Vehicle tracking by detection in UAV aerial video
1. Introduction
2. Tracking by detection
3. Experiments
1. Introduction

UAV video is becoming an effective supplement to fixed monitoring video.

It has obvious advantages in the aspect of traffic information acquisition:

• Unique flexibility and easy maneuverability
• More comprehensive and clear video data
• Fast deployment in emergent situation

The purpose of this paper is to give a method of vehicles tracking in unmanned aerial vehicle (UAV) video to support further traffic information analysis.
1. Introduction

At present, there are mainly three kinds of vehicle detection algorithms:

- Detection based on prior knowledge
- Detection based on motion information
- Detection based on machine learning
  - Not depend on shifting scene
  - Suitable for big traffic
  - High accuracy

Tracking: determine the position of the object in every frame continuously

- Tracking based on Kalman filtering
  - Small storage
  - Small computational cost
- Tracking based on Mean Shift
- Tracking based on particle filter
- on-line boosting tracking method

Characteristics of traffic video shooting by UVA:
Large scene, different background, containing dozens of vehicles with different size and wide-range of speed
1. Introduction
2. Tracking by detection
3. Experiments
As a tracking by detection method, we firstly detect vehicles in an UAV Video. Here we use Faster-RCNN in following steps:

- Extract feature maps from each UAV aerial image using sharable convolutional layers.
- Generate region proposals on feature maps using RPN module.
- Train the RPN module and detector module in Fast R-CNN with shared features.
2. Tracking by detection

To utilize Kalman filtering as a tracking method, we construct a Kalman filter motion model including equations of state and observation. During the tracking process, the time between the adjacent images is very short. We assume that the motion of vehicle in unit time is an uniform motion, the system state and the observed value are linear. So the equation of state is:

\[
S(t) = A(\Delta t)S(t - \Delta t) + \omega(t - \Delta t)
\]

\(S(t)\) represents the state of the system at time \(t\)
\(A(\Delta t)\) expresses the state transform matrix within \(\Delta t\)
\(\omega(t)\) indicates the estimation error

We use four dimensional vectors, containing the vehicle’s position and velocity, to represent the system state and the estimation error.

\[
S(t) = [x(t), y(t), \dot{x}(t), \dot{y}(t)]^T
\]

\[
\omega(t) = [\omega_x(t), \omega_y(t), \omega_{\dot{x}}(t), \omega_{\dot{y}}(t)]^T
\]

\[
A(\Delta t) = \begin{bmatrix}
1 & 0 & \Delta t & 0 \\
0 & 1 & 0 & \Delta t \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

In the UAV images, only the position can be observed for the state vector. We assume that the observed value are linear, so the equation of observation is:

\[
O(t) = H(t)S(t) + e(t)
\]

\(H(t)\) is the observation matrix, represents as:

\[
H(t) = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{bmatrix}
\]

\(O(t)\) is the observation vector
\(e(t)\) is the observation error.
2. Tracking by detection

After constructing the motion model with equations of state and observation, we could track vehicles based on detection results by Kalman filtering algorithm.

First, initialize Kalman filter. Then, run the loop of predicting the state, matching target and updating Kalman parameters.

Under the help of prediction function, Kalman filter gets trajectory for every target very fast.

Figure 6. Processing flow of vehicle tracking by Kalman filtering algorithm
TABLE OF CONTENTS

1. Introduction
2. Tracking by detection
3. Experiments
3. Experiments

Shoot device and scenarios:

We used a Dajiang “Mavic” UAV to get videos for our experiments.

- Shot from different fly altitudes: 76m, 86m and 96 m
- Varied traffic scenarios: crossroad, a three-way intersection, a roundabout and an exit of a main stem

Figure 7. Images from a same crossroad in different seasons with different fly altitudes (from left to right: 76m, 86m and 96m)
3. Experiments

Shoot device and scenarios:

We used a Dajiang “Mavic” UAV to get videos for our experiments.

• Shot from different fly altitudes: 76m, 86m and 96 m
• Varied traffic scenarios: crossroad, a three-way intersection, a roundabout and an exit of a main stem

Figure 8. Images from different places
(from left to right: a roundabout, a three-way intersection, an exit of a main stem)
3. Experiments

Training samples:
★ We use one of the video shot from 76 meters height on a crossroad at winter as our training examples.
★ Totally: 1,314 UAV images

Benchmarks:
★ We labeled the rest videos as our benchmarks (Table 1).
★ Totally: 8 videos, totally over 11 minutes.

Evaluation criterion:
We use the recall (Re) and the precision (Pr) as the evaluation criterions of the detection results.
\[
\text{Re} = \frac{TP}{(TP + FN)} \\
\text{Pr} = \frac{TP}{(TP + FP)}
\]
TP is the correctly detected number of vehicles;
FP is the number of errors being mistaken as the vehicles;
FN is the number of vehicles detected as the background

<table>
<thead>
<tr>
<th>Name</th>
<th>Fly Height</th>
<th>Scene</th>
<th>Season</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR 76 Winter 2</td>
<td>76</td>
<td>A crossroad as same as the one in our training examples</td>
<td>Winter</td>
</tr>
<tr>
<td>CR 76 Winter 3</td>
<td>76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CR 76 Summer</td>
<td>76</td>
<td></td>
<td>Summer</td>
</tr>
<tr>
<td>CR 86 Summer</td>
<td>86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CR 96 Summer</td>
<td>96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RA 76</td>
<td>76</td>
<td>A roundabout</td>
<td></td>
</tr>
<tr>
<td>3WI 76</td>
<td>76</td>
<td>A three-way intersection</td>
<td></td>
</tr>
<tr>
<td>Exit 76</td>
<td>76</td>
<td>An exit of a main stem</td>
<td></td>
</tr>
</tbody>
</table>
3. Experiments

Results of detection:

<table>
<thead>
<tr>
<th>Video Name</th>
<th>Faster RCNN</th>
<th></th>
<th></th>
<th>YOLOv3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pr</td>
<td>Re</td>
<td>Pr</td>
<td>Re</td>
<td></td>
</tr>
<tr>
<td>CR 86 Summer</td>
<td>0.9261</td>
<td>0.9664</td>
<td>0.816</td>
<td>0.7901</td>
<td></td>
</tr>
<tr>
<td>CR 96 Summer</td>
<td>0.9806</td>
<td>0.9641</td>
<td>0.854</td>
<td>0.8071</td>
<td></td>
</tr>
<tr>
<td>CR 76 Summer</td>
<td>0.9694</td>
<td>0.9601</td>
<td>0.9243</td>
<td>0.9338</td>
<td></td>
</tr>
<tr>
<td>RA 76</td>
<td>0.9245</td>
<td>0.9807</td>
<td>0.6106</td>
<td>0.6496</td>
<td></td>
</tr>
<tr>
<td>3WI 76</td>
<td>0.908</td>
<td>0.964</td>
<td>0.8542</td>
<td>0.803</td>
<td></td>
</tr>
<tr>
<td>CR 76 Winter 2</td>
<td>0.9762</td>
<td>0.9586</td>
<td>0.9049</td>
<td>0.9013</td>
<td></td>
</tr>
<tr>
<td>CR 76 Winter 3</td>
<td>0.919</td>
<td>0.8891</td>
<td>0.8543</td>
<td>0.8752</td>
<td></td>
</tr>
<tr>
<td>Exit 76</td>
<td>0.9352</td>
<td>0.9534</td>
<td>0.9019</td>
<td>0.8774</td>
<td></td>
</tr>
<tr>
<td>average</td>
<td>0.942375</td>
<td>0.95455</td>
<td>0.840025</td>
<td>0.829688</td>
<td></td>
</tr>
</tbody>
</table>
3. Experiments

Results of detection:

Figure 9. Detection results for CR 76 Winter 2 and CR 76 Winter 3.
3. Experiments

Results of detection:

Figure 10. Detection results for CR 86 Summer (up) and CR 96 Summer (bottom).
3. Experiments

Results of detection:

Figure 11. Detection results for RA 76 (left), 3WI 76 (middle) and Exit 76 (right).
### 3. Experiments

#### Results of tracking:

<table>
<thead>
<tr>
<th>Video Name</th>
<th>CR 86 Summer</th>
<th>CR 96 Summer</th>
<th>CR 76 Summer</th>
<th>RA 76</th>
<th>3WI 76</th>
<th>CR 76 Winter 2</th>
<th>CR 76 Winter 3</th>
<th>Exit 76</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT</td>
<td>89</td>
<td>110</td>
<td>183</td>
<td>43</td>
<td>96</td>
<td>134</td>
<td>159</td>
<td>119</td>
</tr>
<tr>
<td>Tracking result based on Faster RCNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT</td>
<td>87</td>
<td>107</td>
<td>179</td>
<td>42</td>
<td>93</td>
<td>131</td>
<td>151</td>
<td>115</td>
</tr>
<tr>
<td>PT</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>ML</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>IDs</td>
<td>58</td>
<td>12</td>
<td>138</td>
<td>5</td>
<td>37</td>
<td>13</td>
<td>96</td>
<td>103</td>
</tr>
<tr>
<td>MOTA</td>
<td>90.47%</td>
<td>95.98%</td>
<td>92.66%</td>
<td>95.73%</td>
<td>91.11%</td>
<td>94.92%</td>
<td>81.96%</td>
<td>94.41%</td>
</tr>
<tr>
<td>MOTP</td>
<td>0.011</td>
<td>0.011</td>
<td>0.009</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.014</td>
<td>0.011</td>
</tr>
<tr>
<td>Tracking result based on YOLOv3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT</td>
<td>77</td>
<td>91</td>
<td>174</td>
<td>33</td>
<td>79</td>
<td>124</td>
<td>150</td>
<td>92</td>
</tr>
<tr>
<td>PT</td>
<td>6</td>
<td>12</td>
<td>7</td>
<td>6</td>
<td>13</td>
<td>5</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>ML</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>IDs</td>
<td>63</td>
<td>33</td>
<td>25</td>
<td>41</td>
<td>12</td>
<td>72</td>
<td>70</td>
<td>91</td>
</tr>
<tr>
<td>MOTA</td>
<td>61.31%</td>
<td>67.08%</td>
<td>86.39%</td>
<td>24.68%</td>
<td>69.35%</td>
<td>80.14%</td>
<td>72.70%</td>
<td>76.22%</td>
</tr>
<tr>
<td>MOTP</td>
<td>0.172</td>
<td>0.167</td>
<td>0.163</td>
<td>0.145</td>
<td>0.189</td>
<td>0.152</td>
<td>0.163</td>
<td>0.177</td>
</tr>
</tbody>
</table>
3. Experiments

Results of tracking:

Figure 12. Tracking results of the 1st, 100th, 200th, 300th frame of one video
Thank you