A Knowledge Graph Framework for Interpretable Video-Based Activity Recognition

Radu-Casian Mihailescu

School of Mathematical and Computer Sciences
Heriot-Watt University
R.Mihailescu@hw.ac.uk

Abstract

We propose two approaches for human activity recognition in videos that leverage knowledge graph representations. The first method constructs a Positional Encoding Knowledge Graph (PE-KG) by extracting objects and their spatial relationships from video keyframes, which are then analyzed using association rule mining. The second approach, termed Video KG, augments this representation by incorporating semantic cues from image captioning and affective insights from emotion detection with demographic analysis. The approach employs knowledge graph embeddings to capture spatiotemporal and contextual dependencies, leading to improved classification accuracy and enhanced interpretability on benchmarks such as the Kinetics dataset.

Introduction

The availability and proliferation of video datasets coupled with the advancements in compute power have positioned video data as a key, yet largely untapped resource for modern machine learning (ML) models. Video data provides a rich source of temporal and spatial information, allowing for more comprehensive scene understanding compared to static images. Unlike static images that capture only a single moment in time, video data enables the extraction of dynamic cues such as motion vectors, object trajectories, temporal dependencies, and the evolution of scene elements over time. These temporal features offer critical insights into the progression of actions, subtle changes in motion, and the context of interactions, thereby enriching the scene understanding and context-awareness of ML systems.

In this paper we are focusing on a key ML task in the context of video data, namely human activity recognition (HAR). It is worth mentioning though, that non-visual sensing modalities have also been applied to this end. For example, wearable inertial sensors—such as accelerometers and gyroscopes embedded in smart devices—have become key for continuously monitoring daily activities (De Ramón Fernández et al. 2024). Fluctuations in radio frequency signals from WiFi and Bluetooth networks are now exploited to infer fine-grained motion patterns even in cluttered indoor settings (Engström and Persson 2023). Furthermore, acoustic sensors that analyze ambient sound patterns are emerging as a viable modality for recognizing activities without compromising privacy (Gharib et al. 2023).

Nonetheless, video-based HAR provides the most comprehensive information, as it captures rich spatial and temporal context along with subtle visual cues that no single non-visual modality can fully replicate. By analyzing the dynamic and contextual information inherent in video data, HAR systems can interpret complex social dynamics and interactions that are critical in real-world applications. The ability to accurately recognize activities enables the development of intelligent systems that can assist individuals, enhance security, and improve automation in diverse settings. Thus, HAR has gained significant attention due to its broad applications in surveillance, healthcare, autonomous systems, and human-computer interaction, to name a few.

For instance, in surveillance, HAR systems can automatically detect anomalous or dangerous behaviors in real time, thereby assisting security personnel with prompt threat assessment (Tandel et al. 2024). In healthcare, these systems contribute to patient monitoring by identifying critical events such as falls or abnormal movements, which is particularly valuable for elderly care and rehabilitation (Jamali et al. 2025). Autonomous systems, including self-driving cars and drones, leverage HAR to predict pedestrian trajectories and understand environmental dynamics, thus improving navigational safety (Bi et al. 2019). Moreover, in the realm of human-computer interaction, HAR facilitates intuitive gesture recognition and context-aware interfaces, leading to more immersive and responsive user experiences (Sun et al. 2024). Such diverse applications, underscore HAR's potential to transform multiple industries by offering critical insights into human behavior and interactions.

While deep learning approaches have achieved significant breakthroughs in HAR, relying solely on these methods often results in models that are highly data-intensive and sometimes less robust in challenging environments. Classical techniques, with their domain-specific handcrafted features, offer robustness and interpretability but can fall short in capturing the high-level abstractions necessary for complex video understanding. In this paper we aim to build on the advantages of both type of approaches. By leveraging hybrid models, our research aims to contribute to the development of more accurate and adaptive HAR frameworks that can be deployed in real-world applications.

The remainder of this paper is structured as follows. In Section 2 we review related works for video-based HAR. Section 3 introduces two approaches based on knowledge graph representations. Section 4 presents the experimental setup and discusses results. Finally, we conclude the paper in Section 5.

Related Work

Early approaches primarily relied on handcrafted features for activity recognition, such as histogram of oriented gradients (Dalal, Triggs, and Schmid 2006), optical flow (Wang et al. 2011), and spatiotemporal interest points (Chakraborty et al. 2011), which, while computationally efficient, were limited in performance due to the quality of these features. Onward, the methodologies utilized in HAR have evolved significantly, driven by advancements in deep learning techniques.

CNN-based approaches

Convolutional neural networks (CNNs) have been widely employed for spatial feature extraction from video frames, effectively capturing local patterns, such as edges and textures. In (Simonyan and Zisserman 2014) the authors introduced an important milestone, whereas, a deep learning framework decouples video analysis into two complementary streams: a spatial stream that processes static RGB frames to capture appearance and object cues, and a temporal stream that processes stacked dense optical flow fields to capture motion information. This multi-stream approach has been further developed in various HAR settings. For instance, the application of automated machine learning (AutoML) techniques is investigated in (Popescu, Mocanu, and Cramariuc 2020), where the authors devise several data streams through independent 2D CNNs, focusing on depth data, skeletal information, and contextual objects and evaluate different fusion mechanism, concluding that the late fusion approach consistently outperformed the other techniques. Unlike 2D CNNs, which process spatial information from single frames, 3D CNNs extend convolutions into the temporal dimension, making them well-suited for capturing motion patterns over time. In (Carreira and Zisserman 2017), the authors propose repurposing existing 2D CNN architectures designed for image classification and inflating them into 3D architectures capable of processing video. In general, 3D CNNs are widely used for HAR, particularly for video and depth sensor data, having been shown to deliver high performance.

RNN-based approaches

However, to account for temporal dynamics, recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks, are used to model sequential dependencies, enabling the retention of information from previous frames. In (Inoue, Inoue, and Nishida 2016), the Deep RNN is structured as a multi-layer LSTM network, where each internal layer captures the temporal dynamics. Extensive parameter tuning is performed, managing to decrease the inference time by an order of magnitude compared to previous work. Similarly, in (Ullah et al. 2020), the authors use LSTM and its variants such as multi-layer or deep LSTM and bidirectional LSTM networks, for sequence learnign in the context of HAR. Ensembles of deep LSTM learners are explored in (Guan and Plötz 2017) and shown to outperform individual LSTM networks.

Hybrid approaches

In (Johnson and Uthariaraj 2020), a Restricted Boltzmann machine (RBM) architecture, is combined with CNNs, such that the convolutional part serves as the first layer for extracting low-level spatial-temporal features from video blocks, which are then further processed by the RBM-NN to build a robust, compact representation for human action. In (Joudaki, Imani, and Arabnia 2025), the authors build upon this architecture combining a 2D Conv-RBM with an LSTM network. Due to its efficient and lightweight architecture, the model is particularly beneficial for real-time systems.

Recent studies have shown that hybrid models, integrating CNNs with RNNs, yield robust performance in recognizing activities. (Xia, Huang, and Wang 2020) proposes a two-layer LSTM followed by convolutional layers architecture. A global average pooling layer is used to replace the fully connected layer after convolution for reducing model parameters, in combination with batch normalization. A CNN-LSTM architecture is introduced in (Mutegeki and Han 2020), where the authors report on results on various network configurations, showing state-of-the-art results on several video datasets. ConvLSTM and LRCN representing different variants for integrating CNNs and RNNs, are compared in (Uddin et al. 2024), demonstrating that the LRCN model offers a better trade-off between accuracy and efficiency. In (Mihanpour, Rashti, and Alavi 2020), the authors combine CNNs and deep bidirectional LSTM. First, the approach processes raw video frames through ResNet152 to extract deep, discriminative features that capture the visual content of each frame. These features are then sequentially fed into a DB-LSTM, which processes the data in both forward and backward directions to effectively learn the temporal dependencies and dynamic patterns across frames.

Attention-based approaches

Despite their effectiveness, CNN-LSTM architectures can encounter computational inefficiencies due to high parameter count, particularly with high-resolution or extended video data. This limitation has prompted researchers to explore alternative architectures, such as Vision Transformers (ViTs (Kolesnikov et al. 2021)), which utilize self-attention mechanisms (Lin et al. 2017) to capture global dependencies across spatial and temporal dimensions.

VideoMAE V2 (Wang et al. 2023) proposes a novel framework for scaling video foundation models by integrating a dual masking strategy into the masked autoencoder paradigm. The encoder operates on a highly masked subset of video tokens to extract robust spatiotemporal representations, while a second masking mechanism is applied within the decoder to further reduce computational cost without sacrificing reconstruction quality. This mechanism enables efficient pre-training even for billion-parameter models built on ViTs. ViT-ReT (Wensel, Ullah, and Munir 2023), is a novel transformer-based framework for human activity recognition in videos that replaces the conventional convolutional and recurrent layers with a ViT for spatial feature extraction and a Recurrent Transformer (ReT) for modeling temporal dependencies. The ViT efficiently processes individual video frames, while the ReT captures the sequential dynamics across frames in parallel, overcoming the inherent bottlenecks of traditional CNN-RNN architectures A key innovation in (Bertasius, Wang, and Torresani 2021) consists in introducing the ”divided attention” mechanism, which separately applies temporal and spatial attention within each Transformer block to efficiently capture both local and long-range dependencies, essentially extending the transformer framework to temporal data. The advantage of this approach consists in significantly improving computational efficiency and scalability.

In contrast to traditional deep learning architectures that often function as black boxes, knowledge graph-based approaches offer a promising pathway for achieving both robust performance and enhanced interpretability in video analysis. By representing videos as structured graphs where nodes denote objects and edges capture their relationships, these methods enable explicit reasoning over complex visual scenes. This structured representation facilitates the integration of heterogeneous data sources, thereby enriching contextual understanding. In this paper we introduce two KG-based methods for the video action recognition task, focusing on contextual-awareness and enhanced interpretability.

Proposed Framework

Overview

Formally, the problem of video action recognition assumes a given training set of labeled videos $\mathcal{D}^t = \{\mathcal{V}^t, \mathcal{Y}^t\}$, where each video $v^t \in \mathcal{V}^t$ is associated with an action label $y^t \in \mathcal{Y}^t$ and each video can be represented as a sequence of frames $v^t = \{f_1, f_2, \ldots f_n\}$. Similarly, there is a validation dataset $\mathcal{D}^v = \{\mathcal{V}^v, \mathcal{Y}^v\}$ and the task is to map each video $v^v \in \mathcal{V}^v$ to its corresponding action label $y^v \in \mathcal{Y}^v$, where $\mathcal{Y}^t = \mathcal{Y}^v$. The goal is thus to learn a classifier that can generalize to $\mathcal{D}^v$.

Positional Embedding KG Approach

Activity recognition in video streams requires analyzing continuous sequences of images, while understanding the context of an image requires analyzing its content through high-level semantic concepts. These concepts are represented as objects, which are fundamental for interpreting scene context and play a crucial role in recognizing activities within videos. Starting from this observation we set out to investigate the feasibility of using a high-level image representations for action recognition, by describing images through identifying connections between various objects, resulting in a more detailed and meaningful representation of the visual scene.

Deep neural networks have been established as the best performing ML models for object detection including various architectures, such as region-based or single-stage detectors. Region-based models first generate region proposals and then classify objects within those regions e.g. Faster R-CNN (Ren et al. 2015). Single-stage detectors, e.g. YOLO (Tian, Ye, and Doermann 2025), directly predict object classes and bounding boxes in a single forward pass. While the former category offers precise localization and classification, the latter can provide a good trade-off between speed and accuracy. The output of an object detection model includes coordinates defining object location (bounding boxes), the predicted object category and the confidence score of a correct detection (see Fig. 1).

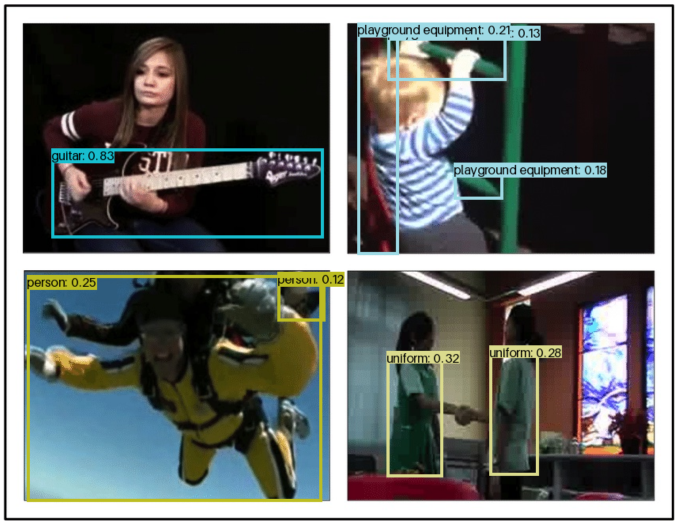

Figure 1: Applying object detection on examples from the Kinetics dataset (Kay et al. 2017)

Knowledge graph generation. In this phase, we begin by extracting from each video the set of unique objects from the scene and determining their relative position based on the information obtained from the bounding box coordinates. We define the following set of mutually exclusive positional relationships between the detected objects: $\mathcal{P}$= $\{$*next to, behind, in front, to the left, to the right, above, under*$\}$. Next, we represent this information using a knowledge graph that captures the given relations between the identified objects. The Positional Encoding Knowledge Graph (PE-KG) $\mathcal{G}_{pos} = \{\mathcal{O}, \mathcal{T}, \mathcal{P}, \mathcal{L}\}$ is defined as a union of a set of nodes $\mathcal{O}$ and a set of directed triples $\mathcal{T} \subseteq \mathcal{O}$ x $\mathcal{P}$ x $\mathcal{O}$, that are constructed over a set of predicates P. The nodes $o \in \mathcal{O}$, from the generated PE-KG, correspond to the classes of objects that can be recognized by the given object detector model. The set of literal values $\mathcal{L}$ denotes the confidence score associated to each triple. Consequently, we process the training set $\mathcal{D}^t$ and obtain a positional encoding knowledge graph for each video $v \in \mathcal{V}^t$.

Activity Mining Next, we adapt the data mining Apriori algorithm (Agrawal and Srikant 1994) and apply it on the PE-KG graphs for deriving association rules that can be used for activity classification.

The Apriori algorithm starts by identifying frequent itemsets based on a minimum support threshold. In our setting, for each video $v \in \mathcal{V}^t$, associated to a given action label $y^t \in \mathcal{Y}^t$, we extract a set of triples representing object relationships, which correspond to an itemset. Frequent sets of relationship are determined based on support calculation for a given itemset X:

$$\text{Support}(X) = \frac{\text{Number of videos containing } X}{\text{Total no. of videos}} \tag{1}$$

In order to account for the probability scores produced by the object detection model, we replace the traditional binary support (1 if a relation appears, 0 otherwise), with the arithmetic mean of the scores of all the objects that appear within the itemset. Thus, we onwards use the updated formula of the support, where instead of just counting the presence of X, we take the average score of the relations in X across all videos.

$$S(X) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} S_v(X) \tag{2}$$

$$S_v(X) = \frac{1}{|X|} \sum_{i \in X} s_i \tag{3}$$

where s$_i$ is the probability score of object i, $|X|$ is the number of object in the itemset X and $S_v(X)$ is the probability-weighted support of X. Itemsets are then filtered based on whether the support is greater than or equal to a given minimum support threshold δ_s. We generate rules of the form $X \Rightarrow Y$, where X is a set of triples and Y represents the activity class, by computing the following metrics:

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)} \tag{4}$$

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Confidence}(X \Rightarrow Y)}{\text{Support}(Y)} \tag{5}$$

Similarly, we used δ_c and δ_l as minimum thresholds for prunning the candidate set based on the confidence and lift scores and obtaining the final association rules. Finally, new video samples can be classified using the generated activity association rules $X \Rightarrow Y$, where X is a structured scene representation. Using association rules instead of relying on single object-object relations provides a more comprehensive and context-aware understanding of a scene.

Video KG Approach

In the following we introduce an additional knowledge graph video representation that builds upon the positional embedding approach, augmenting the type of predicates captured from the scenes. Namely, we further integrate two state-of-the-art pretrained models, one for image captioning and one for emotion detection with demographic information, in order to enrich the expressiveness of the extracted relationships. Note that the proposed framework is not restricted to specific type of relationships, but as we will see, it is general in the sense that it can directly integrate any type of model that outputs relevant triples.

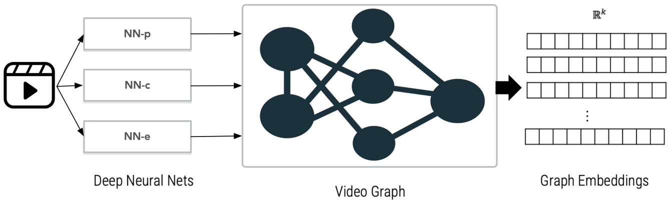

Figure 2: Video graph framework

The first step is video preprocessing, where keyframes are sampled from the video, based on scene change detection, using the histogram-based comparison technique (Cho and Kang 2019). Next, a captioning model is employed to produce detailed descriptions for the identified keyframes. The typical image-to-text architecture for image captioning consists of an image encoder and a language decoder. Specifically, in this study we use VC-GPT(Wang, Huang, and Li 2022), a transformer-based model that leverages the powerful text generation capabilities of GPT-2. The generated captions are then parsed to extract entities such as people, objects, and actions, as well as relationships between them, which are used to generate triples. Emotion detection is integrated by employing a high performing CNN-based facial expression recognition model (Khattak et al. 2022), that classifies emotions such as happiness, sadness, and surprise from detected faces. The detected emotions are then linked to specific individuals and inserted as triples within the knowledge graph, enabling an affective representation of video content. Demographic analysis is also incorporated by applying the same pretrained model, which is designed for multi-task learning, being able to infer attributes such as age, gender, and ethnicity from detected faces.

Triplet Encoding Fig. 2 depicts the video graph frameworks, which brings together the triples generated by the three different neural network models NN_p, NN_c, NN_e, focusing on positional relations, image captioning, emotion recognition and demographics, respectively. The set of triples (h, r, t) are consolidated into the *Video Graph* $\mathcal{G}$, describing facts from the scenes, with h representing the head entity (or subject), r the relation (or predicate), and t the tail entity. Although triples can be used to effectively represent structured data, their symbolic nature makes knowledge graphs challenging to manipulate and use for downstream ML task. However, graph embeddings have emerged as an efficient technique to convert high-dimensional sparse graphs into low-dimensional, dense and continuous feature spaces. The purpose of graph embeddings is to encode nodes into a low-dimensional, highly-informative latent vector space, preserving the graph properties and fostering knowledge inference and fusion.

KG embeddings are typically categorized into three classes: translational, semantic matching, and neural network-based. In this paper, we are focusing on one representative technique from each class, namely *TransE, ComplEx,* and *ConvKB*. The *TransE* model (Bordes et al. 2013) was the first to introduce the idea of translational invariance in the context of graph embeddings, taking inspiration

Activity	**Precision**		**Recall**		**F1-score**	
Approach	I3D	PE-KG	I3D	PE-KG	I3D	PE-KG
Cooking Chicken	0.68	**1.00**	0.54	0.38	**0.60**	0.55
Dunking Basketball	**0.79**	0.62	0.53	**0.78**	0.63	**0.69**
Filling Eyebrows	0.72	**0.86**	**0.85**	0.18	**0.78**	0.30
Golf Putting	0.55	**0.75**	**0.81**	0.45	**0.66**	0.56
Playing Piano	**0.76**	0.71	0.48	**0.88**	0.59	**0.79**
Pushing Cart	0.58	**1.00**	**0.67**	0.45	**0.62**	0.62
Riding Elephant	0.82	**1.00**	0.76	**0.94**	0.76	**0.97**
Roller Skating	**0.83**	0.16	0.32	**0.47**	0.46	0.23
Skateboarding	**0.67**	0.49	**0.95**	0.90	**0.78**	0.63
Smoking	0.68	**0.73**	0.40	**0.44**	0.51	0.55

Table 1: Performance metrics comparing the I3D detector to the PE-KG approach.

from regularities observed in the *word2vec* (Mikolov et al. 2013) linguistic embeddings. Entities and relations are low-dimensional vectors in Euclidean $\mathbb{R}^d$ space, enforcing that relation r acts as a translation vector between the two entities h and t:

$$h + r \approx t \tag{6}$$

Once the embedding is trained, scoring functions are used to measure the likelihood or strength of a relationship between nodes based on their learned embeddings. The scoring function for *TransE* model is computed using the $l1$ or $l2$ norm constraints:

$$f_r(h,t) = \|h + r - t\|_{l_1/l_2} \tag{7}$$

ComplEx (Trouillon et al. 2016), belongs to the class of semantic matching models, where the triple likelihood is quantified using the multiplication operator:

$$f_r(h,t) = h \times r \times t \tag{8}$$

In *ComplEx*, the embeddings are values in the complex space $\mathbb{C}^d$, where the scoring function returns the real value $Re(\cdot)$ of:

$$f_r(h,t) = Re\left(h^\top diag(r)\bar{t}\right) \tag{9}$$

$\bar{t}$ represents the complex conjugate of t and $diag(r)$ restricts r to a diagonal matrix.

Finally, *ConvKB* (Nguyen et al. 2017), exploits a convolutional neural network to train embeddings, where the scoring function is obtained by applying the 1D convolution filters Ω to the matrix obtained by transforming each element of a triple into a three-row matrix $[h,r,t] \in \mathbb{R}^{3\times d}$. Then, the feature maps are concatenated and used to determine the score by performing a dot product with the weight vector w:

$$f_r(h,t) = concat\left(\sigma\left([h,r,t] * \Omega\right)\right) w \tag{10}$$

Activity recognition In order to perform activity recognition, we first proceed to generate an individual knowledge graph for each given activity in our training set $\mathcal{D}^t$. This implies conducting the following sequence of steps:

- Extracting the keyframes for each video $v \in \mathcal{V}^t$;
- Applying the neural network models NN_p, NN_c and NN_e onto the keyframes to generate the set of triples $\mathcal{T}_v$ corresponding to each video v based on extracting positional relations, image captions and emotion detection, plus demographics information;
- For each activity label $y \in \mathcal{Y}^t$, we construct a video graph $\mathcal{G}_y^t = \{\bigcup \mathcal{T}_v | \forall v \in \mathcal{V}^t$, where v, corresp. to act. $y\}$ as the union of all triples extracted from all videos v associated to activity y.
- We train three embedding representations for each $\mathcal{G}_y^t$ based on the *TransE, ComplEx,* and *ConvKB* models.

In the previous section, we reviewed the scoring functions for the different embedding representations. These are used to determine the plausibility of each triple. Training the embeddings consists of solving the optimization problem that maximizes the total plausibility of observed triples, by updating the entity and relation embeddings. It is worth mentioning that the triples extracted from videos represent the positive examples used during training. Consequently, we also generate negative samples, both by corrupting either its tail or head, and by sampling triples from the other classes, in order construct a training set able to learn strong discriminative representations. Finally, activity recognition for an unseen video is carried out by first extracting all the triples, similarly to the training process, followed by comparatively assessing the plausibility of the given set of triples against the previously generated video graphs $\mathcal{G}_y^t$, corresponding to each activity label y.

Experimental Results

In the following, we report on experiments on the two proposed approaches evaluated on the Kinetics dataset. Kinetics (Kay et al. 2017) is a large-scale action recognition dataset used for training and benchmarking machine learning models in video understanding. It was developed by DeepMind[1], it contains short video clips labeled with human actions and it comes in several versions, depending on the size and number of activity classes. In this paper we are using the Kinetics-400 version, consisting in 400,000 videos across

[1] https://deepmind.google

400 action classes. Due to the nature of our approach, in the experiments we are particularly focusing on human-centric activities that involve interacting with different objects.

We choose a 3D-CNN architecture as a baseline for comparison against our proposed PE-KG approach. Specifically, we opt for the efficient I3D network, which can process videos of different lengths by using sliding window mechanisms and 3D global pooling, making it adaptable to real-world video sequences. Additionally, I3D is known to outperform other 3D-CNN models, such as C3D (Tran et al. 2015), having higher accuracy and better transfer learning capabilities, being pretrained on the ImageNet dataset (Deng et al. 2009). In Table 1 we summarize results on a number of activities across key performance metrics. It is interesting to observe that our proposed approach shows improved performance against I3D mainly for object-centric activities, which involve a larger scale object e.g. playing piano, riding elephant. The PE-KG approach also proves to be good in terms of having a low rate of false positives. However, it seems to be struggling with activities where the object involved is not clearly highlighted in the video, such as roller skating. Overall, the PE-KG model has the benefit of providing human-level interpretability of the activity detection process in contrast to black-box neural network models. Moreover, by generating the association rules used during activity recognition, it provides an interface for the human user to inspect and possibly directly adjust these high level activity representations learned during the training process.

The Video Knowledge Graph approach is also evaluated based on the Kinetics dataset. For classifying a given video v, first, the NN_p, NN_c, and NN_e models are applied to the keyframes to generate the set of triples $\mathcal{T}_v$, leveraging extracted positional relationships, image captions, emotion detection, and demographic information. Next, we want to evaluate the plausibility for the set $\mathcal{T}_v$ of belonging to one of the video graphs $\mathcal{G}_y^t$. Recall, that during training we are constructing a separate video graph $\mathcal{G}_y^t$ for each activity label y. In order to compare the plausibility of $\mathcal{T}_v$ against the video graphs $\mathcal{G}_y^t$, we cast the problem of activity recognition to a ranking problem, considering the following information-retrieval metrics:

$$MR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} rank_{(h,r,t)_i} \tag{11}$$

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \frac{1}{rank_{(h,r,t)_i}} \tag{12}$$

$$Hits@N = \sum_{i=1}^{|Q|} 1 \quad if \quad rank_{(h,r,t)_i} \leq N \tag{13}$$

where $|Q|$ denotes the number of triples extracted from a given video.

In Eq. 11, the *Mean Rank* (MR) computes the average of all the ranks of the triples extracted from a given video. The value ranges from 1, the ideal case when all ranks are equal to 1, to the number of corruptions, where all the ranks are last. Note, that the rank is determined based on ordering the scoring functions outlined in the previous section. In this manner, we obtain an aggregated score for the set of triples $\mathcal{T}_v$ across each of the activity labels. The final label is assigned based on the lowest rank. Similarly, the *Mean Reciprocal Rank* (MRR) in Eq. 12, computes the average of the reciprocal ranks of all the triples. The value ranges from 0 to 1; higher the value, better is the model. Finally, the *Hits@N* metric (Eq. 13) gives the percentage of computed ranks that are greater than (in terms of ranking) or equal to a rank of N. The value ranges from 0 to 1, with higher values indicating a better model.

Table 2: Performance Metrics for the Video KG Approach

	TransE	**complEX**	**convKB**
Motorcycling			
Mean Rank	6.45	14.10	10.53
Mean Reciprocal Rank	0.37	0.60	0.46
Hits @1	0.00	0.56	0.36
Hits @10	0.85	0.67	0.64
Hits @100	1.00	1.00	1.00
Number of videos	50	50	50
Number of analyzed triples	604	604	604
UsingComputer			
Mean Rank	4.29	6.26	2.33
Mean Reciprocal Rank	0.37	0.54	0.82
Hits @1	0.00	0.46	0.71
Hits @10	0.92	0.75	0.93
Hits @100	1.00	1.00	1.00
Number of videos	50	50	50
Number of analyzed triples	237	237	237
WalkingDog			
Mean Rank	18.75	25.21	16.72
Mean Reciprocal Rank	0.10	0.12	0.24
Hits @1	0.00	0.06	0.18
Hits @10	0.32	0.22	0.33
Hits @100	1.00	1.00	1.00
Number of videos	25	25	25
Number of analyzed triples	129	129	129
ReadingNewsPaper			
Mean Rank	4.29	7.74	3.69
Mean Reciprocal Rank	0.40	0.65	0.78
Hits @1	0.00	0.62	0.71
Hits @10	0.89	0.69	0.88
Hits @100	1.00	1.00	1.00
Number of videos	30	30	30
Number of analyzed triples	169	169	169
Shot			
Mean Rank	3.94	7.33	1.92
Mean Reciprocal Rank	0.42	0.72	0.84
Hits @1	0.00	0.68	0.76
Hits @10	0.94	0.82	0.98
Hits @100	1.00	1.00	1.00
Number of videos	50	50	50
Number of analyzed triples	1696	1696	1696

In Table 2 we compare the performance of the three embedding representations based on the Video KG approach across various video activity categories highlighting the differences in ranking effectiveness. *ConvKB* consistently outperforms the other models, achieving the lowest MR and the highest MRR in most cases, making it the most effective model for ranking triples correctly. *ComplEx* follows as the second-best model, while *TransE* struggles the most, failing to rank correctly the videos for most activities. The *Hits@k* metric provides valuable insight into how well each model performs within the top k predictions. Across all activities, *ConvKB* consistently achieves the highest *Hits@1* and *Hits@10* scores, indicating its superior ability to rank correct triples in top positions. Overall, although computationally more expensive, results confirm that *ConvKB* is the most effective embedding for activity detection, based on our proposed Video KG approach.

Conclusions and Future Work

By modeling videos as graphs where nodes represent objects and edges capture relationships, these representations enhance classification performance and improve visual understanding. In this paper we have introduced two approaches that employ knowledge graphs for the task of activity recognition, namely the PE-KG approach, which we further develop into the more general Video KG solution. The proposed approaches not only facilitate recognition but also fosters reasoning over the visual world, allowing models to infer meaningful connections between different elements. Additionally, grounding visual concepts in language through knowledge graphs bridges the gap between vision and language, making it easier to interpret and explain model decisions for a human user.

One promising direction for future work is the integration of additional modalities, such as audio cues, textual metadata, and sensor data, to further enrich the contextual representation and improve activity recognition performance. Additionally, exploring graph neural network architectures to dynamically model temporal evolution within knowledge graphs could enhance the capture of complex, time-varying relationships in video data. Finally, real-world applications would benefit from studies on domain adaptation and scalability, as well as the development of real-time implementations to facilitate deployment in surveillance, healthcare, and autonomous systems.

References

Agrawal, R.; and Srikant, R. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*, 487–499.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 813–824. PMLR.

Bi, H.; Fang, Z.; Mao, T.; Wang, Z.; and Deng, Z. 2019. Joint Prediction for Kinematic Trajectories in Vehicle-Pedestrian-Mixed Scenes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10382–10391.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chakraborty, B.; Holte, M. B.; Moeslund, T. B.; Gonzàlez, J.; and Xavier Roca, F. 2011. A selective spatio-temporal interest point detector for human action recognition in complex scenes. In *2011 International Conference on Computer Vision*, 1776–1783.

Cho, S. I.; and Kang, S.-j. 2019. Histogram Shape-Based Scene-Change Detection Algorithm. *Journal of Imaging Science and Technology*, 63(4): 040403.

Dalal, N.; Triggs, B.; and Schmid, C. 2006. Human Detection Using Oriented Histograms of Flow and Appearance. In Leonardis, A.; Bischof, H.; and Pinz, A., eds., *Computer Vision – ECCV 2006*, 428–441. Berlin, Heidelberg: Springer Berlin Heidelberg.

De Ramón Fernández, A.; Ruiz Fernández, D.; García Jaén, M.; and Cortell-Tormo, J. M. 2024. Recognition of Daily Activities in Adults With Wearable Inertial Sensors: Deep Learning Methods Study. *JMIR Medical Informatics*, 12.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Engström, J.; and Persson, J. A. 2023. Accurate indoor positioning by combining sensor fusion and obstruction compensation. In *2023 13th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 1–7.

Gharib, S.; Tran, M.; Luong, D.; Drossos, K.; and Virtanen, T. 2023. Adversarial Representation Learning for Robust Privacy Preservation in Audio. *IEEE Open Journal of Signal Processing*, 5: 294–302.

Guan, Y.; and Plötz, T. 2017. Ensembles of Deep LSTM Learners for Activity Recognition using Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2).

Inoue, M.; Inoue, S.; and Nishida, T. 2016. Deep Recurrent Neural Network for Mobile Human Activity Recognition with High Throughput.

Jamali, M.; Davidsson, P.; Khoshkangini, R.; Mihailescu, R.-C.; Sexton, E.; Johannesson, V.; and Tillström, J. 2025. Video-Audio Multimodal Fall Detection Method. In Hadfi, R.; Anthony, P.; Sharma, A.; Ito, T.; and Bai, Q., eds., *PRICAI 2024: Trends in Artificial Intelligence*, 62–75. Singapore: Springer Nature Singapore.

Johnson, D. R.; and Uthariaraj, V. R. 2020. Research Article A Novel Parameter Initialization Technique Using RBM-NN for Human Action Recognition.

Joudaki, M.; Imani, M.; and Arabnia, H. R. 2025. A New Efficient Hybrid Technique for Human Action Recognition Using 2D Conv-RBM and LSTM with Optimized Frame Selection. *Technologies*, 13(2).

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; Suleyman, M.; and Zisserman, A. 2017. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950.*

Khattak, A.; Asghar, M. Z.; Ali, M.; and Batool, U. 2022. An efficient deep learning technique for facial emotion recognition. *Multimedia Tools Appl.*, 81(2): 1649–1683.

Kolesnikov, A.; Dosovitskiy, A.; Weissenborn, D.; Heigold, G.; Uszkoreit, J.; Beyer, L.; Minderer, M.; Dehghani, M.; Houlsby, N.; Gelly, S.; Unterthiner, T.; and Zhai, X. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Lin, Z.; Feng, M.; Santos, C. N. d.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130.*

Mihanpour, A.; Rashti, M. J.; and Alavi, S. E. 2020. Human Action Recognition in Video Using DB-LSTM and ResNet. In *2020 6th International Conference on Web Research (ICWR)*, 133–138.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Mutegeki, R.; and Han, D. S. 2020. A CNN-LSTM Approach to Human Activity Recognition. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 362–366.

Nguyen, D. Q.; Nguyen, T. D.; Nguyen, D. Q.; and Phung, D. 2017. A novel embedding model for knowledge base completion based on convolutional neural network. *arXiv preprint arXiv:1712.02121.*

Popescu, A.-C.; Mocanu, I.; and Cramariuc, B. 2020. Fusion Mechanisms for Human Activity Recognition Using Automated Machine Learning. *IEEE Access*, 8: 143996–144014.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, 91–99. Cambridge, MA, USA: MIT Press.

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27.

Sun, Q.; Zhang, T.; Gao, S.; Yang, L.; and Shao, F. 2024. Optimizing Gesture Recognition for Seamless UI Interaction Using Convolutional Neural Networks. *arXiv preprint arXiv:2411.15598.*

Tandel, J.; Darak, S.; Desai, K.; Desai, M.; and Kothari, M. 2024. Human Anomaly Detection System Using YOLOv8 and LSTM. *International Journal for Research in Applied Science and Engineering Technology*. Paper ID: IJRASET65191, Published: 2024-11-12.

Tian, Y.; Ye, Q.; and Doermann, D. 2025. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv preprint arXiv:2502.12524.*

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4489–4497.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, E.; and Bouchard, G. 2016. Complex Embeddings for Simple Link Prediction. In Balcan, M. F.; and Weinberger, K. Q., eds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 2071–2080. New York, New York, USA: PMLR.

Uddin, M. A.; Talukder, M. A.; Uzzaman, M. S.; Debnath, C.; Chanda, M.; Paul, S.; Islam, M. M.; Khraisat, A.; Alazab, A.; and Aryal, S. 2024. Deep learning-based human activity recognition using CNN, ConvLSTM, and LRCN. *International Journal of Cognitive Computing in Engineering*, 5: 259–268.

Ullah, A.; Muhammad, K.; Hussain, T.; Lee, M.; and Baik, S. W. 2020. Deep LSTM-based sequence learning approaches for action and activity recognition. In *Deep Learning in Computer Vision*, 127–150. CRC Press.

Wang, H.; Kläser, A.; Schmid, C.; and Liu, C.-L. 2011. Action recognition by dense trajectories. In *CVPR 2011*, 3169–3176.

Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; and Qiao, Y. 2023. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14549–14560.

Wang, X.; Huang, H.; and Li, S. J. 2022. VC-GPT: End-to-End Visual Conditioned GPT for Image Captioning. *arXiv preprint arXiv:2201.12723.*

Wensel, J.; Ullah, H.; and Munir, A. 2023. ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos. *IEEE Access*, 11: 72227–72249.

Xia, K.; Huang, J.; and Wang, H. 2020. LSTM-CNN Architecture for Human Activity Recognition. *IEEE Access*, 8: 56855–56866.