NAIST-IS-MT1551073

修士論文

条件付き確率場とディープニューラルネットワークの組み合わせによる映像中の重要人物識別

西田 篤史

2017 年 3 月 16 日

奈良先端科学技術大学院大学
情報科学研究科

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士 (工学) 授与の要件として提出した修士論文である。

西田 篤史

審査委員：

横矢 直和 教授	（主指導教員）
萩田 紀博 教授	（副指導教員）
佐藤 智和 准教授	（副指導教員）
中島 悠太 客員准教授	（副指導教員/大阪大学）

条件付き確率場とディープニューラルネットワークの組み合わせによる映像中の重要人物識別*

西田 篤史

内容梗概

映像中や画像中の重要領域推定は，小さな画面に合わせて映像の一部を拡大して表示するビデオリターゲティングや映像のコンテンツに応じた圧縮など，広範な応用を持つ．重要領域推定は盛んに研究されており，生物の視覚システムが持つ生物学的な特徴をモデル化した視覚的顕著モデルや，人間は人の顔に注目するという性質に基づいて顔検出を援用するモデルなどが提案されている．顔検出を援用した重要領域推定は前述の応用において有用であると考えられる一方で，偶然通りかかった人物とその映像中において主要な人物を区別することができないという問題があった．

そこで本研究では，複数の人物を含むシーンにおいて，映像中の人物がその映像に必要な重要人物なのか，偶然映り込んだ非重要人物なのかを識別する手法を提案する．一般に，映像中の人物が重要か，非重要かは視聴者によって異なり，一意に決定することはできない．そこで，本研究では，その映像の撮影者の観点から重要人物，非重要人物を区別する．視聴者は撮影者の意図を汲み取ろうとすることから，多くの場合，撮影者，視聴者それぞれにとっての重要人物は一致するものと考えられる．

撮影者は重要人物を撮影する際に，その人物を映像フレーム中の中心付近に配置するように，撮影時のカメラの動かし方に一定の傾向があるものと考えられる．そこで，提案手法では，このようなカメラの動きが反映されると考えられる顔領

*奈良先端科学技術大学院大学 情報科学研究科 修士論文, NAIST-IS-MT1551073, 2017 年 3 月 16 日.

域の大きさ，および軌跡を人物の動きの特徴量として用いる．加えて，顔の向きなど見え方も重要人物の識別において有効であると考え，人物の見え方に関する特徴量として用いる．また，識別には条件付き確率場とディープニューラルネットワークを組み合わせたモデルを利用し，画面中の人物間の位置関係を考慮することで複数の人物を含むシーンでの識別精度の向上を試みる．実験では，ウェブ上で収集したホームビデオを用いてネットワークを学習し，80%を超える精度で重要人物識別が可能であることを示す．また，提案モデルをサポートベクターマシンや条件付き確率場を用いないネットワークと比較することで提案モデルの有効性および条件付き確率場の効果を実験により検証した．

キーワード

ニューラルネットワーク， 条件付き確率場，重要人物推定

Finding Important People in a Video using a Deep Neural Network with Conditional Random Field*

Atsushi Nishida

Abstract

Finding important regions is essential for applications like content-aware video compression and video retargeting, which automatically crops an important region in a video for small screens. Various models for important region estimation have been proposed. Since people are one of the main content of videos, some methods for finding important regions use face detection. However, those existing methods usually do not distinguish important people from passers-by in a video.

This thesis proposes a method to classify people in a video frame into important or non-important ones. Generally, this classification problem is not well designed because who is important or not may differ viewer by viewer. Therefore, instead of the viewers perspective, we use videographers perspective. That is, our method finds people who are important for the videographer. Since viewers try to understand what the videographer wants to express in the video, important people for viewers and videographers may highly correlate. It is considered that videographers have a certain tendency in, e.g, how to move the camera when taking the video, such as placing important people near the center of the video frame. Since videographers' such behavior is reflected in the trajectories and sizes of face regions, we use them as features for the classification. In addition,

*Master's Thesis, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1551073, March 16, 2017.

as visual cues like the orientation of faces are helpful for important person classification, the proposed method exploits visual features such as color histograms. The proposed method uses a conditional random field (CRF) built upon a deep neural network (DNN), which can capture the various types of relationships, such as spatial one, among people in a video frame in order to facilitate the classification. Experimental results demonstrate that our models trained on a dataset of user-generated videos achieve the accuracy of over 80%. Our experiments also verify the effectiveness of the proposed model and the effect of the conditional random field by comparing our model with baselines, such as a support vector machines and a DNN without a CRF.

Keywords:

Neural network, Conditional random field, Important people classification

目 次

図 目 次

表 目 次

1. はじめに

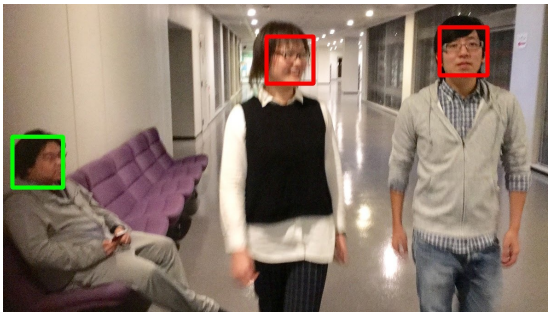
重要領域推定とは，画像中や映像中において視聴者が注目する領域を推定することである．画像や映像中から重要な領域を推定する技術は，重要な領域が変形しないように画像サイズを変更するビデオリターゲティング [3,4] や，映像の各領域の重要度に応じて圧縮率を変えるコンテンツに応じた映像圧縮 [5–8] など，広範な応用がある．重要領域の定義は手法の目的によって異なる．

重要領域推定は盛んに研究されており，画像の輝度や，色相などの低レベルな特徴量を用いる手法 [1,9,10] とオブジェクトから顕著性を推定するなどの高レベルの特徴量を用いる手法 [2,11,12] に分類できる．前者は，コントラストの強い箇所や画像中央に注目しやすいという視覚特性に基づく指標を用いて重要領域を推定する手法であり，その代表的な研究として Itti ら [1] によるものが挙げられる．Itti ら [1] は動物の視覚特性に基づき，色やコントラストなど視覚細胞が反応しやすい低レベル特徴量を組み合わせて重要度を算出する視覚的注意モデルを提案した．一方，後者の手法では，重要なオブジェクトが重要領域と一致するという考えに基づき重要領域を推定する．Yang [2] は画像をパッチとよばれる小領域に分割し，各パッチごとに事前に決められたオブジェクトの有無を推定することで，重要領域を推定する．また，Ma ら [11] は人物の顔は重要領域になりやすいという考えから人物の顔を検出し，その顔の大きさや位置から重要度を算出する．

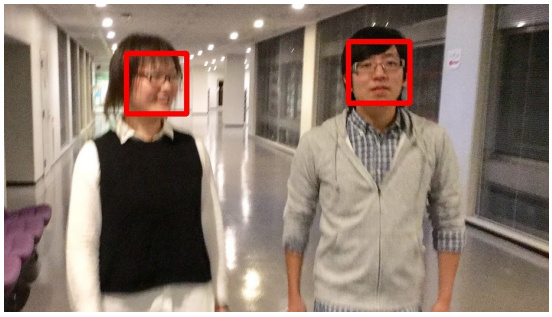
多くの映像は人物を撮影したものであるため，そのような映像では，Ma ら [11] のような人物の顔に基づく重要領域推定が効果的である．しかし，従来のすべての人物を重要領域とする手法は，その人物が映像中において重要かどうかを考慮していないため，複数の人物を含む映像においては偶然映り込んだ人物も重要領域に含む場合がある．例えば，図 1 のような複数の人物が映っている映像の場合，左下の人物のように偶然映り込んだ人物の重要度は低く，画面中央に映っている 2 人は重要度が高いと考えられる．

ここで，この映像にリターゲティングを施すとする．図 2(a) は全ての人物を重要領域とした場合の図 1 のリターゲティング処理例である．一方，図 2(b) は映像中の人物の重要度を考慮したリターゲティング処理例である．このように，図 1 における左下の人物のような重要でない人物が重要領域に含まれると，リターゲ

図 1: 重要人物と非重要人物の例



(a) 全ての人物を重要領域と考えた場合



(b) 人物の重要度を考慮した場合

図 2: 図 1 のリターゲティング処理例

ティングのようなアプリケーションの性能が損なわれる場合がある．

本研究では，このような複数の人物を撮影した映像から重要人物だけを含む重要領域を抽出するために，映像中の人物の重要度推定に取り組む．具体的には，映像中から検出した人物をそれぞれが映像中において重要な人物か，あるいは偶然写り込んだ非重要人物かを判定する識別器を開発する．この識別結果を用いて非重要人物の領域を重要領域の候補から除去することにより，非重要人物を含まない重要領域推定が可能となる．

一般に，映像中の人物が重要か，非重要かは視聴者によって異なり，一意に決

定することはできない．そこで，本研究では，その映像の撮影者の観点から重要人物，非重要人物を区別する．本研究において，重要人物とは撮影者が撮影したい人物のことであり，非重要人物は撮影者の意図と異なり偶然映り込んだ人物である．

映像中の人物の動きとその見え方には，その人物の映像における重要度が反映される．例えば，一般的に撮影者は撮影したい人物を画面の中央に大きく配置する．また，重要人物に対して正面，あるいは顔が見える位置から撮影することが多い．そこで，本研究では映像中の人物をトラッキングした結果得られる軌跡と顔領域の視覚特徴量を人物の重要度識別に利用する．

また，同じような動き，見え方を持つ人物は同程度の重要度を示す可能性が高い．例えば図 1 では，重要人物の顔領域の矩形と人物の軌跡を赤色，非重要人物を緑色で表している．図 1 の重要人物は並んで歩いているため，人物の動きや見え方は類似している．一方，非重要人物は端で座っているため，動きや見え方が重要人物とは異なっている．そこで，提案手法ではこのような人物の特徴間の相関関係を考慮した識別モデルを提案する．

本研究では重要人物識別のため，様々な画像識別タスクで高い性能を発揮しているディープニューラルネットワーク (Deep Neural Network: DNN) を用いて識別器を構築する．また提案手法では複数の人物の相関とその重要度をモデル化するため，条件付き確率場 (Conditional Random Fields: CRF) を取り入れたモデルを設計する．CRF は機械学習におけるモデルの一種であり，設計する事後確率が最大になるようにパラメータを学習する．事後確率を計算する際のエネルギーとして複数の特徴間の相関関係を表現した関数があり，この関数を用いることで，人物同士の相関関係をモデル化することができる．具体的には，DNN の出力に CRF を組み合わせ，End-to-End で学習を行い，同じフレーム内の人物同士の特徴量から識別結果を算出する．

実験では，YouTube 映像のデータセットを用いて提案モデルを学習し，人物の重要，非重要のラベルを持つホームビデオのデータセットを使って識別精度を評価した．提案手法は，映像中の人物の重要度推定が高い精度で可能であることを示した．

本論文は，2 章で重要領域推定における従来研究，CRF と DNN を用いた関連研究，および本研究の位置付けについて述べる．3 章では本論文の提案手法である CRF と DNN を用いた重要人物識別について述べる．4 章では提案モデルを従来モデルと比較するための実験と結果について述べる．最後に，5 章でまとめ及び今後の展望について述べる．

2. 関連研究および本研究の位置付け

本章では関連研究として，重要領域推定に関する従来手法と，提案手法で用いるディープニューラルネットワークおよび条件付き確率場に関する関連研究を概観し，本研究の位置付けについて述べる．

2.1 重要領域推定に関する研究

画像や映像の重要領域推定は広範な応用を持つ技術であるが，どのような領域を重要とするべきかはそれぞれのアプリケーションに依存しており，明確に決定されるものではない．そのため，これまで画像中や映像中において重要な領域を推定するための手法は数多く提案されている．これらの手法の重要領域の定義は手法の目的によって異なる．

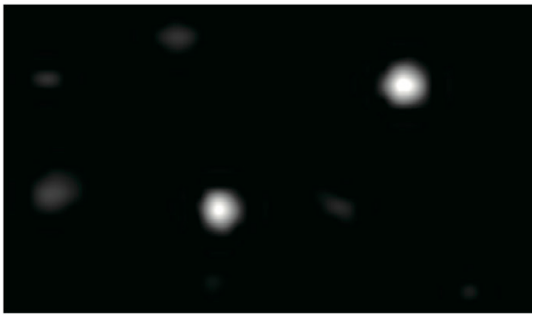
重要領域推定を行うために，視覚的注意モデルを用いた手法が多数提案されている [1,9,10,13]．これらの手法は，人はコントラストの強い箇所や画像中央に注目しやすいという視覚特性に基づいて設計されている．Itti ら [1] は視覚システムが持つ生物学的な特徴に基づいて，画像の輝度，色相，エッジの向きなどの変化に強く反応する性質を模倣する視覚的注意モデルを構築した．また，Itti ら，および Baldi ら [9,10] は，映像中で予測できないような変化をする領域を重要領域と考える Bayesian Surprise モデルを提案した．Achanta ら [13,14] は，Lab 表色系において，入力画像を平滑化し，入力画像の平均画素値との差分を重要度と考えることで，視覚的注意モデルを簡素化したモデルを提案した．図 3(b) は図 3(a) の重要領域を可視化した画像である．しかし，特に映像においては，例えば図 3(c) のような人物を対象とした映像に対して，視聴者は中央付近の人物が重要領域と考えると予想される．一方で、Itti ら [1] の手法は図 3(d) で示すようなコントラストの高い領域を重要領域として推定しており，人物が重要領域に含まれていない．そのため，視聴者の想定とは異なる領域が重要領域として推定される．

一方で，オブジェクトから重要領域を推定するアプローチも提案されている．この手法は，画像あるいは映像中で重要なオブジェクトの占める領域が重要領域と一致するという考えに基づいている．例えば，Yang [2] は画像をパッチに分割

(a) 入力画像

(b) 重要度マップ

(c) 入力画像



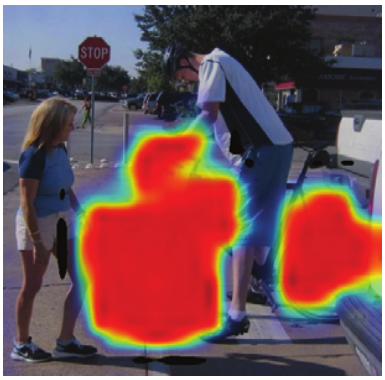
(d) 重要度マップ

図 3: Itti ら [1] の手法による重要領域推定

し，各パッチごとに事前に決められたオブジェクトの有無を推定することで，重要領域を推定する．図 4 は物体らしさを学習したモデルを用いて，それぞれ自転車，車，人物の重要度を高くした重要領域推定結果であり，それらの物体が存在している領域は，重要度が高く算出されている．Ma ら [11] は，多くの映像は人物を撮影したものであり，そのような映像では人物の顔付近が重要領域となる傾向に着目して，検出した人物の顔を重要領域の候補として用いる手法を提案した．また，顔を対象とした重要領域推定では，肌色が検出された領域を重要領域として用いる手法も提案されている [12]．

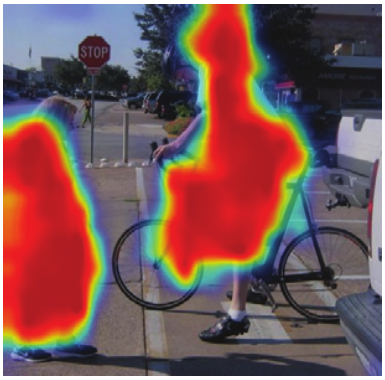
多くの映像は人物を撮影したものであるため，そのような映像では，Ma ら [11] のような人物の顔に基づく重要領域推定が効果的である．しかし，従来の人物の

(a) 入力画像



(b) 自転車を重視した重要度マップ

(c) 車を重視した重要度マップ



(d) 人物を重視した重要度マップ

図 4: Yang ら [2] の手法による重要領域推定

有無を重要領域推定の指標として用いた手法は，その人物の映像中における重要度を考慮しないため，複数の人物を含む映像において重要度の低い人物も重要領域に含む場合がある．重要でない人物が重要領域に含まれると，リターゲティングのようなアプリケーションの性能が損なわれる場合がある．

このような課題を解決するため，Nakashima ら [15] は撮影者の観点に基づき，複数の人物を含む映像の重要人物を識別をする手法を提案した．Nakashima らは，同じフレーム中の重要人物同士は大きさや動きの軌跡に相関があるという考えと，重要人物や非重要人物は短い期間では入れ替わらないという考えのもとに，条件付き確率場を用いたモデルを採用した．本論文では，さらなる精度向上のため，Nakashima ら [15, 16] の手法を拡張し，CRF を取り入れた DNN を用いた識別手

法を提案する．次節では，識別においてCRFとDNNを用いた関連研究について述べる．

2.2 条件付き確率場とディープニューラルネットワークに関する研究

CRFは，マルコフ確率場 (Markov Random Field: MRF) と呼ばれる無向性のグラフィカルモデルの一種であり，入力 x と出力 y が共に構造をもつ条件付き確率分布を表現するモデルである．多くの場合CRFの条件付き確率 $p(y|x)$ は次の形で表現される．

$$p(y|x) = \frac{1}{Z}e^{-E(y,x)} \tag{1}$$

$$E(y,x) = \sum_i f_i(x_i|y) + \sum_{ij} f_{ij}(x_i,x_j|y) \tag{2}$$

ここで，$E(y,x)$ はエネルギー関数と呼ばれる関数であり，多くの場合，入力データ x_i で定まる関数 $f_i(x_i|y)$ と，入力データ x_i, x_j 間の相関関係を表現した関数 $f_{ij}(x_i,x_j|y)$ の和で構成される．Z は分配関数 (Partition function) と呼ばれる正規化定数である．

多くのDNNを用いた識別手法は，特徴間の相関関係を考慮しない．そこで，DNNとCRFを組み合わせることで，識別精度の向上を試みる手法が近年，多数提案されている [17–23]．例えば，Bengioら [17] は手書き文字認識の推定精度を改善するため，畳み込みニューラルネットワーク (Convolutional Neural Networks: CNN) の出力信号を隠れマルコフモデルの入力信号とすることで両者の利点を取り入れたモデルを提案した．自然言語処理の分野では，Yaoら，およびWangら [18, 19] は文章の品詞タグ付け問題の推定精度を，リカレントニューラルネットワークとCRFを組み合わせで改善されることを確認した．また，Maら [24] はLong Short Term MemoryとCRFの組み合わせることで，文章の品詞タグ付け問題の推定精度が改善することを確認した．コンピュータビジョンの分野では，CRFとCNNの組み合わせは，領域分割 [20–23] や人物の姿勢推定 [25]，深度推定 [26] の性能を向上させている．Arnabら [22] は，CNNから得られた特徴量と，物体検出結果とスーパーピクセルによるエネルギーとペアワイズ項の4つのエネ

ルギーを利用し，CRF を用いてピクセル単位の領域分割を提案した．Farabet [27] らは車載画像の領域分割を，CNN と CRF の組み合わせで解決するための手法を提案した．具体的には，スーパーピクセル毎の特徴ベクトルを CNN を用いて抽出し，次にスーパーピクセル間の相関関係を CRF で記述し，推定解を求めた．また，Liu ら [26] は，単一の画像を入力とした深度推定の問題を CNN と CRF の組み合わせで解決する方法を提案した．また，Chanra ら [23] は，領域分割において，CRF と DNN を組み合わせたネットワークを End-to-End で学習する手法を提案した．

しかし，これらの手法の多くは，識別するクラスの数が多く，また入力データが多くなるほど CRF の計算量は膨大になる．そのため，CRF における計算は，Contrastive Divergence 法 [28] などの近似手法を用いる場合が多い．

2.3 本研究の位置付け

本研究では，人物の有無に基づく重要領域推定を改良するための要素技術として人物の重要度推定に取り組む．さらなる精度向上のため，最近，画像識別のタスクで高い性能を発揮している DNN を用いる．これにより，Nakashima ら [15,16] の手法を拡張し，映像中の人物が重要な人物か，あるいは偶然映り込んだ非重要人物か識別する手法を提案する．本研究では，CRF を取り入れた DNN 識別モデルを提案する．CRF を取り入れることで，映像中の複数の人物の相関関係を考慮した識別が可能となり，識別精度の向上が期待できる．

一般に，CRF を用いた手法は計算量が膨大なため，近似手法を用いることが多い．しかし，本研究では，識別するクラスが 2 つと少なく，また映像中に映っている人物の数も限られている．そこで，本手法は近似手法を用いずにモデルを最適化する．この時，途中の計算結果を再利用することで，計算量の抑制が可能であることを示す．

3. 条件付き確率場とニューラルネットワークを用いた重要人物識別

3.1 提案手法の概要

本研究の目的は人物を撮影した映像から，その映像中の各人物の重要度を推定することである．これを実現するため，映像中の人物を重要，あるいは非重要に識別する識別器を構築する．図 5 に提案手法の概要を示す．提案手法ではまず，映像中から人物を検出する．次に検出した各人物の顔領域を追跡し，人物の動きの特徴量を抽出する．加えて，提案手法は見えの特徴量として顔領域の画像特徴を抽出する．こうして得られた特徴量を入力に，提案する識別器は映像中の各人物について重要あるいは非重要のクラスラベルを出力する．

本研究では識別モデルとして CRF を用いた DNN を構築する．映像中の人物が同じような動きや見えの特徴を持つ場合，同程度の重要度を示す可能性が高い．そこで，提案する識別モデルでは，人物から抽出された特徴量の相関関係を考慮するため CRF を取り入れる．

以下，3.2 節では重要人物識別のための特徴量抽出，3.3 節では条件付き確率場とニューラルネットワークによる重要人物の識別，そして 3.4 節ではネットワークの学習方法について述べる．

3.2 重要人物識別のための特徴量抽出

提案手法では，まず映像中から人物を検出し，各人物の動きと見え方に関する特徴量を抽出する．ここで，本研究の関心は人物検出の精度ではなく，重要人物識別である．そこで，本研究では映像中の人物検出は実現されたものとし，人手で付与された顔領域のバウンディングボックスを人物検出結果とした．以下，人物の動きの特徴量と人物の見えの特徴量について詳述する．

人物の動きの特徴量

一般的に，撮影者は重要人物を画面の中央に大きく配置するなど，要人物に関して，構図やカメラの動きには特有の傾向があると考えられる．そこで，本研究で

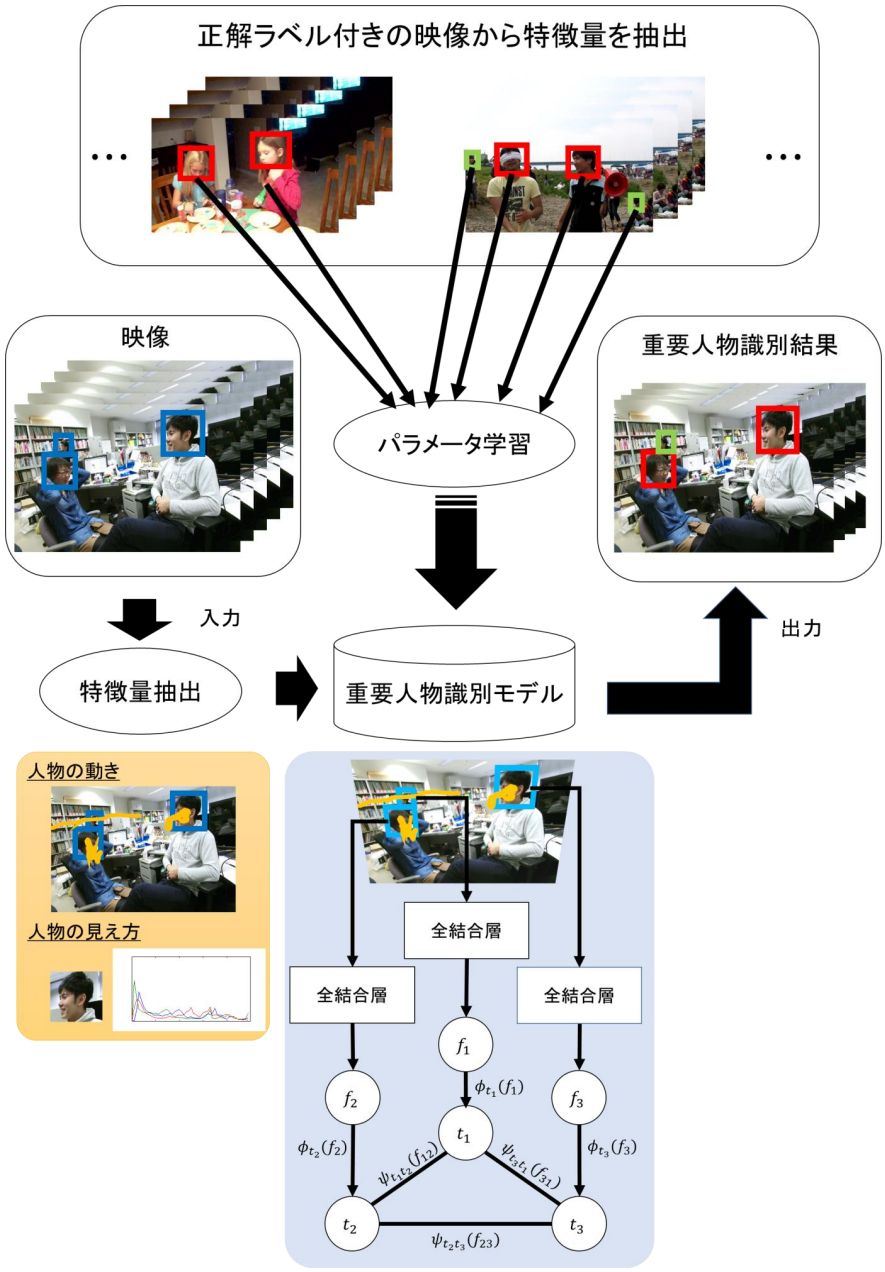
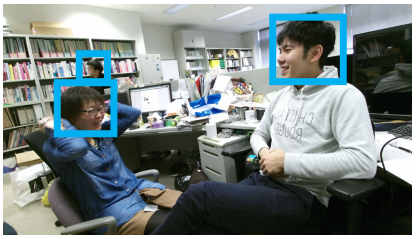

正解ラベル付きの映像から特徴量を抽出
映像
パラメータ学習
重要人物識別結果
入力
出力
特徴量抽出
重要人物識別モデル
人物の動き
人物の見え方
全結合層
全結合層
全結合層
f_1
f_2
f_3
t_1
t_2
t_3
$\phi_{t_1}(f_1)$
$\phi_{t_2}(f_2)$
$\phi_{t_3}(f_3)$
$\psi_{t_1t_2}(f_{12})$
$\psi_{t_3t_1}(f_{31})$
$\psi_{t_2t_3}(f_{23})$

図 5: 提案手法の概要

(a) 注目フレームから 100 フレーム前



(b) 注目フレーム

(c) 注目フレームから 100 フレーム後

(d) トラッキングから得られた人物の軌跡

図 6: トラッキングの例

は人物の重要度は映像中の人物の位置や大きさに反映されるとして，人物の動きから得られる特徴量を重要人物識別に用いる．まず注目フレームから検出された人物を前後 100 フレームの間トラッキングし，その人物の顔領域の大きさと位置の変化を取得する．本手法では，顔領域を追跡するために，KCF トラッカー [29] を採用した．

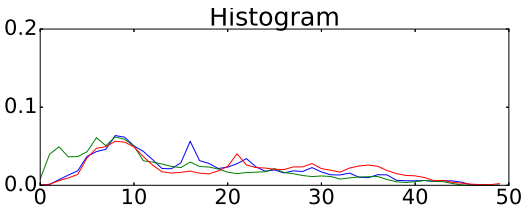
図 6 はトラッキングの例である．図 6(a) は注目フレームから 100 フレーム前，6(c) は注目フレームの 100 フレーム後を表しており，青色の矩形が顔領域である．図 6(d) の黄色の線が顔領域中心の変化を表している．

こうしてある人物 i から得られた，前後 100 フレームにおける顔領域から座標と大きさを抽出し，この 3 次元ベクトルを連結した $x_i^m \in \mathbb{R}^{600}$ を人物の動きの特徴量とする．なお，図 6(b) の奥の人物のように，トラッキング対象の人物が，移動やオクルージョンにより画面上から消失した場合，トラッキングを中止し，残りフレームの顔領域の大きさおよび位置は 0 とする．

人物の見えの特徴量

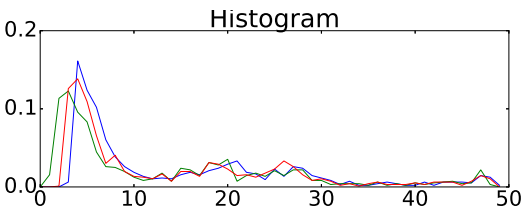
重要人物はカメラに対して正面か，少なくとも顔が見えるように撮影されること

(a) 顔領域



(b) 顔領域 (a) のカラーヒストグラム

(c) 顔領域



(d) 顔領域 (b) のカラーヒストグラム

図 7: 人物の見えの特徴量の例

が多く，顔の見え方に関する特徴量も，動き同様重要人物識別において有効であると考えられる．本研究では，見えに関する特徴量として，カラーヒストグラムと DNN 特徴量 [30] の 2 種を評価する．カラーヒストグラムは顔領域から R，G，B それぞれのチャンネルのヒストグラムを算出し，それぞれのチャンネルから抽出した 50 次元のヒストグラムを連結する．こうして得られたベクトル $x_i^l \in R^{150}$ を人物の見えの特徴量とする．図 7 は取得した顔領域とその顔領域のカラーヒストグラムである．例えば，図 7(a) は顔が見えているため，肌領域が多く，対応するカラーヒストグラムは明部と暗部に大きな偏りは見られない．一方で，図 7(c) は後ろを向いているため，肌領域が少なく，カラーヒストグラムは図 7(d) のように暗部を中心の分布を持つ．これは，主にアジア系の人物に特有の傾向である．このように、顔の向きによって取得されるヒストグラムが異なる．

DNN 特徴量として事前学習済みの DNN に顔画像領域を入力し，隠れ層の出力を抽出する．提案手法では顔認識用に学習された FaceNet [30] を用いて，ネットワークの出力を特徴量として採用した．ここで得られる特徴量 x_i^l は 128 次元ベクトルである．

3.3 条件付き確率場とニューラルネットワークによる重要人物識別

提案する識別モデルを図 8 に示す．提案モデルは 2 層の全結合層と CRF 層で構成される．全結合層では，映像から抽出された人物 i の動きと見え方に関する特徴量 x_i^m, x_i^l から，ベクトル f_i を算出する．提案モデルでは人物の動きと見え方を考慮して重要度推定を行うため，第一層の出力を連結し，第二層の入力とする．

$$h_i^m = \rho(W_m x_i^m + b_m) \tag{3}$$

$$h_i^l = \rho(W_l x_i^l + b_l) \tag{4}$$

$$f_i = \rho(W h_i^{ml} + b_{ml}) \tag{5}$$

ここで，行列 $W_m \in \mathbb{R}^{600\times100}$，$W_l \in \mathbb{R}^{d\times100}$，$W \in \mathbb{R}^{200\times100}$ は識別モデルのパラメータであり，x_i^l がカラーヒストグラムの場合 $d = 150$，DNN 特徴量の場合 $d = 128$ である．また，活性化関数 ρ は Rectificed Linear Unit 関数 [31] とする．式 (5) において，h_i^{ml} は h_i^m と h_i^l の出力を連結して作成した特徴ベクトルである．

CRF 層では，あるフレームに映っている人物 i の特徴量から算出されたベクトル f_i，ただし $(i = 1, \ldots, I)$，からそれぞれの重要度ラベル $t_1, \ldots, t_I$ の事後確率を求める．このとき，人物 i の重要度ラベル t_i は，その人物が重要であるとき $t_i = 1$，それ以外は 0 である．CRF 層は，それぞれの人物についてエネルギーを算出するデータ項と，同一フレームに含まれる人物の特徴間の関係をモデル化するペアワイズ項からなる．

データ項は，各人物ごとにそれぞれのラベルについてエネルギーを算出する．提案モデルでは，データ項のエネルギーを以下のように定義する．

$$\phi_0(f_i) = \rho(v_0^\top f_i + k_0) \tag{6}$$

$$\phi_1(f_i) = \rho(v_1^\top f_i + k_1) \tag{7}$$

ここで，ベクトル v_0，$v_1 \in \mathbb{R}^{100}$ とスカラー k_0, k_1 は識別モデルのパラメータである．データ項 $\phi_0(fi)$，$\phi_1(fi)$ はそれぞれ人物 i を非重要（0），あるいは重要（1）と識別する場合のコストに対応する．例えば，ある人物が非重要人物であり，対応する重要度ラベルが $t_i = 0$ である時，データ項のエネルギーは高い値を示す．

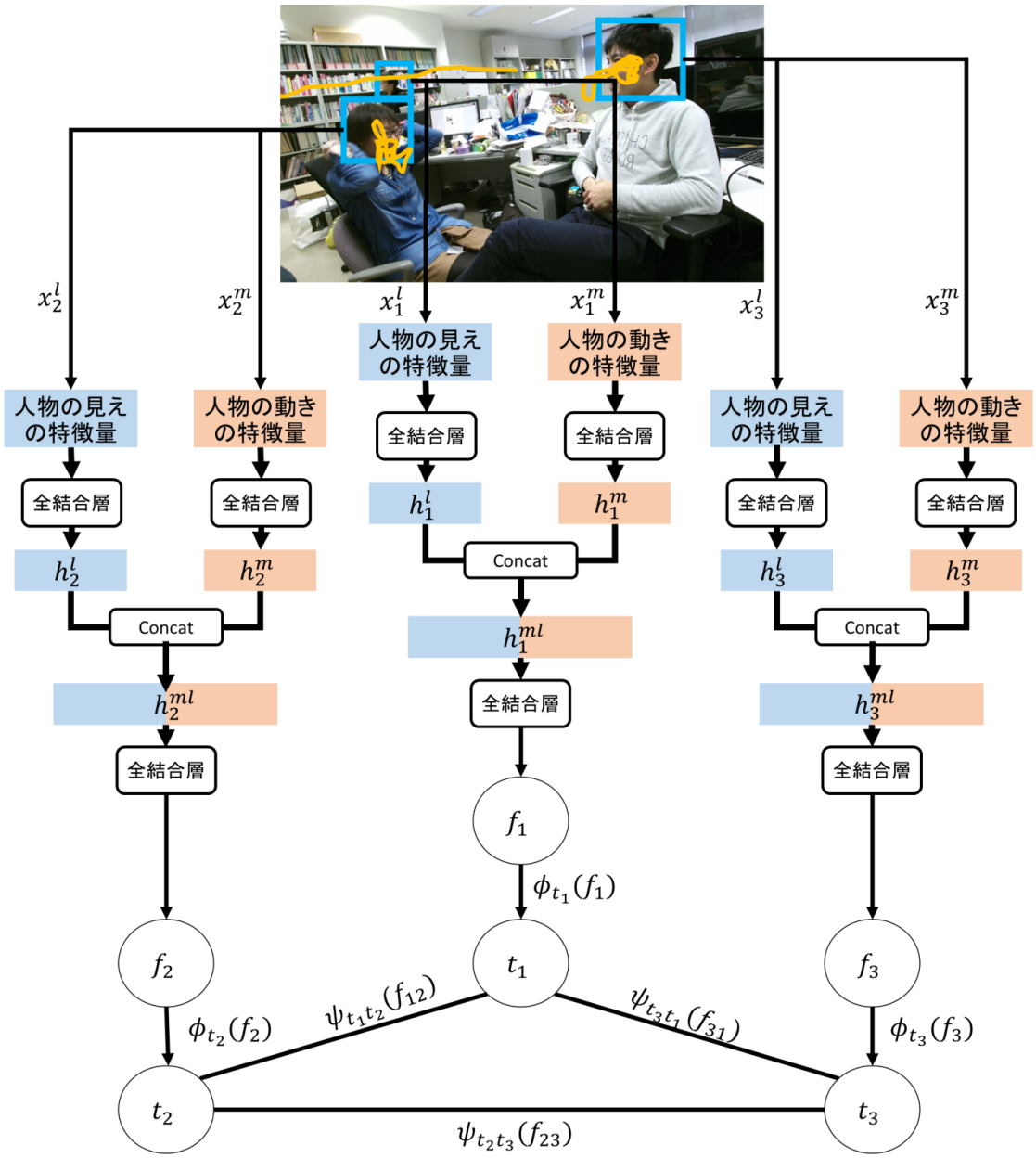


図 8: 提案する識別モデル

ペアワイズ項では，同じフレームに含まれる 2 人の人物の重要度ラベルとそれぞれの特徴量から，エネルギーを算出する．ペアワイズ項のエネルギーを以下のように定義する．

$$\psi_{00}(f_{ij}) = \rho(u_{00}^{\top} f_{ij} + c_{00}) \tag{8}$$

$$\psi_{01}(f_{ij}) = \rho(u_{01}^{\top} f_{ij} + c_{01}) \tag{9}$$

$$\psi_{10}(f_{ij}) = \rho(u_{10}^{\top} f_{ij} + c_{10}) \tag{10}$$

$$\psi_{11}(f_{ij}) = \rho(u_{11}^{\top} f_{ij} + c_{11}) \tag{11}$$

ここで，f_{ij} は式 (5) の出力 f_i, f_j を連結した特徴ベクトルである．ペアワイズ項 $\psi_{00}(f_{ij}), \psi_{01}(f_{ij}), \psi_{10}(f_{ij}), \psi_{11}(f_{ij})$ はそれぞれ，2 人の人物の非重要 (0)，重要 (1) と識別する組み合わせのコストに対応する．

提案する識別モデルは，事後確率を最大化するラベルの組を求めることで重要人物を識別する．フレーム内の全ての人物のラベルを $T = \{t_i | i = 1, \ldots, I\}$，その人物の特徴量から算出したベクトルを $F = \{f_i | i = 1, \ldots, I\}$ とする．ここで，エネルギー関数 $E(T, F)$ をデータ項とペアワイズ項を用いて以下のように定義する．

$$E(T, F) = \sum_i \phi_{t_i}(f_i) + \sum_{ij} \psi_{t_i t_j}(f_{ij}) \tag{12}$$

このエネルギー関数を用いて重要度ラベルの事後確率は次のように定義する．

$$p(T|F) = \frac{1}{Z} e^{-E(T,\, F)} \tag{13}$$

ここで，Z は分布を正規化するための分配関数を表し，次のように定義する．

$$Z = \sum_T e^{-E(T,\, F)} \tag{14}$$

式 (14) に示すように，このとき事後確率 $p(T|F)$ を求めるために，可能なすべての重要度ラベルを評価する必要がある．この分配関数 Z の算出は評価される要素数とクラス数に応じて，膨大な計算を要する．一般に，CRF の学習においては，このような計算を避けるために Contrastive Divergence [28] などの近似手法が採

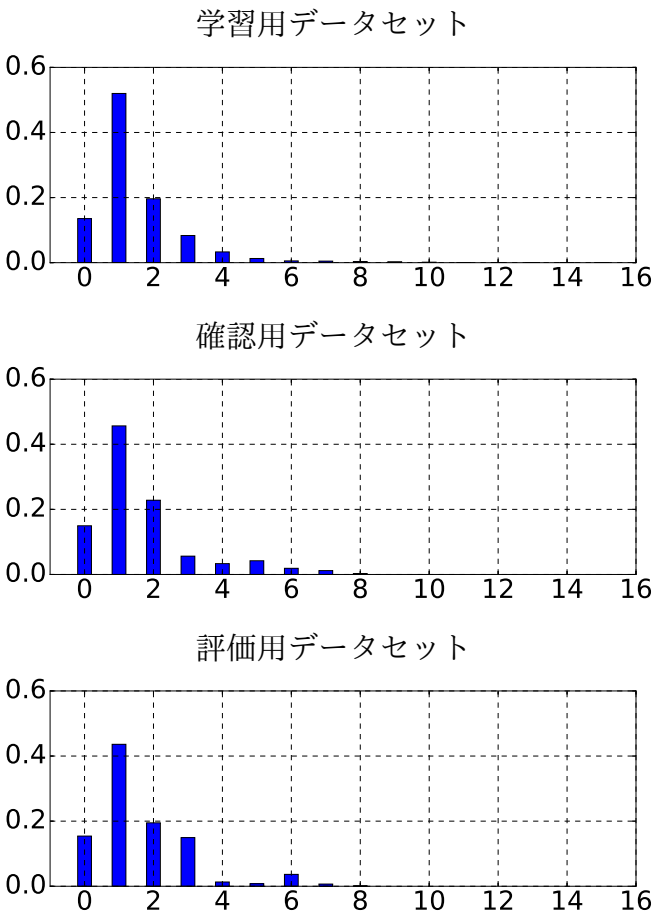


図 9: データセットにおけるフレームに映っている人数の分布

用されている．しかし，本研究が対象とする映像中の重要人物の識別は，1 フレームに含まれる人数が限られており，クラス数も重要と非重要の 2 クラスのみである．図 9 は本研究で用いたデータセットにおける，1 フレームから検出された人数の分布である．横軸は 1 フレームから検出された人数，縦軸は，その人数を含むフレーム数であり，各データセットの全フレーム数で正規化されている．このように，多くのフレームでは，10 人以下の人物しか検出されず，必要な計算量は抑えられている．そのため，本研究では近似手法を用いず，可能な重要度ラベルの組み合わせを全て評価し分配関数を求めることが可能である．

学習の際に，式 (14) に示すように，分配関数 Z を計算する際に同じ計算を何度も行う必要がある．この計算を効率化するため，提案手法では，必要なデータ

項とペアワイズ項を事前に計算し，その結果を再利用して分配関数 Z を求める．ここでデータ項，ペアワイズ項を以下のように求める．

$$\phi(f_i) = Vf_i + K \tag{15}$$

$$\psi(f_{ij}) = Uf_{ij} + C \tag{16}$$

ここで，$V = (v_0\ v_1)^\top$，$K = (k_0\ k_1)^\top$，$U = (u_{00}\ u_{01}\ u_{10}\ u_{11})^\top$，そして $C = (c_{00}\ c_{01}\ c_{10}\ c_{11})^\top$ である．この ϕ,ψ を用いて ϕ では，重要 (1)，非重要 (0) の 2 通り，ψ では 4 通りすべての組み合わせをあらかじめ計算しておく．この事前計算にかかる，計算コストは大きくない．エネルギー E を求める際には，あらかじめ ϕ, ψ を計算した中からラベルと対応する値を取得する．こうすることで，より少ない計算量での分配関数の算出が可能となる．

3.4 ネットワークの学習

ネットワークの学習において，真値ラベルの負の対数尤度を損失関数 L とし，これを最小化することで識別モデルを学習する．

$$L(T_m, F_m) = -\sum_m \log p(T_m|F_m) \tag{17}$$

ここで，T_m と F_m は，フレーム m に含まれる人物の重要度ラベルと特徴ベクトルを表している．学習時では，過学習を避けるために，全結合層に Dropout [32] を適用し，確率的勾配降下法 [33] により識別モデルを最適化する．

4. 評価実験

本章では，提案手法の有効性を検証するために，ホームビデオを収集して作成したデータセットを用いて重要人物識別を行った．本実験では，提案モデルを種々のベースライン手法と比較することで提案モデルの有効性および CRF 層の効果を検証した．

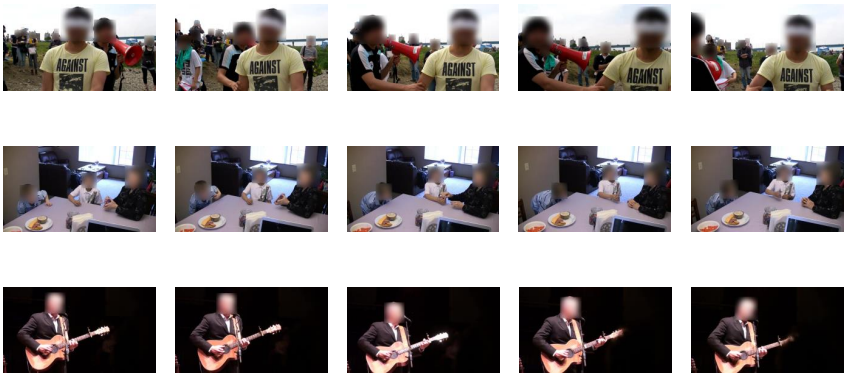
以下，データセットの詳細について述べた後，実験および，その結果と考察について述べる．

4.1 データセット

学習用データセットと評価用データセットは従来手法 [15] と同様，YouTube 映像とホームビデオ映像からなるデータセットを用いた．このデータセットは，一般ユーザが撮影した未編集映像であり主に人物を撮影対象としている．図 10 に学習用データセットと評価用データセットの例を示す．なお，本論文ではプライバシーの観点から画像に保護処理を施している．データセットに含まれる映像は，アノテーションデータとして，顔領域の位置を示すバウンディングボックス，各人物の重要度ラベルが付与されている．重要度ラベルには，重要，非重要の 2 種のラベルが付与されている．データセットは 99 本の YouTube の映像と 20 本のホームビデオ映像に分けられる．それぞれについて詳述する．

YouTube 映像中の人物は各 6 人のアノテータにより重要，非重要のラベルが付与されている．本研究では付与された重要，非重要のラベルを多数決により真値として採用した．この YouTube 映像のデータセットを学習用データ（66 本），確認用データ（33 本）の 2 種類に分け，識別モデルを学習した．学習用データセットのサンプル数は 120, 955 であり，そのうち，82, 079 サンプルが重要人物である．確認用データセットは 67, 655 サンプルのうち，39, 764 サンプルが重要人物である．

ホームビデオ映像は，撮影者自身がアノテータとなり重要，非重要のラベルが付与されている．撮影者がつけた重要度と視聴者にとっての重要度は一致するという知見が得られている [15]．そのため，YouTube 映像のデータセットで識別



(a) 学習用データセットの例

(b) 評価用データセットの例

図 10: データセットの例

モデルを学習した場合でもホームビデオ映像を識別できると考えられる．評価用データセットのサンプル数は 55, 336 であり，そのうち，37, 431 サンプルが重要人物である．データセットについてまとめたのを表 1 に示す．

4.2 実験の詳細

提案手法を検証するために本研究では，まず従来手法と提案手法を比較した．また CRF の効果の検証するため，提案モデルの機能を一部除去したベースライン手法と提案手法を比較した．

表 1: データセットの構成

	ラベル付加方法	映像本数	サンプル数	重要人物の数
YouTube 映像				
学習用データセット	6 人の視聴者	66 本	120,955	82,079
確認用データセット	による多数決	33 本	67,655	39,764
ホームビデオ映像				
評価用データセット	撮影者本人	20 本	55,336	37,431

本実験では従来手法として Nakashima らの手法 [15] のトラッキングに基づく時間方向の平滑化を除いた簡略化手法と比較する．これは，本研究では各フレームを起点として人物の短時間のトラッキングを行うが，フレーム間での人物の対応付けをしていないためである．また，Nakashima らの手法はネットワークの入力が人物の動きの特徴量のみを用いているため，提案手法と純粋な比較はできない．そこで，提案手法のネットワークから，見えに関する特徴量を入力とする層を除去したモデルを作成し，従来手法との比較に用いる．

CRF の効果を検証するために，CRF を使用しないベースラインモデルを学習し，提案手法と識別精度を比較する．また，提案手法と同じモデルを学習し，識別の際には CRF を用いない手法とも比較を行い，CRF がどのような影響を与えるか検証する．この比較実験では，人の見えの特徴量として，カラーヒストグラムを用いた場合と，DNN 特徴量を用いた場合の 2 種を評価する．

2 つの実験では前節で示した 20 本の動画に対し，下記の (1)〜(5) の 5 つの手法を用いて重要人物識別を行う．

(1) Nakashima らのを簡略化した手法 [15]

(2) ペアワイズ項を除去したモデル

(3) CRF 層を除去したモデル

(4) 提案手法 (評価時にペアワイズ項を除去)

(5) 提案手法

ここで，手法 (1) は従来手法と提案手法の比較でのみ用いる．以下，手法 (2)，(3)，(4) について述べる．

(2) ペアワイズ項を除去したモデル

手法 (2) は，提案手法の CRF 層からペアワイズ項を除外したモデルである．このモデルと比較することで，CRF 層が人物の特徴量間の相関を考慮することで識別結果に及ぼす影響を調査する．

(3) CRF 層を除去したモデル

手法 (3) は，CRF 層を重要，非重要の尤度を出力する 1 層の全結合層に置き換えたモデルである．このモデルは損失関数として，次式で定義される Softmax Cross-Entropy を用いた．

$$z_a = \frac{\exp(u_a)}{\sum_{b=0}^{1} \exp(u_b)} \tag{18}$$

ここで，a は出力層のユニットの数 $(a = 0, 1)$ であり，u_0, u_1 は出力層の 1, 2 番目のユニットの入力，z_0, z_1 はその出力を表す．前章で述べたある人物 i を全結合層から求めた特徴ベクトル f_i を入力として取り，その人物が重要人物か否かを表す確率を出力する．出力された確率から重要度ラベル t_i を次のように選択する．

$$t_i = \begin{cases} 1 & (z_1 \geq 0.5) \\ 0 & (otherwise) \end{cases} \tag{19}$$

(4) 提案手法 (評価時にペアワイズ項を除去)

手法 (4) は学習された CRF 層のペアワイズ項がどのように影響を及ぼしているのを調査するのが目的である．手法 (4) は，評価時のみ，CRF のペアワイズ項をエネルギー計算から除外して識別する．

学習では，バッチサイズを 100，学習率を 0.0001，パラメータ更新回数をエポック数を 20 回とした．学習時，確認用データセットにおいて最も高い識別精度を達成したモデルを採用した．また，実装には深層学習のフレームワークである Chainer [34] を用いた．

4.3 実験結果

本節では，手法 (1)〜(5) を用いた評価用データセットの重要人物識別を行い，手動でラベル付けした真値と比較することで評価する．具体的には，真値で付けられたラベルが重要人物であり，ネットワークの識別結果が重要人物である人物の数を TP(True Positive)，真値で付けられたラベルが重要人物であり，ネットワークの識別結果が非重要人物である人物の数を FN(False Negative) とし，重要人物の再現率を求める．

$$REC = \frac{TP}{TP+FN} \tag{20}$$

また，真値で付けられたラベルが非重要人物であり，ネットワークの識別結果が重要人物である人物の数を FP (False Positive)，真値で付けられたラベルが非重要人物であり，ネットワークの識別結果が非重要人物である人物の数を TN (True Negative) とし，非重要人物の誤識別率（FPR: False positive rate）を求める．

$$FPR = \frac{FP}{FP+TN} \tag{21}$$

また同様に，重要人物の適合率 PRE (precision)，重要人物識別の識別精度 ACC (Accuracy)，F 値 (F1-measure) を以下の式で求める．

$$PRE = \frac{TP}{TP+FP} \tag{22}$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{23}$$

$$F_1 = 2 \cdot \frac{PRE \cdot REC}{PRE+REC} \tag{24}$$

表 2 に重要人物識別の定量的評価を示す．表 2 では，それぞれの特徴量の中で最大の識別精度と F 値の値を太字とした．図 11 に人物の動きの特徴量を用いた場合の重要人物識別の結果を示す．図 12-14 に人物の動きの特徴量とカラーヒストグラムを用いた場合の重要人物識別の結果を示す．図 15-17 に人物の動きの特徴量と FaceNet 特徴ベクトルを用いた場合の重要人物識別の結果を示す．ここで，図 11-17 は赤色が重要人物，緑色が非重要人物の識別結果を表す．また，矩形が真値と同じ識別結果，バツ印が真値と異なる識別結果を表す．また，図の下は対応するフレーム数，図の左には識別を行った手法をを示している．

表 2: 手法 (1)〜(5) による定量的評価結果

	REC(%)	PRE(%)	FPR(%)	ACC(%)	F1(%)
人物の動きの特徴量					
(1) Nakashima らの手法 [15]	76.0	83.3	31.8	73.5	79.5
(5) 提案手法	82.4	86.3	27.3	**79.3**	**84.3**
人物の動きの特徴量+カラーヒストグラム					
(2a) ペアワイズ項を除去したモデル	68.5	93.8	9.5	75.7	79.2
(3a) CRF 層を除去したモデル	74.7	90.5	16.4	77.6	81.9
(4a) 提案手法 (ペアワイズ項を除去)	96.8	77.0	60.3	78.3	85.8
(5a) 提案手法	85.9	87.2	26.3	**82.0**	**86.5**
人物の動きの特徴量+FaceNet					
(2b) ペアワイズ項を除去したモデル	75.6	92.5	12.8	79.4	83.2
(3b) CRF 層を除去したモデル	75.0	91.5	14.6	78.3	82.4
(4b) 提案手法 (ペアワイズ項を除去)	96.1	80.6	48.3	**81.7**	**87.6**
(5b) 提案手法	79.9	88.5	21.7	79.4	84.0

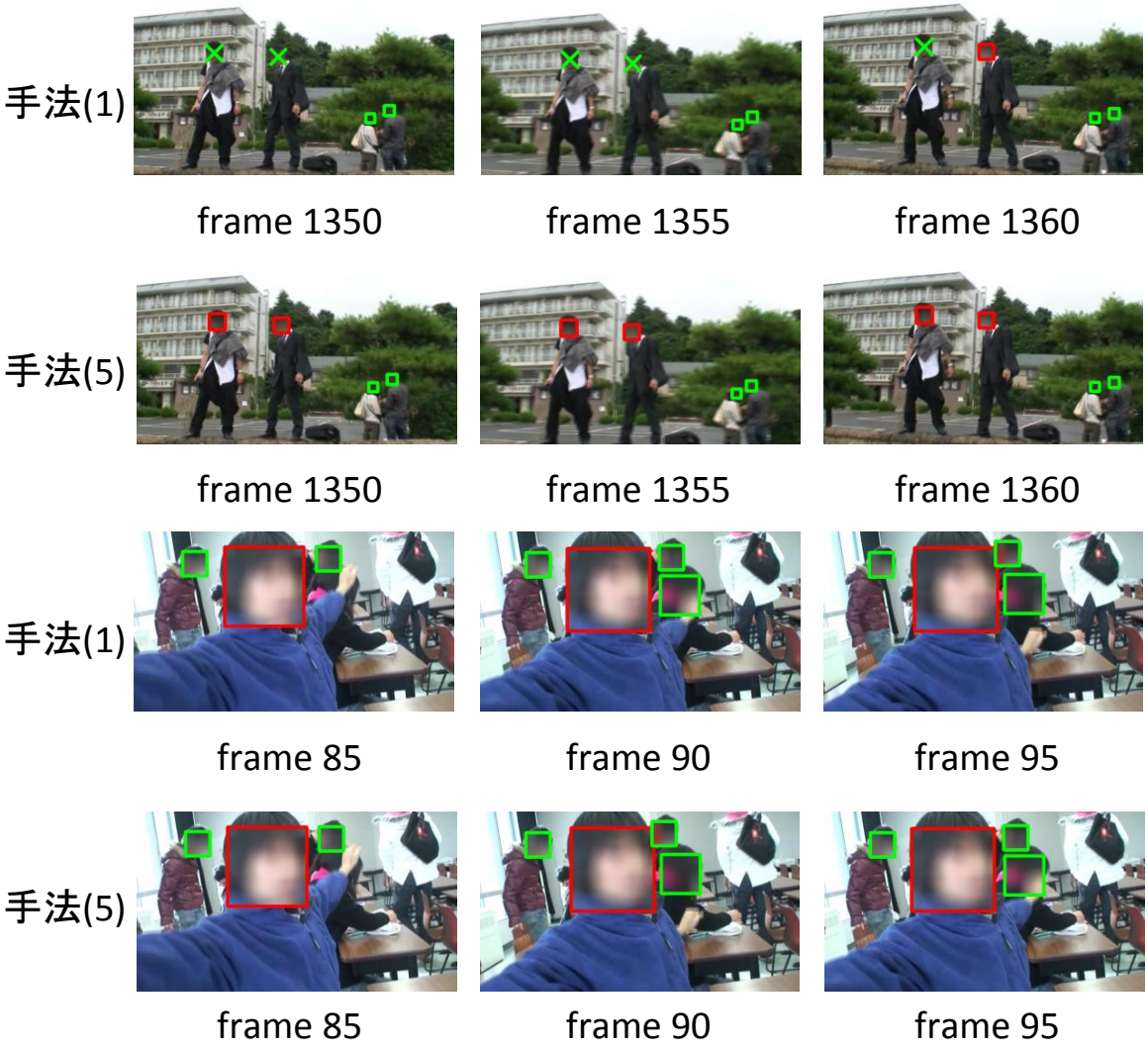


図 11: 手法 (1) と手法 (5) の識別結果の例

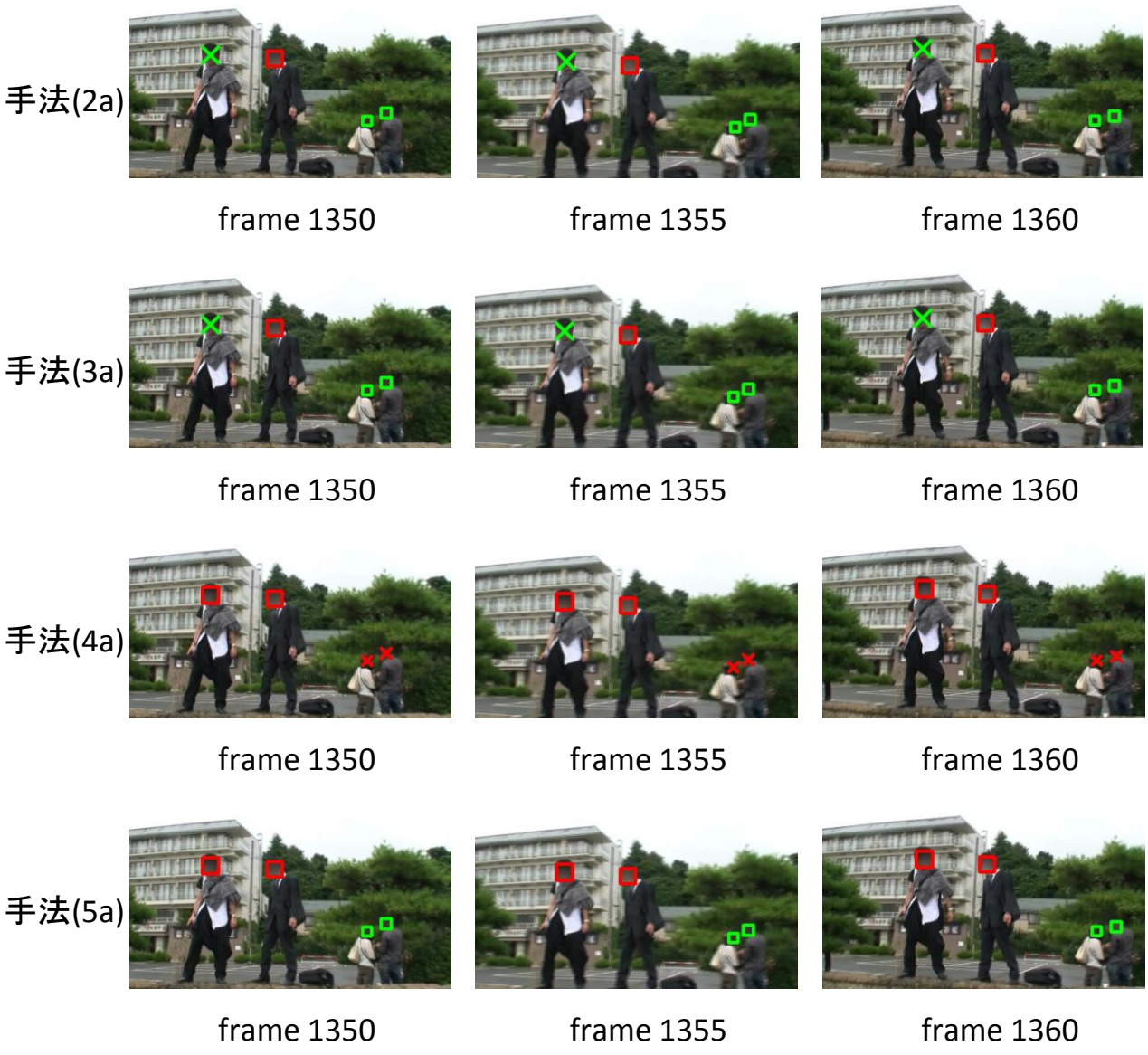


図 12: 手法 (2a) から手法 (5a) における識別結果の例 1

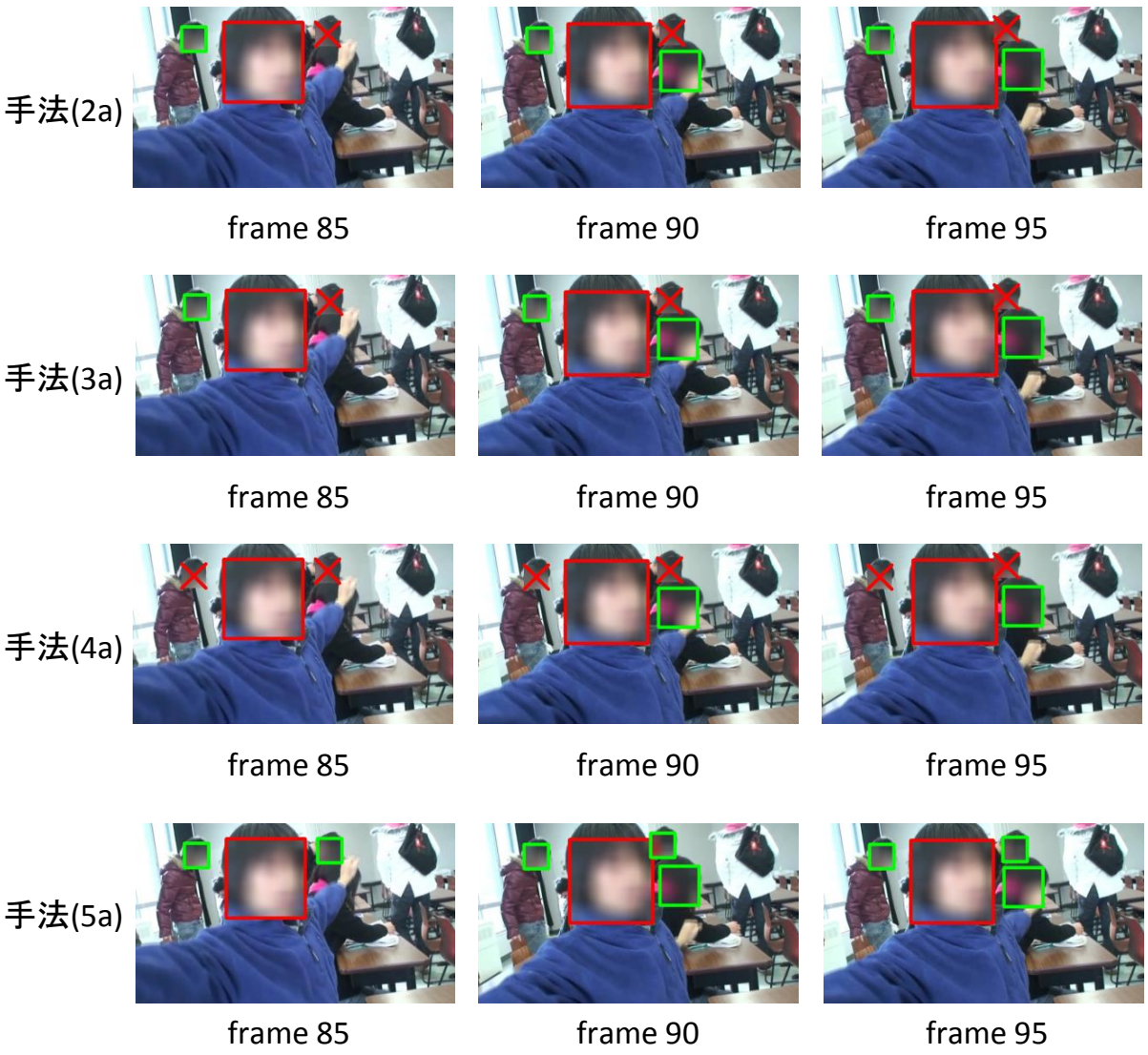


図 13: 手法 (2a) から手法 (5a) における識別結果の例 2

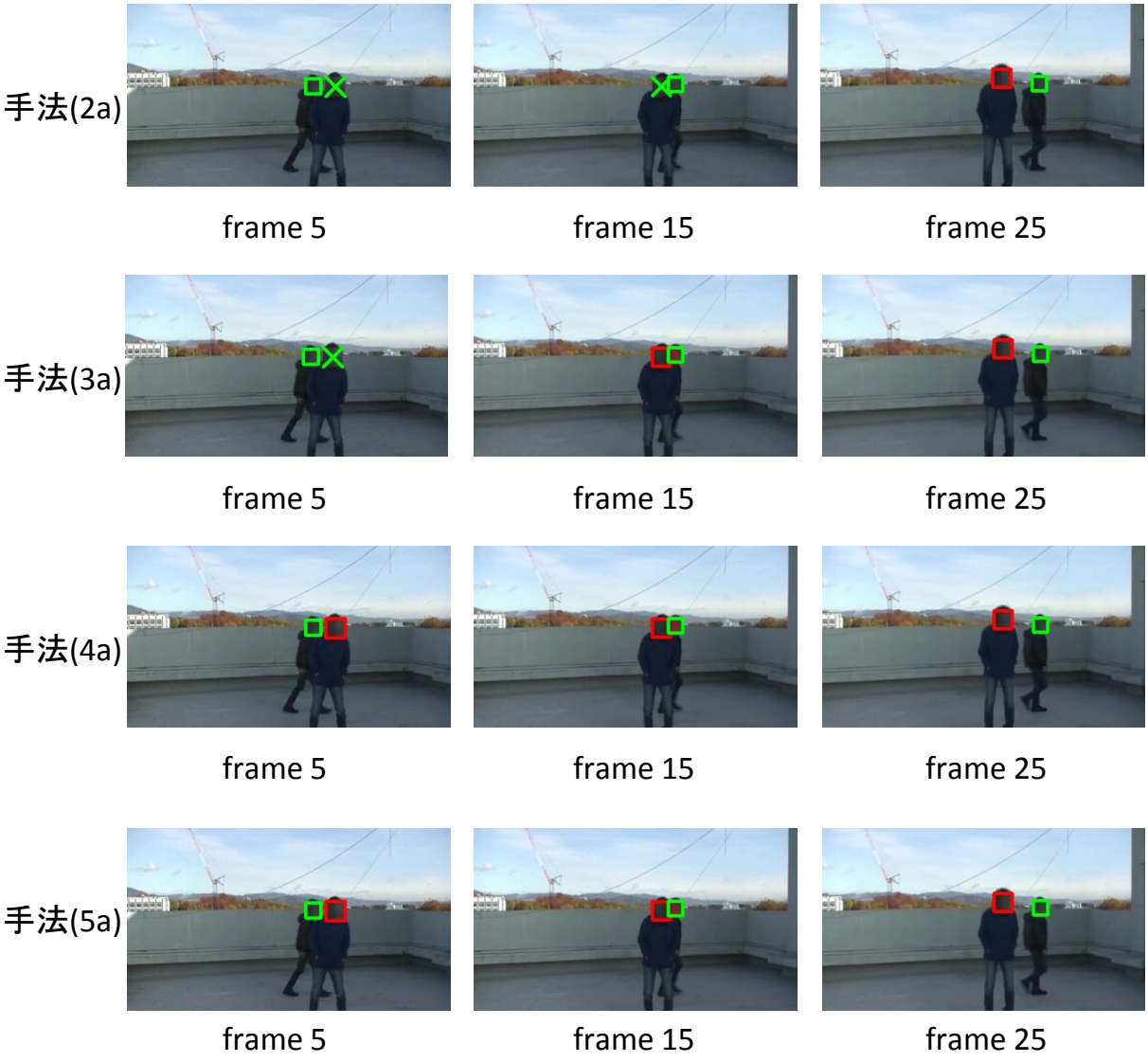


図 14: 手法 (2a) から手法 (5a) における識別結果の例 3

図 15: 手法 (2b) から手法 (5b) における識別結果の例 1

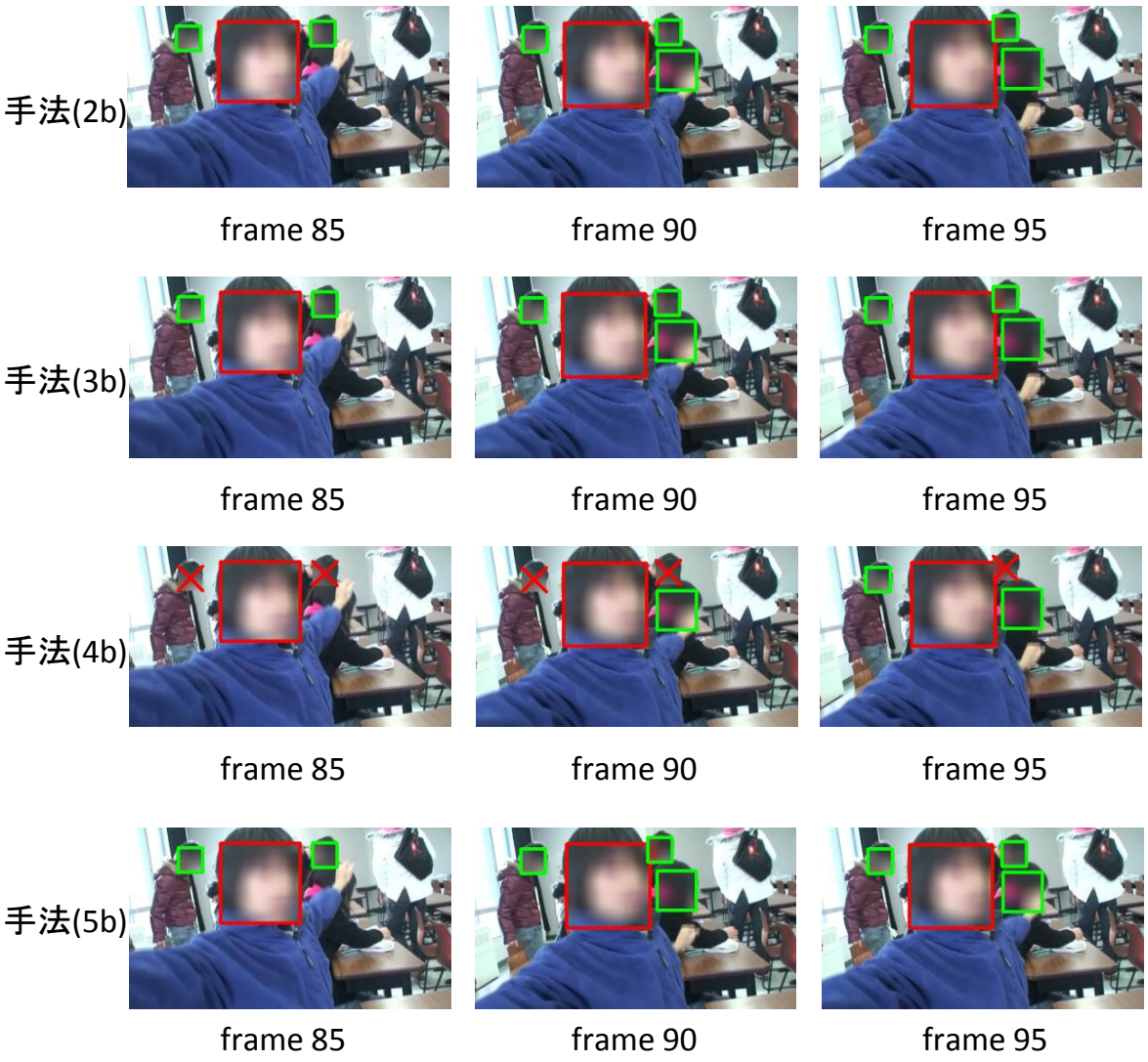


図 16: 手法 (2b) から手法 (5b) における識別結果の例 2

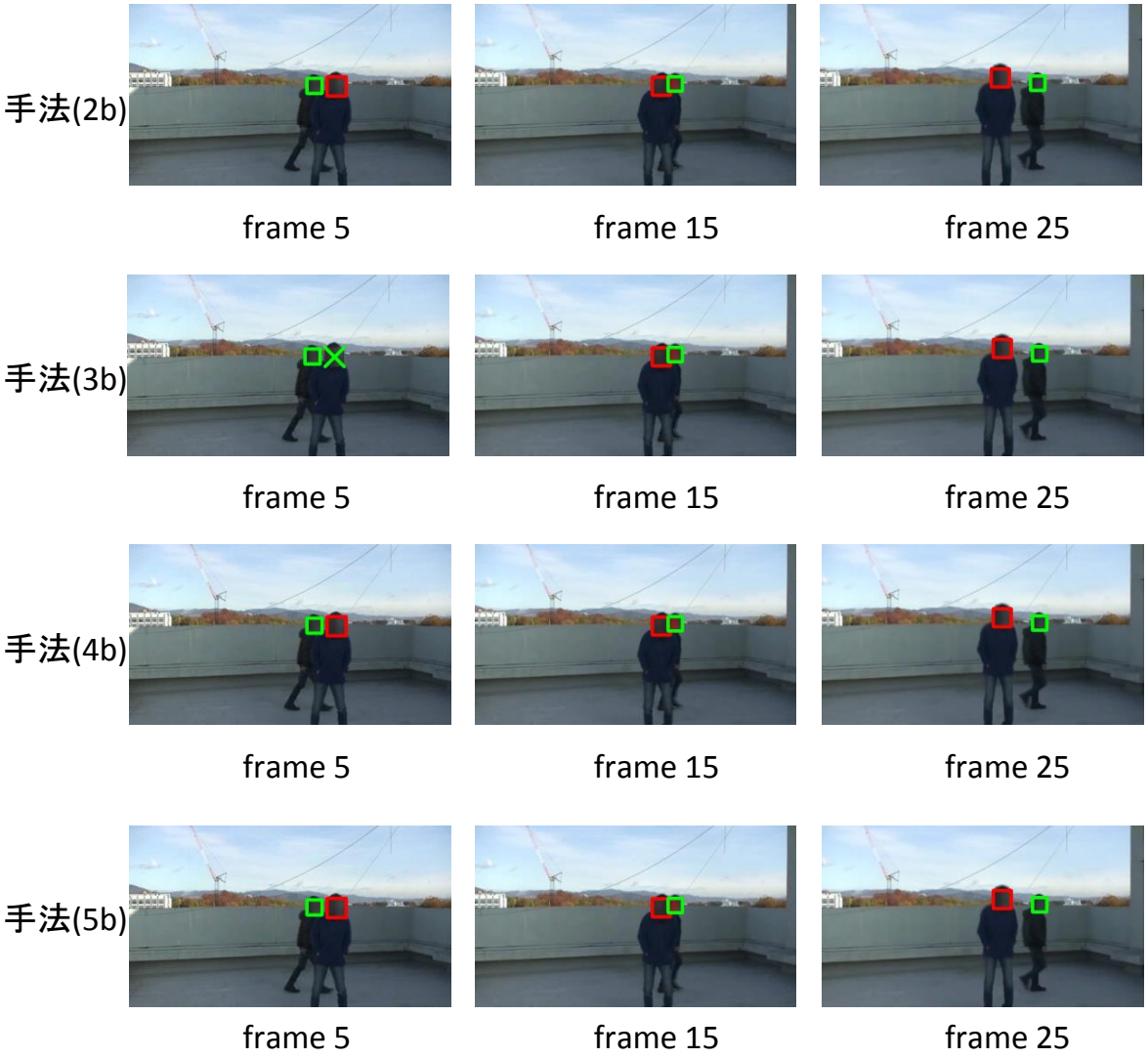


図 17: 手法 (2b) から手法 (5b) における識別結果の例 3

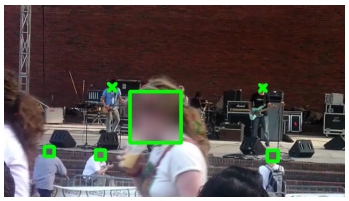
4.4 考察

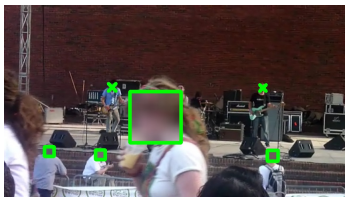
従来手法と提案手法との比較実験では，従来手法より提案手法が識別精度が高いことが示された．これにより，従来手法のモデルより提案手法のモデルが，重要人物識別において有効であることが確認された．

カラーヒストグラムを入力として用いた場合，提案手法が識別精度において最も高い性能を示した．また，提案手法とペアワイズ項を評価時に除去した手法 (4) を比較すると，手法 (4) が再現率においてより高い値を達成した．一方で，適合率，非重要人物の誤識別率，F 値は提案手法の方が優れている．これは，主に手法 (4) では FP が多くなっているためである．例えば，図 13 では，手法 (4a) の奥にいる人物が誤って重要人物と識別されている．一方で，提案手法では図 13 のように，手法 (5a) では手前にいる人物が重要人物，奥に映っている人物が非重要人物と識別されている．このことから，CRF のペアワイズ項は，FP を抑制する効果があると思われる．

提案手法と手法 (2)，(3) を比較すると，識別精度において提案手法が優れている．また，手法 (2)，(3) は，提案手法に比べると，適合率，非重要人物の誤識別率においては優れている．一方で，再現率，F 値では提案手法の方が優れている．これは，手法 (2)，(3) は重要人物と識別されるハードルが高く，一定の重要度を持たなければ，非重要人物のラベルを選択する傾向があるためだと考えられる．一方で提案手法は，フレーム内にいる他の人物による影響をうけるため同じ特徴を持っていたとしても同じフレームにいる人物によっては識別結果が異なることがある．例えば，図 12 において，左端の方にいる人物を，手法 (2a)，(3a) では誤って非重要人物と識別している．一方で，提案手法では，CRF によってフレーム内の人物の特徴間の相関を考慮する．そのため，図 12 において，左端にいる人物はとなりにいる重要人物と顔の大きさ，見えや動きなどの相関が高いため端にいる人物も重要人物と識別される．このように，CRF が人物の特徴間の相関を考慮することで，識別精度を向上することができる．

しかし，提案手法が人物同士の相関関係をモデル化することによって，誤りを生じる例もある．図 18 は手法 (2)，(3) では，真ん中の人物を正しく非重要人物と識別している．しかし，提案手法では，真ん中の人物は他の非重要人物と異な


手法 (2a)


手法 (3a)


手法 (5a)

図 18: 提案手法の失敗例

る見えや動き方をしているため，真ん中の人物が重要人物と誤って識別されている．このように，同じフレーム内で複数の非重要人物が存在する時，誤って識別される場合がある．

一方で，DNN 特徴量である FaceNet [30] を用いた場合では，手法 (4) が一番高い識別精度を示した．これは，ペアワイズ項を評価時に除外した手法では，重要人物と識別されるハードルが低く，重要人物のラベルを選択されやすい傾向がある．さらに，評価用のデータセットは全 $55,336$ サンプルのうち重要人物が $37,431$ と重要人物の割合が高い．そのためペアワイズ項を評価時に除外した手法 (4) が高い識別精度を示したと考えられる．また，FaceNet は顔分類のためのネットワークであるため，人物の顔の変化によって，出力される特徴ベクトルは大きく変化する．しかし，この変化は人物の顔の向きなどの見えの変化によって，重要人物と非重要人物に識別できるような変化が見られない．そのため，CRF 層で人物の特徴間の相関関係を考慮した結果，提案手法は，ペアワイズ項を除外したものより識別精度が下がったと考えられる．そのため，FaceNet のネットワークを重要人物識別に特化するように再学習することで提案手法をさらなる精度向上が期待できる．

また，動きに関する特徴だけでなく，見えに関する特徴量も利用したモデルの方がより高い識別精度や再現率を達成した．このことから人物の画像特徴が重要人物識別において重要であると考えられる．

5. まとめ

本論文では，複数の人物を含むシーンにおいて，映像中の人物がその映像に必要な重要人物なのか，偶然映り込んだ非重要人物なのかを識別する手法を提案した．提案手法では，人手で検出した人物から，人物の動きの特徴量と人物の見えの特徴量を抽出する．こうして得られた特徴量を入力として，CRF と DNN を組み合わせた識別モデルは映像中の各人物について重要，あるいは非重要のクラスラベルを出力する．提案手法では，複数の人物の特徴間の相関関係とその重要度をモデル化するために CRF を取り入れたモデルを設計した．また，CRF を学習するための効率的な損失関数の計算手法を提案した．

実験において，YouTube 映像のデータセットを学習し，ホームビデオ映像のデータセットを用いて識別精度を計測した．CRF と DNN を組み合わせたモデルは，従来手法より識別精度において優れていることを確認した．また，CRF を使用しないベースラインモデルとの比較実験では，CRF が人物の特徴間の相関を考慮することで，識別精度を向上させることも実験により示した．また，人物の動きの特徴を入力とするモデルと，人物の動きと見えの特徴を入力とするモデルを比較することで，人物の動きの特徴だけでなく，見えに関する特徴量も利用したモデルのほうがより高い識別精度であることを確認した．

今後は，より正確に識別を行うために，FaceNet のネットワークパラメータを含めた End-to-End での再学習を行う必要がある．また，人物の検出を人手でなく自動化することも今後の課題である．また，データセットを増やすことで，ネットワークの識別精度の向上や，より定量的な評価を行う必要がある．今後の展望として，リターゲティングなどのアプリケーションなどによる提案手法の有用性の検証が考えられる．

謝辞

本研究の全過程を通して，懇切なる御指導，御鞭撻を賜りました視覚情報メディア研究室 横矢 直和 教授に心より感謝致します．また，本研究の遂行にあたり，有益な御助言，御鞭撻を頂いた環境知能学研究室 萩田 紀博 教授に厚く御礼申し上げます．そして，本研究を進めるにあたり，始終暖かい御指導をして頂いた視覚情報メディア研究室 佐藤 智和 准教授に深く感謝致します．また，本研究を行うにあたり，多大なる御助言，御鞭撻を賜った視覚情報メディア研究室 河合 紀彦 助教に心より感謝致します．さらに，本研究を通じて，的確な御助言，御鞭撻を頂いた視覚情報メディア研究室 中島 悠太 助教 (現 大阪大学データビリティフロンティア機構 准教授) に深く御礼申し上げます．特に，中島 悠太 助教には，本研究の着想およびテーマ設定から研究の遂行，発表練習など，長期にわたり様々なご指導をいただきました．また，研究室生活において様々な支援をして頂いた，視覚情報メディア研究室秘書 石谷 由美 女史，南 あずさ 女史に厚く御礼申し上げます．また，あらゆる面において，多大なるご助言を頂いた視覚情報メディア研究室 大谷まゆ 女史に深く感謝いたします．そして，研究のみならず研究室生活全般においてお世話になりました視覚情報メディア研究室の諸氏に深く感謝いたします．最後に，両親をはじめ，私の二年間の大学院生活に関わった全ての方々に感謝の意を表します．

参考文献

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] J. Yang and M.-H. Yang, "Top-down visual saliency via joint CRF and dictionary learning," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 2296–2303, 2012.

[3] F. Liu and M. Gleicher, "Video retargeting: Automating pan and scan," in *Proc. ACM Int. Conf. Multimedia (MM)*, pp. 241–250, 2006.

[4] X. Fan, X. Xie, H.-Q. Zhou, and W.-Y. Ma, "Looking into video frames on small displays," in *Proc. ACM Int. Conf. Multimedia (MM)*, pp. 247–250, 2003.

[5] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Processing*, vol. 13, no. 10, pp. 1304–1318, 2004.

[6] W. Lai, X.-D. Gu, R.-H. Wang, W.-Y. Ma, and H.-J. Zhang, "A content-based bit allocation model for video streaming," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, vol. 2, pp. 1315–1318, 2004.

[7] M.-H. Hsiao, Y.-W. Chen, H.-T. Chen, K.-H. Chou, and S.-Y. Lee, "Content-aware video adaptation under low-bitrate constraint," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 2, 17 pages, 2007.

[8] M. Sun, A. Farhadi, B. Taskar, and S. Seitz, "Salient montages from unconstrained videos," in *Proc. European Conf. Computer Vision (ECCV)*, pp. 472–488, 2014.

[9] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Proc. Neural Information Processing Systems (NIPS)*, pp. 547–554, 2005.

[10] P. Baldi and L. Itti, "Of bits and wows: A Bayesian theory of surprise with applications to attention," *Neural Networks*, vol. 23, no. 5, pp. 649–666, 2010.

[11] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[12] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[13] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1597–1604, 2009.

[14] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Proc. Int. Conf. Computer Vision Systems*, pp. 66–75, 2008.

[15] Y. Nakashima, N. Babaguchi, and J. Fan, "Intended human object detection for automatically protecting privacy in mobile video surveillance," *Multimedia Systems*, vol. 18, no. 2, pp. 157–173, 2012.

[16] Y. Nakashima, N. Babaguchi, and J. Fan, "Privacy protection for social video via background estimation and CRF-based videographer's intention modeling," *IEICE Trans. Information and Systems*, vol. E99.D, no. 4, pp. 1221–1233, 2016.

[17] Y. Bengio, Y. LeCun, and D. Henderson, "Globally trained handwritten word recognizer using spatial representation, convolutional neural networks, and hidden markov models," in *Proc. Neural Information Processing Systems (NIPS)*, pp. 937–937, 1994.

[18] K. Yao, B. Peng, G. Zweig, D. Yu, X. Li, and F. Gao, "Recurrent conditional random field for language understanding," in *Proc. IEEE Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4077–4081, 2014.

[19] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Recursive neural conditional random fields for aspect-based sentiment analysis," in *Proc. ACL Conf. Empirical Methods Natural Language Processing (EMNLP)*, pp. 616–626, 2016.

[20] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph LSTM," in *Proc. European Conf. Computer Vision (ECCV)*, pp. 125–143, 2016.

[21] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 1529–1537, 2015.

[22] A. Arnab, S. Jayasumana, S. Zheng, and P. H. S. Torr, "Higher order conditional random fields in deep neural networks," in *Proc. European Conf. Computer Vision (ECCV)*, pp. 524–540, 2016.

[23] S. Chandra and I. Kokkinos, "Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian CRFs," in *Proc. European Conf. Computer Vision (ECCV)*, pp. 402–418, 2016.

[24] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. Association for Computational Linguistics (ACL)*, 10 pages, 2016.

[25] X. Chu, W. Ouyang, H. Li, and X. Wang, "CRF-CNN: Modeling structured information in human pose estimation," in *Proc. Neural Information Processing Systems (NIPS)*, pp. 316–324, 2016.

[26] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 5162–5170, 2015.

[27] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Aalysis and Machine Intelligence (PAMI)*, vol. 35, no. 8, pp. 1915–1929, 2013.

[28] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2006.

[29] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. European Conf. Computer Vision (ECCV)*, pp. 702–715, 2012.

[30] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[31] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Machine Learning (ICML)*, pp. 807–814, 2010.

[32] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Jounal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[33] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 13 pages, 2015.

[34] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: A next-generation open source framework for deep learning," in *Proc. Neural Information Processing Systems (NIPS)*, 6 pages, 2015.