CVPR
Computer Vision and Pattern Recognition 2018
Jun 19, 2018 - Jun 21, 2018, Salt Lake City, USA

Reviews For Paper

Paper ID 3790
Title DeepAutoTrack (DAT): Vehicle Trajectory Prediction for Autonomous Driving

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
[Paper Summary] What is the paper about? Please, be concise (3 to 5 sentences)	The authors describe a system to track vehicles in synthetic images. First, they create a dataset of 200,000 continuous frames from Synthia. They then implement existing approaches for detecting, tracking, and segmenting vehicles within the dataset. Those detections are used as input to LSTMs which predict vehicle trajectories.
[Paper Strengths] Please discuss, justifying your comments with the appropriate level of details, the strengths of the paper (i.e. novelty, theoretical approach and/or technical correctness, adequate evaluation, clarity, etc). For instance, a theoretical paper may need no experiments, while a paper with a new approach may require comparisons to existing methods.	Better tracking and prediction approaches would be very useful for the automated driving community.
[Paper Weaknesses] Please discuss, justifying your comments with the appropriate level of details, the weaknesses of the paper (i.e. lack of novelty – given references to prior work-, lack of novelty, technical errors, or/and insufficient evaluation, etc). Note: If you think there is an error in the paper, please explain why it is an error. Also remember that theoretical results/ideas are essential to CVPR (some theoretical papers may not need to have experiments). If the theory is novel and interesting, but the results did not outperform other existing algorithms, it is not necessarily a reason to reject. It is not appropriate to ask for comparisons with unpublished papers and papers published after the CVPR deadline. In all cases, please be polite and constructive. CVPR 2018 policy on dual submission and arxiv appears at: http://cvpr2018.thecvf.com/submission/main_conference/author_guidelines.	The main contributions of the paper, according to the authors, are: 1) The verification of the feasibility of prediction vehicles' future trajectories 2) Novel "tracking-by-detection" framework for robust tracking 3) Design various temporal models for the problem of predicting 4) Demonstrate the effectiveness of "privileged information" (e.g. semantic segmentation) 5) Construct a time-series dataset for vehicle trajectory prediction and formalize the problem into a 3D occupancy grid There are no mentions of prior work for any of those. The related work section focuses on object detection with 7 references to object detection methods, 9 for instance association (which they use for tracking), 4 references to other LSTM usages, and 4 for automated driving datasets. There is no mention of previous frame prediction methods. E.g.: - Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. CoRR, 2015. - Rakesh Chalasani and Jose C. Principe. Deep predictive coding networks. CoRR, 2013. - Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 2013. - Tobias Egner, Jim M. Monti, and Christopher Summerfield. Expectation and surprise determine neural population responses in the ventral visual stream. J Neurosci, 2010. - Chelsea Finn, Ian J. Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. CoRR, 2016. 1) Unlike stated in the paper, vehicle tracking in images has been studied for years. There are summary papers (e.g. VEHICLE DETECTION AND TRACKING TECHNIQUES: A CONCISE REVIEW Raad Ahmed Hadi1, 2 , Ghazali Sulong1 and Loay Edwar George, SIPIJ). There is no mention of prior art. 2) They adopt fDSST[6] and use it to track detections. What is the novelty, except for the implementation with specific parameters? 3) 3 very similar LSTMs with different input 4) The one LSTM with more information performs better. That is not surprising and not a real insight. There are a number of inaccuracies, e.g. line 044, where the authors state that automated driving models are either neural networks or controllers with if-then-else rules, or line 304, that states that the closest related work to their tracking is an ego-motion regression by Xu, Gao and Darrell. Line 471 calls Synthia images photo-realistic. The actual contributions, and evaluation could be clearer. It took some time to understand the different metrics for detection and tracking as they are supposed to be fused. It is not clear if the dataset is released with the paper. There is no comparison to existing methods, the authors mention that this is due to the lack of existing datasets.
[Preliminary Rating] Please rate the paper according to one of the following six choices:	Strong Reject
[Preliminary Evaluation] Please indicate to the AC, your fellow reviewers, and the authors your current opinion on the paper. Please tell the ACs what points you think have the most weight in your reviews and summary, and why.	The actual contribution is not significant enough and could be described more clearly.
[Confidence]	Confident - to stress that you are mostly sure about your conclusions (e.g., you are not an expert but can distinguish good work from bad work in that area).

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
[Paper Summary] What is the paper about? Please, be concise (3 to 5 sentences)	Given a sequence of RGB frames, the authors predict future trajectories of vehicles. They build a pipeline that combines object detection, instance association (tracking), and an RNN. They use an off-the-shelf modern convolutional object detector in order to spot traffic participants (i.e., vehicles) and track the targets of interest with the published fDSST tracker. Then, they use a SEG-LSTM recurrent network to fuse the multiple-streams from past frames and to predict targets' future trajectories.
[Paper Strengths] Please discuss, justifying your comments with the appropriate level of details, the strengths of the paper (i.e. novelty, theoretical approach and/or technical correctness, adequate evaluation, clarity, etc). For instance, a theoretical paper may need no experiments, while a paper with a new approach may require comparisons to existing methods.	I'm glad to see the authors compare against a Kalman Filter baseline. I think it's good for the deep learning community to remember Kalman Filters and always ensure that if we are using a complex convnet that we are actually achieving superior performance. The prediction task is not well-studied and I appreciate the authors' willingness to approach a challenging problem.
[Paper Weaknesses] Please discuss, justifying your comments with the appropriate level of details, the weaknesses of the paper (i.e. lack of novelty – given references to prior work-, lack of novelty, technical errors, or/and insufficient evaluation, etc). Note: If you think there is an error in the paper, please explain why it is an error. Also remember that theoretical results/ideas are essential to CVPR (some theoretical papers may not need to have experiments). If the theory is novel and interesting, but the results did not outperform other existing algorithms, it is not necessarily a reason to reject. It is not appropriate to ask for comparisons with unpublished papers and papers published after the CVPR deadline. In all cases, please be polite and constructive. CVPR 2018 policy on dual submission and arxiv appears at: http://cvpr2018.thecvf.com/submission/main_conference/author_guidelines.	Who else has experimented on the SYNTHIA dataset? Or are the authors the first ones? Are the authors releasing a specific publicly-available prediction challenge based on SYNTHIA? Where are comparisons with other activity forecasting/trajectory prediction works? I think the authors should reimplement SocialLSTM for 2D data (Stanford UAV dataset seems to be a 2D standard prediction task -- just throw away images and just use the xy plane coordinates, e.g. center of detected bounding box). SocialLSTM used "Avg. disp. Error", "Avg. non-linear disp. error" , and "final disp. Error" What was the justification for the metrics the authors chose? In the SocialLSTM paper's Related Work section, there are about 20 sources that focus on "Activity forecasting" The details about the fusion of "location" and "semantic" information into the LSTM hidden state seemed missing (was a bit vague). A figure of the architecture of the SEG-LSTM would go a long way in clarifying this (or are the authors using Xu et al.'s "End-to-end Learning of Driving Models from Large-scale Video Datasets" FCN-LSTM from Trevor Darrell's group?). How is this more than just an engineering pipeline, combining tracking (distilling bounding boxes to coordinates), and then adding an LSTM? SocialLSTM already did the second part in principle. I was a bit confused about why the "tracking-by-detection" framework was novel -- how is it different from "Tracking-learning-detection" (TLD), from Z. Kalal, K. Mikolajczyk, and J. Matas. In TPAMI 2012? Why are scene semantics privileged (which you also call "auxiliary") information? If the pipeline relies on fusing the "location" and "semantic" stream, if privileged information is missing at test time, then won't the pipeline break? (privileged information is usually present at train time, but absent at test time). Can multiple objects be tracked at once, or just one? The video in the supplementary material only showed one object being tracked. What was the average number of objects in each SYNTHIA scene frame? The focus of the paper seems a bit unclear: apparently, the goal is to "predict car's future odometry given previous egomotion visual input". But the paper is about prediction of other vehicles' trajectories, not about predicting one's own car odometry (which is different). Odometry: "Odometry is the use of motion sensors to determine the robot's change in position relative to some known position" (not other robots' positions) https://groups.csail.mit.edu/drl/courses/cs54-2001s/odometry.html "Visual Odometry -- Estimating the motion of a camera in real time using sequential images (i.e., egomotion)" http://www.cs.toronto.edu/~urtasun/courses/CSC2541/03_odometry.pdf (but we are not trying to estimate the motion of a camera-mounted vehicle here). I felt like the detection/tracking theme slightly bogs down the main focus of prediction in the paper. I think refocusing it on prediction would strengthen the argument. Line 124 - The authors discuss a "lack of proper metrics for evaluating the temporal prediction result." How are the current metrics poorly suited to the task? What would be better?
[Preliminary Rating] Please rate the paper according to one of the following six choices:	Weak Reject
[Preliminary Evaluation] Please indicate to the AC, your fellow reviewers, and the authors your current opinion on the paper. Please tell the ACs what points you think have the most weight in your reviews and summary, and why.	My current opinion is "Weak Reject," but if the authors would add significant experiments to compare with papers like SocialLSTM, I would be willing to re-evaluate my opinion.
[Rebuttal Requests] Please pose questions you want to be answered in the	I would appreciate it if the authors would address each of the individual

rebuttal.	weaknesses I've enumerated above.
[Confidence]	Confident - to stress that you are mostly sure about your conclusions (e.g., you are not an expert but can distinguish good work from bad work in that area).

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
[Paper Summary] What is the paper about? Please, be concise (3 to 5 sentences)	The authors propose a novel system to predict future trajectories of other traffic participants for autonomous driving. At the core, the proposed system is an LSTM that takes as an input multiple streams, namely raw RGB images, detection/tracking results and semantic segmentation. It then produces trajectories of other vehicles for a duration that corresponds roughly to the total driver reaction time as an output.
[Paper Strengths] Please discuss, justifying your comments with the appropriate level of details, the strengths of the paper (i.e. novelty, theoretical approach and/or technical correctness, adequate evaluation, clarity, etc). For instance, a theoretical paper may need no experiments, while a paper with a new approach may require comparisons to existing methods.	The authors present a solid idea and explain it in great detail. I really like how this paper combines advancements from different fields (object detection, segmentation, and temporal models) and builds a system that takes advantage of multiple streams with various levels of abstraction. Extensive experiments show the effectiveness of the proposed method.
[Paper Weaknesses] Please discuss, justifying your comments with the appropriate level of details, the weaknesses of the paper (i.e. lack of novelty – given references to prior work-, lack of novelty, technical errors, or/and insufficient evaluation, etc). Note: If you think there is an error in the paper, please explain why it is an error. Also remember that theoretical results/ideas are essential to CVPR (some theoretical papers may not need to have experiments). If the theory is novel and interesting, but the results did not outperform other existing algorithms, it is not necessarily a reason to reject. It is not appropriate to ask for comparisons with unpublished papers and papers published after the CVPR deadline. In all cases, please be polite and constructive. CVPR 2018 policy on dual submission and arxiv appears at: http://cvpr2018.thecvf.com/submission/main_conference/author_guidelines.	There are too many contributions. For example, I would combine 1 and 3 and maybe also 2 and 4. In general, it is better to have a few strong contributions than many weak ones. In a real car accurate depth information might not be available. Also, tracking and semantic segmentation results in the real world will probably be noisier than for synthetic data. Experiments on real data would make the results much stronger. Figure 6: It would be nice to have some state-of-the-art tracking results as reference (e.g. ECO, STAPLE_CA, SiamFC – check: https://github.com/foolwood/benchmark_results) Section 3 is quite short and it seems that the tracking methodology would fit well here. In general a lot of methodology detail is actually in section 4 (implementation details). Maybe consider reorganizing a little bit. Overall, the paper is well-written, but here are a few things that should be fixed. Line 22, 100, 114, 121, 367, 386, 509, 518: missing 'the' Line 34: missing 'to' Line 48, 51, 104, 247, 263, 377, 456: missing 'a' Line 90: if stopping -> whether to stop Line 94: benefits -> benefit Line 247: a mount -> amount Line 262: Rewrite sentence Line 314: can be of -> can be Line 322: boundingbox -> bounding box Line 360: is as -> as Line 397: take -> takes Line 426: rewrite sentence Line 478: we especially concern -> we are especially concerned Line 501: back-projecting -> back-project Line 504: tracking vehicle -> tracked vehicle Line 512: missing 'sets' Line 562: is -> it is Line 564: we concerns mostly -> we are mostly concerned Line 613: undergoing -> undergo Line 627: high way -> high-way
[Preliminary Rating] Please rate the paper according to one of the following six choices:	Poster
[Preliminary Evaluation] Please indicate to the AC, your fellow reviewers, and the authors your current opinion on the paper. Please tell the ACs what points you think have the most weight in your reviews and summary, and why.	The paper presents an approach for vehicle trajectory prediction. The combination of fusing multiple streams corresponding to different levels of abstraction (raw image, object detections, semantic segmentation) and using temporal consistency (LSTM) seems very promising and might open new avenues for solving autonomous driving. Currently, autonomous driving is usually solved with one of two extreme approaches either end-to-end driving where controls are predicted directly from images or parsing a scene into all components and then applying a complex rule catalog. The authors develop their idea logically throughout the paper with the practical use-case in mind (e.g. reaction time of humans). While the results on the synthetic data serve as a good proof-of-concept, it would have been nice to see some results on real data. Overall a solid paper though.
[Rebuttal Requests] Please pose questions you want to be answered in the rebuttal.	Do you think the results you obtained with the Synthia data will transfer to real-world data? Not exactly a request, but maybe something to consider for future work: There are photo-realistic simulators available now (CARLA, Sim4CV, etc.) which allow simulation of autonomous driving with 'real' physics and photo-realism similar to Synthia. It would be interesting to implement this system and evaluate 'real-world' performance.
[Confidence]	Very Confident - to stress that you are pretty sure about your conclusions (e.g., you are an expert who works in the paper's area).