
Guillaume VAUDAUX-RUTH\textsuperscript{1,2}, Adrien CHAN-HON-TONG\textsuperscript{1,3}, Catherine\textsuperscript{2} ACHARD

\textsuperscript{1}ONERA \quad \textsuperscript{2}Sorbonne Université \quad \textsuperscript{3}Université Paris-Saclay
Problem setup

Action Spotting
Problem setup

Action Spotting

Time (s)

20.0s 43.2s 80.2s 90.9s 116.6s 134.8s
Problem setup

Action Spotting
Problem setup

Action Spotting

Time (s)
Problem setup

Action Spotting

$\text{t} = 26.4\text{s}$
Problem setup

Action Spotting

- $t = 26.4s$
- $p = 0.59$
Problem setup

Action Spotting

Time (s)

$\tau = 26.4s$
$p = 0.59$
Clean & Jerk
Problem setup

Action Spotting

\[ t = 26.4 \text{s} \]
\[ p = 0.59 \]
Clean & Jerk
Problem setup

Action Spotting

\[ t = 26.4s \]
\[ p = 0.59 \]
Clean & Jerk
Problem setup

Action Spotting

$t = 26.4s$
$p = 0.59$

Clean & Jerk
Problem setup

Action Spotting

**Clean & Jerk**

$t = 26.4s$
$p = 0.59$

$t = 83.7s$
$p = 0.96$

Clean & Jerk
**Problem setup**

**Action Spotting**

- **Clean & Jerk**
  - $t = 26.4s$
  - $p = 0.59$

- **Clean & Jerk**
  - $t = 83.7s$
  - $p = 0.96$
Problem setup

Action Spotting

$\begin{align*}
\text{Clean & Jerk} & \quad t = 26.4\text{s} \\
p & = 0.59 \\
\text{Clean & Jerk} & \quad t = 83.7\text{s} \\
p & = 0.96
\end{align*}$
Problem setup

Action Spotting

- **Clean & Jerk**
  - $t = 26.4s$
  - $p = 0.59$

- **Clean & Jerk**
  - $t = 83.7s$
  - $p = 0.96$

- **Clean & Jerk**
  - $t = 120.1s$
  - $p = 0.45$
Problem setup

Action Spotting

- **t = 26.4s, p = 0.59**
  - Clean & Jerk

- **t = 83.7s, p = 0.96**
  - Clean & Jerk

- **t = 120.1s, p = 0.45**
  - Clean & Jerk
State-of-the-Art

Alwassel et al., ECCV 2018
State-of-the-Art

Requires a lot of human acquisitions
Asymmetrical problem
ActionSpotter: Approach overview

Asymmetrical problem

Use of reinforcement learning
ActionSpotter: Approach overview

Asymmetrical problem

Use of reinforcement learning

Only need action detection ground-truth
Asymmetrical problem
Use of reinforcement learning
Only need action detection ground-truth
Extracting spot frames while observing as few frames as possible
ActionSpotter: Approach overview

- Asymmetrical problem
- Use of reinforcement learning
- Only need action detection ground-truth
- Extracting spot frames while observing as few frames as possible
- End-to-End training
ActionSpotter: Approach overview

- Asymmetrical problem
- Use of reinforcement learning
- Only need action detection ground-truth
- Extracting spot frames while observing as few frames as possible
- End-to-End training
- Increase efficiency on action spotting task
ActionSpotter: RL Framework
ActionSpotter: RL Framework
ActionSpotter: RL Framework

Javelin Throw

\[ p_n \quad l_n \quad \tau_{n+1} \]

\[ \text{crit} \quad CL \quad SF \quad BROW \]

\[ h_{n-1} \quad f_n \quad h_n \]

GRU

CNN Backbone

\[ \text{GRU} \quad CNN \text{ Backbone} \]

\[ p_{n+1} \quad l_{n+1} \quad \tau_{n+2} \]

\[ \text{crit} \quad CL \quad SF \quad BROW \]

\[ h_{n+1} \]

\[ \pi \]

\[ t \]
ActionSpotter: RL Framework
ActionSpotter: RL Framework
ActionSpotter: RL Framework
ActionSpotter: RL Framework
ActionSpotter: RL Framework
ActionSpotter: RL Framework
ActionSpotter: RL Framework
ActionSpotter: RL Framework
ActionSpotter: RL Framework
Global dynamic: Then, at step $n$, ActionSpotter ($AS$), has the following dynamic:

\[
\begin{align*}
& \quad f_n = BB(v_{\tau(n)}) \\
& h_n = GRU(f_n, h_{n-1}) \\
& l_n = SF(h_n) \\
& p_n = CL(h_n) \\
& \alpha_n = \arg\max_c (p_n, c) \\
& \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{\{\tau_n, l_n, \alpha_n\}\} \\
& \tau_{n+1} = \tau_n + BROW(h_n)
\end{align*}
\]

with $h_{n-1} = GRU(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).
Global dynamic: Then, at step $n$, ActionSpotter (AS), has the following dynamic:

$$
\begin{align*}
    f_n &= BB(v_{\tau(n)}) \\
    h_n &= GRU(f_n, h_{n-1}) \\
    l_n &= SF(h_n) \\
    p_n &= CL(h_n) \\
    \alpha_n &= \arg\max_c (p_n, c) \\
    \mathcal{V}_{n+1} &= \mathcal{V}_n \cup \{ (\tau_n, l_n, \alpha_n) \} \\
    \tau_{n+1} &= \tau_n + BROW(h_n)
\end{align*}
$$

with $h_{n-1} = GRU(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).
Global dynamic: Then, at step $n$, ActionSpotter (AS), has the following dynamic:

$$
\begin{align*}
\mathcal{V}_{n+1} &= \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\
\tau_{n+1} &= \tau_n + BROW(h_n)
\end{align*}
$$

with $h_{n-1} = GRU(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).
Global dynamic: Then, at step $n$, ActionSpotter (AS), has the following dynamic:

$$
\begin{align*}
\text{AS}(v_{\tau(n)}, h_{n-1}) : & \\
& f_n = BB(v_{\tau(n)}) \\
& h_n = \text{GRU}(f_n, h_{n-1}) \\
& l_n = SF(h_n) \\
& p_n = CL(h_n) \\
& \alpha_n = \arg \max (p_n, c) \\
& \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\
& \tau_{n+1} = \tau_n + BROW(h_n)
\end{align*}
$$

with $h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).
Global dynamic: Then, at step $n$, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}) : \begin{cases} f_n = BB(v_{\tau(n)}) \\ h_n = GRU(f_n, h_{n-1}) \\ l_n = SF(h_n) \\ p_n = CL(h_n) \\ \alpha_n = \arg\max (p_n) \\ \tau_{n+1} = \tau_n + BROW(h_n) \end{cases}$$

with $h_{n-1} = GRU(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).
Global dynamic: Then, at step $n$, ActionSpotter (AS), has the following dynamic:

$$\begin{align*}
    f_n &= BB(v_{\tau(n)}) \\
    h_n &= GRU(f_n, h_{n-1}) \\
    l_n &= SF(h_n) \\
    p_n &= CL(h_n) \\
    \alpha_n &= \arg\max_c \{p_{n,c}\} \\
    \mathcal{V}_{n+1} &= \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\
    \tau_{n+1} &= \tau_n + BROW(h_n)
\end{align*}$$

with $h_{n-1} = GRU(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

$$\begin{align*}
    r_{\pi,n} &= mAP(\mathcal{V}_n) - mAP(\mathcal{V}_{n-1}) + \rho \mathcal{H}(\pi(n)) \\
    R_{\pi,n} &= \sum_{k=0}^{N-n-1} \gamma^k r_{\pi,k+n} \\
    \mathcal{L}_{global} &= \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{critic} - \lambda_2 J_{actor} \\
    \nabla J_{actor} &= \nabla \mathbb{E} \left[ \sum_{n=1}^{N} \log(\pi(n))(R_{\pi,n} - \mathbb{E}[R_{\pi,n}|h_n]) \right]
\end{align*}$$
ActionSpotter: RL Framework

**Global dynamic:** Then, at step $n$, ActionSpotter (AS), has the following dynamic:

$$AS(v_{τ(n)}, h_{n-1}) : \begin{cases} f_n = BB(v_{τ(n)}) \\ h_n = GRU(f_n, h_{n-1}) \\ l_n = SF(h_n) \\ p_n = CL(h_n) \\ α_n = \arg \max_{c} \{p_n,c\} \\ V_{n+1} = V_n \cup \{(τ_n, l_n, α_n)\} \\ τ_{n+1} = τ_n + BROW(h_n) \end{cases}$$

with $h_{n-1} = GRU(BB(\{v_{τ(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

$$r_{π,n} = mAP(V_n) - mAP(V_{n-1}) + \rho H(π(n))$$

$$R_{π,n} = \sum_{k=0}^{N-n} γ^k r_{π,k+n}$$

$$L_{global} = L_{cls} + λ_1 L_{critic} - λ_2 J_{actor}$$

$$\nabla J_{actor} = \nabla E \left[ \sum_{n=1}^{N} \log(π(n))(R_{π,n} - E[R_{π,n}|h_n]) \right]$$
ActionSpotter: RL Framework

Global dynamic: Then, at step \( n \), ActionSpotter (AS), has the following dynamic:

\[
\begin{align*}
& f_n = BB(v_{\tau(n)}) \\
& h_n = \text{GRU}(f_n, h_{n-1}) \\
& l_n = \text{SF}(h_n) \\
& p_n = \text{CL}(h_n) \\
& \alpha_n = \arg\max_c (p_{n,c}) \\
& \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\
& \tau_{n+1} = \tau_n + \text{BROW}(h_n)
\end{align*}
\]

with \( h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1})) \), the memory of the past viewed frames (or frame chunks).

\[
r_{\pi,n} = \text{mAP}(\mathcal{V}_n) - \text{mAP}(\mathcal{V}_{n-1}) + \rho\mathcal{H}(\pi(n))
\]

\[
R_{\pi,n} = \sum_{k=0}^{N-n-1} \gamma^k r_{\pi,k+n}
\]

\[
\mathcal{L}_\text{global} = \mathcal{L}_\text{cls} + \lambda_1 \mathcal{L}_\text{critic} - \lambda_2 J_{\text{actor}}
\]

\[
\nabla J_{\text{actor}} = \nabla \mathbb{E} \left[ \sum_{n=1}^{N} \log(\pi(n))(R_{\pi,n} - \mathbb{E}[R_{\pi,n}|h_n]) \right]
\]
**ActionSpotter: RL Framework**

- **Global dynamic:** Then, at step $n$, ActionSpotter (AS), has the following dynamic:

\[
AS(v_{\tau(n)}, h_{n-1}) : \begin{cases} 
  f_n = BB(v_{\tau(n)}) \\
  h_n = GRU(f_n, h_{n-1}) \\
  l_n = SF(h_n) \\
  p_n = CL(h_n) \\
  \alpha_n = \arg\max_c (p_{n,c}) \\
  \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\
  \tau_{n+1} = \tau_n + BROW(h_n)
\end{cases}
\]

with $h_{n-1} = GRU(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

- Reward function:

\[
r_{\pi,n} = \text{mAP}(\mathcal{V}_n) - \text{mAP}(\mathcal{V}_{n-1}) + \rho \mathcal{H}(\pi(n))
\]

\[
R_{\pi,n} = \sum_{k=0}^{N-n-1} \gamma^k r_{\pi,k+n}
\]

- **Loss function:**

\[
\mathcal{L}_{global} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{critic} - \lambda_2 J_{actor}
\]

\[
\nabla J_{actor} = \nabla \mathbb{E} \left[ \sum_{n=1}^{N} \log(\pi(n))(R_{\pi,n} - \mathbb{E}[R_{\pi,n}|h_n]) \right]
\]
Global dynamic: Then, at step $n$, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}):$$

$$\begin{align*}
    f_n &= BB(v_{\tau(n)}) \\
    h_n &= GRU(f_n, h_{n-1}) \\
    l_n &= SF(h_n) \\
    p_n &= CL(h_n) \\
    \alpha_n &= \arg\max_c (p_{n,c}) \\
    \mathcal{V}_{n+1} &= \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\
    \tau_{n+1} &= \tau_n + BROW(h_n)
\end{align*}$$

with $h_{n-1} = GRU(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

$$r_{\pi,n} = mAP(\mathcal{V}_n) - mAP(\mathcal{V}_{n-1}) + \rho \mathcal{H}(\pi(n))$$

$$R_{\pi,n} = \sum_{k=0}^{N-1-n} \gamma^k r_{\pi,k+n}$$

$$\mathcal{L}_{global} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{critic} - \lambda_2 J_{actor}$$

$$\nabla J_{actor} = \nabla \mathbb{E} \left[ \sum_{n=1}^{N} \log(\pi(n))(R_{\pi,n} - \mathbb{E}[R_{\pi,n}|h_n]) \right]$$
Experiments

• **Thumos14**

<table>
<thead>
<tr>
<th>Approach</th>
<th>Detection mAP@</th>
<th>Spotting mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>Glimpses [5]</td>
<td>48.9</td>
<td>44.0</td>
</tr>
<tr>
<td>SMS [30]</td>
<td>51.0</td>
<td>45.2</td>
</tr>
<tr>
<td>M-CNN [31]</td>
<td>47.7</td>
<td>43.5</td>
</tr>
<tr>
<td>CDC [32]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TURN [33]</td>
<td>54.0</td>
<td>50.9</td>
</tr>
<tr>
<td>R-C3D [34]</td>
<td>54.5</td>
<td>51.5</td>
</tr>
<tr>
<td>SSN [35]</td>
<td>66.0</td>
<td>59.4</td>
</tr>
<tr>
<td>A-Search [14]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CBR [36]</td>
<td>60.1</td>
<td>56.7</td>
</tr>
<tr>
<td>BSN + UNet [37]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Re-thinking F-RCNN [38]</td>
<td>59.8</td>
<td>57.1</td>
</tr>
<tr>
<td>D-SSAD [39]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours (TSN backbone)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours (3D backbone)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

• **ActivityNet1.2**

<table>
<thead>
<tr>
<th>Approach</th>
<th>Detection mAP@</th>
<th>Spotting mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
<td>0.75</td>
</tr>
<tr>
<td>W-TALC [45]</td>
<td>37.0</td>
<td>14.6</td>
</tr>
<tr>
<td>SSN-SW [35]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3C-Net [46]</td>
<td>37.2</td>
<td>23.7</td>
</tr>
<tr>
<td>FPTADC [47]</td>
<td>37.6</td>
<td>21.8</td>
</tr>
<tr>
<td>SSN-TAG [35]</td>
<td>39.2</td>
<td>25.3</td>
</tr>
<tr>
<td>BSN [48]</td>
<td>46.5</td>
<td>30.0</td>
</tr>
<tr>
<td>BMN [49]</td>
<td>50.1</td>
<td>34.8</td>
</tr>
<tr>
<td>Ours (TSN backbone)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours (3D backbone)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Conclusion

Key idea: **Use of Reinforcement Learning + End-to-End training**

Key properties:
- Only need action detection ground-truth
- Able to sparsely browse videos
- mAP as training criterion

Guillaume VAUDAUX-RUTH\textsuperscript{1,2}, Adrien CHAN-HON-TONG\textsuperscript{1,3}, Catherine\textsuperscript{2} ACHARD

\textsuperscript{1}ONERA \quad \textsuperscript{2}Sorbonne Université \quad \textsuperscript{3}Université Paris-Saclay