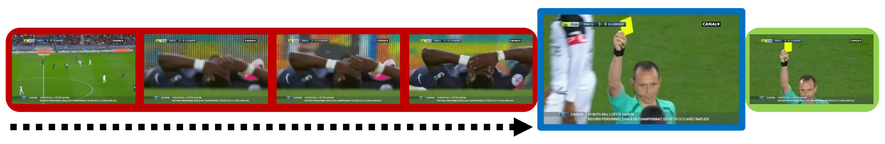
$\mathcal{RMS}$-Net: Regression and Masking for Soccer Event Spotting

Matteo Tomei[1], Lorenzo Baraldi[1], Simone Calderara[1]
Simone Bronzin[2], Rita Cucchiara[1]
[1]University of Modena and Reggio Emilia - name.surname@unimore.it
[2]Metaliquid - name.surname@meta-liquid.com

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:

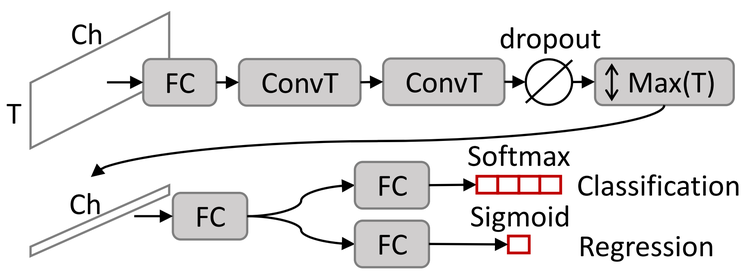
- A temporal offset *regression branch* to predict event temporal location.
- A data *sampling and balancing strategy* to overcome the inherent frequency unbalance of soccer events and their sparsity during training.
- A *masking policy* to make the model focus on the most relevant frames of a clip.



$\mathcal{RMS}$-Net

Given a short video clip $\boldsymbol{X} = (x_1, x_2, ..., x_T)$ 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_e and ground truth event e (including background events).
- 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.



$$\mathcal{L}_{cls} = -\sum_{c=0}^{C} 1_{c=e} \log(p_c)$$

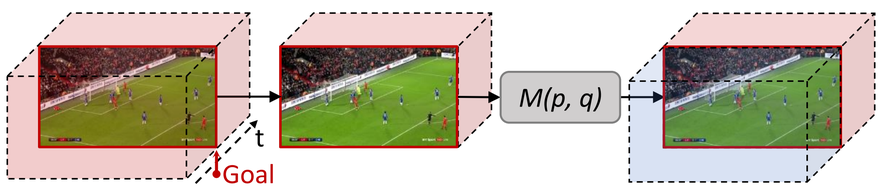
$$\mathcal{L}_{regr} = (\sigma(o) - r)^2$$

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{regr}$$

Masking Strategy

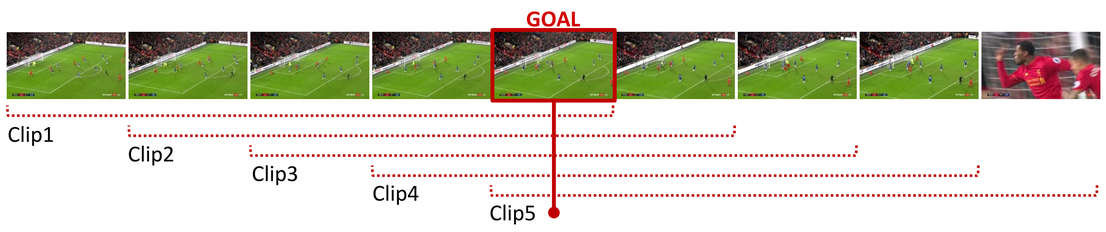
Since the majority of visual cues that contribute to the recognition of an event occur just after the event [1], we propose a masking function which encourages the network to learn robust features after the event. Our function randomly replaces the frames before an event with a background clip, as follows:

$$M(p,q)(\boldsymbol{X}) = \begin{cases} (z_1, ..., z_{t-s-1}, x_{t-s}, ..., x_T) & \text{if } r \leq q, u < p \\ (x_1, ..., x_{t-s-1}, x_{t-s}, ..., x_T) & \text{otherwise,} \end{cases}$$



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

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.

SoccerNet Dataset

The SoccerNet dataset [2] provides:

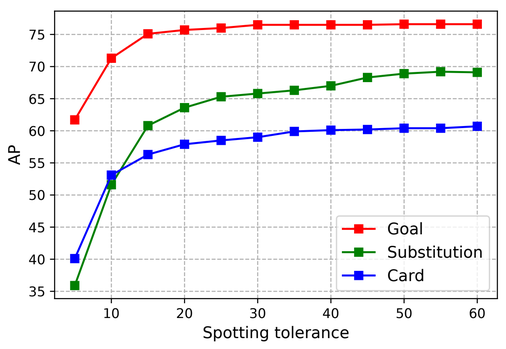
- 500 full broadcast soccer matches (300 train, 100 val, 100 test).
- Annotations of spots belonging to 3 classes (Goal, Card, Substitution)
- One-second resolution annotations
- Pre-computed ResNet-152 frame features released with the dataset.

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

Main Results

Model	Clip length (s)	Features	Val Avg-mAP	Test Avg-mAP
SoccerNet baseline [2]	5	ResNet-152 (PCA)	-	34.5
SoccerNet baseline [2]	60	ResNet-152 (PCA)	-	40.6
SoccerNet baseline [2]	20	ResNet-152 (PCA)	-	49.7
Vanderplaetse *et al.* [3]	20	ResNet-152 (PCA) + Audio	-	56.0
Vats *et al.* [4]	15	ResNet-152 (PCA)	-	60.1
Cioppa *et al.* [1]	120	ResNet-152 (PCA)	-	62.5
Ours	20	ResNet-152 (PCA)	**67.8**	**65.5**

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

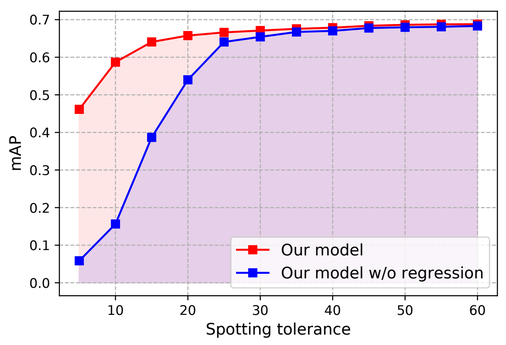


Per-class Average Precision, as a function of spotting tolerance.

Model	Pre-train	Val Avg-mAP	Test Avg-mAP
R18 + Our	ImageNet	73.8	70.9
R50 + Our	ImageNet	76.6	74.9
R152 + Our	ImageNet	**77.5**	**75.1**

Performance when fine-tuning different variants of ResNet with $\mathcal{RMS}$-Net.

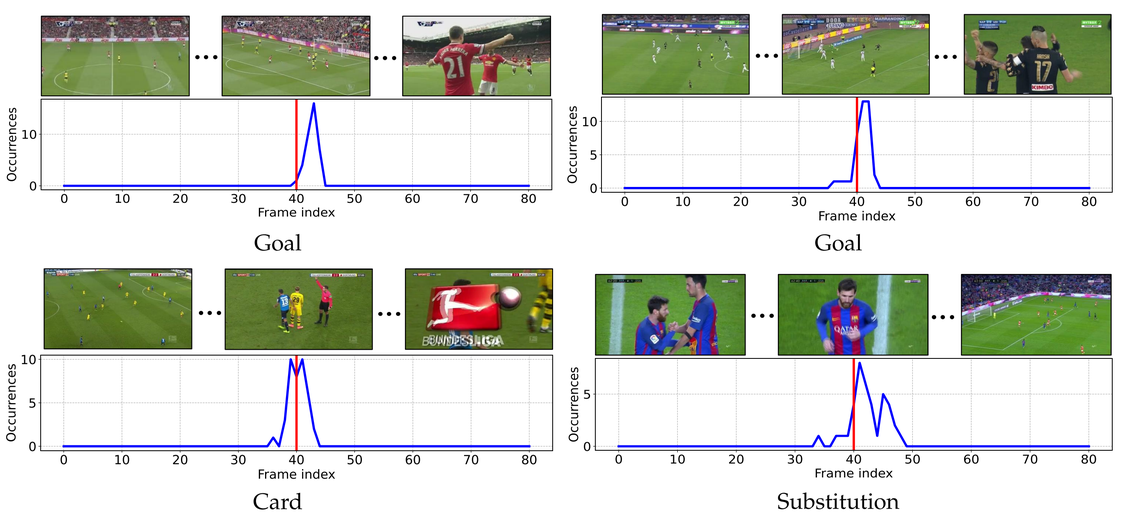
Ablation Study



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

Model	Val Avg-mAP	Test Avg-mAP
Ours	**67.8**	**65.5**
w/o augmented offsets	48.7	46.2
w/o offset regression	58.5	55.7
w/o masking	66.5	64.0

Performance of the proposed model when removing key components.



Qualitative results. The ground truth action timestamp is shown in red, while the blue curve shows the number of times a frame index was predicted as spot, sliding a 41-frames window over a video with 81 frames.

References

[1] A. Cioppa, A. Deliege, S. Giancola, B. Ghanem, M. V. Droogenbroeck, R. Gade, and T. B. Moeslund. A context-aware loss function for action spotting in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[2] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[3] B. Vanderplaetse and S. Dupont. Improved soccer action spotting using both audio and video streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[4] K. Vats, M. Fani, P. Walters, D. A. Clausi, and J. Zelek. Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.