Forecasting Human Fine-grained Behaviours

Simon Romanski
George Adaimi
Alexandre Alahi

Visual Intelligence for Transportation (VITA), EPFL

May 2018

STRC

18th Swiss Transport Research Conference
Monte Verità / Ascona, May 16 – 18, 2018

Visual Intelligence for Transportation (VITA), EPFL

Forecasting Human Fine-grained Behaviours

Simon Romanski, George Adaimi, Alexandre Alahi
Visual Intelligence for Transportation Lab (VITA)
Ecole Polytechnique Federale Lausanne (EPFL)
Route Cantonale, 1015 Lausanne, Schweiz
phone: +41-21-693 26 08
fax: +41-21-693 26 08
{simon.romanski, george.adaimi, alexandre.alahi}@epfl.ch

May 2018

Abstract

For self-driving cars and autonomous delivery platforms, one of the crucial steps to safe and seamless integration of these platforms is a human trajectory prediction module. While self-driving cars reach good performances in urban environments, crowded scenarios require a more accurate prediction of human-human and human-space interactions. Recent approaches perform the motion forecasting by using only coordinates and velocities of the pedestrians. Inherently, some things are impossible to predict with this representation, e.g., when a pedestrian starts walking or if people recognize and consequently walk towards each other. This work adds human pose information and human activity labels as features to allow a new way of forecasting pedestrian movements. For every human, a time series of bounding boxes, poses, and activities are used to train a Long-Short Term Memory (LSTM) network to predict a future time series of bounding box coordinates. Further experiments will be performed to analyze if predictions for activities and poses are feasible. The LSTM is trained and validated with annotated volleyball and basketball images. In the future, this work should be validated for a broad and general use in human trajectory forecasting.

Keywords

Trajectory Prediction, LSTM, Recurrent Neural Networks, Human Activity Forecasting

1 Introduction

Self-driving cars have great potential to avoid accidents and to make mobility accessible for demographic groups as children, elder people and disabled people who currently have to rely on public transport. Autonomous delivery platforms can potentially relieve the stress of carrying heavy luggage while guiding to a specific location simultaneously. For both, self-driving cars and autonomous delivery platforms, one of the crucial steps to safe and seamless integration of these platforms is a human trajectory prediction module.

Self-driving cars reach good performances in urban environments like in Palo Alto where the few pedestrians that cross the roads use crosswalks and traffic lights. The task is much more complicated in cities like Paris where sometimes no lane markings exist, and yet multiple lane roundabouts are used and at the same time a lot of tourist cross the street. Another complex scenario is a university campus, where pedestrians often do not follow any traffic rules and in practice cars have to be compliant with these unwritten rules. Regardless of these difficulties, it is still feasible for a human driver to foresee pedestrian actions in the given scenarios. Hence, a machine should have the same ability to develop intuition.
Consequently, crowded scenarios require a more accurate prediction of human-human and human-space interactions.

2 Related Work

Recent works have already attempted to predict future human actions:

Helbing and Molnar (1995) modeled human-human interactions based on a social-force model that uses attractive and repulsive forces. This approach has been adapted for robotics by Luber *et al.* (2010) and further been used in Leal-Taixe *et al.* (2014), Leal-Taixe *et al.* (2011) and Mehran *et al.* (2009) for scene understanding.

Similar models as developed by Treuille *et al.* (2006) use continuum dynamics to model human behavior, whereas Wang *et al.* (2008) and Tay and Laugier (2008) use Gaussian Processes for the human-human interactions. Antonini *et al.* (2004) predicts human motion behavior based on a discrete choice model. Yi *et al.* (2015) predicts motion with particular consideration of stationary groups. Yamaguchi *et al.* (2011) utilizes an agent-based behavioral model for prediction.

In addition to modeling interactions, a large set of works forecast human movement by clustering trajectories, e.g. Kim *et al.* (2011), Morris and Trivedi (2011) and Zhou *et al.* (2011).

In contrast to the approaches above, Kitani *et al.* (2012) uses optimal control theory to predict human interactions with their surrounding space.

Ziebart *et al.* (2011) predicts movement by a planning-based approach. Turek *et al.* (2010) developed a similar approach using a functional map of a scene.

Trautman *et al.* (2013) attempts to encourage humans to interact with autonomous robots rather than just predicting their movement.

Hochreiter and Schmidhuber (1997) developed the **Long Short-Term Memory (LSTM)** that has been shown to be useful for a variety of sequence predictions: Yunpeng *et al.* (2017), Althelaya *et al.* (2018), Tian and Pan (2015), Troiano *et al.* (2018), Xu *et al.* (2017), Vinayakumar *et al.* (2017).

Alahi *et al.* (2016) use LSTMs to predict individual motion behavior while pooling the information after each step.

All of these approaches represent a pedestrian by its coordinates and velocities. A great advantage of this representation is that the problem is transformed into a sequence prediction problem where priors on the dynamic can be used.

However, one task that is barely solved is, to estimate abrupt non-linear behaviors. Pedestrians suddenly slow down, stop, start to move again etc. In this work, we argue that we need to go beyond modeling pedestrians coordinates but integrate richer visual information such as their poses or the perceived actions in the forecasting framework.

3 Use Richer Visual features

Some behaviors are inherently impossible to predict given just pedestrians coordinates. Richer visual information is needed. For instance, gestures like waving at each other, i.e., *human poses*, might influence the dynamics of people. Additionally, cues like where pedestrians are looking at, i.e., *human attention*, or what they are doing, i.e., *human actions*, could help better predict their behavior.

Consequently, the goal of this project is to study which additional information is relevant to better predict human behavior and how to effectively model it. We will study multiple neural network architectures that learn to predict human behavior given rich visual information.

In this work, we focus on studying how human actions can improve motion predictability. The activity labels are extracted from the dataset created by (Bagautdinov *et al.* (2017)). Given video data, a time series of bounding boxes and activities are extracted for every humans as input (observed features) to train a Long Short-Term Memory (LSTM) network to predict the future coordinates. We still frame our problem as a time series prediction problem. As shown by Alahi *et al.* (2016) LSTM networks are capable of learning human motion behavior and therefore suitable for the given prediction problem.

4 Dataset

One of the challenges of this project is to identify labeled data that can be used for the training and testing of the proposed algorithms. More precisely, labels on human actions are needed in addition to human coordinates. To the best of our knowledge, Bagautdinov *et al.* (2017) is the only dataset that track humans in space and have annotated their actions as well as their collective activities.

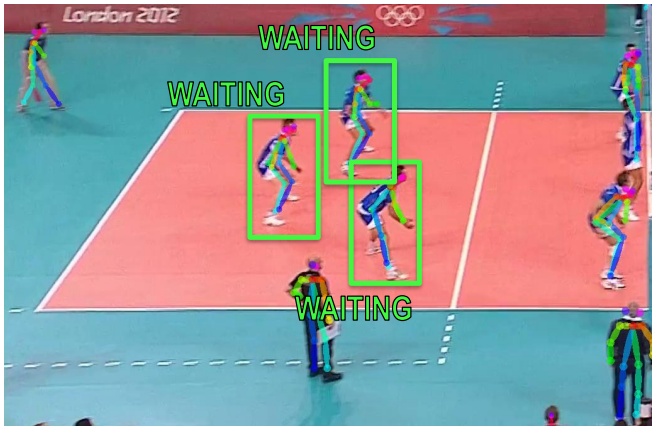
1 shows part of a sample frame of the dataset. The bounding box and action labels are annotated based on the given ground truth information.

The nine different possible labels are: 'blocking', 'digging', 'falling', 'jumping', 'moving', 'setting', 'spiking', 'standing' and 'waiting'.

Furthermore, each frame has been processed by OpenPose a framework developed by Cao *et al.* (2017), Simon *et al.* (2017) and Wei *et al.* (2016) which is capable of extracting human poses from image data.

Although the pose provides more accurate information, it will not be considered in the following. Concepts using only one human action, e.g. generated by a classifier as presented in Bagautdinov *et al.* (2017), have the potential be more efficient as they only add one state variable to the coordinates rather than one variable for each joint in the pose.

Figure 1: Example Frame from Dataset processed with OpenPose



Source: Based on ground truth and data set by Bagautdinov *et al.* (2017) processed with OpenPose by Cao *et al.* (2017)

Although the context is a volleyball game, we can still study our claims and compare our algorithms given ground truth data. Our future work includes collecting a dataset in an urban setting.

5 Potential neural network architectures

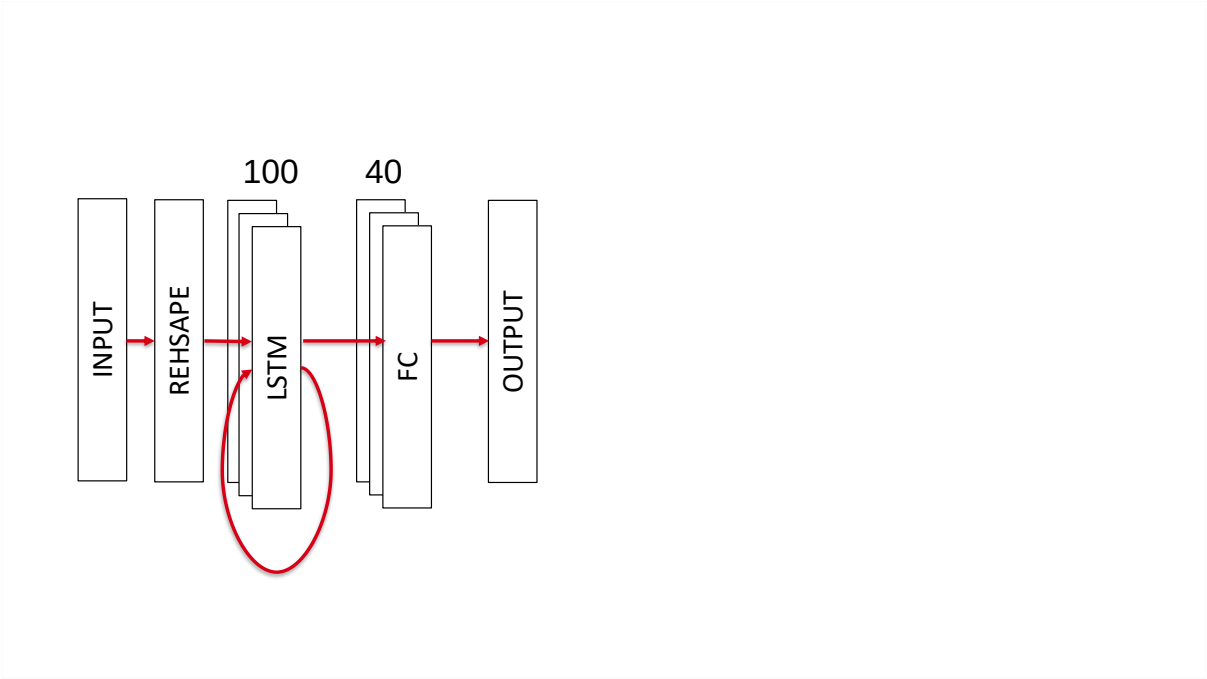
This report presents multiple recurrent neural network architectures based on the Long Short-Term Memory (LSTM) architecture by Hochreiter and Schmidhuber (1997) to handle the contextual information that is required for the desired approach. All architectures use spatial information and activity labels based on the recorded video sequences. The sequences are sliced into two parts - one as LSTM input and the other one as ground truth for the training of the LSTM - consisting of 20 frames each. The scenes are further divided into scenes that are used for training and scenes that are used for validation of the prediction.

We want to predict the position of the player for the next 20 frames based on the informa-

tion of the last 20 frames.

The proposed architectures share a common overall structure: The input is reshaped and then fed into a layer of LSTMs which is followed by a fully connected layer. This architecture is shown in 2

Figure 2: Basic LSTM Architecture

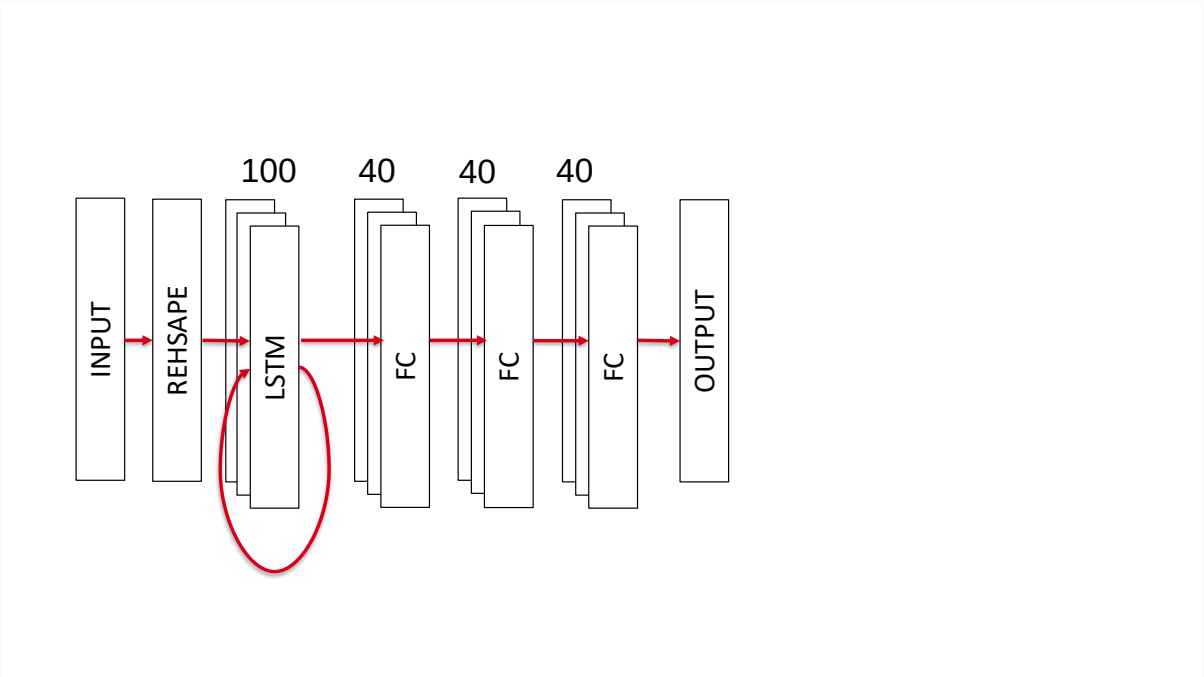


One LSTM can trained for each player in the recorded sequence based on spatial information. As we use x and y coordinates as well as one variable describing the action of the player, we have three input variables per frame and sample.

While in this architecture each LSTM has the ability of learning individual human behavior it will most likely not learn the human-human interactions. In Bagautdinov *et al.* (2017) a **joint LSTM** is used for action classification. A joint LSTM for prediction could reason based on the relative spatial positions and the distance between players. Instead of training the LSTM individually for each player, the network has potential to learn the entire positioning of the players on the field.

In comparison to the individual LSTM approach, it requires more data for training as the network has more connections and additionally every frame can only be used once for the entire team in contrast to the individual LSTM where each frame can be used once for each player.

Figure 3: Deep LSTM Architecture



As a variation of the basic architecture, we can attempt to make the architecture deeper in order to recombine the learned representation from prior layers and create new representations at high levels of abstraction as shown in 3

A different kind of architecture with a similar intention are stacked LSTMs as shown in 4. Stacked LSTMs have been used by Graves *et al.* (2013) and achieved outstanding results for the challenging standard problem of speech recognition. Stacked LSTMs add levels of abstraction of input observations *over time*.

To be able to learn abstract behavior as well as patterns over time, we finally combine the last two approaches which results in a stacked LSTM with multiple fully connected layers. This architecture is shown in 5

All architectures are implemented by using Tensorflow (Abadi *et al.* (2015)) and Keras (Chollet *et al.* (2015)) for Python.

Figure 4: Stacked LSTM Architecture

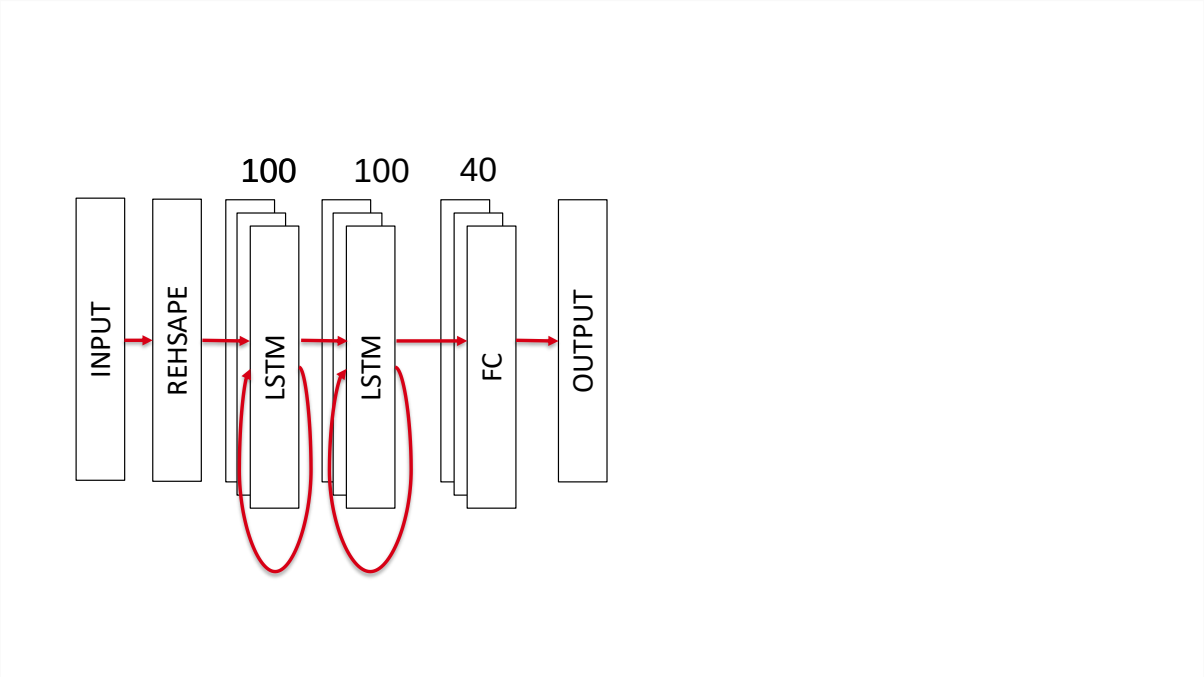
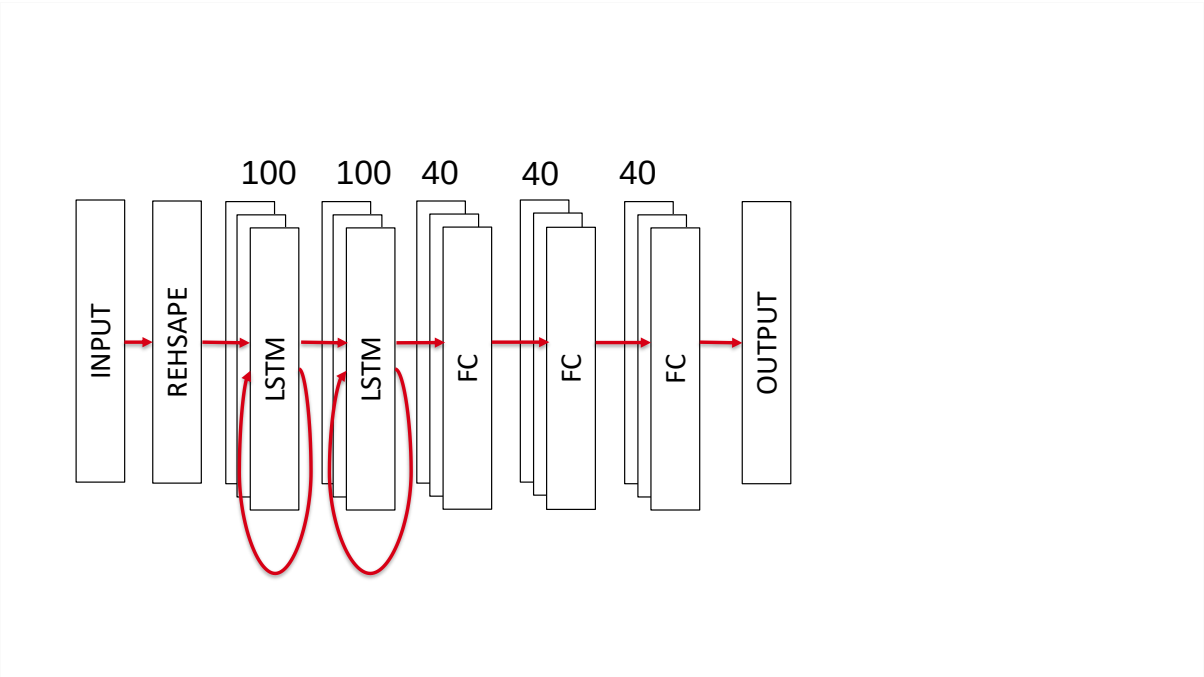


Figure 5: Deep Stacked LSTM Architecture



6 Preliminary Results

The performance is tested for each presented architecture on the volleyball dataset (Bagautdinov *et al.* (2017)). 1 compares the Average Displacement Error in pixels. The MSE loss calculated for the entire sequence by Keras is provided for reference.

To provide insight if rich visual features are beneficial for prediction of human motion the individual LSTM is trained without visual features for comparison. One can obtain that using visual features the average displacement error is already reduced by approximately 21%.

The focus of this work is on the choice which architecture is most beneficial for prediction using the given features. Therefore the basic individual LSTM is compared with basic joint LSTM. The joint LSTM is outperformed by the individual LSTM by almost 53 %. This is very likely a consequence of the size of the dataset.

The deeper architectures beat the basic LSTM architectures by more than 63%.

The stacked LSTM with additional fully connected layers reaches an average displacement error of 15.88 pixels. Given that predictions are performed in the image coordinates, one can only approximate the accuracy in meters.

Table 1: Quantitative Results on the Volleyball Dataset (Bagautdinov *et al.* (2017))

Architecture	**MSE Loss**	**Average Displacement Error (in pixels)**
Stacked Deep LSTM	716.10	15.88
Stacked LSTM	761.02	16.23
Deep LSTM	1239.77	21.59
LSTM (no visual features)	13998.62	74.10
LSTM	6997.50	58.60
Joint LSTM	27620.14	101.79

7 Future Work

The following approaches could be used for further evaluation:

If a fixed amount of frames is used as an input time series and a fixed amount of frames is desired as an output, one could even use a non-recurrent neural network. The LSTM layer is replaced by another fully connected layer and the structure results in a Feedforward Neural Network.

Furthermore, the performance could be compared with the prediction of a Linear Kalman Filter using spatial coordinates.

Poses generated from OpenPose can serve as additional rich visual feature in the future and might improve accuracy.

Future work includes the collection of a data set in urban environments for the task of more generic human activity forecasting.

8 Conclusion

This work reasons why spatial positions are not a sufficient representation for human trajectory prediction. For certain scenarios additional rich visual features are obligatory. A variety of neural network architectures are proposed and their performance is evaluated.

The use of additional visual features is helpful for the improvement of human-human interaction modeling and consequently also beneficial for the success of human activity forecasting.

This evaluation based on the given volleyball dataset aims to be a proof of concept that rich visual features increase the accuracy of prediction.

Although motion prediction in sports might be easier due to its clear rules and strategies, previous work (Alahi *et al.* (2016), Yamaguchi *et al.* (2011)) has shown that unwritten rules and movement strategies in traffic seem to exist and can be learned by LSTM networks. Therefore, it is reasonable to assume that LSTM structures can learn this behavior using the additional visual information about activity as it can perform such predictions in sports.

Pedestrian movement forecasting is one of the crucial modules for self-driving cars and autonomous delivery platforms. The possibility to handle tasks as predicting the start of a movement provides a new way of understanding traffic scenes and allows new means of acting in scenarios for autonomous vehicles.

9 References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng (2015) TensorFlow: Large-scale machine learning on heterogeneous systems, `https://www.tensorflow.org/`. Software available from tensorflow.org.

Alahi, A., K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei and S. Savarese (2016) Social lstm: Human trajectory prediction in crowded spaces, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Althelaya, K. A., E. S. M. El-Alfy and S. Mohammed (2018) Evaluation of bidirectional lstm for short-and long-term stock market prediction, paper presented at the *2018 9th International Conference on Information and Communication Systems (ICICS)*, 151–156, April 2018.

Antonini, G., M. Bierlaire and M. Weber (2004) Discrete choice models of pedestrian behavior.

Bagautdinov, T., A. Alahi, F. Fleuret, P. Fua and S. Savarese (2017) Social scene understanding: End-to-end multi-person action localization and collective activity recognition, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cao, Z., T. Simon, S.-E. Wei and Y. Sheikh (2017) Realtime multi-person 2d pose estimation using part affinity fields, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chollet, F. *et al.* (2015) Keras, `https://keras.io`.

Graves, A., A. Mohamed and G. E. Hinton (2013) Speech recognition with deep recurrent neural networks, *CoRR*.

Helbing, D. and P. Molnar (1995) Social force model for pedestrian dynamics, *Physical review E*.

Hochreiter, S. and J. Schmidhuber (1997) Long short-term memory, *Neural computation*.

Kim, K., D. Lee and I. Essa (2011) Gaussian process regression flow for analysis of motion trajectories, *IEEE International Conference on Computer Vision (ICCV)*.

Kitani, K., B. Ziebart, J. Bagnell and M. Hebert (2012) Activity forecasting, *ECCV*.

Leal-Taixe, L., M. Fenzi, A. Kuznetsova, B. Rosenhahn and S. Savarese (2014) Learning an image-based motion context for multiple people tracking, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Leal-Taixe, L., G. Pons-Moll and B. Rosenhahn (2011) Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker, *ICCV Workshops*.

Luber, M., J. A. Stork, G. D. Tipaldi and K. O. Arras (2010) People tracking with human motion predictions from social forces, *Robotics and Automation (ICRA)*.

Mehran, R., A. Oyama and M. Shah (2009) Abnormal crowd behavior detection using social force model, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Morris, B. T. and M. M. Trivedi (2011) Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach, *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Simon, T., H. Joo, I. Matthews and Y. Sheikh (2017) Hand keypoint detection in single images using multiview bootstrapping, paper presented at the *CVPR*.

Tay, M. K. C. and C. Laugier (2008) Modelling smooth paths using gaussian processes, *Field and Service Robotics*.

Tian, Y. and L. Pan (2015) Predicting short-term traffic flow by long short-term memory recurrent neural network, paper presented at the *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 153–158, Dec 2015.

Trautman, P., J. Ma, R. Murray and A. Krause (2013) Robot navigation in dense human crowds: the case for cooperation, *Robotics and Automation (ICRA)*.

Treuille, A., S. Cooper and Z. Popovic (2006) Continuum crowds, *ACM Transactions on Graphics (TOG) , volume 25*.

Troiano, L., E. M. Villa and V. Loia (2018) Replicating a trading strategy by means of lstm for financial industry applications, *IEEE Transactions on Industrial Informatics*, 1–1, ISSN 1551-3203.

Turek, M. W., A. Hoogs and R. Collins (2010) Unsupervised learning of functional categories in video scenes, *ECCV*.

Vinayakumar, R., K. P. Soman and P. Poornachandran (2017) Applying deep learning approaches for network traffic prediction, paper presented at the *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2353–2358, Sept 2017.

Wang, J. M., D. J. Fleet and A. Hertzmann (2008) Gaussian process dynamical models for human motion, *Pattern Analysis and Machine Intelligence*.

Wei, S.-E., V. Ramakrishna, T. Kanade and Y. Sheikh (2016) Convolutional pose machines, paper presented at the *CVPR*.

Xu, J., R. Rahmatizadeh, L. Boeloeni and D. Turgut (2017) A sequence learning model with recurrent neural networks for taxi demand prediction, paper presented at the *2017 IEEE 42nd Conference on Local Computer Networks (LCN)*, 261–268, Oct 2017, ISSN 0742-1303.

Yamaguchi, K., A. Berg, L. Ortiz and T. Berg (2011) Who are you with and where are you going?, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yi, S., H. Li and X. Wang (2015) Understanding pedestrian behaviors from stationary crowd groups, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yunpeng, L., H. Di, B. Junpeng and Q. Yong (2017) Multi-step ahead time series forecasting for different data patterns based on lstm recurrent neural network, paper presented at the *2017 14th Web Information Systems and Applications Conference (WISA)*, 305–310, Nov 2017.

Zhou, B., X. Wang and X. Tang (2011) Random field topic model for semantic region analysis in crowded scenes from tracklets, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ziebart, B. D., N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey and S. Srinivasa (2011) Planning-based prediction for pedestrian, *Intelligent Robots and System*.