

Literature Review - Project

Paper 1: Jersey Number Recognition using Keyframe Identification from Low-Resolution Broadcast Videos

PURPOSE

Motivation: Current approaches using Spatial Transformer Networks, CNNs, and Vision Transformers struggle with real-world video data, where jersey numbers are not visible in most frames.

→ Sub-problem: identify frames that do contain the jersey number

Goal: propose keyframe identification module that extracts frames containing essential high-level information about the jersey number

Dataset: Soccernet

APPROACH (Pipeline)

1. Keyframe Identification (Kfld) Module:

Identifies instances that capture critical moments: frames containing high-level features for effective JNR

Localizes the jersey number and eliminates outlier jersey detections

→ Robust to occlusions and blurs

Uses: JNL, RoI and Spatial Context Aware filtering (LHC and GHC)

STAGES:

- **JNL: Jersey Number Localization**

off-the-shelf detector: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation

fine-tuned on the dataset

- **RoI: Region of Interest**

To mitigate spurious (erroneous) detections from JNL

Filter the JNL detections: Jersey numbers are always localized in a specific region within a RoI.

RoI is manually preset for each image based on its width and height (corners: $(w/4, h/5)$, $(3w/4, h/2)$)

I^* computed based on the Area of intersection between RoI and detection.

I^* is then compared to a threshold value to determine validity.

- **SCA (Spatial Context Aware) filtering**

Problems to address:

Multiple players visible in the frame

No holistic representation of the jersey number (JNL localizes digits separately)

- Conversion to **HUE** (Hue, Saturation, and Value) color space: to distinguish colours in disruptive conditions
- **LHC** (Local Histogram Correlation): compares detections' correlation scores within each frame of the tracklet. If two frames have high correlation scores, they are likely to correspond to two digits forming a single jersey number → these detections are merged.
- **GHC** (Global Histogram Correlation): cluster histograms of all filtered detections, the one with more detections is chosen (and continues to the next step in the pipeline), the rest are discarded. → to filter only target player.

2. Spatio-temporal neural network:

Extracts spatial and temporal context of the tracklet to identify the jersey number.

STAGES

- Sampling: Avoid processing all filtered keyframes (memory constraints and extraction of redundant features) → sample randomly ensuring any two frames sampled are at least d frames apart from each other (d : fixed number). 40 sequence length.
- ResNet-18 network pretrained on ImageNet → Extract 512 spatial features
- bi-LSTM (bidirectional Long Short-Term Memory) network: input 512 spatial features to capture temporal dynamics in tracklet → Output: 256 temporal features.
- Linear layers with Multi-task loss function: Classify each digit separately (digit-wise classification)

RESULTS

- Drop of 87.65% in the number of frames after Kfld module → Meaning that most of the images in each tracklet does not have a visible jersey number.
- 37.81% and 37.70% increase in accuracies on 2 test sets with domain gaps when adding the Kfld module (for their best-performing model: LSTM)
- A 73.77% challenge accuracy for the model using Kfld and LSTM as the temporal model

geyscale?