Naas, Si-Ahmed; Jiang, Xiaolan; Sigg, Stephan; Ji, Yusheng

Functional Gaze Prediction in Egocentric Video

Published in:
18th International Conference on Advances in Mobile Computing and Multimedia, MoMM2020 - Proceedings

DOI:
10.1145/3428690.3429174

Published: 30/11/2020

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Naas, S.-A., Jiang, X., Sigg, S., & Ji, Y. (2020). Functional Gaze Prediction in Egocentric Video. In P. D. Haghighi, I. L. Salvadori, M. Steinbauer, I. Khalil, & G. Kotsis (Eds.), *18th International Conference on Advances in Mobile Computing and Multimedia, MoMM2020 - Proceedings* (pp. 40-47). ACM.
https://doi.org/10.1145/3428690.3429174

Functional Gaze Prediction in Egocentric Video

Si-Ahmed Naas
Department of Communications and Networking, Aalto University, Finland
si-ahmed.naas@aalto.fi

Xiaolan Jiang
Department of Informatics, The Graduate University for Advanced Studies, Japan
xljiang@nii.ac.jp

Stephan Sigg
Department of Communications and Networking, Aalto University, Finland
stephan.sigg@aalto.fi

Yusheng Ji
Information Systems Architecture Science Research Division, National Institute of Informatics, Japan
kei@nii.ac.jp

ABSTRACT

Streaming 360° videos to a head-mounted display (HMD) client is challenging due to their high network resource consumption and computational load. This is due to the use of gaze point prediction or image saliency features from the field of view (FoV) since, in real-time scenarios, FoV extraction is computationally demanding. We propose a functional gaze prediction system that addresses these issues by relying on a tiling scheme for gaze prediction. We condition gaze point prediction on virtual reality (VR) content and long short-term memory (LSTM)-encoded eye movement history. Further, we encode image flow and saliency maps of RGB images via VGG16, using a convolutional neural network (CNN). Future gaze points are then predicted using a novel sinusoidal encoding technique. In experiments, our tile-based approach outperforms state-of-the-art FoV-based schemes in terms of computational load and predicted gaze position.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing systems and tools*; **Virtual reality**; **Mixed / augmented reality**; **Mobile devices**; • **Computing methodologies** → **Probabilistic reasoning**; **Mixed / augmented reality**.

KEYWORDS

Pervasive HMD interaction, 360° video, convolutional neural network, gaze prediction, machine learning, virtual and augmented reality

ACM Reference Format:
Si-Ahmed Naas, Xiaolan Jiang, Stephan Sigg, and Yusheng Ji. 2020. Functional Gaze Prediction in Egocentric Video. In *The 18th International Conference on Advances in Mobile Computing and Multimedia (MoMM '20), November 30-December 2, 2020, Chiang Mai, Thailand*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3428690.3429174

MoMM '20, November 30-December 2, 2020, Chiang Mai, Thailand

ACM ISBN 978-1-4503-8924-2/20/11…$15.00
https://doi.org/10.1145/3428690.3429174

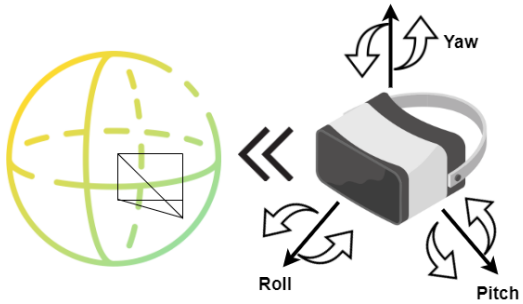


Figure 1: HMD rotation in 360-degree video

1 INTRODUCTION

Driven by technical and algorithmic advances, Augmented Reality (AR) and Virtual Reality (VR) are increasingly important in pervasive and ubiquitous computing domains. Particularly, the AR market is expected to reach 70–75 billion by 2023 [1] while the VR market may exceed 120 billion by 2026 [2]. Many AR and VR applications involve the streaming of 360-degree videos, only part of which is actually presented to an AR/VR user in her field of view (FoV) (cf. Fig. 1). A person explores such scenes through a set of consecutive fixations of the gaze, which indicate interest or attention in a scene. Therefore, gaze point is considered as a key element in first-person vision [1].

In computer vision, first-person view perspectives are prominent. It can be found in applications, such as ubiquitous smart space [2], human–robot interaction (HRI), e-health, social analysis, group identification [3], handled objection recognition [4], video summarizing [5] and editing [6], event recognition [7], augmented reality [8], or virtual reality [9].

In these domains, predicting gaze information (Fig. 2) improves the user experience through the pre-allocation of VR rendering resources.

Recent advances in technology, in particular, head-mounted displays (HMDs), challenge the existing algorithms in the field.

Without knowledge about the field of view (FoV), 360° high-definition video applications require streamed full panorama scenes that contain both visible and invisible parts of the FoV, thus overly

[1] https://www.trendforce.com/presscenter/news/20151204-9123.html
[2] https://www.globenewswire.com/news-release/2020/06/15/2048254/0/en/Virtual-Reality-VR-Market-to-Touch-120-5-Billion-by-2026-Rapid-Advancements-in-Deep-Technology-Domain-to-Brighten-Market-Outlook-Fortune-Business-Insights.html

Figure 2: Illustration of gaze points in a sequential scene of four time instances.

straining the bandwidth-limited network [10]. In addition, real-time field of view (FoV) extraction, such as saliency maps, image flow, and gaze history for gaze prediction, easily exceeds the processing resources of mobile hardware, such as VR HMDs or mobile phones [8, 11, 12].

Pre-processing the FoV into a vector of its features could significantly reduce the computational load of extracting the FoV, salient map, and image flow features. However, since the FoV is not known during the encoding phase, we propose a tile-based approach that divides frames into multiple tiles and pre-encodes them to feature vectors. Only those predicted vectors that correlate with the FoV are then utilized during real-time gaze prediction.

As a result, only tiles in the FoV are streamed at a high quality and the bitrate is significantly reduced, without affecting the quality of experience (QoE).

2 RELATED WORK

The FoV of a person is guided by objects or stimuli of higher relevance to the viewer. It is indirectly related to saliency, the aspect of any stimulus that makes it stand out from its surroundings, such as contrast, movement, or color. Objects of higher relevance further draw the gaze of the viewer. Both saliency and gaze have been predicted in the literature, and in particular, studies on egocentric videos are related to our work.

In pervasive and ubiquitous computing domains, gaze-point prediction has been applied, for instance, in egocentric video [1] by exploiting the implicit behavior cues observed from the wearer of a camera system or for the automatic inference of relevance of real-world objects [13]. To support the experience of head-mounted displays, Gaze prediction was further utilized for the estimation of gaze-estimation error [14] as well as for continuous self-calibration in eye-gaze tracking in head-mounted VR-systems [15]. Recently, Santini et al. presented a slippage-robust and glint-free gaze estimation for pervasive head-mounted displays [16].

2.1 Saliency Prediction

Saliency detection has been extensively investigated in computer vision. It comprises bottom-up approaches that use low-level features such as color and scene orientation [17–20], top-down approaches based on high-level features such as scene context [21–24], and hybrid approaches [25–28]. However, few studies address gaze prediction for VR and, in particular, saliency in 360° video [11]. The authors in [29] studied the relationship between salient objects and the human brain and found that the first cortical visual area (V1) interacts with the bottom-up saliency signals, the secondary visual cortex (V2) interacts with top-down saliency signals, and the fourth human visual field map (hV4) shows a convergence between top-down and bottom-up signals. Hence, saliency maps are efficient information for an accurate gaze prediction system. For instance, for static VR images, [30] studied the viewing behavior and saliency in a VR scene and predicted saliency based on head and gaze data as well as on user behavior. In contrast, [31] exploited a shallow and a deeper convolutional neural network (CNN) for saliency prediction in a regression rather than a classification approach. The authors employed a loss function to measure the distance between predicted and ground-truth saliency maps. The approach was further improved in [32] by addressing the training of the network on an adversarial loss function. The two employed networks predicted the saliency map and identified whether the predicted map resembled the ground truth.

Our approach follows this successful solution in our saliency encoder component. An example for salient object extraction from VR videos is [33]. An agent that follows the scene foreground and viewing angle selects objects in a scene before a recurrent neural network (RNN) predicts the main object and, finally, the FoV.

2.2 Gaze Prediction in Egocentric Videos

Despite the great advancement in image and video processing, only a few authors have focused on gaze prediction in egocentric videos [34].

The majority of existing work focuses on gaze prediction using bottom-up models [35] (color, orientation, intensity, etc.) such as in [36] [37] or top-down models (such as object context) such as in [38] [39]. For instance, [1] proposed a gaze prediction scheme using a random regression forest. This technique combines a user's head motion and hand location and evaluates it on gaze data from eye-tracking glasses. Because the viewer's hand is seldom involved with objects in the FoV, this technique cannot be generalized to HMD domains.

In [12] a novel CNN is proposed for gaze prediction based on combining object position, head velocity, and saliency features. The authors conducted their experimental dataset composed of 43 users for dynamic scenes. However, their evaluation study remains limited in terms of objects considered in a scene. The collected data from users consists of only free-viewing scenarios in a silent environment, and as for moving objects, only animals were considered.

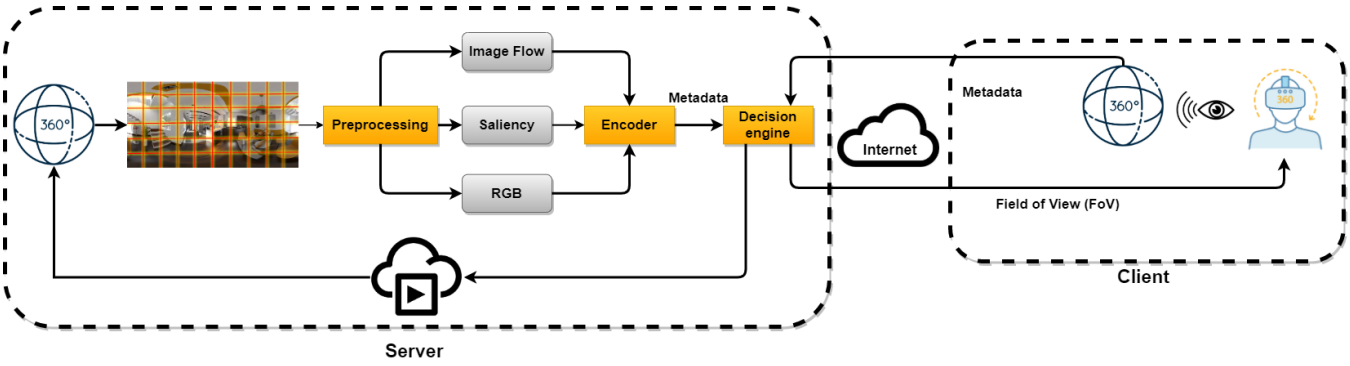


Figure 3: Schematic view of the proposed system in a server architecture. The FoV is selected according to the gaze prediction

In [35], a temporal continuity for visual attention is presented. Authors target accurate temporal continuity as an indicator of accurate gaze prediction. An autocorrelation function is proposed for temporal continuity evaluation. Their work was evaluated on task-oriented conditions and a free-viewing scenario. However, this technique can not be generalizable as it is only evaluated in VR game scenes. In [40], a sparse coding-based saliency method was employed for gaze prediction. The work targeted the noisiness and computational complexity in saliency maps and used canonical correlation analysis (CCA) to merge different image features. However, the experiment is made based on the eye-tracking glasses dataset, which is not enough to validate the effectiveness of this approach.

In addition, based on bottom-up visual saliency, the authors in [41] extracted attention maps based on the camera's rotation velocity and how it moves. However, since the camera's information remains unknown, this approach cannot be applied in our case.

Recently, [8] proposed a generative adversarial neural network-based model (GAN). Their approach consisted of deep future gaze (DFG) to generate future frames based on a single frame, then predicted temporal saliency maps for the upcoming seconds. The GAN network was composed of a generator network (GN) and a discriminator network (D). The GN constructed N frames from the latent representation of the current frame, and N saliency maps. The DN determined fake frames from real frames based on scene semantic and foreground/background coherence.

Another recent work by [11] proposed to predict viewer behavior using a long short-term memory (LSTM) network in the current FoV. Then, they fed saliency maps in the current FoV to a CNN for feature extraction. The authors merged the features of gaze information and saliency maps to estimate the gaze displacement between two consecutive moments. However, due to long processing times and latency, it is not practical for deployment in real-world cases.

We propose a *lightweight* tile-based gaze prediction method using video content features and gaze history to predict gaze displacement. Our system can be utilized for any gaze prediction scheme. We aim to significantly reduce the processing time from FoV extraction. We also propose a novel sinusoidal encoding that shows a significant improvement over angle normalization-based encoding. To make our system realistic and more challenging, we ensure the diversity of scene content in our benchmark dataset (indoor, outdoor, moving objects, etc.).

This is the first work that targets tile-based gaze prediction along with saliency maps. In contrast to other schemes, our system can be deployed on the client side. Experiments show the efficiency of our system in terms of prediction accuracy, computational load, and time cost.

3 FUNCTIONAL GAZE PREDICTION

Extracting the FoV in real time for gaze location anticipation suffers from high processing load. To address this challenge, we propose a tile-based approach exploiting saliency maps and gaze sequence history. Our system is composed of a video content encoder, a gaze sequence encoder, and a gaze displacement module (cf. Fig. 4).

First, we split the 360° video into tiles before extracting saliency maps from each tile where $v = \{\tau_1, \tau_2, ..., \tau_n\}$. Each FoV is composed of set of tiles (τ_i) from v.

The image flow is then derived for each pair of consecutive tiles. Different frame representations are further encoded in the video content encoder using a VGG16 [42] CNN. On the obtained feature vectors, we perform the element-wise product, while the gaze history points are encoded separately using an LSTM.

Finally, features from the CNN and LSTM (gaze module and video content module) are concatenated and fed into a fully connected layer for gaze prediction. From these gaze points, the tiles that cover the FoV are extracted.

When a viewer is watching a 360° video, we assume that videos are uploaded to the server and have been pre-converted to tiles based on equirectangular projection (Fig. 3). Meta-information (video ID and gaze movement) is sent to the server. Based on this information, at time t, the decision engine predicts the gaze at time $t+1$. Only those tiles that cover the field of view (FoV) are streamed with high quality, while tiles outside the FoV will be of low quality.

3.1 Gaze History Encoder

We encode viewer gaze history points by computing the horizontal viewing direction, the yaw (0° to 360°) and the vertical viewing direction, the pitch (0° to 180°) as

$$Yaw = \frac{Yaw_{(r)} + Yaw_{(l)}}{2} \tag{1}$$

$$Pitch = \frac{Pitch_{(r)} + Pitch_{(l)}}{2}, \tag{2}$$

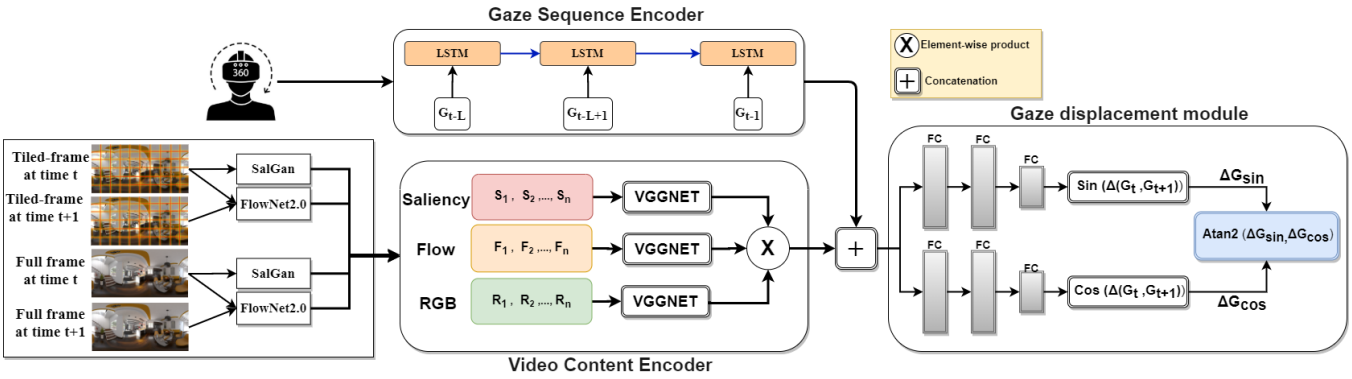


Figure 4: The proposed tile-based gaze prediction architecture. The output of the Gaze sequence encoder and the video content encoder is combined as input for the Gaze displacement module

where r is the right eye and l is the left eye coordinate.

We select 10 frames as the length of our sequence and encode the history of yaw $[Yaw_{t-N+1}, ..., Yaw_t]$ and, likewise, of pitch $[Pitch_{t-N+1}, ..., Pitch_t]$ using two LSTM layers with 128 neurons each. The length of the sequence is selected by experiments.

3.2 Video Content Encoder

The prediction error experienced when only gaze history is used can be reduced by exploiting other features from the scene. Motivated by the results obtained in [11], we consider image saliency, image flow, and RGB to be image features.

Image Saliency: The most salient objects correlate with gaze points [11] [43]. We calculate saliency maps with SalGan [32].

Image Optical Flow: Moving objects usually correlate with gaze points. We compute the image flow and include the use of FlowNet 2.0 [44] on pairs of consecutive frames.

RGB Images: Viewers can freely move their heads to different parts of the scene, results in FoV changing. It is therefore important to maintain whole scene information.

We divide frames into tiles and feed them into the VGG16 [42] network for feature extraction. VGG16 is known for its ability to extract robust image features.

3.3 Sinusoidal Gaze Displacement Module

Provided with the predicted gaze direction, bandwidth consumption can be reduced by using high-quality streaming for only those parts of a scene that are in the FoV. We propose a novel sinusoidal encoding for gaze prediction. The displacement module takes as input the encoded features of video content and gaze sequence history.

Yaw and pitch angles are converted to radian, then encoded to Sin(Yaw), Cos(Yaw), Sin(Pitch), Cos(Pitch). A benefit of this encoding is the reduced angle periodicity (e.g. -90 and 90 refer to the same direction) (cf. Fig. 5), which eases the learning.

Since *arctan* expects angles from quadrants I or IV, we use *atan2* [3] to obtain the original angles in the radian:

$$\Delta(angle_{t+1}, angle_t) = atan2\left[sin(\Delta_{angle}), cos(\Delta_{angle})\right] \quad (3)$$

[3]https://en.wikipedia.org/wiki/Atan2

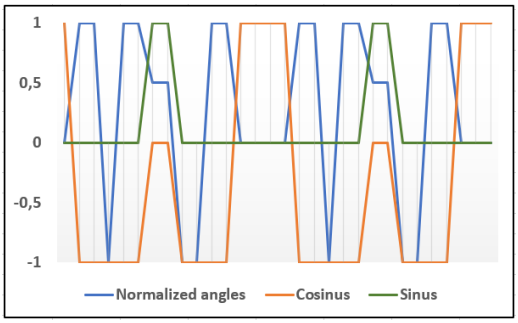


Figure 5: Illustration of gaze movement on yaw direction. Effect of sinusoidal encoding on the angles.

with

$$atan2(y,x) = \begin{cases} 2arctan(\frac{y}{\sqrt{x^2+y^2}+x}) & \text{IF } x > 0 \\ 2arctan(\frac{\sqrt{x^2+y^2}-x}{y}) & \text{IF } x \leq 0, y \neq 0 \\ \pi & \text{IF } x < 0, y = 0 \\ \text{Not defined} & \text{IF } x = 0, y = 0. \end{cases} \quad (4)$$

We then obtain the predicted angles using

$$angle_{t+1} = angle_t + \Delta(angle_{t+1}, angle_{t+1}). \quad (5)$$

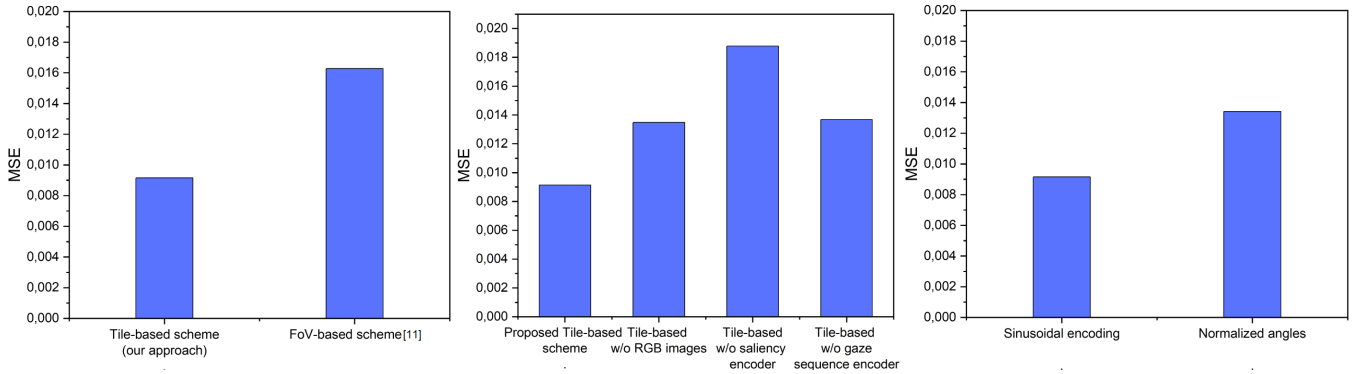
The video content features and gaze sequence encoding are transformed as

$$\Delta\left(angle_{t+1}, angle_t\right) = \xi(\{V^e_{t+1}, angle^e_{t+1}\}). \quad (6)$$

The function $\xi()$ represents two fully connected layers with 128 and 50 neurons, and mean square error(mse) as the loss function. We predict for 1 second; however, the duration can be increased by feeding the output of this module as input.

4 EVALUATION

We performed an experiment to validate the effectiveness of our system where we used the MSE of the predicted gaze to ground



(a) Tile-based and FoV-based approaches **(b) w/o saliecy maps, image flow, RGB images** **(c) Sinusoidal and normalized encoding**

Figure 6: Comparison of errors in the predicted gaze (mean squared error) for various configurations.

truth as our evaluation metric. Our system was implemented and tested with an NVIDIA Tesla V100 Volta GPU Accelerator 32GB Graphics Card and 128GB of RAM.

4.1 Datasets

We employed the dataset presented in [45], which comprised 57 participants (32 males, aged 19–44 yrs) watching 19 360° videos with an unconstrained viewing experience where the videos were displayed in a VR headset. All participants had a normal or corrected vision, checked with the Ishihara test. Each participant watched 4k videos (categories indoor, outdoor, urban, people, water, rural, nature) for a duration of 20 seconds.

4.2 Experimental Setup

We selected 11 videos for training and eight for testing, with no overlap between videos. Videos were processed to equirectangular projection, converted to frames using the FFmpeg software tool, and tiled as 8x4. Flows were converted using Flownet 2.0 [44][4]. We implemented our system using the TensorFlow 2.0 framework. The model, with sigmoid activation in the last dense layer, was trained with stochastic gradient descent (SGD) at a learning rate of 0.01 and batch size 32.

4.3 Evaluation Metric

We computed the prediction error as the MSE of the ground truth displacement angle ($\Delta_i(Y_{t+1}, Y_t)$) and the predicted gaze position ($\Delta_i(\hat{Y}_{t+1}, \hat{Y}_t)$) (where smaller is better).

For N frames, the prediction error was calculated as

$$Prediction\_err = \frac{1}{N}\sum_{i=0}^{N}[\Delta_i(Y_{t+1}, Y_t) - \Delta_i(\hat{Y}_{t+1}, \hat{Y}_t)]^2 \quad (7)$$

4.4 Comparison of System Components

We compare image flow, saliency maps, RGB images, and gaze sequence encoder with their impact on the mean squared prediction error on the yaw direction. Moreover, we compared the tile-based scheme with real-time FoV extraction. For this purpose, we re-implemented an algorithm proposed in [11]. Fig. 6a shows that our system performed better than the FoV-based approach. Similarly, in Table 1, the required preprocessing time of our approach was about 3 ms, while the FoV-based approach required 47 ms for gaze prediction.

Fig. 7a shows the benefit of combining video content features and gaze sequence history by comparing the saliency encoder, gaze sequence encoder, and our system. We further analyzed the saliency encoder, gaze sequence encoder, and RGB frames, with the results shown in Fig. 6b. Eliminating any element from our system raised the prediction error. The saliency encoder performed better than the saliency map and image flow alone (cf. Fig. 7b). In Fig. 7c we show the proof that the encoding of full frames results in large prediction errors.

4.5 Prediction Accuracy

In pose estimation [46], and image classification [47], the prediction of displacement achieved better results than direct values. Gaze prediction in [11] confirmed the effectiveness of the displacement approach. We proposed a novel technique based on sinusoidal encoding, and Fig. 6c indicates the efficiency of our displacement encoding scheme.

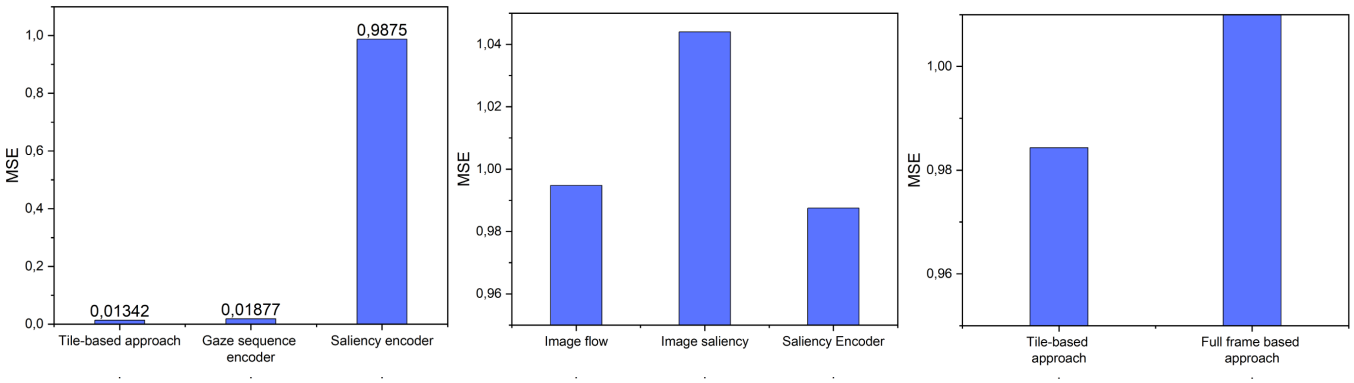
4.6 Computational Load

We investigated the computational load of a server-based deployment (cf. Fig. 3). The reported prepossessing time included the retrieval of future tiles. Table 1 presents the preprocessing time on the server, showing that our scheme required approximately 15x less time than the FoV scheme proposed in [11].

We obtained this figure by running this experiment multiple times and taking the average preprocessing time. These results prove the effectiveness of our system when deployed on the server, and it smoothes the transition of tiles and improves the user experience.

We further evaluated the RAM and CPU usage rates to compare our approach with the FoV in real-time. To realize this experiment, we ran both approaches for simulation of one video that lasts for 66 seconds, then we reported measurements every 3 seconds. Note

[4]Flowiz. Available from https://github.com/georgegach/flowiz



(a) Gaze prediction error **(b) Different components of the model (MSE)** **(c) Tile-based vs. full frame based**

Figure 7: Error in the predicted gaze (mean squared error) for various configurations.

Table 1: Comparison of Processing Time and File Sizes Between Proposed Tile-Based and FoV-Based Approaches

Approach	**Tile-based scheme**	**FoV-based scheme [11]**
Time (ms)	**3.25**	**47**
Size (kB)	**25**	**666**

that our approach retrieves only correspond tiles while the FoV-based approach requires a loading phase of saliency and image flow frameworks. The results of this experiment are summarized in Fig. 8b and Fig. 8c where our approach showed a less computation load (RAM and CPU) compared to the FoV approach. The RAM usage for the FoV-based approach is approximately 6% while our scheme uses only 2% of the RAM. Similarly, for the CPU usage, our scheme remains stable and uses less than 2% of CPU while the FoV-based approach loads the frameworks which explain the increase in CPU usage until it reaches 39%. As a conclusion, the low computational load that is justified by the needless of loading of saliency and image flow frameworks during the gaze prediction process, which makes the proposed scheme practical for real-world deployments.

To reduce the size of image files, we compressed the obtained tiles that were originally in PNG format and converted them into low image resolution (cf. Fig 8a). Our system remains superior even with low image quality; therefore, it optimizes resource consumption and can be easily deployed on the client user equipment (UE).

5 EYE-TRACKING AND VR/AR APPLICATIONS

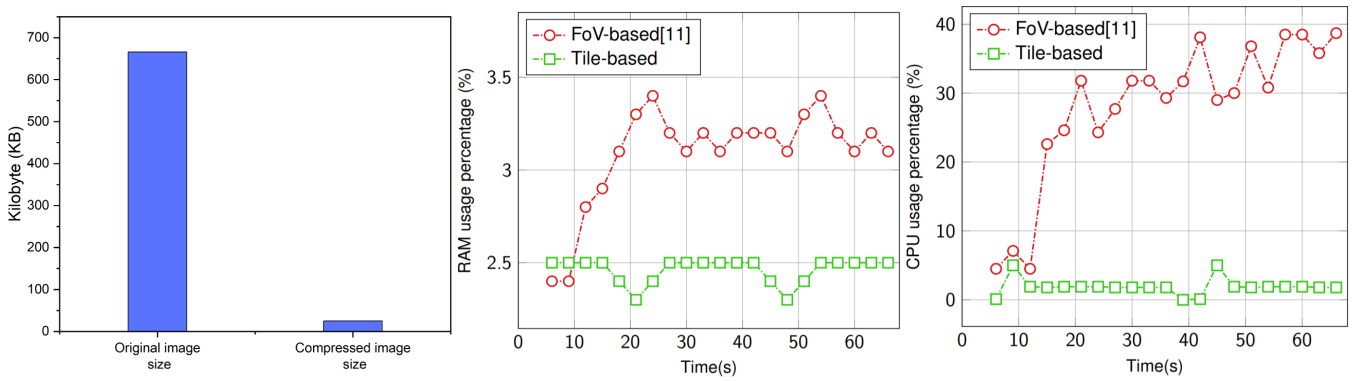
The rapid development in AR/VR technologies and wearable devices makes their involvement in different sectors accompanied by huge market investments. In particular, applying VR in ubiquitous environments has attracted a lot of attention, with the role of expanding human capabilities [2] in a space surrounded by smart objects and smart interactive tools.

During COVID-19 pandemic, AR and VR are increasingly becoming emerging solutions to satisfy user expectation [48] such as in e-learning, virtual shopping, remote socialization for the objective of minimizing physical contact thus minimizing possible infections. With advanced sensing of audio, video, interaction in VR scenes, an immersive experience is provided.

The eye-tracking has gained a lot of attention in VR systems. The gaze information plays a pivotal role in the success of many of its applications. For instance, Hartholt et al. in[49] presented a platform for the development of virtual humans that includes VR and AR technologies. The proposed framework leverages the technologies inside a room for audio-visual sensing. Instead of showing an avatar, the framework produces an animated human being personified by the real user. The framework receives feedback from the user such as eye position through a mobile sensor in the headset. Berton et al. [50] studied the gaze behavior in collision avoidance for the VR environment between a real human and VR character. The authors compared the user's gaze and mobility behavior in both the real scenario and VR setup. The study finds that the collision avoidance patterns are similar in real and virtual environments regardless of VR setup conditions. To investigate the behavior of a pedestrian in a crowded environment, MeerhoffL et al. [51] and in a virtual environment, studied the gaze points captured from HMD along with trajectory and found a relationship of gaze points with movement adjustments between neighbor walkers. Mardanbegi et al. in[52] presented a novel technique to interact with an object in a VR environment based on gaze tracking. Their technique relies on choosing an action, and then apply it to an object at the line-of-sight. Salehin et al. [53] proposed a novel technique for video summarization. At first smooth pursuit is identified, then during smooth pursuit, distance gaze is calculated, finally, keyframes are selected according to a probability score.

6 CONCLUSION

We proposed a solution for gaze prediction in dynamic scenes, often encountered in ubiquitous and pervasive applications using head-worn displays. In particular, we addressed the long processing time on pervasive and mobile displays, which is caused by extracting the Field of View (FoV) in real time by segmenting the video into smaller tiles before conducting the gaze prediction. Our tile-based gaze prediction converts image saliency, image flow, and full scene images to tiles before encoding them using VGGnet. We further



(a) Comparison between the size of the original and the compressed image

(b) RAM usage percentage

(c) CPU usage percentage

Figure 8: Computational load comparison between tile-based and FoV-based approaches

proposed a sinusoidal encoding that results in a more accurate gaze prediction. We implemented and compared our proposal to the FoV-based scheme proposed in [11].

Our approach showed superior performance in terms of gaze prediction accuracy, time cost, and computational load, all of which are of particular practical relevance in Mobile, ubiquitous and pervasive domains. Summarizing, our proposed scheme can be effectively deployed jointly on a server and in collaboration with a pervasive, ubiquitous or wearable client user equipment of limited resources.

ACKNOWLEDGMENTS

The authors appreciate partial funding from Nokia Solutions and Networks.

REFERENCES

[1] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3216–3223, 2013.

[2] Bernadetta Kwintiana Ane, Dieter Roller, and Jagadish Lolugu. Ubiquitous virtual reality: The state-of-the-art. 2019.

[3] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012.

[4] Xiaofeng Ren and Matthai Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2009.

[5] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2244, 2015.

[6] Eakta Jain, Yaser Sheikh, Ariel Shamir, and Jessica Hodgins. Gaze-driven video re-editing. *ACM Transactions on Graphics (TOG)*, 34(2):1–12, 2015.

[7] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE, 2012.

[8] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4372–4381, 2017.

[9] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[10] Mohammad Hosseini and Viswanathan Swaminathan. Adaptive 360 vr video streaming: Divide and conquer. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 107–110. IEEE, 2016.

[11] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze prediction in dynamic 360 immersive videos. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5333–5342, 2018.

[12] Zhiming Hu, Sheng Li, Congyi Zhang, Kangrui Yi, Guoping Wang, and Dinesh Manocha. Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE transactions on visualization and computer graphics*, 26(5):1902–1911, 2020.

[13] Melih Kandemir and Samuel Kaski. Learning relevance from natural eye movements in pervasive interfaces. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 85–92, 2012.

[14] Michael Barz, Florian Daiber, and Andreas Bulling. Prediction of gaze estimation error for error-aware gaze-based interfaces. In *Proceedings of the ninth biennial acm symposium on eye tracking research & applications*, pages 275–278, 2016.

[15] Subarna Tripathi and Brian Guenter. A statistical approach to continuous self-calibrating eye gaze tracking for head-mounted virtual reality systems. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 862–870. IEEE, 2017.

[16] Thiago Santini, Diederick C Niehorster, and Enkelejda Kasneci. Get a grip: slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking. In *Proceedings of the 11th ACM symposium on eye tracking research & applications*, pages 1–10, 2019.

[17] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):1915–1926, 2011.

[18] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[19] Kathryn Koehler, Fei Guo, Sheng Zhang, and Miguel P Eckstein. What do saliency models predict? *Journal of vision*, 14(3):14–14, 2014.

[20] Hadi Hadizadeh and Ivan V Bajić. Saliency-aware video compression. *IEEE Transactions on Image Processing*, 23(1):19–33, 2013.

[21] Fang Guo, Jianbing Shen, and Xuelong Li. Learning to detect stereo saliency. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014.

[22] Selim S Hacisalihzade, Lawrence W Stark, and John S Allen. Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *IEEE Transactions on systems, man, and cybernetics*, 22(3):474–481, 1992.

[23] Alexis Gabadinho, Gilbert Ritschard, Nicolas Séverin Mueller, and Matthias Studer. Analyzing and visualizing state sequences in r with traminer. *Journal of Statistical Software*, 40(4):1–37, 2011.

[24] Risheng Liu, Junjie Cao, Zhouchen Lin, and Shiguang Shan. Adaptive partial differential equation learning for visual saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3866–3873, 2014.

[25] Ali Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *2012 ieee conference on computer vision and pattern recognition*, pages 438–445. IEEE, 2012.

[26] Vidhya Navalpakkam and Laurent Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2049–2056. IEEE, 2006.

[27] Guokang Zhu, Qi Wang, and Yuan Yuan. Tag-saliency: Combining bottom-up and top-down information for saliency detection. *Computer Vision and Image Understanding*, 118:40–49, 2014.

[28] Yin Yan, Li Zhaoping, and Wu Li. Bottom-up saliency and top-down learning in the primary visual cortex of monkeys. *Proceedings of the National Academy of Sciences*, 115(41):10499–10504, 2018.

[29] Lucia Melloni, Sara van Leeuwen, Arjen Alink, and Notger G Müller. Interaction between bottom-up saliency and top-down control: how saliency maps are created in the human brain. *Cerebral cortex*, 22(12):2943–2952, 2012.

[30] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018.

[31] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 598–606, 2016.

[32] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[33] Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun. Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1396–1405. IEEE, 2017.

[34] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, pages 199–207, 2015.

[35] Zhiming Hu, Sheng Li, and Meng Gai. Temporal continuity of visual attention for future gaze prediction in immersive virtual reality. *Virtual Reality & Intelligent Hardware*, 2(2):142–152, 2020.

[36] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[37] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014.

[38] Ali Borji, Dicky N Sihite, and Laurent Itti. Probabilistic learning of task-specific visual attention. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 470–477. IEEE, 2012.

[39] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[40] Yujie Li, Atsunori Kanemura, Hideki Asoh, Taiki Miyanishi, and Motoaki Kawanabe. A sparse coding framework for gaze prediction in egocentric video. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1313–1317. IEEE, 2018.

[41] Kentaro Yamada, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki. Attention prediction in egocentric video using motion and visual saliency. In *Pacific-Rim Symposium on Image and Video Technology*, pages 277–288. Springer, 2011.

[42] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, pages 730–734. IEEE, 2015.

[43] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze prediction in dynamic 360 immersive videos. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5333–5342, 2018.

[44] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[45] Erwan J David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. A dataset of head and eye movements for 360 videos. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 432–437, 2018.

[46] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016.

[47] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European conference on computer vision*, pages 1–16. Springer, 2014.

[48] Sonali Agarwal, Narinder Singh Punn, Sanjay Kumar Sonbhadra, P Nagabhushan, KK Pandian, and Praveer Saxena. Unleashing the power of disruptive and emerging technologies amid covid 2019: A detailed review. *arXiv preprint arXiv:2005.11507*, 2020.

[49] Arno Hartholt, Ed Fast, Adam Reilly, Wendy Whitcup, Matt Liewer, and Sharon Mozgai. Ubiquitous virtual humans: A multi-platform framework for embodied ai agents in xr. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 308–3084. IEEE, 2019.

[50] Florian Berton, Anne-Hélène Olivier, Julien Bruneau, Ludovic Hoyet, and Julien Pettré. Studying gaze behaviour during collision avoidance with a virtual walker: Influence of the virtual reality setup. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 717–725. IEEE, 2019.

[51] Laurentius Antonius Meerhoff, Julien Bruneau, A Vu, A-H Olivier, and Julien Pettré. Guided by gaze: Prioritization strategy when navigating through a virtual crowd can be assessed through gaze activity. *Acta psychologica*, 190:248–257, 2018.

[52] Diako Mardanbegi, Benedikt Mayer, Ken Pfeuffer, Shahram Jalaliniya, Hans Gellersen, and Alexander Perzl. Eyeseethrough: Unifying tool selection and application in virtual environments. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 474–483. IEEE, 2019.

[53] Md Musfequs Salehin and Manoranjan Paul. A novel framework for video summarization based on smooth pursuit information from eye tracker data. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 692–697. IEEE, 2017.