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

Team 48
Zheng Tang, Gaoang Wang, Hao Xiao, Aotian Zheng, Jenq-Neng Hwang
Information Processing Lab, Department of Electrical Engineering
University of Washington
Introduction

• Intelligent Transportation System (ITS)
  – Estimating traffic flow
  – Anomalies detection
  – Multi-camera tracking and re-identification

• Single-Camera Tracking (SCT)
  – Object detection/classification + data association

• Inter-Camera Tracking (ICT)
  – Re-identification of the same object(s) across multiple cameras
Introduction

• Challenges in SCT & ICT
  – Extraction of 3D information
  – Failure/confusion in object detection
  – High similarity among vehicle models
  – Frequent occlusion
  – Large variation in different viewing perspectives
  – Low video resolution (for license plate recognition)
Overview

Camera calibration

Clustering-based SCT

Object detection

Adaptive appearance modeling

Vehicle re-identification/ICT
Camera Calibration

- Minimization of reprojection error solved by EDA

\[
\min_{\mathbf{P}} \sum_{k=1}^{N_{1_{\text{S}}}} \left( \| P_k - Q_k \|_2 - \| \bar{P}_k - \bar{Q}_k \|_2 \right)
\]

s. t. \( \mathbf{P} \in \text{Rng}_\mathbf{P}, p_k = \mathbf{P} \cdot \bar{P}_k, q_k = \mathbf{P} \cdot \bar{Q}_k \)

- \( \mathbf{P} \): Camera projection matrix
- \( \text{Rng}_\mathbf{P} \): Range for optimization
- \( P_k, Q_k \): True endpoints of line segments
- \( \bar{P}_k, \bar{Q}_k \): Estimated endpoints of line segments
- \( p_k, q_k \): 2D endpoints of line segments
- \( N_{1_{\text{S}}} \): Number of endpoints
Object Detection

- YOLOv2 [Redmon et al., CVPR 2017]
  - 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

- Histogram-based adaptive appearance model
  - A history of spatially weighted (kernel) histogram combinations will be kept for each vehicle

The first row respectively presents the RGB, HSV, Lab, LBP and gradient feature maps for an object instance in a tracklet, which are used to build feature histograms.

The second row shows the original RGB color histograms.

The third row demonstrates the Gaussian spatially weighted (kernel) histograms, where the contribution of background area is suppressed.
Clustering-based SCT

\[ l = \sum_{i=1}^{n_v} l_i \]

\[ l_i = \lambda_{sm} l_{i,sm} + \lambda_{vc} l_{i,vc} + \lambda_{ti} l_{i,ti} + \lambda_{ac} l_{i,ac} \]

- Smoothness
- Velocity
- Time interval
- Appearance

- \( n_v \): No. of vehicles in a single camera
- \( l_i \): Loss for the \( i \)-th vehicle
- \( l_{i,sm} \): Smoothness loss
- \( l_{i,vc} \): Velocity change loss
- \( l_{i,ti} \): Time interval loss
- \( l_{i,ac} \): Appearance change loss
- \( \lambda \)'s: Regularization parameters

**Black dots** show the detected locations at time \( t \).

**Red curves** represent trajectories from Gaussian regression.

**Green dots** show \( n_k \) neighboring points on the red curves around the endpoints of the tracklets at \( t_{j,nd} \) and \( t_{j+1,nt} \).
Clustering-based SCT

• Smoothness loss
  – The total distance between the regression trajectory and observed 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
  – (Average) Bhattacharyya distance between each pair of histograms in the adaptive appearance models
Clustering-based SCT

- Clustering operations

\[
\Delta l_j^* = \arg \min_{\Delta l_j} \left( \Delta l_{j,\text{as}}, \Delta l_{j,\text{mg}}, \Delta l_{j,\text{sp}}, \Delta l_{j,\text{sw}}, \Delta l_{j,\text{bk}} \right)
\]

- \(\Delta l_{j,\text{as}}, \Delta l_{j,\text{mg}}, \Delta l_{j,\text{sp}}, \Delta l_{j,\text{sw}}\) and \(\Delta l_{j,\text{bk}}\) 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.
- Convergence is guaranteed.
Clustering-based SCT

- **Assign operation**

\[
\Delta l_{j,as} = \min_i \left( l(S(j) \setminus \tau_j) + l(S_i \cup \tau_j) \right) - \left( l(S(j)) + l(S_i) \right)
\]

- Loss after operation
- Loss before operation

- \( \tau_j \): The tracklet of interest
- \( S(j) \): The trajectory set of \( \tau_j \), noted \( S(j) \)

![Diagram showing before and after tracking operations](image)
Clusterings-based SCT

- Merge operation

\[ \Delta l_{j,mg} = \min_i (l(S(j) \cup S_i)) - \left( l(S(j)) + l(S_i) \right) \]

Loss after operation  Loss before operation

Trajectory 1 \((S(j))\)

Trajectory 2 \((S_i)\)

before

after
Clustering-based SCT

- Split operation

\[ \Delta l_{j,sp} = (l(\tau_j) + l(S(j) \setminus \tau_j)) - l(S(j)) \]

Loss after operation  Loss before operation

Trajectory 1 \((S(j))\)

Trajectory 2 \((S_i)\)

before

after
Clustering-based SCT

- Switch operation

\[ \Delta l_{sw} = \min_i \left( l(S_{bef}(j) \cup S_{i,aft}) + l(S_{aft}(j) \cup S_{i,bef}) \right) - \left( l(S(j)) + l(S_i) \right) \]

- \( S_{bef}(j) \): Tracklets before \( \tau_j \) in \( S(j) \)
- \( S_{aft}(j) \): Tracklets after \( \tau_j \) in \( S(j) \)
Clustering-based SCT

- Break operation

\[ \Delta l_{bk} = \left( l(S_{bef}(j)) + l(S_{aft}(j)) \right) - l(S(j)) \]

Loss after operation  Loss before operation

Trajectory 1 (\(S(j)\))

Trajectory 2 (\(S_i\))

before  after
Vehicle Re-identification/ICT

\[ L = \sum_{i=1}^{N_v} L_i \]

\[ L_i = L_{i,ac} \times L_{i,nn} \times L_{i,lp} \times L_{i,ct} \times L_{i,tt} \]

- **Appearance change loss**
  - (Average) **Bhattacharyya distance** between each pair of histograms in the adaptive appearance models

- **Mis-classified car type loss**
  - Different **detected categories** (majority vote) between vehicles will cause penalty.

\( N_v \): No. of vehicles appeared in all cameras
\( L_i \): Loss for the i-th vehicle
\( L_{i,ac} \): Appearance change loss
\( L_{i,nn} \): Matching loss of DCNN features
\( L_{i,lp} \): License plate comparison loss
\( L_{i,ct} \): Mis-classified car type loss
\( L_{i,tt} \): Traveling time loss
Vehicle Re-identification/ICT

- Matching loss of DCNN features
  - Pre-trained model on the Comprehensive Cars (CompCars) dataset
  - 3 images are chosen for each vehicle for feature extraction
  - The dimension of each feature vector is 1024
  - Comparison given by Bhattacharyya distance
Vehicle Re-identification/ICT

- License plate comparison loss

The confidence of OCR is too low

Finding min. no. of 1’s in bitwise OR: 74.0781%

Random perspective transforms
Vehicle Re-identification/ICT

- 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: \( S_1 = DR \times (1 - NRMSE) \)
  - \( DR \) 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: \( S_3 = 0.5 \times (TDR + PR) \)
  - \( TDR \) is the trajectory detection rate and \( PR \) is the localization precision
Track 1 Experimental Results

DR: 1.0000  RMSE: 4.0963 mi/h
https://youtu.be/_i4numqiv7Y
Table 2. Quantitative comparison of multi-camera tracking on the NVIDIA AI City Dataset [9]

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>S3 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>team48</td>
<td>0.7106</td>
</tr>
<tr>
<td>2</td>
<td>team37</td>
<td>0.2861</td>
</tr>
<tr>
<td>3</td>
<td>team79</td>
<td>0.0785</td>
</tr>
<tr>
<td>4</td>
<td>team18</td>
<td>0.0074</td>
</tr>
<tr>
<td>5</td>
<td>team28</td>
<td>0.0026</td>
</tr>
<tr>
<td>6</td>
<td>team41</td>
<td>0.0024</td>
</tr>
<tr>
<td>7</td>
<td>team53</td>
<td>0.0002</td>
</tr>
<tr>
<td>8</td>
<td>team6</td>
<td>0.0001</td>
</tr>
<tr>
<td>9</td>
<td>team10</td>
<td>0.0000</td>
</tr>
<tr>
<td>10</td>
<td>team31</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

TDR: 3/7    PR: 0.9925

https://youtu.be/Jlvh_KxHl40
Conclusion

- Fusion of visual and semantic features for SCT: motion, temporal and appearance attributes
- Fusion of visual and semantic features for ICT: appearance, license plate, vehicle type and temporal attributes
- Adaptive appearance model to robustly encode long-term appearance change
- Camera calibration based on EDA optimization for reliable 2D-to-3D backprojection
- Top performance in both Track 1 & Track 3 on the challenge dataset
- GitHub: https://github.com/zhengthomastang/2018AICity_TeamUW