コンテキスト情報を用いた複数人の経路予測に関する研究

TP18001 家崎遥　　指導教授：山下隆義

1. はじめに

歩行者の経路予測は，運転手の判断を支援する運転支援システムや防犯カメラ映像を用いた映像解析などで必要とされている．従来の深層学習を用いた経路予測は，1 つのネットワークを用いて，1 人の経路を予測する [1]．そのため，予測対象が複数人の場合，同じネットワークを用いて，複数人の経路を 1 人ずつ予測しなければならない．そこで本研究では，歩行者の矩形情報と移動情報をコンテキスト情報として埋め込んだ画像をコンテキスト画像として用い，複数人の経路を 1 つのネットワークで同時に予測する手法を提案する．

2. LSTM-Bayesian

従来の経路予測手法として，歩行者情報を表す矩形領域と車載カメラ画像および車の移動量を入力する LSTM-Bayesian[1] が提案されている．LSTM-Bayesian は，歩行者の矩形領域推定ネットワークと車の移動量推定ネットワークで構成されている．本手法では，自車と歩行者の状態を個別に推定することで，高精度な経路予測を実現している．しかし，1 人の予測に対して 2 つのネットワークによる処理を繰り返し行うため，複数人を予測する場合には計算コストが高くなる．加えて，入力画像が車載カメラ全体のため，歩行者情報以外の特徴量がノイズとなる問題点がある．

3. 提案手法

従来手法の問題点を解決するために，コンテキスト画像を用いた確率的な経路予測を提案する．まず，複数人の経路を予測するネットワークは，Encoder-Predictor 構造とする．Encoder-Predictor 構造の内部に Convolutional Long Short-Term Memory Network (ConvLSTM) を用いることで，歩行者のコンテキストを考慮した経路予測が実現できる．そして，ネットワークには歩行者の矩形情報と移動情報を埋め込んだコンテキスト画像を入力する．

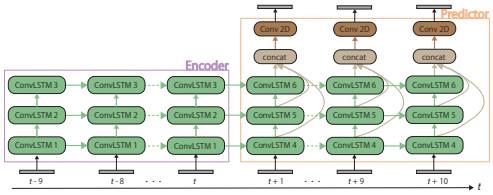
3.1. コンテキスト画像

歩行者の経路予測におけるコンテキストとして，画像上の歩行者を含む周辺の矩形領域をコンテキストとして用いる．本研究では，人の色や形のコンテキストを表す RGB 画像と移動方向のコンテキストを表す Opticacl flow (Flow) 画像の 2 種類を使用する．Flow 画像はフローベクトルであり，本研究では HSV カラー表現した 3 チャンネルの画像として用いる．コンテキスト画像は，矩形領域外にマスク処理を施して作成する．人の位置を表す矩形領域の左上・右下の座標を $(x_1, y_1), (x_2, y_2)$ とした時のコンテキスト画像の作成例を図 1 に示す．

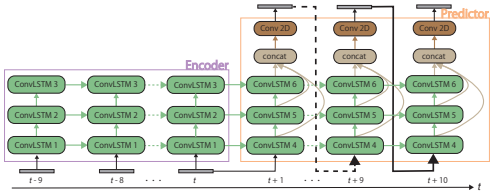
図 1: Flow 画像を用いた動画像のサンプル作成例

3.2. ネットワーク構造

提案手法のネットワークの構造を図 2(a) および 2(b) に示す．Encoder-Predictor は，時系列の画像を扱うことのできる ConvLSTM 層と畳み込み層（Conv 2D）で構成される．はじめに，過去から現在までに取得したコンテキスト画像を 3 層の ConvLSTM 層から成る Encoder へ入力して内部表現を獲得する．そして，Predictor に画素値が 0 のコンテキストがない画像と合わせて内部表現を入力する．Predictor の ConvLSTM 各層は，Encoder で学習された ConvLSTM 各層の最終状態（メモリセルと中間層のユニット値）を引き継ぐ．そして，Predictor は，現在より 1 時刻先から数フレーム先まで，画素値が 0 のコンテキストがない画像を入力する．最後に，ConvLSTM 各層の出力を連結し，畳み込み層で特徴量の畳み込み処理を行い，予測画像を逐次出力する．



(a) モデル 1



(b) モデル 2

図 2: 提案手法のネットワーク構造．Predictor が非再帰構造のモデル 1(a) と再帰構造のモデル 2(b)．

本研究では，このネットワークをモデル 1 とする．さらに，Predictor が異なるモデル 2 は，Predictor の予測の 1 時刻目 $(t+1)$ に Encoder の最終フレームを入力し，2 時刻目以降は，前時刻に予測した画像を入力する再帰的なモデルとする．2 つのモデルを比較実験することで，未来に獲得できるコンテキスト情報による効果を検証する．

3.3. 高速化とメモリ削減

本研究では，提案モデルの高速化およびメモリ削減も行う．ネットワークの層数やフィルタサイズの増加に伴い，内部パラメータ数が増加するため，計算コストが高くなる [2]．ConvLSTM の畳み込み処理は，メモリセルの各ゲートへ入力する入力値と前時刻の中間層に対して行う．そこで，ConvLSTM の入力値に対する畳み込み処理をメモリセルで行わず，メモリセルの外部で行う．

4. 評価実験

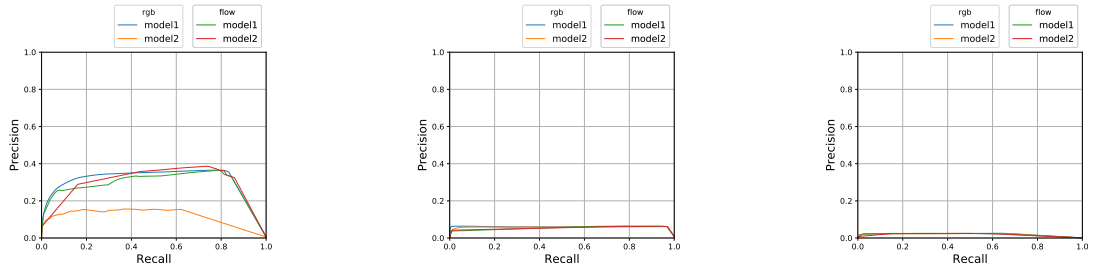
提案手法の有効性を 3 つのデータセットを用いて評価する．1 人称視点のデータセットとして，Cityscapes，3 人称視点のデータセットとして，Town Centre と FDST を使用する．Cityscapes は，学習が 1,978 本，評価が 1,036 本である．Town Centre は，学習が 135 本，評価が 57 本である．FDST は，学習が 7,020 本，評価が 4,420 本である．入出力のフレーム数は，各 10 フレームとする．学習条件はバッチサイズが 2，最適化手法は学習率が 0.001 の Adam とする．損失関数には，Negative log likelihood (NLL) を使用する．本実験では，2 種類のコンテキスト画像と 2 種類のモデルを用いた比較実験を行う．

4.1. 評価方法

評価指標には，NLL と Precision および Recall を用いる．NLL は，予測分布と真値の分布の誤差を算出する．予測分布は，ネットワークから出力された最大値をもとに正規化を行う．Precision と Recall は，Precision-Recall (PR) 曲線と Average Precision (AP) により評価する．PR 曲線は，真値 BB の範囲と予測分布から Precision と Recall を求めて算出する．Precision は，予測分布に対する真値 BB の範囲の割合である．一方，Recall は，真値 BB の範囲に対する予測分布の割合である．AP は，閾値毎に算出した Precision の平均値である．PR 曲線と AP の範囲は，予測した確率分布に対する閾値であるため，取り得る範囲は 0.0〜1.0 である．

4.2. NLL

従来手法として [1] を用いる．表 1 に予測分布の NLL を示す．表 1 より，提案手法は従来手法より約 3.0 誤差を抑



(a) Cityscapes (b) Town Centre (c) FDST

図 3: PR 曲線

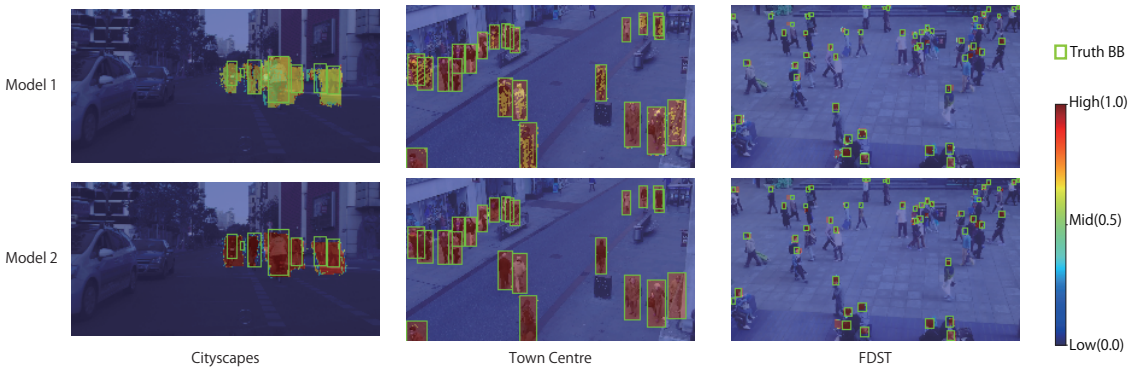


図 4: 異なる視点のデータセットを用いた予測の結果例（$t = t + 10$）

表 1: 予測分布の誤差

Method	Prediction frames	Image	NLL		
			Cityscapes	Town Centre	FDST
従来手法	8	rgb	3.92	-	-
Our model 1	10	rgb	0.692	0.670	0.692
		flow	0.690	0.650	**0.691**
Our model 2	10	rgb	0.694	0.670	0.692
		flow	**0.689**	**0.647**	**0.691**

表 2: AP

Method	Prediction frames	Image	AP		
			Cityscapes	Town Centre	FDST
Our model 1	10	rgb	0.182	0.055	0.018
		flow	0.290	**0.057**	**0.022**
Our model 2	10	rgb	0.127	0.051	0.018
		flow	**0.325**	**0.057**	0.021

制できていることが分かる．従来手法が 8 フレーム予測先までの対し，提案手法は 10 フレーム先まで予測しているため，より長期的な予測が可能であると言える．視点の異なるデータセット毎に比較すると，3 人称視点の Town Centre は Cityscapes や FDST より誤差が低い．Town Centre は固定カメラで撮影された動画像であり，かつ歩行者の全身を矩形領域としているため，歩行者の移動量が少なくコンテキスト情報が多い．そのため，高精度な経路予測が可能であると考えられる．

4.3. Precision と Recall

Cityscapes の PR 曲線を図 3(a)，Town Centre の PR 曲線を図 3(b)，FDST の PR 曲線を図 3(c)，AP の結果を表 2 に示す．図 3(a) より，Recall が低い場合，モデル 1 の方がモデル 2 より Precision が高い．一方，Recall が 0.4 を越えると，Flow 画像を用いたモデル 2 の方が Precision が高く，Flow 画像は RGB 画像より高精度に予測していると考えられる．しかし，図 3(b) と図 3(c) は，Precision が低い．これは，真値の矩形領域は網羅できているが，矩形領域外も予測しているためと考えられる．3 人称視点のデータセットは，1 人称視点と比較して人数が多く，隣人との領域が重なりやすい．また，FDST のような小領域は予測が難しい．そのため，最も Precision が低い結果になったと考えられる．また表 2 より，ほとんどの場合において Flow 画像を用いたモデル 2 が高い AP である．これらの結果より，RGB 画像を用いたモデル 1 は，網羅性が高いが正確性が低い．よって，AP 値が高く PR 曲線で Precision

表 3: モデル 2 の高速化とメモリ削減

ConvLSTM [層数]	入力値の畳込み処理	チャンネル数	NLL	学習時間 [h]	処理速度 [s/sample]	GPU の使用量 [GB]
3	w	128, 64, 64	0.647	21.6	0.871	19
2	w	128, 128	0.648	24.0	1.723	18
		64, 64	0.653	13.3	0.500	18
	w/o	128, 128	0.647	22.2	0.795	17
		64, 64	0.648	**9.5**	**0.440**	**9**

の高い Flow 画像を用いたモデル 2 が，安定した予測モデルであると言える．

4.4. 経路の予測結果

mIoU の評価で高精度であった Flow 画像を用いた予測結果例を図 4 に示す．モデル 1 と 2 を比較すると，モデル 2 はモデル 1 より高精度に予測できていることが分かる．また，異なる視点のデータセットで比較すると，歩行者の変化量が大きい Cityscapes では，Town Centre や FDST より予測が困難であることが分かる．

4.5. 計算コストの比較

4.2節の NLL の評価で，最高精度であった Town Centre の Flow 画像を用いたモデル 2 で計算コストの評価を行う．評価結果を表 3 に示す．まず，ConvLSTM の層数を削減することで誤差が大きくなり，学習時間も増加することが確認できた．一方，提案手法は ConvLSTM の層を削減する前と同等の精度を維持できていることが分かる．さらに，学習時間を約 12 時間短縮，処理速度を高速化し，GPU 使用量を削減した．

5. おわりに

本研究では，Flow のコンテキスト画像を用いることで，高精度に予測できることを確認できた．また，ネットワークは再帰的な構造にすることで，精度が向上することを確認できた．今後は，複数人の経路を同時に予測する際，1 人 1 人の経路に対する評価を検討する．

参考文献

[1] A. Bhattacharyya, *et al.*,“Long-Term On-Board Prediction of People in Traffic Scenes under Uncertainty”, CVPR, 2018.

[2] X. Shi, *et al.*, “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting”, NeurIPS, 2015.

研究業績

[1] H. Iesaki, *et al.*, “Simultaneous Visual Context-aware Path Prediction”, VISAPP, 2020.

(他　学会発表 2 件)