Vol.2015-CG-161 No.21
Vol.2015-CVIM-199 No.21
2015/11/7

歩行者群の注視方向とシーン中の注視点の同時時系列最適化

小嶋 翔太[1,a)]　浮田 宗伯[1]　萩田 紀博[1]　Ming-Hsuan Yang[2]

概要：歩行者群を撮影した低解像度の動画から，各人の頭部位置と頭部方向，およびシーン中の注視点を推定し，それらの同時最適化によって推定精度を向上させる手法を提案する．各人がシーン中のどこに注意を向けているかを知る上で，頭部方向は有用な情報である．しかし，監視映像のような低解像度画像においては，アピアランスのみによる頭部方向の推定は誤差が大きく，注視点の推定も困難である．本研究では，複数の人々の視線が集中する地点は注視点である可能性が高いと考え，注視点を頭部方向推定にフィードバックすることで，両者を同時に最適化する．この同時最適化は，人が注視点から視線を誘引されていると仮定する Social Charge モデルに，個人・人群のそれぞれにおける注視の時間連続性や，歩行進行方向と頭部方向の一致性を表現する項を組み込んで実現する．このモデルによって，一般的な街頭などにおける人群の注視方向とシーン中の注視点を推定する．

Simultaneous Temporal Optimization of Pedestrians' Gaze Orientation and Gaze Point in a Scene

KOJIMA Shota[1,a)]　UKITA Norimichi[1]　HAGITA Norihiro[1]　Ming-Hsuan Yang[2]

Abstract: We present a method to simultaneously optimize pedestrians' head positions, directions and gaze points in a scene altogether. For estimating the positions where people look at in a scene, their head directions are useful. However, in low-resolution images such as surveillance videos, head directions estimated only by their appearances are noisy. Such noisy head directions make it difficult to estimate the gaze points. In this study, we assume that the gaze points are looked at by multiple people simultaneously and/or continuously. Based on this assumption, we optimize both head positions, head directions, and gaze points simultaneously in a video. The simultaneous optimization is realized by formulating the Social Charge Model, where each gaze point attracts people's gazes, with our proposed forces representing the following properties: (1) temporal continuity of gaze directions in each individual and multiple people and (2) consistency between walking and gaze directions in each individual. This model allows us to estimate pedestrians' gaze directions and gaze points in a general scene such as a street.

1. はじめに

人が何に注目し，何に興味を持っているかという情報にはさまざまな有用性がある．例えば，駅や商店街に設置されている広告やデジタルサイネージが，どのような人にどれだけ興味を持たれているかという情報は，広告を出す側にとって非常に有用な情報である．人の興味を理解して行動するようなロボットの開発にも必要不可欠である．また，興味によってだけでなく，違和感によっても人が目を引かれることがある．例えば，一人歩きしている幼児や不審な挙動をする人物に対して，明確に異常だと認識することはなくても，周囲の人々は目を向けることだろう．こうした違和感による注目を異常と判断することで，事件や事故の早期発見につなげることができる．

人の興味を直接，外部のセンサから取得することは不可能であるが，センサで取得した人の振る舞いから，注目している方向を推定することは可能である．この注目している方向を，本研究では注視方向と呼ぶこととし，注視方向に注目する人の振る舞いを特に注視と呼ぶこととする．この注視方向を，近年急速に普及している監視カメラなどから推定する試みがなされている．注視方向を正確に推定す

[1] 奈良先端科学技術大学院大学
[2] University of California, Merced
[a)] kojima.shota.kh1@is.naist.jp

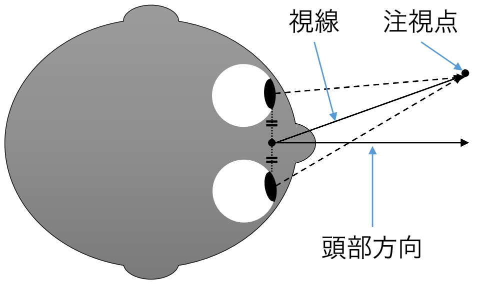


図 1 視線と実際の注視点．本研究では，頭部方向を注視方向の近似として用いるため，注視点は頭部方向の半直線上にある．

図 2 評価実験に使用するデータセットから顔領域をクロップしたもの．

るには，眼球の 3 次元姿勢を計測して視線を求める必要があるが（図 1），環境に取り付けられたカメラから撮影したような，人が低解像度で撮影されているような映像からは，眼球の姿勢を計測することは不可能である（図 2）．このため，低解像度の映像で人の注目を扱う研究では，視線同様に興味の指標である頭部方向の推定を行い，注視方向の近似として使用している [1], [2]．

本研究の興味は，こうした監視カメラなどで撮影された動画に対して行った注視方向推定の誤差を小さくすることにある．この注視方向推定の誤差は，頭部方向推定の誤差と，頭部方向と視線方向の誤差から発生するものである．したがって，画像から完全に頭部方向を求めても，それは視線方向ではなく，シーン中の「多くの人が注視している点」の方向を向いていない．そこで，シーン中の注視点を向くように頭部方向（注視方向）を最適化することで，頭部方向と視線方向の誤差を吸収して，頭部方向推定によって，注視方向を推定することを目指す．

ある人が注視している対象は，何らかの注視を引き起こす性質を持っていると考えられ，そのような対象は，瞬間的または長期的に複数の人から注視されている可能性が高い．本研究では，そうした瞬間的または長期的に複数人の頭部方向前方にある地点を（共通の）注視点と定義する．注視点は以下の性質を持つ:

- 注視点は空間中に複数存在しうる．
- 注視点は時間変化する．すなわち，発生，移動，消滅する．

また，性質ではないが，注視点は空間の静的な対象に発生

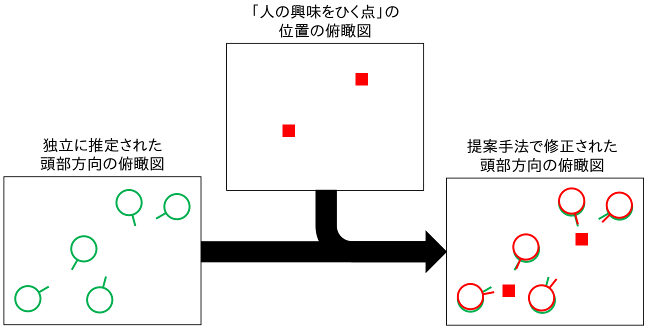


図 3 人の注視を「人の興味をひく点」の方向に修正するアイデアの模式図．俯瞰図における円は人頭部位置を表し，円から伸びる線分は頭部方向を表している．

するとは限らず，走っている車のような動体や，向かい合っている複数人の間の空間上にも発生しうるということにも留意いただきたい．本手法では，この注視点をアピアランスを用いたおおまかな推定頭部方向から求め，注視点を用いて頭部方向を修正することで，両者の同時最適化を行う．本研究では，すべての人は空間中のいずれかの注視点を注視していると仮定して頭部方向の修正を行う．実際の歩行者は，必ずしも意識的な注視を行っているとは限らず，移動方向に頭部方向を向けていることが多い．そのような歩行者は「無限遠の注視点を注視している」と仮定することで，問題を簡単化する．

2. 関連研究

2.1 頭部方向推定

映像中の人の頭部方向を推定する研究は古くから行われている [3], [4], [5], [6], [7]．主な手法としては，頭部方向のラベル付けがされている画像のアピアランスから学習を行い，クラス分類や回帰を行うもの [4], [5] と，目元や鼻，口元といった顔の特徴点を検出することで顔の向きや 3 次元形状を用いて角度を求めるもの [6], [7] がある．これらの方向推定器は，入力する頭部画像の解像度が低いほど精度が低下することがわかっており，特に顔の特徴点を使用する手法は，特徴点の検出に失敗する可能性が高くなる．

このため，低解像度の顔画像に対する頭部方向推定では，頭部方向をいくつかのクラスに分割してクラス識別する問題として扱われることが多い．Robertson ら [8] は，頭部方向 360° を 45° 区切りの 8 クラスに 8 分割し，アピアランスを学習してクラス識別を行う手法を提案した．Benfold ら [9] は，歩行者の頭部方向は移動方向と一致することが多いと仮定し，固定視点から撮影した動画から得られる大量の頭部画像に自動的に移動方向のクラスラベルを付与することで，シーンに特化した学習データを効率的に得られることを示した．本研究でもこれらの手法に倣い，頭部方向最適化に利用する入力頭部方向として，クラス分類器で得られた離散的な頭部方向を利用することとする．

なお，単一の RGB カメラのみを用いた頭部方向推定の

ほかに，複数の RGB カメラを同時に利用して頭部方向推定の精度を高める手法 [10] や，近年市販されるようになったデプスカメラを用いて頭部方向推定を行う手法 [11] も研究されている．本研究の立場は，近年の監視カメラの普及を背景としているため，すでに設置されているカメラのみで頭部方向推定を行うことを目指し，頭部方向推定の精度の向上のために上記の機器を追加で導入することは考えないものとする．

2.2 注視点推定

複数の研究で，映像からの頭部方向推定の結果を利用して，人が注視している対象の推定が行われている [2], [12]．これらの研究では，人の注視行動について，

(1) 人は頭部方向の先にある対象を注視している可能性が高い
(2) 人は他の人が注視している対象を注視している可能性が高い

という仮定を用いることで，効果的な注視点推定を行っている．Fathi ら [2] は，複数人がインタラクションを行っているシーンの映像から，共通の人を注視しているグループを検出する手法を示した．Park ら [12] は，電気的な引力を発生させる電荷のようなもの（Social Charge）が人の注視を引き起こしていると仮定し，複数人に共通する注視から電荷の位置を求め，注視点の推定を行った．しかし，これらの手法は，会議やパーティーのような，複数人が向かい合っていたり共通の点を注視したりする状況を想定し，動画に対しても各フレーム独立に注視点推定を行っている．このため，互いに無関係な多数のグループが存在して共通の注視がまれにしか発生しない一般的な環境には適用が難しい．

本研究では，これらの既存手法で用いられているような仮定を引用しつつ，既存手法では適用が難しい，一般的な環境での注視点推定手法を検討する．具体的な解決方法として，以下の 2 つの仮定を用いてこの問題を解決する:

(1) 個人の注視の時間連続性に着目し，個人がある時刻の前後で注視している点はその時刻においても注視点である可能性が高いと仮定し，注視点推定に前後時刻の頭部方向も用いる．
(2) 移動方向と頭部方向との角度の差が大きい人は何かを注視している可能性が高い（図 4）と仮定し，注視点推定における重みを大きくする．

仮定（1）は複数人による共通の注視が発生しにくい環境に対して，前後の時刻の同一人物を「共通の注視を行っている複数人物」として扱うことで，注視推定の枠組みに組込むことを可能にする．仮定（2）は，多くの人が移動方向を向いている街頭のような環境で，実際に意識的に見ているわけではないにも関わらず移動方向の先を注視点と推定されるのを避けるはたらきをする．

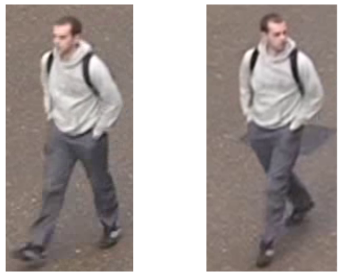


図 4 移動方向を向いている人（左）と移動方向を向いていない人（右）．左が実際に何かを見ているのか，何にも興味を持っていないか不明瞭な一方，右は何かに興味をひかれて注視している可能性が高い．

3. 頭部方向および注視点の最適化手法

本章では，推定頭部方向および推定注視点の同時時系列最適化手法について述べる．人頭部の回転における自由度は 3（yaw，pitch，roll）であるが，本研究では，頭部方向推定の精度の都合上，鉛直方向を軸とする回転（yaw）のみを考慮し，水平面上における頭部方向推定および注視点推定を行う．また，以降の議論では，便宜上動画に現れる人に j などの ID を付与するが，この ID の付与は軌跡に対して唯一であるが，個人に唯一ではなくてよいものとする．すなわち，A という個人が一度動画上から消失して再出現した場合，消失する前の ID を j，再出現後の ID を $k(\neq j)$ と付与することがある．

最適化は，注視点推定のフェーズと頭部方向最適化のフェーズを交互に繰り返すことによって行う．最適化の流れを図 5 に示す．注視点推定のフェーズでは，すべての時刻 t における人 $j \in \{1, 2, ..., J\}$ の水平面上の位置 $\boldsymbol{p}_{j,t} \in \mathbb{R}^2$，頭部方向 $\boldsymbol{o}_{j,t} \in \mathbb{R}^2$，移動方向 $\boldsymbol{v}_{j,t} \in \mathbb{R}^2$ が与えられたとき，時刻 $\hat{t}$ における注視点の個数 $I_{\hat{t}}$ および注視点の位置 $\{\boldsymbol{r}_{i,\hat{t}} \in \mathbb{R}^2\}_{i=1}^{I_{\hat{t}}}$ を求める．また，$(\boldsymbol{p}_{j,t}, \boldsymbol{o}_{j,t}, \boldsymbol{v}_{j,t})$ は人 j が動画中に存在している間のみ定義される．すなわち，人 j が時刻 $t_{emerge,j}$ に動画中に現れ，時刻 $t_{dissolve,j}$ に消失したとき，$(\boldsymbol{p}_{j,t}, \boldsymbol{o}_{j,t}, \boldsymbol{v}_{j,t})$ の定義域は $[t_{emerge,j}, t_{dissolve,j}]$ である．

頭部方向最適化のフェーズでは，時刻 $\hat{t}$ について，$\{(\boldsymbol{p}_{j,\hat{t}}, \boldsymbol{o}_{j,\hat{t}}, \boldsymbol{v}_{j,\hat{t}})\}_{j=1}^{J}$ および $\{\boldsymbol{r}_{i,\hat{t}} \in \mathbb{R}^2\}_{i=1}^{I_{\hat{t}}}$ が与えられたときの，各人の最適な頭部方向 $\{\boldsymbol{o}^*_{j,\hat{t}}\}_{j=1}^{J}$ を求める．次の注視点推定のフェーズでは，頭部方向を頭部方向最適化フェーズで求めたものに更新して推定を行う．

3.1 注視点推定

注視点を推定する上で，水平面座標系の任意の座標 $\boldsymbol{x}$ における「注視度= 注視点らしさ」を求める．時刻 t における座標 $\boldsymbol{x}$ の注視度は，時刻 t で観測されているすべての人の，$\boldsymbol{x}$ を見ている可能性の度合いの合算として算出する．ある人 j が時刻 t で座標 $\boldsymbol{x}$ を見ている可能性の度合い（以

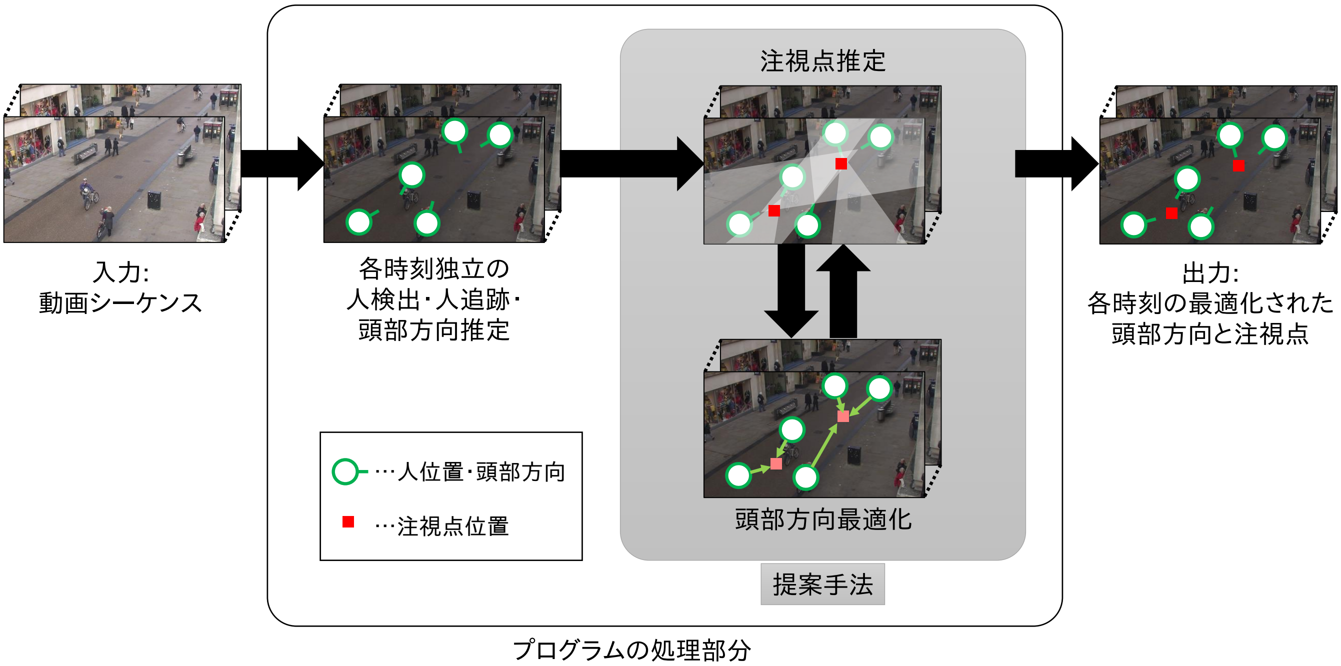


図 5　同時最適化の流れ.

下，個人の注視度と言う）を，以下の 2 つの項

- 空間項
- 移動方向項

の関数として定義する．空間項とは，位置関係に起因する個人の注視度への影響を表す項である．人 j の位置 $\boldsymbol{p}_{j,t}$，頭部方向 $\boldsymbol{o}_{j,t}$ と座標 $\boldsymbol{x}$ の関係として，明らかに，$\boldsymbol{x}$ が人 j の正面にあるほど，見ている可能性が高いと考えられる（図 6）．空間項 $c((\boldsymbol{o}_{j,t}, \boldsymbol{p}_{j,t}), \boldsymbol{x})$ を以下で定義する:

$$c((\boldsymbol{o}_{j,t}, \boldsymbol{p}_{j,t}), \boldsymbol{x}) = \max\left\{\frac{\boldsymbol{o}_{j,t}^{\mathrm{T}} \cdot (\boldsymbol{x} - \boldsymbol{p}_{j,t})}{\|\boldsymbol{o}_{j,t}\| \, \|\boldsymbol{x} - \boldsymbol{p}_{j,t}\|}, 0\right\}. \quad (1)$$

一方，移動方向項とは，移動方向と頭部方向に起因する，個人の注視度への影響を表す項である．2.2 節で仮定したように，人 j の移動方向 $\boldsymbol{v}_{j,t}$ と頭部方向 $\boldsymbol{o}_{j,t}$ との角度の差が大きい人ほど，実際に座標 $\boldsymbol{x}$ を見ている可能性が高いと考えられる．移動方向項 $d(\boldsymbol{v}_{j,t}, \boldsymbol{o}_{j,t})$ を以下で定義する:

$$d(\boldsymbol{v}_{j,t}, \boldsymbol{o}_{j,t}) = \begin{cases} \frac{\|\boldsymbol{v}_{j,t} \times \boldsymbol{o}_{j,t}\|}{\|\boldsymbol{v}_{j,t}\| \|\boldsymbol{o}_{j,t}\|} & \text{if } \boldsymbol{v}_{j,t} \cdot \boldsymbol{o}_{j,t} \geqq 0 \\ 1 & \text{otherwise} \end{cases}. \quad (2)$$

以上の 2 つの項で個人の注視度を定義する．個人の注視度 $e_{j,t}(\boldsymbol{x})$ は，移動方向項と空間項の単純な積で与えることとする:

$$e_{j,t}(\boldsymbol{x}) = d(\boldsymbol{v}_{j,t}, \boldsymbol{o}_{j,t}) \cdot c\left((\boldsymbol{o}_{j,t}, \boldsymbol{p}_{j,t}), \boldsymbol{x}\right). \quad (3)$$

最後に，座標 $\boldsymbol{x}$ における注視度 $E_{\hat{t}}(\boldsymbol{x})$ を定義する．注視度の計算には，2.2 節で述べたように，時刻 $\hat{t}$ の前後時刻も含めて注視度の合算を行うとする．このとき，注視度は以下の式で与えられる:

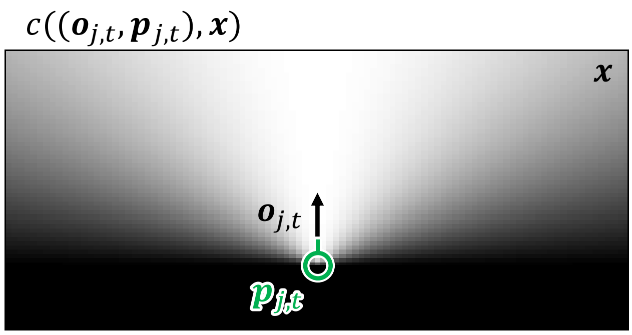


図 6　人位置 $\boldsymbol{p}_{j,t}$ および頭部方向 $\boldsymbol{o}_{j,t}$ に対する，「座標 $\boldsymbol{x}$ が人 j に見られている可能性の度合い」のヒートマップ.

$$E_{\hat{t}}(\boldsymbol{x}) = \sum_{t=\hat{t}-t_W/2}^{\hat{t}+t_W/2} \sum_{j=1}^{J} e_{j,t}(\boldsymbol{x}). \quad (4)$$

ただし，t_W は注視度の合算を行う時刻のウインドウサイズである．

以上で定義した注視度を用いて，各時刻 t における任意の座標における注視度を計算し，注視度が極大となる座標を注視点 $\{\boldsymbol{r}_{i,t}\}_{i=1}^{I_t}$ とする．時刻 t の注視点の個数 I_t は，座標平面上における注視度の極大値の個数によって変化することに注意していただきたい．

3.2　頭部方向最適化

頭部方向の最適化では，本研究における「すべての人は空間中のいずれかの注視点を注視している」という仮定に従い，単純に現在推定されている人の頭部方向から最もずれの小さい位置にある注視点の方向に修正する．すなわち，

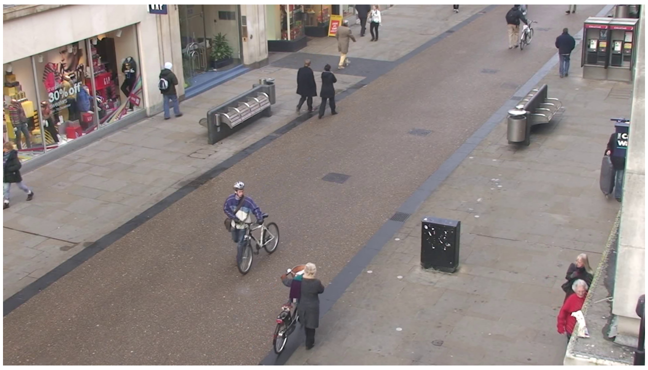


図 7　TownCenter データセット（フレーム番号 50）.

$$\boldsymbol{o}_{j,t} \leftarrow \boldsymbol{o}^*_{j,t} = \frac{\boldsymbol{r}^j_t - \boldsymbol{p}_{j,t}}{\left\|\boldsymbol{r}^j_t - \boldsymbol{p}_{j,t}\right\|}. \tag{5}$$

ただし,

$$\boldsymbol{r}^j_t = \underset{\boldsymbol{r}\in\{\boldsymbol{r}_{i,t}\}_{i=1}^{I_t}}{\arg\max} \frac{\boldsymbol{o}^{\mathrm{T}}_{j,t}\cdot(\boldsymbol{r}-\boldsymbol{p}_{j,t})}{\|\boldsymbol{o}_{j,t}\|\,\|\boldsymbol{r}-\boldsymbol{p}_{j,t}\|}. \tag{6}$$

4. 評価

本章では，提案手法の評価のために行った評価について述べる．評価に使用したデータセットの関係上，本稿では定性的評価のみ述べる．

4.1 データセット

提案手法を適用する動画データとして，Benfold ら [1] が公開している Town Centre データセット*1 を使用した．4501 フレームからなる動画で，ほとんどのフレームで 10 人以上の歩行者が観測される．動画像平面上で右上から左下にかけて道が伸び，多くの歩行者が道に沿って移動しているため，複数人の注視線が平行になることが多い（図 7）．このため，提案手法が解決を目指している，複数人による共通の注視が発生しにくい環境の例として適している．4501 フレームのうち，評価にはフレーム番号が 5 の倍数のもの 901 枚を使用した．

Town Centre のアノテーションとして，動画座標系での身体領域，頭部領域および個人の ID が用意されている．以下の評価用システムでは，これらの情報はあらかじめ与えられているものとして利用している．実際に未知の動画に対して本手法を適用する場合，これらの情報は動画から取得する必要がある．身体領域の検出および個人 ID の追跡については例えば [13] を，身体領域中からの頭部領域の検出には例えば HOG 特徴量 [14] を用いて頭部のアピアランスを学習した検出器を利用できる．

4.2 評価用システム

評価用システムは以下の 3 つのパートからなる:

*1 http://www.robots.ox.ac.uk/ActiveVision/Research/Projects/2009bbenfold_headpose/project.html

図 8　BMVC2009 データセットの一部．大部分の画像がこのような低解像度の画像である．アノテーションの角度は，左から 135°，161°，280°，307°．

(1) 動画の各フレームに対してアピアランスを用いて頭部方向推定を行う
(2) 動画座標系での人位置，頭部方向をホモグラフィ変換を用いて水平面座標に変換する
(3) 提案手法を用いて頭部方向および注視点を推定，同時最適化を行う

これらについて，それぞれ以下で述べる．

4.2.1 アピアランスを用いた頭部方向推定

頭部方向および注視点の最適化を行う初期値として，アピアランスを用いて動画の各フレームの頭部方向を推定する．頭部方向推定には，水平面上 360° を 22.5° ずつ 16 のクラスで分割し学習した識別器を用いる．入力として頭部領域の画像を与えると，16 方向のいずれかを推定頭部方向として出力する．学習の特徴量として HOG[14] を用い，one-versus-one マルチクラス SVM で学習した．

学習データとしては，Benfold ら [1] が公開している BMVC2009*1を用いた．BMVC2009 は，人頭部をやや斜め上方から撮影した画像およそ 1500 枚からなるデータセットである．一辺の大きさが 10〜188 ピクセル，平均 45 ピクセルの低解像度の画像であり，正面顔の画像を 0° として，0° から 359° まで 1° 単位でのアノテーションが与えられている（図 8）．このデータセットのすべての画像を，0° ± 12.25° をクラス 0，22.5° ± 12.25° をクラス 1，…と 16 のクラスに分割して学習した．識別器の精度の確認のために，学習したデータセットそのものでテストを行った結果を図 9 に示す．

4.2.2 ホモグラフィ変換

水平面座標系での注視点推定を行うために，動画座標系上で得られている人位置，頭部方向および移動方向を，水平面座標系に変換する．TownCentre データセットは地面平面に対してカメラキャリブレーションが行われているため，ここでは動画座標系を地面平面座標系に変換する．変換に利用するホモグラフィ行列 H は，カメラキャリブレーション情報を用いてあらかじめ計算しておく．

変換後の水平面座標系における人位置 $\boldsymbol{p} = (p_x, p_y)$ は，動画座標系における人身体領域の下部中央の座標をホモグラフィ変換することで得る．すなわち，動画フレームの身体領域の左上および右下座標がそれぞれ (l, t)，(r, b) で与えられているとき，以下で計算される:

		estimated															
		0°	22.5°	45°	67.5°	90°	112.5°	135°	157.5°	180°	202.5°	225°	247.5°	270°	292.5°	315°	337.5°
ground truth	0°	0.3636	0.1653	0.0083	0.0248	0.0083	0	0.0165	0.0083	0.0248	0.0331	0.0083	0.0248	0.0165	0.0496	0.0744	0.1736
	22.5°	0.3167	0.3667	0.1	0	0.0167	0	0.0333	0	0.0333	0.0167	0.0167	0.0167	0	0	0.0333	0.05
	45°	0.1786	0.1607	0.1964	0.1071	0.0893	0.0179	0.0179	0.0179	0.0357	0	0.0357	0.0357	0	0.0714	0	0.0357
	67.5°	0.0732	0.0976	0.0732	0.2683	0.1463	0.0488	0.0488	0.0244	0	0.0488	0	0.0244	0	0.0732	0	0.0732
	90°	0.0854	0.061	0.0488	0.061	0.3171	0.1463	0.0488	0.0122	0.0366	0.0366	0	0.061	0	0.0366	0.0244	0.0244
	112.5°	0.0156	0.0156	0.0156	0.0313	0.1406	0.3281	0.2813	0.0781	0.0313	0	0	0.0156	0	0.0313	0	0.0156
	135°	0.0217	0.0217	0.0145	0.0072	0.0362	0.1087	0.4493	0.2029	0.0652	0.0217	0.0072	0.0145	0	0	0.0217	0.0072
	157.5°	0.0144	0.0144	0	0.0072	0.0216	0.1223	0.2806	0.3669	0.0576	0.0144	0.0072	0.0504	0.0144	0.0216	0	0.0072
	180°	0.0417	0.0238	0	0	0.006	0.0417	0.1607	0.2917	0.25	0.0833	0.0119	0.0536	0.006	0.0119	0.0119	0.006
	202.5°	0.0233	0.0116	0.0116	0.0116	0.0116	0.0698	0.0349	0.1047	0.1744	0.314	0.0349	0.1047	0.0465	0.0116	0.0233	0.0116
	225°	0.0299	0.0299	0	0.0149	0.0149	0.0299	0.0597	0.1045	0.1045	0.1194	0.0746	0.2537	0.1194	0.0299	0	0.0149
	247.5°	0.0149	0	0	0	0.0448	0	0.1045	0.0597	0.0597	0.0746	0.0597	0.3284	0.1194	0.0896	0.0149	0.0299
	270°	0.0492	0.0328	0	0	0.0328	0.0328	0.0492	0.082	0.0328	0.0492	0.0328	0.1639	0.2623	0.1311	0.0328	0.0164
	292.5°	0.0408	0.0408	0.0204	0.0102	0.0306	0	0.0408	0.0306	0.0204	0	0.0306	0.051	0.1122	0.3265	0.1735	0.0714
	315°	0.0945	0.0551	0	0.0079	0.0157	0	0.0157	0.0157	0	0.0079	0.0157	0.0079	0.0236	0.1654	0.4094	0.1654
	337.5°	0.19	0.15	0.02	0.01	0	0.02	0	0.02	0.01	0	0	0.03	0.01	0.17	0.1	0.27

図 9 データセットでテストを行った結果の confusion matrix．クラスを正しく推定できたものは全体の 32%，正答または隣のクラスに誤識別したものは全体の 63%であった．

$$\begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} = H \begin{bmatrix} (l+r)/2 \\ b \\ 1 \end{bmatrix}. \qquad (7)$$

また，変換後の水平面座標系における人頭部方向 $\boldsymbol{o} = (o_x, o_y)$ は，動画座標系における頭部方向ベクトルをホモグラフィ変換することで得る．ベクトルのホモグラフィ変換は，ベクトルの始点および終点をそれぞれホモグラフィ変換することで与えられる．変換前のベクトルの始点は変換後の人位置になる点，すなわち $((l+r)/2, b)$ である必要があることに注意する．動画座標系での頭部方向ベクトルを (u_x, u_y) とすると，変換後のベクトルの終点 (p_x^+, p_y^+) は，

$$\begin{bmatrix} p_x^+ \\ p_y^+ \\ 1 \end{bmatrix} = H \begin{bmatrix} (l+r)/2 + u_x \\ b + u_y \\ 1 \end{bmatrix}. \qquad (8)$$

となり，変換後の人頭部方向 $\boldsymbol{o}$ は，

$$(o_x, o_y) = \left(\frac{p_x^+ - p_x}{\|\boldsymbol{p}^+ - \boldsymbol{p}\|}, \frac{p_y^+ - p_y}{\|\boldsymbol{p}^+ - \boldsymbol{p}\|} \right). \qquad (9)$$

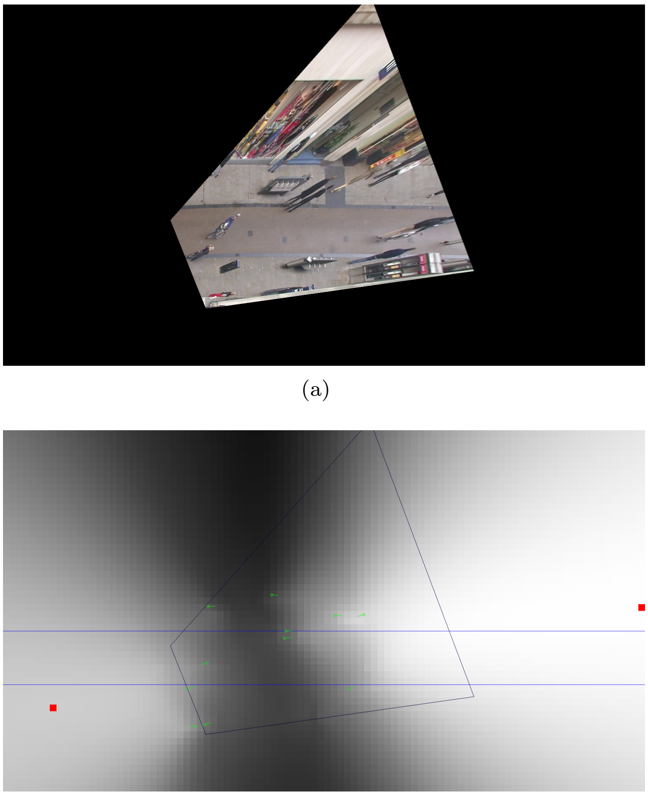
ただし，$\|\boldsymbol{p}^+ - \boldsymbol{p}\| = \sqrt{(p_x^+ - p_x)^2 + (p_y^+ - p_y)^2}$.

最後に，人 j の移動方向 $\boldsymbol{v}_{j,t}$ を求める．$\boldsymbol{v}_{j,t}$ は以下で与える:

$$\boldsymbol{v}_{j,t} = \begin{cases} 0 & \text{if } t = t_{emerge,j} \\ \frac{\boldsymbol{p}_{j,t} - \boldsymbol{p}_{j,t-1}}{\|\boldsymbol{p}_{j,t} - \boldsymbol{p}_{j,t-1}\|} & \text{otherwise} \end{cases}. \qquad (10)$$

4.2.3 同時最適化

4.2.2 節で得た，地面平面座標系での人データ $(\boldsymbol{p}_{j,t}, \boldsymbol{o}_{j,t}, \boldsymbol{v}_{j,t})_{j=1}^{J}$ を用いて，提案手法を適用する．注視点推定のフェーズでは，映像をホモグラフィ変換したもの（図 10(a)）を確認しながら手動で与えた定義域に対して，各格子点上で注視度を計算した．注視度を合算する時刻のウインドウサイズは $t_W = 20$ とした．各格子点の注視度に対して，5×5 の平均値フィルタを適用し，Non-maxima supression を行って注視点を求めた．ただし，平均値フィルタ適用後の注視度の最大値 $\max_{\boldsymbol{x}}\{E_t(\boldsymbol{x})\}$ に対して，注



(a)

(b)

図 10 (a) 映像をホモグラフィ変換したものと (b) 注視度の計算結果を可視化したもの（いずれもフレーム番号 50）．(a) と (b) の画像上の位置は対応している．グリッド状の彩色が注視度であり，明度が高いほど注視度が高く，赤色の座標が注視点を表す．緑色の円と線分は人の位置および頭部方向を表す．

視度が 30%未満である点は注視点から除外した．注視度の計算結果を可視化したものを図 10(b) に示す．頭部方向最適化のフェーズでは，以上で得た注視点を用いて，頭部方向の修正を行った．両方のフェーズを 10 回交互に行った結果，フェーズ間の同時刻で注視点の移動がほとんど見られなくなったのを確認した．

4.3 考察

手法の評価で着目する点は，以下の 2 点である：（1）注視点として推定されるべき場所が正しく注視点として推定

されたか，(2) 推定された注視点によって尤もらしい頭部方向の修正が行われたか．これらの点について，評価用システムの実行結果を手動でチェックし，定性的な評価をまとめる．

4.3.1 注視点として推定されるべき場所が正しく注視点として推定されたか

使用した TownCentre データセットでは，すべての時刻において通路の双方向に向かって歩いている歩行者が確認される．このため，すべての時刻で，注視点が図のように通路の両端に現れることが期待される．実際に注視点の位置が通路の両端に推定されたものは，901 フレーム中 591 フレームであった．通路の両端に注視点が現れなかった時刻の，最適化途中の注視度の変化を確認したところ，いずれの時刻においても，図 11 のようにはじめは通路の両端に注視点が推定されていたものの，最適化のフェーズが進む際に中央付近の注視点に人々の頭部方向が修正された結果，通路端の注視度が減少していることが確認された．実用上は，注視点が両端に現れないことはあまり大きな問題にならない．通路の両端のような，常に注視点が発生しうる地点は，「静的な注視点」としてあらかじめ設定しておけばよい．静的な注視点は，注視点推定フェーズで求めた「動的な注視点」とともに，頭部方向最適化の際に頭部方向を向ける先の注視点候補として扱う．

通路端のほか，複数人が同時に注視している位置も，注視点として推定されるべきである．動画上で複数人による共通の注視が確認されるフレームを確認したところ，注視点が検出されたもの，検出されなかったものがそれぞれ存在した（図 12）．注視点が検出されたものと検出されなかったものの違いは，注視に関係した人の，頭部方向と移動方向との差が大きかったか否か，が影響していると考えられる．提案した注視度計算手法では，移動方向と頭部方向が一致している人は注視度に全く影響を与えない．実際には何らかの注視を行っている可能性があるため，移動方向と頭部方向が一致していても，ある程度は注視度計算に考慮するような関数を設計するべきである．

4.3.2 推定された注視点によって尤もらしい頭部方向の修正が行われたか

アピアランスから推定した頭部方向と，最適化手法適用後の頭部方向の結果を比較した．主に確認された最適化による変化は以下の 2 つである．

- アピアランスからの推定の大幅な誤差は修正されず
  注視点が通路の両端に現れた時刻では，3 点以上の注視点が検出されている場合でも，多くの頭部方向でこのような修正が行われた．これは，頭部方向最適化のフェーズの注視点選択アルゴリズムに，単純に修正する角度が小さい注視点を選択しているため，遠くにある注視点が相対的に選ばれやすくなったためと考えられる（図 13）．またこれが原因で，最適化の初期値であるアピアランスの頭部方向推定結果が大きく誤っている場合に，最適化手法を適用しても推定結果の改善がみられなかった．前者の解決には，角度だけでなく距離の近さも考慮するような，効果的な注視点選択アルゴリズムを選択する必要がある．後者の解決には，アピアランスでの注視点推定に移動方向を考慮し，移動方向 180° 以内から頭部方向を行うことで対処できる．
- 唯一の注視点に向けて誤修正
  最適化の結果，注視点が通路中央付近ただ 1 つになったフレームが 131 フレームあった．この注視点に向けて頭部方向を修正した結果，移動方向を向いている歩行者などについて，最適化前よりも頭部方向推定の精度が悪化した．この問題は，先に挙げた，静的な「無限遠の注視点」を設定することで解決可能である．

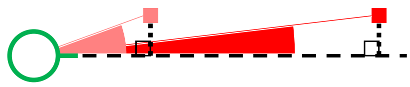


図 13 頭部方向最適化で遠くにある注視点のほうが選ばれやすい原因を可視化したもの．頭部方向の軸からの距離が同じであれば，より遠くにある点のほうが頭部方向からの角度の差が小さい．

5. おわりに

監視カメラなどで撮影された低解像度の動画中の人々の頭部方向推定のために，頭部方向と注視点を同時最適化する手法を提案した．評価実験を行い，定性的評価によって一定の有効性の確認と改善すべき点についての知見を得た．

本研究で検討の余地がある点は 3 点ある．(1) 定量的評価．データセットの都合上，本研究では定性的評価を行うに留めた．将来研究では，頭部方向のアノテーションが与えられているデータセットでの頭部方向推定や，注視点を空間中に明示的に用意したデータセットでの注視点推定の精度について，定量的評価を行いたい．(2) 注視度の計算式の検討．本研究では単位ベクトルの内積および外積という単純な式を用いて注視度を計算した．将来研究では，定量的評価を用いてどのような式がより高精度に最適化を行えるか検討する．例えば，空間項の関数をよりピークが出やすい関数（フォン・ミーゼス分布など）に変更する，空間項を距離で減衰させる，などの変更を行って，最適化の結果の変化を調査する．(3) 注視点の重要度の導入．本研究では，すべての注視点を区別なく扱い，頭部方向最適化の際に最も角度の差が小さい注視点を選んで頭部方向の修正を行った．実際には，注視点ごとに注視のされやすさや重要度が異なるほか，特定の性質を持つ人々に注目されやすい注視点（女性誌の広告は女性により注視されやすい等）もある．将来研究では，このような注視点特有の性質についても考慮する方法を考えたい．

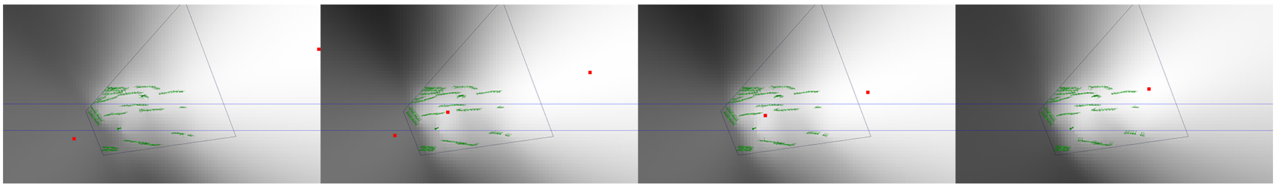


図 11 注視点の個数の減少の可視化
（左から 1 回目，2 回目，3 回目，5 回目の注視点推定フェーズ，フレーム番号 4240）．
赤色で表されている注視点がはじめ通路の両方向にあったものが，フェーズが進むにつれて注視が中央に修正されていき，通路の両方向の注視点が消失している．

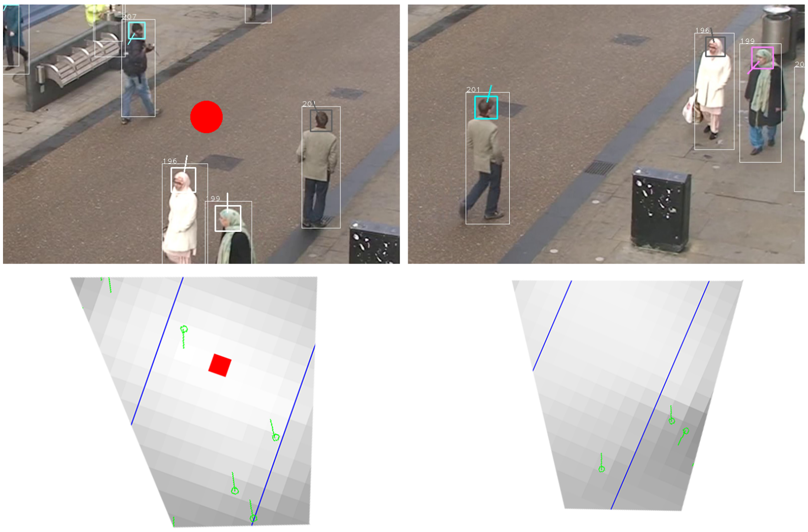


図 12 複数人が向かい合って注視点が検出されたもの（左上，フレーム番号 4145）およびその注視度ヒートマップ（左下）と，検出されなかったもの（右上，フレーム番号 3980）およびその注視度ヒートマップ（右下）．左は男性が別の立ち止まっている男性に話しかけられて振り返っている．右は男性が女性 2 人に話しかけながら前進している．

参考文献

[1] Benfold, B. and Reid, I.: Guiding Visual Surveillance by Tracking Human Attention, *Proceedings of the 20th British Machine Vision Conference* (2009).

[2] Fathi, A., Hodgins, J. K. and Rehg, J. M.: Social interactions: A first-person perspective, *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, pp. 1226–1233 (2012).

[3] Murphy-Chutorian, E. and Trivedi, M. M.: Head Pose Estimation in Computer Vision: A Survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 31, No. 4, pp. 607–626 (online), DOI: 10.1109/TPAMI.2008.106 (2009).

[4] Jones, M. and Viola, P.: Fast Multi-view Face Detection, Technical Report TR2003-96, MERL - Mitsubishi Electric Research Laboratories, Cambridge, MA 02139 (2003).

[5] Li, Y., Gong, S., Sherrah, J. and Liddell, H.: Support vector machine based multi-view face detection and recognition (2004).

[6] Gee, A. H. and Cipolla, R.: Determining The Gaze Of Faces In Images, *Image and Vision Computing*, Vol. 12, pp. 639–647 (1994).

[7] Uřičář, M., Franc, V. and Hlaváč, V.: Detector of Facial Landmarks Learned by the Structured Output SVM, *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications* (Csurka, G. and Braz, J., eds.), Vol. 1, Portugal, SciTePress — Science and Technology Publications, pp. 547–556 (2012).

[8] Robertson, N. and Reid, I.: Estimating gaze direction from low-resolution faces in video, *Proc. 9th IEEE Eur. Conf. on Computer Vision, Graz, Austria* (2006).

[9] Benfold, B. and Reid, I.: Unsupervised Learning of a Scene-Specific Coarse Gaze Estimator, *ICCV* (2011).

[10] Muñoz-Salinas, R., Yeguas-Bolivar, E., Saffiotti, A. and Medina-Carnicer, R.: Multi-camera head pose estimation, *Machine Vision and Applications*, Vol. 23, No. 3, pp. 479–490 (2012).

[11] Fanelli, G., Dantone, M., Gall, J., Fossati, A. and Van Gool, L.: Random Forests for Real Time 3D Face Analysis, *Int. J. Comput. Vision*, Vol. 101, No. 3, pp. 437–458 (2013).

[12] Park, H. S., Jain, E. and Sheikh, Y.: Predicting Primary Gaze Behavior using Social Saliency Fields, *International Conference on Computer Vision* (2013).

[13] Andriyenko, A. and Schindler, K.: Multi-target Tracking by Continuous Energy Minimization, *CVPR* (2011).

[14] Dalal, N. and Triggs, B.: Histograms of oriented gradients for human detection, *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1, pp. 886–893 vol. 1 (online), DOI: 10.1109/CVPR.2005.177 (2005).