Single-camera and Inter-camera Vehicle Tracking and 3D Speed Estimation Based on Fusion of Visual and Semantic Features

Zheng Tang, Gaoang Wang, Hao Xiao, Aotian Zheng, Jenq-Neng Hwang

University of Washington, Seattle, WA 98195, USA

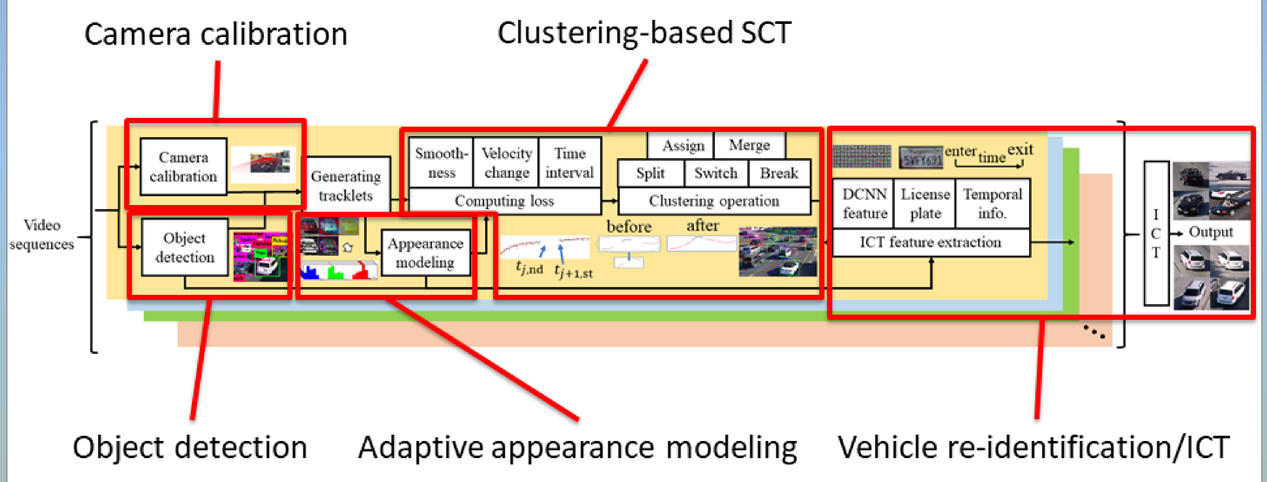
Abstract

Tracking of vehicles across multiple cameras with non-overlapping views has been a challenging task for the intelligent transportation system (ITS). It is mainly because of high similarity among vehicle models, frequent occlusion, large variation in different viewing perspectives and low video resolution. In this work, we propose a fusion of visual and semantic features for both single-camera tracking (SCT) and inter-camera tracking (ICT). Specifically, a histogram-based adaptive appearance model is introduced to learn long-term history of visual features for each vehicle target. Besides, semantic features including trajectory smoothness, velocity change and temporal information are incorporated into a bottom-up clustering strategy for data association in each single camera view. Across different camera views, we also exploit other information, such as deep learning features, detected license plate features and detected car types, for vehicle re-identification. Additionally, evolutionary optimization is applied to camera calibration for reliable 3D speed estimation. Our algorithm achieves the top performance in both 3D speed estimation and vehicle re-identification at the NVIDIA AI City Challenge 2018.

Introduction

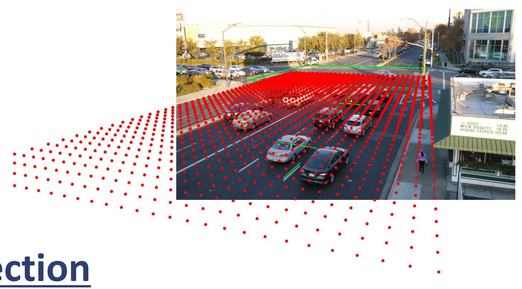
- **Single-Camera Tracking (SCT)**: Object detection /classification + data association
- **Inter-Camera Tracking (ICT)**: Re-identification of the same object(s) across multiple cameras
- **Challenges**: High similarity among vehicle models, frequent occlusion, large variation in different viewing perspectives, etc.

Overview



Camera calibration

- Minimization of reprojection error solved by EDA
- $$\min_{\mathbf{P}} \sum_{k=1}^{N_{\mathrm{ls}}} \left| \|P_k - Q_k\|_2 - \|\widehat{P_k} - \widehat{Q_k}\|_2 \right|$$
  $$\text{s.t. } \mathbf{P} \in \mathrm{Rng}_{\mathbf{P}}, p_k = \mathbf{P} \cdot \widehat{P_k}, q_k = \mathbf{P} \cdot \widehat{Q_k}$$

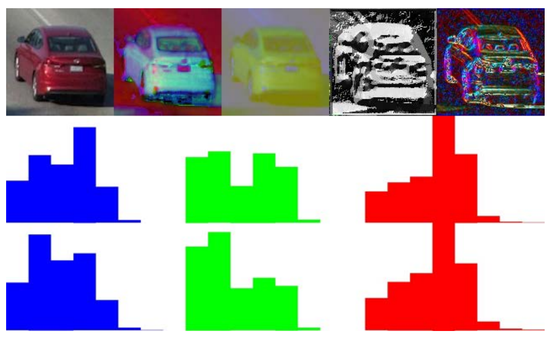


Object detection

- YOLO v2 [1] trained on ~4,500 manually labeled frames
- 8 categories: Sedan, hatchback, bus, pickup, minibus, van, truck and motorcycle
- Initialization: Provided pre-trained weights

Adaptive appearance modeling

- **A history of spatially weighted histogram combinations** will be kept for each vehicle
- **Feature space**: RGB, HSV, Lab, LBP and HOG
- **Spatial weighting** by Gaussian (kernel) distribution
- **Comparison**: (Average) Bhattacharyya distance between each pair of histograms in the adaptive appearance models



Clustering-based SCT

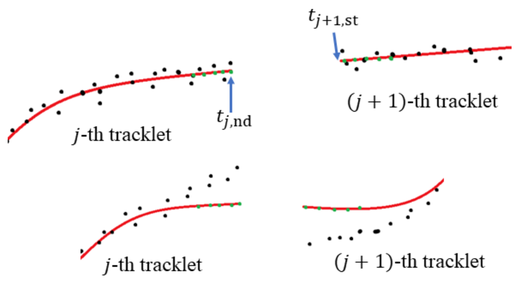
- **Loss function**: $l = \sum_{i=1}^{n_{\mathrm{v}}} l_i$

$$l_i = \lambda_{\mathrm{sm}} l_{i,\mathrm{sm}} + \lambda_{\mathrm{vc}} l_{i,\mathrm{vc}} + \lambda_{\mathrm{ti}} l_{i,\mathrm{ti}} + \lambda_{\mathrm{ac}} l_{i,\mathrm{ac}}$$

- **Smoothness loss**: The total distance between the regression trajectory and observed trajectory

Same trajectory

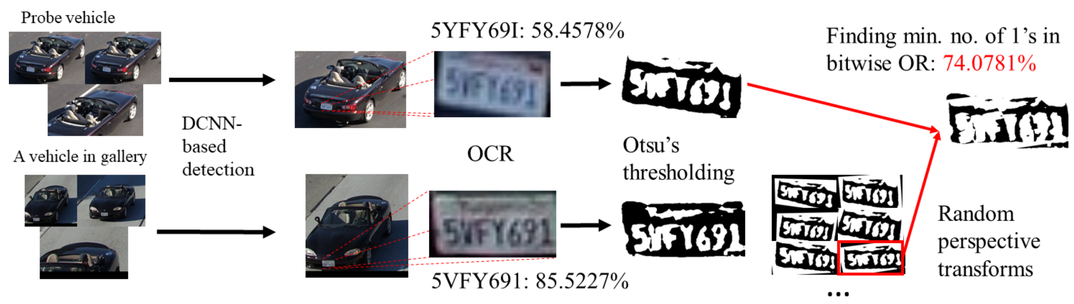
Different trajectory



- **Velocity change loss**: Maximum acceleration around each end point of the tracklets
- **Time interval loss**: Time interval between two adjacent tracklets
- **Appearance change loss**: Distance between adaptive appearance models
- **Clustering operations** $\Delta l_j{}^* = \arg\min_{\Delta l_j} (\Delta l_{j,\mathrm{as}}, \Delta l_{j,\mathrm{mg}}, \Delta l_{j,\mathrm{sp}}, \Delta l_{j,\mathrm{sw}}, \Delta l_{j,\mathrm{bk}})$
- $\Delta l_{j,\mathrm{as}}$, $\Delta l_{j,\mathrm{mg}}$, $\Delta l_{j,\mathrm{sp}}$, $\Delta l_{j,\mathrm{sw}}$ and $\Delta l_{j,\mathrm{bk}}$ respectively stand for the changes of loss for assign, merge, split, switch and break operations.
- The operation with minimum loss-change value is chosen.
- If $\Delta l_j{}^* > 0$, no change is made for this tracklet.

Vehicle re-identification/ICT

- Loss function: $L = \sum_{I=1}^{N_{\mathrm{v}}} L_I \qquad L_I = L_{I,\mathrm{ac}} \times L_{I,\mathrm{nn}} \times L_{I,\mathrm{lp}} \times L_{I,\mathrm{ct}} \times L_{I,\mathrm{tt}}$
- **Appearance change loss**: Distance between adaptive appearance models
- **Matching loss of DCNN features**: Bhattacharyya distance between DCNN features given by pre-trained model on the Comprehensive Cars (CompCars) dataset [2]
- **License plate comparison loss**:



- **Traveling time loss**:

Experimental results

- Track 1 - Traffic flow analysis
  - 27 videos, each 1 minute in length, recorded at 30 fps and 1080p resolution
  - Performance evaluation: $S1 = DR \times (1 - NRMSE)$
  - *DR* is the detection rate and *NRMSE* is the normalized Root Mean Square Error (RMSE) of speed
- Track 3 - Multi-camera vehicle detection and re-identification
  - 15 videos, each around 0.5-1.5 hours long, recorded at 30 fps and 1080p resolution
  - Performance evaluation: $S3 = 0.5 \times (TDR + PR)$
  - *TDR* is the trajectory detection rate and *PR* is the localization precision

Table 1. Quantitative comparison of speed estimation on the *NVIDIA AI City Dataset* [9]

Rank	Team	S1 Score
1	**Ours**	**1.0000**
2	team79	0.9162
3	team78	0.8892
4	team24	0.8813
5	team12	0.8331
6	team4	0.7924
7	team65	0.7654
8	team6	0.7174
9	team40	0.6564
10	team26	0.6547
11	team18	0.6226
12	team45	0.5953
13	team39	0.0000

Table 2. Quantitative comparison of multi-camera tracking on the *NVIDIA AI City Dataset* [9]

Rank	Team	S3 Score
1	**Ours**	**0.7106**
2	team37	0.2861
3	team79	0.0785
4	team18	0.0074
5	team28	0.0026
6	team41	0.0024
7	team53	0.0002
8	team6	0.0001
9	team10	0.0000
10	team31	0.0000

* Bold entries indicate the rank #1 in each comparison.

Source code available at GitHub: https://github.com/zhengthomastang/2018AICity_TeamUW

References

[1] J. Redmon and A. Farhadi. YOLO9000: Better, Faster, Stronger. arXiv preprint arXiv: 1612.08242, 2016.

[2] L. Yang, P. Luo, C. C. Loy and X. Tang. A large-scale car dataset for fine-grained categorization and verification. Proc. IEEE Conf. Comput. Vis. Pattern Recogn., 3973–3981, 2015.