CityFlow-NL: Tracking and Retrieval of Vehicles at City Scale by Natural Language Descriptions

Qi Feng
Boston University
fung@bu.edu

Vitaly Ablavsky
University of Washington
vxa@uw.edu

Stan Sclaroff
Boston University
sclaroff@bu.edu

Abstract

Natural Language (NL) descriptions can be the most convenient or the only way to interact with systems built to understand and detect city scale traffic patterns and vehicle-related events. In this paper, we extend the widely adopted CityFlow Benchmark with natural language descriptions for vehicle targets and introduce the CityFlow-NL Benchmark. The CityFlow-NL contains more than 5,000 unique and precise NL descriptions of vehicle targets, making it the largest-scale tracking with NL descriptions dataset to our knowledge. Moreover, the dataset facilitates research at the intersection of multi-object tracking, retrieval by NL descriptions, and temporal localization of events. The code and data we use in this paper are available at: https://github.com/fredfung007/cityflow-nl.

1. Introduction

Understanding city-wide traffic patterns and detecting specific vehicle-related events in real time has applications ranging from urban planning and traffic-engineering to law enforcement. In such applications, Natural Language (NL) descriptions can be the most convenient (from a user’s standpoint) or the only way to interact with a system (e.g., taking as input an informal description provided by a bystander). Therefore, it becomes necessary to track and retrieve vehicles, potentially seen from multiple spatio-temporally disjoint viewpoints, based on natural-language queries.

The task of Multi-Target Multi-Camera (MTMC) tracking of vehicles is first introduced by Tang et al. [17]. We extend the MTMC tracking problem with natural language (NL) and formulate the vehicle tracking and retrieval problem. The task of vehicle tracking and retrieval by NL at city-scale is distinct from other language-vision tasks. While it shares similarities with the task of tracking a single target by natural-language specification [12, 5, 4] and with natural-language-based video retrieval [1, 21], it is different in that the proposed problem requires localizing both temporal and visual location (in the form of tracking) from the NL description. An added challenge in developing and evaluating algorithms for vehicle tracking and retrieval by NL descriptions is the lack of realistic datasets.

Therefore, we introduce CityFlow-NL, an open dataset designed to facilitate research at the intersection of multi-object tracking, retrieval by natural language specification, and temporal localization of events. Our benchmark is derived from CityFlow [17], which is itself a public dataset that has been at the center of several recent workshop-challenges focused on MTMC tracking and re-identification. As a result, the proposed CityFlow-NL dataset exhibits a diversity of real-world parameters, including traffic density and patterns, viewpoints, ambient conditions, etc. The NL descriptions are provided by at least three human annotators, thus capturing realistic variations and ambiguities that one could expect in such application domains. An example of the NL descriptions and vehicle targets are shown in Fig. 1.

2. Related Works

In the past decade, researchers have started to look into exploiting natural language understanding in computer vision tasks. These models usually consist of two components: a language model and an appearance model to learn a new feature space that is shared between both NL and visual appearance [10, 18]. More recent object detection and vision grounding models [8, 19] jointly exploit vision and NL using Siamese networks and depth-wise convolutional neural networks between the NL representations and visual representations.

The VisualGenome [11], Flickr-30K [20] and Flickr-30K entities [15] benchmarks facilitate research in
understanding language grounded in visual contents. These image datasets provide detailed descriptions of regions in an image but still lack temporal information that can enable systems to better handle the temporal context and motion patterns in the natural language descriptions.

OTB-99-LANG [12] is the first open dataset that provides NL descriptions for single object tracking, followed by LaSOT [3] which is a larger scale single object tracking benchmark annotated with NL descriptions. Both of these datasets are annotated with one NL description for the target object for the entire sequence. The form of the NL descriptions is limited and can be ambiguous when used for tracking, as NL descriptions in the form of a phrase or sentence can cause ambiguity in understanding the motion pattern of the target [4], especially over a longer time interval.

3. CityFlow-NL Benchmark

In this section, we describe the statistics of the proposed CityFlow-NL benchmark and how we extend and annotate the CityFlow Benchmark [17].

3.1. Dataset Overview

We extend the CityFlow Benchmark with Natural Language (NL) Descriptions for each target vehicle. The proposed CityFlow-NL consists of 3,028 tracks of vehicles from 40 calibrated cameras, and 5289 unique NL descriptions. The average number of frames a target vehicle shows in the CityFlow Benchmark is 75.85. The distribution of the number of frames of target vehicles is shown in Fig. 2. CityFlow-NL poses new problems and challenges in the CityFlow Benchmark.

CityFlow-NL makes it possible to build and evaluate systems that can jointly leverage both language and visual modalities. In this paper, we propose the vehicle retrieval by NL descriptions problem for the 2021 AI
4. Vehicle Retrieval by Natural Language Descriptions

We now describe how we use the proposed CityFlow-NL Benchmark for the Vehicle Retrieval by Natural Language Specification Track in the 2021 AI City Challenge.

4.1. The Problem Definition

For the purpose of this task, we utilize the proposed CityFlow-NL Benchmark in a single-view setup. For each single-view vehicle track, we bundle it with a query that consists of three different NL descriptions for training. On the other hand, during testing, the goal is to retrieve and rank vehicles tracks based on NL queries.

This variation of the proposed CityFlow-NL contains 2,498 tracks of vehicles with three unique natural language descriptions each. Additionally, 530 unique vehicle tracks together with 530 query sets each with three descriptions are curated for testing.

The proposed problem is uniquely posed with relation to action recognition problems and content-based image retrieval problems, thus introducing unique challenges in building retrieval systems. Different from prior content-based image retrieval systems [9, 6, 14], retrieval models for this problem need to consider the context of a vehicle track and motion patterns within the track as well.

4.2. Evaluation Metrics

The Vehicle Retrieval by NL Descriptions task is evaluated using standard metrics for retrieval tasks [13]. We use the Mean Reciprocal Rank (MRR) as the main evaluation metric. Recall @ 5, Recall @ 10, and Recall @ 25 are also evaluated for all models.

For each query in the testing split, we rank all 530 candidate tracks by the retrieval system and evaluate the retrieval performance based on the above-mentioned metrics.

4.3. The Baseline Model

We build the baseline model to measure the similarities between a vehicle track $T$ and a natural language description $Q$ by taking ideas from content-based image retrieval systems [20] and image captioning models [10, 18]. The overview of the baseline model is shown in Fig. 4.

The vehicle track is defined as a sequence of frames in a video clip, $T = \{F_1, \ldots, F_t\}$. We use the ground truth bounding box of the target vehicle to crop each frame and reshape the sequence of crops, denoted as $C$, to the same shape of $(256, 256)$. This sequence of

---

3. Figure 3: Screenshot of the website we use to collect the CityFlow-NL Benchmark. Annotators are given detailed instructions on how to annotate NL descriptions for the vehicle targets presented in multi-view GIFs.

3.2. NL Annotation Process

We split multi-view tracks from the CityFlow Benchmark by the timestamps of cameras into multiple multi-view (or single-view) tracks that do not overlap with each other temporally. For example, if a target shows up in camera 1 and camera 2 in two temporal segments that do not overlap with each other, we generate two separate tracks for our annotation process. For each multi-view track, we render video clips from all views with bounding boxes drawn around the target vehicle.

We design a website, shown in Fig. 3 for collecting NL annotations. Crowd-sourcing workers are directed to give a detailed NL description of the target vehicle that can be used to uniquely identify the target vehicle from other vehicles. NL descriptions of the same target can be very different, as NL descriptions are subjective and may focus on different aspects of the target. Examples of such annotations are shown in Fig. 1. We collect annotations for each multi-view track with at least three crowd-sourcing workers from Amazon SageMaker.
resized frame crops of the target is used as the visual representation of the vehicle: \( V = \{ C_1, \ldots, C_t \} \).

We build a siamese embedding model for the retrieval task, where each crop \( C_i \) is embedded by a ResNet-50 [7] pretrained on ImageNet [16], denoted as \( E_{C_i} \).

We use a pretrained BERT [2] to embed the natural language description \( Q \) into a 256 dimensional vector, denoted as \( E_Q \).

The similarity between \( C_i \) and \( Q \) is intuitively defined as the vector similarity:

\[
S(C_i, Q) = \exp\left(-\|E_{C_i} - E_Q\|_2\right). \tag{1}
\]

During training, both positive and negative pairs of \( C_i \) and \( Q \) are constructed. Negative pairs of \( C_i \) and \( Q \) are built by randomly selecting natural language descriptions that describe other targets. The baseline model is trained with a cross entropy loss and an initial learning rate of 0.01 for 20 epochs on 2 GPUs using a stochastic gradient descent optimizer.

For the purpose of inference, we measure the similarity between a test track and a test query as the average of the similarities between all pairs of crops in the track and natural language description in the test query, i.e.,

\[
S(T, \{Q_1, Q_2, Q_3\}) = \frac{1}{3 \cdot t} \sum_{j=1}^{3} \sum_{i=1}^{t} S(C_i, Q_j). \tag{2}
\]

For each test query, we compute this similarity on all test tracks and rank the tracks for evaluation. The baseline model achieves an MRR of 0.0269, Recall @ 5 of 0.0264, Recall @ 10 of 0.0491, Recall @ 25 of 0.1113.

The baseline model does not consider the motion pattern nor the context of vehicle tracks and the retrieval performance is not very competitive. Due to the nature of the CityFlow-NL dataset, the vehicle retrieval by natural language description problem could be better handled by taking additional features to produce a model that attains better performance than that obtained by the simple baseline.

5. Conclusion

This paper has presented CityFlow-NL, a city-scale multi-target multi-camera tracking with natural language descriptions dataset that provides precise natural language descriptions for multi-view ground truth vehicle tracks. Our NL annotations can be used to benchmark tasks like vehicle retrieval by NL, motion pattern analysis with NL, and MTMC with NL descriptions. It is the first large scale multi camera multi target tracking dataset to provide NL descriptions. As the CityFlow-NL is annotated with NL descriptions for vehicle tracks, systems that were built for single object tracking with NL descriptions [12, 5, 4] should also be able to make use of this dataset to continue to make progress on the problem.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In Proceedings of the IEEE international conference on computer vision, pages 5803–5812, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[3] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5374–5383, 2019.

[4] Qi Feng, Vitaly Ablavsky, Qinxun Bai, Guorong Li, and Stan Sclaroff. Real-time visual object tracking with natural language description. In The IEEE Winter Conference on Applications of Computer Vision, pages 700–709, 2020.

[5] Qi Feng, Vitaly Ablavsky, Qinxun Bai, Stan Sclaroff. Robust visual object tracking with natural language region proposal network. arXiv preprint arXiv:1912.02048, 2019.

[6] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In Advances in neural information processing systems, pages 678–688, 2018.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In European Conference on Computer Vision, pages 630–645. Springer, 2016.
[8] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[9] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016. 3, 4

[10] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016. 1, 3

[11] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1

[12] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6495–6503, 2017. 1, 2, 4

[13] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge University Press, 2008. 3

[14] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 3, 4

[15] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649, 2015. 1

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 4

[17] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasian, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019. 1, 2

[18] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 1, 3

[19] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4683–4693, 2019. 1

[20] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1, 3

[21] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019. 1