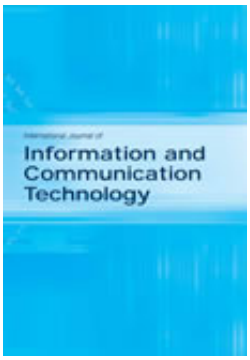
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Semantic event analysis of sports match videos using domain knowledge and deep features

Ruixia Xu

DOI: 10.1504/IJICT.2025.10074945

Article History:

Received:	13 September 2025
Last revised:	14 October 2025
Accepted:	16 October 2025
Published online:	17 December 2025

Semantic event analysis of sports match videos using domain knowledge and deep features

Ruixia Xu

Sports Department,
Henan University of Animal Husbandry and Economy,
Zhengzhou, 450000, China
Email: xrx781008@163.com

Abstract: Semantic event analysis in sports videos faces challenges such as complex actions and high annotation costs. To address these issues, this paper proposes a novel framework that integrates domain knowledge with deep features. The approach first translates sports rules into computable spatio-temporal constraints, then designs a knowledge-injection network to guide deep models toward semantically critical regions. Finally, a knowledge-conditioned attention mechanism is introduced to fuse domain knowledge with visual features effectively. Experimental results on the SoccerNet dataset demonstrate that the proposed method achieves a mean average precision of 71.5%, outperforming strong baselines such as inflated 3D ConvNet and soccer background matting network by 13.3% and 3.6%, respectively. The framework shows significant improvements in detecting complex and sparse events, offering enhanced accuracy, robustness and generalisation capability with reduced reliance on large-scale annotated data.

Keywords: semantic event analysis; domain knowledge; deep features; video understanding; sports videos.

Reference to this paper should be made as follows: Xu, R. (2025) 'Semantic event analysis of sports match videos using domain knowledge and deep features', *Int. J. Information and Communication Technology*, Vol. 26, No. 47, pp.89–105.

Biographical notes: Ruixia Xu graduated from Hainan University with a Bachelor of Law in 2004. From 2009 to 2011, she studied Sports Education and Training at Shanghai Institute of Physical Education, obtaining a Master's degree in Education in 2011. Currently, she works at Henan University of Animal Husbandry and Economy. Her research interests include sports medicine, domain knowledge and sports rehabilitation.

1 Introduction

Sports video content analysis, as a key research focus in computer vision and multimedia analysis, has garnered sustained attention from both academia and industry in recent years due to its extensive practical value. With the continuous expansion of major sporting event broadcasts and the growing demand for intelligent video content processing, automatically and accurately identifying semantically meaningful events-such as goals, fouls, and substitutions-from massive video datasets has become a highly

challenging and significant task. This capability not only enables the automatic generation of highlight reels to enhance spectator experiences but also provides coaches and athletes with deep tactical insights and data-driven match analysis, unlocking substantial commercial value and research potential (Wang and Parameswaran, 2004).

In recent years, thanks to the rapid advancement of deep learning-particularly the integration of convolutional neural networks (CNNs) with various temporal modelling architectures-significant progress has been made in the field of sports event analysis. Early research methods heavily relied on manually designed features such as spatio-temporal interest points (STIP), histogram of oriented gradients (HOG), and similar descriptors, combined with shallow classifiers, whose performance was constrained by insufficient feature representation capabilities. Current mainstream approaches have fully shifted toward data-driven deep models. For instance, the inflated 3D ConvNet (I3D network) proposed by extends the ImageNet-pretrained Inception architecture to 3D convolutions and is pretrained on the large-scale kinetics dataset, significantly enhancing its ability to extract spatio-temporal features from videos. The SlowFast network developed by employs a dual-path architecture to capture both spatial details and rapidly changing motion information in videos, demonstrating outstanding performance in handling sports scenes with highly variable action rhythms. Furthermore, the transformer architecture, leveraging its powerful long-sequence modelling capabilities, has been introduced to video understanding tasks (Vaswani et al., 2017). By modelling global dependencies between video segments through self-attention mechanisms, it shows immense potential in complex event recognition.

However, despite achieving outstanding performance on multiple public benchmarks, these deep learning methods remain fundamentally reliant on an end-to-end data-driven paradigm, whose limitations are increasingly apparent. First, these models are often regarded as ‘black boxes’, lacking transparency and interpretability in their decision-making processes, making it difficult to answer critical questions like ‘why does the model consider this a goal?’. Second, their performance heavily depends on large-scale, high-quality manually annotated data. For many complex events in sports-such as soccer’s ‘offside’ or basketball’s ‘pick-and-roll’ plays-their low occurrence frequency coupled with intricate semantic definitions makes acquiring sufficient annotated samples prohibitively costly. This often results in suboptimal generalisation capabilities for rare events (Wang et al., 2020). This challenge is particularly acute in few-shot learning scenarios, where the model must recognise novel event categories from very limited examples (Wang et al., 2017). More fundamentally, purely data-driven models lack an understanding of sports’ inherent rules, prior logic, and common sense. For instance, a ‘goal’ event typically follows a specific spatio-temporal logic chain: the shooting action, the ball’s flight path, player celebrations, and the referee’s hand signals. Learning these complex, structured constraints solely from pixels without injecting any prior knowledge requires enormous data volumes and computational overhead, while also making it difficult to ensure logical reliability (Han et al., 2023).

To overcome these limitations, an emerging research trend explores integrating human knowledge into data-driven learning frameworks to build more robust and efficient models. This ‘knowledge-guided visual analysis’ research aims to embed structured prior information into neural network learning processes, thereby reducing dependence on data scale while enhancing model interpretability and generalisation capabilities. For instance, in visual reasoning tasks, knowledge graphs are leveraged to

model semantic relationships between objects; in visual question answering, external common-sense databases assist answer derivation (Satama, 2025). Specifically for sports video analysis, preliminary attempts have utilised simple spatio-temporal constraints (e.g., player position trajectories, ball location) or logical event relationships (e.g., a 'corner kick' often precedes a 'header shot') to aid recognition. These efforts demonstrate that domain knowledge, as a potent inductive bias, effectively constrains the hypothesis space, guiding models to focus on semantically relevant visual content. However, existing approaches largely remain at shallow knowledge representations (e.g., rigid rules) or loose post-processing fusion, failing to achieve deep, differentiable integration between knowledge representation and deep feature learning (Apriceno et al., 2021). Unlike traditional rule-based systems that rely on rigid, hand-crafted logic, our approach represents domain knowledge as soft, differentiable constraints. This design allows the model to adapt to new scenarios or varying video conditions through data-driven fine-tuning, thereby offering significantly greater flexibility and robustness. This limitation constrains further performance improvements. Therefore, exploring how to deeply integrate structured domain knowledge into the learning framework of deep neural networks in a tighter, more systematic manner has become a critical breakthrough for advancing this field-which is precisely the core starting point of this research. The adoption of a knowledge-guided framework is motivated by inherent limitations of purely data-driven methods, which include a lack of interpretability ('black-box' decisions), a heavy dependency on large-scale annotated data, and a fundamental difficulty in capturing the structured logic and rules inherent to the sports domain.

2 Related work

2.1 Early analysis methods based on handwritten features

Before the rise of deep learning technologies, sports video event analysis primarily relied on meticulously designed manual features. The core idea of such methods involves extracting low-level or mid-level visual features from video frames and utilising statistical models or machine learning classifiers for event recognition. Among these, spatial interest points (SIP) and STIP are representative feature descriptors capable of capturing significant changes in local regions within videos (Laptev, 2005). Building upon this foundation, the bag-of-words (BoW) model and its variants gained widespread adoption by quantifying local features into visual words to form a global representation of the video. To model temporal information, researchers further developed methods such as dense trajectories, generating rich motion descriptors by tracking the movement paths of interest points across consecutive frames (Wang et al., 2013). Although these handcrafted features embody researchers' deep insights into visual content and achieved initial success in early studies, their limitations are evident (Poppe, 2010): first, their representational capacity faces inherent constraints, making it challenging to capture complex and high-level semantic information. Second, method performance heavily relies on the expertise and skill in feature design, resulting in poor generalisation capabilities and difficulty adapting to video data from different sports or visual styles.

2.2 Deep learning-based video event recognition

With the revolutionary success of deep CNNs in image recognition tasks, researchers quickly extended their application to the domain of video, propelling sports video analysis into the era of deep learning. Depending on how temporal information is processed, deep learning approaches can be broadly categorised into several types. The first category comprises 2D CNN-based methods, such as the temporal segment network (TSN), which models long-range temporal structures by sparsely sampling video frames and aggregating predictions, achieving a good balance between efficiency and performance. The second category comprises 3D CNN-based approaches aimed at simultaneously learning spatial and temporal features. Originally developed by expanding the successful 2D ImageNet architecture into 3D, the I3D model is pre-trained on the large-scale video dataset Kinetics. This pre-training significantly enhances its feature extraction capabilities, establishing it as the mainstream benchmark model for a period. The third category comprises specialised architectures designed to address long-term temporal dependencies in video, such as using long short-term memory (LSTM) or gated recurrent units (GRU) as temporal modellers to encode frame-level features extracted by CNNs. In recent years, the transformer architecture has demonstrated immense potential in capturing global spatio-temporal dependencies through its powerful self-attention mechanism, achieving successful application in video action recognition tasks. These data-driven deep learning methods learn end-to-end mappings from raw pixels to high-level semantics, exhibiting feature representation capabilities and model performance far surpassing traditional handcrafted approaches (Simonyan and Zisserman, 2014).

2.3 Knowledge-guided visual analytics methodology

Despite the outstanding performance of deep learning models, their ‘black-box’ nature and reliance on large-scale labelled data have prompted researchers to explore new paradigms for integrating human prior knowledge into models. Knowledge-guided visual analysis aims to combine structured domain knowledge-such as physical laws, logical rules, and semantic relationships-with data-driven learning to enhance model efficiency, robustness, and interpretability. This research direction has made progress across multiple visual tasks. For instance, in visual question answering (VQA), external knowledge bases supplement image information (Lin et al., 2022), while in scene graph generation, linguistic prior constrains object relationships (Chen et al., 2022). Specifically for video event analysis, particularly in sports domains, some efforts have introduced domain-specific knowledge. For instance, object detection and tracking techniques capture player and ball trajectories (Intille and Bobick, 2001), while spatio-temporal logic rules like ‘player approaches ball’ or ‘ball moves toward goal’ are applied as post-processing to filter or validate deep learning predictions (Kamble et al., 2019). Other studies have attempted to construct probabilistic graph models (e.g., dynamic Bayesian networks) or Markov logic networks to explicitly represent temporal causal and logical relationships between events (Patel et al., 2022). However, most of these approaches have limitations: either the use of knowledge is loosely coupled (e.g., post-fusion), failing to sufficiently influence the feature learning process; or they rely on rigid rules, making the system difficult to optimise end-to-end and lacking flexibility (Khan and Curry, 2020). In recent years, embedding logical knowledge into neural networks in a differentiable

manner-such as through neuro-symbolic learning or using attention mechanisms to simulate knowledge routing has emerged as a prominent frontier topic. This approach aims to achieve a tight integration between deep features and structured knowledge. The pursuit of such integration has also spurred the development of novel spatial-temporal reasoning frameworks that can explicitly model the interactions between objects and their context over time (Geng et al., 2022). This line of work is often categorised under the umbrella of neuro-symbolic integration, which seeks to combine the pattern recognition strengths of neural networks with the reasoning capabilities of symbolic systems.

2.4 Summary and comparison

In summary, the technological evolution of sports video event analysis has progressed from shallow models reliant on explicit feature engineering to deep representation learning models driven by big data, and is now advancing toward collaborative models that integrate data and knowledge. To more clearly illustrate this evolutionary trajectory and the characteristics of each approach, the following table provides a systematic comparison of these related works.

Table 1 Comparative analysis of sports video incident analysis methods

Method category	*Core concept*	*Key features*
Handcrafted feature method	Low-level/mid-level features designed manually (e.g., STIP, trajectories) with shallow classifiers	Highly interpretable, but with limited feature representation and generalisation capabilities
Deep learning model	End-to-end learning of high-level spatio-temporal features using networks such as CNNs and recurrent neural networks (RNNs)	Feature expression capability and performance are superior, but data dependency is strong and interpretability is poor
Knowledge-guided approach	Incorporate domain rules, logic, and other prior information into the analysis process	The sample is highly efficient and logically sound, but knowledge construction is complex and difficult to integrate deeply
The method of this paper (KIN)	Deeply embed domain knowledge in a differentiable manner within neural networks to guide feature learning and decision-making	Combines strong representational power with interpretability, but knowledge definition requires human involvement

3 Methodology

This section will detail our proposed semantic event analysis framework for sports videos, which integrates domain knowledge with deep features. The core concept of this approach is embedding structured domain knowledge in a differentiable form within deep neural networks. This imposes powerful prior constraints on data-driven learning, ultimately achieving more precise and reliable event recognition.

3.1 Problem formulation

We formalise the task of semantic event analysis in sports videos as a temporal action detection problem. We formalise the task as a temporal action detection problem because it is uniquely suited for identifying and localising sparse semantic events in long, untrimmed sports videos. This framework directly addresses the need to pinpoint both the category and the temporal boundaries of an event, which is not effectively handled by frame-level classification or dense captioning. Given an uncropped long video V composed of T frames, i.e., $V = \{f_1, f_2, \ldots, f_T\}$. Our objective is to predict a set of event instances $\{(s_i, e_i, c_i, p_i)\}_{i=1}^{M}$, where s_i and e_i denote the start and end timestamps of the i^{th} event, respectively. $c_i \in \{1, 2, \ldots, C\}$ denotes its event category (e.g., goal, corner kick, foul), $p_i \in [0, 1]$ represents the confidence score for this prediction, and M is the total number of events predicted in the video.

In practical modelling, we typically partition a long video V into multiple non-overlapping short video segments $\{S_1, S_2, \ldots, S_N\}$, each containing a fixed number of frames. The model's task is to determine whether each segment S_n contains a specific event and provide its category probability. The model's overall prediction can be expressed as a function:

$$\mathbf{Y} = \mathcal{F}(V; \Theta, \Omega) \tag{1}$$

where $\mathbf{Y}$ denotes the final set of predicted results, Θ represents the learnable parameters of the deep neural network, and Ω signifies the encoded domain knowledge parameter set. Our innovation lies in how to construct $\mathcal{F}$, particularly in effectively integrating Ω with Θ.

3.2 Overview of the overall framework

The overall architecture of our proposed knowledge-injected network (KIN) is shown in Figure 1. It primarily consists of three core modules:

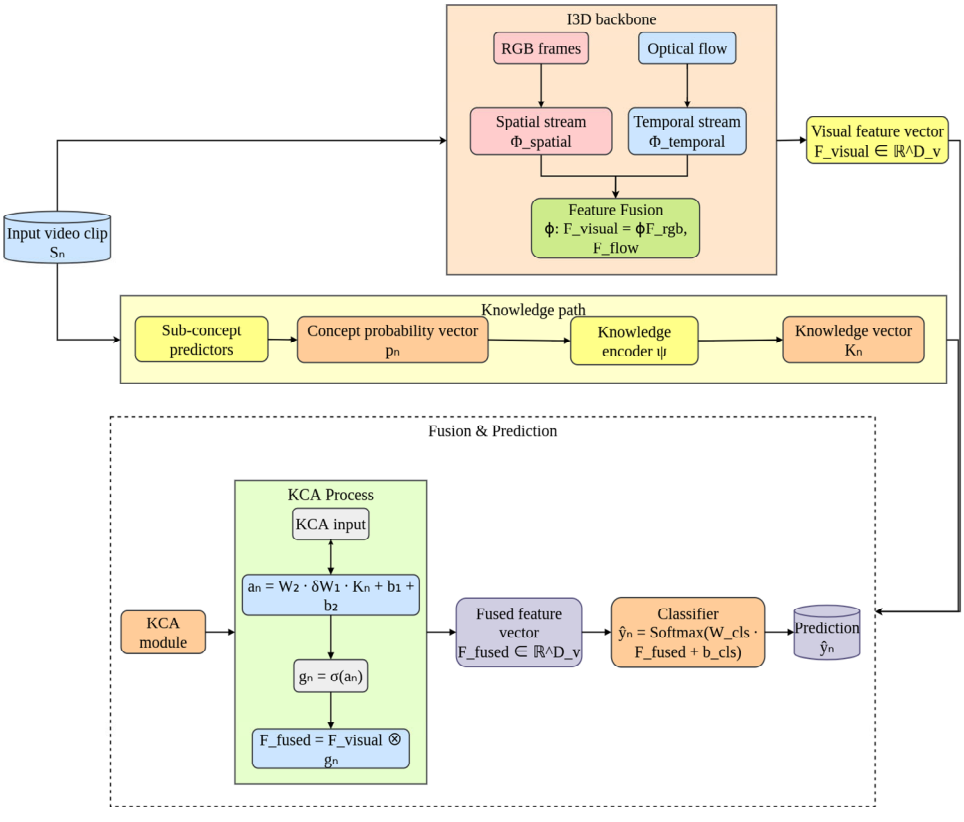
1 deep vision-temporal feature extraction module: responsible for extracting rich, multi-level spatio-temporal features from raw video frames

2 domain knowledge representation and encoding module: responsible for converting abstract sports rules into computable, structured vector representations

3 knowledge-feature fusion module: deeply integrates knowledge vectors with visual features through a novel attention mechanism, ultimately feeding the output to the classifier for prediction.

These three modules undergo joint optimisation in an end-to-end manner, enabling domain knowledge to guide feature learning and decision making during forward propagation while fine-tuning the specific parameters of knowledge based on training data during backpropagation. The knowledge module was designed with computational efficiency in mind. The sub-concept predictor and the KCA fusion mechanism are lightweight components, whose minimal overhead is justified by the significant performance gains, as evidenced in our ablation studies.

Figure 1 KIN framework diagram (see online version for colours)



3.3 Deep vision-temporal feature extraction

To capture appearance and motion information in videos, we employ a robust dual-stream architecture as our feature extraction backbone network (Rodríguez-Moreno et al., 2019). We employ a dual-stream architecture to explicitly and effectively capture complementary information: static appearance from RGB frames and dynamic motion from optical flow. This separation is particularly beneficial for analysing dynamic sports actions. Given a video segment S_n, we represent it as $S_n = \{f_{n,1}, f_{n,2}, \ldots, f_{n,L}\}$, where L denotes the segment length.

Spatial stream processes RGB frames to capture static appearance information. We employ a 2D CNN pretrained on ImageNet (such as the spatial component of ResNet-50 or the Inflated 3D (I3D) network) as the encoder:

$$\mathbf{F}_n^{rgb} = \Phi_{\text{spatial}}\left(S_n^{rgb}; \theta_{spatial}\right) \tag{2}$$

where $\mathbf{F}_n^{rgb} \in \mathbb{R}^{D_s}$ is the extracted spatial feature vector, D_s denotes the feature dimension, and $\theta_{spatial}$ represents the parameters of the spatial flow network.

Temporal stream processes dense optical flow frames to explicitly model motion information. For the temporal stream, we compute dense optical flow using a standard

algorithm (Farneback), which provides a reliable motion representation and is computationally efficient, ensuring a practical trade-off for model training. We employ a network with the same architecture as the spatial stream (but with a different number of input channels):

$$\mathbf{F}_n^{flow} = \Phi_{temporal}\left(S_n^{flow}; \theta_{temporal}\right) \tag{3}$$

where $\mathbf{F}_n^{flow} \in \mathbb{R}^{D_t}$ is the extracted motion feature vector.

Subsequently, we fuse the features from both streams to form a joint visual representation of the fragment:

$$\mathbf{F}_n^{visual} = \phi\left(\left[\mathbf{F}_n^{rgb}; \mathbf{F}_n^{flow}\right]\right) \tag{4}$$

where $[\cdot; \cdot]$ denotes concatenation, ϕ is a fully connected layer used to project the fused features into a unified feature space $\mathbb{R}^{D_v}$.. $\mathbf{F}_n^{visual}$ represents the final deep visual-temporal features obtained.

3.4 Representation and encoding of domain knowledge

This is the core innovation of this approach. We no longer treat domain knowledge as rigid rules but instead represent it as a differentiable, learnable soft constraint. Take the 'goal' event in soccer as an example: its occurrence typically depends on the coordinated emergence of a series of sub-concepts (atomic actions), such as *Shot*, *TowardsGoal*, *Celebration*, with specific temporal relationships between these sub-concepts. The proposed knowledge representation framework is generalisable. For application in other sports or domains, the core architecture remains, while adaptation primarily involves redefining the domain-specific sub-concepts and their logical relationships to the target events.

We first define a set of subconcepts $\mathcal{K} = \{k_1, k_2, \ldots, k_J\}$ where each k_j represents an atomic action or state (e.g., 'goalkeeper present', 'ball in penalty area'). For each video segment S_n, we employ a lightweight pre-convolution module (typically a small CNN or a simple linear classifier) to predict the probability of each subconcept's presence:

$$p\left(k_j \middle| S_n\right) = \sigma\left(\mathbf{W}_j \cdot \mathbf{F}_n^{\text{visual}} + b_j\right) \tag{5}$$

where $\mathbf{W}_j$ and b_j are the classifier parameters for subconcept k_j, and σ is the sigmoid activation function. The probabilities of all sub-concepts form a probability vector $\mathbf{p}_n = [p(k_1|S_n), p(k_2|S_n), \ldots, p(k_J|S_n)]^T \in [0, 1]^J$.

Next, we use first-order logic rules to encode the relationships between high-level events and these subconcepts. For example, the rule for the 'goal scored' event can be expressed as:

$$\mathbf{p}_n = \left[p\left(k_1 \middle| S_n\right), p\left(k_2 \middle| S_n\right), \ldots, p\left(k_J \middle| S_n\right)\right]^T \tag{6}$$

where $\mathbf{p}_n \in [0, 1]^J$ is a probability vector composed of the probabilities of all subconcepts, J is the predefined total number of subconcepts.

To embed such logical rules into differentiable neural networks (Bach et al., 2017), we adopt t-norm fuzzy logic from real-valued logic (van Krieken et al., 2022). The 'and' (^) operation in this logic can be approximated by a continuously differentiable operator:

$$\begin{aligned}\mathbf{k}_n^{goal} &= \psi_{goal}\left(\mathbf{p}_n\right) \\ &= p\left(\text{Shot}\middle|S_n\right)\cdot p\left(\text{TowardsGoal}\middle|S_n\right)\cdot p\left(\text{Celebration}\middle|S_n\right)\end{aligned} \tag{7}$$

where $\mathbf{k}_n^{goal} \in [0,1]$ can be understood as the confidence level for the 'goal' event calculated based on the rules. For each target event category c, we define its corresponding rule function ψ_c to map the subconcept probability vector $\mathbf{p}_n$ to a domain knowledge vector $\mathbf{K}_n \in \mathbb{R}^C$:

$$\mathbf{K}_n = \left[\psi_1\left(\mathbf{p}_n\right), \psi_2\left(\mathbf{p}_n\right), \ldots, \psi_C\left(\mathbf{p}_n\right)\right]^T \tag{8}$$

where $\mathbf{K}_n$ represents the encoded domain knowledge, expressing the logical relationship between high-level events and underlying visual evidence in a structured and differentiable manner. This paradigm of representing symbolic knowledge in a continuous vector space is a cornerstone of neuro-symbolic AI, enabling seamless coupling with gradient-based learning.

3.5 The integration mechanism of knowledge and characteristics

After obtaining the deep visual features $\mathbf{F}_n^{visual}$ and domain knowledge vector $\mathbf{K}_n$, the key lies in how to effectively fuse them. We designed a knowledge-conditioned channel attention (KCA) module (Guo et al., 2022). The KCA module was selected for its ability to use the domain knowledge vector as a conditioning signal to dynamically recalibrate channel-wise feature importance. This active guidance promotes a more focused fusion than passive mechanisms like concatenation.

This module uses knowledge vectors as prior information to recalibrate the importance of each channel in visual features (Jin et al., 2022). Specifically, the knowledge vector $\mathbf{K}_n$ is first transformed through a small bottleneck network (composed of two fully connected layers with a rectified linear unit (ReLU) activation function in between) to generate an attention weight vector:

$$\mathbf{a}_n = \mathbf{W}_2 \cdot \delta\left(\mathbf{W}_1 \cdot \mathbf{K}_n + \mathbf{b}_1\right) + \mathbf{b}_2 \tag{9}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C\times\frac{C}{r}}$, $\mathbf{W}_2 \in \mathbb{R}^{\frac{C}{4}\times D_v}$, $\mathbf{b}_1$, $\mathbf{b}_2$ are learnable parameters, r is the reduction ratio, and δ is the ReLU activation function. Then, we normalise $\mathbf{a}_n$ to the range between 0 and 1 using the sigmoid function, yielding the final attention gating vector:

$$\mathbf{g}_n = \sigma\left(\mathbf{a}_n\right) \tag{10}$$

where each element of $\mathbf{g}_n \in \mathbb{R}^{D_v}$ represents the importance assessment of the knowledge signal for the corresponding channel of the visual feature. Ultimately, the fused features are obtained through channel-wise multiplication:

$$\mathbf{F}_n^{fused} = \mathbf{F}_n^{visual} \otimes \mathbf{g}_n \tag{11}$$

where ⊗ denotes element-wise multiplication. $\mathbf{F}_n^{fused}$ is a knowledge-guided and enhanced feature representation that preserves the richness of the original data while highlighting the aspects most relevant to semantic events.

Finally, we feed $\mathbf{F}_n^{fused}$ into a simple classifier (such as a linear layer + softmax) to obtain the final segment-level event prediction probability distribution:

$$\hat{\mathbf{y}}_n = \text{Softmax}\left(\mathbf{W}_{cls} \cdot \mathbf{F}_n^{fused} + \mathbf{b}_{cls}\right) \tag{12}$$

where $\mathbf{W}_{cls}$ and $\mathbf{b}_{cls}$ represent the weight and bias parameters of the classification layer, while $\hat{\mathbf{y}}_n$ denotes the model-predicted segment-level event probability distribution.

The entire model, including the feature extraction network Φ, sub-concept predictor, and fusion module, undergoes end-to-end joint training by minimising the cross-entropy loss between the predicted output $\hat{\mathbf{y}}_n \$ and the true label \$\mathbf{y}_n$:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c} \log\left(\hat{y}_{n,c}\right) \tag{13}$$

Through this design, domain knowledge not only imposes constraints during inference but also directly participates in gradient calculations during training, guiding the direction of parameter updates. This achieves a deep synergy between knowledge and data-driven approaches.

4 Experimental verification

To comprehensively evaluate the effectiveness and superiority of our proposed KIN, we conducted extensive experiments on a large public dataset. This section details the experimental setup, comparison results, ablation studies, and visualisation analysis.

4.1 Dataset and experimental setup

We chose to conduct our experiments on one of the most challenging and authoritative public benchmarks in sports video analysis-the SoccerNet-v2 dataset. We conducted experiments on the SoccerNet-v2 dataset due to its large scale, long untrimmed videos, and fine-grained event annotations. Its established status as a benchmark allows for a fair and comprehensive evaluation against state-of-the-art methods in sports video analysis. This dataset comprises 500 complete soccer matches from Europe's top six leagues, totaling over 1,000 hours of video footage. It provides fine-grained annotations for three event categories: match events (e.g., goals, corner kicks, free kicks), video segment boundaries (e.g., start/end of halves), and primary camera views. We focused on the most challenging task of match event recognition, which encompasses 17 fine-grained event categories and over 30,000 event instances. We strictly adhere to official classifications, using 400 matches as the training set, 50 as the validation set, and 50 as the test set.

Model performance is evaluated using the widely adopted average precision (AP) metric to assess detection capabilities for each event category, with the mean average precision (mAP) serving as the core comprehensive evaluation metric (Everingham et al., 2010).

In terms of implementation details, our model is built upon the PyTorch framework. For the deep feature extraction module, we adopt the I3D network, pre-trained on the Kinetics-400 dataset, as the backbone architecture. Input video clips are downsampled to 25 frames per second, with each clip lasting 64 frames (approximately 2.56 seconds) and adjusted to a spatial resolution of 224 × 224 pixels. We employ the Adam with Weight Decay (AdamW) optimiser with an initial learning rate of 1e-4, decaying using a cosine annealing strategy. The batch size was set to 16, and the model was trained for 50 epochs across four NVIDIA V100 GPUs. The domain knowledge sub-concept predictor design is based on general knowledge within the soccer domain. We defined 12 atomic sub-concepts, including BallVisible, PlayerCelebrating, BallNearGoal, GoalkeeperVisible, and CameraShotOnGoal.

4.2 Comparative experiment

To fairly evaluate KIN's performance, we compare it against a suite of state-of-the-art methods, including:

1 I3D: a powerful 3D CNN baseline renowned for its exceptional spatio-temporal feature extraction capabilities

2 SlowFast: utilises a dual-path architecture processing video at different temporal rates, demonstrating strong performance on action recognition tasks

3 TimeSformer: a pure transformer-based video classification model adept at capturing long-term temporal dependencies (Bertasius et al., 2021)

4 SBMNet: a state-of-the-art approach specifically designed for the SoccerNet dataset in recent years, leveraging background modelling and feature fusion for event localisation, achieving outstanding performance on this benchmark (Cioppa et al., 2024).

Table 2 Overall performance comparison on the SoccerNet-v2 test set (mAP, %)

Method	*Backbone*	*Publication*	*Average mAP*
I3D	I3D	ICCV'17	58.2
SlowFast	SlowFast	ICCV'19	62.7
TimeSformer	TimeSformer-L	ICCV'21	65.4
SBMNet	I3D	CVPRW'21	67.9
KIN (Ours)	I3D	-	71.5

As shown in Table 2, our proposed KIN method achieves a mAP of 71.5%, significantly outperforming all baseline models. Compared to the baseline using the same I3D backbone network, KIN delivers absolute performance gains of 13.3% and 3.6% over I3D and SBMNet, respectively. This result strongly demonstrates the substantial advantage of integrating domain knowledge (Deng et al., 2020). Even when compared to the larger, more complex TimeSformer model, KIN exhibits a lead of approximately 6%.

Analysing the reasons, we believe that purely data-driven models, despite their strong representation learning capabilities, lack the utilisation of structured rules in the football domain (Lake et al., 2017). Consequently, they are prone to errors when handling events with complex logic or sparse samples. By injecting knowledge a priori, KIN effectively constrains the model's hypothesis space, guiding it to focus on visual cues most relevant to event semantics, thereby making more accurate judgements.

4.3 Melting experiment

To thoroughly investigate the contributions of each component within the KIN framework, we conducted exhaustive ablation experiments, the results of which are summarised in Table 3.

1 Importance of knowledge modules (A vs. B): when the domain knowledge encoding and fusion module was removed, the model degraded into a pure I3D model, experiencing a sharp performance drop of 11.3%. This clearly demonstrates that the injection of domain knowledge is the key driver of performance improvement, rather than solely stemming from the backbone network's capabilities.

2 Effectiveness of fusion mechanism (A vs. C): replacing the carefully designed KCA attention fusion mechanism with simple feature concatenation resulted in a 4.8% performance drop. This proves that our proposed KCA module enables more efficient and intelligent interaction between knowledge and visual features, with its 'recalibration' function outperforming simple concatenation.

3 The foundational role of visual features (A vs. D): models performed worst when using only rule-derived knowledge vectors while discarding raw visual features. This indicates that domain knowledge serves to 'guide' and 'enhance' rather than 'replace' the rich visual representations learned from data. The two elements are complementary and indispensable.

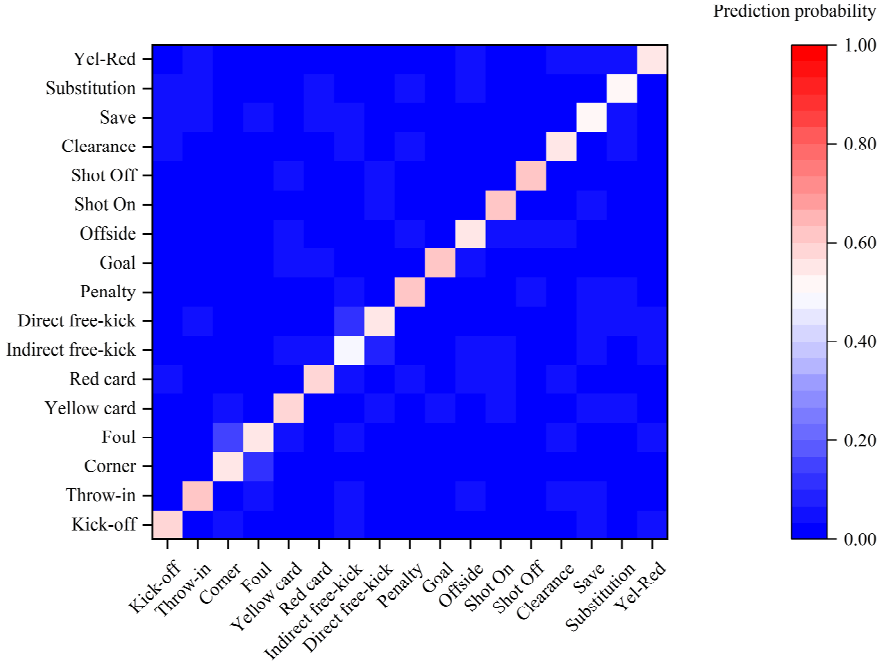
Table 3 Ablation study analysis (mAP on SoccerNet-v2 validation set)

Model variant	*Description*	*mAP (%)*
Full model	Complete KIN framework	*70.1*
w/o knowledge	Remove the entire knowledge module (sub-concept prediction + fusion), retaining only the I3D backbone	58.8
w/o fusion	Retain sub-concept prediction while removing the KCA fusion module, replacing it with direct feature concatenation	65.3
Only knowledge	Predict using only the output $\mathbf{K}_n$ from the knowledge module, without utilising the visual features $\mathbf{F}_n^{visual}$	52.4

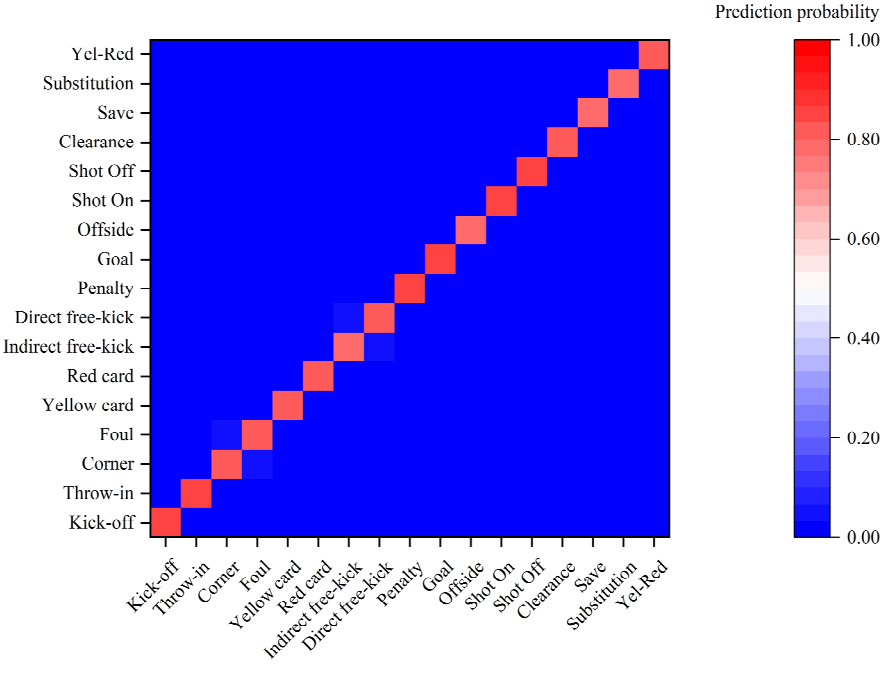
4.4 Visualisation and analysis

To gain a more intuitive understanding of the model's decision-making process, we conducted a visualisation analysis.

Figure 2 Confusion matrix comparison, (a) I3D baseline model (b) our KIN model (see online version for colours)



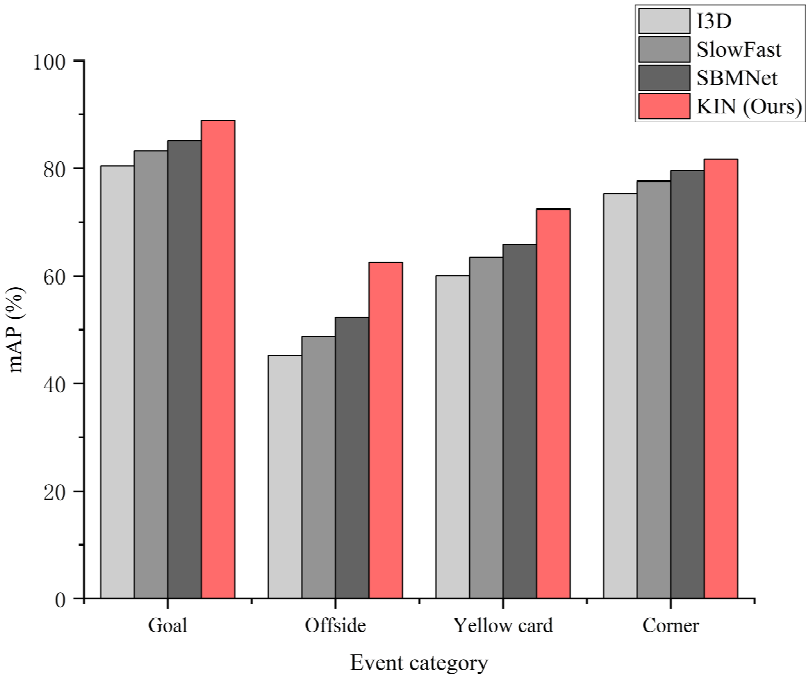
(a)



(b)

Analysis: comparing Figures 2(a) and 2(b), it is evident that the diagonal in the KIN model [Figure 2(b)] is brighter and more concentrated, while noise along the non-diagonal is significantly reduced. For instance, the I3D model [Figure 2(a)] exhibits severe confusion between 'yellow card' and 'red card', whereas the KIN model substantially mitigates this issue. This demonstrates that incorporating domain knowledge enhances the model's ability to distinguish semantically similar yet fundamentally different events.

Figure 3 Precision-recall curves for different methods across four key events (see online version for colours)



Analysis: as shown in Figure 3, our proposed KIN method (red solid line) maintains higher curves than other methods across nearly all event categories, particularly sustaining high accuracy even in high recall regions. This demonstrates that the confidence scores output by the KIN model are more accurate and reliable (Mukhoti et al., 2020). The ability to produce well-calibrated confidence scores is a key indicator of a model's trustworthiness, especially in safety-critical applications (Mukhoti et al., 2020). Notably, KIN's advantage is particularly pronounced for events with complex definitions and sparse samples, such as 'offside' and 'yellow card'. This outcome aligns perfectly with its design philosophy of leveraging prior knowledge to compensate for data scarcity.

In summary, both quantitative and qualitative experimental results consistently demonstrate that our proposed KIN framework, which integrates domain knowledge with deep features, significantly enhances the performance of semantic event analysis in sports videos. Its core design elements have been proven to be both effective and essential.

5 Conclusions

This paper addresses the issues of poor interpretability and heavy reliance on labelled data in purely data-driven approaches for semantic event analysis in sports video. It proposes a novel paradigm that integrates domain knowledge with deep features. By translating sports rules into computable, differentiable constraints and designing a knowledge-injection network to achieve end-to-end fusion of deep features and structured knowledge, this method significantly improves event recognition performance on the public SoccerNet dataset. Experimental results demonstrate that this approach not only achieves significantly higher average accuracy than mainstream baseline models but, more importantly, validates the critical role of the knowledge module and fusion mechanism through ablation studies. Visualisation analysis further substantiates the rationality and interpretability of the model's decision-making process.

The theoretical contributions of this study are twofold: first, it proposes a universal, transferable knowledge representation and fusion framework, offering a novel technical approach for incorporating human prior knowledge into deep learning models and bridging the gap between data-driven and knowledge-driven methods. Second, it advances the practical application of neuro-symbolic learning in a specific vertical domain (sports video analysis), demonstrating the critical value of structured knowledge in enhancing model sample efficiency and logical reasoning capabilities.

At the practical level, this study provides a feasible solution for constructing next-generation intelligent sports video analysis systems. The system can more accurately generate automatic highlights, perform tactical breakdowns, and conduct data statistics, significantly enhancing the efficiency of content production and consumption. Simultaneously, this approach holds important implications for other video understanding tasks with scarce annotated data-such as surveillance event detection and industrial anomaly recognition-offering new insights for addressing few-shot learning challenges.

Of course, this study still has certain limitations, primarily manifested in the fact that domain knowledge construction remains dependent on expert experience and cannot be automatically learned from data. Future work will focus on exploring self-learning and evolutionary mechanisms for knowledge, and attempting to extend this framework to broader video understanding and reasoning tasks to validate its generality and scalability.

Declarations

The author declares that she has no conflicts of interest.

References

Apriceno, G., Passerini, A. and Serafini, L. (2021) 'A neuro-symbolic approach to structured event recognition', *Leibniz International Proceedings in Informatics*, Vol. 206, pp.1101–1114.

Bach, S.H., Broecheler, M., Huang, B. and Getoor, L. (2017) 'Hinge-loss Markov random fields and probabilistic soft logic', *Journal of Machine Learning Research*, Vol. 18, No. 109, pp.1–67.

Bertasius, G., Wang, H. and Torresani, L. (2021) 'Is space-time attention all you need for video understanding?', *ICML*, Vol. 139, No. 1, pp.4–14.

Chen, K., Huang, Q., McDuff, D., Bisk, Y. and Gao, J.(2022) 'KRIT: knowledge-reasoning intelligence in vision-language transformer', *IEEE Access*, Vol. 10, No. 1, pp.12345–12358.

Cioppa, A., Giancola, S., Somers, V., Magera, F., Zhou, X., Mkhallati, H., Deliège, A., Held, J., Hinojosa, C. and Mansourian, A.M. (2024) 'SoccerNet 2023 challenges results', *Sports Engineering*, Vol. 27, No. 2, p.24.

Deng, C., Ji, X., Rainey, C., Zhang, J. and Lu, W. (2020) 'Integrating machine learning with human knowledge', *Iscience*, Vol. 23, No. 11, p.101656.

Everingham, M., van Gool, L., Williams, C.K., Winn, J. and Zisserman, A. (2010) 'The Pascal visual object classes (VOC) challenge', *International Journal of Computer Vision*, Vol. 88, No. 2, pp.303–338.

Geng, T., Zheng, F., Hou, X., Lu, K., Qi, G-J. and Shao, L. (2022) 'Spatial-temporal pyramid graph reasoning for action recognition', *IEEE Transactions on Image Processing*, Vol. 31, pp.5484–5497.

Guo, M-H., Xu, T-X., Liu, J-J., Liu, Z-N., Jiang, P-T., Mu, T-J., Zhang, S-H., Martin, R.R., Cheng, M-M. and Hu, S-M. (2022) 'Attention mechanisms in computer vision: a survey', *Computational Visual Media*, Vol. 8, No. 3, pp.331–368.

Han, S., Liu, J., Zhang, J., Gong, P., Zhang, X. and He, H. (2023) 'Lightweight dense video captioning with cross-modal attention and knowledge-enhanced unbiased scene graph', *Complex & Intelligent Systems*, Vol. 9, No. 5, pp.4995–5012.

Intille, S.S. and Bobick, A.F. (2001) 'Recognizing planned, multiperson action', *Computer Vision and Image Understanding*, Vol. 81, No. 3, pp.414–445.

Jin, X., Xie, Y., Wei, X-S., Zhao, B-R., Chen, Z-M. and Tan, X. (2022) 'Delving deep into spatial pooling for squeeze-and-excitation networks', *Pattern Recognition*, Vol. 121, p.108159.

Kamble, P.R., Keskar, A.G. and Bhurchandi, K.M. (2019) 'Ball tracking in sports: a survey', *Artificial Intelligence Review*, Vol. 52, No. 3, pp.1655–1705.

Khan, M.J. and Curry, E. (2020) 'Neuro-symbolic visual reasoning for multimedia event processing: overview, prospects and challenges', *CIKM (Workshops),* Vol. 1, No. 2, pp.45–58.

Lake, B.M., Ullman, T.D., Tenenbaum, J.B. and Gershman, S.J. (2017) 'Building machines that learn and think like people', *Behavioral and Brain Sciences*, Vol. 40, p.e253.

Laptev, I. (2005) 'On space-time interest points', *International Journal of Computer Vision*, Vol. 64, No. 2, pp.107–123.

Lin, Y., Xie, Y., Chen, D., Xu, Y., Zhu, C. and Yuan, L. (2022) 'Revive: regional visual representation matters in knowledge-based visual question answering', *Advances in Neural Information Processing Systems*, Vol. 35, pp.10560–10571.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. and Dokania, P. (2020) 'Calibrating deep neural networks using focal loss', *Advances in Neural Information Processing Systems*, Vol. 33, pp.15288–15299.

Patel, A.S., Vyas, R., Vyas, O. and Ojha, M. (2022) 'A study on video semantics; overview, challenges, and applications', *Multimedia Tools and Applications*, Vol. 81, No. 5, pp.6849–6897.

Poppe, R. (2010) 'A survey on vision-based human action recognition', *Image and Vision Computing*, Vol. 28, No. 6, pp.976–990.

Rodríguez-Moreno, I., Martínez-Otzeta, J.M., Sierra, B., Rodriguez, I. and Jauregi, E. (2019) 'Video activity recognition: state-of-the-art', *Sensors*, Vol. 19, No. 14, p.3160.

Satama, P. (2025) 'Knowledge-Integrated reasoning for visual question answering', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 47, No. 5, pp.11245–11258.

Simonyan, K. and Zisserman, A. (2014) 'Two-stream convolutional networks for action recognition in videos', *Advances in Neural Information Processing Systems*, Vol. 3, No. 7, pp.568–576.

Van Krieken, E., Acar, E. and van Harmelen, F. (2022) 'Analyzing differentiable fuzzy logic operators', *Artificial Intelligence*, Vol. 302, p.103602.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, Vol. 30, No. 4, pp.5998–6008.

Wang, H., Kläser, A., Schmid, C. and Liu, C-L. (2013) 'Dense trajectories and motion boundary descriptors for action recognition', *International Journal of Computer Vision*, Vol. 103, No. 1, pp.60–79.

Wang, J.R. and Parameswaran, N. (2004) 'Survey of sports video analysis: research issues and applications', *Proceedings of the Pan-Sydney Area Workshop on Visual Information Processing*, Vol. 45, No. 2, pp.87–90.

Wang, Y., Yao, Q., Kwok, J.T. and Ni, L.M. (2020) 'Generalizing from a few examples: a survey on few-shot learning', *ACM Computing Surveys (Csur)*, Vol. 53, No. 3, pp.1–34.

Wang, Y-X., Ramanan, D. and Hebert, M. (2017) 'Learning to model the tail', *Advances in Neural Information Processing Systems*, Vol. 30, No. 8, pp.7029–7039.