Overview

Event spotting consists in finding the exact timestamp in which an event occurs and to recognize the event type. We propose a modular framework for soccer event spotting enriched with:
- A temporal offset regression branch to predict event temporal location.
- A data sampling and balancing strategy to overcome the inherent frequency unbalance of soccer events and their sparsity during training.
- A masking policy to make the model focus on the most relevant frames of a clip.

RMS-Net

Given a short video clip \( X = (x_1, x_2, \ldots, x_T) \) from a soccer match, our network predicts a probability over action classes \( p \) and a temporal offset \( o \):
- We minimize the cross-entropy loss between predicted event class \( p_i \) and ground truth event \( c \).
- We minimize the squared-error loss between predicted relative offset \( o \) and ground truth relative offset \( r \) (excluding background events).
- At prediction stage: convert relative timestamps to absolute timestamps.

\[
\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{C} y_i \log(p_i)
\]

\[
\mathcal{L}_{\text{reg}} = (o - r)^2
\]

\[
\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{reg}}
\]

Masking Strategy

Since the majority of visual cues that contribute to the recognition of an event occur just after the event [1], we propose a masking function which encourages the network to learn robust features after the event. Our function randomly replaces the frames before an event with a background clip, as follows:

\[
M(p, q, l) = \begin{cases} 
(z_1, \ldots, z_{c-1}, x_{r}, \ldots, x_T) & \text{if } r \leq q, u < p \\
(x_1, \ldots, x_{c-1}, z_{r}, \ldots, x_T) & \text{otherwise.}
\end{cases}
\]

- \( p \) is a fixed masking probability.
- \( q \) is the maximum relative temporal offset in the clip to allow masking.
- \( r \) is the starting absolute timestamp of the video clip.
- \( u \) is the relative timestamp of the event in the clip.
- \( \{z_i\}_{i=1}^{c} \) is a sequence of frames selected from a random background clip.
- \( v \) is a random value sampled from the uniform distribution \([0, 1] \).

Data Sampling and Balancing

- Given an event, extract all clips with length \( T \) containing the event, sliding a window along the time axis with stride 1.
- Slice a window of size \( T \) with stride \( T \) over the remaining parts of the matches, to obtain background clips.
- Balance the number of clips per class.
- During inference, extract and process non-overlapping clips.

RMS-Net: Regression and Masking for Soccer Event Spotting

Matteo Tomei¹, Lorenzo Baraldi¹, Simone Calderara¹
Simone Bronzin², Rita Cucchiara¹
¹University of Modena and Reggio Emilia - name.surname@unimore.it
²Metaliquid - name.surname@meta-liquid.com

SoccerNet Dataset

The SoccerNet dataset [2] provides:
- 500 full broadcast soccer matches (300 train, 100 val, 100 test).
- Annotations of spots belonging to 3 classes (Goal, Card, Substitution).
- One-second resolution annotations.
- Pre-computed ResNet-152 frame features released with the dataset.

Average-mAP: given a tolerance \( \delta \), the AP for each class is computed considering a prediction as positive if the distance from its closest ground truth spot is less than \( \delta \). The mAP is the average of the AP of each class. The Avg-mAP is the area under the mAP curve obtained by varying \( \delta \) from 5 to 60 seconds.

Main Results


<table>
<thead>
<tr>
<th>Model</th>
<th>Clip length (s)</th>
<th>Features</th>
<th>Val-Avg-mAP</th>
<th>Test-Avg-mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoccerNet baseline [2]</td>
<td>5</td>
<td>ResNet-152 (PCA)</td>
<td>-</td>
<td>34.5</td>
</tr>
<tr>
<td>SoccerNet baseline [2]</td>
<td>20</td>
<td>ResNet-152 (PCA)</td>
<td>-</td>
<td>49.7</td>
</tr>
<tr>
<td>Vanderplas et al. [1]</td>
<td>20</td>
<td>ResNet-152 (PCA) + Audio</td>
<td>-</td>
<td>56.7</td>
</tr>
<tr>
<td>Vats et al. [4]</td>
<td>15</td>
<td>ResNet-152 (PCA)</td>
<td>-</td>
<td>60.1</td>
</tr>
<tr>
<td>Ours</td>
<td>120</td>
<td>ResNet-152 (PCA)</td>
<td>-</td>
<td>62.5</td>
</tr>
<tr>
<td>Ours</td>
<td>20</td>
<td>ResNet-152 (PCA)</td>
<td>-</td>
<td>67.8</td>
</tr>
<tr>
<td>Ours</td>
<td>120</td>
<td>ResNet-152 (PCA)</td>
<td>-</td>
<td>68.5</td>
</tr>
</tbody>
</table>

Comparison with other approaches using ResNet-152 features released with SoccerNet.

Ablation Study

mAP when varying the spotting tolerance, with and without the offset regression branch.

References