BiRNN を用いたアテンションモデルによる歩行者軌道予測

呉 家旭，Hanwool Woo
田村 雄介，Alessandro Moro, Stefano Massaroli，山下 淳，淺間 一

Pedestrian trajectory prediction using BiRNN attention model

Jiaxu WU, Hanwool WOO
Yusuke TAMURA, Alessandro MORO, Stefano MASSAROLI,
Atsushi YAMASHITA and Hajime ASAMA

Department of Precision Engineering, The University of Tokyo
7-3-1Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

This paper presents a novel approach for pedestrian trajectory prediction. In particular, we developed a novel attention model using bidirectional recurrent neural networks (BiRNNs). The difficulty of incorporating social interactions into the model has been addressed. Thanks to the special structure of BiRNNs enhanced by the attention mechanism, a proximity-independ model of the relative importance of each pedestrian. The main difference between our and the previous approaches is that BiRNN allows us to employ information on the future state of the pedestrians. We tested the performance of our method on several public datasets. The proposed model outperforms the current state-of-the-art approaches on most of these datasets. Furthermore, we analyze the learnt attention scores to prove the advantages of BiRNNs on recognizing social interactions

Key Words :Pedestrian trajectory prediciton; artificial neural networks; attention model

1. はじめに

近年，ショッピングモールや空港などの公共の場における移動ロボットの実用化が進んでおり，ロボットと人間が共存する環境における安全問題が注目されている．移動ロボットは掃除や運搬などのタスクを実行すると共に，周囲の歩行者との衝突を自律的に回避することが強く求められる．歩行者の軌道を予測し，妨げにならないように移動計画を自律的に行う機能は，ロボットと人間が共存する社会の実現のために必須であるといえる．

複雑な人間の動きを予測するためには，行動パターンを理解する必要がある．混雑した環境における歩行者の行動パターンは，自らの状態や目的だけでなく，周囲の歩行者の行動にも依存すると考えられる．すなわち，周囲の歩行者とのインタラクションを考慮することで, 適切な歩行者の軌道予測が可能となる．Lewin らは人間の行動をモデリングするため，Field theory を提案した[1]．Field theory は人間の行動 B をその周囲の環境に対する理解 P と環境を表す E の関数$B = f(P, E)$として考える．

東京大学大学院工学系研究科精密工学専攻（〒113-8656 東京都文京区本郷 7-3-1）wujiaxu@robot.t.u-tokyo.ac.jp

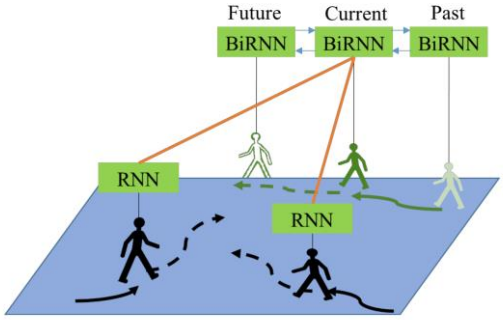
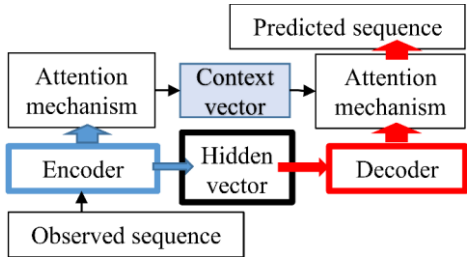


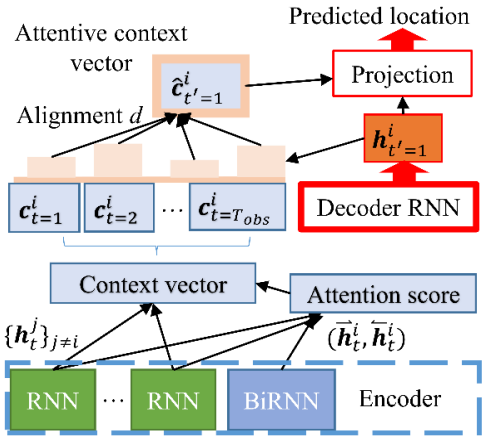
Fig. 1 Illustration of our approach, BiRNN encoder-decoder framework with attention.

歩行者の軌道を解析的に予測するため, Helbing らは Social force model（SFM）を用いてその関数を定式化した[2]．SFM は歩行者とその目的地との間に引力を，他者との間に斥力を発生させ，歩行者のダイナミックスを計算し，未来の位置を予測する．その後，SFM は Tamura ら[3]によって改善されてきたが，力を計算する関数の中のパラメータは試行錯誤によって決められるため，モデルの構築には膨大な労力を必要とし，様々なシナリオに対応する汎化能力は低いと考えられる．

一方，Kitani ら[4]や Pedllegrini [5]らは機械学習を用いて，歩行者の軌道データに基づいて歩行者の過去の軌道から未来の軌道にマッピング可能な手法を提案した．



(a) Overview of the model



(b) Illustration of attention mechanism

Fig. 2 Illustration of attention model

しかし，歩行者間のインタラクションに対する考慮が不十分であり，混雑した環境下では適切な予測が行われない可能性がある．Alahi らや Gupta らは人工ニューラルネットワークに Social pooling[6]や Max pooling[7]などのレイヤーを追加し，歩行者間のインタラクションをネットワークに組み込む手法を提案した．しかし，ここでのインタラクションは歩行者間の相対距離のみに依存し，歩行者が周囲の人の行動をどのように理解し自分の行動に反映するかを学習させることはできなかった．

本研究では，Bi-directional Recurrent Neural Network（BiRNN）を用いたアテンションモデルによる軌道予測手法を新規に提案する．BiRNN は Recurrent Neural Network（RNN）と違い，時系列データを順番（forward）に入力する RNN と逆順（backward）に入力するもう 1 つの RNN が存在する[8]（図 1 参照）．RNN で歩行者の軌道を入力時系列データとして処理する際，出力される計算結果は hidden vector と言い、過去軌道のメモリを持つ．ゆえに，ある時刻 t において backward の RNN が出力した hidden vector は未来の軌道の情報を持つ．アテンションモデルは歩行者の過去と未来の情報から歩行者間のインタラクションを学習し，ネットワークに組み込む．

本論文では，まず問題設定を説明する．次に BiRNN を用いたアテンションモデルの構造と各要素について説明する．その後，実験の方法を説明し，テストにより提案手法の有効性を示す．最後に提案手法の性能評価を行い，考察を述べる．

2. 問題設定

本研究では，人間以外のオブジェクトが歩行者の軌道に与える影響は小さく，俯瞰視点からすべての歩行者の軌道が追従できるシナリオを想定する．任意の時刻 t における i 番目の歩行者の位置を(x_t^i, y_t^i)とする．

以上の設定から，歩行者の軌道予測問題は以下のように記述できる．歩行者の総数を N とすると，$t = 1$ から $t = T_{obs}$ までの軌道 $Q = \{(x_t^i, y_t^i) | i = 1, 2, \dots, N; t = 1, 2, \dots, T_{obs}\}$が与えられ，$t = T_{obs} + 1$から$t = T_{obs} + T_{pred}$までの軌道を予測する．ここで，$T_{obs}$と$T_{pred}$はそれぞれ観測時間と予測時間を表す．

3. BiRNN を用いたアテンションモデル

3・1 提案手法の概要 本研究では，観測された歩行者の軌道と周辺環境における他者とのインタラクション（歩行者に対する周辺環境の影響）を用いて、歩行者の未来の軌道を予測するアテンションモデルを構築する．アテンションモデルは encoder-decoder モデルの一種であり，アテンションメカニズム（attention mechanism）という関連付けの方法を用いて，異なる入力の間の関係を調べることができる[9]．本研究で提案するモデルにおいて，encoder は BiRNN を用いて観測した歩行者の軌道を処理し，時系列データ（軌道 Q）を hidden vector にマッピングする．また，時刻ごとにアテンションメカニズムを用いて，個々の歩行者と周辺の歩行者とのインタラクションを推定し，context vector を計算する（図 2(a)参照）．decoder は encoder から各歩行者の軌道に対応する hidden vector を入力とし，context vector を歩行者に対する周辺環境の影響として参照しながら歩行者のもう 1 つの RNN で軌道を予測する．

3・2 BiRNN と RNN による入力軌道の処理 毎時刻における歩行者の位置(x_t^i, y_t^i)は，全結合層により状態ベクトルs_t^iに埋め込まれる．そして，シナリオにいる N 名の歩行者のうち，i 番目の歩行者を予測対象とする際，encoder は BiRNN を用いて i 番目の歩行者の時系列状態ベクトルの$\{s_t^i | t = 1, 2, \dots, T_{obs}\}$を処理し，毎時刻における forward と backward の hidden vector $\overrightarrow{h_t^i}$と$\overleftarrow{h_t^i}$を算出する．一方，他者の時系列状態ベクトル

$\{\boldsymbol{s}_t^j | j \neq i;\ t = 1, 2, \ldots, T_{obs}\}$が RNN によって処理され，毎時刻において，他者の hidden vector $\boldsymbol{h}_t^j$が得られる．BiRNN の機能により，観測時間中の任意の時刻tにおいて，観測の開始から時刻tまでの推定対象者と周辺他者の軌道のメモリが$\overrightarrow{\boldsymbol{h}}_t^i$と$\boldsymbol{h}_t^j$に保存されるだけではなく，時刻$t$から観測の終了までの推定対象の軌道がメモリとして$\overleftarrow{\boldsymbol{h}}_t^i$に保存される．

3・3　アテンションメカニズムによる歩行者インタラクションの推定　先行研究とは異なり，本研究では同時に$\overrightarrow{\boldsymbol{h}}_t^i$と$\boldsymbol{h}_t^j$，$\overleftarrow{\boldsymbol{h}}_t^i$と$\boldsymbol{h}_t^j$の間にアテンションメカニズムを用い，forward と backward のアテンションスコア$a_t^{i,j}$と$b_t^{i,j}$を計算する（図 2(b)参照）．

$$a_t^{i,j} = \frac{\exp(\langle \overrightarrow{\boldsymbol{h}}_t^i, \boldsymbol{h}_t^j \rangle)}{\sum_{j \neq i} \exp(\langle \overrightarrow{\boldsymbol{h}}_t^i, \boldsymbol{h}_t^j \rangle)} \quad (1)$$

$$b_t^{i,j} = \frac{\exp(\langle \overleftarrow{\boldsymbol{h}}_t^i, \boldsymbol{h}_t^j \rangle)}{\sum_{j \neq i} \exp(\langle \overleftarrow{\boldsymbol{h}}_t^i, \boldsymbol{h}_t^j \rangle)} \quad (2)$$

ここで，$a_t^{i,j}$と$b_t^{i,j}$はそれぞれ$\overrightarrow{\boldsymbol{h}}_t^i$と$\boldsymbol{h}_t^j$，$\overleftarrow{\boldsymbol{h}}_t^i$と$\boldsymbol{h}_t^j$の間の類似度を計算しており，$a_t^{i,j}$は推定対象者と周辺他者の軌道の類似度を評価する．また，$b_t^{i,j}$は推定対象者の軌道と周辺他者の軌道の依存関係を評価する．これら 2 つのスコアを使い，各時刻における推定対象者に対する周辺環境の影響を代表する context vector $\boldsymbol{c}_t^i$を次のように計算する．

$$\boldsymbol{c}_t^i = [\textstyle\sum_{j \neq i} a_t^{i,j} \boldsymbol{h}_t^j, \sum_{j \neq i} b_t^{i,j} \boldsymbol{h}_t^j] \quad (3)$$

ここで$[*,*]$は２つのベクトル*を連結する操作を表す．以上のように，encoder は観測時間全体に渡って推定対象者のコンテキストを$\{\boldsymbol{c}_t^i |\ t = 1, 2, \ldots, T_{obs}\}$に集約することが可能となる．

3・4　decoder による軌道予測　decoder は Bottle neck から連結した encoder の最終 hidden vector $[\overrightarrow{\boldsymbol{h}}_t^i, \overleftarrow{\boldsymbol{h}}_t^i]$ $(t = T_{obs})$を受け取り，decoder の RNN の初期状態とする．観測された推定対象者の最終位置$(x_{T_{obs}}^i, y_{T_{obs}}^i)$を decoder の RNN に入力し，各時刻における decoder の hidden vector $\boldsymbol{h}_{t'}^i$を計算する（図 2(b)参照）．ここで，$t' \in \{1, 2, \ldots, T_{pred}\}$は予測時間における各時刻を表す．さらに，過去のコンテキスト$\{\boldsymbol{c}_t^i |\ t = 1, 2, \ldots, T_{obs}\}$の推定対象者の未来の軌道に対する影響を予測に組み込むため，提案手法はアテンションメカニズムを用いて観測時間における各時刻の context vector $\boldsymbol{c}_t^i$と hidden vector $\boldsymbol{h}_{t'}^i$の関連性を評価する．また，attentive context vector $\hat{\boldsymbol{c}}_t^i$を計算し，推定対象者に対する周辺環境の影響を代表する．

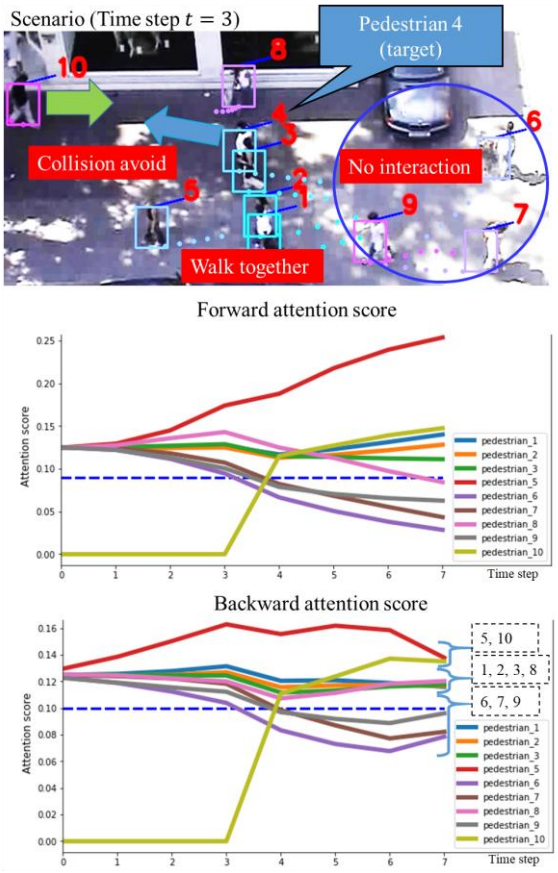


Fig. 3 Visualization of attention score.

$$\hat{\boldsymbol{c}}_t^i = \textstyle\sum_{t=1}^{12} d_{t',t}^i \boldsymbol{c}_t^i \quad (4)$$

$$d_{t',t}^i = \frac{\exp(\langle \boldsymbol{h}_{t'}^i, \boldsymbol{c}_t^i \rangle)}{\sum_{t=1}^{12} \exp(\langle \boldsymbol{h}_{t'}^i, \boldsymbol{c}_t^i \rangle)} \quad (5)$$

ここで，$d_{t',t}^i$は alignment vector であり，観測時間の各時刻における推定対象者に対する周辺環境からの影響の度合いを表す．最後に，attentive context vector $\hat{\boldsymbol{c}}_t^i$と decoder の hidden vector $\boldsymbol{h}_{t'}^i$が連結され，１つの全結合層により予測位置$(\hat{x}_{t'}^i, \hat{y}_{t'}^i)$に反映される．以上の操作を$t' = 1$から$t' = T_{pred} - 1$まで繰り返すと推定対象者の予測軌道$\{(\hat{x}_{t'}^i, \hat{y}_{t'}^i) | t' = 1, 2, \ldots, T_{pred}\}$を計算することができる．ただし，$t'$において予測された位置を$t' + 1$の RNN の入力とする．

3.　実験

提案したアテンションモデルの性能を評価するため，歩行者軌道の公開データセットを用いて人工ニューラルネットワークを構築し，性能評価を行った．

使用したデータセットには ETH, Hotel, Student, Zara の 4 つのシナリオがある[5][10]．合計 2027 名の歩行者の

軌道データは手動のラベリングにより動画から抽出した.

モデルの汎化能力を評価するため，全データセットのうちの 1 つのシナリオから抽出された軌道データをテストデータとして扱い，残りのデータで訓練とバリデーションを行い，その方法を 4 つのシナリオに対して繰り返した．訓練と評価では，観測された軌道として 8 time step の座標を入力し，12 time step の座標を予測軌道として出力した.

4. 実験結果と考察

性能評価では，Average Displacement Error（ADE）と Final Displacement Error（FDE）を指標として予測誤差を評価した.

$$\boldsymbol{ADE} = \frac{\sum_{t'=1}^{12}\sqrt{(x_{t'}-\hat{x}_{t'})^2+(y_{t'}-\hat{y}_{t'})^2}}{12} \quad (6)$$

$$\boldsymbol{FDE} = \sqrt{(x_{12}-\hat{x}_{12})^2+(y_{12}-\hat{y}_{12})^2} \quad (7)$$

ここでt'は予測における各時刻を表す．$(x_{t'}, y_{t'})$と$(\hat{x}_{t'}, \hat{y}_{t'})$は時刻$t'$における decoder の真値と予測結果である．提案手法と従来手法の評価結果との比較を表 1 に示す．全データセットにおいて提案手法の予測誤差はベースラインの Social-LSTM より小さかった．Hotel と Student の 2 つのシナリオに対しては，Social-GAN より誤差が小さかった．Zara における提案手法の性能は従来手法の予測誤差と同等となった．ETH における提案手法の予測誤差は大きくなったが，その原因としては地面の積雪が歩行者の動きに影響したと考えられる.

また，提案手法の性能を分析するため，歩行者のアテンションスコア$a_t^{i,j}$と$b_t^{i,j}$（3 章　参照）をプロットし，アテンションスコアの変化を可視化した上で，アテンションメカニズムの機能を分析した（図 3 参照）．Forward と backward におけるアテンションスコアの分布を比較すると，backward におけるアテンションメカニズムは異なるタイプのインタラクションを分ける効果がより顕著である．スコアの値の大小により，推定対象者周辺の他者が 3 つのグループに分けられた．推定対象者が前を歩く他者を避ける必要があるため，アテンションメカニズムが前を歩く他者に与えるスコアが高かった．一方，推定対象者の視野外にいる他者は推定対象の軌道に影響しないため，与えられたスコアが低かった（図 3 参照）．通常の RNN に backward を追加した BiRNN を使うことにより，モデルは異なるタイプのインタラクションを学習できると考えられる.

Table 1 Quantitative result of ADE and FDE (meter) on those approach (ADE/FDE)

Dataset	Social-LSTM[6]	Social-GAN[7]	Our model
ETH	1.09/2.35	**0.87/1.62**	1.26/2.32
Hotel	0.79/1.76	0.67/1.37	**0.59/1.31**
Student	0.67/1.40	0.76/1.52	**0.63/1.31**
Zara01	0.47/1.00	0.35/0.68	0.42/0.87
Zara02	0.56/1.17	0.42/0.84	(Merged)

5. おわりに

本研究では，歩行者の軌道予測問題に取り組み，歩行者のインタラクションを考慮できる BiRNN アテンションモデルを提案した．公開されている歩行者軌道データセットを用いた実験により，従来手法より高い予測精度が達成された．また，アテンションスコアの可視化により，提案したニューラルネットワークは歩行者間のインタラクションを学習できることを示した.

参 考 文 献

(1) Lewin K., Field theory in social science, *Harper & Brothers*, (1951)

(2) Helbing D., et al., “Social force model for pedestrian dynamics”, *Physical Review*, Vol.2, No.5(1995), pp.4282-4386.

(3) Tamura Y., et al., “Development of pedestrian behavior model taking account of intention”, *Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robot and Systems*, (2012), pp.382-387.

(4) Kitani K.M., et al., “Activity forecasting”, *Proceedings of the 2012 European Conference on Computer Vision*, (2012), pp.201-214.

(5) Pedllegrini S., et al., “You’ll never walk alone: Modeling social behavior for multi-target tracking”, *Proceedings of the 2009 IEEE International Conference on Computer Vision*, (2009), pp.261-268.

(6) Alahi A., et al., “Social LSTM: Human trajectory prediction in crowded spaces”, *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, (2016), pp. 961-971.

(7) Gupta A., et al., “Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks”, *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, (2018), pp. 2255–2264.

(8) Schuster M., et al.., “Bidirectional recurrent neural networks”, *IEEE Transactions on Signal Processing*, Vol. 45, No. 11(1997), pp.2673-2681.

(9) Vaswani A., et al., “Attention is all you need, *Advances in Neural Information Processing Systems*”, (2017), pp. 5998-6008.

(10) Lerner A., et al., “Crowds by example”, *Computer Graphics Forum*, Vol. 26, No. 3(2007), pp. 655-664.