Real-Time Certified Probabilistic Pedestrian Forecasting
Author's Response

Henry O. Jacobs, Owen Hughes,
Matt Johnson-Roberson, and Ram Vasudevan

May 7, 2017

We would like to begin by thanking the reviewers and the editor for their careful reading and review of our paper. The comments provided were incredibly helpful and insightful, and have strengthened the paper considerably. Based on the recommendations of the reviewers and the editor, we have made several modifications to the paper. The changes made are summarized in detail below and a highlighted manuscript is attached that indicates additions in red. **This letter is largely self contained, so references here are to the bibliography in this file**.

Reviewer 1

1. **This paper presents a novel real-time probabilistic forecasting method for pedestrian trajectories via observing the historical trajectories in a particular scene. The problem is well motivated, formulated and handled. The reported experimental evaluations show pivotal improvement against the current state-of-the-art.**

   Thank you for the positive assessment.

2. **With the current form of the results it is hard to compare the improvements in terms of accuracy gained. This can be overcome by including results such as physical distance between predicted paths and ground truth observations. Such matrices are provided in the baseline models [1] [2] [3] with accuracy values in meters.**

   Thank you for noticing the difficulty in comparing the improvements. We addressed this by adding a comparison to the MHD from . Specifically, we calculated the MHD from each point in the trajectory to sampled points from each distribution. The results aresummarized in Figure 1 below.

   We feel that both plots add something to the analysis, so we've included the following passage to justify both our metrics.

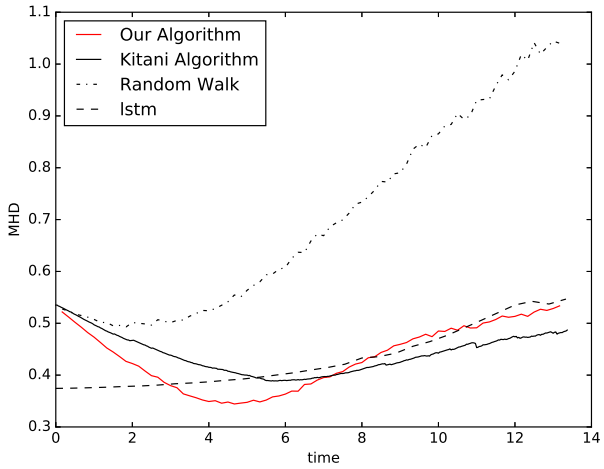

Figure 1: A comparison of the Modified Hausdorff Distance from the ground truth of the pedestrian to a 1000 point sample from each distribution. The method from [4] does very well at short time scales due to its confidence, but we outperform all other methods at intermediate times. At long timescales the MHD to the trajectory of most algorithms converges. Our method increases positional uncertainty with time, and so [1] outperforms us because they know the end point and we do not, and [4] places too much significance on the initial positions of other pedestrians at large time scales.

In our analysis we sought a metric that measured the similarity between the predictor and the ground truth as well as the "safety" of the prediction. We used the area under the curve of the ROC curve as our measure of this. The ROC curve plots the rate of false positive detections against the probability of detecting a pedestrian, or in essence the quality of the detection versus the safety. The area under this curve is a standard measure of the quality of a predictor. Figure 2 shows the analysis of the AUC of each algorithm versus time. In addition, we used the Modified Hausdorff Distance (MHD) from the ground truth trajectory to a sample from the predictions at each time in order to provide a geometric measure of how accurate the predictions are. Figure 1 shows MHD plotted against time.

3. **How the model can incorporate the interactions among pedestrians (i.e group motion) and the influences from neighbouring pedestrians (for instances such as collision avoidance)?** Thank you for this insightful question. Social forces are a natural avenue for future work, and so we included the following paragraph in the paper.

   Though including social forces into the model is out of the scope of this paper, its incorporation could begin as follows. In the current motion model with a single vector field, the acceleration of the ith agent is given by $\ddot{x}_i = s^2 DX(x_i) \cdot X(x_i)$. Incorporating a social force F_i acting on the ith agent can be done by instead asserting $\ddot{x}_i = s^2 DX(x) \cdot X(x) + F_i$. Usually $F_i = \sum_j \nabla U(x_j - x_i)$ where U is an interaction potential [5] .

4. **Presentation and organisation The motivation behind choosing start/end points based clustering is not clear. Does clustering based on start/end points help to**

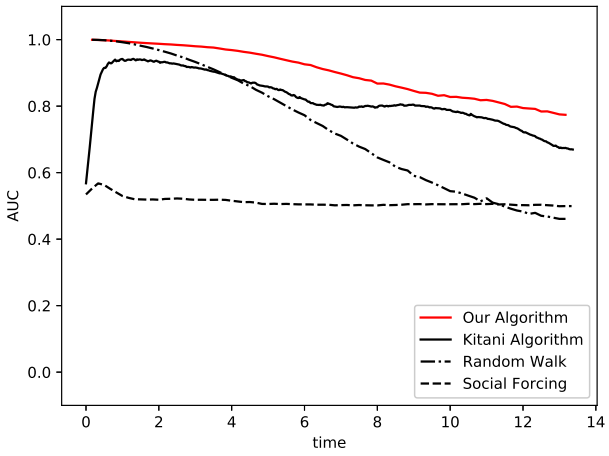

Figure 2: A comparison of the AUC of the various algorithms. Note that the initial dip in the performance of [1] is due to their confidence in their initial estimate. We sampled the S-LSTM [4] model 100 times in order to give them the best chance in this analysis, but their confidence combined with the over-reliance on social forces at moderate-to-large timescales lowered their performance.

identify different motion models? Isn't it better to cluster considering the entire trajectory? This is a relevant question, thank you. As stated in Morris and Trivedi (below), our choice of clustering technique is not nearly as important as the chosen distance metric. Our success with distance functions that took a whole trajectory was limited on the data we tested, particularly when compared against our method. We have several hypotheses as to why this is true. For one, the spatial scale of the data we test against showcases important features (e.g. curb cuts and storefronts), which lead to clustered start points and end points. Similarly, we tested against high quality data in which trajectories with irreconcilable occlusions were rare and thus could be omitted. See the comment below for the text we added to the paper.

5. **Do you cluster different scenes provided in campus dataset [4] (i.e Valley, Fangron) separately or all together? When motion models are learnt, are they learnt in scene specific manner or altogether?** Thank you for noticing that this is unclear. Our algorithm trains on each scene independently, so clustering, vector field learning, and learning the potential functions are all done for each Valley, Fangron, etc. We've added the following text to the paper:

   > We did a 2-fold cross validation by using 20% of the data for testing and the remainder for training within each scene. We learned separate collections of vector fields and model parameters for each fold on all of the four scenes on which we tested.

6. **The motivation behind choosing the affinity propagation clustering algorithm is not clearly stated. What is the distance measure used (i.e Euclidian, Ö) ? Not specifying such information makes reproduction of this work not feasible.**

   This was an oversight on our part, thank you for noticing it. The distance used is actually a custom distance function on $\mathbb{R}^4$, which disregards the involution $(x_1, x_2, x_3, x_4) \to (x_3, x_4, x_1, x_2)$,

and so doesn't care about parity of trajectories. We have included the following clarification:

> We then cluster in $\mathbb{R}^4$ using Affinity Propagation [6] and a custom distance measure defined by $d((x_1, x_2, x_3, x_4), \mathbf{y}) : \mathbb{R}^4 \times \mathbb{R}^4 \to \mathbb{R} = \min\{d_e((x_1, x_2, x_3, x_4), \mathbf{y}), d_e((x_3, x_4, x_1, x_2), \mathbf{y})\}$, for the euclidean distance d_e. This function measures the distance between the endpoints irrespective of their ordering. The scale of the datasets we tested on had large enough spatial scale that clustering based on endpoints captured people moving from destination to destination, e.g. from a storefront to the sidewalk at the edge of a scene. On our data, other distance measures from [7] and [8] didn't identify coherent motion models. This is partly due to the fact that oftentimes pedestrians would take slightly different routes to get to their destination, and the cumulative effect on the distance measures is enough to hinder the clustering.

7. **Adequacy of Citation I believe literature review section can be improved using a sub section on social force models, which is extensively applied for pedestrian trajectory forecasting. Helbing and Molnar, 1995; Koppula and Saxena, 2013; Pellegrini et al., 2010; Yamaguchi et al., 2011; Xu et al., 2012; Wang et al. 2008; Hospedales et al. 2009; Emonet et al. 2011; Yi et al. 2015.**

   We agree that this adds to the paper, thank you for the suggestion. We added a section to our literature review to reflect the relevant work.

   > On the other hand, many authors have approached pedestrian forecasting by deriving their motion model from interactions between pedestrians. Early work by [5] and [9] describe pedestrians interactions using physically motivated methods. Several models such as [10], [11], and [12] derive their motion models from [5] who incorporate collision avoidance through an interaction potential which repels pedestrians from each other, leading to forms of what is called Linear Trajectory Avoidance (LTA). However LTA suffers from not planning for other pedestrians positions at future times. [13], [14] , and [15] all take optical flow as input. [15] and [13] both use variants of Hierarchical Dirichlet Processes on discretized optical flow to determine temporal motifs (i.e. classes of motion within the scene), where [14] substitutes a Markov model. These models are not agent based, and the lack of an explicit motion model limits their predictive power in situations where we know more about the data. [16] predict trajectories by introducing and sampling Anticipatory Temporal Conditional Random Fields which incorporate learned affordances for humans and objects based on observed objectives within the scene. [17] , [18], and [19] create agent-based models based on Gaussian Processes, though they suffer issues when trained on discretized trajectories. [4] uses Long Short-Term Memory to learn pedestrian motion models without making implicit (as in the case of CRFs) or explicit (in the case of social forces) assumptions about the manner in which agents will interact. [4] outperforms LTA, social forcing based on [10], and IGP from [19]. Along with its quick run time, this establishes [4] as state-of-the-art for socially based models, and so we our algorithm to theirs.

8. **Furthermore a comprehensive review on trajectory clustering algorithms is required. This should be used to motivate the reason of choosing affinity propagation algorithm in section 4. Morris and Trivedi 2009; Giannotti et al. 2007; Lee et al. 2007; Ester et al. 1996**

   Thank you for this suggestion and for the list of relevant titles. We included a note about different algorithms as well as a justification for why we chose our metric. We were unable to test against the methods of Giannotti and Ester. The note can be seen in point 6 above.

Reviewer 2

1. **Keep the good work going. Interesting approach.**

   Thank you for the kind comments.

2. **Need to be clearer and dive a little bit deeper into the technical approach.** Thank you for this criticism. We hope we adequately addressed it by strengthening the description of our method for learning the motion model, as well as quantifying the way that we tested our data. We also changed the language below to better reflect that the specific implementation we chose is an example of how to use the larger framework.

   > Given the model established in the previous section, we describe an implementation to showcase one way the model can be applied to observational data.

3. **Need to use and apply more deep learning approaches such as social LSTM or other machine methods.** Thank you for the suggestion. We hope that we have addressed it by including a review of literature of social motivated models, including S-LSTM, as well as a section on how social factors can be integrated into our model. In addition, we compared against S-LSTM in addition to [1] from the paper. The figures corresponding to the predictions of the S-LSTM model are ommitted from the paper because as happened for many of the agents, the S-LSTM predictions quickly exit the bounds of the scene, and so the visualization is identically zero. We changed the subtitle of the figure to reflect this, which is shown below.

   > An illustration of the predictions generated by the various algorithms. In this figure, the dot is the start point of the test trajectory, the diamond is the position at time t, and the X is the end of the trajectory. The likelihood of detection is depicted using the virdis color palette. Notice that the Random Walk is imprecise, while the predictions generated by the algorithm in [1] suffer from the inability of their motion model to adequately match the speed of the agent. The algorithm from [4] quickly passes beyond the boundaries of the scene, so the plots are ommitted here.

   > An illustration of the predictions generated by the various algorithms. In this figure, the dot is the start point of the test trajectory, the diamond is the position at time t, and the X is the end of the trajectory. The likelihood of detection is depicted using the virdis color palette. Notice that the Random Walk is imprecise, while the predictions generated by the algorithm in [1] are unable to match the speed of the agent and choose the wrong direction to follow the agent around the circle. The algorithm in [4] passes beyond the boundaries of the scene, and so the plots are ommitted here.

Reviewer 5

1. **It would be more interesting if the authors show whether the prediction algorithm would transfer the learned "knowledge" of forecasting in novel scenes.** Thank you for your comment. As it is, this is a shortcoming of our model, but we see it as an immediate avenue for future work. As can be read below, the environmental decomposition that the segmentation algorithm in [20] returns as well as an analogue of the methods [21] proposes for characterizing and transferring vector fields from one scene to another could reasonably be applied to our method..

We also hope to do learning transfer with this model using scene segmentation from [20], as well as the semantic context descriptors and routing scores from [21] to show how vector fields can be transferred to novel scenes. It appears that the low-order parameterization of our model, and unit-length vector field assumption make it particularly amenable to the methods developed in [21].

2. **More statistically result is expected rather than analyzing "four different scenes", and comparing run-time by "averaged across several agents and scenes" (how many?).**

   Thank you for noticing this. We have clarified the paper in the passages listed below to clarify the list of scenes we ran in comparison, and how many agents total were run. We clarified that the set of agents used to evaluate were the same that were used to determine run times. Note that we attempted to compare on all of the scenes in the Stanford Drone Dataset, but the difficulty in getting the other models to converge in a reasonable amount of time prevented us from doing so.

   Our analysis was conducted on the Coupa, Bookstore, Death Circle, and Gates scenes from the dataset from [3], with a total of 142 trajectories analyzed.

   The run time per frame for each algorithm was generated using the mean run time for 400 frames, averaged across all of the trajectory data used in the quality analysis.

3. **Besides, it is better to explain the meaning of ROC and AUC in section IV.**

   This is a necessary clarification, thank you for the suggestion. We have included a description of ROC and AUC, as well as a justification for why we chose to evaluate the methods using this criterion. You can read the text that we added to address this in the second point made by Reviewer 1.

References

[1] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, *Activity Forecasting*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 201–214.

[2] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto, "Intent-aware long-term prediction of pedestrian motion," *Proceedings of the International Conference on Robotics and Automation (ICRA)*, May 2016.

[3] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, *Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes*. Cham: Springer International Publishing, 2016, pp. 549–565. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46484-8_33

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

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

[6] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[7] B. Morris and M. Trivedi, "Learning trajectory patterns by clustering: Experimental studies and comparative evaluation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 312–319.

[8] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data.* ACM, 2007, pp. 593–604.

[9] Y. Xu and H.-J. Huang, "Simulation of exit choosing in pedestrian evacuation with consideration of the direction visual field," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 4, pp. 991–1000, 2012.

[10] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* IEEE, 2011, pp. 1345–1352.

[11] S. Yi, H. Li, and X. Wang, "Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance," *IEEE transactions on image processing*, vol. 25, no. 9, pp. 4354–4368, 2016.

[12] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Computer Vision, 2009 IEEE 12th International Conference on.* IEEE, 2009, pp. 261–268.

[13] R. Emonet, J. Varadarajan, and J.-M. Odobez, "Extracting and locating temporal motifs in video scenes using a hierarchical non parametric bayesian model," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* IEEE, 2011, pp. 3233–3240.

[14] T. Hospedales, S. Gong, and T. Xiang, "A markov clustering topic model for mining behaviour in video," in *Computer Vision, 2009 IEEE 12th International Conference on.* IEEE, 2009, pp. 1165–1172.

[15] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 31, no. 3, pp. 539–555, 2009.

[16] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2016.

[17] M. Tay and C. Laugier, "Modelling smooth paths using gaussian processes," in *Field and Service Robotics.* Springer, 2008, pp. 381–390.

[18] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 283–298, 2008.

[19] P. Trautman, J. Ma, R. M. Murray, and A. Krause, "Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 335–356, 2015.

[20] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," in *Computer Vision and Pattern Recognition*, 2014.

[21] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese, "Knowledge transfer for scene-specific motion prediction," in *Proc. of European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, October 2016. [Online]. Available: http://arxiv.org/abs/1603.06987