Correlation-based ConvNet for Small Object Detection in Videos

Brais Bosquet and Manuel Mucientes and Víctor M. Brea

25th International Conference on Pattern Recognition (ICPR)
10 - 15 January 2021
Small Object Detection

- Objects **without definitive visual cues**
  - Sizes **under 16 × 16 in pixel area**

- **Accuracy lags behind** that of larger objects

- Critical in several **applications**:
  - Automated vehicle systems or sense and avoid problems
  - Satellite image analysis
  - Medical imaging, etc.
Spatio-temporal STDnet (STDnet-ST)

- STDnet-ST ConvNet
  - Processes two frames
  - Computes detections and correlations

- STDnet-ST tubelet linking
  - Compose high-quality tubelets
    1. Correlation-based tubelet linking
      - Links object detections over time by finding the most likely sequences
      - Based on the Viterbi algorithm
    2. Tubelet suppression algorithm
      - Identify and remove incorrect data associations
STDnet-ST ConvNet Architecture

- **Two STDnet** branches
- **Correlation** module
  - RCN regions pairs
  - Propagate correlation score to final detections

- **Outputs:**
  - **Detections** $D_t$ and $D_{t-1}$
  - **Correlation** scores
    - $D_t$ and $D_{t-1}$
    - *Free RCN regions*
STDnet-ST ConvNet Architecture

Region Context Network (RCN)\(^1\)

- A fully convolutional network
- Scans a **shallow feature map**
- Fixed-size promising regions

**RoI Collection Layer** (RCL)
- Unify the feature map parts selected by the RCN
- Compose a **disjoint single feature map**

86.5% saved memory: \[
RCL_{\text{output size}} = \left(\frac{r_h n + p_d(n - 1)}{\text{width}}\right) \times \frac{r_h}{\text{height}}
\]

STDnet-ST tubelet linking

Correlation-based linking

- Update the confidences of the current detections using previous frames
- Baseline linking drawbacks:
  - IoU is limited for object linking when small objects
  - Viterbi algorithm generates every possible tubelets

1. Compute score matrix for pair of frames
   - Replaces IoU with the correlation score
   - $s_i^j = p_{i-1}^i + p_t^j + \lambda \cdot c_t^{ij}$
   - $c_t^{ij} = \rho(r_{i-1}^k, r_t^j)$

2. Generate tubelets using the Viterbi algorithm
   - $p_{i(\hat{v})}^{i(\hat{v})} = \frac{1}{\tau} \sum_{i=1}^{\tau} p_t^{i(\hat{v})}$

3. Update confidences using confidence variability
   - Maximum: low variability -> true positives
   - Average: high variability -> some false positives
   - $p_{i(\hat{v})}^{i(\hat{v})} = \max_{i=1}^{\tau} p_t^{i(\hat{v})}$ if $\sigma\{p_t^{i(\hat{v})}\}_{i=1}^{\tau} \leq \kappa$
   - $\frac{1}{\tau} \sum_{i=1}^{\tau} p_t^{i(\hat{v})}$ otherwise
STDnet-ST tubelet linking

**Tubelet suppression**

- **The goal** is to avoid generating unlikely tubelets
  - Remaining detections

- **Dummy nodes** using free RCN regions
  - Aims to associate remaining detections
  - Higher level of abstraction

\[
 s_t^{ij} = p_{t-1}^i + p_t^j + \lambda \cdot c_t^{ij}
\]

- **Tubelets** with dummy nodes are **discarded**
Experiments

Datasets and state-of-the-art

- State-of-the-art approaches:
  - **FPN-based**: FPN-t and Cascade-FPN-t
  - **Spatio-temporal**: FGFA, RDN and MEGA

- Databases:
  - **UAVDT**
    - More than 76,000 extremely small objects
  - **VisDrone2019-VID**
    - More than 27,000 extremely small objects
  - **USC-GRAD-STDdb**
    - More than 56,000 extremely small objects

---

Experiments

Ablation study

- **Temporal information: 2.5% AP@[5:.95]**
  - Baseline linking: **1.2% AP@[5:.95]**
  - Correlation-based linking + confidence variability: **0.5% AP@[5:.95]**
  - Tubelet suppression procedure: **0.8% AP@[5:.95]**

- **Correlation module importance:**
  - Improve the IoU-based baseline **without spatial information**
  - Allows to build the **tubelet suppression** algorithm
    - Higher level of abstraction

|                     | Baseline linking | Confidence variability | Correlation linking | Tubelet suppression | AP@[5:.95] | AP@[.5]
|---------------------|------------------|------------------------|--------------------|---------------------|-----------|-------
| —                   | 18.9             | 59.1                   |                    |                     |           |       
| ✓                   |                  |                        | ✓                  |                     | 20.1      | 61.4  
| ✓                   | ✓                |                        |                    |                     | 20.3      | 61.8  
| ✓                   | ✓                | ✓                      |                    |                     | 20.4      | 61.6  
| ✓                   | ✓                | ✓                      | ✓                  |                     | 20.6      | 62.0  
| ✓                   | ✓                | ✓                      | ✓                  | ✓                   | 20.9      | 62.6  
| ✓                   | ✓                | ✓                      | ✓                  | ✓                   | **21.4**  | **63.4** |
## Experiments

### UAVDT

<table>
<thead>
<tr>
<th>Method</th>
<th>AP&lt;sub&gt;10[.5,.95]&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;5&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN [27]</td>
<td>6.6</td>
<td>26.0</td>
</tr>
<tr>
<td>R-FCN [27]</td>
<td>9.2</td>
<td>32.5</td>
</tr>
<tr>
<td>RON [27]</td>
<td>3.7</td>
<td>19.7</td>
</tr>
<tr>
<td>SSD [27]</td>
<td>6.0</td>
<td>23.5</td>
</tr>
<tr>
<td>FGFA [154]</td>
<td>6.3</td>
<td>20.7</td>
</tr>
<tr>
<td>RDN [23]</td>
<td>9.3</td>
<td>27.9</td>
</tr>
<tr>
<td>MEGA [16]</td>
<td>9.2</td>
<td>26.6</td>
</tr>
<tr>
<td>FPN [78]</td>
<td>11.8</td>
<td>29.7</td>
</tr>
<tr>
<td>FPN-t</td>
<td>12.0</td>
<td>30.3</td>
</tr>
<tr>
<td>Cascade-FPN-t</td>
<td>12.3</td>
<td>31.2</td>
</tr>
<tr>
<td>STDnet [9]</td>
<td>12.5</td>
<td>35.1</td>
</tr>
<tr>
<td>STDnet++</td>
<td>12.6</td>
<td>35.4</td>
</tr>
<tr>
<td>STDnet-ST</td>
<td>13.1</td>
<td>36.0</td>
</tr>
<tr>
<td>STDnet-ST++</td>
<td><strong>13.3</strong></td>
<td><strong>36.4</strong></td>
</tr>
</tbody>
</table>

### VisDrone2019- VID

<table>
<thead>
<tr>
<th>Method</th>
<th>AP&lt;sub&gt;10[.5,.95]&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;5&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGFA [154]</td>
<td>3.8</td>
<td>16.8</td>
</tr>
<tr>
<td>RDN [23]</td>
<td>4.7</td>
<td>20.7</td>
</tr>
<tr>
<td>MEGA [16]</td>
<td>4.8</td>
<td>21.0</td>
</tr>
<tr>
<td>FPN [78]</td>
<td>6.2</td>
<td>19.9</td>
</tr>
<tr>
<td>FPN-t</td>
<td>6.3</td>
<td>20.2</td>
</tr>
<tr>
<td>Cascade-FPN-t</td>
<td>6.2</td>
<td>20.4</td>
</tr>
<tr>
<td>STDnet [9]</td>
<td>7.2</td>
<td>21.4</td>
</tr>
<tr>
<td>STDnet++</td>
<td>7.3</td>
<td>22.0</td>
</tr>
<tr>
<td>STDnet-ST</td>
<td>7.5</td>
<td>21.9</td>
</tr>
<tr>
<td>STDnet-ST++</td>
<td><strong>7.5</strong></td>
<td><strong>22.4</strong></td>
</tr>
</tbody>
</table>

### USC-GRAD-STDdb

<table>
<thead>
<tr>
<th>Method</th>
<th>AP&lt;sub&gt;10[.5,.95]&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;5&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGFA [154]</td>
<td>11.7</td>
<td>37.5</td>
</tr>
<tr>
<td>RDN [23]</td>
<td>15.5</td>
<td>48.6</td>
</tr>
<tr>
<td>MEGA [16]</td>
<td>17.4</td>
<td>53.1</td>
</tr>
<tr>
<td>FPN [78]</td>
<td>17.3</td>
<td>54.5</td>
</tr>
<tr>
<td>Cascade-FPN [13]</td>
<td>17.4</td>
<td>55.9</td>
</tr>
<tr>
<td>FPN-t</td>
<td>18.7</td>
<td>57.2</td>
</tr>
<tr>
<td>Cascade-FPN-t</td>
<td>19.1</td>
<td>58.9</td>
</tr>
<tr>
<td>STDnet [9]</td>
<td>18.3</td>
<td>57.8</td>
</tr>
<tr>
<td>STDnet++</td>
<td>18.9</td>
<td>59.1</td>
</tr>
<tr>
<td>STDnet-ST</td>
<td>20.1</td>
<td>62.1</td>
</tr>
<tr>
<td>STDnet-ST++</td>
<td><strong>21.4</strong></td>
<td><strong>63.4</strong></td>
</tr>
</tbody>
</table>

- STDnet-ST achieves state-of-the-art results:
  - Outperforms spatial **STDnet**
  - Outperforms **FPN-based** approaches
  - Outperforms **spatio-temporal** approaches
Correlation-based ConvNet for Small Object Detection in Videos

Brais Bosquet and Manuel Mucientes and Víctor M. Brea

25th International Conference on Pattern Recognition (ICPR)
10 - 15 January 2021