Delay Compensation for a Telepresence System With 3D 360 Degree Vision Based on Deep Head Motion Prediction and Dynamic FoV Adaptation

Tamay Aykut, *Student Member, IEEE*, Mojtaba Karimi, Christoph Burgmair, Andreas Finkenzeller, Christoph Bachhuber, and Eckehard Steinbach, *Fellow, IEEE*

***Abstract*—The usability of telepresence applications is strongly affected by the communication delay between the user and the remote system. Special attention needs to be paid in case the distant scene is experienced by means of a Head Mounted Display. A high motion-to-photon latency, which describes the time needed to fully reflect the user's motion on the display, results in a poor feeling of presence. Further consequences involve unbearable motion sickness, indisposition, and termination of the telepresence session in the worst case. In this letter, we present our low-cost MAVI telepresence system, which is equipped with a stereoscopic 360° vision system and high-payload manipulation capabilities. Special emphasis is placed on the stereoscopic vision system and its delay compensation. More specifically, we propose velocity-based dynamic field-of-view adaptation techniques to decrease the emergence of simulator sickness and to improve the achievable level of delay compensation. The proposed delay compensation approach relies on deep learning to predict the prospective head motion. We use our previously described head motion dataset for training, validation, and testing. To prove the general validity of our approach, we perform cross validation with another independent dataset. We use both qualitative measures and subjective experiments for evaluation. Our results show that the proposed approach is able to achieve mean compensation rates of around 99.9% for latencies between 0.1 and 0.5 s.**

***Index Terms*—3D vision, telepresence, virtual reality, remote reality, omnidirectional vision.**

I. Introduction

Telepresence can be defined as the mediated perception of a temporally or spatially distant real environment [1] and is a timeless and still open research field. Its applications range from remote healthcare, such as telesurgery or telerehabilitation, over rescue and exploration missions in hazardous environments to applications in the private sector. A typical telepresence system such as the one depicted in Fig. 1 comprises a master (i.e. the user or operator) and a client system (here a mobile platform), which exchange video and audio signals over a communication network. In case the platform is equipped with manipulation capabilities, additionally haptic information is transmitted and the telepresence system becomes a teleoperation system. Several studies revealed a strong correlation between the feeling of presence and the availability and quality of visual, auditory, and haptic feedback [2]–[4]. Particularly important for an immersive experience is the visual feedback. A higher quality visual representation of the remote scene results in a better sense of presence [4], [5].

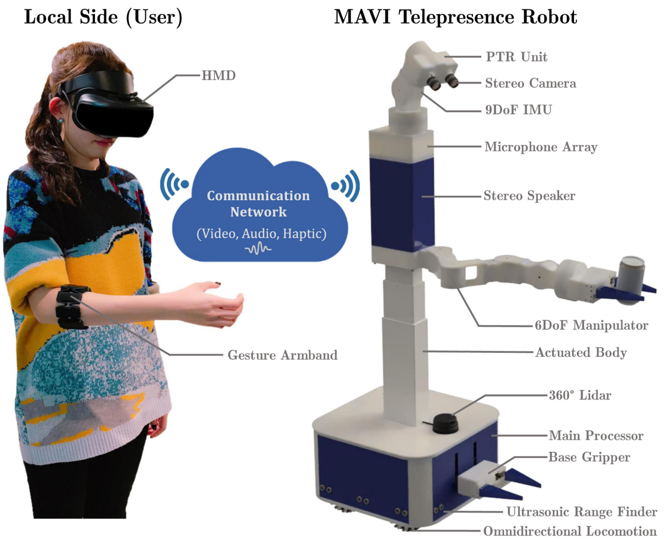

Fig. 1. Studied telepresence scenario, where the user controls our MAVI robot in a remote environment.

The bidirectional transmission of data imposes strong demands on the communication network, as it is part of a global control loop between the user and the teleoperator. The end-to-end delay between the operator and the remote side has a critical impact on the stability of the system and strongly affects the quality of experience (QoE). While delayed audio feedback may diminish the feeling of presence, a delayed visual response, especially when wearing an HMD, strongly deteriorates the QoE and provokes motion sickness. The time needed to completely reflect the user's motion and display the corresponding view on the screen is referred to as the motion-to-photon (M2P) latency. If the M2P exceeds a threshold of around 20 ms [6], this will

Manuscript received February 24, 2018; accepted July 1, 2018. Date of publication August 8, 2018; date of current version August 31, 2018. This letter was recommended for publication by Associate Editor C. Secchi and Editor Y. Yokokohji upon evaluation of the reviewers' comments. *(Corresponding author: Tamay Aykut.)*

The authors are with the Department of Electrical and Computer Engineering, Technical University of Munich, 80333, Munich, Germany (e-mail: tamay.aykut@tum.de; mojtaba.karimi@tum.de; christoph.burgmair@tum.de; andreas.finkenzeller@tum.de; christoph.bachhuber@tum.de; Eckehard.Steinbach@tum.de).

Digital Object Identifier 10.1109/LRA.2018.2864359

lead to motion sickness and the user will experience visual discomfort. To mitigate the effect of motion sickness, the user's ego-motion must comply with the expectations and the sensory inputs [7]–[10]. Sophisticated solutions are, hence, required to counteract these negative impacts.

In this letter, we present an advanced binocular omnidirectional vision system for our MAVI telepresence robot [11] to enable a three-dimensional delay-free perception of a distant environment. To this end, we equipped the MAVI platform with an actuated stereo-camera system with three degrees-of-freedom (DoF) to mimic the user's head motion at the client side. The vision system provides a stereoscopic visual representation of the distant scene, i.e. separate imagery from different vantage points for each eye, to enable the sensation of depth. Even though this system is able to cover viewing directions of 360°, in its raw setup it suffers from the lag between head motion and the display of the corresponding content. In [12], [13], we introduced our concept of a delay compensating vision system (DCVS). Applying this approach allows us to compensate the perceived latency of the user. The resulting reduction of the M2P latency prevents the user from experiencing visual discomfort. The DCVS deploys fisheye cameras to capture a larger FoV than actually displayed on the HMD to the user. We thereby create a cache zone around the displayed content and leverage the remaining imagery for local compensation until the updated frame arrives. The compensation rate is introduced as a metric to describe the achievable level of compensation. The amount of compensation is subject to the current head rotation speed, the present delay, and the available buffer. In case of a small buffer and large latency, fast head rotations and sudden direction changes cannot be fully compensated.

In this work, we extend the DCVS concept by incorporating a velocity-based FoV adaptation. Depending on the current angular head motion velocity, we temporarily decrease the displayed FoV of the user. In doing so, we are able to momentarily enlarge the cache area, which results in a higher level of compensation for fast rotations. This approach is motivated by the characteristics of the human eye. The macula is a part of the eye's retina that provides high-acuity vision. Only a minor portion of our vision is in high resolution. Around 90% of our vision lies outside of the central gaze and is denoted as the peripheral vision. The discrimination of (fine) details, color, and shape is limited for the peripheral vision. Assuming a non-stationary situation, e.g., during head rotation, the perception and differentiation becomes even worse both for the central and peripheral vision. Given these facts, we claim that a temporarily reduction of the FoV during rapid head motions does not influence the feeling of presence. Instead, it has positive implications on the achievable level of compensation and supports the reduction of simulator sickness. We use qualitative measures and subjective experiments to prove our hypothesis.

To further optimize the degree of compensation, we propose a deep learning-based head motion prediction (HMP) approach. We designed and investigated various deep architectures and compare them to state-of-the-art head motion predictors. Real head motion datasets are used to evaluate our deep head motion prediction approach. The key contributions can be summerized as follows:

- We extend our previous delay compensation approach in [13] by a velocity-based dynamic FoV adaptation.
- We further apply a deep learning-based head motion prediction approach and compare our method to state-of-the-art prediction techniques.
- We deploy both qualitative and subjective measures to verify our approach.

II. Related Work

Stereoscopic 360 ° Remote Vision: Acquiring and sending the entire 360° visual representation of the remote scene provides the user with the capability to casually turn the head with immediate visual feedback and thereby to keep the M2P latency low. While capturing monoscopic full panorama videos is considered state-of-the-art, the real-time acquisition and processing of the complete 360° scene in 3D, i.e., stereoscopically, still poses major challenges (e.g. [12], [13]). There are typically three main approaches to acquire stereoscopic panoramas: catadioptric systems [14], sequential acquisition [15], or the application of multiple cameras [16]. Available systems are usually characterized by a limited stereoscopic budget and provide only a constrained 3D impression. Moreover, they need costly equipment as most approaches rely on stitching of multiple high-resolution image contents and are, hence, computationally complex. This is why the 3D 360° footage is often created in a post-processing step and may take from several minutes to hours. These systems need precise calibration and are deemed to be sensitive to external influences. All of these approaches are subject to a stitching process. Depending on the available features in the scene, distortions are mostly unavoidable. Erroneous image content on HMDs is acutely critical, as the footage is shown to the eye from a distance as small as a few centimeters. Artifacts and stitching errors are magnified and result in visual discomfort. We exploit the benefits of an actuated stereo camera system in combination with a delay compensation approach and provide vision on demand. Such a system is lean, low-cost, provisions a large stereoscopic budget, and is real-time capable. Combined with a delay compensating approach, it allows 3D visual impressions in any viewing direction, while keeping the M2P latency low. Leveraging the residual imagery from the buffer allows the visual and vestibular sensory information to be aligned.

Adaptive FoV: Prior art revealed a strong correlation between the perceived level of presence and simulator sickness [17], [18]. The right choice of the FoV is thereby a unique etiological factor. Wider FoVs improve the feeling of presence [19], but also stimulate larger portions of the peripheral vision, which is especially sensitive to stimuli induced by fast motion, fostering the onset of motion sickness [20]. Reducing the FoV, on the contrary, proved to lower the effect of motion sickness, but simultaneously deteriorates the degree of presence [21]. In this letter, we aim to keep the realism high, while reducing visually induced motion sickness. Therefore, we introduce a velocity-based head motion FoV adaptation. We investigate circular and asynchronous rectangular FoV restriction techniques. Although changing the FoV has been examined before, previous studies either restricted the FoV throughout the whole session or modified it in discrete steps [22]. The most similar research

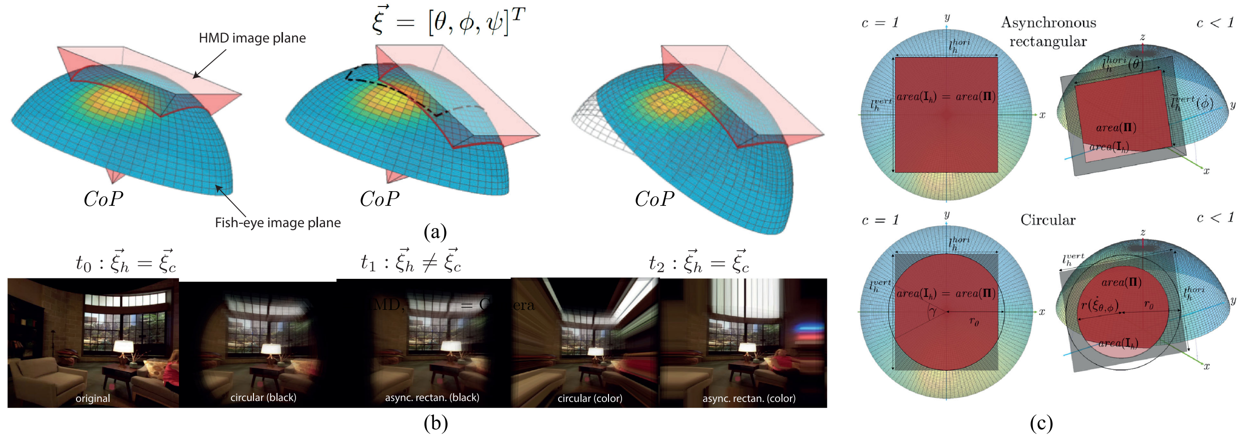

Fig. 2. a) Concept of the delay compensation approach (DCVS) with equidistant fisheye cameras. The blue hemisphere corresponds to the image plane of the equidistant fish-eye camera. In red, we visualize the image plane of the user that is eventually displayed to the HMD. By capturing more imagery than actually showing to the user, we are able to leverage the remaining image content for local delay compensation and instantaneous visual response. b) FoV adaptation techniques with high $\mathcal{R}_{\max} = 0.5$ for visualization purpose and c) Algebraic description of the dynamic FoV adaptation approaches.

investigating adaptive FoVs is reported in [22], [23]. Fernandes *et al.* modified the FoV depending on the translation speed of joy pad controllers [22], which is helpful for screen-based applications. In our work, we leverage the filtered internal orientation sensors of the HMD to modify the FoV more intuitively and naturally. For that, we propose two velocity-based circular and asynchronous rectangular modification strategies. Kala *et al.*, instead, utilized the visual content to change the FoV [23]. Their approach is applicable for virtual environments, where simulator sickness might emerge through vection. In contrast, we consider a real telepresence scenario, where the occurrence of vection is less probable. Our approach targets to lessen visually induced motion sickness, triggered by the temporal lag between ego-motion and visual response.

Head motion prediction: HMP is a research field that has started already in the early 90's. At that time, researchers aimed to compensate the local M2P latency. Nowadays, HMP is utilized for 360° (tile-based) streaming applications to predict the future head position after the communication delay. Many prior approaches rely on past traces and apply averaging or (Weighted) Linear Regression ((W)LR) techniques to predict the future head orientation [24]. Another frequently applied method is to first compute the optimal state estimate of the current head orientation (i.e. orientation, angular velocity and acceleration) by means of filters (e.g., Kalman Filter (KF), Particle Filter, etc.) and then deploy polynominal extrapolation with respect to the adopted motion model [25]. Although linear fitting approaches profit from being simple and accurate for smooth and homogeneous motions, they are sensitive to abrupt orientation changes and, hence, tend to dramatically overshoot. KF-based extrapolation are deemed to be robust against fast fluctuations, but suffer from susceptibility to noise sensory data. Beside its computational complexity, KF-based approaches highly depend on the validity of the motion model. In this letter, however, we propose and investigate various deep neural network architectures to exploit their benefits of being robust against fast fluctuations, resistant to noise, and perform well for homogeneous motions.

III. Dynamic FoV Adaptation

In [12], [13], we introduced the theory behind the delay compensating vision system. Its conceptional modes of operation are illustrated in Fig. 2a. We use equidistant fish-eye cameras to capture a larger FoV than displayed to the user. By using only a subset of the cameras' FoV, we are able to leverage the residual imagery for local delay compensation until the updated frame acquired at the present head position arrives. We extend this concept in this letter by velocity-based FoV adaptation. Adjusting the visual field as a function of the actual head rotation speed aims to decrease the onset of motion sickness during rapid head movements. In this way, we are able to optimize the level of compensation, especially for quick direction changes and fast head motions. For that, we propose a circular and asynchronous rectangular FoV modification approach, both being a function of the directed head motion velocity. The constraint portions are either filled with black pixels or with artificial colored pixels, which are generated by making a round-analysis and taking the outermost available color value (see Fig. 2b). A subsequent fading layer is superimposed to smooth the transition. The resulting compensation rate $\widetilde{c}_{\text{rpy}}$, describing the achievable level of delay compensation, can be reformulated to

$$c_{\text{rpy}} = \frac{area(\mathbf{\Pi})}{l_h^{\text{hori}} \cdot l_h^{\text{vert}}} \Rightarrow \widetilde{c}_{\text{rpy}} = \frac{area(\mathbf{\Pi})}{area(\mathbf{I}_h)}, \qquad (1)$$

where $\mathbf{\Pi}$ represents the set of all image points $\vec{\pi}_i \in \mathbf{\Pi}$ that are available to be displayed. $l_h^{\text{hori}} = 2 \cdot f \cdot \tan(\frac{fov_h^{\text{hori}}}{2})$ and $l_h^{\text{vert}} = 2 \cdot f \cdot \tan(\frac{fov_h^{\text{vert}}}{2})$ depend on the displayed horizontal and vertical FoV $(fov_h^{\text{hori}}, fov_h^{\text{vert}})$ and refer to the width and height of the original image plane of the HMD. Note, that the subscript h indicates any variable that is related to the head/HMD motion. The subscript c, instead, refers to the quantities of the camera system.

As we temporarily restrict this image plane, we use $area(\mathbf{I}_h)$ to denote the area of the HMD's image plane that is to be considered for the present head velocity. We utilize the auxiliary measure $\mathbb{H}$ to define the permitted height of an arbitrary image point $\vec{p_i} \in \mathbf{I}_h$ to be in $\mathbf{\Pi}$ [13]

$$\mathbb{H}(\vec{p_i}, fov_c^{\text{hori}}) = |\vec{p_i}| \cdot \cos(fov_c^{\text{hori}}/2). \tag{2}$$

An image point $\vec{p_i}$ is deemed to be in the set $\mathbf{\Pi}$ if the following condition is met

$$\mathbf{\Pi} \cup \{\vec{p_i}\} \; \forall \vec{p_i} \in \mathbf{I}_h \iff p_{z,i} \geq \mathbb{H}\left(\vec{p_i}, fov_c^{\text{hori}}\right). \tag{3}$$

During quick head rotations the amount of image points $\vec{p_i}$ that need to be taken into account gets confined. We denote λ_{th} [rad/s] as the threshold rotation velocity that needs to be exceeded to initiate the FoV adaptation and λ_{max} [rad/s] as the maximum rotation speed after which the maximum restriction $\mathcal{R}_{\text{max}} \in (0, 1)$ is reached. The degree of restriction as a function of the current head velocity λ can be expressed as

$$\mathcal{R}(\lambda) = \mathcal{R}_{\text{max}} \cdot \min\left\{1, \max\left\{0, f_{\{\text{exp, lin}\}}(\lambda)\right\}\right\}, \tag{4}$$

where $f_{\{\text{exp, lin}\}}(\lambda)$ represents a specific adaptation behavior. We propose linear and exponential adaptation characteristics

$$f_{\text{lin}}(\lambda) = (\lambda - \lambda_{\text{th}})/(\lambda_{\text{max}} - \lambda_{\text{th}}), \text{or}$$
$$f_{\text{exp}}(\lambda) = (\mathrm{e}^{\lambda^2} - \mathrm{e}^{\lambda_{\text{th}}^2})/(\mathrm{e}^{\lambda_{\text{max}}^2} - \mathrm{e}^{\lambda_{\text{th}}^2}). \tag{5}$$

Asynchronous Rectangular Restriction: For the asynchronous rectangular restriction, we regard the pan $\dot{\theta}$ and tilt $\dot{\phi}$ rotation speed independently. The roll rotation is neglected for the FoV adaptation. The area of the HMD's adaptive image plane can therefore be expressed as $area(\mathbf{I}_h) = \widetilde{l}_h^{\text{hori}}(\dot{\theta}) \cdot \widetilde{l}_h^{\text{vert}}(\dot{\phi})$ with a dynamic length and width

$$\widetilde{l}_h^{\text{hori}}(\dot{\theta}) = (1 - \mathcal{R}(\dot{\theta})) \cdot l_h^{\text{hori}}$$
$$\widetilde{l}_h^{\text{vert}}(\dot{\phi}) = (1 - \mathcal{R}(\dot{\phi})) \cdot l_h^{\text{vert}}. \tag{6}$$

The modified set of image points $\widetilde{\vec{p_i}}$ that is utilized to calculate the set of available imagery $\mathbf{\Pi}$ results to

$$\widetilde{\vec{p_i}} = \mathbf{R} \cdot \begin{bmatrix} \pm\widetilde{l}_h^{\text{hori}}(\dot{\theta})/2 & \pm\widetilde{l}_h^{\text{vert}}(\dot{\phi})/2 & f \end{bmatrix}^T, \tag{7}$$

with $\mathbf{R}$ being the rotation matrix that describes the current rotation $\vec{\xi_t} = [\theta_t \; \phi_t \; \psi_t]^T$ for pan, tilt, and roll. *Circular Restriction:* For the circular restriction, we consider the velocity of pan and tilt jointly $\dot{\xi}_{\theta,\phi} = \sqrt{\dot{\theta}^2 + \dot{\phi}^2}$. The area of the HMD's image plane thereby amounts to $area(\mathbf{I}_h) = r(\dot{\xi}_{\theta,\phi})^2 \cdot (\pi - \gamma + \sin(\gamma))$, where γ is the angle of the excluded circular segments (see Fig. 2c). $r(\dot{\xi}_{\theta,\phi})$ alters with respect to the current joint head motion velocity

$$r(\dot{\xi}_{\theta,\phi}) = (1 - \mathcal{R}(\dot{\xi}_{\theta,\phi})) \cdot r_0, \tag{8}$$

where $r_0 = fov_h^{\text{vert}}/2$ corresponds to the original size of the circle without any adaptation. $area(\mathbf{\Pi})$ can be computed by considering the constraint set of available image points after a round analysis $\forall \alpha \in [0, 2\pi)$

$$\widetilde{\vec{p_i}} = \mathbf{R} \cdot \begin{bmatrix} r(\dot{\xi}_{\theta,\phi}) \cdot \cos(\alpha) & r(\dot{\xi}_{\theta,\phi}) \cdot \sin(\alpha) & f \end{bmatrix}, \text{with}$$
$$|r(\dot{\xi}_{\theta,\phi}) \cdot \cos(\alpha)| \leq l_h^{\text{hori}}/2 \wedge |r(\dot{\xi}_{\theta,\phi}) \cdot \sin(\alpha)| \leq l_h^{\text{vert}}/2, \tag{9}$$

For a dynamic FoV adaptation the considered image points $\vec{p_i}$ are amended to $\widetilde{\vec{p_i}}$. Fig. 2c illustrates the constraint set of image points that are taken into account. The available region of image content $area(\mathbf{\Pi})$ that can be used for visualization can then be computed in analogy to [13].

The additional benefit that is gained by means of the dynamic FoV adaptation is subject to the appropriate value selection for λ_{th}, λ_{max}, and $\mathcal{R}_{\text{max}}$. They both influence the level of compensation and the degree of visual comfort. Depending on the task at hand or the target application, the focus can be set accordingly. As our goal was to simultaneously provide a high level of compensation and visual comfort, we conducted a pilot pre-study to find suitable trade-off parameters. For that group of participants, we found out that a circular restriction technique with $\lambda_{\text{th}} = 1$ rad/s, $\lambda_{\text{max}} = 2$ rad/s, and $\mathcal{R}_{\text{max}} = 0.2$–$0.4$ yielded the best results. Consequently, we used this set of parameters for our experimental validation.

IV. Deep Head Motion Prediction

To the best of our knowledge, this is the first work that applies deep learning for head motion prediction. For that reason, we investigated numerous architectures ranging from simple dense feed forward networks (FFN), over deep recurrent neural networks such as the well established Long Short-Term Memory (LSTM) to more sophisticated joint structures (see Fig. 3). We use real head motion datasets for training and testing, where participants (P) watched 360° videos wearing an HMD. Our dataset [12], [13], which we will refer to as the LMT dataset in the following, contains the head movements of 30 participants aged between 22–40 watching three different 360° videos with an average video length of 120 s and varying dynamics in the respective scene. The filtered orientation data from the IMU was acquired at a frequency of 80 Hz (i.e., $\Delta t = 12.5$ ms). We separated the LMT dataset into a training set (81 profiles, P = 27), a validation set (3 profiles, P = 1), and a test set (6 profiles, P = 2). We further examine the general validity by cross-validating the deep head motion predictions on another, completely independent dataset [26] (called the IMT dataset in the following) with different participants and sensory information acquired under dissimilar conditions. The IMT dataset contains the head movement recordings of 58 subjects, each watching five 70 s long 360° video content.

We take the present pan (θ_t), tilt (ϕ_t), and roll (ψ_t) orientations as well as past values thereof within a certain time window (W) as input for the network. W = 0 ms would imply to consider only the current orientations. We found out empirically that considering the 20 last values for each orientation direction (W = 250 ms) delivers the best results. And rather than taking the absolute orientation values into account, we initially subdivide the inputs into their respective orientation groups and compute their normalized differences

$$\hat{\vec{\xi}}_{\text{diff, t}} = \frac{\vec{\xi}_{\text{diff, t}}}{\max(|\vec{\xi}_{\text{diff}}|)}, \text{with } \vec{\xi}_{\text{diff, t}} = \vec{\xi}_{\text{t}} - \vec{\xi}_{\text{t}-\Delta t}. \tag{10}$$

This step is essential to ensure a generalization for other datasets. Omitting this step leads to erroneous inferences. The subsequent convolutional layer with a kernel size of ten acts as low pass filter for the noisy differences. The mean absolute error (MAE) was deployed as loss function and showed superior

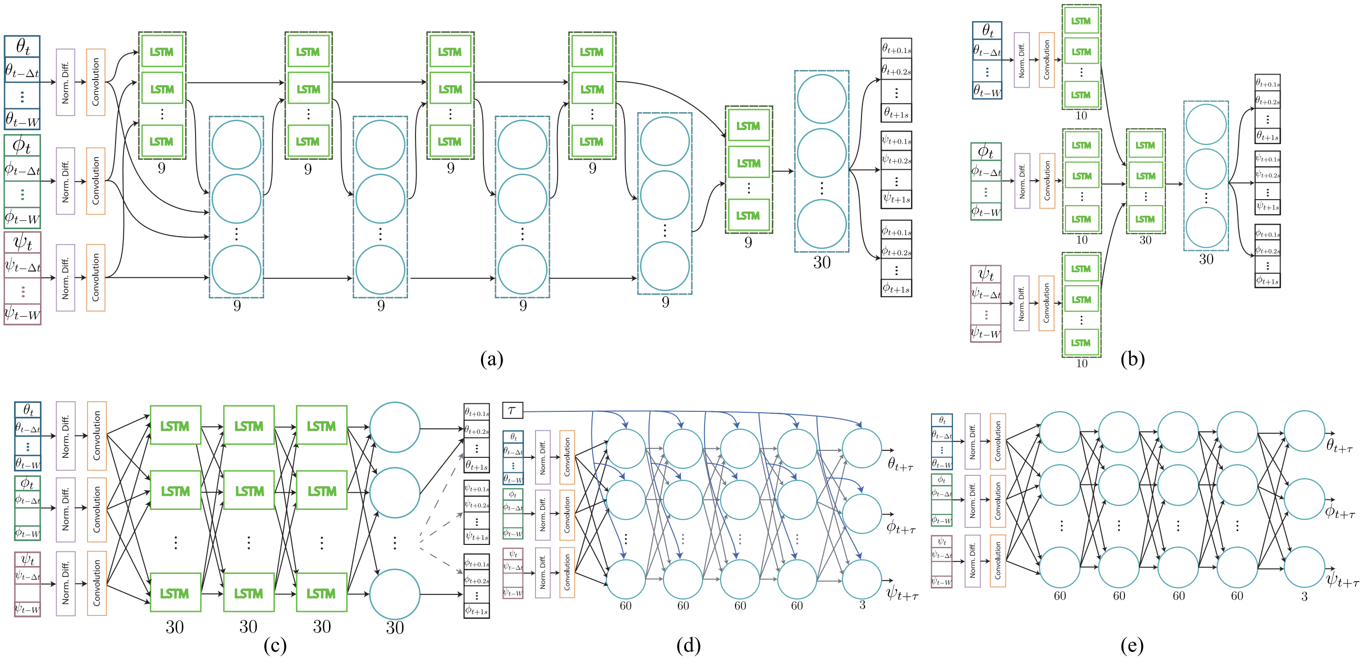

Fig. 3. Investigated neural network architectures: a) FFN + LSTM (interleaved), b) LSTM (subdivided inputs), c) LSTM (dense), d) FFN + Delay Shortcuts, e) FFN (dense).

performance compared to the mean squared error (MSE). Except for dense FFN we predicted a whole sequence of future orientations ($\vec{\xi}^{\text{pred}}_{t+n\cdot 100\text{ ms}}\ \forall n \in \mathbb{Z}, n = 1, \ldots, 10$). As we actually predict the future normalized differences $\hat{\vec{\xi}}^{\text{pred}}_{\text{diff, t}}$, we remap to the absolute orientation values as follows

$$\vec{\xi}^{\text{pred}}_{\text{t}} = \vec{\xi}_{\text{t}} + \sum_{\text{k}=\text{t}-\tau/\Delta t}^{\text{t}} \hat{\vec{\xi}}^{\text{pred}}_{\text{diff, t}} \cdot \max(|\vec{\xi}_{\text{diff}}|). \tag{11}$$

Our initial approach was to examine simple dense FFNs (Fig. 3e), although typically not applied for time series forecasting, to experience the prediction behavior for head motions. As the input of orientation data is equal for all future values the network had difficulties to learn for different delays. Thus, we extended the FFN by another input, being the present delay, and injected it as shortcuts to all other nodes (Fig. 3d). A more intuitive way, however, is to apply deep recurrent neural networks as they are characterized by a feedback loop and thereby establish a way of memory and share weights over time. LSTMs are the most famous and effective representatives and provide a solution for the vanishing gradient problem. Fig. 3b and c illustrate two different LSTM-based architectures we investigated. In Fig. 3a, we designed an interleaved architecture of LSTMs and dense FFN. The objective was to increase the number of weights for an improved learning behavior and still maintain the memorization characteristics by the LSTM layers.

We use the Adam optimization algorithm as an extension to stochastic gradient descent, combining both advantages of RMSProp and AdaGrad. The maximum number of epochs is set to 1000. We applied the early stopping technique (patience = 2, min. delta = 0) to avoid overfitting. Besides that, a learning rate decay scheduler was used to decrease the initial learning rate of 0.001 every 30 epochs by 70%. The batch size was set to 2^{11}. We further used the Rectified Linear Unit (ReLU) as activiation function for the FFNs. The deep learning approaches were implemented in Keras and Tensorflow. The time for inference of future orientation data was in the domain of single-digit microseconds (Core i7 (x64)) and, hence, suitable for real-time applications.

V. Experimental Setup

We investigated our approach on our MAVI telepresence platform [11]. This semi-autonomous low-cost platform is characterized by a 4 mecanum-wheel based omnidirectional locomotion system, high manipulation capabilities at a budget price, and a 3D 360° vision system. We used self-designed light-weight 3D printed parts to construct the platform. Fig. 4 illustrates the robot's main components and the dimensions thereof. In contrast to most state-of-the-art monoscopic (360°) vision systems, we provide the user with the 3D impression in every viewing direction at real-time. We built a binocular camera system mounted on an actuated 3 DoF Pan-Tilt-Roll (PTR) Unit to mirror the user's head motion. Together with the presented delay compensation approach, we aim to provision a delay-free perception. The PTR-Unit follows the yaw, pitch, and roll (ZYX) euler convention. Any desired head orientation triggered by the user is converted to quaternions. The correct conversion of a head orientation $q = [q_x, q_y, q_z, q_w]^T$ to the angles θ_c, ϕ_c, and ψ_c subject to the PTR-Unit design is obtained with

$$\begin{aligned}\theta_c &= \text{atan2}(-2(q_{\text{x}}q_{\text{y}} - q_{\text{w}}q_{\text{z}}), q_{\text{w}}^2 + q_{\text{x}}^2 - q_{\text{y}}^2 - q_{\text{z}}^2))\\ \phi_c &= \text{asin}(2(q_{\text{x}}q_{\text{z}} + q_{\text{w}}q_{\text{y}}))\\ \psi_c &= \text{atan2}(-2(q_{\text{y}}q_{\text{z}} - q_{\text{w}}q_{\text{x}}), q_{\text{w}}^2 - q_{\text{x}}^2 - q_{\text{y}}^2 + q_{\text{z}}^2).\end{aligned} \tag{12}$$

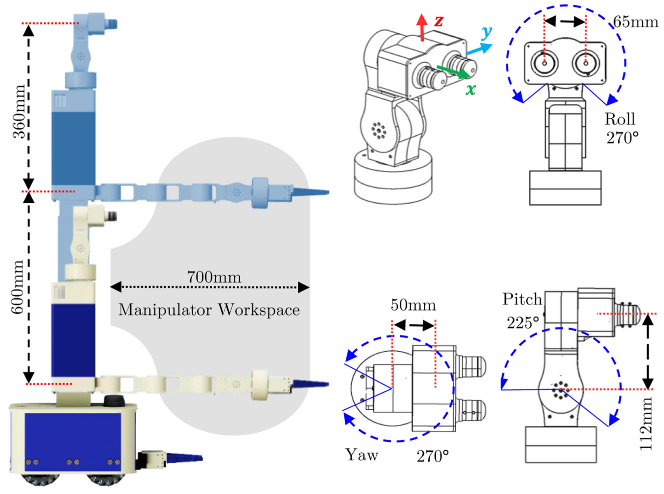

Fig. 4. Design of our MAVI platform and its actuated stereoscopic vision system.

VI. Experiments

Qualitative Measures: For both, the dynamic FoV adaptation as well as for the deep learning-based head motion prediction, we exploited the LMT and IMT dataset. Qualitative measures used to evaluate the performance are the mean error, the root mean square error (RMSE) and the mean compensation rate $\bar{c}_{rpy}$. The mean compensation rate averages the mean for all videos V, participants P, and available samples $S = \frac{t_{\text{end}}^{\text{V, P}} - t_{\text{start}}^{\text{V, P}}}{f_{\text{IMU}}}$ for a present delay τ [13]. One drawback of this metric is that it averages over all samples; even for situations where no motion happens. That is why we propose a more meaningful version to which we will refer as the event-based mean compensation rate $\hat{\bar{c}}_{rpy}$

$$\hat{\bar{c}}_{rpy} = \frac{1}{V \cdot P \cdot \sum_{k=1}^{S} \delta(a_k)} \cdot \sum_{i=1}^{P} \sum_{j=1}^{V} \sum_{k=1}^{S} \delta(a_k) \cdot c_{rpy}^{i,j,k}, \quad (13)$$

where the dirac delta function $\delta(a_k)$ returns one when its argument (a_k) becomes zero

$$a_k = \begin{cases} 0, & \text{if } \{|\dot{\theta}_k| \text{ or } |\dot{\phi}_k| \text{ or } |\dot{\psi}_k|\} \geq 0 + \epsilon \\ \neq 0, & \text{otherwise.} \end{cases} \quad (14)$$

In this way, we accumulate only those compensation rates where at least one of the head rotation velocities is above the IMU's sensor noise ($\epsilon = 5°$).

For the stereo camera system on our MAVI telepresence robot, we used equidistant fisheye cameras ($fov_c^{\text{hori}} = fov_c^{\text{vert}} = 150°$) and investigated different buffer sizes for latencies $\tau \in [0.1 - 1s]$. The cache area was determined by considering the available diagonal buffer [13]

$$b^{\text{diag}} = \tfrac{1}{2}(fov_c^{\text{diag}} - fov_h^{\text{diag}}) \quad (15)$$

with $fov_c^{\text{diag}} (= fov_c^{\text{vert}} = fov_c^{\text{hori}})$ and fov_h^{diag} describing the camera's and HMD's diagonal FoV, respectively [13]

$$fov_h^{\text{diag}} = 2 \cdot \operatorname{atan}\left(\sqrt{\tan^2\left(\frac{fov_h^{\text{hori}}}{2}\right) + \tan^2\left(\frac{fov_h^{\text{vert}}}{2}\right)}\right). \quad (16)$$

Subjective Experiments: We conducted subjective experiments to assess our approach with respect to the degree of presence, simulator sickness, and the overall opinion. After removing outliers, we had N = 13 healthy subjects with normal or corrected-to-normal visual acuity, who participated in the study (female = 31%, male = 69%, average age = 25.38, SD = 5.18). For a meaningful and reproducible comparison, we used an existing 3D 360° footage kindly provided by the Fraunhofer HHI [27] to conduct the subjective experiments. The subjects were exposed to the virtual scene six times for 45s. For each iteration we changed the visualization method according to Fig. 2b and Fig. 7. The subjects where asked to find artificially generated objects whose position and color was changed randomly. After each session the subjects filled a form that contained established questions to evaluate the level of presence, simulator sickness, and the overall opinion.

To measure the level of presence, we deployed the Igroup Presence Questionnaire (IPQ) [28]. The IPQ self-report questionnaire is very reliable (Cronbach's $\alpha = .87$) and consists in total of 14 items based on a seven-point Likert scale (0–6). It measures three different subscales of presence (Spatial Presence, Involvement, Experienced Realism) and one additional item (the general "sense of being there") [28].

To examine symptoms of simulator sickness, we applied the Simulator Sickness Questionnaire (SSQ) [29]. The SSQ is an established measure, contains 16 items rated on a four-point Likert scale (0–3), and conveys the symptoms of three major subscales (nausea, oculomotor, and disorientation). A total score determines the aggregated severity.

In a final step, we were interested in the general opinion of the subjects experiencing the FoV adaptation. To this end, we asked the participants to rate the FoV reduction compared to the reference FoV by means of the Mean Opinion Score (MOS). We used the Absolute Category Rating scale, mapping ratings between *Bad* and *Excellent* to numbers between 1 and 5. The final MOS value is a single rational number and is calculated as the arithmetic mean over all individual ratings.

VII. Discussion

In Fig. 5a and b, we illustrate the MAE and RMSE of the examined deep learning architectures. The associated event-based mean compensation rate is depicted in Fig. 6a. The results show that the dense FFN's exhibit only a slight improvement compared to applying no prediction at all. The LSTM-based architecture, however, leads to a significant improvement of the MAE and RMSE. This is also reflected in the event-based compensation rate. The best performance is achieved by the interleaved architecture of LSTM and dense FFN blocks. In Fig. 5c and d, and Fig. 6b, we contrast the best deep head motion predictor to state-of-the-art head motion prediction approaches. As there are various prior art methods that apply head motion prediction, we re-implemented the most widely used approaches like the Linear Regression approach and the polynominal extrapolation technique, which is premised on a constant acceleration

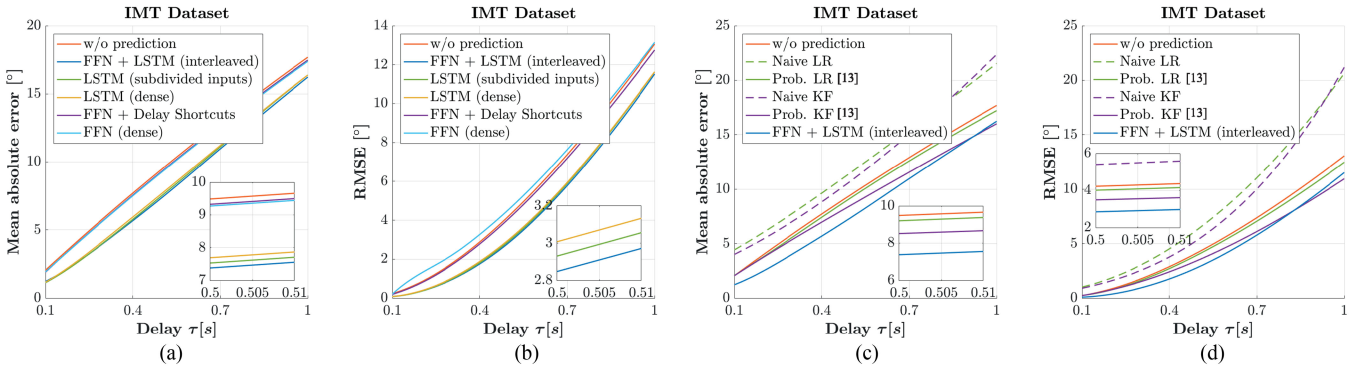

Fig. 5. Mean absolute error and RMSE shown for horizontal rotations and tested delays between 0.1–1 s. The interleaved FFN and LSTM architecture leads to a substantial improvement in mean absolute error and root mean squared error. a) and b) compare the investigated neural network architectures. c) and d) compare the FFN + LSTM (interleaved) predictor to state-of-the-art deterministic predictors.

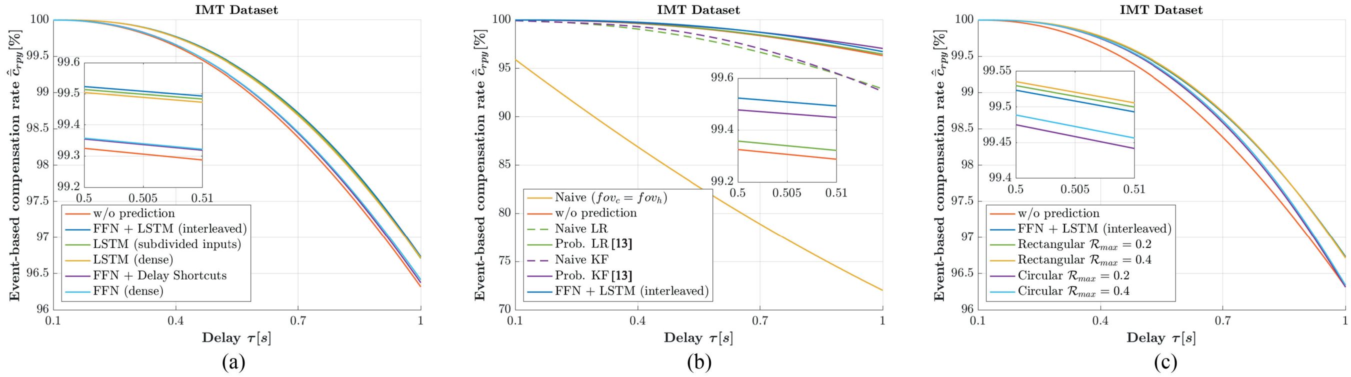

Fig. 6. Event-based mean compensation rate $\hat{\bar{c}}_{rpy}$ for tested delays between 0.1–1 s with $b_{diag} = 34°$. a) compares the investigated neural network architectures. The interleaved FFN and LSTM architecture leads to a substantial improvement. b) compares the FFN + LSTM (interleaved) predictor to state-of-the-art deterministic predictors. Compared to the naive approach, where no compensation is applied, the realization of the 3D compensation approach results in significantly more imagery that is available for visualization. c) shows the improvement of the dynamic FOV adaptation.

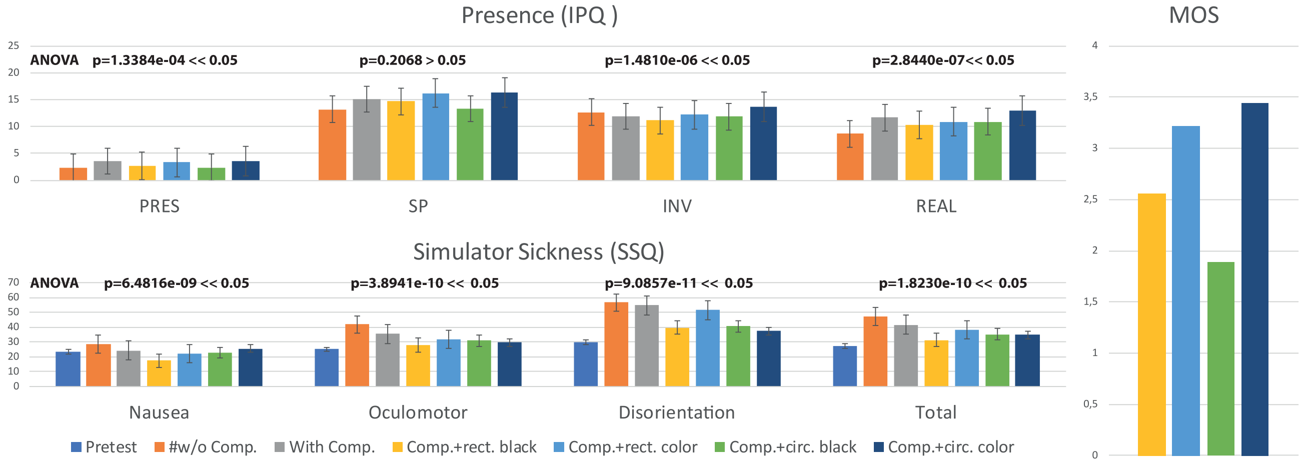

Fig. 7. Subjective experiments to assess the degree of presence, simulator sickness and the overall opinion (MOS) with ($\lambda_{\text{th}} = 1$ rad/s, $\lambda_{\text{max}} = 2$ rad/s). Compared to the naive approach, where no compensation is applied, the realization of the 3D compensation approach results in a significant improvement. Simulator sickness is reduced by the dynamic FOV adaptation while the degree of presence remains constant or in some cases even improves.

motion model and a Kalman Filter based optimal state estimate. We tag them as *naive LR* and *naive KF*. We further include our probabilistic modification approach proposed in [13] (*Prob. LR*, *Prob. KF*), which uses a probabilistic error model to weigh any forecasted orientation with its probability to be there. The results demonstrate a substantial improvement of the deep predictor for latencies in the range of 0.1–0.9 s. For delays higher than 0.9 s the deep approach achieves similar performance as *Prob. KF*. Identical behavior can be observed in the event-based compensation rate, where the deep interleaved network is superior compared to the other approaches. The deep neural network learns the physical limitations of the user and can approximate a better and more advanced motion model. In this way, it is robust against fast fluctuations but still performs well for homogeneous motions.

Fig. 6c shows that applying the dynamic FoV adaptation can further improve the achievable level of (event-based) compensation. The asynchronous rectangular adaptation technique outperforms the circular restriction technique. The results from the subjective experiments depicted in Fig. 7 clearly demonstrate that the compensation itself exhibits a superior performance compared to the naive version without any compensation. The results further indicate that the dynamic FoV adaptation is capable of reducing the effect of simulator sickness while maintaining the feeling of presence. Statistical significance was verified with the ANOVA test. It is interesting to see that a FoV restriction partially leads to an improved sense of presence. This is particularly visible for the colorized restriction approaches. This outcome is likewise confirmed by the MOS values.

VIII. Conclusions

In this letter, we present our low-cost MAVI telepresence robot and tackle the challenge of decreasing the perceived latency of a remote reality experience while providing an immersive 3D 360° visual impression. To this end, we propose two velocity-based FoV adaptations techniques with either black or colorized restriction strategies. Deep learning architectures are investigated for HMP and compared to state-of-the-art approaches. The results of the qualitative measures and subjective experiments verify that a temporary FoV constriction can help to reduce the emergence of simulator sickness and improve the level of achievable delay compensation.

The dynamic FoV adaptation combined with the deep learning-based HMP further improves the achievable degree of delay compensation. Nonetheless, greater scrutiny is required to find the optimal parameters. Additionally, we believe that the HMP by means of deep learning has high potential. That is why we plan to put further effort in improving the deep neural networks. We also consider to use the compensation rate itself as loss function to allow the network a certain error tolerance as long as the compensation rate is kept at its maximum.

References

[1] J. Steuer, “Defining virtual reality: Dimensions determining telepresence,” *J. Commun.*, vol. 42, no. 4, pp. 73–93, 1992.

[2] M. Slater and M. Usoh, “Presence in immersive virtual environments,” in *Proc. IEEE Annu. Int. Symp. Virtual Reality*, 1993, pp. 90–96.

[3] W. Barfield and S. Weghorst, “The sense of presence within virtual environments: A conceptual framework,” *Adv. Human Factors Ergonomics*, vol. 19, pp. 699–699, 1993.

[4] M. Slater and M. Usoh, “Representations systems, perceptual position, and presence in immersive virtual environments,” *Presence: Teleoperators Virtual Environ.*, vol. 2, no. 3, pp. 221–233, 1993.

[5] C. Hendrix and W. Barfield, “Presence within virtual environments as a function of visual display parameters,” *Presence: Teleoperators Virtual Environ.*, vol. 5, no. 3, pp. 274–289, 1996.

[6] S. M. LaValle, A. Yershova, M. Katsev, and M. Antonov, “Head tracking for the oculus rift,” in *Proc. IEEE Int. Conf. Robot Autom.*, 2014, pp. 187–194.

[7] J. T. Reason and J. J. Brand, *Motion Sickness*. New York, NY, USA: Academic, 1975.

[8] R. H. So, W. Lo, and A. T. Ho, “Effects of navigation speed on motion sickness caused by an immersive virtual environment,” *Human Factors*, vol. 43, no. 3, pp. 452–461, 2001.

[9] R. S. Allison, L. R. Harris, M. Jenkin, U. Jasiobedzka, and J. E. Zacher, “Tolerance of temporal delay in virtual environments,” in *Proc. IEEE Virtual Reality*, 2001, pp. 247–254.

[10] M. A. Watson and F. Black, “The human balance system: A complex coordination of central and peripheral systems,” Vestibular Disorders Assoc., Portland, OR, USA, Publ. no. S-7, 2008.

[11] M. Karimi, T. Aykut, and E. Steinbach, “MAVI: A research platform for telepresence and teleoperation,” *CoRR*, vol. abs/1805.09447, 2018, [Online]. Available: http://arxiv.org/abs/1805.09447

[12] T. Aykut, S. Lochbrunner, M. Karimi, B. Cizmeci, and E. Steinbach, “A stereoscopic vision system with delay compensation for 360 remote reality,” in *Proc. Thematic Workshops ACM Multimedia*, 2017, pp. 201–209.

[13] T. Aykut, C. Burgmair, M. Karimi, J. Xu, and E. Steinbach, “Delay compensation for actuated stereoscopic 360 degree telepresence systems with probabilistic head motion prediction,” in *Proc. IEEE Winter Conf. Appl. Comput. Vision*, 2018, pp. 2010–2018.

[14] R. Aggarwal, A. Vohra, and A. M. Namboodiri, “Panoramic stereo videos with a single camera,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 3755–3763.

[15] C. Richardt, Y. Pritch, H. Zimmer, and A. Sorkine-Hornung, “Megastereo: Constructing high-resolution stereo panoramas,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 1256–1263.

[16] Facebook surround 360, 2016. [Online]. Available: https://code.facebook.com/posts/1755691291326688/introducing-facebook-surround-360-an-open-high-quality-3d-360-video-capture-system/

[17] K. W. Arthur and F. P. Brooks, Jr., “Effects of field of view on performance with head-mounted displays,” Ph.D. dissertation, Dept. Comput. Sci., Univ. North Carolina at Chapel Hill, Chapel Hill, NC, USA, 2000.

[18] C. Jerome, R. Darnell, B. Oakley, and A. Pepe, “The effects of presence and time of exposure on simulator sickness,” *Proc. Human Factors Ergonomics Soc. Annu. Meeting*, vol. 49, no. 26, pp. 2258–2262, 2005.

[19] A. F. Seay, D. M. Krum, L. Hodges, and W. Ribarsky, “Simulator sickness and presence in a high field-of-view virtual environment,” in *Proc. Extended Abstr. Human Factors Comput. Syst.*, 2002, pp. 784–785.

[20] W. IJsselsteijn, H. d. Ridder, J. Freeman, S. E. Avons, and D. Bouwhuis, “Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence,” *Presence: Teleoperators Virtual Environ.*, vol. 10, no. 3, pp. 298–311, 2001.

[21] P. DiZio and J. R. Lackner, “Circumventing side effects of immersive virtual environments,” in *Proc. ACM Human Comput. Interact.*, 1997, pp. 893–896.

[22] A. S. Fernandes and S. K. Feiner, “Combating VR sickness through subtle dynamic field-of-view modification,” in *Proc. IEEE 3D User Interfaces*, 2016, pp. 201–210.

[23] N. Kala *et al.*, “P-218: An approach to reduce VR sickness by content based field of view processing,” *SID Symp. Dig. Tech. Papers*, vol. 48, no. 1, pp. 1645–1648, 2017.

[24] F. Qian, L. Ji, B. Han, and V. Gopalakrishnan, “Optimizing 360 video delivery over cellular networks,” in *Proc. 5th Workshop All Things Cellular: Operations, Appl. Challenges*, 2016, pp. 1–6.

[25] R. Azuma and G. Bishop, “A frequency-domain analysis of head-motion prediction,” in *Proc. 22nd Annu. Conf. Comput. Graph. Interactive Techn.*, 1995, pp. 401–408.

[26] X. Corbillon, F. De Simone, and G. Simon, “360-degree video head movement dataset,” in *Proc. ACM Multimedia Syst. Conf.*, 2017, pp. 199–204.

[27] Fraunhofer heinrich hertz institut, 2016. [Online]. Available: https://www.hhi.fraunhofer.de/

[28] T. Schubert, F. Friedmann, and H. Regenbrecht, “The experience of presence: Factor analytic insights,” *Presence: Teleoperators Virtual Environ.*, vol. 10, no. 3, pp. 266–281, 2001.

[29] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, “Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness,” *Int. J. Aviation Psychol.*, vol. 3, no. 3, pp. 203–220, 1993.