Delay Compensation for Actuated Stereoscopic 360 Degree Telepresence Systems with Probabilistic Head Motion Prediction

Tamay Aykut      Christoph Burgmair      Mojtaba Karimi      Jingyi Xu      Eckehard Steinbach
Chair of Media Technology, Technical University of Munich
{tamay.aykut, christoph.burgmair, mojtaba.karimi, jingyi.xu, eckehard.steinbach}@tum.de

Abstract

Communication delay is a major challenge for the acceptance of telepresence applications. It is particularly critical when the user experiences the remote environment via a Head-Mounted-Display. The lag between head motion and display response results in motion sickness, indisposition, and, at worst, abortion of the telepresence session. In this paper, we propose a delay compensation approach for 3D 360° telepresence systems realized with a mechanically actuated stereoscopic vision system. We further introduce a novel metric to evaluate the achievable level of delay compensation. We investigate state-of-the-art head motion predictors and propose a novel probabilistic prediction paradigm, which can half the mean prediction error and improve the level of delay compensation by up to 26%. The general validity of our approach is shown by means of two independent real head motion datasets. The experimental results verify that average compensation rates of more than 99% can be achieved for communication delays between 100-500ms.

1. Introduction

Telepresence systems allow users to immerse themselves into a remote environment. The quality-of-experience (QoE) is crucial for the acceptance of these types of systems. 3D perception of the remote scene is desirable as it enhances the level of immersion compared to monocular views and in many cases also improves the task performance. For this, a stereoscopic vision system needs to be deployed to provide separate image content from different vantage points for each eye. This allows for the perception and sensation of depth. Fig. 1 shows a telepresence scenario, where the user is equipped with a Head-Mounted-Display (HMD) at the local side and a robotic camera system (e.g., a Pan-Tilt-Roll Unit (PTRU)) at the remote side that follows the user’s head movement. In between is a communication network that transmits the video and audio signals. The communication network introduces unavoidable delay that has negative implications on the QoE. If the so-called motion-to-photon (M2P) latency, which describes the amount of time that is required to fully reflect the user’s motion and display it on the screen is higher than 20ms, the user will experience visual discomfort. It is indispensable that the user perceives its ego-motion to be consistent with the sensory impressions from the visual system, the vestibular system, and the non-vestibular proprioceptors [21]. If this premise is not met and the user’s expectation based on prior experiences is not in accordance with these inputs, the visual discomfort of the user will lead to motion sickness and/or nausea [5, 16, 19]. The resulting deterioration of the QoE provokes the discontinuation of the telepresence session.

An intuitive solution to prevent the negative implications of delay would be to use a 360° x 180° video acquisition system. The immediate access to the complete visual representation of the remote scene would allow the user to gaze around keeping the M2P latency low. While recording monoscopic full panorama videos is considered state-of-the-art, the real-time acquisition of 3D panorama videos still poses major challenges. Either catadioptric or multicamera approaches are used to approximate binocular vision for 360° [1–4, 6, 11, 15, 17, 23]. These systems are...
usually bulky, expensive, and often not real-time capable. They are computationally demanding and need to be precisely calibrated and positioned for correct stitching. The stitching process is highly dependent on the features in the image and, hence, error prone. Distortions in the video content are magnified by HMDs due to the small display-to-eye distance and are, thus, even more critical. Besides that, sending two complete monocular 360° videos requires substantial communication capacity, even though large portions of the imagery are not shown to the user. In view of these facts, we decided to exploit the benefits of an actuated stereoscopic camera system (e.g., a PTRU), which is able to provide the user with a 360° x 180° stereoscopic visual impression. When the user rotates his/her head, the current orientation data of the HMD is sent to the remote side to mimic the head motion. Depending on the network delay, it takes a certain amount of time to replicate the head motion and send the updated image frames back to the user. The accumulation of delays causes incongruities between ego-motion and visual response. Simply using a PTRU-based setup is, hence, not appropriate, as the introduced M2P latency will provoke the user to suffer from motion sickness and/or nausea. The Delay Compensation Vision System (DCVS) proposed in [7] compensates the perceivable latencies for horizontal (pan) motions by capturing a larger field-of-view (fov) than actually displayed. The resulting image buffer is used for local delay compensation. In this paper, we extend this approach to 3 dimensions and compensate for joint pan, tilt, and roll rotations. We use equidistant fisheye cameras for a larger fov. The so-called 3D compensation rate is presented as a novel metric to describe the achievable level of delay compensation. To further improve the delay compensation, we investigate commonly used head motion predictors. A probability-based prediction strategy is additionally proposed, which can be applied to existing methods and shows substantial improvement.

Our contributions can be summarized as follows:

- We extend the delay compensation approach in [7] for pan, tilt, and roll rotations and introduce in this context the so-called 3D compensation rate as a qualitative measure to evaluate the level of latency compensation. We show its superior performance in contrast to the naive approach, where no compensation is applied.

- We equip the camera setup with equidistant fisheye cameras to increase the image buffer, and hence the achievable level of compensation.

- Additionally, we investigate different head motion prediction methods and propose a novel prediction paradigm based on a probabilistic error model. We validate our approach by means of two real and independent head motion datasets to prove its general validity.

2. Related Work

Sending the whole 360° visual representation of the scene would allow the user to freely rotate his/her head to keep the M2P latency low. Beside the fact that the omni-stereoscopic real-time acquisition of full panorama videos is still an open research topic [11, 23], these systems also entail certain disadvantages [7]. Some of them were previously discussed in Section 1. Deploying the smallest possible number of 2 cameras in combination with a mechanically actuated element provides instead a lean, low-cost, and real-time capable system. In contrast to available 3D 360° acquisition systems, where the 3D perception is limited to certain dominant directions, such an approach allows for 3D perception in every viewing direction [1–4, 6, 15, 17]. These systems are also characterized by a flexible stereoscopic budget, that is, the inter-ocular distance can easily be adapted to each user.

Simply using a stereoscopic PTRU, however, is not sufficient. The present total delay, which is an accumulation of various contributing latencies (see Section 3), has a great (negative) impact on the QoE. If the M2P is too high and the head is turned, the screen will first stay static or frozen, until the motion is eventually reflected. The QoE is thereby highly dependent on the underlying control or prediction method. Applying, however, a compensation approach allows for certain error tolerance and improves the visual comfort.

Former prediction approaches aimed to compensate especially the local lag between head motion and display response. Such local latencies originate from the time needed for tracking the head and rendering the imagery onto the HMD [8, 12–14, 18, 22]. Even nowadays, it may take more than 20ms to render image content onto the HMD [9]. That is why those prediction techniques usually aimed to compensate delays in the range of 10-100ms. There are typically two standard practices. One comprises the consideration of past values and subsequently fitting a (weighted) first-order linear function. The second way is to integrate filters like the (Extended) Kalman Filter or Particle Filter to first obtain a better state estimate and use a polynomial function for extrapolation. We tested both techniques and show their limitations especially for large prediction times. To counteract this, we propose a probabilistic error model that can half the errors of existing prediction methods and increase thereby the delay compensation by 26%.

3. 1D Delay Compensation

In the following section, we will briefly recapitulate the concept behind the DCVS proposed in [7]. The considered telepresence scenario comprises a user equipped with an HMD, a stereoscopic vision system on the remote side, and a communication network in between. In [7], a stereo
camera system mounted on a Pan-Tilt-Unit is used to follow the user’s head motion. The main objective was to compensate the overall aggregated delay $t_{\text{total}}$ for the rotation in horizontal direction. During a global operating cycle ranging from the head movement of the user to the telepresence camera setup and back again involves several delays that contribute to the total latency $\tau$ [7]:

$$\tau = t_{\text{total}} = t_s + t_m + t_c + t_p + t_r + 2 \cdot t_n. \quad (1)$$

t_s is determined by the sampling rate of the orientation sensor in the HMD, $t_m$ reflects the mechanical delay of the actuators and the deployed hardware setup, $t_c = \frac{1}{f_c}$ is defined by the available camera frame rate. The processing delay $t_p$ conveys the required time for rectifying, encoding, and processing the captured imagery. The time that is needed to extract the received frames, decode them, and render them to the HMD is summarized by the rendering delay $t_r$. A crucial portion of the latency is given by the network delay $t_n$. In general, the total latency $\tau = t_{\text{intr}} + t_r t_{\text{rtt}}$ can be divided into the round trip network delay $t_{\text{rtt}}$, which is assumed to be approximately $2 \cdot t_n$, and the intrinsic delay $t_{\text{intr}}$. $t_{\text{intr}}$ is determined by the system’s characteristics and is inherently given.

The key concept of the DCVS is to capture more imagery than actually shown to the user. The field-of-view of the user ($fov_h$) is thereby a subset of the camera’s field-of-view ($fov_c$): $fov_h \subset fov_c$. The remaining image content $\hat{\theta}_h = \hat{\theta}_h^{\text{hori}} = \frac{1}{2}(fov_c^{\text{hori}} - fov_h^{\text{hori}})$ (2)
can be used for local delay compensation. When the user changes his/her viewing direction, the residual imagery $b$ can be leveraged to locally adapt the view orientation without noticing the present latency until the updated frame arrives. In this way, the user’s perception of ego-motion can match the sensory inputs from the visual system, the vestibular system, and the non-vestibular proprioceptors mitigating the effect of cyber sickness [16].

The level of compensation is thereby dependent on the time $t_{\text{comp}}$ it takes to turn through the provided buffer $b$ for a given (horizontal) head velocity $\hat{\theta}_h^{\text{hori}}$ [7]:

$$t_{\text{comp}} = \frac{\hat{\theta}_h^{\text{hori}}}{\dot{\theta}_h^{\text{hori}}} = \frac{fov_c^{\text{hori}} - fov_h^{\text{hori}}}{2 \cdot \hat{\theta}_h^{\text{hori}}} \cdot t_n. \quad (3)$$

Hence, a full compensation can be guaranteed as long as $t_{\text{comp}} \geq \tau$. Subject to this condition the roundtrip network delay that can be fully compensated can be estimated as follows

$$t_{\text{rtt}} \approx 2 \cdot t_n \leq \frac{fov_c^{\text{hori}} - fov_h^{\text{hori}}}{2 \cdot \hat{\theta}_h^{\text{hori}}} - t_{\text{intr}}. \quad (4)$$

Along with the provided $fov_c$ and the present communication delay $t_{\text{rtt}}$, the current angular head velocity $\dot{\theta}_h^{\text{hori}}$ of the user strongly affects the achievable delay compensation. The compensation rate $c \in [0, 1]$ describes the amount of compensation where $c = 1$ is equivalent to full (100%) compensation [7]

$$c(\dot{\theta}_h^{\text{hori}}, t_{\text{rtt}}) = \frac{1}{2}(fov_c^{\text{hori}} + fov_h^{\text{hori}}) - \frac{\dot{\theta}_h^{\text{hori}} \cdot (t_{\text{rtt}} + t_{\text{intr}})}{fov_h^{\text{hori}}}. \quad (5)$$

### 4. 3D Delay Compensation

In [7], a stereo camera system mounted on a Pan-Tilt-Unit (PTU) was used to prove the concept behind DCVS. The compensation was, however, only tested for the horizontal, so the yaw or pan, rotation. In our work, we add another degree-of-freedom (DoF) – the roll rotation – and apply a combined compensation for all three DoFs. Our telepresence setup with its 3-DoF Pan-Tilt-Roll-Unit (PTRU)-based binocular camera system is shown in Fig. 1. From Eq. 5 it can be inferred that a greater $fov_c$ facilitates a larger buffer and would, hence, yield a higher overall compensation rate. There is, however, a trade-off between the level of compensation and the perceivable depth. As a result, we decided to increase the buffer size by equipping the cameras with fisheye lenses. In the following, we will describe the extension of the DCVS concept to 3 DoFs and provide a detailed system description.

#### 4.1. Three Dimensional Rotation Description

We denote the angular rotation around the x-axis with $\phi$ (pitch, tilt) and around the y-axis with $\theta$ (yaw, pan) (see also Fig. 4). The expressions for yaw and pan as well as tilt and pitch are considered to be identical and are interchanged throughout this work.

The overall rotation matrix $R \in \mathbb{R}^{3 \times 3}$ is expressed as the product of the individual rotations for roll ($R_\psi$), pitch ($R_\phi$), and yaw ($R_\theta$)

$$R = R_\psi \cdot R_\phi \cdot R_\theta, \quad (6)$$

where the pan and tilt rotations correspond to the orientation around their respective axes

$$R_\theta = \begin{bmatrix} \cos(\Delta \theta) & 0 & \sin(\Delta \theta) \\ 0 & 1 & 0 \\ -\sin(\Delta \theta) & 0 & \cos(\Delta \theta) \end{bmatrix}, \quad (7)$$

$$R_\phi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\Delta \phi) & -\sin(\Delta \phi) \\ 0 & \sin(\Delta \phi) & \cos(\Delta \phi) \end{bmatrix}. \quad (8)$$

The present angular deviation between the head orientations and the PTRU orientation is expressed as $\Delta \psi = |\psi_h - \psi_c|$, $\Delta \phi = |\theta_h - \theta_c|$, and $\Delta \theta = |\phi_h - \phi_c|$. In contrast to the yaw and pitch rotation, we assume the head’s roll rotation $R_\theta$ to be around the optical axis instead of the z-axis. The optical axis can change dynamically and can be any arbitrary axis.
through the origin. The direction vector of the optical axis \( \vec{v} \) is thereby dependent on the present pan and tilt orientation and the focal length \( f \)

\[
\vec{v} = \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = R_\phi \cdot R_\theta \cdot \begin{bmatrix} 0 \\ 0 \\ f \end{bmatrix}.
\]  

(9)

By considering the rotation of an arbitrary vector \( \vec{x} \) around the view direction vector \( \vec{v} \) by the amount of \( \Delta \psi \), the rotation matrix for the roll orientation \( R_\psi \) can be computed as follows [20]

\[
R_\psi \cdot \vec{x} = \left[ \cos(\Delta \psi) + (1 - \cos(\Delta \psi)) \cdot \frac{\vec{v} \otimes \vec{v}}{||\vec{v}||^2} \right] \cdot \vec{x} + \frac{\sin(\Delta \psi)}{||\vec{v}||} \cdot (\vec{v} \times \vec{x}),
\]

(10)

\[
R_\psi = \begin{bmatrix}
    k_1 v_x^2 + k_2 & k_1 v_x v_y - k_3 v_z & k_1 v_x v_z + k_3 v_y \\
    k_1 v_x v_y + k_3 v_z & k_1 v_y^2 + k_2 & k_1 v_y v_z - k_3 v_x \\
    k_1 v_x v_z - k_3 v_y & k_1 v_y v_z + k_3 v_x & k_1 v_z^2 + k_2
\end{bmatrix},
\]

with \( k_1 = \frac{1 - \cos(\Delta \psi)}{||\vec{v}||^2}, k_2 = \cos(\Delta \psi), \) and \( k_3 = \frac{\sin(\Delta \psi)}{||\vec{v}||} \).  

4.2. Three Dimensional Compensation

The images captured with a fisheye camera provide a larger \( \text{fov}_c \) and, thus, more residual imagery for local compensation. In [7], the compensation rate \( c \) was defined as the ratio of available image content and the horizontal \( \text{fov}_{\text{hori}} \) of the imagery that is shown to the user. Following the conventions from [7], we propose the 3D compensation rate \( c_{\text{rpy}} \) for a combined roll, pitch, and yaw orientation

\[
c_{\text{rpy}} = \frac{\text{area}(\Pi)}{l_{\text{hori}} l_{\text{vert}}},
\]

(12)

\( l_{\text{hori}} \) and \( l_{\text{vert}} \) represent the width and height of the HMD’s image plane and are computed as follows

\[
l_{\text{hori}} = 2 \cdot f \cdot \tan \left( \frac{\text{fov}_{\text{hori}}}{2} \right), l_{\text{vert}} = 2 \cdot f \cdot \tan \left( \frac{\text{fov}_{\text{vert}}}{2} \right).
\]

(13)

The general concept of the delay compensation approach is depicted in Fig. 2. While the image plane of the HMD \( I_h \) is considered as a 2D plane in the 3D space, which is equivalent to the one of a perspective camera, the image surface of a camera equipped with an equidistant fisheye lens is hemispherical. The radius of the hemispherical surface is equivalent to the focal length \( f \) of the fish-eye camera. Depending on the user’s view orientation, we map the requested hemispherical image portion onto the HMD’s 2D image plane to simultaneously rectify the imagery before displaying it on the HMD. The area of the available rectified image content from the fisheye camera for the present view orientation of the user is denoted as \( \text{area}(\Pi) \). \( \Pi \) corresponds to the set of all available image points \( \vec{p}_i \in \Pi \).

To calculate the overlapping area, we first determine the position of the corner points \( P_c = [\vec{p}_{00}, \vec{p}_{01}, \vec{p}_{10}, \vec{p}_{11}] \) of the image plane \( I_h \)

\[
\vec{p}_{00} = R \cdot \begin{bmatrix} -l_{\text{hori}}/2 \\ -l_{\text{vert}}/2 \end{bmatrix}, \vec{p}_{01} = R \cdot \begin{bmatrix} l_{\text{hori}}/2 \\ -l_{\text{vert}}/2 \end{bmatrix}, \vec{p}_{10} = R \cdot \begin{bmatrix} -l_{\text{hori}}/2 \\ l_{\text{vert}}/2 \end{bmatrix}, \vec{p}_{11} = R \cdot \begin{bmatrix} l_{\text{hori}}/2 \\ l_{\text{vert}}/2 \end{bmatrix}.
\]

(14)

(15)

We introduce the auxiliary measure \( h \) as a function of the camera’s \( \text{fov}_{\text{hori}} \) to describe the permitted height of an arbitrary image point \( \vec{p}_i \in I_h \) to be in \( \Pi \)

\[
h(\vec{p}_i, \text{fov}_{\text{hori}}) = |\vec{p}_i| \cdot \cos \left( \frac{\text{fov}_{\text{hori}}}{2} \right).
\]

(16)

For a better understanding, the derivation of \( h \) is illustrated in Fig. 3 (left). Any image point \( \vec{p}_i \) is considered to be in the set \( \Pi \) as long as the following condition is met

\[
\Pi \cup \{ \vec{p}_i \} \forall \vec{p}_i \in I_h \iff p_{z,i} \geq h(\vec{p}_i, \text{fov}_{\text{hori}}).
\]

(17)

Depending on the images’ resolution the amount of pixels can increase significantly. To avoid iterating over all image points, we compute the \( \text{area}(\Pi) \) by determining the set of respective corner and edge points \( \Pi_E \). We first use Eq. 17 to check if the four corner points \( \vec{p}_m \in P_c \forall m = 1...4 \) of
Iₖ are in Π and add them to Πₑ if satisfied. If |Πₑ| = 0 (or 1), we can conclude a zero (or full) compensation. Otherwise, we compute the edge points and construct a line
g : x'_{ij} = \bar{p}_i + \lambda(\bar{p}_j - \bar{p}_i), \quad (18)
with \bar{p}_i and \bar{p}_j being opposite points each between two neighbouring corner points. For λ ∈ [0, 1], g returns a point on the respective line segment through Iₖ. We then iterate this line from the left edge to the right. For a point to be an edge point we formulate the following condition
\[ x_{z,ij} = \frac{1}{h(\bar{x}_{ij}, f_{ov}^{\text{hori}})}, \]
\[ p_{z,i} + \lambda(p_{z,j} - p_{z,i}) = h(\bar{p}_i + \lambda(\bar{p}_j - \bar{p}_i), f_{ov}^{\text{hori}}). \quad (19) \]
We compute λ by means of the Newton Raphson algorithm. If λ ∈ [0, 1], \( \bar{x}_{ij} \) is considered to be an edge point and added to set Πₑ (\( Πₑ \cup \{ \bar{x}_{ij} \} \)).

After complementing the set of edge points, we transform the set Πₑ into the center of the xy-plane to ease the computation of area(Π)
\[ \Pi_{E}^{xy} = R^T \cdot (\Pi_{E} - \vec{v}), \quad (20) \]
where R is the orthogonal \( R^{-1} = R^T \) rotation matrix from Eq. 6. The resulting transformation of the polygon is shown in Fig. 3 (right). The area of Π is calculated by integrating over Πₑ. We therefore define auxiliary top and bottom border curves \( t \) and \( b \), which enclose Π in the range from \( x_{\min} \) to \( x_{\max} \). The correct points to be assigned to \( t \) and \( b \) are calculated from the set \( \Pi_{E}^{xy} \). Finally, the area of Πₑ can be computed as follows
\[ \text{area}(\Pi) = \int_{x_{\min}}^{x_{\max}} (t - b) \, dx, \quad (21) \]
with \( x_{\min} \) and \( x_{\max} \) being the bounds of integration
\[ x_{\min} = \{ x \in \mathbb{R} \mid \min\{p_{x,i}\}, \forall \bar{p}_i = (p_{x,i}, p_{y,i})^T \in \Pi_{E}^{xy} \}, \]
\[ x_{\max} = \{ x \in \mathbb{R} \mid \max\{p_{x,i}\}, \forall \bar{p}_i = (p_{x,i}, p_{y,i})^T \in \Pi_{E}^{xy} \}. \quad (22) \]

5. Head Motion Prediction

In Section 7, we demonstrate the performance of the 3D compensation approach. To further enhance the compensation effect, we propose a novel head motion prediction strategy based on a probabilistic error model.

The core idea behind our approach is to take an arbitrary predictor \( \bar{I}_{t+\tau} \) as baseline and weight it with its probability to be at that position. For this purpose, we use the real head motion dataset from [7] to determine the error distribution at every time instance. We assumed a Gaussian distribution and fitted the probability density function accordingly. We trained the mean \( \mu(\tau) \) and standard deviation \( \sigma(\tau) \) for all delays \( \tau \in [0, 1, 2s] \). This was done for all 3 DoFs with \( \bar{\xi} = [\theta, \phi, \psi]^T \). The resulting solution space for \( \mu(\tau) \) and \( \sigma(\tau) \) as a function of the delay \( \tau \) is defined as
\[ \mu(\tau) = \kappa \cdot \tau + \zeta, \]
\[ \sigma(\tau) = \eta \cdot \ln(\tau + \chi) + \vartheta, \quad (23) \]
with \( \kappa, \zeta, \eta, \chi, \) and \( \vartheta \) being empirically determined parameters. The error probability density function can, thus, be expressed as
\[ \bar{P}(\Delta\bar{\xi}, \tau) = \frac{1}{\sqrt{2\pi} \cdot \sigma(\tau)^2} \cdot \exp \left( -\frac{(\Delta\bar{\xi} - \mu(\tau))^2}{2 \cdot \sigma(\tau)^2} \right), \quad (24) \]
with \( \Delta\bar{\xi} = [\Delta\theta, \Delta\phi, \Delta\psi]^T \). The resulting probabilistic prediction for \( \bar{\xi}_{t+\tau}^{\text{pred}} \) after the latency \( \tau \) can then be specified as
\[ \bar{\xi}_{t+\tau}^{\text{pred}} = \bar{\xi}_t + \bar{P} \left( (\bar{I}_{t+\tau} - \bar{\xi}_t), \tau \right) \cdot (\bar{I}_{t+\tau} - \bar{\xi}_t). \quad (25) \]
The initial predictor \( \bar{I}_{t+\tau} \) can be any desired forecasting method (e.g., Linear Regression (LR) or a Kalman Filter (KF)). In our system, we use the Kalman Filter for a proper state estimation. With this state estimate at hand, we apply a constant acceleration motion model (CAMM) to extrapolate the orientation at time \( t + \tau \).
which uses the state estimates from the Kalman Filter for the constant acceleration motion model.

Figure 4: Overview of the overall applied control scheme. We train a probabilistic error model to modify the initial predictor, from Eq. 25

The final prediction values that are sent to the PTRU can be obtained by applying the probability based error model from Eq. 25

\[
\hat{\xi}_{t+\tau}^{\text{pred,KF}} = \hat{\xi}_t + \hat{P} \left( \left( \xi_{t+\tau}^{\text{pred,KF}} - \hat{\xi}_t \right), \tau \right) \cdot \left( \xi_{t+\tau}^{\text{pred,KF}} - \hat{\xi}_t \right).
\] (27)

The overall control graph is depicted in Fig. 4.

6. Experiments

Our general experimental telepresence setup is depicted in Fig. 1. For the server side, we prepared two scenarios. One involves a PTRU, where actuators are used to mirror the users head motion. In the second one, we use a 360° stereo panorama video as input to prove that the underlying concept can be applied for both scenarios. For a representative validation, we used the real head motion dataset from [7] (called LMT dataset in the following). The 30 participants were aged between 22-40 and watched 3 dissimilar 360° videos with changing dynamics in the respective scene. The entire dataset contains 30 \cdot 3 = 90 subsets, each with an average video length of 120s. With the IMU sensor providing the filtered orientation data at a frequency of \( f_{\text{IMU}} = 80 \text{Hz} \), we had approximately 864000 data points at hand. We exploited 81 profiles to train the probabilistic error model and used the remaining 9 for validation. We proved its general validity by cross-validating the proposed approach on another independent 360° head motion dataset (denoted as IMT dataset) with completely different participants [10]. In [10], the head motion profile of 58 subjects was recorded, each watching five 70s long 360° videos.

We use the mean-error, the root mean square error (RMSE), and the 3D compensation rate \( c_{\text{rpy}} \) to investigate the quality of our probabilistic compensation approach compared to others. We utilize the mean compensation rate \( \bar{c}_{\text{rpy}} \), which averages the mean for all videos \( V \), participants \( P \), and available samples \( S \) for a present delay \( \tau \)

\[
\bar{c}_{\text{rpy}} = \frac{1}{V \cdot P \cdot S} \sum_{i=1}^{P} \sum_{j=1}^{V} \sum_{k=1}^{S} c_{i,j,k}.
\] (28)

with \( S = \frac{t_{\text{end}} - t_{\text{start}}}{t_{\text{imu}}} \). The evaluation was done for a circular fisheye camera \( f_{\text{ov}}^{\text{hor}} = f_{\text{ov}}^{\text{ver}} = 150^\circ \) with different buffer sizes and changing delays \( \tau \in [0.1 - 28] \). Therefore, we varied the horizontal \( f_{\text{ov}} \) of the user \( f_{\text{ov}}^{\text{hor}} \in \{60^\circ, 90^\circ, 110^\circ\} \), but kept a ratio of 8:9 to be in accordance with the HMD’s expected aspect ratio. The corresponding vertical \( f_{\text{ov}} \) of the HMD can be calculated with

\[
f_{\text{ov}}^{\text{ver}} = 2 \cdot \arctan \left( \frac{1}{\text{ratio}} \cdot \tan \left( \frac{f_{\text{ov}}^{\text{hor}}}{2} \right) \right).
\] (29)

However, the horizontal and vertical \( f_{\text{ov}} \) are not representative figures for the determination of the available buffer. A smaller \( f_{\text{ov}}^{\text{hor}} \) used for visualization compared to the equidistant fisheye camera's \( f_{\text{ov}} \) does not imply a larger buffer. This phenomenon is demonstrated in Fig. 5. A more indicative measure is given by the diagonal \( f_{\text{ov}}^{\text{diag}} \)

\[
f_{\text{ov}}^{\text{diag}} = 2 \cdot \arctan \left( \sqrt{\tan^2 \left( \frac{f_{\text{ov}}^{\text{hor}}}{2} \right) + \tan^2 \left( \frac{f_{\text{ov}}^{\text{ver}}}{2} \right)} \right).
\] (30)

The available (diagonal) buffer can consequently be computed as follows

\[
b^{\text{diag}} = \frac{1}{2} \left( f_{\text{ov}}^{\text{diag}} - f_{\text{ov}}^{\text{diag}} \right).
\] (31)

Table 1 summarizes the examined buffer sizes.

Table 1 summarizes the examined buffer sizes.
Table 1: $fov_{\{\text{hori, vert, diag}\}}$ and the resulting diagonal buffer sizes used for evaluation.

<table>
<thead>
<tr>
<th>$f_{fov_c}$</th>
<th>hori</th>
<th>vert</th>
<th>diag</th>
<th>$y_{\text{diag}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$60^\circ$</td>
<td>150°</td>
<td>150°</td>
<td>81.98°</td>
<td>34.01°</td>
</tr>
<tr>
<td>$80^\circ$</td>
<td>66.01°</td>
<td>103.26°</td>
<td>23.37°</td>
<td></td>
</tr>
<tr>
<td>$100^\circ$</td>
<td>106.56°</td>
<td>121.72°</td>
<td>14.14°</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Front view of the camera’s image surface and the borders of the image content, which is mapped onto the HMD’s image plane.

7. Discussion

In Fig 7, we contrast the naive approach ($h = 0$), where no compensation is applied, to our proposed 3D compensation approach (3D DCVS) for a diagonal buffer size of $34.01^\circ$ and varying delays. It is clearly observable that applying the 3D compensation significantly improves the 3D $360^\circ$ visual experience. Even for large delays, a major portion of the HMD’s display can be filled with image content. To allow the user’s ego-motion to comply with the expectations and the sensory inputs from the visual system, we provide instantaneous response by showing any available video content. The residual parts are filled with black pixels. The considerable positive impact of the 3D DCVS compared to the naive way can be seen for both datasets. To further improve the already high compensation rates, we investigated different head motion prediction methods in combination with the 3D DCVS. Besides the respective compensation rates illustrated in Fig. 7, we use the mean error and the root mean squared error (RMSE) as qualitative measures for a fair comparison. Fig. 6 depicts the present errors for the dominant, that is the horizontal, rotations. The results demonstrate that simply using Linear Regression ($\approx$ naive LR) or a Kalman Filter ($\approx$ naive KF) is not enough to improve the performance. One can even see that these predictors lead to a discernible deterioration. To address this concern, we propose a novel probabilistic modification. We leveraged our trained probabilistic error model for LR and KF (+ CAMM). Both the mean error and the RMSE in Fig. 6 as well as the resulting compensation rates in Fig. 7 reveal that the incorporation of our proposed probabilistic model leads to a substantial improvement compared to their naive versions. In some instances, the errors are halved and the compensation rate increased by more than 26% for large delays. The best outcomes are achieved by 3D DCVS + prob. KF. Superior performance is identifiable for any buffer size and amount of delay. The results for the IMT dataset are even better and thereby prove the general validity of our proposed concept.

Fig. 8 indicates that the delay compensation rate increases with a larger buffer size. However, there is a trade-off between perceivable depth sensation and level of delay compensation for horizontal stereoscopic camera configurations as depicted in Fig. 1 due to the changing effective baseline.

8. Conclusion

In this paper, we propose a 3D delay compensation approach, which extends the approach in [7] to 3 dimensions. We provide a system description for equidistant fisheye camera based binocular vision systems. The proposed delay compensation approach, which extends the approach in [7] to 3 dimensions. We provide a system description for equidistant fisheye camera based binocular vision systems. $c_{\text{top}}$ is introduced as a metric to describe the amount of compensation for a joint head rotation. State-of-the-art prediction techniques are investigated and re-implemented to further enhance the level of compensation. While existing methods tended to deliver inferior results, we proposed a novel probability-based prediction strategy. The 3D DCVS in combination with a Kalman Filter based polynomial extrapolation approach and the proposed probabilistic modification thereof featured the best performance. We used real head motion datasets [7, 10] for meaningful evaluation. We used 90% of [7] for training and the remaining 10% for evaluation. To prove our concept’s general validity, we additionally conducted a cross-validation with the head motion dataset from [10]. Meaningful qualitative measures verify the superior performance of our approach.

In future work, we plan to use neural networks to enhance the quality of prediction. Moreover, we aim to conduct extensive subjective tests. As mentioned previously, we display the user all available image content and pad the residual parts with black pixels. We claimed that this technique facilitates the sensory inputs to be in accordance with the user’s ego-motion and, thus, mitigates the effect of motion sickness. This statement is premised on numerous positive feedbacks we received during experimentation and validation. Our next goal is to verify our hypothesis by means of subjective experiments.
Figure 6: Mean error and RMSE shown for horizontal rotations and tested delays between 0.1-2s. Applying the probabilistic error model leads to a substantial improvement, especially for large delays.

Figure 7: Mean compensation rate \( \bar{\tau}_{\text{rpy}} \) for tested delays between 0.1-2s. 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. The incorporation of the probabilistic modification further improves the overall level of compensation.

Figure 8: Comparison of mean compensation rates for different diagonal buffer sizes and tested delays between 0.1-2s. The level of compensation strongly depends on the amount of available buffer.
References