A Hybrid Visual Transformer for Efficient Deep Human Activity Recognition

Youcef Djenouri
University of South-Eastern Norway, Norway
NORCE Norwegian Research Centre, Norway
youcef.djenouri@usn.no, yodj@norceresearch.no

Ahmed Nabil Belbachir
NORCE Norwegian Research Centre
Norway
nabe@norceresearch.no

Abstract

Human Activity Recognition (HAR) has gained significant attention in recent years due to its wide-ranging applications. This paper introduces a novel hybrid visual transformer methodology designed to enhance the robust analysis and comprehension of activities. CVTN (Convolution Visual Transformer Network) leverages sensor data represented jointly in spatial and temporal dimensions to enhance the resilience of the HAR process. The proposed technique employs a hybrid model that integrates Convolutional Neural Networks (CNNs) and Visual Transformers (VTs). Initially, the CNN component learns spatial visual features from diverse sensor data. Subsequently, these acquired visual features are inputted into the transformer segment of the model. VT captures temporal insights by observing sensor statuses across different time points. The efficacy of the CVTN methodology is assessed using the Kinetics dataset, which emulates real-world human activity recognition scenarios. The experimental results reveal clear superiority compared to the recent baseline HAR solutions, reaffirming its potential for advancing activity analysis.

1. Introduction

The recognition of human activities has emerged as a crucial research area driven by the increasing ubiquity of wearable sensors, smartphones, and IoT devices [33, 32]. HAR involves detecting and classifying various activities, such as walking, running, sitting, and complex actions, based on sensor data collected from accelerometers, gyroscopes, and other sources [36, 38, 47]. The ability to automatically identify these activities has applications in health monitoring [12], personalized fitness tracking [3], context-aware computing [25], and more. This paper delves into the challenges and opportunities of *HAR*, emphasizing the integration of data from multiple sensors to enhance recognition accuracy. Sensor fusion is the process of combining data from multiple sensors in order to obtain a more accurate and complete picture of a system or environment [5, 21]. This technique is commonly used in *HAR* to gather data from different types of sensors and devices in order to provide a more comprehensive view of a person's health and well-being [35, 26]. As an illustration, sensor fusion might involve mixture information from a wearable heart rate monitor, a blood pressure monitor, and a pedometer. This combination offers a more comprehensive overview of an individual's cardiovascular well-being. Through the integration of data derived from these diverse sensors, it becomes feasible to identify patterns and deviations that might perturb the recognition process when relying solely on a solitary sensor. Most *HAR*-based solutions from sensor fusion are recently developed, including:

1. **Computer vision-based solutions**: Computer vision based solutions can be used for *HAR* to analyze data from visual sensors, such as cameras or motion sensors [28, 22, 16]. This phenomenon can yield significant insights into an individual's kinematic patterns, bodily alignment, and actions, thereby facilitating an evaluation of their physiological condition and overall welfare. To illustrate, computational visual processes can be harnessed to scrutinize visual information procured through cameras, enabling the identification of deviations in an individual's manner of walking or bodily orientation, which could potentially signify impediments in locomotion or equilibrium. Such applications hold considerable promise in the surveillance of aged or differently-abled persons who are susceptible to instances of stumbling or other bodily harm. Moreover, solutions grounded in computational vision methodologies can be employed to oversee an individual's routine activities encompassing fundamental tasks like culinary pursuits, domestic tidying, and self-maintenance undertakings. By leveraging sophisticated deep learning algorithms for the examination of visual data, it becomes feasible to unearth underlying patterns within an individual's behavioral repertoire that might serve as indicators of alterations in their medical state or general well-being. Solutions based on convolution neural networks [34], generative adversarial networks [30], and autoencoders [31] are few examples of these solutions.

2. **Sequence-based solutions**: Sequence-based representation for *HAR* involves analyzing time series data, such as sensor data or physiological data, to extract meaningful information about a person's movements, activities, and behaviors over time. This can be used to assess their health and well-being, and to detect changes that might indicate a health problem or other issue. For instance, extracting relevant features from the time series data, such as the mean, standard deviation, or frequency components of the signal. These features can then be used as inputs to a machine learning model to detect patterns or changes in the data. The most popular deep learning algorithms for sequential data are RNNs. Traditional RNNs can only detect short-term dependencies because of gradient vanishing or explosion problems. Because they fixed this issue, variations like LSTM's (Long Short-Term Memory) [17] and GRU's (Gated Recurrent Units) [8] were considered the most popular. Although LSTM and GRU perform equally well on a variety of tasks, GRU's structure becomes less complex and can be trained more quickly. The main drawback of LSTM, and GRU is the computationally expensive which are considered more complex than traditional RNNs and require more memory resources to train. Another important factor is the limited context of LSTMs and GRUs where they have a fixed memory size and are only able to capture a limited amount of context from the input sequence. To overcome the contextual issue, transformer [15] has been developed. Its main strength lies in its ability to efficiently capture long-range dependencies using the self-attention mechanism. Transformer is also able to process multiple positions in a sequence simultaneously during both training and inference, which allows to learn and process massive amount of data in real time. Solutions suggested by Li's, and Xiao's teams [24, 41] are few examples of these solutions.

Motivations All the above solutions are effectively used in real-world applications of *HAR* and achieved promising results in identifying human activities in real time. However, these solutions suffer from accuracy performance for several reasons:

1. **Variability**: Human behavior is highly variable and can differ from person to person and even from instance to instance for the same person. For example, walking can vary based on the individual's stride length, walking speed, and walking style.

2. **Sensor precision**: Sensor data can be noisy and contain artifacts that make activity recognition challenging. Different sensors can also vary in their accuracy and precision, which can affect the quality of the data and the accuracy of the recognition.

3. **Ambiguity**: Certain activities can be ambiguous and difficult to distinguish from one to another. For example, walking and jogging can have similar sensor data patterns, making it challenging to differentiate between them.

Figure 1 presents an illustrative example of the difficulty of *HAR* problem. The images are retrieved from the EPIC Kitchen dataset [11, 9, 10]. This figure reveals the necessity of handling both spatial and temporal information of the detected frame. Indeed, if we consider one image in the "cleaning" activity, and another image in the "preparing coffee" activity, it will be hard to distinguish between both activities. However, if we consider the entire sequence of both activities, it will be more evident to separate the two activities.

> *We assume that the integration of both spatial and temporal information features holds the potential to induce a profound enhancement in the HAR performance. By incorporating the inherent spatial features, which pertain to the physical arrangement and distribution of sensors or data sources, along with the temporal dynamics that capture the sequential evolution of activities.*

Our hypothesis is grounded in the belief that a more comprehensive representation of human actions can be achieved. Motivated by the success of CNN in learning the spatial visual features, and VT in learning the temporal features, we propose a hybrid CNN and transformer based model to learn spatiotemporal interconnections between sensor readings for addressing the *HAR* challenges.

Contributions This research work introduces a novel approach for addressing the challenges in *HAR* systems. The primary contributions of this research are outlined as follows:

1. We introduce an innovative idea aimed at addressing the challenges of the *HAR* system. This idea is referred to as CVTN (Convolution Visual Transformer Network), which leverages insights from both spatial and temporal data captured by sensors to enhance the process of *HAR*.

2. We design a hybrid model that combines CNNs, and VTs. Initially, the CNN component focuses on acquiring spatial features from various sensor data sources. These acquired features are subsequently integrated into the visual transformer component. Within the transformer, the model captures temporal information by analyzing the sensor states at different timestamps.

3. We conduct an evaluation and analysis of the performance of CVTN within a real-world human activity recognition scenario. A comparative assessment is carried out against established baseline *HAR*-solutions. The findings highlight the superior capabilities of the CVTN model, showcasing an recognition rate performance of 95%.

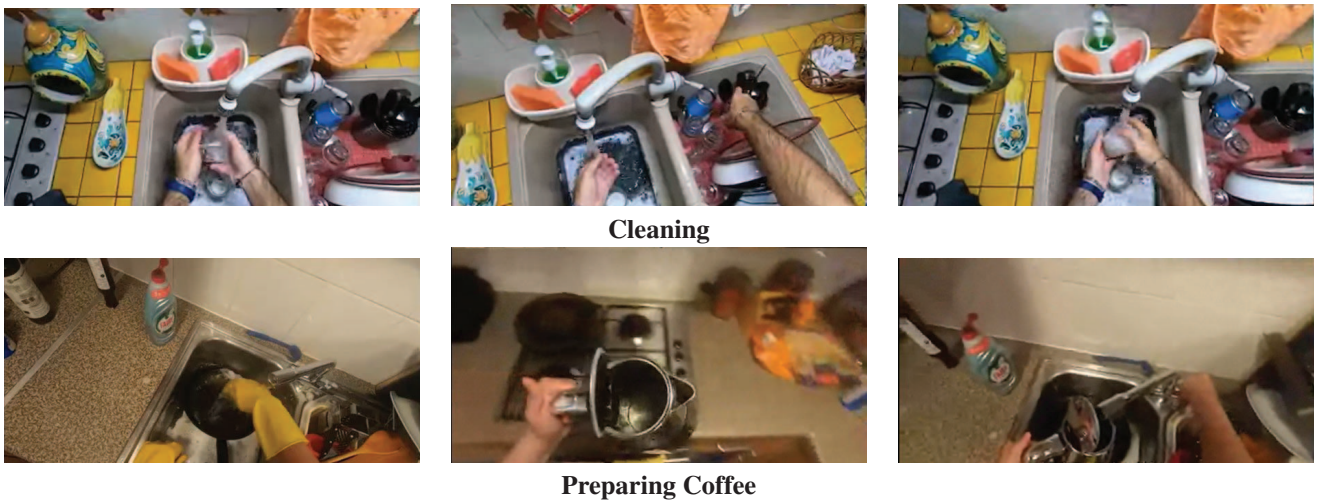

Figure 1: Illustration of the difficulty of *HAR* problem: The first sequence of images are related to the cleaning activity, and the second sequence of images are related to the preparing coffee activity.

2. Background

Human Activity Recognition (HAR) is a field of research and application within the broader domain of artificial intelligence and machine learning. Its primary objective is to develop algorithms and models that can automatically detect and classify human activities based on data collected from various sensors and sources. In the context of home-based monitoring, HAR involves using sensors and devices to collect data about the movements and behaviors of individuals within their own homes. This data can include information from sources like: accelerometers and gyroscopes, cameras, environmental sensors, and smart home devices. For instance, HAR in the home-based monitoring setting aims to classify and identify the specific activities that an individual is engaging in within its home. These activities can vary widely and might include: cooking, Watching TV, exercising, sleeping, working, socializing, personal hygiene, and moving between rooms. HAR in a home-based monitoring setting has various benefits, including assisting elderly or disabled individuals to live independently, providing insights for healthcare professionals, and even enabling more efficient energy usage in smart homes. This study focuses on solving HAR problem in home-based monitoring setting.

Definition 1 *Given a dataset D consisting of n instances, each represented as a sequence of l consecutive time steps, where $X_i = \{x_{i,1}, x_{i,2}, ..., x_{i,l}\}$, where $i \in [1, n]$ and l is the sequence length. Each $x_{i,t}$ represents the feature vector of the sensor readings at time step t for instance i. We define the HAR by the process of learning a mapping function f that can accurately classify the activity label y_i for each sequence X_i. Mathematically, the problem can be defined as finding the function f that maps the sensor readings to activity labels:*

$$f : X_i \rightarrow y_i, \quad \forall i$$

where, y_i belongs to a predefined set of activity classes, denoted as $\mathcal{Y}$.

Definition 2 *To train the model f, we define a loss function to quantify the difference between the predicted activity probabilities and the true activity labels.*

$$\mathcal{L}(y_i, \hat{y}_i) = -\sum_{y \in \mathcal{Y}} y_i \cdot \log(\hat{y}_i), \tag{1}$$

where, y_i is the one-hot encoded true activity label, and $\hat{y}_i$ is the predicted probability distribution over activity classes obtained from the model f.

The task involves training the model f using a labeled training dataset $D_{\text{train}} = \{(X_i, y_i)\}$, where each instance X_i is associated with its true activity label y_i. The goal is to minimize the average cross-entropy loss over the training instances:

$$\min_f \frac{1}{N_{\text{train}}} \sum_{(X_i, y_i) \in D_{\text{train}}} \mathcal{L}(y_i, f(X_i)). \tag{2}$$

3. Related Work

CTVN is a hybrid model which consider the benefits of the best models for solving HM task. Existing works can be roughly grouped into two families, convolution neural networks, and visual transformers. In the following, we will give insights of using CTVN compared to studies belong to both families.

3.1. Convolutional Neural Networks

Several works have been trying to design effective CNN architectures for solving *HAR* [45, 4, 20, 18, 6, 43]. Yang et al. [45] introduced a method for systematically learning features to address the *HAR* problem. This approach utilized CNN to automatically extract features from raw inputs in a structured manner. The deep architecture enables the acquired features to serve as elevated abstract representations of the initial low-level time series signals. By incorporating labeled information through supervised learning, these acquired features gain enhanced discriminative capability. Bevilla et al. [4] suggested the utilization of CNN by employing raw data gathered from a collection of inertial sensors. They investigated various permutations of both activities and sensors, illustrating the process of adapting motion signals for input into CNN through the utilization of diverse network architectures. Instead of manually analyzing predetermined features in time-series sensor data, Jiang et al. [20] organized sequences of accelerometer and gyroscope signals into a novel representation called an "activity image". This innovative approach allows CNN to autonomously derive the most suitable features directly from the activity image, enhancing their capability for accurately recognizing different activities. Huang et al. [18] introduced a channel equalization for *HAR* to reignite inactive channels. Through strategic whitening or decorrelation operations, this approach reengages each channel, rebalancing their contributions to enhance feature representation. It also fosters a harmonious environment, amplifying the significance of every channel and uncovering nuanced patterns within data.

3.2. Visual Transformers

VT-based solutions for *HAR* has been heavily investigated in the recent literature [2, 37, 44, 7, 42]. Ahn et al. [2] proposed a deep learning transformer capable of representing two pass functionalities as a distinguishable vector. First, frames are output as global grid tokens and skeletons are output as joint location tokens from the input video and skeleton sequences, respectively. These tokens are then combined to form multi-class tokens, which are then fed into the designed transformer. The encoder includes a full spatiotemporal attention module as well as a suggested zigzag spatiotemporal attention module. Truong et al. [37] presented a DirecFormer, an innovative end-to-end Transformer-based framework for robust action recognition. It introduced ordered temporal learning, a Directed Attention mechanism, and conditional dependency for accurate sequence modeling. Yang et al. [44] designed the Recurrent Vision Transformer (RViT) for video action recognition. It combined the power of VTs with recurrent processing to capture both spatial and temporal features. RViT utilized a self-attention mechanism from VT and introduced an attention gate. This gate connects the current frame to the previous hidden state, allowing the model to gather global features across frames. This creates a temporal context for capturing patterns and relationships between frames. Xing et al. [42] introduced SVFormer, which employs the steady pseudo-labeling framework (EMATeacher) for handling unlabeled video samples. They also presented Tube TokenMix, a novel augmentation strategy for video data. This involves blending video clips using a temporally consistent mask of tokens. A temporal warping augmentation is investigated to accommodate complex temporal variations in videos by stretching selected frames to varying temporal durations within the clip.

3.3. Discussion

Existing studies lack comprehensive incorporation of both spatial and temporal features within the realm of *HAR*. In response to this gap, we introduce an innovative hybrid model that amalgamates the strengths of CNN and VT. This strategic approach harnesses the spatial feature extraction capabilities of CNNs and the adeptness of VT in recognizing temporal patterns. By leveraging the successful fusion of these techniques, which has proven effective in handling spatiotemporal data, our proposed methodology takes a step forward in refining HAR. Its primary objective is to concurrently capture the complex relationships of spatial and temporal features, thereby contributing significantly to the evolution of recognition methodologies.

4. CVTN: Convolution Visual Transformer Network

4.1. Principle

CVTN (Convolutional Vision Transformer Network) commences by utilizing raw sensor data, which is subsequently transformed into a series of frames. Each individual frame encapsulates a collection of images captured at distinct timestamps. A combined approach involving CNN and VT models is employed for training. The initial phase employs CNN to extract and comprehend visual attributes from each specific image within a frame. Subsequently, a VT-based analysis is carried out, concentrating on the acquisition of temporal characteristics and patterns from a sequence of these image frames. This dual-process approach contributes to the comprehensive understanding of both spatial and temporal information inherent in the sensor data. Figure 2 illustrates the overall design of CVTN.

4.2. Data Collection and Preprocessing

HAR encompasses the acquisition of data from individuals within their personal environments. The nature of the gathered data is contingent upon the distinct objectives of the

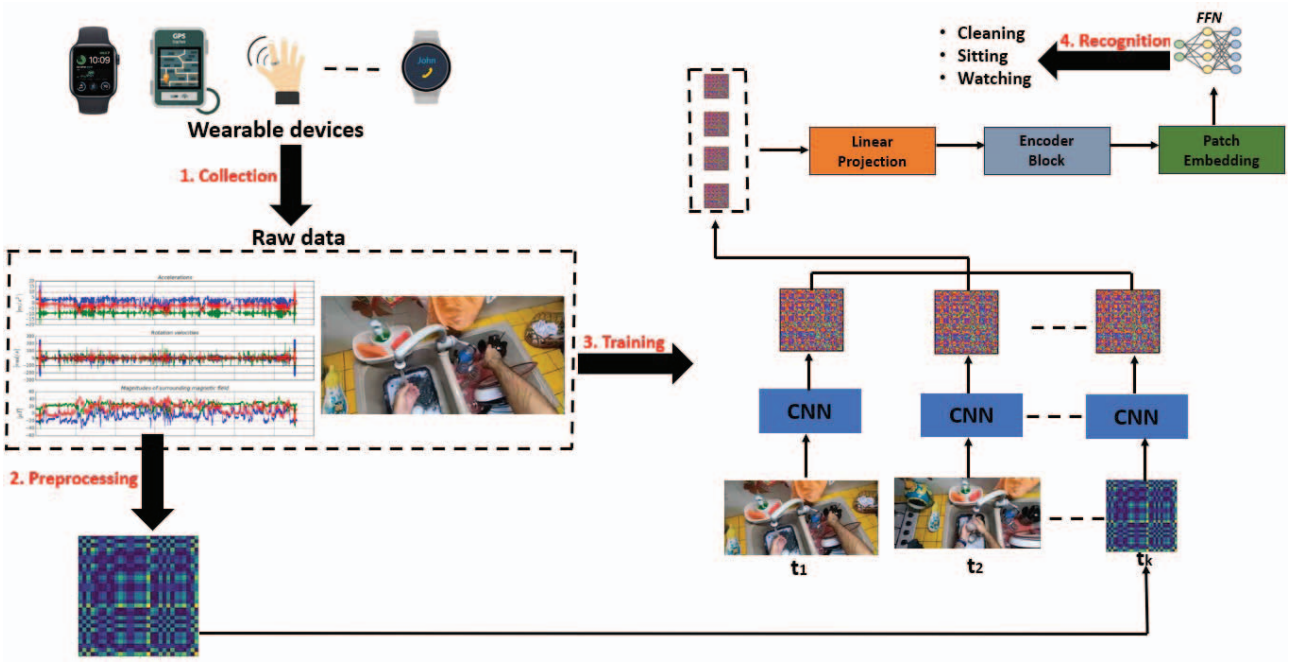

Figure 2: *CVTN: Initially, the data from sensors undergoes a transformation into a collection of images. A training process ensures employing a blend of CNN and VT models. This involves an initial application of CNN to grasp spatial attributes from each individual images. Subsequently, a VT process takes over, focusing on comprehending temporal attributes gleaned from a sequence of these images.*

recognition endeavor. We shall combine diverse methodologies for data collection, including: 1) Wearable sensors: We intend to utilize fitness trackers or smartwatches for the purpose of gathering information pertaining to physical activity. These sensors have the capability to be worn consistently, thereby offering an uninterrupted flow of data. 2) Remote monitoring devices: We employ blood pressure monitors, glucose monitors, and pulse oximeters to offer up-to-the-minute information concerning vital signs and various health markers. 3) Smart home devices: We leverage smart home devices, such as intelligent scales, to amass information about weight, body composition, and additional indicators of user behavior. These devices can seamlessly integrate with other recognition systems, facilitating a more holistic understanding of an individual's actions. Upon completion of this stage, we will possess a heterogeneous dataset comprising time series and images acquired at distinct timestamps $(t_1, t_2, \ldots, t_n)$. At each of these timestamps, we will possess the corresponding representative data $\hat{d}(t)$, which has been captured by various sensors. In this transformation, each time series is converted into a matrix where the values are derived from the pairwise angles between the data points. These matrices can be interpreted as images, enabling CVTN to analyze temporal patterns. The last step of this stage is to normalize the raw data using Min-Max normalization [19]. It is a widely used technique in data preprocessing to transform numerical data into a specific range, typically [0, 1]. This normalization ensures that all features have the same scale, which can be crucial for the CTVN that rely on distance or magnitude comparisons between features. Given a dataset $\mathbf{D} = [x_1, x_2, \ldots, x_n]$, where x_i represents a data point, the Min-Max normalization process involves the following operations:

1. We identify the minimum and maximum values in $\mathbf{D}$: Let $x_{\text{min}} = \min(\mathbf{D})$ and $x_{\text{max}} = \max(\mathbf{D})$.

2. We normalize each data point using the Min-Max formula:
$$x_{\text{normalized}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{3}$$
where x is an original data point, and $x_{\text{normalized}}$ is its normalized counterpart.

3. We repeat the normalization process for all data points in the dataset.

The Min-Max normalization scales each data point linearly between 0 and 1. The minimum value in the dataset (x_{min}) is transformed to 0, and the maximum value (x_{max}) is transformed to 1. The values in between are scaled proportionally based on their distance from the minimum and maximum.

4.3. Spatial Visual Learning

This step aims to capture the visual features of each data in $d(t)$ using CNN. The process of computing visual features with a CNN can be broken down in the following: Let d_i be the normalized input data, which can be represented as a three-dimensional tensor of size (H, W, C), where H is the height, W is the width, and C is the number of channels. For instance, if the captured data is an image then it will directly injected to the tensor. However, if it is a time series then a conversion is required before injecting it to the tensor. We used the Gramian Angular Fields (GAF) [39, 27] to encode the time series as images. GAF encapsulates the correlation structures and uses the output to generate 2D images. It is particularly useful for representing temporal patterns in a visual format that can be processed by image-based algorithms. A time series signal is presented in GAF as a polar coordinated system, and the angles of every data point are transformed into matrices. Let K be a set of learnable kernels of size (K_h, K_w, C, O), where K_h and K_w are the kernel height and width, and O is the number of output channels. Each kernel K_j in K is convolved with the input data d_i to produce a feature map F_l, where l ranges from 1 to O. This operation is defined as:

$$F_l = K \bigodot d_i \bigoplus b_l \tag{4}$$

where $\bigodot$ denotes the convolution operation, $\bigoplus$ is a matrix addition, b_l is a learnable bias term for the l^{th} filter, and the output feature map f_l is of size (H', W', 1), where H' and W' are the height and width of the output feature map. The output feature map F_l is then passed through a non-linear activation function g, which introduces non-linearity into the model. We use the commonly known activation function, named ReLU, and which is defined as:

$$g(x) = max(0, x) \tag{5}$$

where x is the input to the activation function. After each convolutional layer, a pooling layer is often used to down-sample the output feature map. We used the most commonly pooling operation which is max-pooling. It aims to extract the maximum value within a pooling window of size (F'_h) from the input feature map. This operation is defined as:

$$F'_i = max_{j=1}^{F'_h} F_{i+(j-1)s} \tag{6}$$

where F'_i is the output of the pooling operation at i^{th} position, s is the stride (i.e., the distance between adjacent pooling windows), and $F_{i+(j-1)s}$ is the input feature map value at position $i+(j-1)s$. At the end of this step, we have the set of data features F', each element $F'_i \in F'$ represents the visual features of the image I_i.

4.4. Temporal Learning

To detect temporal dependencies in the series of spatial embeddings produced by the CNN network, we design a transformer-based network. Transformer is a type of neural network architecture that are able to effectively capture long-range dependencies in sequential data. The key innovation of the Transformer is the self-attention mechanism, which allows the model to attend to different parts of the input sequence when making predictions. A transformer used for machine translation receives a one-dimensional input and is followed by a succession of transformer layers described as lightweight feed-forward networks that project each element in the input independently. The existing transformers' self-attention mechanism computes the similarity score between all representations, linearly aggregates the feature representation, and changes the input representation accordingly. Even while this combination produces excellent results in machine translation, it results in a decrease in computing time and is regarded as a needless procedure in some computer vision applications, since nearby spatial representations frequently correspond to the same visual concept. Indeed, our methodology entails using the visual CNN features learned in the previous stage directly in the visual transformer. Overall, the function that the suggested transformer targets to optimize is defined as:

$$VT(x) = softmax\left(\frac{CNN(x)A^T}{\sqrt{c}}\right) \tag{7}$$

$CNN(x)$ (F') is the visual features of the input x obtained using the CNN model. c is a normalization constant, A^T is the attention layer that acts across the entire input. In the following, we present our adaptation of transformer for analyzing the sequence of features F' derived in the previous step:

1. **Patch Embeddings**: Each input features $F'_i \in F'$ is first divided into a set of non-overlapping patches, each of which is then flattened into a vector representation using a linear projection. These patch embeddings $\mathcal{P}$ are then fed into the Transformer encoder as input sequences.

2. **Encoder**: The Transformer encoder consists of a stack of several identical layers, each of which contains multi-head attention and feedforward neural networks. The multi-head attention mechanism allows the model to attend to different parts of the input patches $\mathcal{P}$ and capture long-range dependencies. In our transformer, we will not use the decoder since the output in *HAR* will be a fixed-length vector.

3. **Positional Encoding**: Since the transformer does not use recurrent connections, our model needs to incorporate positional information about the input patches

to capture the spatial structure of the features. This is done by adding positional encodings to the patch embeddings before they are fed into the transformer encoder. The positional encodings are learned and represent the spatial location of each patch in the image.

4. **Output Head**: At the end of the encoder, an output head will be added to the architecture to map the final hidden states to the final output of the model. We will use a fully connected layer integrated with softmax function.

5. Performance Evaluation

Intensive experiments have been carried out to evaluate the performance of CVTN solution.

5.1. Dataset and Metrics

We use the Kinetics human action video dataset, a pivotal asset in the realm of *HAR* research. Comprising a vast collection of more than 650,000 succinct video clips, each lasting about 10 seconds, this dataset encapsulates a comprehensive spectrum of 400 distinct action categories. These categories encompass a wide array of human engagements such as sports, culinary endeavors, and musical performances, sourced meticulously from YouTube and meticulously filtered to ensure exclusivity to human actions. We will use the latest version, Kinetics-700, launched in 2019, wherein the action taxonomy has been expanded to include 700 classes. Renowned as a benchmark standard, the Kinetics dataset plays a pivotal role in the evaluation of the efficacy of human activity recognition algorithms. To evaluate the performance of CVTN, several factors have been take into account, including:

1. **Accuracy**: This is a measure of how well the solution correctly identifies the human activity being performed. This measure is determined by model accuracy. It is often the most important metric for evaluating the performance of a HAR model. Higher accuracy means that the model is able to correctly identify the activity being performed more often, and lower accuracy indicates that the model is making more incorrect predictions.

2. **Recognition rate**: This refers to the ability of the solution to perform well on new, unseen data. It refers to the proportion of instances in a dataset that are correctly classified by a human activity recognition model. It is often used as a measure of the model's performance and is typically expressed as a percentage. For example, if a model correctly identifies 100 activities out of a total of 150 instances in the test set, the recognition rate would be 67%. The higher the recognition rate, the better the performance of the model.

Models	20%	50%	80%	100%
DST-LSTM	44	58	68	71
Hybridnet	57	63	74	75
CVTN	62	75	89	95

Table 1: Recognition rate performance of CTVN compared to baseline methods (DST-LSTM, and Hybridnet).

Category	VT	CNN	VT + CNN
Standing in the room	94	91	**95**
Sitting on the floor	92	89	**93**
Sitting with stretched legs	89	88	**90**
Sitting cross-legged	88	87	**89**
Lying in bed, leg raised	87	85	**93**
average	85	88	**92**

Table 2: Impact of each network model on CVTN in the top five frequent activities.

5.2. Baselines

We used these two recent baseline solutions for comparisons:

1. **DST-LSTM** [40]: It aims to categorize five common activity states (standing, sitting, walking on the floor, descending stairs, and ascending stairs) using collected data. The data's diverse information will be harnessed to develop an LSTM model. This LSTM model's strength in capturing sequential patterns will ensure accurate classification of activity states based on the provided data, enhancing the reliability of activity recognition.

2. **Hybridnet** [46]: It intelligently leveraged contextual information using CNN, and graph neural architecture enhancing its ability to make informed predictions and decisions based on the broader context. The robust combination of intricate structural modeling and contextual awareness in Hybridnet resulted in an accurate outcome, showcasing the model's advanced capabilities in handling complex data scenarios.

5.3. Accuracy Performance

Numerous experiments have been conducted to evaluate the precision performance of CTVN. We systematically manipulated the number of iterations (epochs) within VT and CNN across a range of 100 to 1,000, and additionally adjusted the loss rate within the range of 0.01 to 0.09. The outcomes of these experiments are visually depicted in Figure 3. Our observations reveal that the accuracy of the model exhibits improvement with an escalation in the number of epochs for both the CNN and VT. For instance, when employing 100 epochs for VT, 100 epochs for the CNN, and a loss rate of 0.01, the model's accuracy remains below 35%.

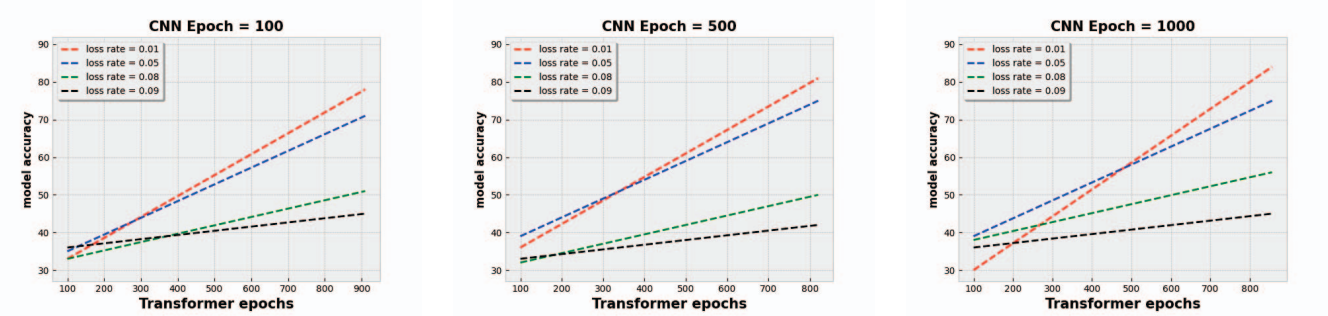

Figure 3: Accuracy performance of CVTN for different CNN epochs (from 100, 500 to 1,000), different transformer epochs (from 100 to 1,000), and for different loss rate {0.01, 0.05, 0.08, 0.09}.

In contrast, when VT is subjected to 1,000 epochs, the CNN also undergoes 1,000 epochs, combined with a loss rate of 0.01, the model's accuracy surpasses 84%. Additionally, our findings suggest that elevating the loss rate leads to a reduction in model accuracy. Specifically, the model's accuracy is high when the loss rate is set at 0.01 and 0.05, whereas it diminishes notably for loss rates of 0.08 and 0.09. Guided by these results, we have established the subsequent parameters for the remaining experiments: the VT's epoch count is set to 1,000, the CNN's epoch count is set to 1,000, and the loss rate is maintained at 0.01.

5.4. Recognition Rate Performance

Numerous tests have been conducted to assess the efficacy of recognition rate performance involving CVTN alongside the baseline solutions. Parameter such as the proportion of selected input features (ranging from 10% to 100%) was subjected to variation. The outcomes of these experiments are depicted in Table 1. The findings distinctly highlight the advantageous standing of CVTN when contrasted with the baseline techniques (DST-LSTM and Hybridnet), irrespective of the specific experimental conditions employed. These observations affirm the viability of the proposed methodologies in the realm of discerning human activities from unobserved instances. The achievement of these outcomes can be attributed to the effective fusion of both VT and CNN architectures, enabling the acquisition of spatiotemporal visual information from divergent data sensors exhibiting substantial interdependence. CVTN emerges as a viable option for data fusion strategies, wherein a sequence of images is synthesized and trained using information from multiple data sources.

5.5. Ablation Study

In this last experiment, we conduct an ablation study that involves systematically removing the combined models to analyze their impact on the overall CVTN outcome. From this standpoint, we perform the analysis of three distinct model configurations within the context of the CVTN (VT, CNN, and a hybrid CNN with VT). The results of our ablation study, shown in Table 2 serve to illustrate the performance of each configuration. Notably, the hybrid CNN with VT reaches the highest point of accomplishment across all defined configurations. It showcases a clear improvement in recognition accuracy, showcasing a 2% increase when compared to using only VT, and an even more remarkable rise of 4% when contrasted with using just a CNN.

6. Conclusion

This paper introduces CVTN (Convolution Visual Transformer Network) for *HAR* that empowers examination, and recognizing human activities. It is a new technique for robustifying the *HAR* process by learning from sensor data that is collectively represented in space and time. CNN begins by learning visual spatial features from sensor data. The visual features that have been learned are then injected into VT, which encapsulates temporal data by observing the sensor status at different timestamps. Kinetics was used to test CVTN. The results show that the CVTN is clearly superior to the most recent baseline *HAR* solutions. In order to enhance spatiotemporal visual learning, future research efforts should be directed towards the incorporation of various optimization techniques, such as hyperparameter optimization [14, 13, 29]. Expanding the applicability of the developed CVTN, and its adaptation on diverse scenarios, notably in the domain of home elderly care [23], where its potential to facilitate health monitoring holds promise. Furthermore, considering the incorporation of CVTN into the context of home-gait monitoring [1] presents a fascinating path for CVTN's investigation to improved safety and overall well-being.

Acknowledgement This work is co-funded by the Research Council of Norway under the project entitled "Next Generation 3D Machine Vision with Embedded Visual Computing" with grant number 325748.

References

[1] Hajar Abedi, Ahmad Ansariyan, Plinio P Morita, Alexander Wong, Jennifer Boger, and George Shaker. Ai-powered non-contact in-home gait monitoring and activity recognition system based on mm-wave fmcw radar and cloud computing. *IEEE Internet of Things Journal*, 2023. 8

[2] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. Star-transformer: A spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3330–3339, 2023. 4

[3] Hymalai Bello, Luis Alfredo Sanchez Marin, Sungho Suh, Bo Zhou, and Paul Lukowicz. Inmyface: Inertial and mechanomyography-based sensor fusion for wearable facial activity recognition. *Information Fusion*, page 101886, 2023. 1

[4] Antonio Bevilacqua, Kyle MacDonald, Aamina Rangarej, Venessa Widjaya, Brian Caulfield, and Tahar Kechadi. Human activity recognition with convolutional neural networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III 18*, pages 541–552. Springer, 2019. 4

[5] Yunus Celik, Samuel Stuart, Wai Lok Woo, Ervin Sejdic, and Alan Godfrey. Multi-modal gait: A wearable, algorithm and data fusion approach for clinical and free-living assessment. *Information Fusion*, 78:57–70, 2022. 1

[6] Sravan Kumar Challa, Akhilesh Kumar, and Vijay Bhaskar Semwal. A multibranch cnn-bilstm model for human activity recognition using wearable sensor data. *The Visual Computer*, 38(12):4095–4109, 2022. 4

[7] Jiawei Chen and Chiu Man Ho. Mm-vit: Multi-modal video transformer for compressed video action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1910–1921, 2022. 4

[8] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 2

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141, 2021. 2

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 2

[12] Bhat Dittakavi, Divyagna Bavikadi, Sai Vikas Desai, Soumi Chakraborty, Nishant Reddy, Vineeth N Balasubramanian, Bharathi Callepalli, and Ayon Sharma. Pose tutor: An explainable system for pose correction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3540–3549, 2022. 1

[13] Youcef Djenouri and Marco Comuzzi. Combining apriori heuristic and bio-inspired algorithms for solving the frequent itemsets mining problem. *Information Sciences*, 420:1–15, 2017. 8

[14] Youcef Djenouri, Gautam Srivastava, and Jerry Chun-Wei Lin. Fast and accurate convolution neural network for detecting manufacturing data. *IEEE Transactions on Industrial Informatics*, 17(4):2947–2955, 2020. 8

[15] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 2

[16] Fabian Herold, Paula Theobald, Thomas Gronwald, Navin Kaushal, Liye Zou, Eling D de Bruin, Louis Bherer, and Notger G Müller. Alexa, let's train now!—a systematic review and classification approach to digital and home-based physical training interventions aiming to support healthy cognitive aging. *Journal of Sport and Health Science*, 2023. 1

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[18] Wenbo Huang, Lei Zhang, Hao Wu, Fuhong Min, and Aiguo Song. Channel-equalization-har: a light-weight convolutional neural network for wearable sensor based human activity recognition. *IEEE Transactions on Mobile Computing*, 2022. 4

[19] Ali Jahan and Kevin L Edwards. A state-of-the-art survey on the influence of normalization techniques in ranking: Improving the materials selection process in engineering design. *Materials & Design (1980-2015)*, 65:335–342, 2015. 5

[20] Wenchao Jiang and Zhaozheng Yin. Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1307–1310, 2015. 4

[21] Sevvandi Kandanaarachchi, Hideya Ochiai, and Asha Rao. Honeyboost: Boosting honeypot performance with data fusion and anomaly detection. *Expert Systems with Applications*, 201:117073, 2022. 1

[22] Athul Krishnan, Subham Das, and Mitradip Bhattacharjee. Flexible piezoresistive pressure and temperature sensor module for continuous monitoring of cardiac health. *IEEE Journal on Flexible Electronics*, 2023. 1

[23] Pravin Kulurkar, Chandra kumar Dixit, VC Bharathi, A Monikavishnuvarthini, Amol Dhakne, and P Preethi. Ai based elderly fall prediction system using wearable sensors: A smart home-care technology with iot. *Measurement: Sensors*, 25:100614, 2023. 8

[24] Bing Li, Wei Cui, Wei Wang, Le Zhang, Zhenghua Chen, and Min Wu. Two-stream convolution augmented transformer for human activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 286–293, 2021. 2

[25] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2017. 1

[26] Zhonghai Ma, Yugang Geng, Songlin Nie, Hui Ji, Xiaopeng Yan, and Haitao Liao. Snif-dfa: A signal processing and information fusion method for smart gua sha device. *IEEE Sensors Journal*, 22(24):24176–24185, 2022. 1

[27] Edson F Luque Mamani and Cristian Lopez del Alamo. Gaf-cnn-lstm for multivariate time-series images forecasting. In *LatinX in AI Research at ICML 2019*, 2019. 6

[28] Hongying Meng, Nick Pears, and Chris Bailey. A human action recognition system for embedded computer vision application. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–6. IEEE, 2007. 1

[29] Tinhinane Mezair, Youcef Djenouri, Asma Belhadi, Gautam Srivastava, and Jerry Chun-Wei Lin. A sustainable deep learning framework for fault detection in 6g industry 4.0 heterogeneous data environments. *Computer Communications*, 187:164–171, 2022. 8

[30] RS Nancy Noella and J Priyadarshini. Diagnosis of alzheimer's, parkinson's disease and frontotemporal dementia using a generative adversarial deep convolutional neural network. *Neural Computing and Applications*, pages 1–10, 2022. 1

[31] M Panagiotou, A Zlatintsi, PP Filntisis, AJ Roumeliotis, N Efthymiou, and P Maragos. A comparative study of autoencoder architectures for mental health analysis using wearable sensors data. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1258–1262. IEEE, 2022. 1

[32] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2669–2676, 2020. 1

[33] Michael S Ryoo and Jake K Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *2009 IEEE 12th international conference on computer vision*, pages 1593–1600. IEEE, 2009. 1

[34] Hamidreza Sadreazami, Miodrag Bolic, and Sreeraman Rajan. Contactless fall detection using time-frequency analysis and convolutional neural networks. *IEEE Transactions on Industrial Informatics*, 17(10):6842–6851, 2021. 1

[35] Dinesh Kumar Sah, Korhan Cengiz, Yasser Alshehri, Noha Alnazzawi, Nikola Ivković, et al. Early alert for sleep deprivation using mobile sensor data fusion. *Computers and Electrical Engineering*, 102:108228, 2022. 1

[36] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 1

[37] Thanh-Dat Truong, Quoc-Huy Bui, Chi Nhan Duong, Han-Seok Seo, Son Lam Phung, Xin Li, and Khoa Luu. Direcformer: A directed attention in transformer approach to robust action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20030–20040, 2022. 4

[38] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 1

[39] Zhiguang Wang, Tim Oates, et al. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, volume 1. AAAI Menlo Park, CA, USA, 2015. 6

[40] Hao Wu, Zhichao Zhang, Xiaoyong Li, Kai Shang, Yongming Han, Zhiqiang Geng, and Tingrui Pan. A novel pedal musculoskeletal response based on differential spatio-temporal lstm for human activity recognition. *Knowledge-Based Systems*, 261:110187, 2023. 7

[41] Shuo Xiao, Shengzhi Wang, Zhenzhen Huang, Yu Wang, and Haifeng Jiang. Two-stream transformer network for sensor-based human activity recognition. *Neurocomputing*, 512:253–268, 2022. 2

[42] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18816–18826, 2023. 4

[43] Shige Xu, Lei Zhang, Wenbo Huang, Hao Wu, and Aiguo Song. Deformable convolutional networks for multimodal human activity recognition using wearable sensors. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022. 4

[44] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14063–14073, 2022. 4

[45] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Ijcai*, volume 15, pages 3995–4001. Buenos Aires, Argentina, 2015. 4

[46] Wenjie Yang, Jianlin Zhang, Jingju Cai, and Zhiyong Xu. Hybridnet: Integrating gcn and cnn for skeleton-based action recognition. *Applied Intelligence*, 53(1):574–585, 2023. 7

[47] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018. 1