社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

[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† 山下　隆義† 玉木　徹†† 藤吉　弘亘†

† 中部大学　〒 487–8501 愛知県春日井市松本町 1200
†† 広島大学大学院 工学研究科　〒 739–8527 広島県東広島市鏡山 1–4–1
E-mail: †hirakawa@mprg.cs.chubu.ac.jp, ††{yamashita,hf}@cs.chubu.ac.jp, †††tamaki@hiroshima-u.ac.jp

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し，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監視カメラ映像の解析や自動車の自動運転技術，ロボットの自律制御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入力は動画像のみである場合が多い．経路予測についても同様であり，周囲の環境や向きなどの予測対象の状態などの情報を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 [1], [2] や属性推定 [3]，セマンティックセグメンテーション [4] 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，動画像を用いた行動予測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．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対象の位置を解析する問題である．一方，行動予測では未来の時刻の情報は取得することができないため，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を行なっている場合 [5] や，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のシーンの 3 次元データを使用使用する場合が多く，さらにロボットが走行する環境が限定的な場合，シーン全体の環境が所与とされていることが多い．一方，本サーベイでは動画像および動画像中に存在する予測対象の位置を入力とするような経路予測問題を扱う．

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 [6]〜[8]．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，行動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課題である．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し，この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) の経路予測部である．特徴抽出部では主に，シーンの環境を理解するための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．経路予測部では経路を予測するために様々な手法が提案されているが，推論のアプローチにより予測手法を 4 種類に分類することができる．

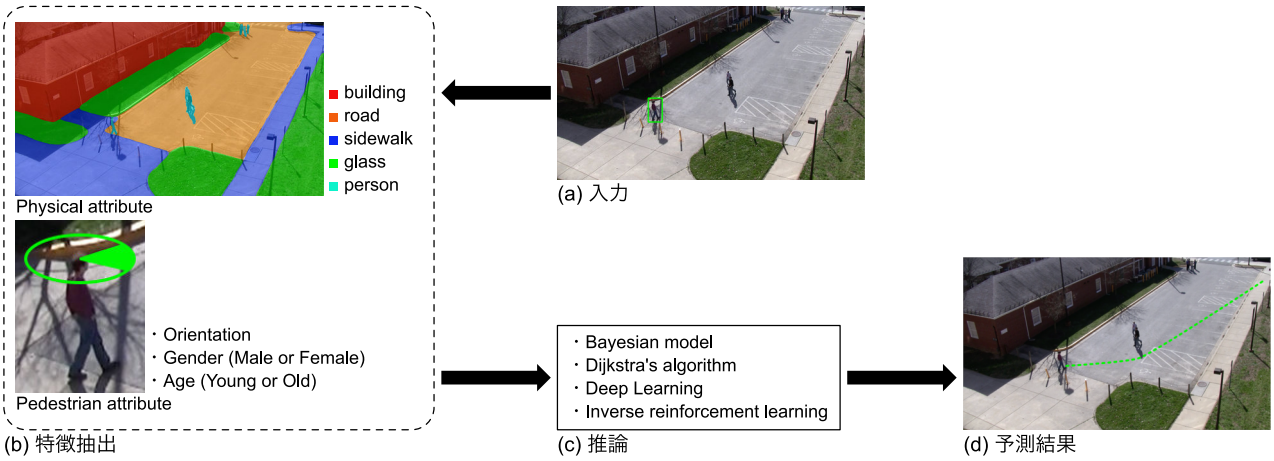

図 1　経路予測手法の処理の流れ．文献 [9] より改変．

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 [10]
		Superpixel-based MRF [11]
		Fully Convolutional Networks [12], [13]
	コスト	Bag-of-Visual Words
		Spatial Matching Network [14]
	シーン全体の特徴ベクトル	Pretrained AlexNet [15]
		Siamese Network [16]
予測対象	歩行者の位置	HOG + SVM detector [17]
	対象の向き	Bayesian orientation estimation [18]
		Orientation Network [14]
	身体的属性	AlexNet-based multi-task learning [19]
	対象物の特徴ベクトル	Mid-level patch features [20]

以上を踏まえ，本稿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し，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．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 節では経路予測に用いられる特徴抽出について述べる．3 節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 節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 節で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．動画像による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．抽出される特徴には，表 1 に示すような様々なものが存在するが，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，本節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自動車などの障害物を避けつつできるだけ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 [21]〜[24]．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がある領域などを推定することが可能である．Kitani ら [21] 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，図 2 に示すように，階層的に領域を分割するセグメンテーション手法 [10] を用いて各ラベルの事後確率を推定しており，各ラベルの事後確率を特徴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．また，Rehder ら [23] は，Fully Convolutional Network (FCN) [12], [13] を用いて推定したセグメンテーション結果を予測に用いている．

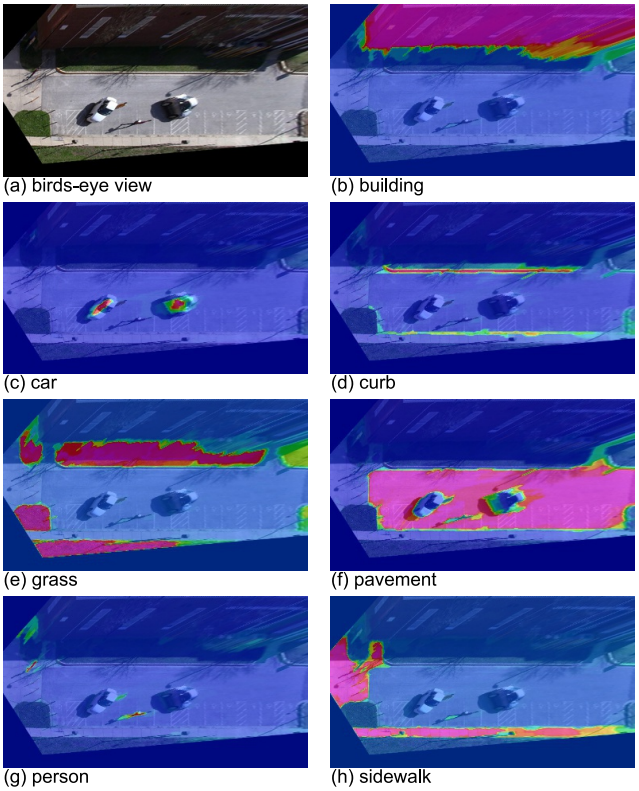

図 2　セマンティックセグメンテーションを用いた環境属性の例．文献 [21] より引用．

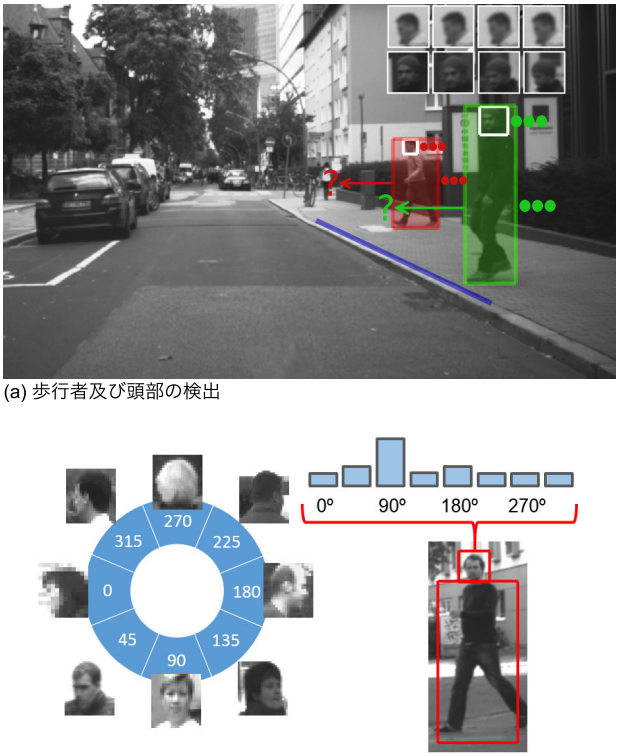

図 3　歩行者の頭部の向きの推定例．文献 [28] より引用．

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 [14], [25]．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る．Walker ら [25] は，最近傍探索により学習データから類似したテクスチャのパッチを検索し，その報酬を微小領域にマッピングす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 [14] は，Spatial Matching Network と呼ばる 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いう仮定のもと，シーンを表現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．そのためには，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があり，近年の深層学習の発達により CNN が主に用いられている．Park ら [26] 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 [15] 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することで，学習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 [27] は AlexNet ベースの Siamese Network [16] 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 [14], [19], [28]．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 [28] 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 [17] 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 [18]．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，車道まえで停止するような予測を実現している．

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．Wei ら [19] は

表 2　経路予測手法の分類.

分類	文献	年代	手法	映像視点	入力	出力	特徴抽出 環境	特徴抽出 対象
Bayesian	Schneider and Gavrila [29]	2013	KF	車載	座標	座標		
	Kooij et al. [28]	2014	DBN	車載	動画	座標	✓	✓
	Ballan et al. [22]	2016	DBN	鳥瞰	動画	座標	✓	
Energy minimization	Xie et al. [30]	2013	Dijkstra	鳥瞰	動画	確率分布	✓	
	Walker et al. [25]	2014	Dijkstra	監視カメラ	動画	確率分布	✓	
	Huang et al. [14]	2016	Dijkstra	監視カメラ	画像	確率分布	✓	✓
DL	Shuai et al. [31]	2016	CNN	監視カメラ	座標	座標		
	Alahi et al. [32]	2016	LSTM	鳥瞰	座標	座標		
	Fernando et al. [33]	2017	LSTM	鳥瞰	座標	座標		
	Fernando et al. [34]	2017	LSTM	鳥瞰	座標	座標		
	Lee et al. [24]	2017	RNN Enc.-Dec.	車載	動画	座標	✓	
IRL	Kitani et al. [21]	2012	IRL	鳥瞰	動画	確率分布	✓	
	Lee and Kitani [35]	2016	IRL	鳥瞰	動画	確率分布	✓	
	Bokhari and Kitani [36]	2016	IRL	一人称	動画	確率分布		✓
	Rhinehart and Kitani [37]	2017	IRL	一人称	動画	確率分布		✓
	Ma et al. [19]	2017	IRL	監視カメラ	画像	確率分布		✓
	Rehder et al. [23]	2017	IRL	車載	動画	確率分布	✓	
Others	Keller and Gavrila [38]	2014	optical flow	車載	動画	座標		
	Rehder and Koloeden [39]	2015	Markov process	車載	動画	座標	✓	
	Park et al. [26]	2016	Data driven	一人称	動画	座標	✓	
	Su et al. [27]	2017	Data driven	一人称	動画	座標	✓	
	Yamaguchi et al. [40]	2011	Social force	鳥瞰	動画	座標		
	Robicquet et al. [41]	2016	Social force	鳥瞰	動画	座標		

AlexNet [15] 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 [25] 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 [20].

3. 予 測 手 法

動画像からの特徴抽出が終了すると，その特徴を用いて経路予測が行われる．表 2 に示すような経路予測手法が提案されているが，これらはベース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．そこで本節では，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や粒子フィルタ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の更新する 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逐次的にを推定する．すなわち，内部状態が各時刻

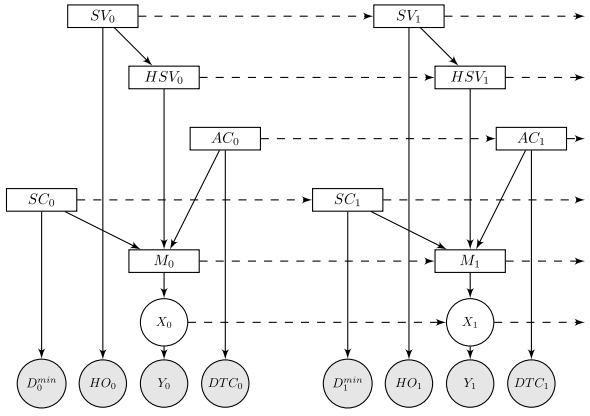

図 4　SLDS を導入した DBN のグラフィカルモデル．文献 [28] より引用.

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される．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．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出力している．

Schneider ら [29] 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することを目的として，経路予測を行なっている．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．これは経路予測問題の初期に取り組まれた研究であり，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) を用いた予測手法も存在する [22], [28]．Kooij ら [28] 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いる．彼らは図 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，歩行者の頭の向きや車両までの距離，歩行者と縁石間の距離などの動画像から抽出した特徴を観測として用いており，これらの特徴を用いて内部状態を更新することで，人検出法で推定した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では，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，シーン全体または一定期間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として扱うことができるためダイクストラ法を用いて解を推定しており，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 [14] 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 [25] 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を比較することで，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 [30] 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，Convolutional Neural Network (CNN) や 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が直接出力されるような構造となっている．そのため，2 節で述べたような特徴抽出は明示的に行われず，特徴抽出と予測が一貫して行われるような構造となっている．

移動経路は 2 次元座標の系列データであるため，LSTM を用いた経路予測手法がいくつか提案されている．Alahi らは [32] 複数人の歩行者の移動経路を予測する際に，歩行者同士の衝突を避けるために Social-pooling (S-Pooing) Layer を提案しており，自身の周辺に存在する歩行者の LSTM の中間層出力を S-Pooling Layer に入力することで空間的関係を保存し次の時刻の LSTM に入力する．これにより，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．Fernando ら [33] 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 [24] は RNN Encoder-Decoder および Conditional Variational Autoencoder 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複数の予測軌跡を生成する．生成された軌跡は重み付けされ，シーンの環境に対して妥当な移動軌跡かどうかを決定する．その後，重み付けされた複数の軌跡を改善することで，最終的な単一の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，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 [34] 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・連結する 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 [31] 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，decomvolution 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．

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

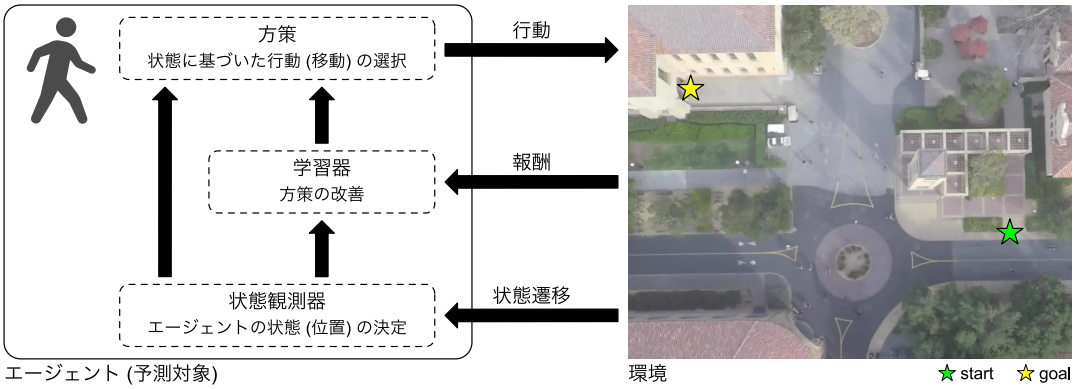

図 5　強化学習の処理の流れ．文献 [41] より改変．

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に現実の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．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．この報酬をもとに，予測対象の行動を連続して推定することで経路予測が実現される．

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 [5] として用いられているが，Kitani ら [21] 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．Kitani らの手法では，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 [21] 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が決定されるが，その最適な重みベクトルを学習

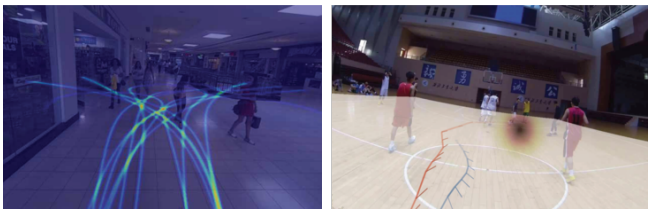

図 6　一人称視点映像からの経路予測．左: 文献 [26]，右: 文献 [27] より引用．

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 [35] 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，Wei ら [19] 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 [23] 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 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 [36] 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 [37] はより広範囲での経路予測を実現している．

3.5　その他のアプローチ

大半の経路予測手法が上記の 4 種類のアプローチにされるが，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 [42] 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

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	歩行者以外の対象物	追加情報
UCY [43]	2007	✓(注1)	786	鳥瞰	3	–	–
ETH [44]	2009	✓(注2)	750	鳥瞰	2	–	–
Edinburagh Informatics Forum [45]	2009	✓(注3)	95,998	鳥瞰	1	–	–
Stanford Drone [41]	2016	✓(注4)	11,216	鳥瞰	8	bikers, skateboarders cars, buses, golf carts	–
VIRAT [9]	2011	✓(注5)	4,021	監視カメラ	11	car, bike	物体の座標 行動カテゴリ
Town Centre [46]	2011	✓(注6)	230	監視カメラ	1	–	頭部の座標
Grand Central Station [47]	2015	✓(注7)	12,600	監視カメラ	1	–	–
Daimler [29]	2013	✓(注8)	68	車載	–	–	ステレオカメラ
KITTI [48]	2012	✓(注9)	6,336	車載	–	car	ステレオカメラ LIDAR 地図情報
EgoMotion [26]	2016		–	一人称	26	–	ステレオカメラ
First-person Continuous Activity [37]	2017		–	一人称	17	–	物体情報

た予測手法として，Yamaguchi ら [40] は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．Ballan ら [41] は，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行うことを目的として，他クラスの Social Force Model を提案している．この手法では，予測対象ごとに他者との衝突を回避する際の距離などの情報を使用し “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する．予測対象の特徴量から回避行動のクラスタを決定し，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 [38] 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 [39] 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 [26] 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．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 [27] 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マッピングすると同時に，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，マッピングした軌跡の座標等から目的関数を定義し，最適な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，学習に必要な軌跡の数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，全ての手法において統一的なデータセットの使用を行うことが難しいためである．そこで本節では，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 俯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，最もよく用いられるのは駅構内や市街地の歩行者を監視カメラ等を用いて撮影した俯瞰

(注1)：https://graphics.cs.ucy.ac.cy/research/downloads/crowd-data
(注2)：http://www.vision.ee.ethz.ch/en/datasets/
(注3)：http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING/
(注4)：http://cvgl.stanford.edu/projects/uav_data/
(注5)：http://www.viratdata.org/
(注6)：http://www.robots.ox.ac.uk/~lav/Papers/benfold_reid_cvpr2011/benfold_reid_cvpr2011.html
(注7)：http://www.ee.cuhk.edu.hk/~xgwang/grandcentral.html
(注8)：http://www.gavrila.net/Datasets/Daimler_Pedestrian_Benchmark_D/daimler_pedestrian_benchmark_d.html
(注9)：http://www.cvlibs.net/datasets/kitti/

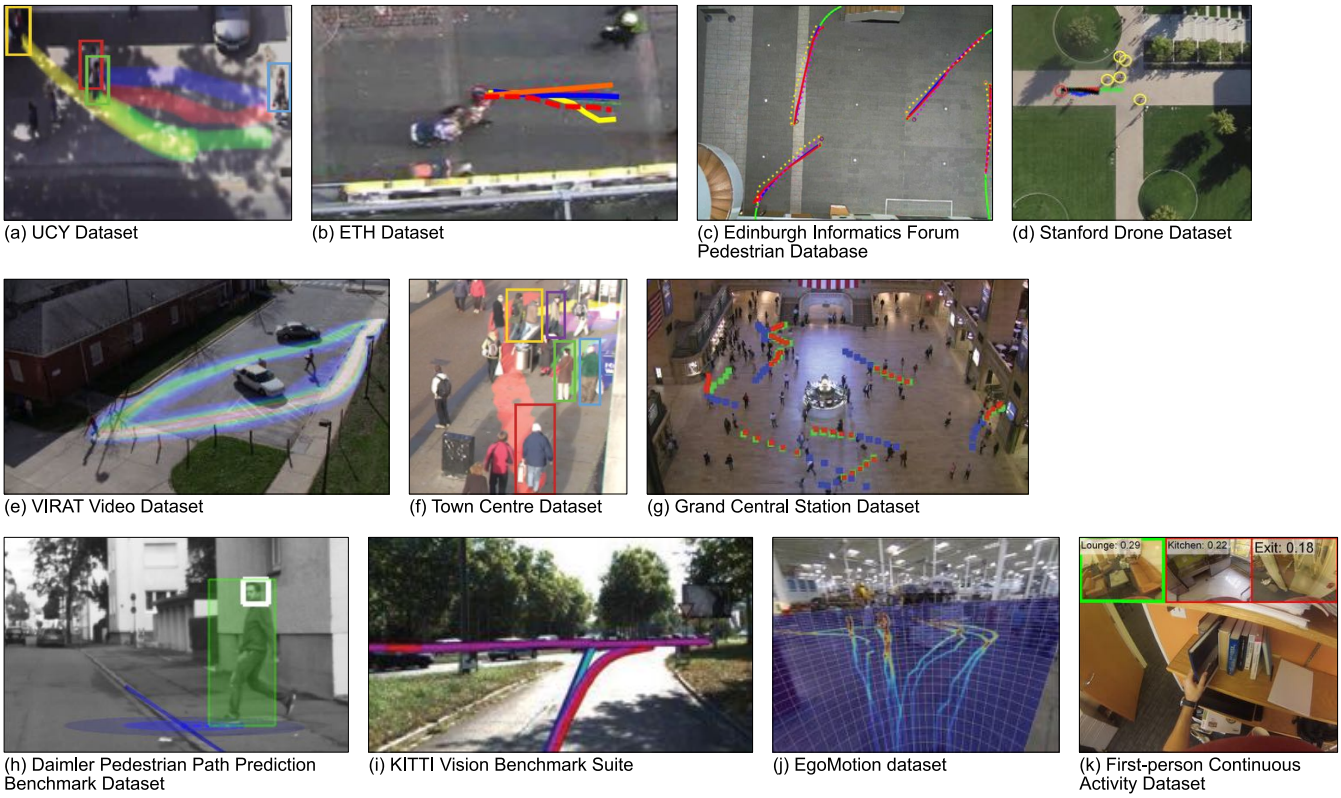

図 7　経路予測のデータセット及び経路予測結果の例．文献 [19], [21], [24], [26], [28], [31], [32], [34], [37], [41] より引用及び改変．

視点映像の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たものであるが，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．

鳥瞰視点映像のデータセット

UCY Dataset [43] 及び ETH Dataset [44] 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，歩行者のみが存在しているような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 [45] 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 と同じような環境で撮影されたデータセットである．このデータセットの特徴は 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 [41] 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 などの複数種類の移動物体の情報が公開されている．

監視カメラ映像のデータセット

上記のデータセットではカメラをシーンのほぼ真上から撮影したデータセットであった．一方，VIRAT Video Dataset [9] 及び Town Centre Dataset [46] は図 7(e, f) に示すように，監視カメラを用いて斜め上から撮影された動画像から構成されている．これらのデータセットでは真上からの撮影とは異なり，歩行者の身体的特徴等が観測できるため，歩行者の属性を考慮した経路予測を行うことが可能となる．VIRAT Video Dataset 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，動画像が撮影シーンは 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 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 [47] 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おり非常に複雑なデータセットとなっている．

4.2 車載カメラ映像のデータセット

経路予測は自動車の自動運転支援を目的としても研究されており，車載カメラ映像を用いたデータセットも用いられている．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．

Daimler Pedestrian Path Prediction Benchmark Dataset [29] 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そのまま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，車両前方を横切るような歩行者の映像が含まれている貴重なデータセットである．

KITTI Vision Benchmark Suite [48] 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 LIDER 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おり，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ことから車載カメラ映像での経路予測に有用とされている．

4.3 一人称視点映像のデータセット

上記の鳥瞰視点映像や車載カメラ映像では映像中に存在する予測対象の移動経路を予測していたのに対し，撮影者自身の移動経路を目的として一人称視点映像からの経路予測も行われている．Park ら [26] は屋内及び屋外を移動する際の一人称視点映像を用いて経路予測を行なっている．このために市街地や店内などの 26 のシーン撮影された一人称視点映像を用いて評価を行なっている．Rhinehart ら [37] 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. おわりに

本稿では，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 [24].

また，経路予測手法の評価に用いられるデータセット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 Pedestrina Path Prediction Benchmark Dataset が存在する．

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] 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.
[2] 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.
[3] 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.
[4] 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.
[5] 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, Oct. 2009.
[6] M.S. Ryoo, “Human activity prediction: Early recognition of ongoing activities from streaming videos,” International Conference on Computer Vision, pp.1036–1043, 2011.
[7] M. Hoai and F.D. laTorre, “Max-margin early event detectors,” Computer Vision and Pattern Recognition, pp.2863–2870, å 2012.
[8] 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, Sept. 2016.
[9] 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, June 2011.

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

[11] 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.

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

[13] 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, April 2017.

[14] 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, Dec. 2016.

[15] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, eds. by F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, pp.1097–1105, 2012.

[16] 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.

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

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

[19] 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. http://arxiv.org/abs/1604.01431

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

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

[22] 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.

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

[24] 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. http://arxiv.org/abs/1704.04394

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

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

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

[28] 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.

[29] 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.

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

[31] 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.

[32] 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, June 2016.

[33] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft + hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection," CoRR, 2017. http://arxiv.org/abs/1702.05552

[34] T. Fernando, S. Denman, A. McFadyen, S. Sridharan, and C. Fookes, "Tree memory networks for modelling long-term temporal dependencies," CoRR, 2017. http://arxiv.org/abs/1703.04706

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

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

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

[38] 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, April 2014.

[39] E. Rehder and H. Kloeden, "Goal-directed pedestrian prediction," Workshop on International Conference on Computer Vision, pp.139–147, Dec. 2015.

[40] K. Yamaguchi, A.C. Berg, L.E. Ortiz, and T.L. Berg, "Who are you with and where are you going?," CVPR 2011, pp.1345–1352, 2011.

[41] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," European Conference on Computer Vision, eds. by B. Leibe, J. Matas, N. Sebe, and M. Welling, pp.549–565, Springer International Publishing, Cham, 2016.

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

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

[44] 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.

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

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

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

[48] 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.