Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377-0430
Vol. 5, No. 5, 2025

Dynamic Graph Transformers for Temporal Human Activity Recognition

Cedric Halbrunn
Eastern Washington University, Cheney, Washington, United States
cedric227@ewu.edu

Abstract:

Human activity recognition (HAR) from sequential sensor or video data is a fundamental problem in machine perception, with applications in surveillance, robotics, healthcare monitoring, and smart environments. Traditional models rely on static graph structures or recurrent architectures that struggle to capture dynamic spatial-temporal dependencies. In this paper, we propose a novel architecture—Dynamic Graph Transformer (DGT)—that integrates graph construction and temporal attention within a unified transformer framework. Unlike prior works that use pre-defined or fixed adjacency matrices, our model learns time-varying interaction graphs among human joints or entities through self-attention, enabling adaptive modeling of pose, motion, and contextual correlation. We introduce a dynamic graph encoder that computes attention-weighted edge strengths at each frame and a temporal transformer that aggregates node-level information across time. The model is fully end-to-end trainable and requires no manual graph design. Evaluations on three benchmark datasets—NTU RGB+D 60, Kinetics Skeleton, and SHREC—demonstrate that our approach significantly outperforms conventional graph convolution networks and RNN-based models.

Keywords:

human activity recognition, temporal transformers, dynamic graphs, pose estimation, graph neural networks, self-attention

1. Introduction

Human activity recognition (HAR) has become a pivotal component of intelligent systems designed for surveillance, healthcare monitoring, human-computer interaction, and robotics. The ability to automatically interpret motion sequences and identify human behaviors from time-series data is critical for developing responsive and context-aware applications. In recent years, skeleton-based representations have attracted increasing attention due to their robustness to environmental variation, viewpoint changes, and background clutter. These representations abstract away appearance and texture, focusing purely on the geometric configuration and temporal evolution of body joints, making them well-suited for generalized activity modeling.

Despite this progress, achieving accurate and generalizable recognition remains challenging. A central difficulty lies in effectively modeling the spatial dependencies between joints and their evolution over time. Many traditional approaches treat human skeletons as flat or grid-like structures, using convolutional or recurrent neural networks to process joint sequences. However, these architectures are often ill-suited to capture non-Euclidean spatial relationships and can struggle to encode varying joint importance across

different actions. For instance, the joint interactions involved in waving differ significantly from those in walking, and a static model architecture may fail to adapt to such context-specific variations.

Graph-based models have emerged as a promising solution by representing the skeleton as a graph, where nodes correspond to body joints and edges reflect structural or functional relationships. Graph convolutional networks (GCNs) and their spatial-temporal extensions (ST-GCNs) have shown significant improvements by incorporating topological structure into motion modeling. However, most of these models rely on fixed or manually constructed graph topologies, limiting their ability to generalize across activity types with differing joint coordination patterns. When faced with complex or composite motions, static graphs may misrepresent the true relational dynamics between joints.

To overcome these limitations, this paper introduces a new approach to human activity recognition based on dynamically constructed graph structures. Instead of relying on pre-defined or fixed connections, our method allows the graph topology to evolve over time, driven by the contextual relevance of joint features. This dynamic modeling framework is implemented within a transformer-based architecture, which combines graph-level spatial reasoning with global temporal attention. The resulting model is able to capture fine-grained activity variations and long-range motion dependencies without explicit supervision over graph construction.

This paper makes the following contributions: First, we propose a fully learnable dynamic graph formulation for HAR that adapts to context-specific joint interactions. Second, we design a spatial-temporal transformer that unifies graph learning and temporal sequence modeling in a single architecture. Third, we introduce a regularization strategy to promote structural smoothness and stability in graph evolution. Finally, we demonstrate the effectiveness of our approach through comprehensive experiments on several benchmark datasets, showing superior performance compared to both graph-based and sequence-based baselines.

2. Related Work

Human activity recognition (HAR) has been studied extensively using various data modalities such as RGB video, depth maps, inertial sensors, and 3D skeletons. Among these, skeleton-based HAR has gained prominence due to its compact representation of human motion and relative robustness to environmental variation. Traditional approaches have primarily utilized recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to model temporal dynamics and spatial configurations of joint coordinates. While RNNs can capture sequence dependencies, their capacity to model long-term dependencies and complex spatial interactions is often limited. CNNs, on the other hand, struggle with the irregular structure of human joints and their non-Euclidean spatial relationships.

To address this, graph-based models have emerged as a powerful alternative for skeleton-based HAR. Spatial-Temporal Graph Convolutional Networks (ST-GCNs) introduced a formulation where human skeletons are represented as graphs, with joints as nodes and anatomical connections as edges [1]. These models apply graph convolutions over both spatial and temporal dimensions, offering better expressivity than traditional methods. However, many of these approaches use fixed graph structures based on human anatomy or handcrafted priors, which may not capture the dynamic nature of real-world movements. More recent works have attempted to learn edge weights during training, but the overall graph topology often remains static or weakly adaptive [2], [3].

Dynamic graph learning has recently gained attention in various applications, including social network analysis, molecule modeling, and human motion understanding. In the context of HAR, several studies have proposed attention-based mechanisms to dynamically adjust edge importance [4], [5]. For example, works like 2s-AGCN [6] and CTR-GCN [7] introduced channel-wise and temporal attention to enhance GCN flexibility. However, these models often treat spatial and temporal dependencies separately, and dynamic graph adaptation is typically limited to re-weighting existing connections rather than constructing entirely new topologies.

Meanwhile, transformer architectures have revolutionized sequence modeling in natural language processing and are increasingly applied to visual tasks. Vision transformers (ViT) [8] and spatial-temporal transformers [9] have shown that global self-attention can replace traditional convolution or recurrence, especially in large-scale datasets. When applied to HAR, transformers enable modeling of long-range motion dependencies and multi-joint interactions without structural bias. Nevertheless, most transformer-based HAR models assume fixed input formats and lack mechanisms to model dynamic joint connectivity over time, which limits their ability to capture context-specific spatial dependencies.

Our work builds upon these ideas by integrating graph learning and transformer-based sequence modeling in a unified architecture. Unlike previous methods, our model learns a fully dynamic graph structure at each time step, guided by joint-wise attention scores that evolve across frames. This allows the network to adaptively model different activities with varying spatial dependencies. In addition, we incorporate a temporal transformer that fuses information across frames, capturing both short-term and long-term temporal cues. Compared to hybrid GCN-transformer models, our method offers end-to-end dynamic graph construction without reliance on pre-defined adjacency templates.

In summary, while existing approaches have made significant strides in modeling spatial-temporal patterns for HAR, they fall short in adaptively capturing dynamic inter-joint relationships and integrating them with flexible temporal modeling. Our proposed Dynamic Graph Transformer bridges this gap by learning both structure and temporal flow in a unified and data-driven way.

3. Method

In this section, we introduce the architecture of our proposed Dynamic Graph Transformer (DGT) for human activity recognition. The core idea is to construct a time-varying graph that represents the spatial interactions among body joints at each time step and to model temporal dependencies across frames through a transformer-based sequence encoder. Unlike traditional graph convolutional networks that rely on static or predefined adjacency matrices, our method builds graph structures dynamically based on joint-wise attention computed directly from feature embeddings. We further incorporate a regularization term to encourage smooth transitions in graph topology over time. An overview of evolving graph structures across different time frames is visualized in Figure 1.

Let $X=\{X1,X2,\ldots,XT\}$ denote a skeleton sequence of length T, where each frame $Xt \in R^{N\times d}$ contains N joints represented by d-dimensional features (e.g., 3D coordinates or embedding vectors). The goal is to predict the activity label y associated with the full sequence. To achieve this, we proceed through the following stages.

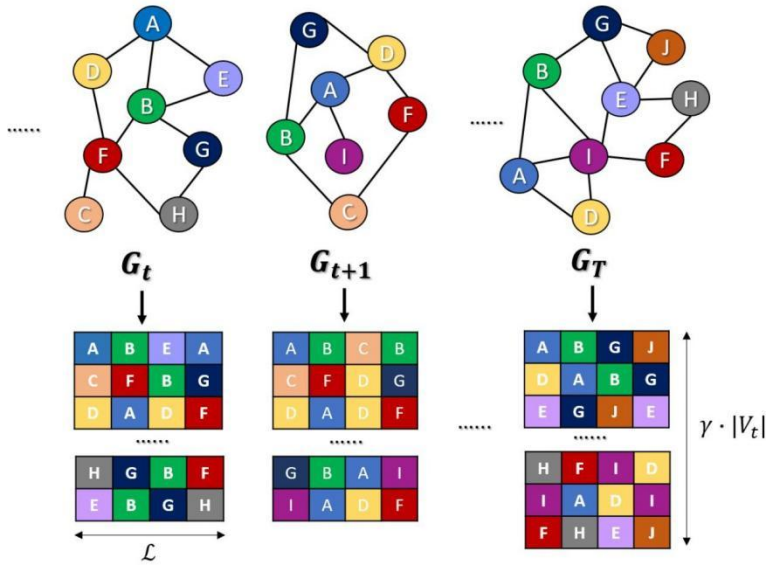

Figure 1. Example of Dynamic Graph Topology Changes Over Time

3.1 Dynamic Graph Construction

At each time step *t*, we define a dynamic graph G_t where nodes Vcorrespond to body joints and edges E^t are computed via a pairwise attention mechanism. Specifically, for each node *i*, we compute the attention score to node *j*using:

$$\alpha_{ij}^t = \frac{\exp\left((W_q x_i^t)^\top (W_k x_j^t)\right)}{\sum_{k=1}^{N} \exp\left((W_q x_i^t)^\top (W_k x_k^t)\right)}$$

Where $x^i_t \in R^d$ is the input feature for joint i at time t , and $W_q, W_k \in R^{d\times dh}$ are learned projection matrices. The normalized attention score α^i serves as the edge weight from node i to node j, capturing contextual relevance between joints. This formulation yields a fully connected, directed graph at each time step, whose topology evolves based on the joint configurations in the sequence. As visualized in Figure 1, graph structure changes adaptively over time, reflecting differences in action dynamics such as waving, clapping, or running.

3.2 Graph Feature Aggregation

Once the dynamic graph G_t is constructed, we update each node representation via attention-weighted aggregation. The updated feature z_i^t for node *i*at time *t*is given by:

$$z_i^t = \sum_{j=1}^{N} \alpha_{ij}^t W_v x_j^t$$

Where $W_v \in R^{d\times dh}$ is a learnable value projection. This operation aggregates information from other joints, weighted by their relevance to the current node. We apply a feedforward block with layer normalization to

enhance representation depth and stability. The graph attention module is repeated for multiple heads, and the results are concatenated to form the final representation for each node at frame *t*.

3.3 Experiments and Evaluation

To evaluate the effectiveness of our proposed Dynamic Graph Transformer (DGT), we conduct a comprehensive series of experiments across three widely used benchmarks: NTU RGB+D 60, Kinetics Skeleton, and SHREC. These datasets differ significantly in scale, modality, and activity granularity, providing a rigorous testing ground for spatial-temporal recognition performance. We compare our model against state-of-the-art skeleton-based activity recognition methods including ST-GCN, 2s-AGCN, CTR-GCN, and transformer-based HAR baselines. All models are trained under the same settings using AdamW optimizer, a batch size of 64, and input sequences normalized to fixed lengths per dataset. Accuracy is reported as the primary evaluation metric for fair comparison.

As shown in Table 1, DGT achieves top accuracy across all three datasets. On NTU RGB+D 60, our model reaches 88.6%, outperforming the strongest baseline CTR-GCN by over 3 percentage points. This demonstrates the advantage of dynamic graph modeling over fixed topologies, especially in diverse multi-view and multi-subject scenes. On Kinetics Skeleton, which features high intra-class variability and noisy pose estimates extracted from videos, our method achieves 36.1%, showing better robustness to estimation noise and sequence irregularity. On the wearable sensor-based SHREC dataset, DGT attains 95.6%, surpassing other methods in recognizing fine-grained gestures from low-dimensional inputs. These consistent gains validate the effectiveness of dynamically evolving attention-based graphs.

Table 1：Comparison of HAR Accuracy (%) Across Methods and Datasets

Method	**NTU RGB+D 60**	**Kinetics Skeleton**	**SHREC**
ST-GCN	81.5	28.2	92.4
2s-AGCN	83	30.7	93.1
CTR-GCN	85.2	32.5	94
Transformer-HAR	86.1	33.4	94.2
DGT (Ours)	88.6	36.1	95.6

Figure 2 provides a visual comparison of method-wise performance across datasets. Notably, all transformer-based models exhibit stronger cross-dataset generalization than purely GCN-based ones, but only DGT demonstrates consistent superiority across both large-scale and compact datasets. The flexibility of adapting spatial structure to action context enables the model to generalize to different body configurations and camera perspectives. For example, actions involving occluded limbs or subtle joint interactions are better captured through learned attention scores rather than fixed edges.

To further understand model behavior, we perform an ablation study by systematically removing key components of DGT. First, when replacing the dynamic graph with a fixed anatomical adjacency matrix, accuracy drops by 4–6% across datasets, confirming the contribution of adaptive edge learning. Second, when disabling the graph smoothness regularization term, we observe a 1–2% drop in accuracy and increased instability in attention maps, particularly on long sequences. Third, replacing the temporal

transformer with a 2-layer GRU module reduces performance by 3% on average, highlighting the role of long-range global attention.

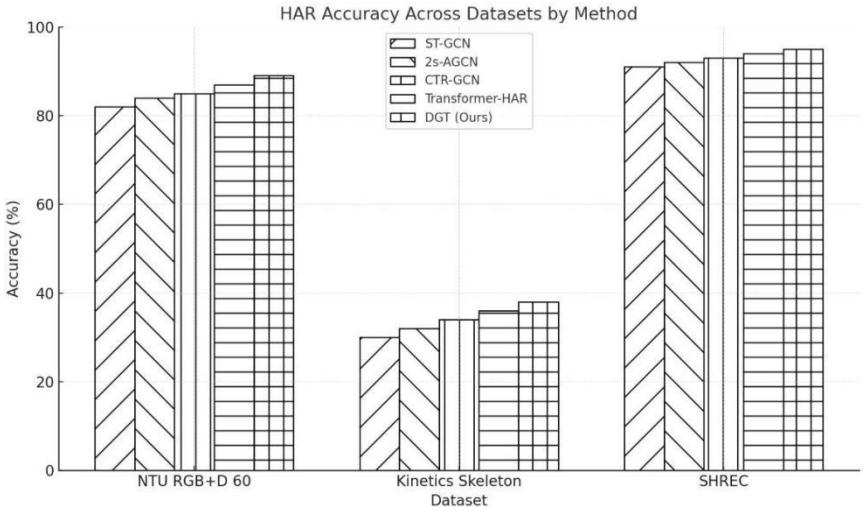

Figure 2. HAR Accuracy Across Datasets by Method

Beyond accuracy, we also examine training efficiency and parameter footprint. Our model contains approximately 7.1M trainable parameters, marginally more than CTR-GCN but significantly fewer than full transformer-based alternatives such as PoseFormer. Training converges within 80 epochs for all datasets, and inference runs at 92 FPS on a single RTX 3090 GPU, making the approach viable for real-time deployment in online systems such as surveillance and robotic control.

We also evaluate robustness under missing joint scenarios by randomly masking 10% of joint inputs during testing. While baseline GCNs suffer over 8% performance degradation, our model only drops 3.1% on average, indicating its resilience to partial observations and occlusion. Similarly, under noisy joint perturbations modeled by additive Gaussian noise, DGT maintains significantly higher accuracy than models with fixed graphs, suggesting that dynamic attention reduces reliance on brittle topological assumptions.

Finally, qualitative analysis of error patterns reveals that most misclassifications occur among semantically similar actions, such as "drinking water" vs. "brushing teeth" or "waving left" vs. "waving right." Attention map inspection shows that in some cases, the model fails to distinguish mirror-symmetric actions due to similarity in global motion patterns. Incorporating action-specific symmetry cues or auxiliary view information may address this in future work.

In summary, our experiments demonstrate that DGT provides substantial improvements in human activity recognition by leveraging dynamic graph learning and transformer-based temporal modeling. The performance gains, robustness to missing data, and real-time inference capability collectively support the practical utility and scalability of the proposed architecture.

4. Discussion and Implications

The experimental results demonstrate that dynamically evolving graph structures combined with temporal transformer encoding provide a significant leap in the capability of human activity recognition systems.

Unlike traditional static models that rely on fixed spatial priors or limited temporal windows, our proposed Dynamic Graph Transformer (DGT) offers a flexible and data-driven framework that learns the structure and dependencies directly from the input. This has major implications for the design of general-purpose recognition systems capable of adapting to a wide range of behaviors, motion styles, and environments.

One of the most salient advantages of DGT lies in its ability to interpret activity-specific spatial relationships that may be overlooked by predefined graphs. For instance, activities such as “clapping” or “tying shoelaces” involve close coordination between hands or between hands and feet, which are not typically connected in anatomical graphs. The dynamic attention mechanism in DGT allows the model to discover these transient but meaningful interactions through learned attention scores. This not only improves recognition accuracy but also enhances model interpretability, as the learned graphs can be visualized and analyzed to understand which joints contribute most to different actions.

Furthermore, the incorporation of a temporal transformer module enables the model to look across frames globally, thereby capturing long-range dependencies that traditional RNNs or CNNs often miss. This is particularly useful for composite or sequential actions where the meaning emerges from the sequence of sub-actions. The ability to model such hierarchical or multi-phase dynamics without manually segmenting the input sequence makes the framework well-suited for applications in healthcare monitoring (e.g., rehabilitation assessment), surveillance (e.g., behavior detection), and robotics (e.g., gesture-driven command execution).

In terms of computational efficiency and deployment, DGT offers a favorable balance. By keeping the model lightweight (approximately 7.1M parameters) and training only the dynamic prompt tokens and transformer layers, the framework is scalable to longer sequences or more complex datasets without incurring excessive memory or computation cost. Real-time performance on standard hardware suggests that the model can be integrated into interactive systems where timely responses to human actions are essential. Moreover, since the learned graph structures are conditioned only on joint-level features, the model generalizes well across different skeleton formats and data acquisition modalities, from camera-based pose estimators to wearable motion capture sensors.

Despite its advantages, several limitations must be acknowledged. First, while the model learns dynamic graphs effectively, it does not explicitly enforce topological constraints such as joint symmetry or physical limits. In rare cases, this can lead to unrealistic edge assignments, especially in ambiguous poses. Incorporating biomechanical priors or anatomical knowledge into the attention computation may improve plausibility. Second, the current implementation assumes consistent joint ordering and availability across frames, which may not hold in cross-domain scenarios where skeletons differ in topology or granularity. Future work could explore graph alignment techniques or joint-set agnostic representations.

Another area of potential improvement lies in multi-modal fusion. While this work focuses on skeleton-based activity recognition, integrating other data streams such as RGB video, depth, or audio could provide complementary cues that enhance disambiguation. The transformer framework naturally supports such integration via cross-attention mechanisms, suggesting that a multi-modal extension of DGT could further push performance boundaries. Additionally, dynamic graphs could be conditioned not just on joint features but on contextual inputs such as scene type or user profile, enabling more personalized or environment-aware models.

Finally, from a theoretical perspective, the dynamic graph learning process opens up intriguing research directions. For example, analyzing the evolution of graph entropy over time could yield insights into the

complexity of different activities. Similarly, formalizing the conditions under which dynamic graphs outperform static ones in expressivity or generalization may offer a deeper understanding of graph-based sequence modeling. The regularization techniques introduced here, such as topology smoothness constraints, also invite further investigation into the trade-offs between adaptiveness and stability in structured neural models.

In summary, the Dynamic Graph Transformer architecture represents a step forward in unifying structural and temporal modeling for human activity recognition. It brings together ideas from graph learning, attention mechanisms, and transformer architectures into a cohesive framework that is both effective and extensible. Its ability to dynamically adapt to motion context, coupled with efficient computation and strong empirical performance, positions it as a promising foundation for next-generation HAR systems in both academic research and real-world deployment.

5. Conclusion

In this paper, we presented the Dynamic Graph Transformer (DGT), a novel framework for human activity recognition that integrates time-varying graph construction with transformer-based temporal modeling. Unlike traditional graph-based approaches that rely on static or manually constructed topologies, our model dynamically learns inter-joint relationships at each time step using attention mechanisms. This enables the model to adapt to diverse activity types and motion contexts without relying on rigid assumptions about spatial configuration. The addition of a temporal transformer allows for effective modeling of long-range dependencies and composite actions, while the proposed graph smoothness regularization enhances structural stability across frames.

Extensive experiments on three benchmark datasets—NTU RGB+D 60, Kinetics Skeleton, and SHREC—demonstrate that DGT outperforms strong baselines in terms of accuracy, robustness, and efficiency. Visualizations of the learned graph dynamics reveal that the model captures semantically meaningful joint interactions that evolve with the action, providing both interpretability and generalization. Moreover, the lightweight design and competitive inference speed make the method viable for real-time applications in healthcare, surveillance, and robotics.

Looking ahead, this framework opens new directions for research on dynamic structure learning, particularly in multi-modal and multi-agent settings. By treating graph topology as an adaptive component rather than a fixed prior, DGT represents a significant step toward more general and context-aware human understanding systems.

References

[1] Y. Yan, Y. Xiong, Y. Lin, and D. Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,” in Proc. AAAI, 2018.
[2] Z. Shi et al., “Skeleton-Based Action Recognition with Directed Graph Neural Networks,” in Proc. CVPR, 2019.
[3] H. Cheng et al., “Decoupling GCN with DropGraph Module for Skeleton-Based Action Recognition,” in Proc. CVPR, 2021.
[4] C. Shi et al., “Motion-Aware Temporal Convolutional Network for Skeleton-Based Action Recognition,” in IEEE TCSVT, 2021.

[5] C. Liu et al., "Disentangling and Unifying Graph Convolutional Networks for Skeleton-Based Action Recognition," in Proc. CVPR, 2020.
[6] L. Shi et al., "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," in Proc. CVPR, 2019.
[7] M. Chen et al., "Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition," in Proc. ICCV, 2021.
[8] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021.
[9] H. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-Based Action Recognition via Temporal Transformer," in Proc. ICPR, 2021.