Pedestrian Action Recognition in Self-driving System Using Video Vision Transformer with Variable Frame Lengths

Hyowon Lee
Student Number *230861592*
Supervisor Name: Professor Georgios Tzimiropoulos
MSc. Artificial Intelligence, QMUL

***Abstract*— This research suggests a method that utilizes a video vision transformer model for pedestrian action recognition in autonomous driving systems. By slicing video at a random number of frames, the model can analyse the action pattern of pedestrians more accurately. It leads to improvement of model accuracy. Due to using transfer learning with a pre-trained model, it is hard to modify the model directly. Thus, this research uses padding to match the number of frames for training well. In the experiments, the JAAD dataset and PIE dataset which have pedestrian video dataset and annotation of pedestrian's action are used. The dataset results are divided into training with a fixed number of frames (32 frames) and a random number of frames (between 16 frames and 32 frames). Based on comparing these results, training the model with various frames demonstrates higher accuracy of pedestrian action classification than training the model with a fixed number of frames.**

Keywords—Video Vision Transformer, autonomous driving system, transfer learning, pedestrian action recognition,

I. INTRODUCTION (*HEADING 1*)

Recognising human actions is one of the core subjects of computer vision research. It can apply to security, sports, crime, and so on. Traditional machine-learning techniques and hand-designed features were the mainstays of early research. However, since deep learning techniques were developed, they adapted to studying and developing human action recognition. (Morshed, M.G et al, 2023) For instance, the Qatar World Cup in 2022, is used for determining off-side by tracking the location of a ball and a player by using a sensor that is in the ball and 12 cameras. (Patel, B. N., 2023) Human action recognition aims to extract one action label from a continuous point of view. (Morshed, M.G et al, 2023) In this research, especially pedestrian action recognition is the focus. In other words, this research aims to analyse pedestrian actions such as walking, crossing, and standing.

The self-driving car market has been increasing and many international businesses are working feverishly to develop self-driving systems. This leads to the prediction that the market value of self-driving cars will reach $615 billion by 2026. (Padmaja, B et al, 2023) As the self-driving system market grows, many studies are being conducted on the development of self-driving assistance systems related to the safety of pedestrians. (Quintero, R et al, 2017) According to the World Health Organization's Global Status Report on Road Safety published in 2015, traffic fatalities all over the world have risen to an all-time high of 1.25 million fatalities per year. (Varytimidis, D et al, 2018) Furthermore, according to Chougule, A et al (2023), 50% people select 'feel unsafe' as the reason why cannot trust self-driving car. This means that people think the limitation of self-driving system is safety. This means that people worry about collision with pedestrian or other cars. Therefore, it is essential to improve the safety of system. To assist the self-driving car in avoiding collisions with pedestrians, the system of self-driving cars should recognise the actions of pedestrians accurately and provide the results to the system for determining the car's motion. (Coelingh, E et al, 2010)

To improve the system, it has to make a model in which video data can be trained because the model should classify one action through several frames of a pedestrian's action. Therefore, recently, research using video data has been actively conducted, which makes it possible to recognise pedestrian action more accurately by utilizing temporal information. (J. Selva, et al, 2023) Therefore, the existing approach of pedestrian action recognition research focuses on 3D CNN or the combination of Vision Transformer and RNN to train spatial-temporal information of video data. However, dding to Arnab, A, et al (2021), Although 3D CNN can be used to train video data, the model needs a lot of training data and computational power because there are more parameters for training time and data space. (Arnab, A, et al, 2021)

Thus, this paper analyses the performance of pedestrian action recognition using a pre-trained video vision transformer (ViViT) model which is based on transformer architecture. By using a pre-trained model for transfer learning, the model can adapt quickly and effectively to a new dataset for a new task. In addition, it can provide reliable results even though the training dataset is limited. (Kim, Y., 2020) JAAD and PIE datasets which are filmed for studying autonomous driving are used for training the model in this paper. However, even though a pre-trained model is used, for adapting to a new dataset, a transformer-based model is still needed enough amounts of dataset for training. Thus, by slicing frames at a random rate, this approach can expect to not only make enough datasets but also data regularisation to prevent overfitting.

II. RELATED WORK

A. Human Action Recognition

The four categories of human activity are gesture, action, interaction, and group activities. These categories are determined by the complexity of the action and the body parts involved. In more detail, a gesture is a body action that contains a particular message such as a thumbs up. Action is a series of physical actions carried out by a single person like running. Interaction is a series of actions that interact with another person or an item such as handshaking. Lastly, group activity is a mash-up of the other three categories. Human Action Recognition is focused on 'action' among human

activity categories. It can recognise and classify behaviour by collecting and processing human action information from video frames. Due to requirements in various industries, human action recognition has received interest from the academic community and real-world business community. (Jegham, I., et al, 2020)

B. 3D CNN for human action recognition

Before 3D CNN was proposed, 2D CNN was proposed for video classification. However, 2D CNN is designed for training image data and appropriately for processing spatial information. (Sun, Z, et al, 2022) Therefore, 3D CNN is proposed for processing not only spatial information but also temporal information as shown in Fig 1. (Boualia, S. N., et al, 2021)

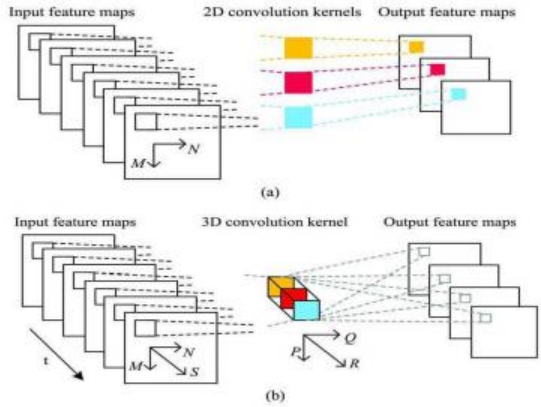

Figure 1. The processes of 2D convolution (a) and 3D convolution (b) (Boualia, S. N., et al, 2021)

According to the result of Boualia, S. N., et al (2021), 3D CNN demonstrates high performance (78% in the KTH dataset, 90% in the J-HMDB dataset) with minimum pre-processing steps. However, it has a high complexity of the model and a high cost of calculation. Because it uses a 3-dimensional kernel for processing temporal information of video data. (Boualia, S. N., et al, 2021)

As illustrated in Fig 2, an architecture that they proposed is consisted of two Conv3D layers with a ReLU activation function. Moreover, the MaxPool layer is used for downsampling, and dropout is set to prevent overfitting. The softmax layer is used for the activation function. This structure

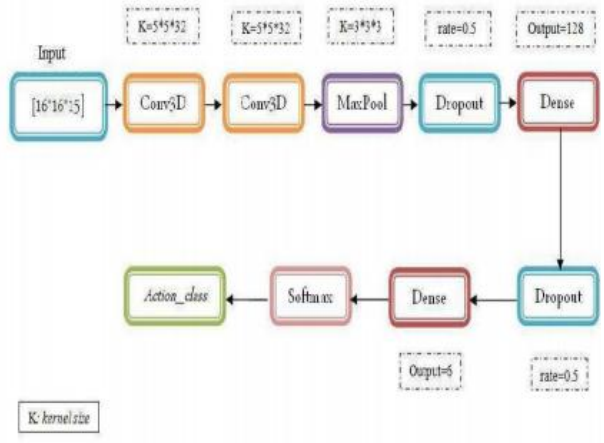

Figure 2. 3D CNN proposed model architecture (Boualia, S. N., et al, 2021)

allows the 3D CNN to learn from the convolution filter itself the temporal learning of continuous frames. That is, with this structure, specific spatio-temporal learning is possible for the short term (a small number of consecutive frames). (Boualia, S. N., et al, 2021) However a limitation of the 3D CNN approach is that, because of the 3D filter, there are a lot of parameters, making it challenging to build a deep structure. (Sun, Z, et al, 2022)

C. Vision Transformer

In the field of human activity recognition, models that use CNN have been mainly used. However, these models have disadvantages of high complexity and difficult real-time processing in resource-constrained environments. (Sun, Z, et al, 2022)

Nonetheless, there has been a lot of interest in the field of computer vision for Transformer-based models since its introduction in the natural language processing domain in 2017. (Dosovitskiy, A., et al, 2020; Zhou, D., et al, 2021) Because it has shown outstanding performance, even processing in image data. Thus, Vision Transformer (ViT) which trains image data based on a transformer model was proposed in 2020. As illustrated in Fig. 3, a Vision Transformer (ViT) specifically splits the image into small patches and processes each patch by turning it into an embedding vector. (Dosovitskiy, A., et al, 2020)

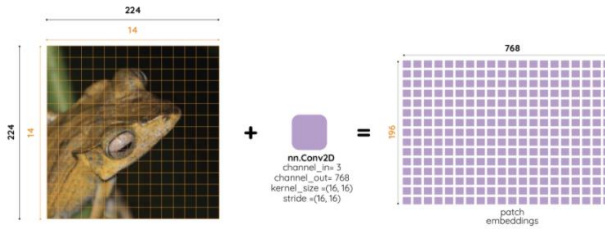

Figure 3. Patch embedding (Hongl, 2021)

Because the Transformer accepts 1-dimensional sequences as an input value, image patches should be flattened on 1-dimensional sequences. This means that an image that has H(height) × W(width) × C(channel) shape transforms the shape to N(number of sequences) × P(size of the patch) × P × C. Using linear projection, the transformed image changes d-dimensional vector to make it appropriate Transformer input shape. Each patch has positional embedding applied to it after linear projection in order to provide positional information. Patches with positional embedding are put into the transformer encoder as an input value. Equation (1) represents the input of Transformer encoder. (Dosovitskiy, A., et al, 2020)

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \cdots ; x_p^N E] + E_{pos} \quad (1)$$

(Dosovitskiy, A., et al, 2020)

In equation (1), x_{class} represents classification token (CLS), $x_p^N E$ means each image sequence divided by patch, and E_{pos} means positional embedding.

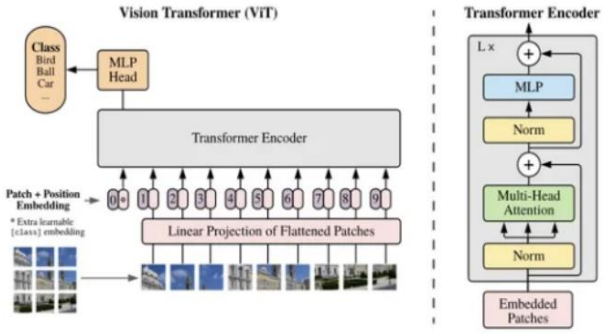

Figure 4. Vision Transformer architecture

The part of the Transformer encoder, as shown in Fig 4, consists of MSA (Multi-head Self Attention) and MLP (Multi-Layer Perceptron). This stacks with L number of layers. Furthermore, by performing layer normalization, which is one of RNN regularisation, before attention and MLP, the model is capable of achieving effective training outcomes.

$$z'_\ell = MSA(LN(z_{\ell-1})) + z_{\ell-1} \qquad \ell = 1 \cdots L \qquad (2)$$

(Dosovitskiy, A., et al, 2020)

$$z_\ell = MSA(LN(z'_\ell)) + z'_\ell \qquad \ell = 1 \cdots L \qquad (3)$$

(Dosovitskiy, A., et al, 2020)

Equation (2) and equation (3) are equations of Transformer. Equation (2) means that applying MSA after LN (Layer Normalisation) to the previous input. Moreover, by doing a skip connection the output of MSA, the output of one layer of the neural network is transmitted to a deeper layer to solve the problem of performance degradation that can occur as the neural network deepens. Equation (3) represents the MLP head.

$$y = LN(z_L^0) \qquad (4)$$

(Dosovitskiy, A., et al, 2020)

Equation (4) represents the output of model. In other words, it means the class of classification result. (Dosovitskiy, A., et al, 2020)

After introducing Vision Transformer, transformer-based models have received significant attention in computer vision area. Because These approaches are more flexible and scalable than traditional CNN-based models. (Zhou, D., et al, 2021)

D. Vision Transformer for human action recognition

Since the vision transformer model was proposed, it indicates better performance in image data than CNN or another model. (Dosovitskiy, A., et al, 2020) Thus, the study of the human action recognition approach to the vision transformer model. Especially, there are some studies for training video data with the combination model vision transformer architecture and RNN. (Zhou, D., et al, 2021)

In the Wensel, J., et al, (2023) paper, they suggest the combination Vision Transformer model with the Recurrent Transformer for classifying human action recognition as illustrated in Fig 5. Firstly, they pre-process the input video to make the subsampled frames resize as 224 x 224. Then Vision Transformer accepts pre-processed frames as input and the output of the Vision Transformer is delivered to the Recurrent Transformer to train the pattern of action along the time and predict the action. To process temporal and spatial information of input data effectively, the model employed multi-head attention and positional encoding. (Wensel, J., et al, 2023)

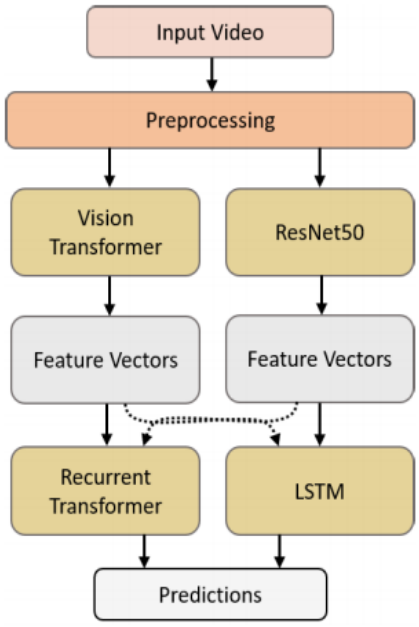

Figure 5. ViT-ReT proposed architecture (Wensel, J., et al, 2023)

They use the HMDB51 dataset for training, which consists of 6,766 video clips across 51 action categories sourced from movies and web videos. According to Wensel, J. et al(2023), the proposed model achieves 78.4% accuracy which is around 5% higher than the combination of Vision Transformer and LSTM(73.7%). Furthermore, they train another dataset and get higher accuracy than another model. A Vision Transformer model proves to be a strong alternative to traditional CNN and RNN-based human action recognition models, particularly when it comes to processing speed and scalability. (Wensel, J., et al, 2023)

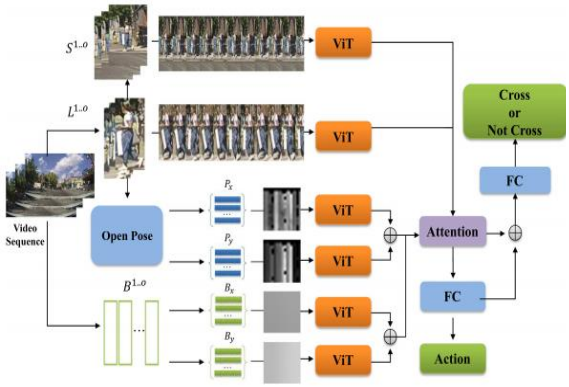

Figure 6. Action-ViT proposed architecture (Zhao, S., et al, 2021)

In Zhao, S., et al (2021) paper, the JAAD dataset and PIE dataset are used for training the model for classifying whether a pedestrian crosses the road or not. The two components of the suggested model architecture are feature extraction and classification, as seen in Fig 6. They use the Vision Transformer model as a feature extractor, and a fully connected layer is used for the classification of the actions of pedestrians. In more detail, insert each frame data and annotations such as bounding box as Vision Transformer input

data to extract feature. These features are time-connected and merged into the final feature vector. Before inputting a feature vector into a fully connected network, to assign a weight differently to each feature of the final feature vector based on its significance, an attention mechanism is utilized. (Zhao, S., et al, 2022)

They train with 15 frames to classify the intention of crossing. In addition, adjust the pedestrian action annotation provided by the dataset to a two-dimensional one-hot vector, and employ it before the output layer to provide the final intention on prediction in depth. According to ZHAO et al (2021), the model based on Vision Transformer demonstrates more accurate results when processing the image by integrating local information and surrounding information of the image. In particular, the Vision Transformer base model improves the accuracy of local images by 2% and 7%. However, for the PIE dataset, all metrics are decreased. This is presumed to be because the surrounding image has more noise, and the Vision Transformer works better when processing images with a single piece of content. This indicates that the feature fusion by integrating two models is effective. Furthermore, by utilizing action annotations, the overall performance of pedestrian intention prediction is improved by providing additional context for pedestrian action. (Zhao, S., et al, 2022)

III. METHODOLOGY

In this paper, training two datasets which are the JAAD dataset and the PIE dataset. These datasets are used for studying pedestrian action in traffic situations. (Rasouli, A., et al, 2017; Rasouli, A., et al, 2018; Rasouli, A., et al, 2019) However, because of cropping pedestrians, the dataset is not enough to train the Video Vision Transformer (ViViT) model. Therefore, by cutting frames at random ranges, data augmentation and data regularisation are expected. Four experiments are conducted with different datasets (JAAD dataset and PIE dataset) and different settings of frame amount (fixed number of frames and variable number of frames).

A. Dataset

Two open datasets the JAAD dataset and the PIE dataset are used. These datasets consist of videos and annotations. In more detail, the JAAD dataset, which is shown in Fig 7, contains 346 videos with from 150 to 300 frames and annotation files, which include a bounding box and action label for each pedestrian. (Rasouli, A., et al, 2017; Rasouli, A., et al, 2018) The PIE dataset, which is illustrated in Fig 8, contains 1,842 pedestrians with pedestrian action annotations. (Rasouli, A., et al, 2019)

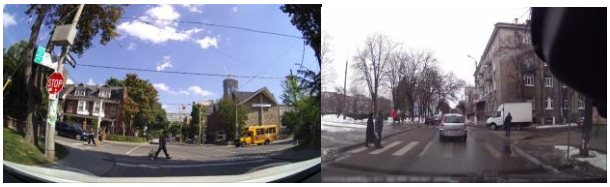

Figure 7 PIE dataset (Left), JAAD dataset (Right)

B. Pre-processing

- Pre-processing

  For training the Video Vision Transformer (ViViT) model, cropping each pedestrian using a bounding box which is defined in XML annotation files. Thus, the model can train by focusing on pedestrians. Furthermore, as shown in Fig 8, when cropping each pedestrian, by enlarging the bounding box at 10%, the model can train with more background information. In addition, resize the frames to fit the Video Vision Transformer model as 224 × 224 × 3 (width × height × channel). (Pantophlet, D., 2023) Due to cropping pedestrians and resizing to make frames appropriate input of the Video Vision Transformer model, the quality of frames is reduced. To enhance frame quality by reducing noise, sharpening techniques are employed using the OpenCV library developed by Bradski and Kaehler (2000). Additionally, when cropping pedestrians, if the number of pedestrian frames is not at least 16 × 4(minimum length of frame × frame rate), pre-processing is not performed in order to match the minimum number of frames to prevent training too short frame. All these pre-processing is demonstrated in detail in Algorithm I.

Figure 8 Cropped frame

Algorithm I pre-processing the frames (cropping and resize)

```
%Parse XML data and extract frames from videos%
Procedure PARSE_AND_PROCESS(video_path, xml_path)
        video_data ←READ_VIDEO(video_path)
        xml_data ← PARSE_XML(xml_path)
         bounding_boxes, occlusions, actions ←
PARSE_ANNOTATIONS(xml_data)
        needed_frames ← FRAMES(bounding_boxes)
        processed_data ← EXTRACT_FRAMES(video data,
needed_frames, occlusions, actions)
        return processed_data

%Crop bounding box%
Procedure CROP_BBOX(frame, bbox, expand_scale)
        xtl, ytl, xbr, ybr ← bbox
         width ← xbr – xtl
         height ← ybr – ytl
         new_xtl ← max(0, int(xtl - width * expand_scale))
         new_ytl ← max(0, int(ytl - height * expand_scale))
         new_xbr ← min(frame.shape[1], int(xbr + width *
         expand_scale))
         new_ybr ← min(frame.shape[0], int(ybr + height *
         expand_scale))
         cropped_bbox ← frame[new_ytl : new_ybr,
         new_xtl : new_xbr]
         improved_frame ←
         IMPROVE_QUALITY(cropped_bbox)
         cropped_bbox_rgb ←
         CONVERT_RGB(improved_frame)
         cropped_bbox_rgb ← RESIZE(cropped_bbox_rgb,
         to dimensions (224, 224))
return cropped_bbox_rgb
```

C. Data augmentation & Data regularisation

- Slicing frames

  For data augmentation and improving data regularisation, cut the frames at a random number of frames between 16 and 32. However, the Video Vision Transformer (ViViT) pre-trained model by Hugging Face should have a fixed number of frames. (Huggingface, 2022a) Hence, padding is used for processing variable frame numbers. When attaching padding to the video, the model should classify the padding frame as an empty frame. Thus, set the padding frame as 0 to make the model ignore the frame. Due to using a pre-trained model, the model architecture cannot be modified for using an attention mask to ignore the padding frame. Accordingly, this research uses adjusted positional encoding dynamically which is provided the model as an option to process the padding indirectly. In more detail, the Video Vision Transformer (ViViT) model by Hugging Face provides an 'interpolate_pos_encoding' parameter to set positional encoding dynamically. (Huggingface, 2022a) By setting this parameter as True, the model can adapt various input sizes and focus on relevant data. However, the padding frame does not have any relevant data because it is an empty(black) frame. Thus, the model can ignore the padding frame indirectly.

Algorithm 2 Slicing frames in various number by applying padding

```
% Multiply dataset size by 3 for data augmentation %
Procedure LEN()
   Return 3 * length(video_files)

% Process frames for training with the ViViT model %
Procedure GET_ITEM(index, video_files, xml_files, transform)
   % Adjust index for augmented data access %
   adjusted_index ← index // 3
   video_path ← video_files[adjusted_index]
   xml_path ← xml_files[adjusted_index]
   container ← OPEN_VIDEO(video_path)
   total_frames ← GET_FRAME_COUNT(container)

   % Determine the number of frames to sample %
   selected_clip_length ←
RANDOM_CLIP_LENGTH(Minimum_frame,
Maximum_frame)
   frame_indices ← SAMPLE_FRAMES(total_frames,
selected_clip_length, frame_rate)

   % Extract and optionally transform frames %
   frames ← EXTRACT_FRAMES(container, frame_indices)
   if transform is not None then
      frames ← APPLY_TRANSFORM(frames, transform)

   % Convert frames to tensor and apply padding if necessary %
   frame_tensor ← CONVERT_TO_TENSOR(frames)
   if GET_FRAME_COUNT(frame_tensor) <
max_frame_count
   then
      frame_tensor ← APPLY_PADDING(frame_tensor,
max_frame_count - GET_FRAME_COUNT(frame_tensor))

   % Generate labels for each frame %
   labels ← GENERATE_LABELS(xml_path, frame_tensor)
   label_tensor ← CONVERT_TO_TENSOR(labels)
   return frame_tensor, label_tensor
```

D. Video Vision Transformer

Video Vision Transformer (ViViT) is a model that has significant performance in video classification by applying it to a video dataset by adding a temporal attention encoder based on Vision Transformer architecture. Each frame of the video is divided into n_w(width) × n_h(height) patches, and each patch is contextually made in the Transformer encoder. This means that the total number of patches is Equation (5). (Arnab, A, et al, 2021)

$$n_{time} \times n_{height} \times n_{width} \quad (5)$$

(Arnab, A, et al, 2021)

Thus, the amount of attention calculated is large because the calculation becomes Equation (6)

$$n_{time}^2 \times n_{height}^2 \times n_{width}^2 \quad (6)$$

(Arnab, A, et al, 2021)

Due to large calculations, instead of calculating all frames, the video clip is used for embedding. To do mapping sequence tokens and videos while deriving from the concept of the ViT model, Tubelet Embedding is used. This method is a method of extracting 'Tube' that is spatio-temporal from the input volume as shown in Fig 9. After extracting, by linear projection in d-dimension, and putting it into a sequence token. This technique expands ViT embedding as three-dimensional (t × h × w), and token extracts from dimensions of each temporal, height, and width. On the contrary to uniform frame sampling, spatio-temporal information is combined during tokenisation. Therefore, using Tubelet embedding uses more temporal information than uniform frame sampling. (Arnab, A, et al, 2021)

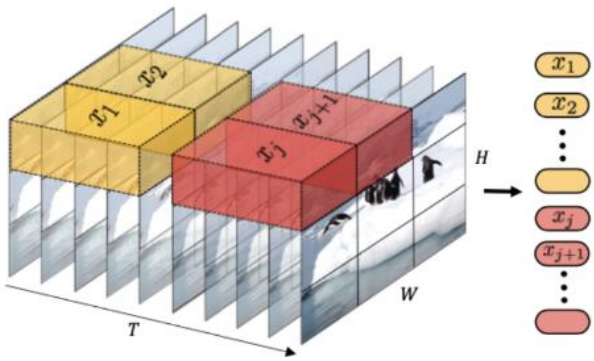

Figure 9 Tubelet Embedding (Arnab, A, et al, 2021)

As illustrated in Fig 10, the model which uses spatio-temporal attention is used in this research. In more detail, after tubelet embedding, positional embedding is attached to the token. Compared to Vision Transformer, video data needs one more axis which is a temporal axis. Accordingly, as indicated in Equation (5) positional embedding is needed additionally. The output of these processes is inserted into the transformer encoder as input. This model is that each Transformer layer models the pair interaction between all the spatio-temporal tokens unlike the CNN network, where the receptive field

increases linearly with the number of layers. (Arnab, A, et al, 2021)

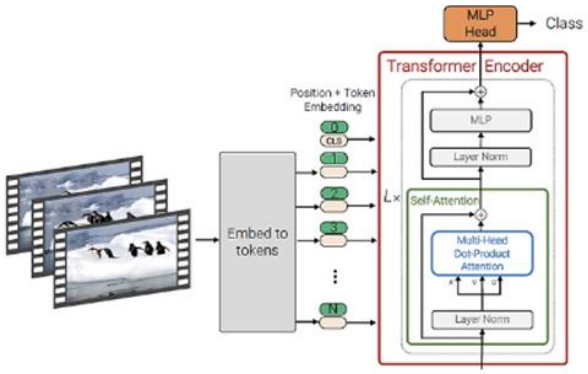

Figure 10 Video Vision Transformer architecture using Spatio-temporal attention method

E. Training

These experiments train using a pre-trained Video Vision Transformer model with this architecture provided by Hugging Face. (Huggingface, 2022a)

To evaluate how much data augmentation and training with a random variable number of frames improve, this research trains the model with three different settings. The dataset is divided into 70% of the training dataset and 30% of the test dataset. Additionally, AdamW is used for optimizer with learning rate 1e-5, and accuracy_score by sklearn library is used for calculating accuracy.

In addition, the ViViT pre-trained model by Hugging Face provides the configuration for transfer learning. (Huggingface, 2022a) By editing the configuration of a model such as tubelet size, video size, and number of frames, this research can use a pre-trained model for training two datasets.

- Experiment 1

  Train the model with a fixed number of 16 frames without applying data augmentation. To compare how data augmentation and slicing frames randomly impact the model's accuracy, Experiment 1 is set without modifying the dataset. However, for transfer learning, some settings of model parameters are adjusted.

  Length of training dataset : 285 clips (JAAD), 1092(PIE)

  Length of test dataset : 123 clips (JAAD), 469 clips (PIE)

  Model tubelet size : (2, 16, 16)

  Model video size : (16, 224, 224)

  Model num frames : 16

- Experiment 2

  Train the model with a fixed number of frames which is the same as Experiment 1, but with applying data augmentation(four times of dataset). In this experiment, we can expect how data augmentation affects to the model performance.

  Length of training dataset : 1142 clips (JAAD), 4370 clips (PIE)

  Length of test dataset : 490 clips (JAAD), 1874 clips (PIE)

  Model tubelet size : (2, 16, 16)

  Model video size : (16, 224, 224)

  Model num frames : 16

- Experiment 3

  Train the model with a variable number of frames, from 16 frames to 32 frames. In addition, apply the data augmentation(four times of dataset) to the model to select different numbers of frames in one video.

  Through this experiment, we can identify how the various number of frames impact the model performance.

  Length of training dataset : 1142 clips (JAAD), 4370 clips (PIE)

  Length of test dataset : 490 clips (JAAD), 1874 clips (PIE)

  Model tubelet size : (2, 32, 32)

  Model video size : (32, 224, 224)

  Model num frames : 32

IV. RESULT

A. JAAD dataset

- Experiment 1. Fixed frame (16 frames) without data augmentation

  The top 1 accuracy of training data and test data is 49.47% and 48.07% as illustrated in Table 1. This is the lowest accuracy of training data and test data. In addition, as shown in Fig 11, the shape of the train dataset and test dataset loss graph fluctuates. This means that it needs to make more data which means data augmentation.

- Experiment 2. Fixed frame (16 frames) with data augmentation

  With data augmentation, the accuracy of training data and test data increased by around 15%p and 14%p over the results of Experiment 1 as indicated in Table 1. Additionally, as illustrated in Fig 11, training data loss and test data loss have a decreasing trend.

- Experiment 3. Variable frame rates (16 - 32 frames) with data augmentation

  Experiment 3 trains with variable frame rates randomly between 16 frames and 32 frames. The result of this experiment indicates slightly higher performance compared to the previous experiment. For the training dataset, it increases by 1.4%p, and for the test dataset, it increases by 1.4%p compared to Experiment 2. Furthermore, the loss trend of training data and test data is decreasing as demonstrated in Fig 11.

TABLE I. TABLE TYPE STYLES

Top-1 Accuracy				
Epoch 30	***JAAD dataset***		***PIE dataset***	
	Training data	***Test data***	***Training data***	***Test data***
Fixed frame (16 frames), No data augmentation	49.47%	48.07%	68.59%	67.07%
Fixed frame (16 frames), Data augmentation	65.15%	62.24%	78.51%	78.71%
Variable frame rates (16 – 32 frames), Data augmentation	66.54%	63.67%	82.47%	81.16%

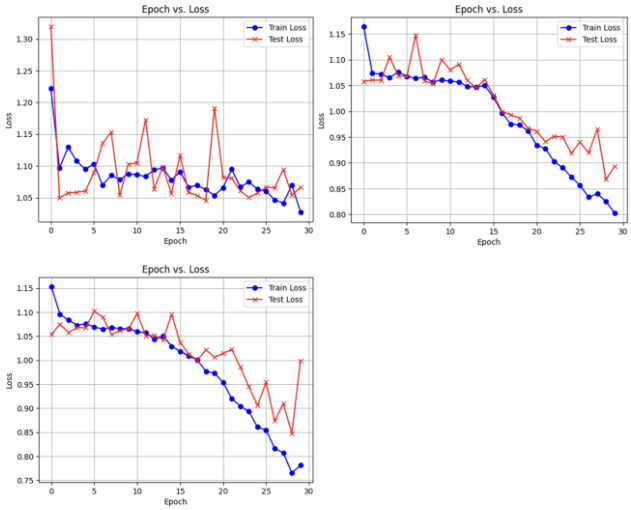

Figure 11 The loss of train dataset and test dataset (JAAD dataset) First – Experiment 1, Second – Experiment 2, Third – Experiment 3

B. *PIE dataset*

- Experiment 1. Fixed frame (16 frames) without data augmentation

  The top 1 accuracy of training data and test data is 68.59% and 67.07% as illustrated in Table 1. This is the lowest accuracy of training data and test data among all experiments. In addition, as shown in Fig 12, the shape of the train dataset and test dataset loss graph fluctuates even though it shows a decreasing trend. Especially, the loss of the test dataset significantly fluctuates.

- Experiment 2. Fixed frame (16 frames) with data augmentation

  With data augmentation, the accuracy of training data and test data increases around 10%p and 11%p over the results of Experiment 1 as indicated in Table 1. In addition, as illustrated in Fig 12, training data loss and test data loss have a decreasing trend. Moreover, it is decreased gradually compared to Experiment 1.

- Experiment 3. Variable frame rates (16 - 32 frames) with data augmentation

  Experiment 3 trains with variable frame rates randomly between 16 frames and 32 frames. The result of this experiment shows the highest performance among previous experiments. For the training dataset, it increases by 3.9%p, and for the test dataset, it increases by 2.5%p compared to Experiment 2. Furthermore, the loss trend of training data and test data is decreasing as demonstrated in Fig 12.

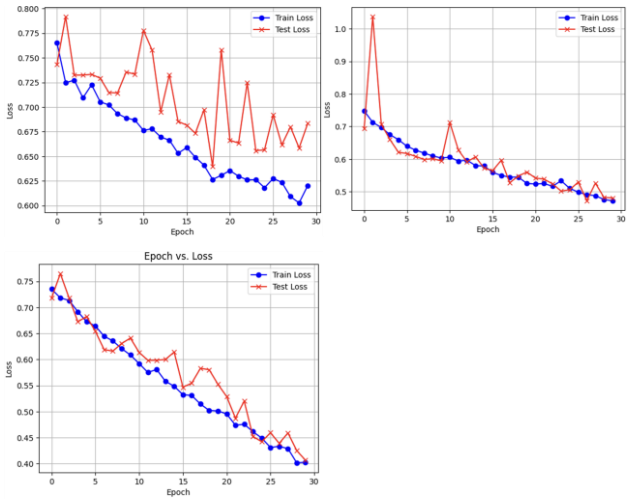

Figure 12 The loss of train dataset and test dataset (PIE dataset) First – Experiment 1, Second – Experiment 2, Third – Experiment 3

V. EVALUATION

A. *JAAD dataset*

- Accuracy

  Compared to Experiment 1, Experiment 2 and Experiment 3 indicate higher accuracy. This demonstrates that data augmentation is related to accuracy. Furthermore, Experiment 3 indicates slightly higher accuracy than Experiment 2 even though it is not a big difference. However, it is assumed that if this model trains more and more, the result of Experiment 3 will be much higher accuracy than Experiment 2.

- Loss

  The loss graph of the train dataset and test dataset of Experiment 1 fluctuates This means that the model cannot train well due to the lack of dataset. Therefore, compared to Experiment 1, Experiment 2 and Experiment 3 indicate that the pattern of graphs in which the loss gradually decreases. This indicates Experiment 2 and Experiment 3 train well rather than

Experiment 1. In other words, data augmentation influences making the model train well.

Additionally, the speed of loss of Experiment 3 shows a bit slower decrease than Experiment 2. This is probably because Experiment 3 has to process more frames than other experiments. However, as illustrated in Fig 12, the loss of Experiment 3 is the lowest among all experiments. This indicates that Experiment 3 has better performance than others. Furthermore, it trains well.

B. *PIE dataset*

- Accuracy

  Compared to Experiment 1, Experiment 2 and Experiment 3 demonstrate significantly higher accuracy. This means that data augmentation affects the accuracy of the model. Moreover, the accuracy of Experiment 3 shows higher accuracy than Experiment 2. This indicates that training the model with various numbers of frame influences the performance of the model.

- Loss

  Compared to Experiment 1, Experiment 2 and Experiment 3 show a more gradual decrease. This means that these models of the two experiments train stably. This is because Experiment 1 has the smallest amount of dataset. Additionally, Experiment 3 has the lowest loss among all experiments due to training various numbers of frames. It is evident that training not only with data augmentation but also with a variety of frame amounts influences the performance of the model.

C. *JAAD dataset & PIE dataset*

The PIE dataset demonstrates significantly higher performance than the JAAD dataset(around 15% difference). This can be estimated for two reasons. Firstly, the PIE dataset has much more video clips. Video Vision Transformer is sensitive to the amount of dataset even though a pre-trained model is used. Secondly, the PIE dataset has better-quality of video clips even after cropping the pedestrian. Thus, even though making the JAAD dataset higher quality using the OpenCV library, it remains the lower quality than the PIE dataset. Therefore, the PIE dataset demonstrates better performance than the JAAD dataset.

VI. DISCUSSION

This research suggests an approach to perform pedestrian action recognition in autonomous driving systems by training Video Vision Transformer pre-trained model from Hugging face with various random numbers of frames. (Huggingface, 2022a) This approach indicates that the model performance is increased.

- Compared to original paper model

  Even though using a different dataset(the original paper used the Kinetic 400 dataset), the result of this research shows slightly better performance than the original paper(80.0% in original paper as shown in Fig 13) even though this research trains only 30 epochs. (Arnab, A, et al, 2021)

	K400	EK	FLOPs ($\times 10^9$)	Params ($\times 10^6$)	Runtime (ms)
Model 1: Spatio-temporal	80.0	43.1	455.2	88.9	58.9

Figure 13. Original paper of Video Vision Transformer result (Top 1 accuracy) (Arnab, A, et al, 2021)

- Limitation

  Even though it indicates better performance, it still has limitations. The label can be changed in one video. In other words, the action of a pedestrian can be changed in one video such as from standing to walking. This leads to making the performance of the model lower. Therefore, in future work, processing labels should be considered.

VII. CONCLUSION

In conclusion, this approach shows better performance in both datasets. Especially, as shown in Fig 12 and Fig 13, the loss shows the lowest loss among other experiments. This indicates that this approach to training the video dataset has been effective in training the model.
For the JAAD dataset, this approach shows only slightly better performance than the result of the dataset with data augmentation. However, for the PIE dataset, this approach shows quite a better performance than the result of the dataset with data augmentation. As demonstrated in the Evaluation, because PIE dataset has more video clips and better-quality of video. Therefore, in future work, not only the approach to train the model but also how to deal with the existing dataset should be considered to make the dataset appropriate to train the model. Furthermore, as discussed in the Discussion section, future work should consider the label of pedestrian of each frame. Additionally, this research is only focused on pedestrian action. However, in the future, it is expected that this method can be used in other video recognition models by expanding the scope.

REFERENCES

Morshed, M.G., Sultana, T., Alam, A., & Lee, Y.-K., 2023. Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities. Sensors, 23(4), p.2182. Available at: https://doi.org/10.3390/s23042182.

Patel, B. N. TECHNOLOGY AND SPORTS: AN ANALYSIS OF QATAR'S USE OF ADVANCED TECHNOLOGIES TO IMPROVE THE WORLD CUP EXPERIENCE.

Padmaja, B., Moorthy, C. V., Venkateswarulu, N., & Bala, M. M. (2023). Exploration of issues, challenges and latest developments in autonomous cars. Journal of Big Data, 10(1), 61.

Quintero, R., Parra, I., Lorenzo, J., Fernández-Llorca, D., & Sotelo, M. A. (2017, October). Pedestrian intention recognition by means of a hidden markov model and body language. In 2017 IEEE 20th international conference on intelligent transportation systems (ITSC) (pp. 1-7). IEEE.

Varytimidis, D., Alonso-Fernandez, F., Duran, B., & Englund, C. (2018, November). Action and intention recognition of pedestrians in urban traffic. In 2018 14th International conference on signal-image technology & internet-based systems (SITIS) (pp. 676-682). IEEE.

Chougule, A., Chamola, V., Sam, A., Yu, F. R., & Sikdar, B. (2023). A comprehensive review on limitations of autonomous driving and its impact on accidents and collisions. *IEEE Open Journal of Vehicular Technology*.

Coelingh, E., Eidehall, A., & Bengtsson, M. (2010, September). Collision warning with full auto brake and pedestrian detection-a practical example of automatic emergency braking. In 13th International IEEE Conference on Intelligent Transportation Systems (pp. 155-160). IEEE.

J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund and A. Clapés, "Video Transformers: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 11, pp. 12922-12943, 1 Nov. 2023, doi: 10.1109/TPAMI.2023.3243465.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6836-6846).

Kim, Y. (2020). Pedestrian behavior detection using YOLOv4 and optical flow (Doctoral dissertation, Sungkyunkwan University).

Jegham, I., Khalifa, A. B., Alouani, I., & Mahjoub, M. A. (2020). Vision-based human action recognition: An overview and real world challenges. Forensic Science International: Digital Investigation, 32, 200901.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human action recognition from various data modalities: A review. IEEE transactions on pattern analysis and machine intelligence, 45(3), 3200-3225.

Boualia, S. N., & Amara, N. E. B. (2021, March). 3D CNN for human action recognition. In 2021 18th International Multi-Conference on Systems, Signals & Devices (SSD) (pp. 276-282). IEEE.

Hongl (2021) Vision Transformer (1). Available at: https://hongl.tistory.com/232 (Accessed: 24 June 2024).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q. and Feng, J., 2021. Deepvit: Towards deeper vision transformer. arXiv preprint arXiv:2103.11886.

Wensel, J., Ullah, H. and Munir, A., 2023. Vit-ret: Vision and recurrent transformer neural networks for human activity recognition in videos. IEEE Access.

Zhao, S., Li, H., Ke, Q., Liu, L. and Zhang, R., 2021. Action-vit: Pedestrian intent prediction in traffic scenes. IEEE Signal Processing Letters, 29, pp.324-328.

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017). Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 206-213).

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2018). It's not all about size: On the role of data properties in pedestrian detection. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops (pp. 0-0).

Rasouli, A., Kotseruba, I., Kunic, T., & Tsotsos, J. K. (2019). Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In

Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6262-6271).

Pantophlet, D. (2023). Vision Transformers for Pain Recognition on Thermal Image Frames (Master's thesis).

Bradski, G. and Kaehler, A., 2000. OpenCV. Dr. Dobb's journal of software tools, 3(2).

Huggingface (2022a). Video vision transformer (vivit). https://huggingface.co/docs/transformers/main/en/model_doc/vivit. Accessed on Jun 24, 2024.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6836-6846).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.