりはるかに小さいと仮定する．この場合に，歩容シルエット画像から主成分を抽出したとすると，その成分は上記の 4 要素の一つ以上と関係することが期待され，誤差項 ϵ^g とは無関係であることが期待される．このとき，抽出された主成分が上記の 4 要素とどのように相関しているかは明示的に明らかとはならないが，これは問題ではない．重要なのは，このような仮定の場合 4 要素すべてから独立した次元は存在せず，従って，ϵ^g とは無関係であるという点である．

以上の考察から，提案手法では，$\mathcal{Q}^g$ を用いて自己符号化器 (Auto Encoder; AE) を学習し，そのエンコーダ部を E，デコーダ部を D として用いる．具体的には，図 5.2 に示すように，まず $\mathcal{Q}^g$ をいくつかのサブセットに分割したのち，それぞれのサブセットを個別に学習データとして用いることにより，複数の AE を学習する．なお，これらのサブセットには重複が存在してもよい．その後，学習した各 AE に各 $s^g_{i,k}$ を入力し，全ての i，k に対して $s^c_{i,k} = D\big[E\big[s^g_{i,k}\big]\big]$ を計算する．以上により，$\mathcal{Q}^c = \{V^c_i | i = 1,\ 2,\ \cdots\}$ を得る．上記の AE の学習に際し，損失関数 L_2 としては平均二乗誤差を採用し，

$$L_2[E, D] \ = \ \sum_i \sum_k \Big|\Big|\ s^g_{i,k} - D\big[E\big[s^g_{i,k}\big]\big]\ \Big|\Big|^2 \tag{5.8}$$

と定める．このとき，中間層のユニット数，すなわち，ベクトル $E[s^g]$ の次元数は，s^g の画素数よりはるかに小さく設定する．この結果，E を用いて s^g から抽出される $E[s^g]$ は，上述の 4 要素のうち一つ以上と関係し，誤差項 ϵ^g とは無関係な主成分を表すこととなる．以上の手順で生成された $\mathcal{Q}^c$ は，前節で述べた要件を満たしている．以降では，E により抽出されたベクトル $E[s^g]$ の各次元を潜在特徴と呼ぶ．

$\mathcal{Q}^g$ をサブセットに分割する方法としては，ブートストラップサンプリングを採用する．具体的には，$\mathcal{Q}^g$ から所定の枚数の歩容シルエット画像をランダムに選択し，それにより 1 つのサブセットを得る．これを十分な回数繰り返し，複数のサブセットを得る．このブートストラップサンプリングにより得られた各サブセットは互いに少しずつ異なるため，それを用いて学習した AE もまた互いに多少異なり，その結果，生成される $\mathcal{Q}^c$ はより多様なものとなる．つまり，様々なサブセットを用いて学習した複数の AE を用いることで，識別器 $\mathcal{R}$ を学習するための GSC をより多様化することができ，提案手法の汎化能力の向上が期待される．

5.3.2 具体的な真偽識別手法

前節の手法で生成した $\mathcal{Q}^g$ および $\mathcal{Q}^c$ を用いて，最終的に識別器 $\mathcal{R}$ を学習する．5.2.2 節において，$\mathcal{R}$ は ϵ^g と ϵ^c の違いを識別すべきであると述べたが，実際には両者はわずかな違いしかないため，その識別は必ずしも容易ではない．むしろ，ϵ^g や ϵ^c そのものよりも，その時間変化の方が識別に有効な情報をより多く含んでいると考えられる．このことから，提案手法では，所与の歩容シルエット動画をフ

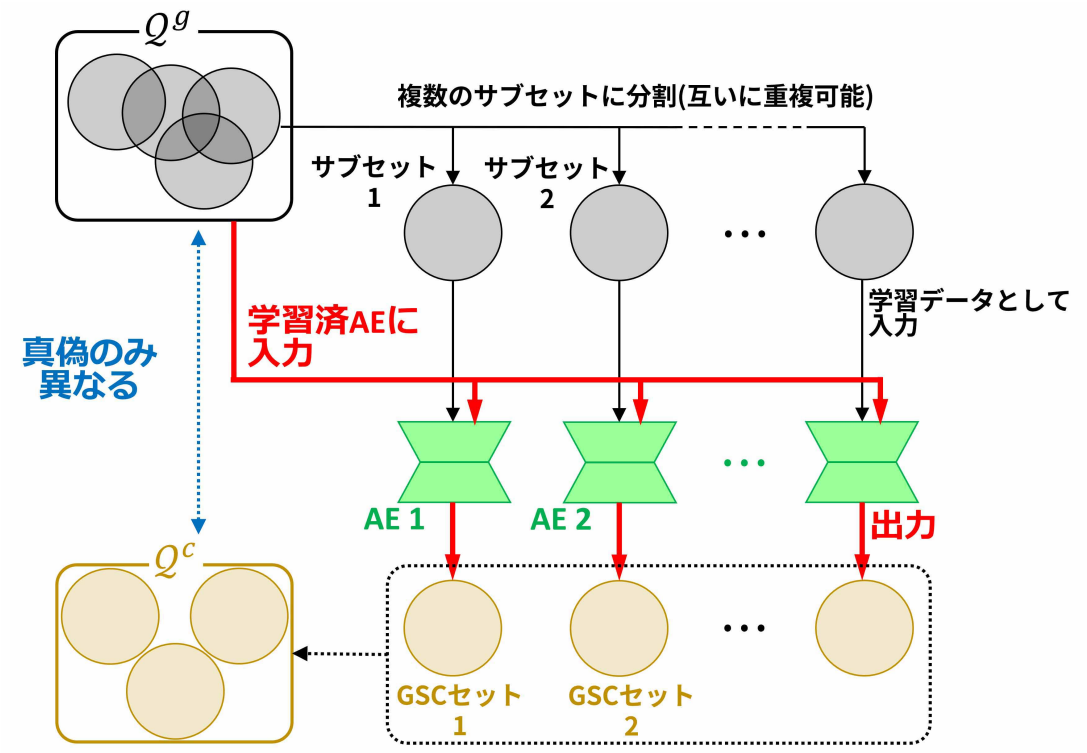


図 5.2: AE による学習データセット構築の概要

レーム単位で処理するのではなく，まず，動画をフレーム長 N_{ch} のセグメントに分割し，セグメントごとに真偽を識別することとする（以下，$\mathcal{R}$ はこのための識別器を指すものとする）．その上で，セグメントごとの識別結果を最終的に多数決ルールで統合し，単一の識別結果（GGS か GSC か）を定める．具体的には，T を動画の総フレーム数とするとき，セグメントの総数（投票数）は $T_{\mathrm{vote}} = T/N_{\mathrm{ch}}$ であり，このうち T_{cloned} 本のセグメントが GSC と識別されたとすると，

$$\frac{T_{\mathrm{cloned}}}{T_{\mathrm{vote}}} > 0.5$$

のときのみ，元の歩容シルエット動画が GSC であったと識別する．

提案手法では，上記の各セグメントを N_{ch} チャネルの画像とみなして $\mathcal{R}$ に入力する．$\mathcal{R}$ は，図 5.3 に示すような CNN として設計し，$\mathcal{R}$ の損失関数には交差エントロピーを採用する．

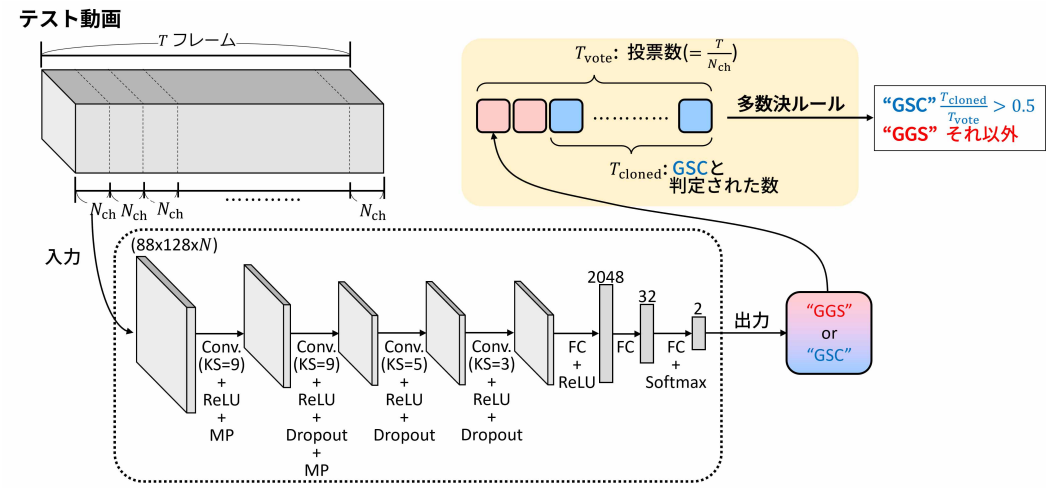


図 5.3: セグメント単位での真偽識別器 $\mathcal{R}$ のネットワーク構造（Conv: 畳み込み層，MP: 最大値プーリング層，FC: 全結合層，KS: 畳み込み層のカーネルサイズ）

5.4 評価実験

提案手法によりどの程度の精度で真偽識別が可能であるかを評価するため，実験を行った．本節では，まず 5.4.1 節で実験設定（主に実験用データセットの準備）について説明したのち，続いて，5.4.2 節で実験結果とその考察を述べる．また，追加的な検証として，提案手法により得られた識別器や AE の特性を実験的に調査した．この結果について，5.4.3 節および 5.4.4 節で報告する．

5.4.1 実験設定

本実験で用いたデータセットは，第 3 章および第 4 章で言及した実験と同様，OU-ISIR Gait Database [47] であり，より具体的には，その中に含まれるサブデータセット Treadmill-dataset-(A)，Treadmill-dataset-(B)，Treadmill-dataset-(C) の 3 つである．これら 3 つのサブデータセットには，それぞれ 612, 2176, 370 本の歩容シルエット動画が含まれている．本実験では，3 つのサブデータセットを一つにまとめた上で，全 3158 本の動画すべてを GGS とみなし，これを 3 つに分割して使用した．以降では，これらをそれぞれ *Train-GGSs*，*Train-GGSs-Another*，および *Test-GGSs-Closed* と呼ぶ．*Train-GGSs* には全体（3158 本）の 1/2 の動画が含まれており，*Train-GGSs-Another* および *Test-GGSs-Closed* には全体の 1/4 の動

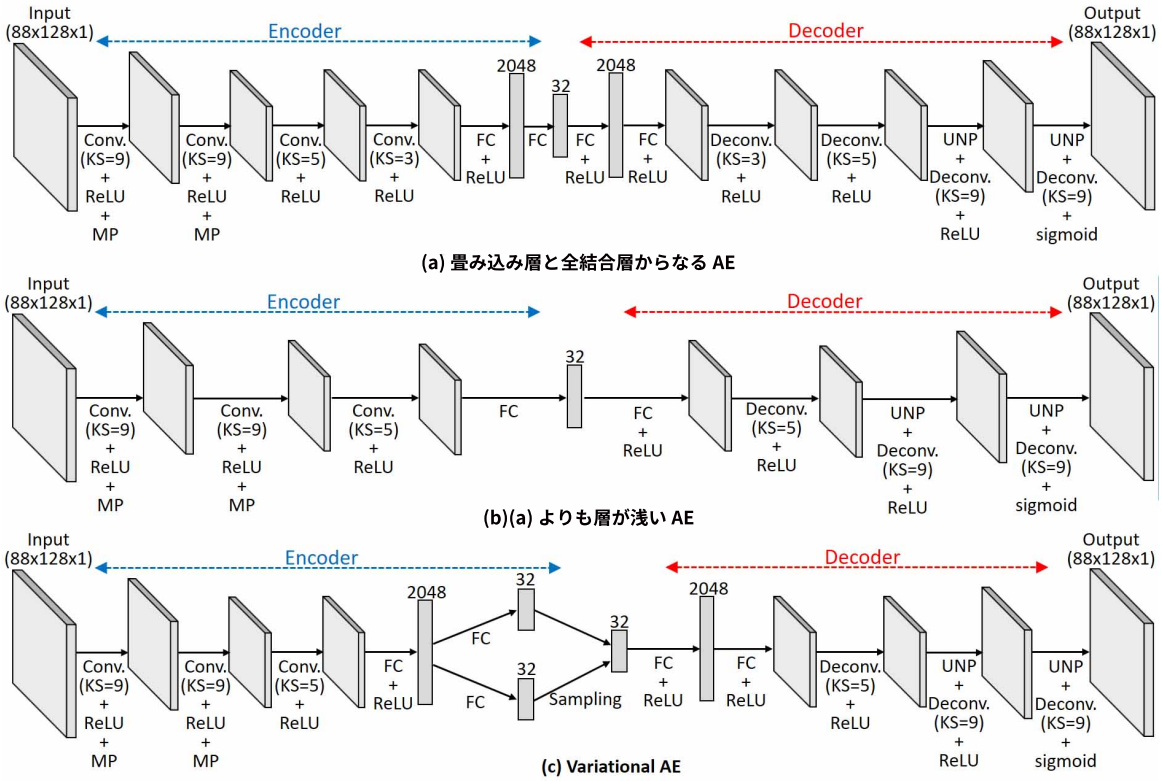


図 5.4: GGS から GSC を生成する際に用いた 3 つの AE のネットワーク構造（Conv，MP，FC，KS は図 5.3 と同様，Deconv: 逆畳み込み層，UNP: 逆プーリング層）

画がそれぞれ含まれている．これら 3 つの用途については後述する．

次に，*Train-GGSs*，*Train-GGSs-Another*，*Test-GGSs-Closed* に含まれる動画をそれぞれ AE に入力して対応する出力を取得し，それらを GSC として使用した．本実験で使用した AE は全 3 種類あり，その各々のネットワーク構造を図 5.4 に示す．これら 3 つの AE は，それぞれ別々に 3 回ずつ学習させた．つまり，本実験では AE の学習を 9 回試行し，合計 9 種類の AE を用意したことになる．なお，AE の学習に用いたのは *Train-GGSs* のみであり，その中からブートストラップサンプリングにより 9 万フレーム分のシルエット画像を選択して使用した．図 5.5 に，これら 9 種類の AE によって生成された GSC の一例を，その元となった GGS と共に示す．以下では，*Train-GGSs*，*Train-GGSs-Another*，*Test-GGSs-Closed* から生成された GSC の集合をそれぞれ *Train-GSCs*，*Train-GSCs-Another*，*Test-GSCs-*

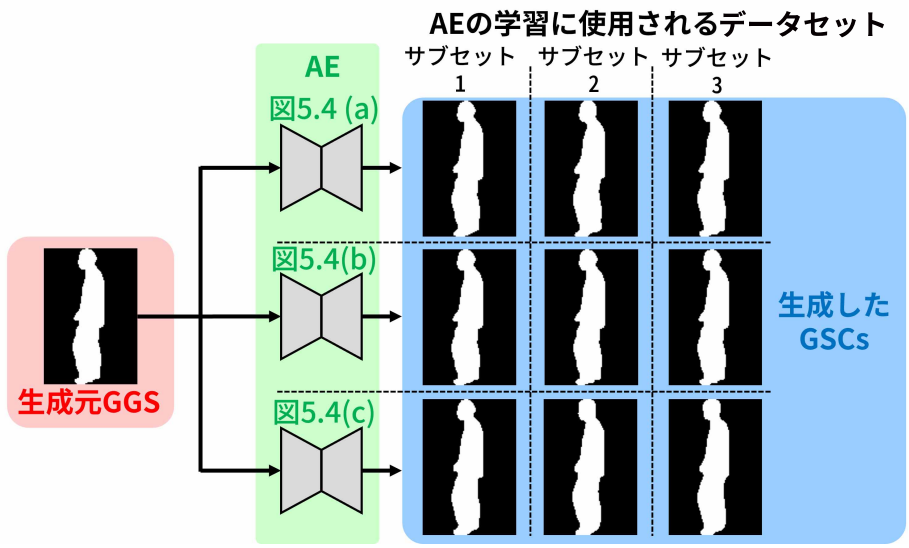


図 5.5: AE により生成した GSC およびその元となった GGS の例

Closed と呼ぶことにする．

提案手法の主眼点は，真偽以外の面で互いに一致している $\mathcal{Q}^g$ と $\mathcal{Q}^c$ を用いることであるため，提案手法における識別器 $\mathcal{R}$ の学習に際しては，$\mathcal{Q}^g$ として *Train-GGSs* を，$\mathcal{Q}^c$ として *Train-GSCs* を用いた．この $\mathcal{R}$ とは別に，$\mathcal{Q}^g$ として *Train-GGSs*，$\mathcal{Q}^c$ として *Train-GSCs-Another* を用いた場合の識別器 $\mathcal{R}'$ も学習し，その性能を $\mathcal{R}$ の性能と比較した．*Train-GGSs* と *Train-GSCs-Another* は，真偽の面に加え，人物や姿勢の面でも違いがあるため，それにより学習された $\mathcal{R}'$ は，5.2.1 節で述べた理由により適切に動作しないと考えられる．この点を実験的に検証するため，テストデータセットとして *Test-GGSs-Closed* と *Test-GSCs-Closed* を用いて，$\mathcal{R}$ と $\mathcal{R}'$ の性能を比較検討した．

さらに，提案手法の汎用性を評価するため，上記とは全く別のテストデータセットも用意した．まず GGS については，14 人の人物の歩行の様子を側面から撮影し，それにより得られた画像にクロマキー処理を施すことによりシルエットを抽出した．上述の OU-ISIR Gait Database とは異なり，この撮影の際にはトレッドミルを用いていないため，抽出されたシルエットの位置は，撮影された動画の中で時間経過とともに移動する．そこで，このシルエットの位置を正規化するために，頭部領域の中心が固定されるようにシルエットを平行移動させた．この点で，本データセットは OU-ISIR Gait Database の撮影環境とは大きく異なる．以

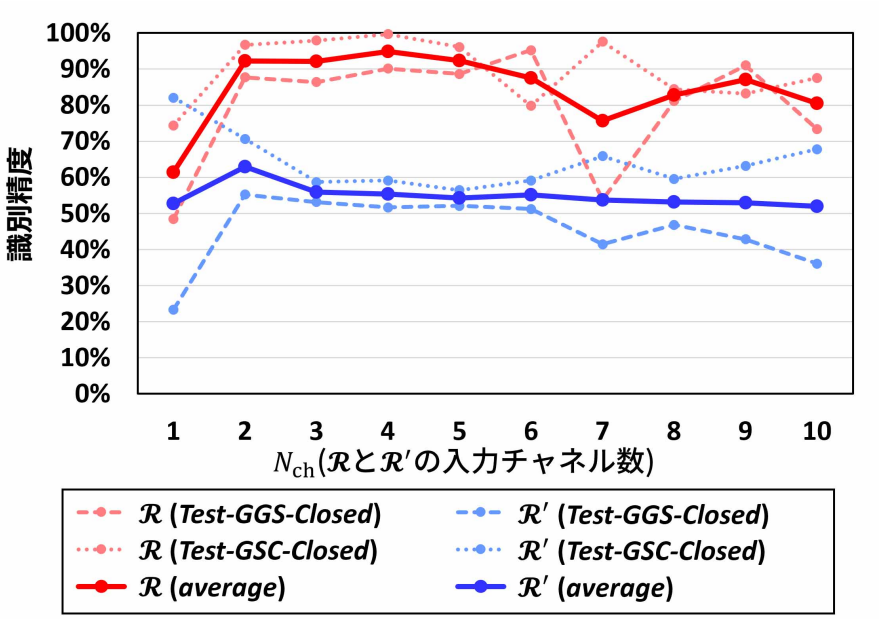


図 5.6: *Test-GGSs-Closed* と *Test-GSCs-Closed* に対する識別精度（赤線は提案手法 $\mathcal{R}$，青線は比較手法 $\mathcal{R}'$ の結果）

上の処理を 1 人あたり 13 回繰り返し，$13 \times 14 = 182$ 本の動画を撮影した．この動画セットを *Test-GGSs-Open* と呼ぶ．

一方，GSC については，3.3.4 節で述べたニューラルネットワークを転用して生成した．3.3.4 節では，歩容シルエット画像はその位相成分と形状成分によって再構成することができるとの考えで，位相コードと形状コードからシルエット画像を生成するデコーダを設計した．このときの形状コードは人物，服装，視点の側面を表すと考えられることから，これに摂動を加えることなく（摂動を 0 にして）シルエット画像を生成することにより，元のシルエットの人物，服装，視点の要素を可能な限り保持した GSC を生成することが可能である．そこで，Treadmill-dataset-(A) の中からランダムに選択されたフレームに対し，実際にそのような手順を適用し，306 本の動画を生成した．この動画セットを *Test-GSCs-Open* と呼ぶ．

5.4.2 実験結果

図 5.6 は，N_{ch} を変化させた際の *Test-GGSs-Closed* と *Test-GSCs-Closed* に対する $\mathcal{R}$ および $\mathcal{R}'$ の識別精度を示している（N_{ch} は 5.3.2 節で定義したセグメント長

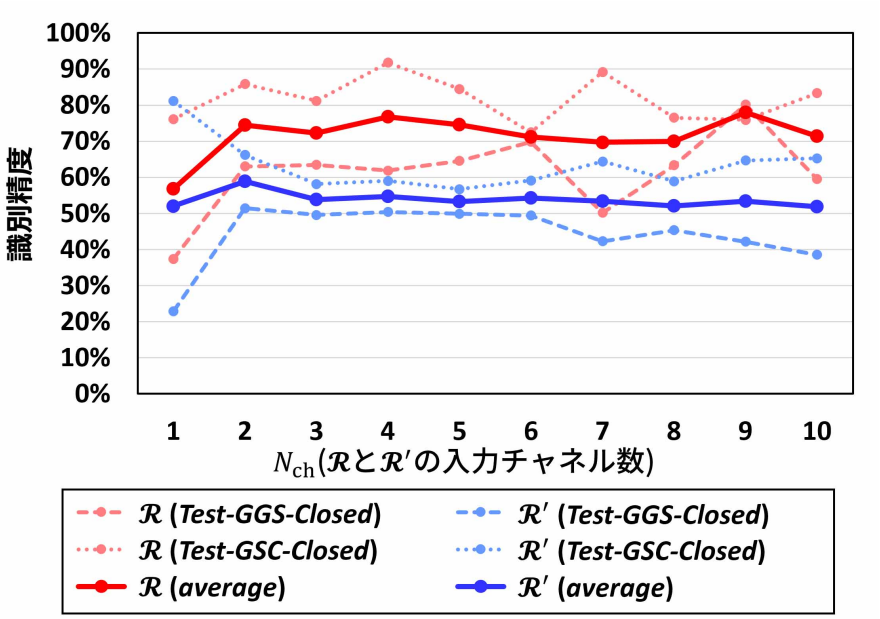


図 5.7: *Test-GGSs-Closed* と *Test-GSCs-Closed* に対する多数決ルール適用前の識別精度

である）．$N_{\mathrm{ch}} \geq 2$ のとき，$\mathcal{R}$ では平均 80%以上の精度で GGS と GSC の識別に成功している．特に $N_{\mathrm{ch}} = 4$ では，94.9%の識別精度が得られた．これに対し，$\mathcal{R}'$ の識別精度はほとんどの場合で 60%以下である．この結果は，提案手法の有効性を示すものと言える．

ここで，図 5.6 の結果では，N_{ch} が大きくなるにつれて $\mathcal{R}$ の精度がやや低くなる傾向が見受けられる．これは，5.3.2 節で述べた多数決の過程において，N_{ch} が大きくなると投票数の総和が小さくなるためである．例えば，$N_{\mathrm{ch}} = 2$ では，長さ 100 フレームの歩容シルエット動画を長さ 2 フレームのセグメントに分割することで合計 50 セグメントを得ることになるが，$N_{\mathrm{ch}} = 10$ では，合計 10 セグメントしか得ることができない．従って，N_{ch} が大きくなると多数決の信頼性が低くなり，最終的な識別精度が低下する．このことは，図 5.7 に示すように，多数決前の $\mathcal{R}$ においては N_{ch} が大きくてもそれほど識別精度の低下が見られないことからも確認できる．

$N_{\mathrm{ch}} = 1$ の場合，$\mathcal{R}'$ だけでなく，$\mathcal{R}$ も適切に機能しない．つまり，1 フレームの情報のみから GGS と GSC を識別することは困難であるということが分かる．5.2.2 節で述べたように，ϵ^g と ϵ^c の分布は異なるが，その違いはごくわずかであ

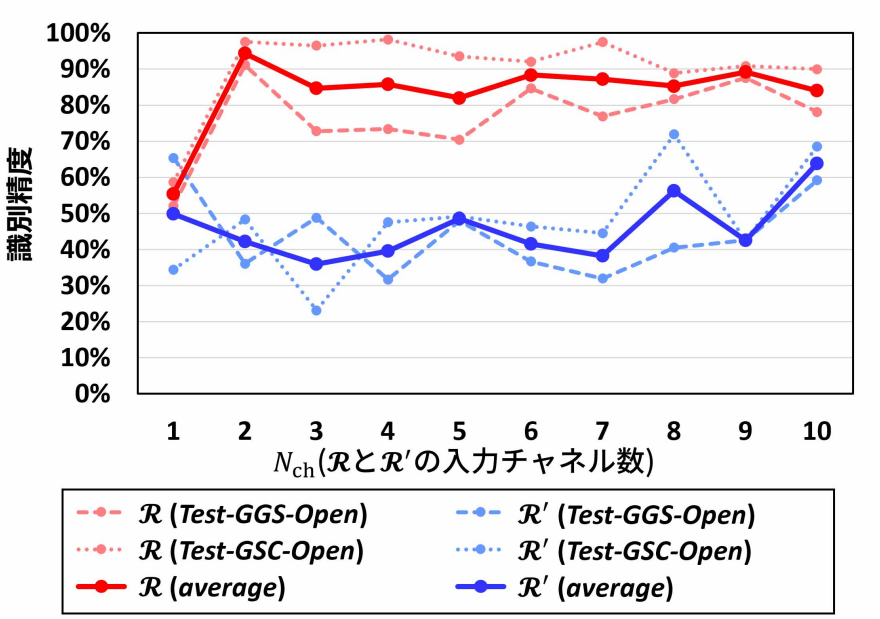


図 5.8: *Test-GGSs-Open* と *Test-GSCs-Open* に対する識別精度

るため，DNN ベースの識別器でも捉えることは困難である．一方，ϵ^g の時間変化は ϵ^c の時間変化とは十分に異なるため，$N_{\mathrm{ch}} \geq 2$ の場合，$\mathcal{R}$ は GSC と GGS を適切に識別することができる．この理由は以下のように考えられる．GGS において誤差 ϵ^g が発生する主な要因の一つは画像ノイズであり，フレームごとに独立性が高いと思われる．そのため，ϵ^g が時間経過とともに滑らかに変化することはあまり考えられない．一方，GSC における誤差 ϵ^c は，完全にコンピュータ処理に起因するため，このようなフレームレベルのランダム性に欠けると思われる．従って，シルエット形状の類似した連続する 2 つのフレームがデコーダ D により生成された場合，この 2 つのフレームに対応する誤差項も類似することになる．その結果，ϵ^c は ϵ^g とは逆に時間経過とともに滑らかに変化する．この違いを $\mathcal{R}$ は適切に捉えられるものと考えられる．

次に，*Test-GGSs-Open* と *Test-GSCs-Open* における $\mathcal{R}$ および $\mathcal{R}'$ の識別精度を図 5.8 に示す．図 5.8 においても図 5.6 と同様の傾向が見られる．*Test-GGSs-Open* および *Test-GSCs-Open* の収集方法は学習データ *Train-GGSs*，*Train-GSCs* の収集方法とは全く異なるにもかかわらず，$\mathcal{R}$ の精度は $N_{\mathrm{ch}} \geq 2$ において平均 80%以上を達成している．これに対し，$\mathcal{R}'$ はほとんどの場合 50%程度であることが分か

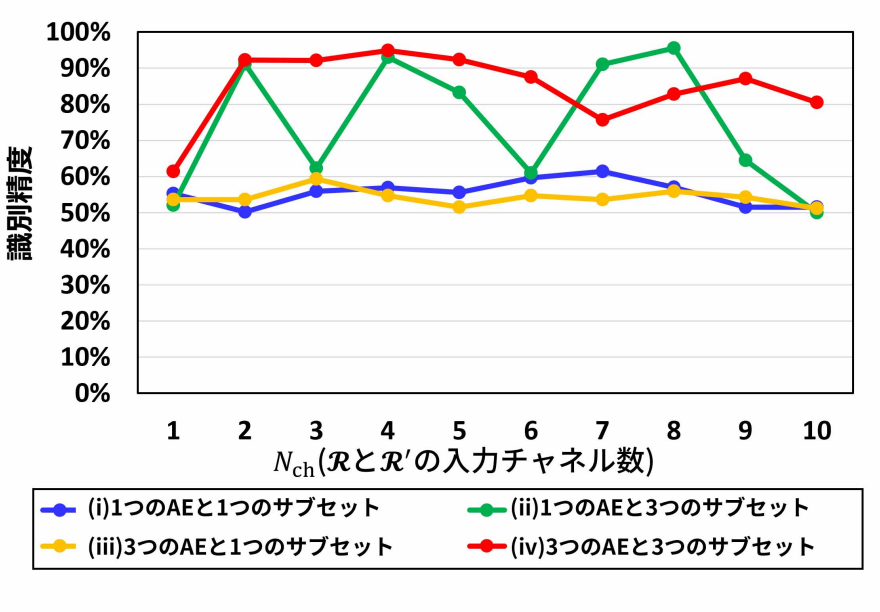


図 5.9: 4 種類の条件の下での *Test-GGSs-Closed* および *Test-GSCs-Closed* に対する $\mathcal{R}$ の識別精度

る．$\mathcal{R}$ は $N_{\mathrm{ch}} = 2$ において 94.3%の最高精度を達成しているが，これは，図 5.6 における *Test-GGSs-Closed* および *Test-GSCs-Closed* の最高精度とほぼ同等である．この結果は提案手法の汎用性の高さを示している．

5.3.1 節で述べたように，提案手法は，$\mathcal{Q}^g$ の様々なサブセットで学習させた複数の AE を用いて $\mathcal{Q}^c$ を構成している．この処理の有効性を実験的に検証するため，以下の 4 条件で $\mathcal{R}$ の識別精度を評価し，その結果を比較した．

(i) 1 つの AE を 1 つのサブセットで学習させた場合

(ii) 1 つの AE を 3 つの異なるサブセットで 3 回学習させた場合

(iii) 3 つの AE を 1 つのサブセットで 3 回学習させた場合

(iv) 3 つの AE を異なるサブセットで 3 回学習させた場合

その結果を図 5.9，図 5.10 に示す．この結果から，条件 (i)，(iii) よりも条件 (ii)，(iv) の方が良好な性能が得られていることがわかる．また，条件 (iv) で最も高い識別精度が得られている．この事実は，様々なサブセットを用いることの有効性を示している．

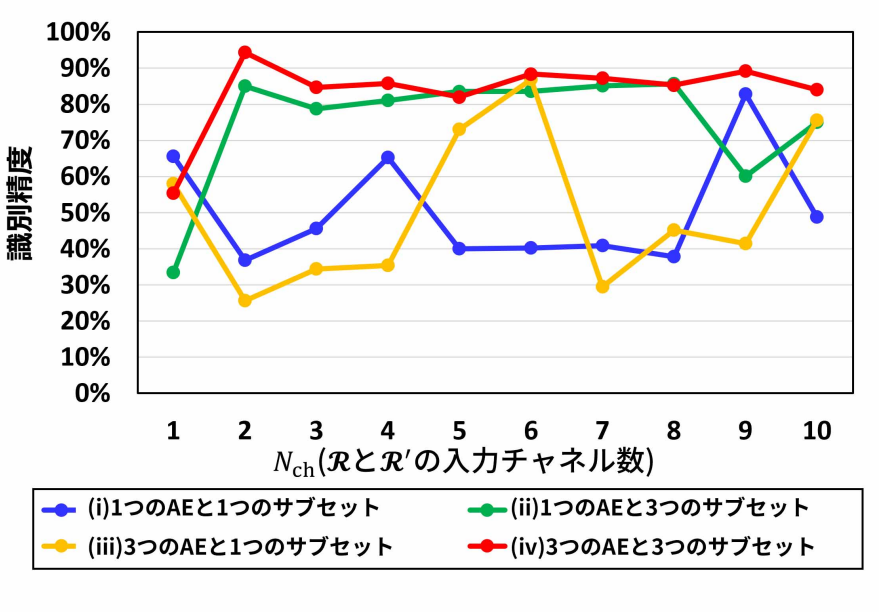


図 5.10: 4 種類の条件の下での *Test-GGSs-Open* および *Test-GSCs-Open* に対する $\mathcal{R}$ の識別精度

これまでに述べた実験の結果から，次の結論を得ることができる．第一に，GSC と GGS を識別するためには，真正性以外は可能な限り $\mathcal{Q}^g$ と同一の $\mathcal{Q}^c$ を生成することが重要である．第二に，$\mathcal{Q}^g$ の様々なサブセットで学習した複数の AE を用いて $\mathcal{Q}^c$ を構成することにより，識別器の汎用性が大幅に向上する．

5.4.3　学習済み自己符号化器の特性

5.2 節，5.3 節において，次の 2 つの仮説を提示した．一つは，デコーダに起因する ϵ^c の分布が ϵ^g の分布とは異なる，というものである．もう 1 つは，AE により抽出される潜在特徴は，主要な 4 要素（人物，服装，姿勢，視点）のうち少なくとも 1 つと相関がある，というものである．本節では，これらの仮説を定量的に検証した結果について述べる．

誤差項の解析

式 (5.2)，式 (5.3) に示すように，誤差項は，理想的なデコーダ D^{ideal} の出力と実際のシルエット画像との差分として定義される．しかし，実環境では D^{ideal} が存

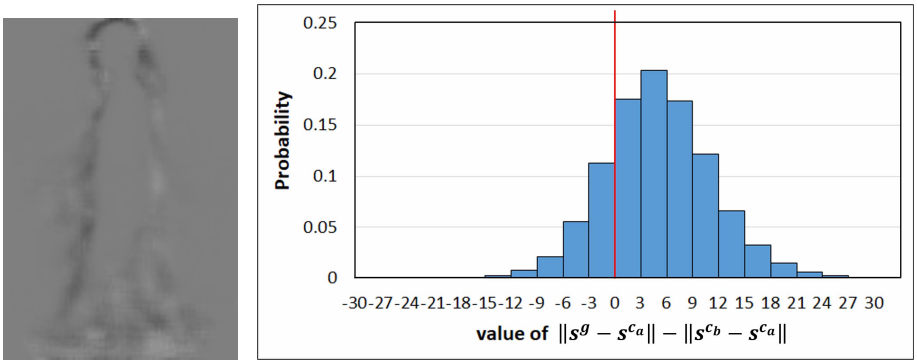


図 5.11: (左) $\mathbb{E}[s^g - s^{c_a}]$ の可視化画像，(右) $||s^g - s^{c_a}|| - ||s^{c_b} - s^{c_a}||$ の分布

在しないため，ϵ^g および ϵ^c を計算することは不可能であり，誤差項を直接解析することはできない．そこで，直接的な解析ではなく，GGS とそれを AE で再構成した結果との差分を解析した．具体的には，Treadmill-dataset-(B) に含まれるすべてのシルエット画像 s^g を図 5.4(a)，(b) に示す AE に入力し，GSC を生成した．それらを s^{c_a}，s^{c_b} とすると，s^g に含まれる主要な 4 要素は s^{c_a} にも含まれていることが予想されるため，

$$\begin{aligned} s^g - s^{c_a} &= (D^{\text{ideal}}[\boldsymbol{f}] + \epsilon^g) - (D^{\text{ideal}}[\boldsymbol{f}] + \epsilon^{c_a}) \\ &= \epsilon^g - \epsilon^{c_a} \end{aligned} \tag{5.9}$$

が満たされる．これは，s^{c_b} の場合にも同様のことが言える．

以上のことを踏まえ，まず，$s^g - s^{c_a}$ の平均，すなわち $\mathbb{E}[s^g - s^{c_a}] = \mathbb{E}[\epsilon^g - \epsilon^{c_a}]$ を解析した．これが有意に 0 でない場合には，ϵ^g の分布と ϵ^{c_a} の分布に違いがあると判断することができる．その結果を図 5.11 の左側に示す．この図の黒，灰，白の領域はそれぞれ $\mathbb{E}[s^g(x) - s^{c_a}(x)] < 0$，$\mathbb{E}[s^g(x) - s^{c_a}(x)] = 0$，$\mathbb{E}[s^g(x) - s^{c_a}(x)] > 0$ を示している．ただし，$s^g(x)$ および $s^{c_a}(x)$ はそれぞれ，s^g および s^{c_a} における画素 x の画素値である．視認性を上げるために上記の各画素値を 8 倍して表示しているため，s^g と s^{c_a} の差は実際よりも強調されているものの，この結果は，$\mathbb{E}[s^g - s^{c_a}]$ が有意に 0 でないことを明確に示している．

次に，$||s^g - s^{c_a}|| - ||s^{c_b} - s^{c_a}||$ の分布について解析した．もし，$||s^{c_b} - s^{c_a}||$ が $||s^g - s^{c_a}||$ よりも小さい傾向となるならば，別の異なるデコーダ D で生成した GSC の誤差項の分布同士は GSC の誤差項の分布と GGS の誤差項の分布を比べた場合よりも類似していると判断することができる．この解析結果を図 5.11 の右側に示す．

この結果から，$||s^{g} - s^{c_{\mathrm{a}}}|| - ||s^{c_{\mathrm{b}}} - s^{c_{\mathrm{a}}}||$ の値が 0 より大きくなる傾向を持っていることが分かる．これは $||s^{c_{\mathrm{b}}} - s^{c_{\mathrm{a}}}||$ が $||s^{g} - s^{c_{\mathrm{a}}}||$ よりも小さくなる傾向があることを明示している．以上の 2 つの解析結果は，誤差項に関する仮説の正しさを示していると言える．

AE により抽出された潜在特徴の解析

AE により抽出された潜在特徴の解析に際しても，Treadmill-dataset-(B) を用いた．Treadmill-dataset-(B) には，68 人の人物がそれぞれ 32 種類の服装を着用して歩行した際の歩容シルエット動画が 1 本ずつ含まれており，各動画は様々な姿勢の画像系列で構成されているため，今回の解析に適している．なお，本データセットでは，Treadmill-dataset-(B) に含まれるすべての動画が同じ視点から撮影されているため，視点要素に関しては無視し，人物，服装，姿勢の 3 つの要素のみに着目して解析を行っている．

前述したように，Treadmill-dataset-(B) は $68 \times 32 = 2176$ 本の動画を含んでおり，68 は人物クラス，32 は服装クラスの数である．これは，人物クラスと服装クラスの直積集合である人物・服装クラスが 2176 種類存在することを意味する．この性質に基づき，図 5.4 (a) に示す AE を用いて 2176 本の動画の各フレームから 32 次元の潜在特徴を抽出し，各特徴について，全分散 $v_{\mathrm{t}}(d)$，人物クラスのクラス内分散 $v_{\mathrm{p}}(d)$，服装クラスのクラス内分散 $v_{\mathrm{c}}(d)$，人物・服装クラスのクラス内分散 $v_{\mathrm{pc}}(d)$ という 4 種類を算出した．ここで，d $(1 \leq d \leq 32)$ は，各潜在特徴の ID（何次元目の特徴であるかを示す番号）である．

上述の $v_{\mathrm{pc}}(d)$ は，人物クラスのクラス間分散と服装クラスのクラス間分散を含まないので，d 番目の特徴が姿勢要素と相関がある場合にのみ大きくなる．一方で，$v_{\mathrm{c}}(d)$ には人物クラスのクラス間分散が含まれるので，d 番目の特徴が姿勢要素と人物要素に相関がある場合に大きくなる．同様に，$v_{\mathrm{p}}(d)$ は，d 番目の特徴が姿勢および服装の要素と相関がある場合に大きくなる．従って，これら 4 つの分散の値を比較することで，d 番目の特徴が 3 つの要素それぞれと相関があるかどうかを調べることができる．

以上の検証実験の結果を図 5.12 に示す．この図から，全分散が小さい特徴は存在しないことが分かる．これは，AE で抽出されたすべての潜在特徴が少なくとも 1 つの要素と相関がある，という仮説を実証している．特に，姿勢要素はすべての特徴量と相関がある．これは，歩容のシルエットの形状を変化させる最大の要因

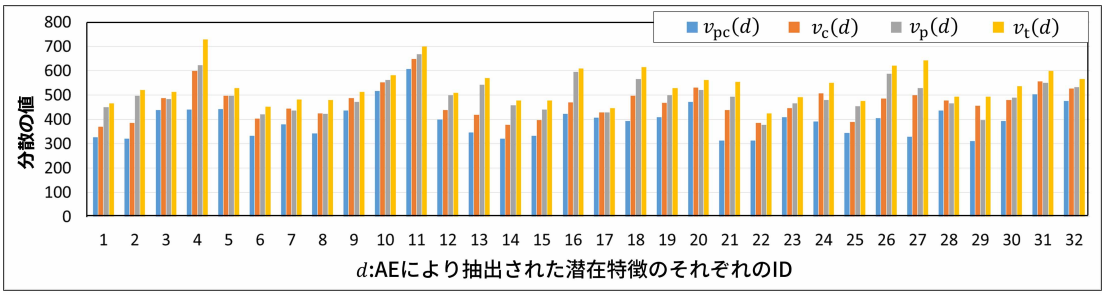


図 5.12: AE により抽出された潜在特徴に関する 4 種類の分散値

が姿勢の変化であるためである．また，ほとんどの特徴量において，$v_{\mathrm{p}}(d)$ は $v_{\mathrm{c}}(d)$ よりも大きいことが分かる．これは，例えばロングスカートとズボンなど，服装の違いが歩容シルエットの形状に大きな影響を与えるケースがあるためである．

図 5.12 に示す結果から，潜在特徴は以下のようにいくつかのグループに分類することができる．

- **潜在特徴 ID 3, 5, 9, 10, 11, 17, 20, 23, 28, 31, 32**

  これらの特徴では，$v_{\mathrm{pc}}(d)$ は $v_{\mathrm{t}}(d)$ よりもわずかに小さい．すなわち，$v_{\mathrm{pc}}(d) \approx v_{\mathrm{t}}(d)$ となっている．これは，姿勢要素としか相関がないことを意味する．

- **潜在特徴 ID 1, 2, 6, 12, 13, 14, 15, 16, 18, 19, 21, 25, 26**

  これらの特徴では，$v_{\mathrm{pc}}(d) < v_{\mathrm{p}}(d) \approx v_{\mathrm{t}}(d)$ となっている．これは，姿勢要素だけでなく，服装要素とも相関があることを意味する．

- **潜在特徴 ID 7, 22, 24, 29**

  これらの特徴では，$v_{\mathrm{c}}(d) > v_{\mathrm{p}}(d)$ となっている．これは，服装要素よりも人物要素と相関があることを意味する．

- **潜在特徴 ID 4, 8, 27, 30**

  これらの特徴量では，$v_{\mathrm{pc}}(d) < v_{\mathrm{c}}(d) \approx v_{\mathrm{p}}(d) < v_{\mathrm{t}}(d)$ となっている．これは，3 つの要素すべてと相関があることを意味する．

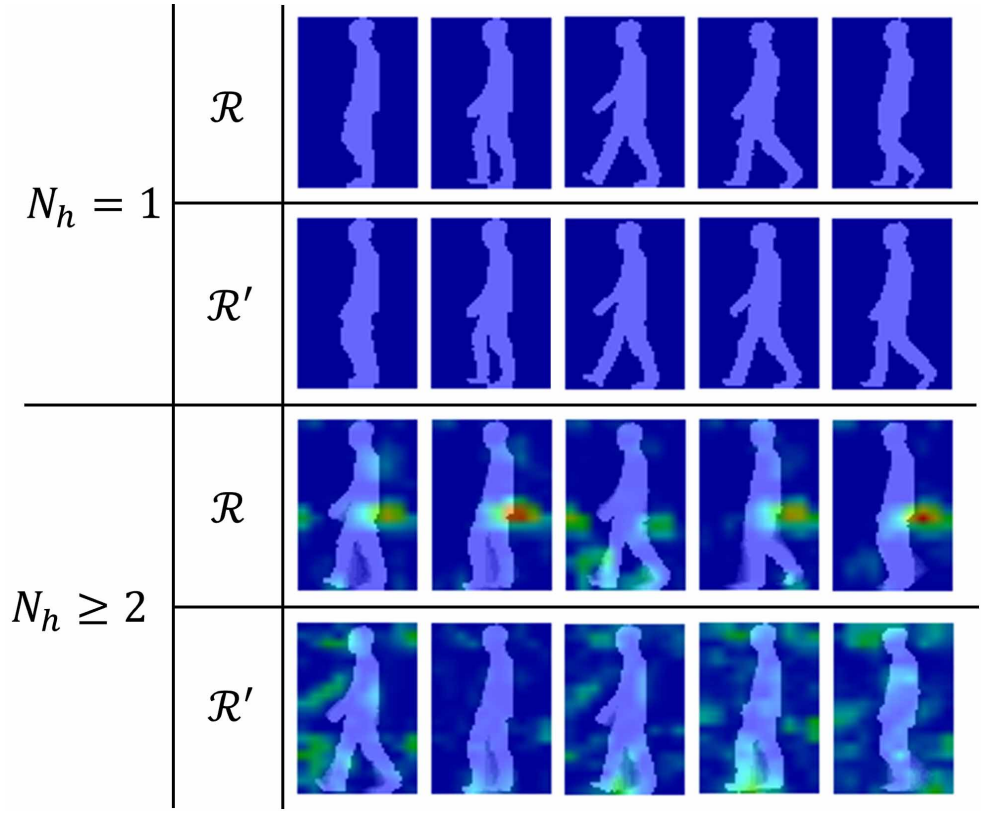


図 5.13: Grad-CAM に基づく視覚的評価結果の例

5.4.4 学習済み識別器の視覚的評価

歩容シルエットのどの部分がより GGS と GSC の識別に寄与しているかを調べるため，Grad-CAM [52] を用いて $\mathcal{R}$ と $\mathcal{R}'$ を視覚的に評価した．Grad-CAM は，評価対象の識別器に何らかの画像を入力した際，その識別に寄与していると判断された部分を強調表示する可視化手法である．なお，Grad-CAM の可視化の対象としては，図 5.3 に示したネットワークにおける最終畳み込み層を用いた．図 5.13 は，*Test-GGSs-Open* におけるいくつかのサンプルに対する可視化結果を示している．赤と緑の領域が識別に強く寄与していると判定された領域，青の領域は寄与していないと判定された領域である．

図 5.13 に見られるように，$N_{\mathrm{ch}} = 1$ の場合，$\mathcal{R}$ と $\mathcal{R}'$ の両方において，シルエット画像内の全領域が青色になっている．これは，識別器が GGS と GSC を全く識別できないことを意味する．一方，$N_{\mathrm{ch}} \geq 2$ の場合では，赤と緑の領域が存在して

おり，その特性は$\mathcal{R}$と$\mathcal{R}'$で異なっている．$\mathcal{R}'$の結果では，どの領域が赤か緑かは個々のサンプルに依存する．このため，$\mathcal{R}'$では，どのような歩容シルエット動画の真偽識別に対しても共通して有効となる領域を発見できていないと言える．一方で，$\mathcal{R}$の結果では，腰および右隣の領域は常に赤または緑となっており，$\mathcal{R}$はこれらの領域から共通して有用な情報を抽出することができる．このため，図 5.6，図 5.7，図 5.8 に見られるように，$\mathcal{R}$は$N_{\mathrm{ch}} \geq 2$の場合に$\mathcal{R}'$よりも高い識別性能を発揮することができる．

側面から撮影された歩容シルエットにおいては，腰部およびその近傍は上着の裾の形状を強く反映し，5.2.2 節で述べた 4 つの主要素だけでなく，風圧という弱い力によっても変化する可能性が考えられる．この力は時間とともに複雑に変化するため，GGS では腰部周辺のϵ^gは滑らかに変化しない．これは 5.4.2 節で述べた GSC のϵ^cの性質とは逆である．このため，腰部とその近傍領域が GGS と GSC の識別に共通して有用な領域として抽出されたものと考えられる．提案手法のようななりすまし防止手法を欺くためには，攻撃者は衣服（特に上衣）の詳細な形状や時間変化を模倣する必要がある．しかし，衣服に依存しない特徴は，歩容認証において望ましい特徴であり，しばしば採用されるため，歩容認証システムを欺くためには，衣服の形状や時間変化の模倣は重要とはならない．これらをまとめると，攻撃者はなりすまし防止手法を欺くには衣服の形状や時間変化の模倣に着目し，歩容認証を欺くには衣服に依存しない特徴の模倣に着目することとなる．従って，攻撃者が歩容認証システムとなりすまし防止技術の両方を同時に欺く場合にはこれらの両特徴に着目し，それらの模倣を行う必要があることから，どちらか一方を欺くよりもはるかに困難であると言える．このことは，提案手法のようななりすまし防止手法の有益性を示している．

5.5 結言

本章では，偽歩容動画の生成によるなりすまし攻撃のリスクが無視できない程度に存在するという第 4 章の結論を踏まえて，それを防御する仕組みとして，真正歩容シルエット動画と偽歩容シルエット動画を識別する手法を提案した．

提案手法の主眼は，識別器学習のための学習データセットとして，真偽以外の条件が統制された GGS と GSC の対を収集する点にある．この点について有効性を検証するため，条件の統制を考慮せずに GGS と GSC を別々に収集する比較手

法と提案手法を比較する実験を行った．その結果，提案手法は最大 94%という精度で比較手法の最大 62%の精度を大きく上回り，真正歩容シルエット動画と偽歩容シルエット動画を識別できることが判明した．このことから，識別器学習のためのデータセットにおいては，真偽以外の要素が統制されていることが重要であり，提案手法の有効性が確かめられた．また，そのような学習データセットの構築法として AE を利用すること，特に，GGS の集合から様々なサブセットを作成して複数の AE を学習し利用することが，識別精度の向上に有益であるという結果も得られた．さらに，提案手法は識別器学習に用いたデータセットとは別種のデータセットに対しても適用可能であり，汎用性の高い手法であることが確認できた．

第6章 結論

本論文では，真正な歩容情報の流通に伴うプライバシー情報流出，および，偽歩容情報の流通がもたらす「なりすまし」，という二つのリスクに対処する方法について論じた．近年のSNSの普及やマルチメディアデータ生成技術の進歩により，真正な生体情報とそれを模倣した偽の生体情報がともにWeb上で大量に流通する時代が訪れようとしている．生体情報の中でも歩容は，近年注目されるようになった比較的新しい情報であるため，プライバシー情報流出のリスクやなりすまし攻撃のリスクへの対策手法が十分に検討されていない．そこで，第2章において，歩容による個人認証，歩容情報の匿名化，歩容に対するなりすまし攻撃とその防御について，現在の研究動向を他の生体情報とも比較しながら議論し，具体的なリスクとして以下の攻撃の可能性が十分に起こり得ることを示した．第一に，Web上の動画に含まれる真正歩容情報から歩容認証技術により個人を同定し，その人物に紐づくプライバシー情報を取得する攻撃の危険性がある．第二に，マルチメディアデータ生成技術により特定個人の偽歩容情報を作成して当該個人になりすまし，それにより偽情報を拡散する攻撃の危険性がある．本論文では，これら二つの攻撃に対する防御手法について議論した．

第3章では，第一の攻撃に対する防御法として，現在主流の歩容認証手法であるシルエットベース認証を想定した歩容動画の匿名化手法について論じた．歩容シルエットから静的な特徴と動的な特徴を抽出し，それぞれに摂動を加えることにより，歩容シルエットを変形させる手法を提案した．さらに，変形前後のシルエット同士から計算される変位ベクトル場を元の歩容動画に対して適用することにより，変形後のシルエットへとテクスチャ情報を転写した．以上により，見た目の自然さを保持したまま歩容シルエットからの個人同定を困難化し，歩容情報の匿名化を実現した．実験では，歩容認証精度を100%から1.57%まで大きく低下させることに成功した．また，見た目の自然さについても，3D-ResNetによる動作認識精度が匿名化前後でほとんど変化しないことから（匿名化前：75.6%，匿名化後：73.0%），大きな劣化は起こらず，動作認識による尺度では自然な見た目が保持されることを確認した．このことから，Web動画のように視聴されることが前提の動画であってもユーザエクスペリエンスを下げず，かつ，個人の同定が困難

な匿名歩容動画を生成可能であるという結果を得た．

第 4 章では，第二の攻撃の実行可能性を具体的に検証するために，一枚の歩容画像のみから当該個人の偽歩容シルエット動画を生成する手法について論じた．一枚の歩容画像だけでは歩行リズムなどの動的な側面を完全には捉えきれず，歩容情報に一部欠落が生じる．そこで，歩容画像から姿勢に非依存な特徴量を抽出した上で，その特徴量に含まれる個人性を強調することにより欠落した情報を補完し，その後，その特徴量を歩容シルエット動画へと変換する手法を提案した．実験の結果，提案手法により生成した偽歩容動画では歩容認証精度が 78.0%となり，個人性強調前と比較して 10%程度の向上が確認された．このことから，第二の攻撃，すなわち，なりすまし攻撃のリスクが無視できない程度に存在するということが明らかになった．

第 5 章では，第 4 章で述べたなりすまし攻撃に対する防御法として，真正な歩容動画と偽の歩容動画を識別する手法について論じた．一般に，真正か偽かに起因する歩容シルエットの差異は，個人性や姿勢などに起因する差異よりも小さい．このため，真正歩容動画と偽歩容動画を独立に収集して識別器を機械学習する手法では，真正か偽かによる差異を正確に捉えることはできない．そこで，真偽以外の条件が統制された歩容動画の対をオートエンコーダにより多数作成し，これを学習データセットとして用いることにより，真正か偽かを正しく区別可能な識別器を学習する手法を提案した．複数のテストデータセットを対象に識別実験を行ったところ，最大 94%の精度を達成し，高い精度で真正歩容シルエットと偽歩容シルエットを識別できるという結果を得た．

以下に本論文の結論を述べる．本論文では，Web 上に存在する真正と偽の両方の歩容情報に対するリスクについて言及した．真正な歩容情報に対するリスクについては，既存の歩容認証手法 [11] において既に 90%以上の高い精度で認証が行えることから，それを悪用したプライバシー情報詐取のリスクが従前より指摘されてきた．一方で，偽の歩容情報がもたらすリスクについては，深層学習によるマルチメディアデータ生成技術の向上が目覚ましいにも関わらず，それを悪用したなりすましの可能性に関する議論は顔や声を対象としたもののみにとどまっており，歩容における同様のリスクはこれまで言及されてこなかった．第 4 章の内容は，偽歩容動画を用いたなりすまし攻撃について具体的に検討し，これが現実的に可能であることを明らかにしたものである．特に，一枚の真正歩容画像があれば偽歩容動画を生成できることを示した点には重大な意味がある．この事実は，

真正な歩容情報を含み得る画像をWeb上で取り扱うだけでも注意が必要であることを示唆する．以上のことから，Web上で歩容情報を安心な形で利用できるようにするためには，従来から指摘されてきたプライバシー情報詐取攻撃だけでなく，なりすまし攻撃のリスクにも備える必要があることが判明した．

上記二つのリスクのうち，プライバシー情報詐取攻撃のリスクに対しては，歩容の持つ静的な特徴と動的な特徴の双方に着目した新たな歩容動画匿名化手法を提案したことにより，従来より安全度の高い防御が実現された．提案手法は見た目の自然さを保持することも可能であるため，Web動画を視聴するユーザのユーザエクスペリエンスを下げずに，匿名化の恩恵のみを享受することができる．一方，偽歩容動画を用いたなりすまし攻撃のリスクに対しては，真正歩容動画と偽歩容動画を識別する手法を提案したことにより，偽歩容動画の生成自体は起こり得るとしても，それを事後的に検知することが可能となった．この手法は，特定の個人になりすました偽歩容動画に限らず，AEのような本研究で取り扱った今後も使用されるであろうマルチメディアデータ生成技術に対しては，なりすまし目的か否かに関係なく識別可能なため，今後類似した攻撃が行われた際にも適用することが可能である．さらに言えば，歩容以外の生体情報の真偽識別に対しても同様の仕組みが適用可能であり，汎用性の高い手法となっている．以上のことから，Web上で歩容情報を扱う際に存在する上記二種類のリスクについて，その各々に対処する手法を実現でき，歩容の取り扱いに関する安全性が向上したとの結論を得た．

今後の課題として，本論文では実際のWeb動画から個々の人物領域が切り出された後に防御処理を行うことを前提としたが，人物領域の切り出し前も考慮に入れて防御法を設計することも考えられ，そのための手法を検討することが挙げられる．例えば真正歩容情報からのプライバシー情報流出リスクに対しては，人物切り出し処理自体を困難化することも一つの方策であり，それを本論文で実現した手法と組み合わせることにより更なる安全性の向上が見込める可能性がある．他方で，人物領域切り出し処理の扱いは攻撃法にも影響を与え得る．例えば，人物領域ではなくその周囲の背景領域のテクスチャパターンをマルチメディアデータ生成技術により操作することにより，特定個人のシルエット形状に近い形で人物領域が切り出されるように誘導できる可能性も，今後の技術の進展次第では否定できない．そのような可能性も考慮の上でなりすまし対策を検討することも重要と考える．

また，本研究で述べた歩容情報という単一の生体情報についてのみ着目せず，他の複数の生体情報に関してもそのリスクについての検討が必要である．例えば，Web 動画におけるプライバシー情報詐取攻撃の場合，他の生体情報を匿名化せず歩容情報のみ匿名化するといった防御策では匿名化されていない生体情報から生体認証によって個人を同定でき，攻撃者にプライバシー情報を入手されてしまう．このため，Web 動画に存在する歩容以外の生体情報についても匿名化する必要がある．ここで，それぞれの生体情報を独立に匿名化した場合，Web 動画における見た目の自然さの低下やそれぞれの生体認証器に対する匿名化性能の低下などが生じる可能性がある．よって，独立に匿名化を検討するのではなく，それぞれの生体情報を複合的に匿名化することも重要な課題であると考えられる．これらの点についても実験的検討を加えることにより，生体情報のより安全な利活用に貢献できる．

謝 辞

本論文は，筆者が大阪大学大学院工学研究科電気電子情報通信工学専攻博士後期課程在学中において研究した成果をまとめたものである．大阪大学大学院工学研究科電気電子情報工学専攻博士前期課程，大阪大学大学院工学研究科電気電子情報通信工学専攻博士後期課程において研究を遂行する中で，多くの方々にご指導を賜った．ここにこれまでお世話になった方々に心からの感謝を申し上げる．

本論文の主査を務めていただいた，大阪大学大学院工学研究科 田中雄一教授が着任されたのは筆者が博士後期課程3年次の秋であり，短い期間ではあったが，研究を進める上で有益なご助言を頂き，懇切丁寧な指導を賜った．また，貴重なお時間を割いて本論文を確認して頂き，執筆上のアドバイスを多数いただいた．ここに深く感謝申し上げる．

本論文の審査委員会の委員である，大阪大学大学院工学研究科 丸田章博教授には研究について数々のご指導と懇切なる御助言を賜った．また，筆者が研究生活を続けていく上で重要となる生活面に関してお力添えいただいたこと，心より厚く御礼申し上げる．

本論文の審査委員会の委員である，大阪大学産業科学研究所 駒谷和範教授には自身では気づかぬ視点からの数々の執筆上の助言をいただいた．また，御多忙の中本論文を確認していただき，研究に関する議論を通じて有用な御助言を賜った．深厚な謝意を表する．

本論文の審査委員会の委員であり，約 6 年間にわたって直接指導していただいた，東京理科大学工学部 中村和晃准教授には研究生活のやり方，プログラミングの実装方法，研究者としての心構え，論文の執筆方法など多様な面から丁寧かつ論理的にご指導いただいたこと，心より深く御礼申し上げる．

福井工業大学環境情報学部 馬場口登教授には，2022 年 3 月に大阪大学大学院工学研究科をご退職されるまでの間，筆者の指導教員として大変お世話になった．また，充実した研究環境を用意していただき，研究活動に対する考え方や人生における自身の生き方について，ご指導ご鞭撻いただいたことに感謝の意を表する．

武庫川女子大学生活環境学部 新田直子教授には本論文における提案手法に対し，様々な側面から御助言，御指摘を賜ったこと，厚くお礼申し上げる．

関西大学社会安全学部 河野和宏准教授には提案手法を実装するにあたって，必要な技術について御助言を賜った．

大阪大学産業科学研究所 八木研究室には本論文の実験において使用するデータセットを提供して頂いた．深くお礼申し上げる．

最後に，長きにわたって筆者の人生を見守っていただいた両親に心より感謝する．

略語一覧

AE	Auto Encoder	自己符号化器
GEI	Gait Energy Image	
GEnI	Gait Entropy Image	
FDF	Frequency Domain Feature	
SFDEI	Signed Frame Difference Energy Image	
CNN	Convolutional Neural Network	畳み込みニューラルネットワーク
GAN	Generative Adversarial Networks	敵対的生成ネットワーク
VC	Voice Conversion	音声変換
TTS	Text-To-Speech	テキスト音声合成
DNN	Deep Neural Networks	深層ニューラルネットワーク
VAE	Variational Autoencoder	変分オートエンコーダ
HRTT	Human Region Texture Transfer	人物領域テクスチャ転写
GGS	Genuine Gait Silhouettes	真正歩容シルエット動画
GSC	Gait Silhouette Clones	偽歩容シルエット動画

参 考 文 献

1) B. Hesney and D. Citron: “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,” California Law Review, Vol.107, pp.1753–1820, 2019.

2) R. Delfino: “Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn’s Next Tragic Act,” Fordham Law Review, Vol.88, No.3 pp.887–938, 2019.

3) B. Dixon: “Deepfakes: More Frightening Than Photoshop on Steroids,” Judges Journal, Vol.58, pp.35–37, 2019.

4) F. Kreuk, Y. Adi, M. Cisse, and J. Keshet: “Fooling End-To-End Speaker Verification With Adversarial Examples,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp.1962–1966, 2018.

5) Y. Zhang, F. Jiang, and Z. Duan: “One-Class Learning Towards Synthetic Voice Spoofing Detection,” IEEE Signal Processing Letters, Vol.28, pp.937–941, 2021.

6) D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar: “Face Swapping: Automatically Replacing Faces in Photographs,” ACM Transactions on Graphics, Vol.27, No.3, 8 pages, 2008.

7) Q. Jianwei, D, Haohua, H. Jiahui, C. Linlin, J. Taeho and L. Xiang-Yang: “Speech Sanitizer: Speech Content Desensitization and Voice Anonymization,” IEEE Transactions on Dependable and Secure Computing, Vol.18, No.6, pp.2631–2642, 2019.

8) B. Chen, X. Liu, Y. Zheng, G. Zhao, and Y. Shi: “A Robust GAN-Generated Face Detection Method Based on Dual-Color Spaces and an Improved Xception,” IEEE Transactions on Circuits and Systems for Video Technology, Vol.32, No.6, pp.3527–3538, 2022.

9) S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui: "Voice Liveness Detection Algorithms Based on Pop Noise Caused by Human Breath for Automatic Speaker Verification," in Proceedings of 16th Annual Conference of the International Speech Communication Association, pp.239–243, 2015.

10) Y. Makihara, D. S. Matovski, M. S. Nixon, J. N. Carter, and Y. Yagi: "Gait Recognition: Databases, Representations, and Applications," John Wiley & Sons, Inc., pp.1–15, 2015.

11) K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi: "GEINet: View-Invariant Gait Recognition Using a Convolutional Neural Network," in Proceedings of International Conference on Biometrics, pp.1–8, 2016.

12) H. Chao, Y. He, J. Zhang, and J. Feng: "GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition," in Proceedings of the AAAI Conference on Artificial Intelligence, Vol.33, pp.8126–8133, 2019.

13) H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi: "Gait Verification System for Criminal Investigation," Information and Media Technologies, Vol.8, No.4, pp.1187–1199, 2013.

14) W. Zeng, C. Wang, and Y. Li: "Model-Based Human Gait Recognition via Deterministic Learning," Cognitive Computation, Vol.6, No.2, pp.218–229, 2014.

15) A. Kale, A. N. Rajagopalan, N. Cuntoor, and V. Kruger: "Gait-Based Recognition of Humans Using Continuous HMMs," in Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pp.336–341, 2002.

16) J. Man and B. Bhanu: "Individual Recognition Using Gait Energy Image," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.28, No.2, pp.316–322, 2006.

17) Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi: "Gait Recognition Using a View Transformation Model in the Frequency Domain," in Proceedings of European Conference on Computer Vision, pp.151–163, 2006.

18) K. Bashir, T. Xiang, and S. Gong: “Gait Recognition Using Gait Entropy Image,” in Proceedings of International Conference on Crime Detection and Prevention, pp.1–6, 2009.

19) N. Setoguchi, K. Nakashima, W. Tong, Y. Iwashita, and R. Kurazume: “A CNN-based Gait Recognition Robust to Low Resolution Images Using Inter-Image Difference,” in Proccedings of 14th Joint Workshop on Machine Perception and Robotics, 2018.

20) 鈴木温之, 村松大吾, 槇原靖, 八木康史: “輝度値の共起に対する計量学習による荷物所持に頑健な歩容認証,” 情報処理学会研究報告, コンピュータビジョンとイメージメディア, 2016-CVIM-203，pp.1–8, 2016

21) 馬場口登: “プライバシーを考慮した映像サーベイランス,” 情報処理, 特集 安全と安心のための画像処理技術, Vol. 48, No. 1, pp. 30–36, 2007.

22) K. Chinomi, N. Nitta, Y. Ito, and N. Babaguchi: “PriSurv: Privacy Protected Video Surveillance System Using Adaptive Visual Abstraction,” in Proceedings of the 14th International Conference on MultiMedia Modeling, pp. 144–154, 2008.

23) M. Boyle, C. Edwards, and S. Greenberg: “The Effects of Filtered Video on Awareness and Privacy,” in Proceedings of ACM Conference on Computer Supported Cooperative Work, pp.1–10, 2000.

24) C. Neustaedter, S. Greenberg, and M. Boyle: “Blur Filtration Fails to Preserve Privacy for Home-Based Video Conferencing,” ACM Transactions on Computer-Human Interaction, Vol.13, No.1, pp.1–36, 2006.

25) E. Newton, L. Sweeney, and B. Malin: “Preserving Privacy by De-Identifying Face Images,” IEEE Transactions on Knowledge and Data Engineering, Vol.17, No.2, pp.232–243, 2005.

26) R. Gross, E. Airoldi, B. Malin, and L. Sweeney: “Integrating Utility into Face De-identification,” in Proceedings of the 5th International Conference on Privacy Enhancing Technologies, pp.227–242, 2005.

27) R. Gross, L. Sweeney, F. de la Torre, and S. Baker: "Model-Based Face De-identification," in Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop, p. 161, 2006.

28) Y. Nakashima, T. Koyama, N. Yokoya, and N. Babaguchi: "Facial Expression Preserving Privacy Protection Using Image Melding," in Proceedings of IEEE International Conference on Multimedia and Expo, pp.1–6, 2015.

29) P. Agrawal and P. J. Narayanan: "Person De-Identification in Videos," IEEE Transactions on Circuits and Systems for Video Technology, Vol.21, No.3, pp.299–310, 2011.

30) I. Mitsugami, M. Mukunoki, Y. Kawanishi, H. Hattori, and M. Minoh: "Privacy-Protected Camera for the Sensing Web," in Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp.622–631, 2010.

31) N. T. Tieu, H. H. Nguyen, H. Nguyen-Son, J. Yamagishi, and I. Echizen: "An Approach for Gait Anonymization Using Deep Learning," in Proceedings of IEEE Workshop on Information Forensics and Security, 6 pages, 2017.

32) I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio: "Generative Adversarial Nets," in Proceedings of the 27th International Conference on Neural Information Processing Systems, pp.2672–2680, 2014.

33) N. T. Tieu, H. H. Nguyen, H. Nguyen-Son, J. Yamagishi, and I. Echizen: "Spatio-Temporal Generative Adversarial Network for Gait Anonymization," Journal of Information Security and Applications, No.46, pp.307–319, 2019.

34) N. T. Tieu, H. H. Nguyen, H. Nguyen-Son, J. Yamagishi, and I. Echizen: "An RGB Gait Anonymization Model for Low Quality Silhouette," in Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp.1686–1693, 2019.

35) D. Muramatsu, Y. Makihara, and Y. Yagi: “Gait Regeneration for Recognition,” in Proceedings of International Conference on Biometrics, pp.169–176, 2015.

36) K. Patel, H. Han, and A. K. Jain: “Secure Face Unlock: Spoof Detection on Smartphones,” IEEE Transactions on Information Forensics and Security, Vol.11, No.10, pp.2268–2283, 2016.

37) A. Anjos, M. M. Chakka, and S. Marcel: “Motion-Based Counter-Measures to Photo Attacks in Face Recognition,” IET Biometrics, Vol. 3, No.3, pp.147–158, 2014.

38) P. Cheng and U. Roedig: “Personal Voice Assistant Security and Privacy — A Survey,” Proceedings of the IEEE, Vol.110, No.4, pp.476–507, 2022.

39) S. Kumar, S. Singh, and J. Kumar: “A Comparative Study on Face Spoofing Attacks,” in Proceedings of International Conference on Computing, Communication and Automation, pp.1104–1108, 2017.

40) M. Toshpulatov, W. Lee, and S. Lee: “Generative Adversarial Networks and Their Application to 3D Face Generation: A Survey,” Image and Vision Computing, Vol.108, 18 pages, 2021.

41) J. Galbally and R. Satta: “Three-Dimensional and Two-and-a-Half-Dimensional Face Recognition Spoofing Using Three-Dimensional Printed Models,” IET Biometrics, Vol.5, No.2, pp.83–91, 2015.

42) U. Muhammad, Y. Zitong, and J. Komulainen: “Self-Supervised 2D Face Presentation Attack Detection via Temporal Sequence Sampling,” Pattern Recognition Letters, Vol.156, pp.15–22, 2022.

43) Q. Wang, X. Lin, M. Zhou, Y. Chen, C. Wang, Q. Li, and X. Luo: “VoicePop: A Pop Noise based Anti-spoofing System for Voice Authentication on Smartphones,” in Proceeding of the IEEE Conference on Computer Communications, pp.2062–2070, 2019.

44) J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang: "Generative Image Inpainting with Contextual Attention," in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp.5505–5514, 2018.

45) H. Nguyen, H. Nguyen-Son, T. Nguyen, and I. Echizen: "Discriminating Between Computer-Generated Facial Images and Natural Ones Using Smoothness Property and Local Entropy," in Proceeding of 14th International Workshop on Digital Forensics and Watermarking, pp.39–50, 2015.

46) Y. Moon, J. Chen, K. Chan, K. So, and K. Woo: "Wavelet Based Fingerprint Liveness Detection," Electronics Letters, Vol.41, No.20, pp.1112–1113, 2005.

47) Y. Makihara, H. Mannami, A. Tsuji, M. A. Hossain, K. Sugiura, A. Mori, and Y. Yagi: "The OU-ISIR Gait Database Comprising the Treadmill Dataset," IPSJ Transactions on Computer Vision and Applications, Vol.4, pp.53–62, 2012.

48) L. Ma, Q. Sun, S. Georgoulis, L. V. Gool, B. Schiele, and M. Fritz: "Disentangled Person Image Generation," in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp.99–108, 2018.

49) O. Gafni, O. Ashual, and L. Wolf: "Single-Shot Freestyle Dance Reenactment," in Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp.882–891, 2021.

50) K. Hara, H. Kataoka, and Y. Satoh: "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp.6546–6555, 2018.

51) J. Redmon, S. Divvala, R. Girshick, and A. Farhadi: "You Only Look Once: Unified, Real-Time Object Detection," in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp.779–788, 2016.

52) R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra: "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proceeding of IEEE International Conference on Computer Vision, pp.618–626, 2017.

53) 馬場口登: “メディアクローン攻撃を防御するコミュニケーション系 —構想・チャレンジ・アプローチ—,” 電子情報通信学会技術研究報告, Vol.116, No.497, CQ2016-115, pp.25–30, 2017.

54) M. Une, A. Otsuka, and H. Imai: “Wolf Attack Probability: A New Security Measure in Biometric Authentication Systems,” in Proceedings of International Conference on Biometrics, pp.396–406, 2007.

55) T. Ohki, S. Hidano, and T. Takehisa: “Evaluation of Wolf Attack for Classified Target on Speaker Verification Systems,” in Proceedings of 12th International Conference on Control Automation Robotics and Vision, pp.182–187, 2012.

56) H. H. Nguyen, J. Yamagishi, I. Echizen, and S. Marcel: “Generating Master Faces for Use in Performing Wolf Attacks on Face Recognition Systems,” in Proceedings of IEEE International Joint Conference on Biometrics, 10 pages, 2020.

57) D. P. Kingma and M. A. Welling: “Auto-Encoding Variational Bayes,” in Proceedings of the International Conference on Learning Representations, 14 pages, 2014.

本論文に関する原著論文

A. 学術論文

1. Y. Hirose, K. Nakamura, N. Nitta, and N. Babaguchi: “Anonymization of Human Gait in Video Based on Silhouette Deformation and Texture Transfer,” IEEE Transactions on Information Forensics and Security, Vol.17, pp.3375–3390, 2022.

2. Y. Hirose, K. Nakamura, N. Nitta, and N. Babaguchi: “Discrimination between Genuine and Cloned Gait Silhouette Videos via Autoencoder-based Training Data Generation,” IEICE Transactions on Information and Systems, Vol.E102-D, No.12, pp.2535–2546, 2019.

B. 国際会議論文

1. Y. Hirose, K. Nakamura, N. Nitta, and N. Babaguchi: “An Experimental Consideration on Gait Spoofing,” International Conference on Computer Vision Theory and Applications, 8 pages, 2023.

2. Y. Hirose, K. Nakamura, N. Nitta, and N. Babaguchi: “Anonymization of Gait Silhouette Video by Perturbing Its Phase and Shape Components,” Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 7 pages, 2019.

C. 口頭発表

1. 廣瀬雄基, 中村和晃, 新田直子, 馬場口登: “シルエット変化に基づく歩容動画の匿名化,” 第 23 回画像の認識・理解シンポジウム, 4 pages, 2020.

2. 廣瀬雄基, 中村和晃, 新田直子, 馬場口登: “歩容情報保護のための歩容シルエット動画の匿名化,” 第 21 回画像の認識・理解シンポジウム, 4 pages, 2018.

3. 廣瀬雄基, 中村和晃, 新田直子, 馬場口登: “映像中の歩容情報保護を目的とした匿名歩容シルエットの生成,” 電子情報通信学会 2018 年総合大会, p.42, 2018.