Unsupervised Moving Object Detection through Background Models for PTZ Camera

Kimin Yun, Hyungil Kim, Kangmin Bae, Jongyoul Park
Overview: Moving Object Detection in PTZ camera

• Related research area
  ≈ video object detection / segmentation
  • supervised method / human intervention for first frame

• Ours: Background-Centric approach
  • Visual surveillance and monitoring
  • Strength: Unsupervised method / Real-time operation without GPU
Overview

• Naive approach
  • Background modeling with Gaussians (mean, variance)
  • Apply the affine/projective transform for Moving Camera
  • Problem: Too many false positives

• Conventional approach
  • Reduce the false positives
  • Apply the Spatio-Temporal background modeling
  • Problem: Foreground loss caused by background contamination
• The overall framework of the proposed method
  • Build the fine background models by extending ViBe*
    • spatio-temporal update is applied
    • model initialization and update rule is changed by camera movement
  • Combine the coarse and fine foreground using Watershed segmentation

Proposed method

- Fine background model
- reduce the background contamination

**Algorithm 1:** Updating the fine background model

\[
\text{for Each pixel } i \text{ on current frame } I^{(t)} \text{ do}
\]
\[
\text{if } t = 0 \text{ then}
\]
\[
M_i \text{ is initialized to } \{I_i^{(0)}, I_i^{(0)}, \ldots, I_i^{(0)}\}
\]
\[
\text{else}
\]
\[
\text{Motion compensation is applied to } M_i.
\]
\[
\text{if } t < N \text{ then}
\]
\[
\tilde{v}_i^{(t)} \text{ is removed from } M_i
\]
\[
I_i^{(t)} \text{ is inserted to } M_i
\]
\[
\text{Compute } C_i \text{ in equation (8)}
\]
\[
\text{if } C_i \geq \#_{\text{min}} \text{ then}
\]
\[
P \leftarrow \min(\alpha_i, \phi)
\]
\[
p \sim \text{Uniform}(0, P - 1)
\]
\[
\text{if } p = 0 \text{ then} \quad \triangleright \text{ update for pixel } i
\]
\[
n \sim \text{Uniform}(0, N - 1)
\]
\[
\tilde{v}_i^{(n)} \text{ is removed from } M_i
\]
\[
I_i^{(t)} \text{ is inserted to } M_i
\]
\[
p_2 \sim \text{Uniform}(0, P - 1)
\]
\[
\text{if } p_2 = 0 \text{ then} \quad \triangleright \text{ update for neighbor pixel } j
\]
\[
k \sim \text{Uniform}(0, K)
\]
\[
j \leftarrow S_i(k)
\]
\[
n \sim \text{Uniform}(0, N - 1)
\]
\[
\tilde{v}_j^{(n)} \text{ is removed from } M_j
\]
\[
I_i^{(t)} \text{ is inserted to } M_j
\]

\[
C_i = \sum_{j \in S_i} \sum_{n=1}^{N} 1(D(I_i, \tilde{v}_j^{(n)}) < R), \quad (8)
\]
Experiments

• Compared methods
  • Object-centric methods
    • uNLC (unsupervised version of NLC)
    • OSVOS (video object segmentation without finetuning)
    • CIS
    • BASNet (Salient object detector)
  • Background-centric methods
    • Background modeling + Naive extension (ViBe*, FIC*, BMRI-ViBe*)
    • Conventional methods: MCD NP, MCD 5.8ms, Stochastic approx, FP Sampling, , SC MCD

• Dataset
  • Moving camera dataset from SC MCD

< Reference >
Experiment results

- F-score measure

<table>
<thead>
<tr>
<th>Method</th>
<th>walking</th>
<th>skating</th>
<th>woman</th>
<th>woman2</th>
<th>fence</th>
<th>ground1</th>
<th>ground2</th>
<th>ground3</th>
<th>ground4</th>
<th>ground5</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViBe* [6]</td>
<td>0.0375</td>
<td>0.2229</td>
<td>0.0375</td>
<td>0.0929</td>
<td>0.1042</td>
<td>0.5656</td>
<td>0.4733</td>
<td>0.4118</td>
<td>0.0299</td>
<td>0.1309</td>
<td>0.2107</td>
</tr>
<tr>
<td>FIC* [8]</td>
<td>0.0613</td>
<td>0.2373</td>
<td>0.0361</td>
<td>0.1345</td>
<td>0.0954</td>
<td>0.4543</td>
<td>0.4108</td>
<td>0.1538</td>
<td>0.0453</td>
<td>0.1319</td>
<td>0.1761</td>
</tr>
<tr>
<td>BMRI-ViBe* [9]</td>
<td>0.0438</td>
<td>0.2402</td>
<td>0.0400</td>
<td>0.0921</td>
<td>0.1104</td>
<td>0.4249</td>
<td>0.3868</td>
<td>0.2161</td>
<td>0.0383</td>
<td>0.1377</td>
<td>0.1730</td>
</tr>
<tr>
<td>MCD NP [25]</td>
<td>0.4351</td>
<td>0.4164</td>
<td>0.4935</td>
<td>0.5791</td>
<td>0.2691</td>
<td>0.2773</td>
<td>0.3750</td>
<td>0.1222</td>
<td>0.1969</td>
<td>0.3540</td>
<td>0.3519</td>
</tr>
<tr>
<td>MCD 5.8ms [26]</td>
<td>0.7349</td>
<td>0.2447</td>
<td>0.3395</td>
<td>0.3448</td>
<td>0.7357</td>
<td>0.6573</td>
<td>0.7177</td>
<td>0.1531</td>
<td>0.5274</td>
<td>0.0678</td>
<td>0.4523</td>
</tr>
<tr>
<td>Stochastic approx [28]</td>
<td>0.8335</td>
<td>0.6543</td>
<td>0.3986</td>
<td>0.8783</td>
<td>0.8788</td>
<td>0.2221</td>
<td>0.2792</td>
<td>0.0181</td>
<td>0.0111</td>
<td>0.2181</td>
<td>0.4392</td>
</tr>
<tr>
<td>FP Sampling [27]</td>
<td>0.7058</td>
<td>0.8539</td>
<td>0.7268</td>
<td>0.5828</td>
<td>0.7654</td>
<td>0.7977</td>
<td>0.8306</td>
<td>0.1396</td>
<td>0.4226</td>
<td>0.8212</td>
<td>0.6646</td>
</tr>
<tr>
<td>SC MCD [29]</td>
<td>0.7496</td>
<td>0.8560</td>
<td>0.6650</td>
<td>0.6311</td>
<td>0.7637</td>
<td>0.8965</td>
<td>0.9118</td>
<td>0.8843</td>
<td>0.8824</td>
<td>0.9326</td>
<td>0.8173</td>
</tr>
<tr>
<td>uNLC [32]</td>
<td>0.0158</td>
<td>0.1419</td>
<td>0.0178</td>
<td>0.0487</td>
<td>0.0346</td>
<td>0.0570</td>
<td>0.0342</td>
<td>0.0216</td>
<td>0.0031</td>
<td>0.0143</td>
<td>0.0389</td>
</tr>
<tr>
<td>OSVOS [1]</td>
<td>0.3397</td>
<td>0.5344</td>
<td>0.0121</td>
<td>0.1260</td>
<td>0.7033</td>
<td>0.7697</td>
<td>0.5447</td>
<td>0.9696</td>
<td>0.0050</td>
<td>0.1224</td>
<td>0.4127</td>
</tr>
<tr>
<td>CIS [33]</td>
<td>0.0538</td>
<td>0.3036</td>
<td>0.1522</td>
<td>0.4681</td>
<td>0.1180</td>
<td>0.1545</td>
<td>0.0862</td>
<td>0.0581</td>
<td>0.0046</td>
<td>0.0184</td>
<td>0.1418</td>
</tr>
<tr>
<td>BASNet [34]</td>
<td>0.3433</td>
<td>0.9379</td>
<td>0.0205</td>
<td>0.2289</td>
<td>0.2119</td>
<td>0.6039</td>
<td>0.9564</td>
<td>0.9586</td>
<td>0.9439</td>
<td>0.9829</td>
<td>0.6188</td>
</tr>
<tr>
<td>Proposed method</td>
<td>0.7809</td>
<td>0.9600</td>
<td>0.7269</td>
<td>0.7065</td>
<td>0.9037</td>
<td>0.9032</td>
<td>0.8700</td>
<td>0.9080</td>
<td>0.9793</td>
<td>0.8546</td>
<td></td>
</tr>
</tbody>
</table>

(a) Results of the skating sequence.
Experiment results

(a) Results of the fence sequence.

(b) Results of the woman sequence.
Experiment results

• Measure from video object segmentation

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean $\mathcal{J}$</th>
<th>Recall $\mathcal{J}$</th>
<th>Mean $\mathcal{F}$</th>
<th>Recall $\mathcal{F}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViBe* [6]</td>
<td>0.2095</td>
<td>0.1364</td>
<td>0.1717</td>
<td>0.0773</td>
</tr>
<tr>
<td>FIC* [8]</td>
<td>0.1701</td>
<td>0.0607</td>
<td>0.2256</td>
<td>0.1337</td>
</tr>
<tr>
<td>BMRI-ViBe* [9]</td>
<td>0.1640</td>
<td>0.0553</td>
<td>0.1703</td>
<td>0.0817</td>
</tr>
<tr>
<td>MCD NP [25]</td>
<td>0.2634</td>
<td>0.0580</td>
<td>0.5569</td>
<td>0.7090</td>
</tr>
<tr>
<td>MCD 5.8ms [26]</td>
<td>0.3736</td>
<td>0.3756</td>
<td>0.5427</td>
<td>0.6100</td>
</tr>
<tr>
<td>Stochastic approx [28]</td>
<td>0.3398</td>
<td>0.3789</td>
<td>0.4003</td>
<td>0.4245</td>
</tr>
<tr>
<td>FP Sampling [27]</td>
<td>0.4294</td>
<td>0.5009</td>
<td>0.6031</td>
<td>0.7156</td>
</tr>
<tr>
<td>SC MCD [29]</td>
<td>0.5213</td>
<td>0.5952</td>
<td>0.7021</td>
<td>0.8200</td>
</tr>
<tr>
<td>uNLC [32]</td>
<td>0.1073</td>
<td>0.1002</td>
<td>0.1416</td>
<td>0.1181</td>
</tr>
<tr>
<td>OSVOS [1]</td>
<td>0.2547</td>
<td>0.2259</td>
<td>0.4129</td>
<td>0.3068</td>
</tr>
<tr>
<td>CIS [33]</td>
<td>0.1583</td>
<td>0.0591</td>
<td>0.2356</td>
<td>0.1253</td>
</tr>
<tr>
<td>BASNet [34]</td>
<td>0.5540</td>
<td>0.6204</td>
<td>0.6696</td>
<td>0.6880</td>
</tr>
<tr>
<td>Proposed method</td>
<td><strong>0.5603</strong></td>
<td><strong>0.6541</strong></td>
<td><strong>0.7214</strong></td>
<td><strong>0.8378</strong></td>
</tr>
</tbody>
</table>

$\mathcal{J}$: region-based segmentation similarity; $\mathcal{F}$: contour-based accuracy

• Synergy effect of two backgrounds

<table>
<thead>
<tr>
<th>Method</th>
<th>precision</th>
<th>recall</th>
<th>$F$-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>coarse BG model</td>
<td>0.9084</td>
<td>0.7655</td>
<td>0.8248</td>
</tr>
<tr>
<td>fine BG model</td>
<td>0.5669</td>
<td>0.7833</td>
<td>0.6095</td>
</tr>
<tr>
<td>combined model</td>
<td>0.9286</td>
<td>0.8041</td>
<td>0.8546</td>
</tr>
</tbody>
</table>

• Computations

<table>
<thead>
<tr>
<th>Module</th>
<th>Time (millisecond)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion estimation</td>
<td>2.207</td>
</tr>
<tr>
<td>Motion compensation</td>
<td>5.117</td>
</tr>
<tr>
<td>Age map update</td>
<td>0.595</td>
</tr>
<tr>
<td>Background model update</td>
<td>13.902</td>
</tr>
<tr>
<td>Foreground combining</td>
<td>0.671</td>
</tr>
<tr>
<td>Total</td>
<td>22.492</td>
</tr>
</tbody>
</table>

CPU only, 320 x 240, 45.5fps
Experiment results

• Combined with supervised method (AMNet*)

<table>
<thead>
<tr>
<th>Method</th>
<th>$F$-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMNet [37] using MCD 5.8ms [26]</td>
<td>0.8789</td>
</tr>
<tr>
<td>AMNet [37] using SC MCD [29]</td>
<td>0.9175</td>
</tr>
<tr>
<td>AMNet [37] using Proposed BG</td>
<td>0.9529</td>
</tr>
</tbody>
</table>

• Robustness test to image noise

Performance change according to image noise intensity: F-measure, Mean J, Mean F

(a) Input  (b) SC MCD  (c) Proposed

Conclusion

• Moving Object detection in PTZ Camera
  • Find moving object region in an unsupervised manner
  • Combine the characteristics of two background models
  • Fine background: reduce foreground loss
  • Robust to image noise and can combine the supervised method
  • Real-Time operation without GPU
  • Suitable for pre-processing and surveillance application

• Future work
  • Combine the powerful appearance model such as salient object detector
  • Extend the method to video object segmentation or video inpainting