THE IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS (JAPANESE EDITION)

電子情報通信学会論文誌 D

情報・システム

DOI:10.14923/transinfj.2018JDR0003
早期公開日:2018/11/01

本PDFは，早期公開版である．本論文を引用する場合には，電子情報通信学会和文論文誌投稿のしおり(情報・システムソサイエティ)の「８.早期公開」を参照すること．

情報・システムソサイエティ

一般社団法人 電子情報通信学会

THE INFORMATION AND SYSTEMS SOCIETY
THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS

サーベイ論文

動画像を用いた経路予測手法の分類

平川　翼[†a)]　山下　隆義[†b)]　玉木　徹[††c)]　藤吉　弘亘[†d)]

Survey on Vision-based Path Prediction

Tsubasa HIRAKAWA[†a)], Takayoshi YAMASHITA[†b)], Toru TAMAKI[††c)], and Hironobu FUJIYOSHI[†d)]

あらまし　経路予測とは，歩行者や自動車などの予測対象が未来にどのような経路を移動するかを推定する技術である．コンピュータビジョンにおける経路予測問題では，主に動画像のみを入力とするため，移動経路の予測のみならず予測対象の状態や周囲の環境なども動画像から推定する必要がある．予測が行われるシーンの環境理解や予測手法には，これまでに多くのアプローチが提案されている．本稿では動画像を入力とする経路予測手法についてサーベイを行い，動画像からの特徴抽出法と予測手法について体系的にまとめる．また，経路予測手法を定量的に評価するために使用されるデータセットおよび評価指標についても紹介する．

キーワード　経路予測, 行動予測, データセット, サーベイ

1. はじめに

経路予測とは歩行者や自動車などのある動画像中の予測対象が未来の時刻にどのような経路で移動するかを推定する技術である．動画を用いた経路予測は異常行動の検出 [1], [2] や自動車の自動運転技術 [3]〜[5], ロボットの自律制御 [6]〜[8] などの様々な分野へ応用できる可能性を秘めているため，近年盛んに研究が行われている問題設定の一つである．

動画像を用いた経路予測は他の画像認識の問題と比べて挑戦的な問題設定である．一般的なコンピュータビジョンの問題では，画像または動画像のみを入力とする場合が多い．経路予測も同様に，周囲の環境や予測対象の向きなどの情報を動画像から推定する必要がある．これらの情報の推定には人物検出 [9], [10] や属性推定 [11]，セマンティックセグメンテーション [12] などの技術が用いられており，経路予測は既存のコンピュータビジョン技術の上に存在する問題となっている．また，動画像を用いた経路予測をはじめとする予測という問題の難しさは，予測時点での観測が取得できないことに大きく起因している．例えば，上記の人物検出や追跡といった問題では，過去から現在までの観測，すなわち動画像から抽出された情報を用いてその時刻に対応する対象の位置を推定する．一方，経路予測では未来の時刻の対象の位置を推定するため，推定する時刻に対応する情報は与えられない．そのため，現在までの観測に加えて周囲の環境や対象者の行動規則などの事前情報をうまく活用する必要がある．

経路予測はロボティクスの分野において古くから扱われている．駅や空港などの不特定多数の歩行者が存在する公共の場において，周囲の歩行者の動きを妨げることなくロボットを自律走行させるために歩行者の経路予測が行われており [6]，ロボット自身が周囲の環境を考慮しながら無駄なく移動するための経路を推定する経路計画 [13]〜[15] も経路予測の問題と捉えることができる．ロボティクスにおける経路予測では，ロボットに搭載されたカメラの映像に加えて，レーザレンジファインダ等で計測されたシーンの 3 次元データを使用することが多い．また，屋内などの限定的な環境でロボットを動作させる場合，シーン全体の環境が所与とされていることが多い．これを踏まえ，本論文では以下のような経路予測手法をサーベイの対象と

† 中部大学，春日井市
Chubu University, 1200 Matsumoto, Kasugai, 487-8501 Japan
†† 広島大学，東広島市
Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima, 739-8527 Japan
a) E-mail: hirakawa@mprg.cs.chubu.ac.jp
b) E-mail: takayoshi@isc.chubu.ac.jp
c) E-mail: tamaki@hiroshima-u.ac.jp
d) E-mail: fujiyoshi@isc.chubu.ac.jp

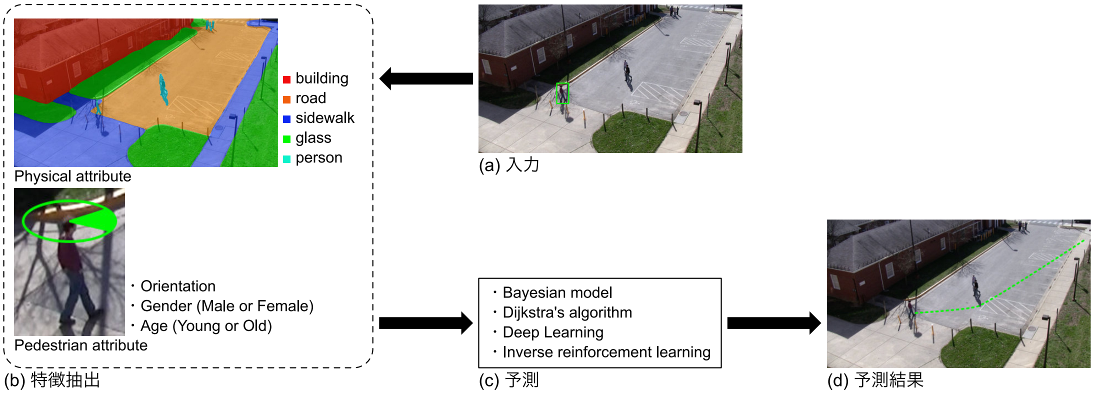


図 1　経路予測手法の処理の流れ．文献 [20] より改変．

する．

（1）　予測を行うシーンを撮影した動画像，静止画像，または予測対象が予測開始時点までに存在していた位置（座標または座標の系列）を入力とする手法．動画像または静止画像を用いることなく，LIDAR 等で計測した 3 次元データのみを入力とする経路予測手法 [16] は本サーベイの対象としない．

（2）　予測結果として予測対象の未来のある時刻での位置（座標）や未来の各時刻に対応する座標の系列，または予測対象がその位置を訪れる可能性を表現した確率分布を出力する手法．

未来の行動を予測する問題は経路予測だけではない．動画像から人物の動作カテゴリを推定する動作認識を時間的に遷移させた問題として早期認識と呼ばれる問題設定が存在する [17]～[19]．早期認識は動画像を入力として，未来の時刻にどのような行動が起こるかを推定する問題であるが，動作カテゴリなどの離散的なラベルを推定する問題設定である．したがって，早期認識は経路予測とは異なる問題設定であるため，本サーベイの対象としないことに注意されたい．

このように，コンピュータビジョンにおける動画像を用いた経路予測は，従来扱われてきた問題と比較して難しく挑戦的な問題である．この問題を解決するために，様々な予測手法が提案されている．提案されている手法は多岐に渡っているが，処理の流れは一貫している．図 1 に動画像を用いた経路予測手法の一般的な処理の流れを示す．動画像を用いた経路予測では，動画像または静止画像とそのシーン中の予測対象の位置または過去の数秒間の移動軌跡を入力とする．まず，入力された動画像から経路予測に有用な特徴を抽出する．その後，抽出した特徴を用いて予測手法を適用することで予測結果を出力する．この処理の流れの中で重要となるのは，図 1(b) 動画像からの特徴抽出部と図 1(c) 経路予測部である．特徴抽出部では主に，シーンを理解するための環境属性に対する特徴と歩行者などの予測対象に関する特徴抽出が行われる．経路予測部では様々な手法が提案されているが，基礎となるアプローチに基づき予測手法をベイズモデル，エネルギー最小化，深層学習，逆強化学習の 4 種類に分類することができる．

以上を踏まえ，本論文ではコンピュータビジョン分野における動画像を用いた経路予測手法についてサーベイを行い，経路予測に用いられる特徴抽出及び予測のアプローチについて体系的にまとめる(注1)．また，経路予測手法の実験評価に用いられるデータセットについても調査する．2 章では経路予測に用いられる特徴抽出について述べる．3 章では経路を予測する手法についていくつかのカテゴリに分類し，各カテゴリの予測手法の特徴について述べる．4 章にて経路予測の定量的評価に使用されるデータセットについて述べ，5 章で経路予測の結果を定量的に評価するための指標を紹介する．6 章にて本サーベイをまとめる．

2. 動画像からの特徴抽出法

歩行者が移動する際には，周囲の環境や自身の状態などの何らかの要因が暗に影響しており，その要因によって歩行経路が決定されると考えられる．動画像を

(注1)：本論文は文献 [21] の内容について，2018 年 4 月時点までに発表されている経路予測手法および定量的な評価指標を追加したものとなっている．

表 1　経路予測に用いる特徴抽出法の分類.

抽出対象	特徴の種類	代表的な手法
環境	シーンラベル	Stacked hierarchical labeling [22] Superpixel-based MRF [23] Fully Convolutional Networks [24], [25]
	コスト	Bag-of-Visual Words [26] Spatial Matching Network [27]
	シーン全体の特徴ベクトル	Pretrained AlexNet [28] Siamese Network [29]
予測対象	歩行者の位置	HOG + SVM detector [30] Faster R-CNN [31], [32]
	対象の向き	Bayesian orientation estimation [33] Orientation Network [27]
	対象の姿勢	OpenPose [34]
	身体的属性	AlexNet-based multi-task learning [35]
	対象物の特徴ベクトル	Mid-level patch features [36]

用いた経路予測においても，移動経路の決定に大きく影響するような情報を用いることで予測精度の向上が期待できる．しかし，動画像のみを入力とする経路予測では，このような情報を動画像から推定する必要があり，予測の前処理として動画像から何らかの特徴が抽出される．表 1 に経路予測に用いる特徴抽出法を示す．このように様々な特徴が抽出されており，これらは 1) 環境属性と 2) 予測対象に対する特徴抽出に大別することができる．そこで，本章ではシーンの環境属性及び予測対象の 2 つの特徴抽出について述べる．

2.1　環境属性に対する特徴抽出

予測対象が移動する際には，その周辺環境が大きく影響している．例えば，歩行者が移動する際にはシーンに存在する建物などの障害物を避けつつ，歩道を移動することが考えられる．また，自動車や自転車が予測対象であれば，交通ルールなどの社会的規範により車道のみを移動するであろう．このように，予測対象の移動は周囲の環境に依存しており，予測を行うシーンの環境に関する特徴を動画像から抽出する必要がある．

動画像からの環境を推定する技術として最も一般的なものはセマンティックセグメンテーションである [37]〜[40]．セマンティックセグメンテーションは画像中の各画素にオブジェクトのクラスを割り当てる技術である．これにより予測対象が移動可能な領域や障害物領域などを推定することが可能である．Kitani ら [37] は，歩行者の移動経路は車道や歩道，花壇，建造物などの物理的環境に大きく影響されていると仮定し，階層的に領域を分割するセグメンテーション手法 [22] を用いて特徴抽出を行なっている．この手法では，図 2 に示すように，各ラベルの事後確率を特徴

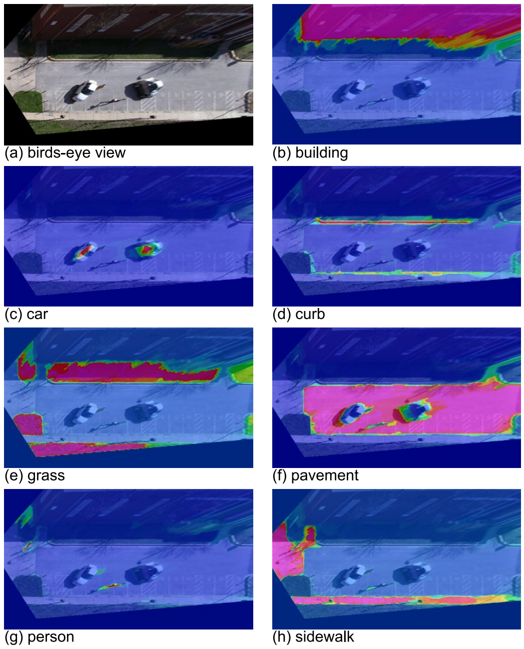


図 2　セマンティックセグメンテーションを用いた環境属性の例．文献 [37] より引用．

ベクトルの要素として用いることで，シーン全体の特徴マップを作成し，予測に用いている．一方，Ballan ら [38] は，微小領域中に含まれるシーンラベルから作成したヒストグラムをその領域に対する特徴ベクトルとして使用している．

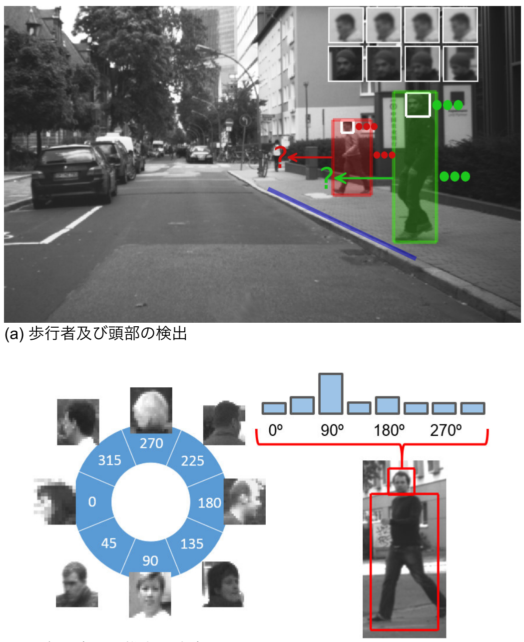
歩行経路に影響を与える環境属性を明示的に用いることなく，予測対象が移動する可能性をコストとして暗に表現する方法も存在する [27], [41]．この方法では，シーン中の微小領域毎に推定したコストからシーン全体のコストマップを作成し，経路予測に用いてい

表 2　経路予測手法の分類.

分類	文献	年代	手法	映像視点	座標系	入力	出力	特徴抽出 環境	特徴抽出 対象
Bayesian	Schneider and Gavrila [3]	2013	KF	車載	地上	座標	座標		
	Kooij et al. [5]	2014	DBN	車載	地上	動画	座標	✓	✓
	Ballan et al. [38]	2016	DBN	鳥瞰	画像	動画	座標	✓	
Energy minimization	Xie et al. [44]	2013	Dijkstra	監視カメラ	地上	動画	確率分布	✓	
	Walker et al. [41]	2014	Dijkstra	監視カメラ	画像	動画	確率分布	✓	
	Huang et al. [27]	2016	Dijkstra	監視カメラ	画像	画像	確率分布	✓	✓
DL	Shuai et al. [45]	2016	CNN	監視カメラ	画像	座標	座標		
	Alahi et al. [46]	2016	LSTM	鳥瞰	画像	座標	座標		
	Bartoli et al. [47]	2017	LSTM	鳥瞰	画像	座標	座標	✓	
	Fernando et al. [48]	2017	LSTM	鳥瞰	地上	座標	座標		
	Fernando et al. [49]	2017	LSTM	鳥瞰	地上	座標	座標		
	Lee et al. [40]	2017	RNN Enc.-Dec.	鳥瞰	画像	動画	座標	✓	
	Yagi et al. [50]	2018	CNN	一人称	画像	動画	座標		✓
IRL	Kitani et al. [37]	2012	IRL	監視カメラ	地上	動画	確率分布	✓	
	Lee and Kitani [51]	2016	IRL	鳥瞰	画像	動画	確率分布	✓	
	Bokhari and Kitani [52]	2016	IRL	一人称	地上	動画	確率分布		✓
	Rhinehart and Kitani [53]	2017	IRL	一人称	地上	動画	確率分布		✓
	Ma et al. [35]	2017	IRL	監視カメラ	画像	画像	確率分布		✓
	Rehder et al. [39]	2017	IRL	車載	地上	動画	確率分布	✓	
Others	Keller and Gavrila [4]	2014	optical flow	車載	地上	動画	座標		
	Rehder and Koloeden [54]	2015	Markov process	車載	地上	動画	座標	✓	
	Park et al. [42]	2016	Data driven	一人称	地上	動画	座標	✓	
	Su et al. [43]	2017	Data driven	一人称	地上	動画	座標	✓	
	Yamaguchi et al. [55]	2011	Social force	鳥瞰	地上	動画	座標		
	Robicquet et al. [56]	2016	Social force	鳥瞰	地上	動画	座標		

る．Walker ら [41] は，最近傍探索を用いて学習データから類似したテクスチャのパッチを検索し，そのコストを微小領域に割り当てることでシーン全体でのコストマップを作成している．Huang ら [27] は，Spatial Matching Network と呼ばれる畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) を提案しており，予測対象が存在しているパッチ画像と周辺の微小領域のパッチ画像の類似度を比較することで局所的な領域の報酬を推定している．

上記の 2 つのアプローチではシーンの微小領域に対する特徴を抽出していたが，シーン全体を一つの特徴ベクトルとして表現するアプローチも存在する．このアプローチでは，類似したシーンでは類似した経路が予測されると仮定し，シーン全体を表現する特徴ベクトルを抽出する．特徴ベクトルを用いて，類似したシーンを学習データから検索し，検索された学習データの軌跡を予測に用いている．そのため，シーンを効果的に表現する特徴ベクトルを抽出する必要である．2016 年以降では深層学習の発達の影響を受け，CNN による特徴抽出が用いられている．Park ら [42] は一人称視点映像における歩行者の経路予測を行うために，AlexNet [28] を用いて抽出した特徴ベクトルと学習データの特徴ベクトルとの類似度を比較し，学習



(a) 歩行者及び頭部の検出

(b) 頭部の向きの推定 (8方向)

図 3　歩行者の頭部の向きの推定例．文献 [5] より引用．

データの中から類似した移動軌跡のデータを予測シーンに転移させることで予測を実現している．また，Su

ら [43] は AlexNet ベースの Siamese Network [29] を用いて類似した学習データを検索している．

2. 2　予測対象に対する特徴抽出

移動経路を決定する要因は周囲の環境などの外的要因だけではない．予測対象自身が保有している内的要因も大きく影響しており，予測対象に対する特徴抽出も様々行われている．具体的には，歩行者の年齢や性別，潜在的要求などの属性情報が移動経路の決定に大きく影響していると考えられる．そこで，以下では予測対象に対する特徴抽出について述べる．

最も頻繁に用いられるのは，予測対象の向きである [5], [27], [35]．対象がどの方向を向いているのかを推定することで，どの方向に移動しようとしているかを考慮することができる．すなわち，経路予測手法を適用する際の移動方向に関する制約を導入することができ，間違った予測を抑制することが可能である．Kooij ら [5] は，車載カメラ映像中の車道を横切る歩行者の移動経路および車道の手前で停止するかどうかを予測することを目的とし，HOG+SVM による歩行者検出法 [30] による歩行者位置の推定に加えて，歩行者の頭の向きを推定している [33]．この時，頭部がカメラ方向を向いている場合，歩行者は接近する車両の存在に気づいていると仮定し，移動速度を緩めたり，車道の手前で停止する動きの予測を実現している．また，Yagi ら [50] は人物の姿勢推定法 [34] を用いており，対象の向きや動きを考慮した予測を行なっている．

年齢や性別などの身体的属性も経路を予測する際の重要な情報である．不特定多数が行き交うような環境において歩行者の経路を予測する場合，他者との衝突を避けつつ移動を行うため，回避行動を考慮した経路予測を行う必要がある．回避行動を起こす位置や時刻は歩行者の移動速度によって変化する．その際，歩行者の移動速度は同じではなく，高齢者や若者などの年齢，性別によって異なることが考えられる．Ma ら [35] は AlexNet [28] を用いて歩行者の向きに加えて，性別や年齢の身体的属性をマルチタスク学習の枠組みで推定しており，推定した歩行者属性を用いて歩行者毎の移動速度を決定し，経路予測を行なっている．

Walker ら [41] は教師なし学習での経路予測を目的とし，上記のような明示的な属性情報ではなく予測対象の特徴ベクトルを直接用いており，予測対象が含まれているパッチ画像から mid-level の特徴ベクトルを抽出している [36]．

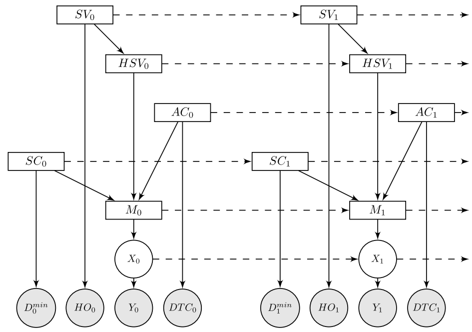


図 4　SLDS を導入した DBN のグラフィカルモデル．文献 [5] より引用．

3. 予 測 手 法

動画像から抽出した特徴を用いて経路予測が行われる．表 2 に経路予測手法の一覧を示す．様々な手法が提案されているが，基礎となるアプローチに基づき，いくつかのカテゴリに分類することができる．そこで本章では，各カテゴリに属する経路予測手法および，それぞれの手法の特徴について述べる．

3. 1　ベイズモデルに基づく手法

一つ目の経路予測のアプローチはカルマンフィルタ [57] や粒子フィルタ [58], [59] などのベイズモデルに基づく推論によって経路を予測する手法である．これらの手法では，内部状態および観測と呼ばれる変数を用いて，内部状態にノイズが付与されたものが観測として現れるという確率的なモデルを定義する．この確率モデルを用いて，時刻間での内部状態を更新する predict および，観測から内部状態を更新する update を繰り返すことで各時刻の内部状態を逐次的に推定する．すなわち，内部状態が各時刻での歩行者の存在する座標，観測が人検出法などで動画像から推定した歩行者の座標となる．この処理を過去から現在まで時刻で行う場合，人物追跡の問題と捉えることができる．一方，経路予測では未来の時刻の状態は観測することができないため，predict のみを繰り返すことで歩行者の座標を推定し，推定した座標列を経路予測の結果としている．

Schneider ら [3] は車載カメラ映像に映った車両前方の歩行者の移動経路を予測している．彼らは拡張カルマンフィルタを用いて歩行者の状態を更新することで移動経路を予測している．この手法では，歩行者の速度や加速度などの予測に用いる情報をいくつか変更

することで歩行者の移動経路を予測するためにはどのような情報が有用かを検討している．

上記のような逐次ベイズフィルタリングの他に，動的ベイジアンネットワーク (Dynamic Bayesian Netowrk; DBN) [60] を用いた予測手法も存在する [5], [38]．Kooij ら [5] もまた車載カメラ映像からの歩行者の経路予測を目的として経路予測を提案している．彼らはより限定的なシーンを想定し，車両の前方に存在する歩行者が車道を横切る際の移動，すなわち車道を横断するか停止するかという経路予測を行なっており，図 4 に示すような，DBN に Swiching Linear Dynamical System (SLDS) を導入したモデルを定義している．このモデルでは，歩行者の座標に加えて，歩行者の頭の向きや車両までの距離，歩行者と縁石（歩道と車道の境界）の距離を観測として用いている．これら複数の観測を用いて内部状態を逐次的に更新することで，座標のみを観測として用いる場合よりも高精度な予測を実現している．

3. 2　エネルギー最小化に基づく手法

前節のベイズモデルによる予測手法は，各時刻での歩行者の座標を逐次的に推定するオンライン処理であった．一方で，現在地から目的地までの経路を一度に予測するバッチ処理の予測手法として，エネルギー最小化に基づく予測手法が存在する．エネルギー最小化に基づく経路予測では，主に予測を行うシーン，すなわち動画像フレームを 2 次元の格子状のグラフとして定義し，動画像から推定した移動コストをグラフのエッジに付与する．このグラフを用いて，コストが最小になるようなエッジの組み合わせを推定することで経路予測が実現される．すなわち，この問題は最短経路問題であり，ダイクストラ法を用いて経路を推定している．そのため，コストの作成方法が予測精度に大きな影響を与える．

Huang ら [27] は一枚の静止画像を入力として，そのシーン中の予測対象の移動経路を予測する手法を提案している．この手法では，まず予測対象を含むパッチ画像を抽出する．抽出したパッチ画像から対象の向きを推定し，さらにパッチ画像とシーン中の他の領域のパッチ画像とのテクスチャを比較することで，対象がその地点を移動するコストを推定する．このコストをグラフのエッジの重みとして使用し，さらに予測対象の方向から移動方向に対する制約としてコストを追加している．Walker ら [41] も同様に微小領域のテクスチャを比較することでコストを推定しているが，予測対象が過去に移動した領域のパッチ画像を用いて各微小領域のテクスチャの類似度を求めることで，学習を行うことなく経路予測を実現している．

テクスチャなどのシーンの見えに関する情報からコストを推定するだけではなく，シーン中に存在する物体からコストを作成する方法も存在する．Xie ら [44] は，歩行者は潜在的な要求 (hunger) によって目的地 (food trunk) が決まっていると仮定し，シーン中のオブジェクトに引き寄せられるようなコストマップを作成し，目的地までの移動経路を予測している．

3. 3　深層学習に基づく手法

近年の深層学習の技術の発展により，CNN や再帰型ニューラルネットワーク (Recurrent Neural Network; RNN)，長期短期記憶 (Long Short-Term Memory; LSTM) を用いた予測手法も存在する．深層学習による経路予測手法では，過去の数フレームの移動軌跡を入力とし，その後の数フレームの移動軌跡を出力するネットワーク構造を定義している．そのため，2 章で述べたような明示的な特徴抽出は行われず，特徴抽出と予測が一貫して行われる構造となっている．

移動経路は 2 次元座標の系列データであるため，主に再帰的な構造を持つ RNN や LSTM を用いた経路予測手法が提案されている．Alahi ら [46] は複数人の歩行者の移動経路を予測しており，歩行者同士の衝突を避ける動きを予測するために Social-pooling (S-Pooing) Layer を提案している．S-Pooling では，予測対象周辺に存在する歩行者の LSTM の中間層出力とその空間的関係を保存している．S-Pooling の出力を次の時刻の LSTM へ入力することにより，衝突を避けるような動きを考慮した経路予測を実現している．Bartoli ら [47] は S-Pooling による歩行者間のインタラクションに加えて，シーンに存在する静的な物体の位置情報を導入した経路予測手法を提案している．Fernando ら [48] も同様に複数人の歩行者が存在するような環境での経路予測を目的としており，Attention モデルを用いて自身の周囲の歩行者の情報を抽出している．Lee ら [40] は RNN Encoder-Decoder [61] および Conditional Variational Autoencoder [62] を用いた予測手法を提案している．この手法では，過去の数フレームの移動軌跡から予測軌跡を生成する．その後，周囲の環境や移動物体の情報を用いて生成した軌跡を改善することで，最終的な予測結果を出力している．

LSTM を用いて経路予測を行う場合，LSTM の長期記憶の性能の限界により長期での経路予測すなわち，

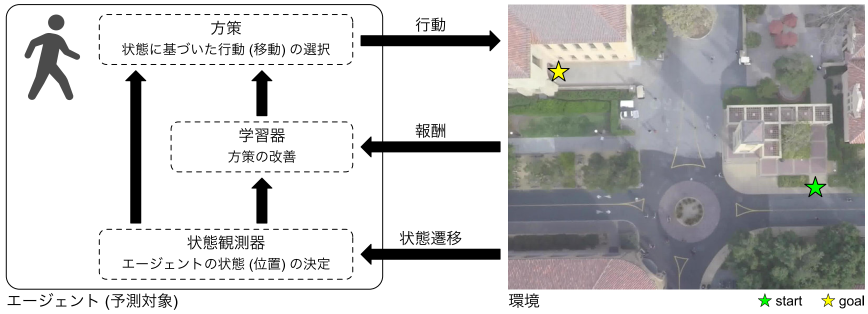


図 5　強化学習の処理の流れ．文献 [56] より改変．

遠い未来の時刻の予測結果は不十分という問題点がある．そのため，長期の経路予測を行う際にはさらなる長期記憶が必要になるという仮定のもと，Fernando ら [49] はメモリセルに格納された過去の各時刻の情報から階層的に有用な情報を選択する Tree Memory Network を提案しており，他の LSTM を用いた予測手法に比べて長期での経路予測を実現している．

経路予測では LSTM が用いられることが多いが，CNN を用いて予測結果を直接推定する手法も存在する．Yi ら [45] は過去の移動経路から予測経路を直接出力するような Behavior-CNN を提案している．この手法では，歩行者の過去の数フレームの移動座標を各チャネルに保存したスパースな 3 次元データを作成し，Behavior-CNN の入力とする．その後，畳み込み及び Max Pooling を適用することで入力データを符号化し，deconvolution を適用し復号化することで，予測結果を出力している．また，入り口や障害物などの存在により，特定のシーン中の位置によって異なる歩行者の振る舞いを考慮するために，location bias map と呼ばれるバイアスをエンコーダによって符号化されたデータの各チャネルに加えることで，予測精度を向上させている．Yagi ら [50] は一人称視点映像中の歩行者の移動経路の予測手法を提案している．この手法では，Faster R-CNN [31], [32] によって推定した対象者の位置と画像中でのスケールに加えて，撮影者自身の自己運動量および OpenPose [34] で推定した対象者の姿勢を入力として用いている．観測開始までの数フレームに対応するこれらの情報を畳み込みによって符号化し，deconvolution を用いて復号化することで，未来の時刻での対象者の位置を出力している．一人称視点映像中の歩行者の経路を予測する場合，撮影者が移動することで画像中の対象者の位置が変化するため，カメラ自身の移動量を考慮する必要がある．この手法では撮影者の移動量を入力の一つとして用いることでカメラの移動を考慮した予測を実現している．

3.4　逆強化学習に基づく手法

前述の 3 つのアプローチは主に，教師あり学習または教師なし学習として分類することができる．これに対して，強化学習の枠組みを用いた経路予測手法が存在する．強化学習とは，ある環境においてエージェントが現在の状態を考慮し，取るべき行動を選択するような問題および行動を決定するための方策を学習することを指す．通常，強化学習は有限の状態数を持つマルコフ決定過程として定義され，報酬が最大化するような行動を取ることができるように最適な方策を試行錯誤しながら学習する．図 5 に示すように，強化学習の枠組みにおいてエージェントは予測対象，環境は動画像から観測されるシーン，状態は歩行者の位置，行動は歩行者の移動と捉えることができる．

強化学習では，エージェントが取った行動の良さの指標として，選択した行動によって遷移した状態に対しての報酬を定義する必要がある．しかし，今回の経路予測のような問題に強化学習を適用する場合，明示的に報酬を決定することが難しい．このように，強化学習における報酬を決定する問題を報酬設計問題と呼び，この問題を解決するためのアプローチの一つに逆強化学習が存在する [63]．逆強化学習では，学習で最適な行動系列のデータから同じような行動を取ることができるような報酬を推定し，テスト時にその報酬を用いてエージェントの行動選択を行う．逆強化学習を経路予測に用いる場合，実際の歩行者の歩行軌跡を最適な行動系列として学習に使用し最適な報酬を推定する．

逆強化学習はロボティクスの分野においてロボットの最適な動きを教師データから学習し制御する技術 [6]

として用いられているが，Kitani ら [37] は動画像からの経路予測に逆強化学習を初めて導入した．彼らは対象の位置を求めるのではなく，ある時刻またはある地点で対象が取るであろう行動を推定しており，推定した行動によって遷移した対象の位置を連続して求めることで移動経路を予測している．座標を直接推定する path prediction に対して，彼らはこの問題設定を activity forecasting と呼び，コンピュータビジョンにおける新たな問題設定として提案している．Activity forecasting は予測対象や環境に応じて取りうる行動を定義し推定できることから，様々な問題に応じた予測を行う可能性を秘めているが，扱うモデルが複雑になるため path prediction に比べて挑戦的な問題設定となっている．

Kitani ら [37] はシーンの物理的環境属性が歩行者の経路に大きく影響していると仮定し，セマンティックセグメンテーションによって推定したシーンの環境属性を特徴マップとして用いている．その特徴ベクトルと重みベクトルとの一次結合によって各領域での報酬を定義しており，最適な重みベクトルを学習データ，すなわち実際に対象が移動した軌跡データから推定する．テスト時には，予測対象の歩行者の現在地および目的地を所与として，目的地に到達するまでの行動系列を生成することにより移動経路を予測する．Lee ら [51] は，同様のアプローチを用いて，アメリカンフットボールのプレイ映像におけるプレイヤーの移動経路を予測している．また，Ma ら [35] は複数の歩行者が衝突を避けつつ目的地へ移動する際の経路を予測することを目的として，逆強化学習に仮想プレイ (Fictitious Play) と呼ばれるゲーム理論を導入することで，複数人の経路を同時に予測している．

前述の逆強化学習に基づく手法では歩行者の目的地を所与としていたが，Rehder ら [39] は Destination Network と呼ばれるネットワークを提案しており，過去の数フレームの移動軌跡から対象者の目的地を推定し，推定した目的地及び FCN [24], [25] で推定した周囲の環境属性を用いて歩行者の経路予測を行なっている．

このほかの逆強化学習を用いた研究として，Bokhari ら [52] は一人称視点映像において対象者が把持している物体やその状態を考慮することで将来の目的地を考慮した経路予測を行なっている．この研究では，キッチンスペース内での移動という非常に狭く限定的なシーン中での移動を予測していたが，Rhinehart ら [53] はキッチンやバスルーム，ベッドルームなどのより広範囲での経路予測を実現している．

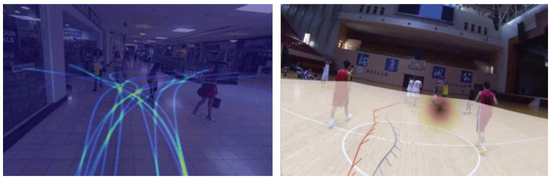


図 6　一人称視点映像からの経路予測．左: 文献 [42]，右: 文献 [43] より引用．

3.5　その他のアプローチ

多くの経路予測手法が上記の 4 種類のアプローチに分類されるが，これらに属さないアプローチもいくつか存在する．

Social Force Model [64] は歩行者間や歩行者と何かしらの物体との間に “social force” と呼ばれるエネルギーが存在していると仮定することで物体とのインタラクションを考慮した歩行者の動きのモデルである．この Social Force Model を元にした予測手法として，Yamaguchi ら [55] は歩行者の状態に歩行者の好みの速度や，目的地，他の歩行者と一緒に移動しているかという状態を追加したモデルを提案している．この研究では主に人物追跡の精度向上が目的とされているが，提案モデルの妥当性を評価するための実験として経路予測を行なっている．Robicquet ら [56] は，Social Force Model に基づき，他の移動物体との衝突を避けるような動きを考慮した経路予測を提案している．この手法では，予測対象が回避行動を起こし始める距離や回避の動きが異なることを仮定しており，対象毎に異なる回避行動を予測している．具体的には学習に用いる移動対象毎に “social sensitivity features” と呼ばれる特徴量を求め，この特徴量に対して K-means clustering を適用し回避行動の種類をいくつかのクラスタに分割し，各クラスタでの平均的な回避行動を求める．予測時には，対象の social sensitivity features からどのクラスタに割り当てられるかを決定し，そのクラスタの行動軌跡をシーンにマッピングすることで経路予測を実現している．

Keller ら [4] は車載カメラ映像から抽出したオプティカルフローを用いた経路予測手法を提案している．この手法では，予測開始までの過去の数フレームからオプティカルフローを抽出し，歩行者の動きの特徴量として方向ヒストグラムを作成する．この方向ヒストグ

表 3　データセットの比較.

	年代	入手	歩行者数	映像視点	シーン数	歩行者以外の対象	追加情報
UCY [65]	2007	✓(注2)	786	鳥瞰	3	–	–
ETH [66]	2009	✓(注3)	750	鳥瞰	2	–	–
Edinburagh Informatics Forum [67]	2009	✓(注4)	95,998	鳥瞰	1	–	–
Stanford Drone [56]	2016	✓(注5)	11,216	鳥瞰	8	bikers, car, bus, cart, skateborder	–
VIRAT [20]	2011	✓(注6)	4,021	監視カメラ	11	car, bike	物体の座標 行動カテゴリ
Town Centre [68]	2011	✓(注7)	230	監視カメラ	1	–	頭部の座標
Grand Central Station [69]	2015	✓(注8)	12,600	監視カメラ	1	–	–
Daimler [3]	2013	✓(注9)	68	車載	–	–	ステレオカメラ
KITTI [70]	2012	✓(注10)	6,336	車載	–	car	ステレオカメラ LIDAR 地図情報
EgoMotion [42]	2016		–	一人称	26	–	ステレオカメラ
First-person Continuous Activity [53]	2017		–	一人称	17	–	物体情報
First-person Locomotion [50]	2018		5,000	一人称	–	–	–

ラムの系列データを用いて学習データから類似した経路データを検索し，マッピングすることで経路予測を実現している．

Rehder ら [54] はマルコフ過程の枠組みに基づき，歩行者の状態（位置）及び移動速度をそれぞれ正規分布とフォン・ミーゼス分布を用いて定義し，各時刻でこれらの分布の積を取ることで歩行者の状態を逐次的に推定し，経路を予測している．その際，歩行者の目的地をシーンの環境属性から推定した結果を用いることで，移動方向に対して制約が与えられ，予測精度を向上させている．

また，Park ら [42] は図 6 に示すように，一人称視点映像から撮影者自身の将来の移動経路を予測することを目的として，検索ベースのアプローチを取っている．この手法では，AlexNet で一人称視点映像からシーンの特徴量を抽出し，学習データの特徴量と比較することで類似した学習シーンを検索，抽出する．この手法では撮影者の頭部に装着した 2 台のウェアラブルカメラをステレオカメラとして用いることで深度画像を取得している．さらに，撮影者を中心とした対数極座標表現を用いて地上座標系へと変換し予測を行なっている．これにより，撮影者の頭部の動きや撮影者の違いによるカメラ位置のオフセットを抑制し，対象者に近い空間に対応する画像情報を考慮した予測を実現している．また，映像中の壁や障害物などで隠された背後に存在する領域を推定することで，オクルージョンが存在する領域に対しても経路の予測を可能としている．Su ら [43] は Park らの手法を同一シーン中の複数人の経路予測へと発展させ，バスケットボールの試合中のプレイヤーの動きを予測している．この手法では，先ほどのように AlexNet を用いて類似シーンの移動軌跡を複数取得すると同時に，複数の一人称視点映像から “joint attention” と呼ばれるプレイヤーに共通する注目領域を推定している．推定した joint attention やプレイヤーの位置，取得した複数の軌跡をもとにコスト関数を定義し，コストを最小にする各プレイヤーの予測軌跡の組み合わせを求めることで，複数人の経路予測を実現している．

4. 経路予測に用いられるデータセット

経路予測手法を定量的に評価するために，表 3 及び図 7 に示すような様々なデータセットが用いられている．これは，予測を行う際の映像視点やシーンの数，経路予測の応用先などの様々な条件によって使用できるデータセットが異なるためであり，全ての手法において統一的なデータセットを使用し評価することが難しい．そこで本章では，経路予測に用いられるデータセットとその特性について述べる．

4.1 鳥瞰視点映像のデータセット

動画像を用いた経路予測において，最も用いられる

(注2)：https://graphics.cs.ucy.ac.cy/research/downloads/crowd-data
(注3)：http://www.vision.ee.ethz.ch/en/datasets/
(注4)：http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING/
(注5)：http://cvgl.stanford.edu/projects/uav_data/
(注6)：http://www.viratdata.org/
(注7)：http://www.robots.ox.ac.uk/~lav/Papers/benfold_reid_cvpr2011/benfold_reid_cvpr2011.html
(注8)：http://www.ee.cuhk.edu.hk/~xgwang/grandcentral.html
(注9)：http://www.gavrila.net/Datasets/Daimler_Pedestrian_Benchmark_D/daimler_pedestrian_benchmark_d.html
(注10)：http://www.cvlibs.net/datasets/kitti/

のは駅構内や市街地をドローンや監視カメラを用いて上部から撮影したデータセットである．これらのデータセットは主に人物追跡を目的として作成されているが，歩行者の座標列，すなわち移動軌跡が教師ラベルとして与えられているため，経路予測の評価実験にも使用されている．本論文ではこのようなデータセットのうち，シーンをほぼ真上から撮影したデータセットを鳥瞰視点映像のデータセットと定義し，以下では鳥瞰視点映像のデータセットについて述べる．

UCY Dataset [65] 及び ETH Dataset [66] は市街地の歩行者を撮影されたシーンからなるデータセットである．このデータセットでは，歩行者のみが存在するシーンを撮影した動画像から構成されている．そのため，経路予測手法の評価に用いられるデータセットの中では比較的シンプルな環境のデータセットとなっている．Edinburgh Informatics Forum Pedestrian Database [67] はエディンバラ大学構内に設置した定点カメラによって歩行者を撮影したデータセットであり，UCY Dataset 及び ETH Dataset と類似した環境で撮影されたデータセットである．このデータセットには，90,000 本以上もの軌跡データが記録されており，非常に大規模なデータセットとなっている．

上記のデータセットは主に経路予測や群衆行動解析を目的として作成されたものである．一方，Stanford Drone Dataset [56] は経路予測を目的として作成されたデータセットである．このデータセットではスタンフォード大学構内の 8 つの地点をドローンに装着したカメラを用いて撮影している．さらに，シーン中の移動物体は歩行者のみではなく，biker や skateborder, car などの複数種類の予測対象の情報が公開されている．

4. 2　監視カメラ映像のデータセット

鳥瞰視点映像と同様にシーンを上部から撮影したデータセットとして，監視カメラ映像のデータセットが存在する．これらの分類の違いは映像視点である．鳥瞰視点映像ではシーンをほぼ真上から撮影しているが，監視カメラ映像のデータセットは図 7(e, f, g) に示すようにシーンを斜め上から撮影している．シーンを斜め上から撮影することにより，障害物などのシーンに存在する物体や歩行者の身体的特徴等を動画像から推定することが可能である．また，鳥瞰視点映像では対象の座標として画像座標を用いていたが，監視カメラ映像では画像座標系および射影変換によって得られた地上座標系を用いる場合が存在する．

VIRAT Video Dataset [20] は監視カメラで撮影された駐車場の映像からなるデータセットであり，歩行者の位置に加えて，自動車やシーン中に存在する物体の位置情報が用意されている．さらに，人物の自動車への乗降りやトランクの開閉などの行動に対するラベルも付与されている．また，動画像の撮影シーンは 11 シーンであり，表 3 に記載した他の鳥瞰視点映像のデータセットよりも多くのシーンを含んだデータセットとなっている．Town Centre Dataset [68] は移動物体は歩行者のみであるが，人物の位置を示すバウンディングボックスに加えて，歩行者の頭の位置に対するラベルも付与されている．

Grand Central Station Dataset [69] は駅構内に設置された監視カメラで撮影されたデータセットである．データセットは図 7(g) に含まれる 1 シーンのみから構成されているが，多数の歩行者を解析する群集行動解析を主な目的として作成されており，広範囲を撮影している．そのため，一度に大量の歩行者が存在している非常に複雑なデータセットとなっている．

4. 3　車載カメラ映像のデータセット

自動車の自動運転支援を目的として，車載カメラ映像のデータセットが経路予測に用いられている．車載カメラ映像では，主に自動車の前方を撮影した映像中の歩行者の移動経路を予測することを目的としている．車載カメラ映像のデータセットにはステレオカメラや LIDAR などの追加情報が含まれており，これらの情報を活用してシーンを地上座標系に変換して予測が行われる．

Daimler Pedestrian Path Prediction Benchmark Dataset [3] は車載カメラ映像を用いて作成されたデータセットである．このデータセットでは，歩行者が車道を横断する際に，立ち止まることなく横断する場合や自動車との衝突を避けるために横断しない場合などの 4 つのクラスに分類されている．また，動画像はステレオカメラで撮影されているため，距離情報を用いることが可能である．このデータセットは車載カメラ映像からの経路予測を目的として作成されたデータセットである．経路予測の初期に作成されたデータセットのため歩行者数は他のデータセットに比べると少ないが，車両前方を横切る歩行者の映像が含まれているデータセットである．

KITTI Vision Benchmark Suite [70] は高度道路交通システム (Intelligent Transport System; ITS) 向けに作成されたデータセットであり，歩行者や車両の

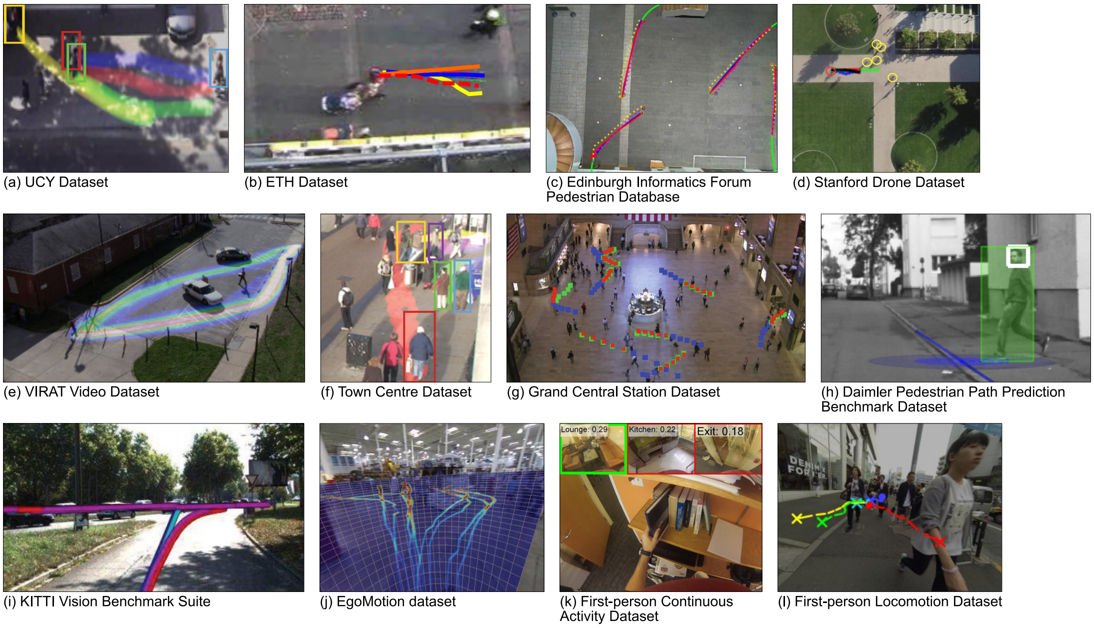


図 7　経路予測のデータセット及び経路予測結果の例．文献 [5], [35], [37], [40], [42], [45], [46], [49], [50], [53], [56] より引用及び改変．

検出，白線検出などの様々な問題の評価に用いられる．KITTI Vision Benchmark Suite では RGB 画像に加えてステレオ画像や LIDAR の 3 次元点群データ，GPS を用いた世界座標系での車両の位置や地図情報が公開されている．そのため，周囲の環境を理解するための多数のデータが利用可能なデータセットである．

4.4　一人称視点映像のデータセット

一人称視点映像を用いた経路予測も行われている．Park ら [42] は一人称視点映像の撮影者の移動経路を予測するために，屋内及び屋外を移動する際の一人称視点映像のデータセットを作成している．このデータセットでは，市街地や店内などの 26 のシーンで合計 9 時間以上の一人称視点映像を撮影している．これに対して，Yagi ら [50] は一人称視点映像中の歩行者の経路を予測しており，屋外で長時間撮影した一人称視点映像のデータセットを作成している．このデータセットでは 5,000 人以上の予測対象となる歩行者が観測されており，他の一人称視点映像のデータセットと比較して大規模なデータセットとなっている．Rhinehart ら [53] はオフィスなどの屋内における人物の経路予測を行うための一人称視点映像のデータセットを作成している．この際，対象者が把持しているマグカップやタオルなどの物体にキッチンやバスルームなどの目的地が依存しているとして，日常生活の行動に準じた一人称視点映像データセットを作成している．

しかし，これらの定量的評価には独自のデータセットが使用されており，公開されていない．そのため，一人称視点映像での行動予測手法の評価を行うためには，自身でデータセットの作成を行う必要がある．

5.　定量的な評価指標

経路予測手法を評価するためには，データセットに加えて統一的な評価指標を用いる必要がある．しかし，表 2 に示したように，経路予測手法の出力は予測手法のアプローチに依存して，座標または確率分布の 2 種類が存在する．そこで，本章では経路予測結果を定量的に評価するための評価指標について紹介する．

5.1　座標に対する評価指標

座標（座標の系列データ）として予測結果を出力する場合に最も用いられる評価指標は displacement error である．Displacement error は，現在までに取得した観測から次時刻の予測対象の座標を逐次的に出力する予測手法で主に用いられる．Displacement error には average displacement error (ADE) と final displacement error (FDE) の 2 種類が存在する．こ

こで，y_t を時刻 t における予測結果，$\hat{y}_t$ を時刻 t における真値としたとき，ADE および FDE は次のように定義される．

$$\mathrm{ADE} = \frac{1}{T}\sum_{t=1}^{T}\|\hat{y}_t - y_t\|_2 \tag{1}$$

$$\mathrm{FDE} = \|\hat{y}_T - y_T\|_2 \tag{2}$$

ADE では，予測を開始する時刻から終了する時刻までの各時刻の真値と予測値のユークリッド距離の平均を求めている．一方，FDE では予測終了時刻での真値と予測値のユークリッド距離を誤差として算出している．ADE では予測期間全体での精度を評価しているのに対して，FDE では遠い未来の時刻での予測精度のみを評価している．評価対象となる時刻が異なるため，これらは同時に評価指標として使用されることが多い．

座標に対する別の評価指標として，modified Hausdorff distance (MHD) [71] が存在する．ここで，$A = \{a_1, \ldots, a_{N_a}\}$ および $B = \{b_1, \ldots, a_{N_b}\}$ をそれぞれ座標の集合とする．座標の集合 A から B までの有向距離は次のように定義される．

$$d(A, B) = \frac{1}{N_a}\sum_{a \in A} d(a, B) \tag{3}$$

ここで，$d(a, B) = \min_{b \in B} \|a - b\|_2$ とする．MHD は A および B の間の有向距離 $d(A, B)$ と $d(B, A)$ を用いて次式により定義される．

$$\mathrm{MHD}(A, B) = \max\left(d(A, B), d(B, A)\right) \tag{4}$$

MHD を経路予測の評価指標として用いる場合，A および B にそれぞれ予測結果の座標列と真値の座標列を用いて計算を行う．MHD は三次元点群で表現された物体間の類似度として提案された指標であるが，経路の座標列を 2 次元の曲線と見なすことで，経路予測の評価指標に用いられている．MHD は上記の displacement error に比べて，長期間の経路予測結果や目的地があらかじめ設定されている経路予測結果に対して使用されることが多い．

5.2 確率分布に対する評価指標

予測結果が確率分布として出力される場合には，negative log-loss (NLL) が評価指標として広く用いられている．真値の座標列を $\hat{\boldsymbol{y}} = \{\hat{y}_1, \ldots, \hat{y}_T\}$，予測手法により生成した分布を $p(y)$ とするとき，NLL は次のように定義される．

$$\mathrm{NLL} = \sum_{t=1}^{T} -\log p(\hat{y}_t) \tag{5}$$

これは分布 $p(y)$ のもとで，真値 $\hat{\boldsymbol{y}}$ を取りうる負の対数尤度を表している．すなわち，NLL の値が小さいほど真値に近い経路が予測されていると見なすことができる．MHD は主に逆強化学習に基づく予測手法で主に用いられている．

6. おわりに

本論文では，動画像を入力とする経路予測手法のサーベイ及び，経路予測手法に用いられるデータセットについて報告した．まず，経路予測に用いられる動画像からの特徴抽出法について分類した．特徴抽出については，シーンの環境に関する特徴抽出及び歩行者などの予測対象に関する特徴抽出について述べた．予測手法については，ベースとなるアプローチに基づき，予測手法を 4 つのグループに大別した．1 つ目のベイズモデルに基づく手法では，対象物が移動する経路についての確率モデルを定義し，各時刻での内部状態を逐次的に求めることで経路予測を実現している．2 つ目のエネルギー最小化に基づく手法では，シーン中の微小領域毎に対象が移動する可能性をコストとして計算し，2 次元の格子状のグラフを作成したのちに，ダイクストラ法を用いて最短経路を推定することで予測経路を生成している．3 つ目の深層学習に基づく手法では，予測開始前の数秒間の対象物の移動軌跡を観測としてネットワークに入力し，その後の数秒間の移動経路を出力することで経路を予測している．4 つ目の逆強化学習に基づく手法では，教師データから推定した行動選択の基準となる報酬及び方策を用いて，予測対象の行動を連続して選択することで経路を予測している．また，これらの 4 種類の予測手法は独立して用いられる場合のみではなく，複合的な手法も提案されている [40]．

また，経路予測手法の評価に用いられるデータセットおよび評価指標についても調査した．調査したデータセットは主に歩行者検出やトラッキングを目的として構築されている．経路予測を目的としたデータセットとして，Stanford Drone Dataset や Daimler Pedestrian Path Prediction Benchmark Dataset が存在する．定量的な評価指標として，座標の予測結果に対して displacement error および MHD が広く用

いられており，確率分布の予測結果に対して NLL が使用されている．

謝辞　本研究は科研費 (JP16H06540) の補助を受けたものである．

文　献

[1] R. Mehran, A. Oyama, and M. Shah, “Abnormal crowd behavior detection using social force model,” Computer Vision and Pattern Recognition, pp.935–942, 2009.

[2] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” Computer Vision and Pattern Recognition, pp.1975–1981, 2010.

[3] N. Schneider and D.M. Gavrila, “Pedestrian path prediction with recursive bayesian filters: A comparative study,” German Conference on Pattern Recognition, pp.174–183, 2013.

[4] C.G. Keller and D.M. Gavrila, “Will the pedestrian cross? a study on pedestrian path prediction,” IEEE Transactions on Intelligent Transportation Systems, vol.15, no.2, pp.494–506, 2014.

[5] J.F.P. Kooij, N. Schneider, F. Flohr, and D.M. Gavrila, “Context-based pedestrian path prediction,” European Conference on Computer Vision, pp.618–633, 2014.

[6] B.D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J.A. Bagnell, M. Hebert, A.K. Dey, and S. Srinivasa, “Planning-based prediction for pedestrians,” International Conference on Intelligent Robots and Systems, pp.3931–3936, 2009.

[7] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto, “Intent-aware long-term prediction of pedestrian motion,” International Conference on Robotics and Automation, pp.2543–2549, 2016.

[8] A. Vemula, K. Muelling, and J. Oh, “Modeling cooperative navigation in dense human crowds,” International Conference on Robotics and Automation, pp.1685–1692, 2017.

[9] 山内悠嗣，山下隆義，藤吉弘亘，“画像からの統計的学習手法に基づく人検出，” 電子情報通信学会論文誌 D，vol.96，no.9，pp.2017–2040，2013.

[10] 福井宏，山下隆義，山内悠嗣，藤吉弘亘，“Deep learning を用いた歩行者検出の研究動向，” 電子情報通信学会技術研究報告 (PRMU) 技術報告，vol.116，no.366，pp.37–46，2016.

[11] 川西康友，新村文郷，出口大輔，村瀬洋，“画像からの歩行者属性認識，” 電子情報通信学会技術研究報告 (PRMU) 技術報告，vol.115，no.388，pp.117–127，2015.

[12] H. Zhu, F. Meng, J. Cai, and S. Lu, “Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation,” Journal of Visual Communication and Image Representation, vol.34, pp.12–27, 2016.

[13] Y.F. Chen, S.-Y. Liu, M. Liu, J. Miller, and J.P. How, “Motion planning with diffusion maps,” International Conference on Intelligent Robots and Systems, pp.1423–1430, 2016.

[14] Y.F. Chen, M. Liu, M. Everett, and J.P. How, “Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning,” International Conference on Robotics and Automation, pp.285–292, 2017.

[15] Y.F. Chen, M. Everett, M. Liu, and J.P. How, “Socially aware motion planning with deep reinforcement learning,” International Conference on Intelligent Robots and Systems, pp.1343–1350, 2017.

[16] L. Sun, Z. Yan, S.M. Mellado, M. Hanheide, and T. Duckett, “3dof pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data,” International Conference on Robotics and Automation, pp.5942–5948, 2018.

[17] M.S. Ryoo, “Human activity prediction: Early recognition of ongoing activities from streaming videos,” International Conference on Computer Vision, pp.1036–1043, 2011.

[18] M. Hoai and F.D. laTorre, “Max-margin early event detectors,” Computer Vision and Pattern Recognition, pp.2863–2870, 2012.

[19] M.S. Ryoo and L. Matthies, “First-person activity recognition: Feature, temporal structure, and prediction,” International Journal of Computer Vision, vol.119, no.3, pp.307–328, 2016.

[20] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.C. Chen, J.T. Lee, S. Mukherjee, J.K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, “A large-scale benchmark dataset for event recognition in surveillance video,” Computer Vision and Pattern Recognition, pp.3153–3160, 2011.

[21] T. Hirakawa, T. Yamashita, T. Tamaki, and H. Fujiyoshi, “Survey on vision-based path prediction,” International Conference on Human-Computer Interaction, 2018.

[22] D. Munoz, J.A. Bagnell, and M. Hebert, “Stacked hierarchical labeling,” European Conference on Computer Vision, pp.57–70, 2010.

[23] J. Yang, B. Price, S. Cohen, and M.H. Yang, “Context driven scene parsing with attention to rare classes,” Computer Vision and Pattern Recognition, pp.3294–3301, 2014.

[24] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” Computer Vision and Pattern Recognition, pp.3431–3440, 2015.

[25] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.39, no.4, pp.640–651, 2017.

[26] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," European Conference on Computer Vision Workshops, pp.1–22, 2004.

[27] S. Huang, X. Li, Z. Zhang, Z. He, F. Wu, W. Liu, J. Tang, and Y. Zhuang, "Deep learning driven visual path prediction from a single image," IEEE Transactions on Image Processing, vol.25, no.12, pp.5892–5904, 2016.

[28] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, pp.1097–1105, 2012.

[29] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," Advances in Neural Information Processing Systems, pp.737–744, 1994.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Computer Vision and Pattern Recognition, pp.886–893, 2005.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in Neural Information Processing Systems, pp.91–99, 2015.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.39, no.6, pp.1137–1149, 2017.

[33] M. Enzweiler and D.M. Gavrila, "Integrated pedestrian classification and orientation estimation," Computer Vision and Pattern Recognition, pp.982–989, 2010.

[34] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," Computer Vision and Pattern Recognition, pp.7291–7299, 2017.

[35] W. Ma, D. Huang, N. Lee, and K.M. Kitani, "Forecasting interactive dynamics of pedestrians with fictitious play," Computer Vision and Pattern Recognition, pp.774–782, 2016.

[36] S. Singh, A. Gupta, and A.A. Efros, "Unsupervised discovery of mid-level discriminative patches," European Conference on Computer Vision, pp.73–86, 2012.

[37] K.M. Kitani, B.D. Ziebart, J.A. Bagnell, and M. Hebert, "Activity forecasting," European Conference on Computer Vision, pp.201–214, 2012.

[38] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese, "Knowledge transfer for scene-specific motion prediction," European Conference on Computer Vision, pp.697–713, 2016.

[39] E. Rehder, F. Wirth, M. Lauer, and C. Stiller, "Pedestrian prediction by planning using deep neural networks," arXiv preprint, arXiv:r1706.05904, 2017.

[40] N. Lee, W. Choi, P. Vernaza, C.B. Choy, P.H.S. Torr, and M.K. Chandraker, "DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents," Computer Vision and Pattern Recognition, pp.336–345, 2017.

[41] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," Computer Vision and Pattern Recognition, pp.3302–3309, 2014.

[42] H.S. Park, J.J. Hwang, Y. Niu, and J. Shi, "Egocentric future localization," Computer Vision and Pattern Recognition, pp.4697–4705, 2016.

[43] S. Su, J.P. Hong, J. Shi, and H.S. Park, "Predicting behaviors of basketball players from first person videos," Computer Vision and Pattern Recognition, pp.1502–1510, 2017.

[44] D. Xie, S. Todorovic, and S.C. Zhu, "Inferring 'Dark Matter' and 'Dark Energy' from videos," International Conference on Computer Vision, pp.2224–2231, 2013.

[45] S. Yi, H. Li, and X. Wang, "Pedestrian behavior understanding and prediction with deep neural networks," European Conference on Computer Vision, pp.263–279, 2016.

[46] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," Computer Vision and Pattern Recognition, pp.961–971, 2016.

[47] F. Bartoli, G. Lisanti, L. Ballan, and A.D. Bimbo, "Context-aware trajectory prediction," arXiv preprint, arXiv:1705.02503, 2017.

[48] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft + hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection," arXiv preprint, arXiv:1702.05552, 2017.

[49] T. Fernando, S. Denman, A. McFadyen, S. Sridharan, and C. Fookes, "Tree memory networks for modelling long-term temporal dependencies," arXiv preprint, arXiv:1703.04706, 2017.

[50] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," Computer Vision and Pattern Recognition, pp.7593–7602, 2018.

[51] N. Lee and K.M. Kitani, "Predicting wide receiver trajectories in american football," Winter Conference on Applications of Computer Vision, pp.1–9, 2016.

[52] S.Z. Bokhari and K.M. Kitani, "Long-term activity forecasting using first-person vision," Asian Conference on Computer Vision, pp.346–360, 2016.

[53] N. Rhinehart and K.M. Kitani, "First-person activity forecasging with online inverse reinforcement learning," International Conference on Computer Vision, pp.3696–3705, 2017.

[54] E. Rehder and H. Kloeden, "Goal-directed pedestrian prediction," International Conference on Computer

Vision Workshops, pp.139–147, 2015.

[55] K. Yamaguchi, A.C. Berg, L.E. Ortiz, and T.L. Berg, "Who are you with and where are you going?," Computer Vision and Pattern Recognition, pp.1345–1352, 2011.

[56] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," European Conference on Computer Vision, pp.549–565, 2016.

[57] R.E. Kalman, "A new approach to linear filtering and prediction problems," Journal of basic Engineering, vol.82, no.1, pp.35–45, 1960.

[58] N.J. Gordon, D.J. Salmond, and A.F.M. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," IEE Proceedings F - Radar and Signal Processing, vol.140, no.2, pp.107–113, 1993.

[59] G. Kitagawa, "Monte carlo filter and smoother for non-gaussian nonlinear state space models," Journal of Computational and Graphical Statistics, vol.5, no.1, pp.1–25, 1996.

[60] J.W. Robinson and A.J. Hartemink, "Non-stationary dynamic bayesian networks," Advances in Neural Information Processing Systems, pp.1369–1376, 2009.

[61] K. Cho, B. vanMerrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," Empirical Methods in Natural Language Processing, pp.1724–1734, 2014.

[62] D.P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," Advances in Neural Information Processing Systems, pp.3581–3589, 2014.

[63] S. Russell, "Learning agents for uncertain environments (extended abstract)," Annual Conference on Computational Learning Theory, pp.101–103, 1998.

[64] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," Physical review E, vol.51, no.5, pp.4282–4286, 1995.

[65] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," Computer Graphics Forum, vol.26, no.3, pp.655–664, 2007.

[66] S. Pellegrini, A. Ess, K. Schindler, and L. vanGool, "You'll never walk alone: Modeling social behavior for multi-target tracking," International Conference on Computer Vision, pp.261–268, 2009.

[67] B. Majecka, "Statistical models of pedestrian behaviour in the forum," PhD thesis, MSc Dissertation, School of Informatics, University of Edinburgh, 2009.

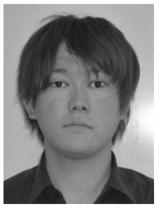
[68] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," Computer Vision and Pattern Recognition, pp.3457–3464, 2011.

[69] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," Computer Vision and Pattern Recognition, pp.3488–3496, 2015.

[70] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," Computer Vision and Pattern Recognition, pp.3354–3361, 2012.

[71] M.P. Dubuisson and A.K. Jain, "A modified hausdorff distance for object matching," International Conference on Pattern Recognition, vol.1, pp.566–568, 1994.

（平成 xx 年 xx 月 xx 日受付）



平川　　翼

2017 年 広島大学大学院博士課程後期課程情報工学専攻修了，2017 年 中部大学研究員．2014 年 独立行政法人日本学術振興会特別研究員 DC1．コンピュータビジョン，パターン認識，医用画像処理の研究に従事．

山下　隆義　（正員）

2002 年 奈良先端科学技術大学院大学博士前期課程修了，2002 年 オムロン株式会社入社，2011 年 中部大学大学院博士後期課程修了，2014 年 中部大学講師，2017 年 中部大学准教授．人の理解に向けた動画像処理，パターン認識・機械学習の研究に従事．

玉木　　徹　（正員）

2001 年 名古屋大学大学院工学研究科博士課程後期課程情報工学専攻修了．同年新潟大学工学部情報工学科助手．2005 年広島大学大学院工学研究科准教授．コンピュータビジョン，画像認識の研究に従事．

藤吉　弘亘　（正員）

1997 年 中部大学大学院博士後期課程修了，1997 年 米カーネギーメロン大学ロボット工学研究所 Postdoctoral Fellow，2000 年 中部大学工学部情報工学科講師，2004 年 中部大学准教授，2005 年 米カーネギーメロン大学ロボット工学研究所客員研究員（～2006 年），2010 年 中部大学教授，2014 年名古屋大学客員教授．計算機視覚，動画像処理，パターン認識・理解の研究に従事．

Abstract Path prediction is a fundamental task for estimating how pedestrians or vehicles are going to move in a scene. Because path prediction as a task of computer vision uses video as input, various information used for prediction, such as the environment surrounding the target and the internal state of the target, need to be estimated from the video in addition to predicting paths. Many prediction approaches that include understanding the environment and the internal state have been proposed. In this survey, we systematically summarize methods of path prediction that take video as input and and extract features from the video. Moreover, we introduce datasets and evaluation metrics used to path prediction methods quantitatively.

Key words Path prediction, Activity Forecasting, Dataset, Survey