Attentional Convolutional Neural Networks for Object Tracking

Xiangdong Kong
(kongxiangdong@buaa.edu.cn)
Prof. Xianbin Cao
Beihang University, Beijing, China
National Key Lab of ATM/CNS

Outline

1. Motivation
2. Problem Analysis
3. Related Work
4. Proposed Model
5. Results and Conclusion
6. Future Work
Outline

• 1. Motivation
• 2. Problem Analysis
• 3. Related Work
• 4. Proposed Model
• 5. Results and Conclusion
• 6. Future Work
Motivation

- As low-altitude airspace opens up, aeronautical surveillance based Unmanned Aerial Vehicle (UAV) has started to be widely used in the aeronautical surveillance.

- Visual object tracking plays an important role in aeronautical surveillance for its accuracy and timeliness. More specifically, UAV-based surveillance systems could track random objects of interest in specified regions.

- We will focus on attentional convolutional neural works (ACNN) which work in conjunction with offline training and online learning for object tracking.
Motivation
Outline

1. Motivation
2. Problem Analysis
3. Single Transformation Model
4. Proposed Model
5. Results and Conclusion
6. Future Work
Problem Analysis

Difficulties in object tracking:

- The kind of object would be unknown before tracking, which means we do not have any prior knowledge about the object to track.
- Visual tracking has to deal with huge variations caused by deformation, illumination variation, blur, rotation, scale variation, fast motion, occlusion, etc.
- The object tracked is usually interfered by similar objects around it.
- Limited training data.
Problem Analysis

Available supplement information for visual object tracking:

- Training datasets of detection and classification are very huge.
- Existing changes are not obvious in a short-term period of time.
- Deep features extracted by CNNs are more conducive to distinguish different targets than traditional features.

Solution:

- We can make use of sufficient training datasets on detection and classification to increase prior knowledge.
- The model remains the same in most situations, and requires only dynamic adjustments in the event of dramatic changes.
- Using deep features instead of traditional features.
Outline

1. Motivation
2. Problem Analysis
3. Related Work
4. Proposed Model
5. Results and Conclusion
6. Future Work
Related Work

Generative algorithms

- Intro: modelling the object in the current frame and regarding the most similar area as the tracked object in the next frame.
- Representative: Kalman filtering, sparse representation, mean-shift.
- Weakness: the background information cannot be used when modelling.
Related Work

Discriminative algorithms

- Intro: Using both object’s appearance information and background information as training data to distinguish the target and the background.
- Category I representatives: Struck, TLD.
- Weakness: These trackers fail to construct effective representation in some special cases.

- Category II representatives: discriminative correlation filters (DCF).
- Weakness: "false" samples; boundary effects; inadequate shallow features.
Related Work

Tracking by deep learning

- Intro: Extracting effective features by Convolutional Neural Networks
- Representatives: SiamFC, CFNet, C-COT
- Weakness: without online adjustment for “pure” deep learning; suffering from boundary effects for deep learning united DCF.
Outline

• 1. Motivation
• 2. Problem Analysis
• 3. Related Work
• 4. Proposed Model
• 5. Results and Conclusion
• 6. Future Work
Proposed Method

Difficulties & Solutions:

- Little prior knowledge —— Offline training by tracking and other datasets.
- Huge variations of object and backgrounds —— Online adjustment.
- Confusions caused by similar objects —— Deep features.

Our method:

we propose a novel architecture called attentional convolutional neural networks (ACNN) in conjunction with offline training and online learning for object tracking. Our method is the improvement of MDNet, so we treat MDNet as our baseline.
Proposed Method

The architecture of ACNN

Conv Layers: 128*51*51, 256*11*11, 512*3*3

Positive Samples

Input: 3*108*108

Negative Samples

Attention Block

FC Layers: 512, 512

Sequence-1

Sequence-2

Sequence-n
Proposed Method

The architecture of attention block

- Input
- Global pooling
- FC
- FC
- Sigmoid
- Output
Proposed Method

Training stage:

- We use a pretrained VGG network on ImageNet for initialization.
- Several positive and negative samples are generated surrounding by ground truth in every frame.
- We only update the trunk and only one branch matching the sequence in every iteration.
- Repeat above procedures until iteration through all epochs or the precision is greater than the specified threshold.
Proposed Method

Tracking stage:

Input: 3*108*108

Conv Layers: 128*51*51, 256*11*11, 512*3*3

FC Layers: 512, 512

Positive Samples

Conv Layers: 128*51*51, 256*11*11, 512*3*3

Attention Block

Negative Samples

Sequence-1

Sequence-new

Sequence-n

Input: 3*108*108
Proposed Method

Tracking stage:

- Removing all branches, and adding a new fc (Fully connected) layer.
- Initializing the new added layer and bounding box regression.
- Start tracking:
  
  Generating samples around the location in the last frame.
  
  Treating the sample owing the max probability as the tracked object.
  
  If max probability < given threshold (0.8)
    
    undating all fc layer using the set M

    Meanwhile, undating all fc layer using the set N every 50 frames.
    
    Set **M**: positive and negative samples collected from adjacent 10 frames.
    
    Set **N**: only positive samples collected from adjacent 50 frames.

Hard negative mining
Outline

• 1. Motivation
• 2. Problem Analysis
• 3. Related Work
• 4. Proposed Model
• 5. Results and Conclusion
• 6. Future Work
Results and Conclusion

Datasets

Results and Conclusion

Evaluation metrics

- **Precision**: which represents the percentages of frames whose estimated location is within the given threshold distance (20 pixels) of the ground truth.

- **Success rates**: which counts the number of successful frames whose overlap $C$ is larger than the given threshold (0.5).

$$C = \frac{|A \cap B|}{|A \cup B|}$$

- **One-pass evaluation (OPE)**, which initializes the tracker from the ground truth position in the first frame and report the average precision or success rate.
Results and Conclusion

Baseline (MDNet) comparison and state-of-art comparison

Precision and Success Plots on OTB2013
Results and Conclusion

Performance on 11 tracking challenges
Results and Conclusion

Performance on 11 tracking challenges
ACNN has advantages over the baseline on five challenges (faster motion, background clutter, in-plane rotation, occlusions, out-of-plane rotation), perform consistently with the baseline on four challenges, and performs slightly poorer than the baseline only in two aspects.
Results and Conclusion

ACNN
Baseline
VID
Struck
VTS
CXT
LSK
## Results and Conclusion

### Influences of Network Layers

<table>
<thead>
<tr>
<th>Numbers</th>
<th>OTB2013</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>AUC</td>
</tr>
<tr>
<td>1</td>
<td>0.703</td>
<td>0.505</td>
</tr>
<tr>
<td>2</td>
<td>0.874</td>
<td>0.662</td>
</tr>
<tr>
<td>3</td>
<td>0.915</td>
<td>0.692</td>
</tr>
<tr>
<td>4</td>
<td>0.832</td>
<td>0.597</td>
</tr>
<tr>
<td>5</td>
<td>0.790</td>
<td>0.576</td>
</tr>
</tbody>
</table>
Results and Conclusion

Conclusions

- ACNN extracts spatial features, and strengthens relationships between channels.
- ACNN equipped with characteristics of offline training and online learning performs outstandingly against state-of-the-art methods.
- We have studied relations between the performance of the tracker and the number of layers. The moderate network achieves a good balance.
Outline

• 1. Motivation
• 2. Problem Analysis
• 3. Related Work
• 4. Proposed Model
• 5. Results and Conclusion
• 6. Future Works
Future Works

- We aim to add more useful attention blocks to extract more discriminative features.
- We are going to speed up the tracking process.
- We will re-initialize the tracker when tracking failure.
Thanks!

Q & A