Automated Event Detection and Classification in Soccer: The Potential of Using Multiple Modalities

Olav Andre Nergård Rongved, Markus Stige , Steven Alexander Hicks, Vajira Lasantha Thambawita , Cise Midoglu, Evi Zouganeli, Dag Johansen, Michael Alexander Riegler and Pål Halvorsen

Outline

- Introduction
- Tested Models
- Model fusion
- Experiment and results
- Discussion
- Conclusions

Introduction

- The video summaries and highlights from sports games cost too much.

  →**<u>decided to make them automatic</u>**

- One of the key components of this is the detection and classification of significant events in real-time

- The two main purposes
  - to develop an intelligent soccer event detection and classification system using machine learning
  - to evaluate the potential of using multiple modalities (video and audio) for event detection

Tested Models

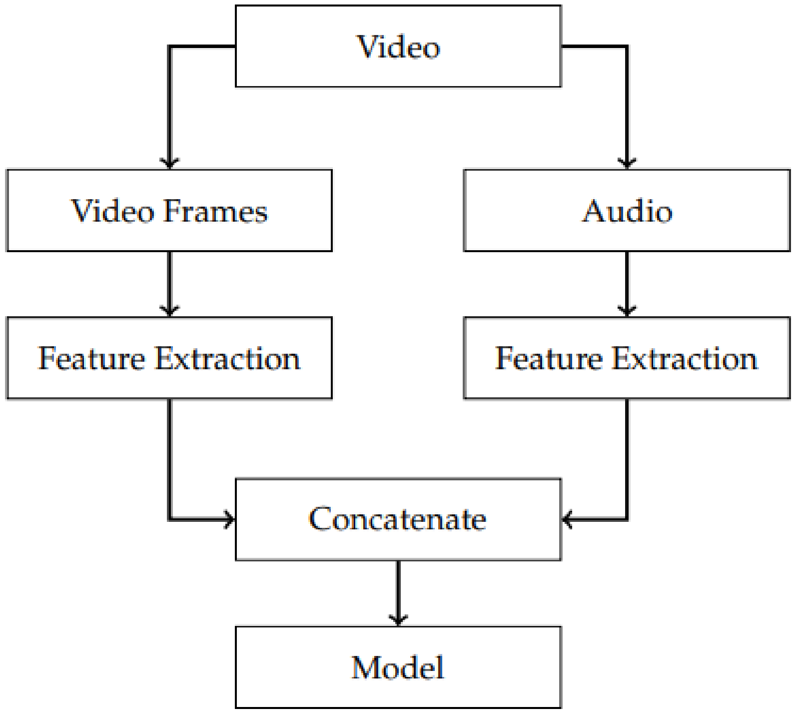
- **Visual Model**(existing models)
  - CALF(called context-aware loss function)
    - The inputs to the model are the ResNet features provided with the SoccerNet dataset.
  - 3D-CNN
    - They used an 18-layered 3D-ResNet on the video frame inputs.
  - 2D-CNN
    - They used a 2D-CNN model that uses the pre-extracted ResNet features provided by SoccerNet.
- **Audio Model**
  - Their audio model is based on transforming the audio into Log-Mel spectrograms.

Model Fusion

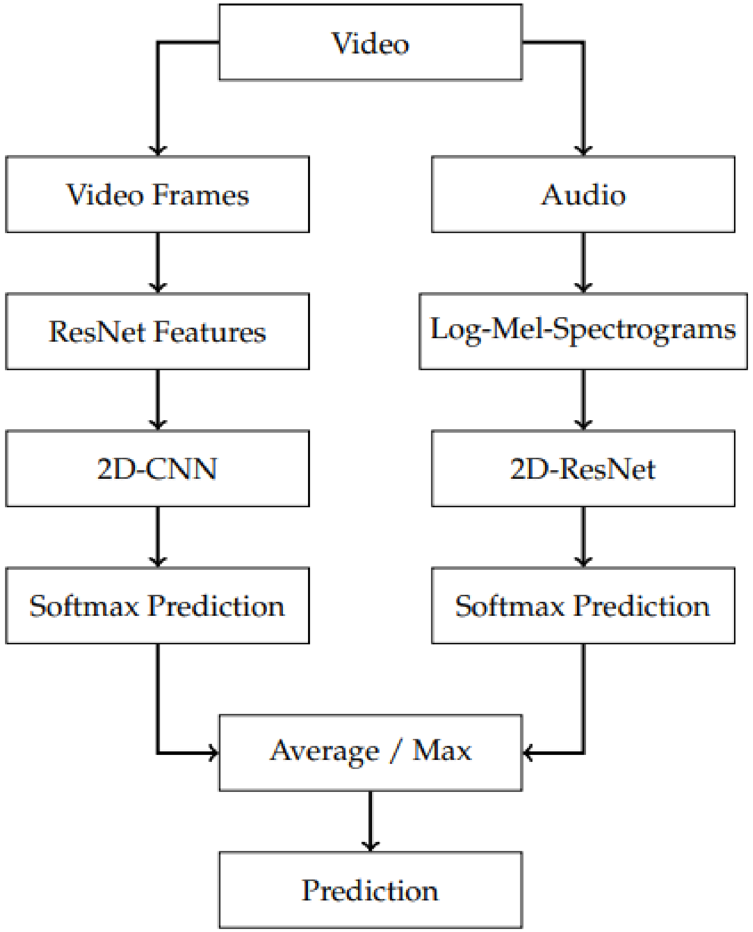
- **Early Fusion**
  - referred to as data-level fusion or input-level fusion, is a traditional way of fusing data before conducting an analysis

- **Late Fusion**
  - referred to as decision-level fusion, data sources are used independently until fusion at a decision-making stage

Model fusion



Early fusion



Late fusion

Experiments and Results

- Dataset
- Training and Implementation Details
- Input Window
  - Window Size
  - Window Position
- Classification Performance
- Spotting Performance

Experiments and Results

- **Dataset (input)**
  - 500 soccer games from 2014 to 2017 with games from six European elite leagues. It has a total duration of 764 h and includes 6637 annotations of the event types **goal**, (yellow/red) **card**, and **substitution**. This gives a frequency of an event happening every 6.9 min on average.
  - They added a **background** class by sampling in between events. If the time distance between two consecutive events is larger than 3 min, then a new background sample is added in the center, such that a background sample will never be within 90 s of another event.

EXAMPLE FRAMES OF EACH EVENT IN DATA SET

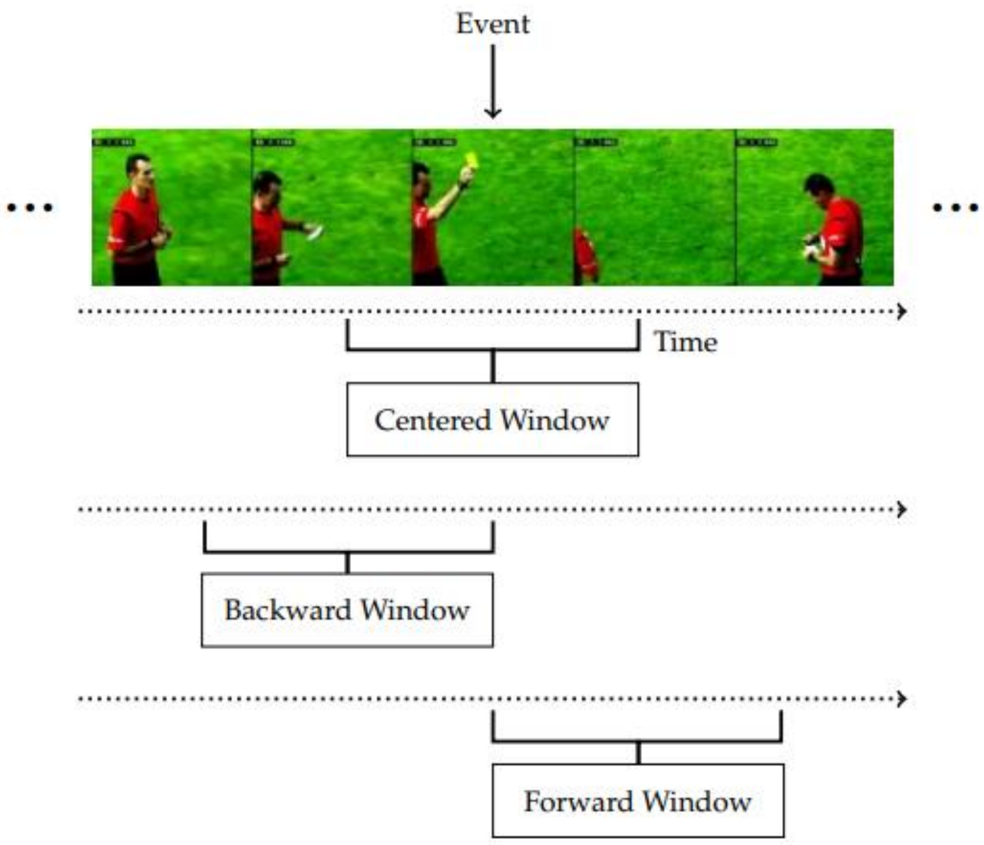
(a) Card.

(b) Substitution.

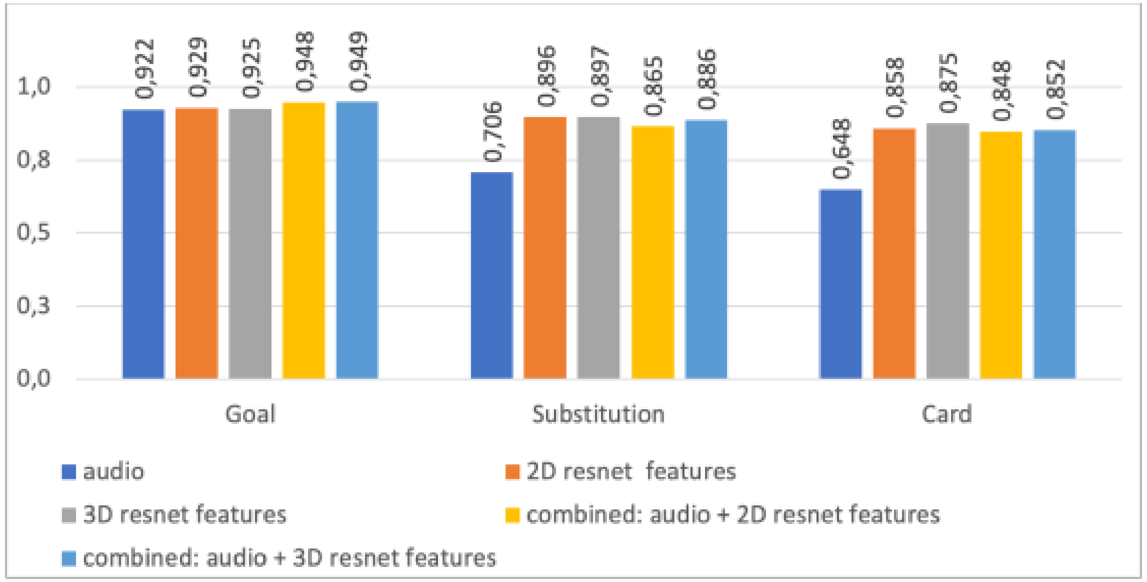
(c) Goal.

Experiments and Results

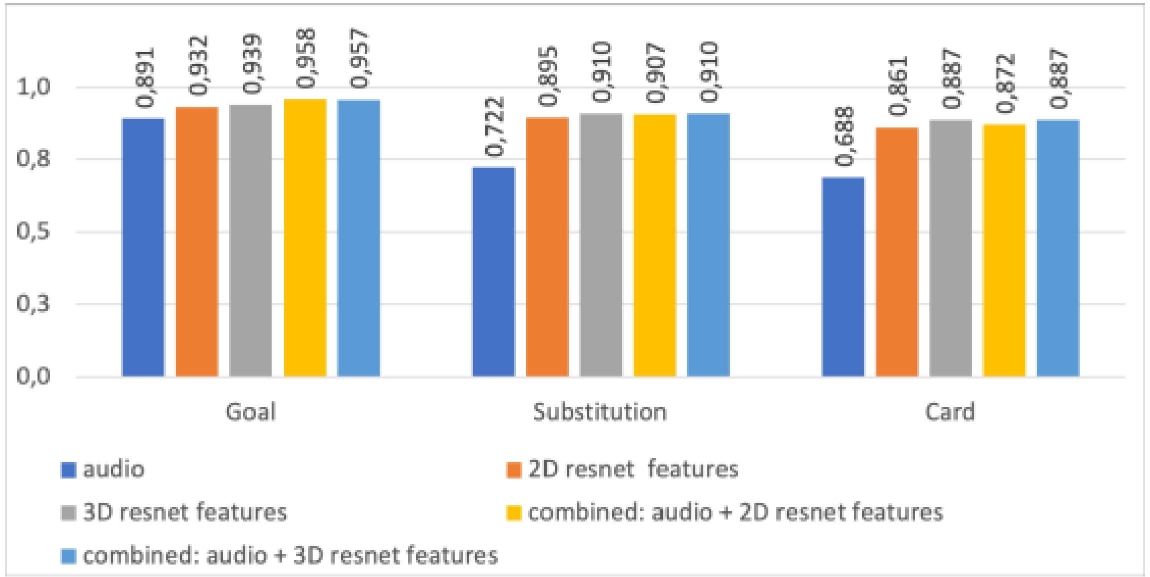
- **Input Window**
  - **Window Size**
    - For classification accuracy, a larger window is better. However, as large windows can also have drawbacks, so they experimented with different windows sizes in this experiment.
  - **Window Position**
    - **Centered window -> is used**
    - Backward window
    - Forward window



- Classification Performance

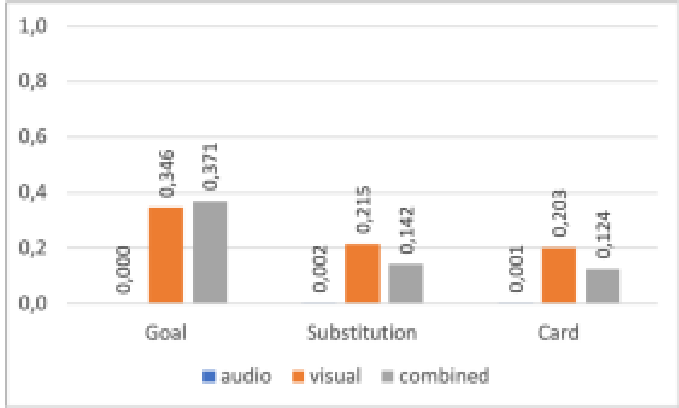


(a) Window size = 8

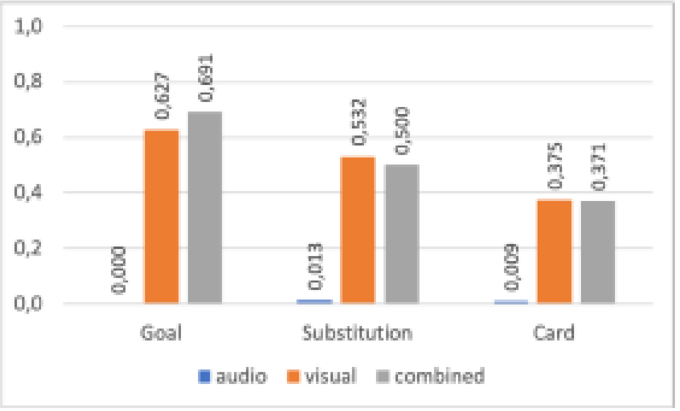


(b) Window size = 16

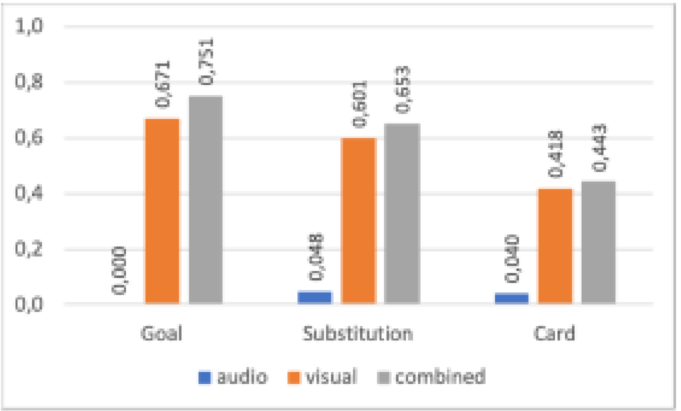
- Detection Performance



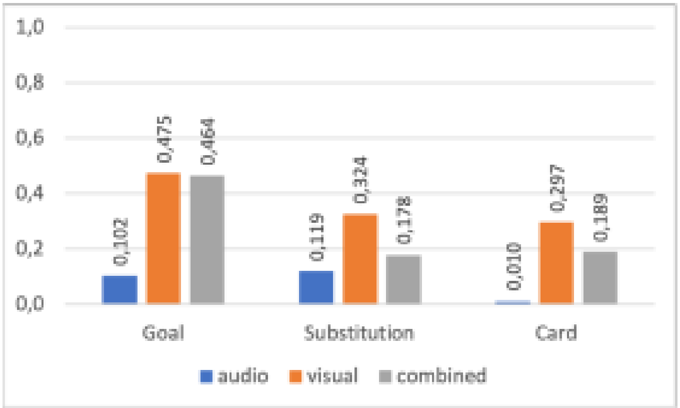
(a) CALF-60-5, tolerance = 5



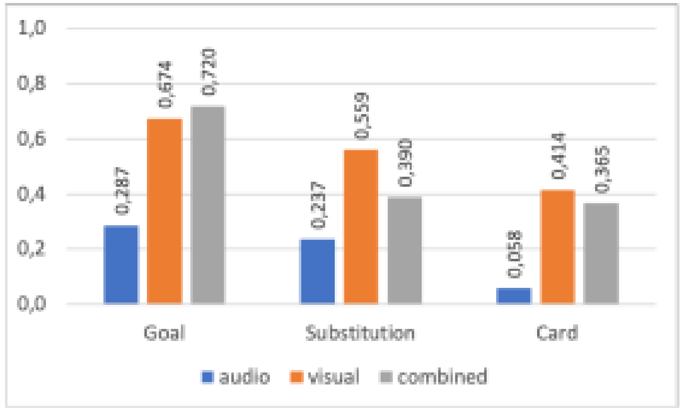
(b) CALF-60-5, tolerance = 20



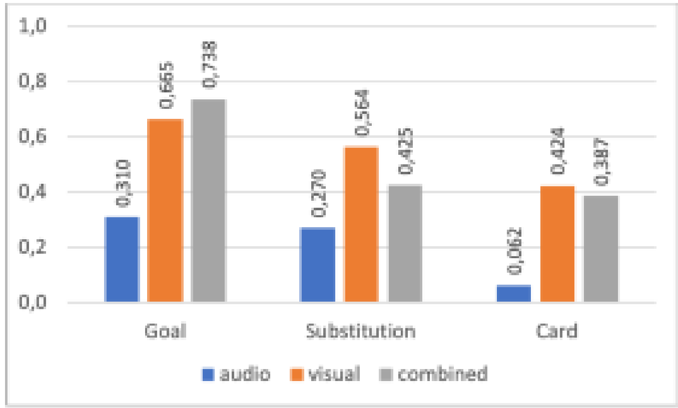
(c) CALF-60-5, tolerance = 60



(d) CALF-60-40, tolerance = 5



(e) CALF-60-40, tolerance = 20



(f) CALF-60-40, tolerance = 60

Discussion

- Experimental results show that the benefit of analyzing audio information alone, or in addition to the visual information, is dependent on the context or the type of event.

- The visual CNN models they have experimented with are meant as examples of state-of-the-art models and they showed a highest AP(average precision) of 84% for goal events in new models which are currently being developed and tested.

Conclusions

- Experimental results demonstrate the potential of using multiple modalities as the performance of detecting events increases in many of the selected configurations when features are combined.
- However, there is a difference in the benefits gained from the multimodal approach with respect to different event types.
  - Ex) the combination of audio and visual features proved more beneficial for the Goal events than for Card and Substitution events.
- In summary, an ML-based event detection component utilizing several available data modalities can be an important component of future intelligent video processing and analysis systems.