IEEE Vehicular Technology Magazine. Available online 26 January 2022
DOI: 10.1109/MVT.2021.3140047

Towards Human-Centered Automated Driving: A Novel Spatial-Temporal Vision Transformer-Enabled Head Tracker

Zhongxu Hu, Nanyang Technological University, Singapore
Yiran Zhang, Nanyang Technological University, Singapore
Yang Xing, Cranfield University, UK
Yifan Zhao, Cranfield University, UK
Dongpu Cao, University of Waterloo, Canada
Chen Lv, Nanyang Technological University, Singapore

Accurate and dynamic driver head pose tracking is of great importance for driver-automation collaboration, intelligent co-pilot, head-up display (HUD), and other human-centered automatedt driving applications. To further advance this technology, this paper proposes a low-cost and mark-less head tracking system using a deep learning-based dynamic head pose estimation model. The proposed system only requires an RGB camera without other hardware or markers. To enhance the accuracy of the driver's head pose estimation, a spatial-temporal vision transformer (ST-ViT) model, that takes an image pair as the input instead of a single frame, is proposed. Compared to the standard transformer, this contains a spatial convolutional vision transformer and a temporal transformer, which can improve the model performance. To handle the error fluctuation of the head pose estimation model, this paper proposes an adaptive Kalman filter (AKF). By analyzing the error distribution of the estimation model and user experience of the head tracker, the proposed AKF includes the adaptive observation noise coefficient; this can adaptively moderate the smoothness of the curve. Comprehensive experiments show that the proposed system is feasible and effectiveness, and it achieves a state-of-the-art performance.

Background

Intelligent driving is currently a hot research topic that requires a combination of multiple disciplines and algorithms. Developing and testing algorithms for real intelligent vehicles is an expensive and time-consuming process. The development of simulation technology provides an alternative way as it can offer physically and visually realistic simulations for several research goals and can also collect a large number of annotated samples to leverage deep learning and machine learning [1]. The driving simulator cockpit is a widely used experimental platform. Immersion is one of the key characteristics. One way to improve visual realism is to use virtual reality (VR) devices that will result in two problems: 1. The dizziness caused by the serious mismatch between the fixed seat and the dynamic virtual graphics; and 2. VR glasses will cover the driver's face, making it impossible to conduct research on the driver's state [2]. Therefore, this study proposed a vision-based driver head tracking system to improve immersion and interaction, as shown in Figure 1. This technique can also be used to improve the user experience of HUD for intelligent vehicles and other driver-in-the-loop applications.

The head pose is an important clue that has been used in several human–machine interaction fields. [3] proposed an orientation sensor-based head tracking system to monitor the behavior of drivers engaging in various non-driving activities. [4] presented a sensor fusion method that integrates the IMU, IR LED, CCD camera, and other sensors. [5] developed a low-cost head tracking device based on the SteamVR tracking technology for a VR system. These methods typically adopt different types of sensors to build the system. There are several similar products in flight simulators. They typically require special devices or optical markers, such as an infrared camera. Although some devices require only an RGB camera, they all require the user to manually adjust the relative parameters, and they typically use certain traditional head estimation methods. Therefore, this study proposed a low-cost and mark-less solution that is only dependent on the RGB sensor as the input device, and a dynamic head pose estimation model based on deep learning was developed to improve the accuracy of the system.

Currently, several types of head pose estimation models combined with multimodality inputs have been proposed to achieve a state-of-the-art performance. These methods can be divided into model-based and model-free methods [6]. Model-based methods typically use a deformable head model to fit the input image. They also locate the facial landmarks to align with the predefined model. Generally, these methods are time-consuming. The model-free approaches are more popular; they train a regression model to map the head image to the pose manifold, and deep learning-based models are basically adopted. To improve the model performance, facial landmarks are also leveraged in certain model-free methods that can be used with vision geometry algorithms or multi-task learning to estimate the head pose [7]. To eliminate the influence of the illumination intensity, the depth image is explored to obtain more robust head poses under poor illumination or large illumination variations. The depth image can also provide additional depth information to improve the model accuracy [8]. These methods estimate the head pose independently for each frame. As a dynamic head tracker, this study focused on leveraging the prior frame to improve the performance of the model. A recurrent neural network (RNN) is a widely used model to handle sequential data and can be combined with a convolutional neural network (CNN) to handle video-based tasks. Recently, self-attention-based models, particularly vision transformers, have shown great potential in multiple tasks [9]. They outperformed inductive bias methods, including the CNN and RNN models based on a large dataset. However, these transformers typically focus only on either the spatial information of the image or the temporal features of the sequential data. Therefore, this study proposed a novel spatial-temporal vision transformer structure that can achieve better performance. It was compared and analyzed using a CNN–RNN-based model.

The estimated curve of consecutive frames fluctuated owing to the error variance of the model. A Kalman filter was used for post-processing to address this problem. By analyzing the error distribution of the estimation model, the AKF improved the performance of filtering, which includes an adaptive observation noise coefficient; it adaptively moderated the smoothness and maintained the curve stable near the initial position.

The main contributions of this study are as follows: 1. A low-cost and effective system for dynamic driver head tracking is proposed, which uses only a normal RGB camera; 2. A novel ST-ViT, which uniquely integrates a spatial vision transformer and a temporal transformer, is proposed, and to the best of our knowledge, this is the first time that a vision transformer is used in the dynamic head pose estimation; and 3. According to the characteristics of the head pose estimation model, an AKF is proposed to improve the stability and continuity of the dynamic head tracking system.

Spatial-Temporal Vision Transformer Based Head Tracker

The purpose of this study is to develop a vision-based dynamic head tracker and implement it on a driving simulator whose view can be automatically aligned with the driver's head pose using a frontal RGB camera. The benefits are as follows: 1. This can improve the immersion and interaction of the simulator. The driver's view will be unconstrained and non-fixed, and the virtual camera will be synchronized with the driver's head pose; 2. The extracted head pose can also be used to monitor the driver's multi-state and further improve their experience in human-centric automotive applications; and 3. This is a low-cost solution that uses a non-invasive camera sensor.

The development of deep learning and computer vision technology provides the basis for the proposed method. Current state-of-the-art head pose estimation methods typically use a single frame as the input. In this study, the prior frame is leveraged, which is combined with the current frame as the input to improve the performance of the model. A novel ST-ViT is proposed to achieve this task. To smooth the inconsistency and volatility of the estimation, this study also proposes an AKF. The overall proposed architecture is illustrated in Figure 1.

Dynamic head pose estimation

Estimating the head pose, a crucial problem that has several applications, is a task that must infer the 3D pose (*pitch, yaw, roll*) of the head from the input image. There are several different methods that use multi-modal input data, including depth images, RGB images, and video clips. Considering the trade-off between system performance and cost, this study investigates the dynamic head pose estimation approach based on the RGB image.

With the development of deep learning, research on head

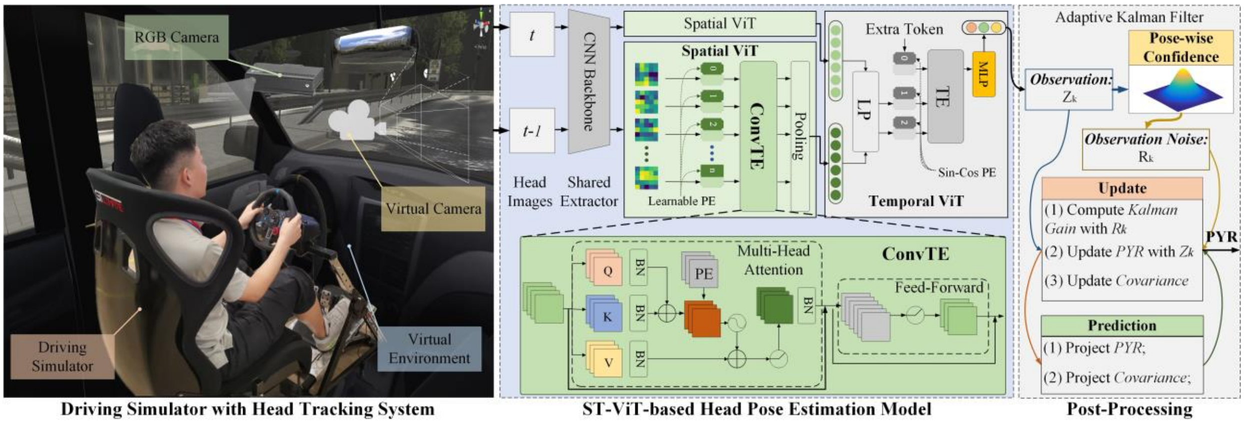


***Figure 1** Framework of the proposed dynamic driver head pose tracking system. The left picture shows the used driving cockpit which includes the input devices, computing server, and RGB camera. The proposed ST-ViT model is adopted as the measurer to estimate the pose, and its result as the observation. The proposed AKF is used to optimize the estimation. Finally, the virtual camera of the simulator is aligned with the output of the framework.*

pose estimation has also achieved good results, but they typically use a single frame as the input. In this study, the prior frame is leveraged, which is combined with the current frame as an input pair. Generally, an RNN is a widely used model to handle this type of sequential data, and it can use a CNN as the feature extractor to handle video-based tasks. We adopt this type of structure model in this study. Notably, the transformer model has shown significant potential, particularly in natural language processing (NLP). Some researchers have also begun to apply it to computer vision tasks and have proposed some vision transformer models [9]. Compared to inductive bias models, such as CNN and RNN, the transformer can better handle a large amount of data and achieve better performance on large datasets.

To handle the dynamic driver head pose, this study proposes a ST-ViT architecture, as illustrated in Figure 1. The input pair includes loosely cropped images from a face detector. It allows the model to focus on the head area and is easy to train. The ST-ViT adopts a pre-trained feature extractor as the CNN backbone, rather than the standard vision transformer which requires large-size datasets for training. The feature extractor shares the weight between the input pair. The extracted feature maps are input into the spatial vision transformer modul.

In the spatial vision transformer (S-ViT) module, the positional embedding (PE) is learnable, and the transformer encoder is convolutional, as shown in the Convolutional Transformer Encoder (ConvTE) module of Figure 1. The *Query*, *Key*, and *Value* (QKV) are calculated through the convolutional layer, rather than the linear connection layer of the standard transformer.

$$\begin{aligned} QKV_{(u,v)} &= BN(Conv(x, W_{QKV}) \\ &= BN\left(\sum_i \sum_j w_{qkv_{u-i,v-j}} \cdot x_{i,j}\right) \end{aligned} \tag{1}$$

where BN denotes the batch normalization layer, $Conv$ denotes the convolutional layer without bias, and W denotes the corresponding weight kernel. Then, the QKV is used to extract the spatial attention information using the multi-head attention mechanism as follows:

$$\begin{aligned} x_{out} &= Conv_{out}(Attention(Q,K,V)) \\ &= Conv_{out}\left(softmax\left(\frac{QK^T + Pos}{\sqrt{d_k}}\right)V\right) \end{aligned} \tag{2}$$

where Pos denotes the position bias that is learnable and d_k denotes the dimension of the *Key*. The attention mechanism leverages the *Query* and *Key* to obtain the similarity or correlation of the feature maps or vectors, then a weighted sum with the *Value* are implemented. The convolution layer, rather than the linear layer, can determine the formal consistency of the feature maps, allowing the residue connection to be used to avoid network degradation. Using the convolutional multi-head attention module with the two convolutional feed-forward layers, the spatial dependency and relationship of the feature maps are expected to be obtained.

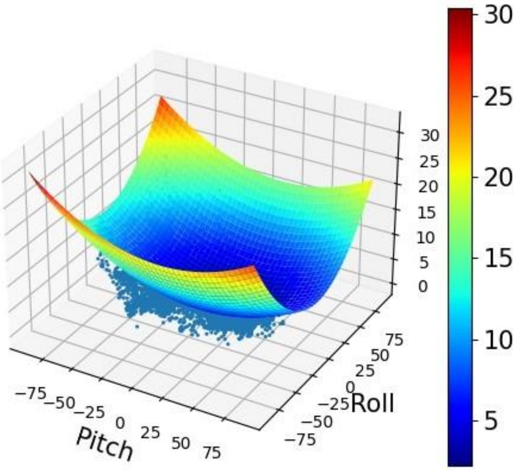


Figure 2 *Error of the proposed head pose estimation model on the pitch and roll axes tested on the BIWI dataset. The 3D blue point represents each sample, and the curved surface is the result of 2D Gaussian fitting.*

The temporal vision transformer (T-ViT) receives the feature vectors of the image pair through an average pooling layer, and a linear projection (LP) layer is used to embed the feature vectors. In this module, the positional embedding, which uses the sine-cosine function, calculates the position encoding, as shown in Equation 3. And an extra token is used to concatenate with the embedding vector that is designed to obtain the prediction of the last frame through the final multilayer perceptron (MLP) head.

$$\begin{cases} PE_{(k,2i)} = sin\left(k/10000^{2i/d_{vector}}\right) \\ PE_{(k,2i+1)} = cos\left(k/10000^{2i/d_{vector}}\right) \end{cases} \tag{3}$$

where $PE_{(k,2i)}$ represents the position encoding of the $2i-th$ position of the feature vector of the $k-th$ frame, and d_{vector} denotes the dimension of the feature vector. The temporal transformer module employs the standard transformer encoder (TE) with the linear layers for calculating the QKV.

The classic transformer is used to handle the sequential task to obtain the temporal dependency. The typical ViT is usually used to learn the spatial information by splitting an image into several patches. The proposed ST-ViT uniquely uses a spatiotemporal architecture to tackle the image sequences, and it also leverages a pre-trained CNN backbone to alleviate the dependence of the large dataset. In the spatial ViT, the convolutional layer, rather than the linear layer, is utilized, allowing the residue connection to be used for avoiding network degradation. Overall, the proposed ST-ViT can obtain the spatiotemporal attention, which is advantageous over the single dimension attention of the standard transformer.

Adaptive Kalman filter

Although the current head pose estimation method exhibits good performance, there is still a certain error. When it is applied to the simulator, its flaws of fluctuation and discontinuity are highlighted. From a practical perspective, the smoothness and continuity of view changes are more important than accuracy. To address this problem, a Kalman filter (KF) is adopted. Kalman filtering is an algorithm that provides the estimates of certain unknown variables that tend to be more accurate, given the measurements observed over time, and contain statistical noise and other inaccuracies. KFs have been demonstrated to be useful in various applications, such as the guidance, navigation, and control of vehicles. The KF has a relatively simple form and requires a small amount of computational power.

Due to the characteristics of approximate uniform and low-speed, the head pose motion can be described by using a linear coordination transformation model, which only involves the pose and velocity, as shown in the post-processing module of Figure 1. This is also beneficial for reducing the computing complexity. The output of the head pose estimation model is used as the observation of the AKF model. R_k is the observation noise covariance that is related to the estimation model and affects the performance of the filter. The results of the head pose estimation model are studied to determine R_k. Statistics revealed that the model usually has different performances at different intervals of the head pose. The accuracy is higher when the pose angle is small; otherwise, the error is higher, particularly on the pitch and roll axes. For example, the BIWI dataset [10] was used to evaluate the proposed head pose estimation model, and the results are shown in Figure 2. The error on the pitch and roll is taken as the X and Y axes, and the error on what is taken as the Z-axis. The blue 3D points represent different samples. A 2D Gaussian function is used to fit the points, as shown in the curved surface. Therefore, adaptive R_k, which can be adaptively adjusted in the iterative process, is proposed in this study. It can make the filtered value close to the observed value when the rotation angle is small, whereas the filtered value becomes smoother when the rotation angle is large.

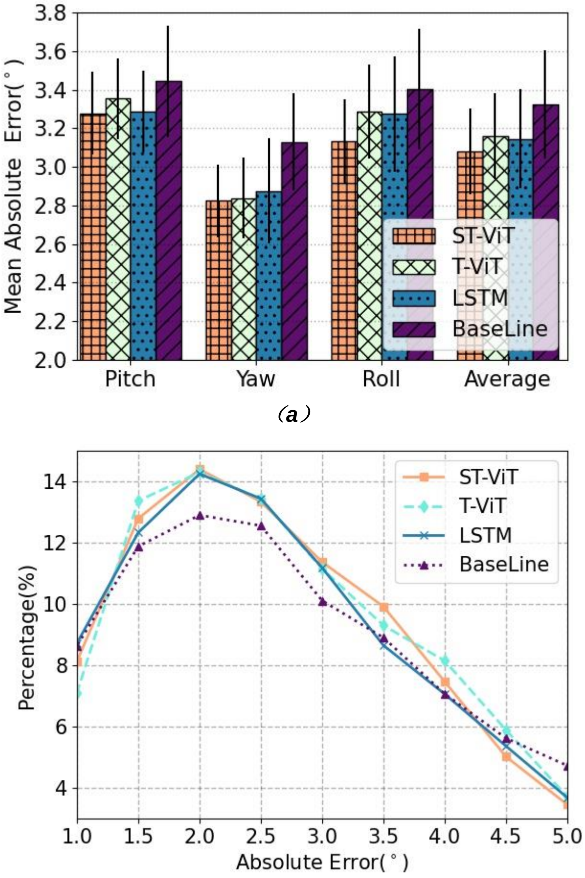
Evaluation and Results

Dynamic head pose dataset

The BIWI dataset, the only dataset suitable for our task, was chosen to evaluate the proposed method. The other datasets did not contain sequential images. The BIWI dataset contained 24 videos, which are over 15 K images of 20 subjects (14 males and 6 females). For each sample, an RGB image and the corresponding annotation were provided. The head pose range covered approximately ±75° yaw, ±60° pitch, and ±60° roll [10]. The ground truth was provided in the form of the 3D location of the head and its rotation, which can be converted to (*pitch, yaw, and roll*).

Model comparison and results

In this study, EfficientNet-B0 [11], a popular backbone, was used as the CNN backbone to extract the feature maps. It developed a new baseline network by performing a neural architecture search and optimized both accuracy and efficiency.



(a)

(b)

Figure 3 *Comparison of the different head estimation models. (a) The MAE* (°) *between different models. (b) The absolute error* (°) *distribution of different models.*

To verify the proposed method, four paradigms were designed as follows: 1. **Baseline** An MLP was leveraged to handle the feature maps extracted from the CNN backbone; the number of hidden layers was 512. 2. **LSTM** Compared to **the baseline**, an image pair was used as the input rather than a single image, and a long short-term memory (LSTM) module

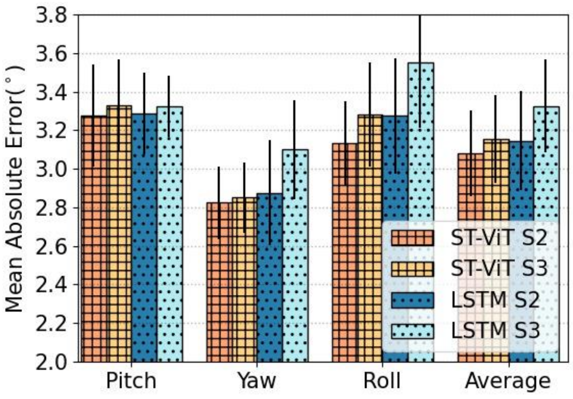


Figure 4 *Comparison of the different lengths of the sequence input. Sn indicates the length of sequence n.*

Model	Input	Pitch(°)	Yaw(°)	Roll(°)	Avg(°)
DeepHeadPose [13]	RGB	5.18	5.67	-	-
DeepHeadPose [13]	RGB+ Depth	4.76	5.32	-	-
SSR-Net-MD[14]	RGB	4.35	4.24	4.19	4.26
VGG16[14]	RGB	4.03	3.91	3.03	3.66
FSA-Caps-Fusion[14]	RGB	4.29	2.89	3.60	3.60
MultiLossResNet50[12]	RGB	3.39	3.29	3.00	3.23
FDNNet[15]	RGB	3.98	3.00	2.88	3.29
Martin[14]	RGB+ Depth	2.50	3.60	2.60	2.90
Baseline	RGB	3.44	3.12	3.40	3.32
LSTM	RGB	3.28	2.87	3.12	3.14
T-ViT	RGB	3.35	2.84	3.28	3.16
ST-ViT	RGB	3.27	**2.82**	3.12	**3.07**

Table 1 *Comparison of the state-of-the-art sequential-based models on the BIWI dataset.*

was used instead of the MLP to handle the image pair. The number of hidden layers was 512. 3. **T-ViT** In this paradigm, the LSTM module was replaced by a T-ViT. The depth of the transformer module was one, the number of heads was eight, the embedding dimension and the number of hidden layers of the MLP head were 512, and the dimensions of Q, K, and V were 64. 4. **ST-ViT** This is our proposed spatial-temporal vision transformer model. This included a spatial convolutional vision transformer to handle the feature maps first compared to the T-ViT. The depth of the spatial transformer and number of heads were same as that of the temporal transformer, and the dimensions of Q and K were 32, whereas that of V was 64.

We followed the common three-fold cross-evaluation experimental protocol proposed earlier in [12] that splits the dataset into 70% (16 videos) for training and 30% (8 videos) for testing. In the training process, the batch size was 16, Adam was used as the optimizer, and the learning rate was $1 \times e^{-4}$. The mean absolute error (MAE) was used as the metric, which is the same as in other studies. The average results are presented in Figure 3. The results indicated that the image pair effectively improved the performance, compared to the single image, especially on the *Yaw* axis whose value range is large. The changes in features caused by this axis are also more significant, so the overall error is smaller. Compared to the LSTM model, the T-ViT model was not always competitive; it had a smaller error only in the *Pitch* axis. Compared with others, the ST-ViT can effectively reduce the error variance, resulting in a significant improvement on the *Roll* axis which usually has a larger error variance. To further evaluate the performance of the models, the error distribution is also displayed, which can reflect the error percentage and distribution under different thresholds. Notably, the overall performance of the LSTM was better than that of the T-ViT because the used dataset did not have sufficient samples, and it could not completely reflect the learning ability of the transformer, especially the used T-ViT is not deep. However, the proposed ST-ViT can still outperform others and achieve the best performance. It has a great potential to handle a larger number of samples.

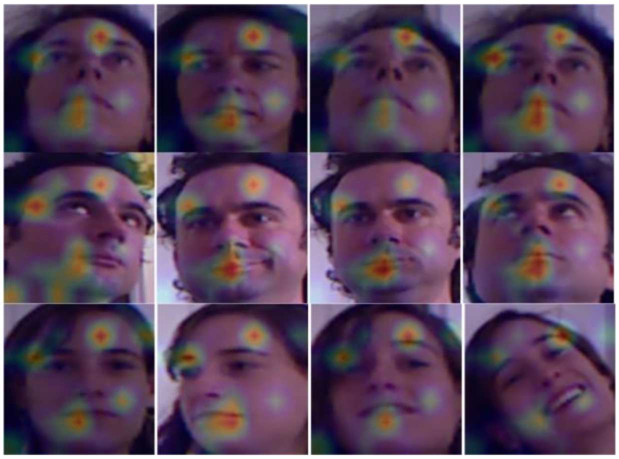


Figure 5 *Visualization of the attention map learned by the proposed ST-ViT model.*

The above results demonstrated that the image pair improved the performance of the models owing to the extra sequential information. To further analyze the effect of sequence, we used three consecutive frames as the input to train the ST-ViT and LSTM models, and the results are shown in Figure 4. The comparison indicated that a longer sequence degenerated the model that had a higher MAE in all three axes. This was because longer sequential information could not solve

Videos	Mean\Std	Method	Pitch (°)	Yaw (°)	Roll (°)
1-8	Mean	Original	3.795	3.206	3.619
		KF	3.809	3.253	3.631
		AKF	3.803	3.196	3.619
	Std	KF	3.649	3.556	4.342
		AKF	3.632	3.495	4.351
9-16	Mean	Original	2.552	2.313	2.568
		KF	2.618	2.371	2.595
		AKF	2.565	2.329	2.571
	Std	KF	2.473	2.003	3.061
		AKF	2.368	1.924	3.054
17-24	Mean	Original	3.485	2.955	3.218
		KF	3.586	3.121	3.248
		AKF	3.496	3.034	3.213
	Std	KF	3.105	2.595	4.306
		AKF	3.019	2.546	4.304

Table 2 *Comparison of the different filtering methods.*

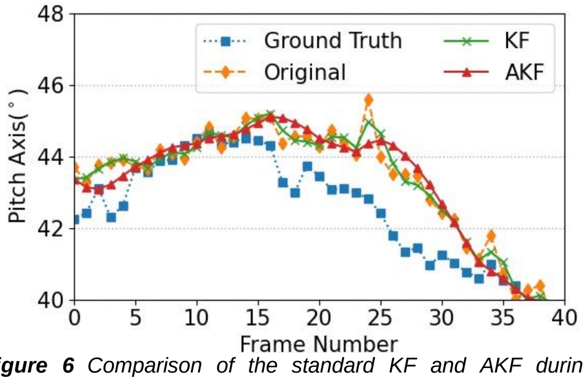


***Figure 6** Comparison of the standard KF and AKF during different angle ranges.*

the problem of cross-subject evaluation. Besides, the framerate of the BIWI videos was not high, thus the deviation between consecutive frames was large. If the framerate can be increased, the performance of longer sequences might be better. Further, the number of sequences can be easily adjusted in the proposed model according to different situations. Compared to ST-ViT, the degradation of LSTM was more serious. For the same length of input, ST-ViT outperformed LSTM. This demonstrated that ST-ViT was more robust than LSTM in handling the sequence data. ST-ViT learnt the relationship between consecutive frames and achieved better performance.

To comprehensively evaluate the proposed method, it was compared to other sequential-based methods, as shown in Table 1. These models also adopted the same training protocol. Because the BIWI dataset contained depth images, certain methods leveraged depth information to improve performance. Table 1 also demonstrates that the benefit of the RGB image is combined with the depth information. To improve the performance of the head pose estimation model, these methods designed different types of models and loss functions from different perspectives. Compared to RGB-based methods, our proposed method achieved state-of-the-art performance. Even on the yaw axis, our proposed method was superior to the depth-based methods. The performance of the baseline and LSTM methods also demonstrated the importance of the backbone, which provides guidelines for future research. A comparison with other methods demonstrated the effectiveness of the proposed method.

Figure 5 illustrates the visualization of the attention map learned by the ST-ViT model; the attention map for the input image was visualized through the attention score of self-attention. It was observed that the method could pay attention on the representative region of the face; they are coincidentally similar to the face landmarks, which can represent the facial expression and orientation. It demonstrated the learning ability of the proposed model. This helps us understand the mechanism of the spatial vision transformer.

Post-processing and results

To evaluate the proposed pipeline, ST-ViT was used to estimate the head pose on the BIWI dataset, and the results are shown in Figure 2 and Figure3. The head pose estimation model inevitably had an error and variance, and hence, it could not be directly used for dynamic head tracking. A reasonable method was to leverage a filter to smooth the curve. Considering that the head pose model had different performance under different angle ranges, this study proposed the AKF. We chose a sequence under large angle range of the pitch axis to illustrate, and the results are shown in Figure 6. The use of the KF smoothened the curve and reduced volatility. Notably, the ground truth also has measurement errors and deviations. This shows that it is necessary and reasonable to use a filter. For the standard KF, R_k is a constant value, which is the mean error of the head pose estimation model. To further improve the performance, constant R_k is replaced as the adaptive one, as mentioned above, and the related parameters are the results of the Gaussian fitting on the dataset. The comparisons are presented in Table 2. The standard KF and the filter with adaptive R_k almost coincided at a low angle range. However, the filter with adaptive R_k has better performance in the high-angle range, and the curve is smoother. The filtering algorithm is a compromise between accuracy and smoothness. The increase in smoothness inevitably loses a certain degree of accuracy. The proposed AKF can maintain accuracy while reducing variance. This is an advantage of adaptive R_k.

Conclusion

To improve the immersion and interaction of the driving simulator and related applications, this study proposed a dynamic head pose tracking system. The proposed system used only an RGB camera without other hardware or markers. To enhance the accuracy of the dynamic head pose estimation, this study proposed a ST-ViT model that used an image pair as the input instead of a single frame. Compared to the standard transformer, this contained a spatial convolutional vision transformer and a temporal vision transformer, which improved the effectiveness of the model. A comprehensive experimental comparison demonstrated that the proposed method outperformed the state-of-the-art methods. Another challenge to deploy the head tracking system was that the head pose estimation models still had certain errors, and hence, could not be directly adopted. To address this problem, this study proposed post-processing of the raw estimation. By analyzing the error distribution of the estimation model and user experience, an AKF was proposed that included the adaptive observation noise coefficient that makes the curve smoother in the area where the estimation model has a large error. The experiments showed that the proposed method was feasible and could be deployed into the driving simulator.

This paper proposed a reasonable low-cost vision-based solution for head tracking, which can be further optimized as the algorithm of head pose estimation improves. It can also be used in other driver-in-the-loop applications, and the source code of this paper will be open-sourced.

References

[1] S. Shah, D. Dey, C. Lovett and A. Kapoor, “AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles,” in *Advanced Robotics*, 621-635, 2018.

[2] Z. Hu, C. Lv, P. Hang, C. Huang and Y. Xing, “Data-driven estimation of driver attention using calibration-free eye gaze and scene features,” in *IEEE Trans. Ind. Electron.*, 1-1, doi: 10.1109/TIE.2021.3057033.

[3] Y. Zhao, L. Görne, I.-M. Yuen, D. Cao, M. Sullman, D. Auger, C. Lv, H. Wang, R. Matthias, L. Skrypchuk, and A. Mouzakitis, “An Orientation Sensor-Based Head Tracking System for Driver Behaviour Monitoring,” *Sensors*, vol. 17, no. 12, p. 2692, 2017.

[4] C. H. Kang, C. G. Park and J. W. Song, “An adaptive complementary Kalman filter using fuzzy logic for a hybrid head tracker system, ” in *IEEE Trans. Instrum. Meas.*, vol. 65, no. 9, pp. 2163-2173, Sept. 2016.

[5] A. K. T. Ng, L. K. Y. Chan and H. Y. K. Lau, “A low-cost lighthouse-based virtual reality head tracking system” in *Proc. IC3D*, Brussels, 2017, pp. 1-5.

[6] Z. Hu, Y. Xing, C. Lv and P. Hang and J. Liu, “Deep convolutional neural network-based Bernoulli heatmap for head pose estimation,” *Neurocomputing*, vol. 436, pp. 198-209, Jan. 2021.

[7] R. Valle, J. M. Buenaposada and L. Baumela, “Multi-task head pose estimation in-the-Wild,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2874-2881, Aug. 2021.

[8] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara and R. Cucchiara, “Face-from-depth for head pose estimation on depth images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 596-609, Mar. 2020.

[9] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou and M. Douze “LeViT: a Vision Transformer in ConvNet’s Clothing for Faster Inference,” *arXiv:2104.01136*, 2021.

[10] G. Fanelli, J. Gall and L. Van Gool, “Real time head pose estimation with random regression forests,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, 2011, pp. 617-624.

[11] M. Tan, Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Int. Conf. Mach. Learn.*, 2019, pp. 6105-6114.

[12] N. Ruiz, E. Chong and J. M. Rehg, “Fine-grained head pose estimation without keypoints,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 2074-2083.

[13] S. S. Mukherjee and N. M. Robertson, “Deep head pose: Gaze-direction estimation in multimodal video,” in *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2094-2107, Nov. 2015.

[14] T. Yang, Y. Chen, Y. Lin and Y. Chuang, “FSA-Net: Learning finegrained structure aggregation for head pose estimation from a single image,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1087-1096.

[15] H. Zhang, M. Wang, Y. Liu and Y. Yuan, “FDN: Feature decoupling network for head pose estimation,” in, *AAAI Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 12789-12796, 2020.

Acknowledgements

This work was supported in part by in part by A*STAR National Robotics Programme under Grant SERC 1922500046, the SUG-NAP under Grant M4082268.050, Nanyang Technological University, Singapore, and the State Key Laboratory of Automotive Safety and Energy under Project (no. KF2021).