Freie Universität Berlin
Fachbereich Mathematik und Informatik
Takustraße 9, 14195 Berlin

MASTER THESIS

USER POSITION PREDICTION IN 6-DOF MIXED REALITY APPLICATIONS USING ARTIFICIAL RECURRENT NEURAL NETWORK

VORHERSAGE DER BENUTZERPOSITION IN 6-DOF-MIXED-REALITY-ANWENDUNGEN UNTER VERWENDUNG EINES KÜNSTLICHEN REKURRENTEN NEURONALEN NETZWERKS

Oleksandra Baga

Freie Universität Berlin
Matrikelnummer 5480722
Master Computer Science
E-Mail: oleksandra.baga@gmail.com

Prof. Dr. First Supervisor
Fachbereich Mathematik und Informati
Freie Universität Berlin

Prof. Dr. Second Supervisor
Fachbereich Mathematik und Informati
Freie Universität Berlin

Statutory Declaration

I herewith formally declare that I have written the submitted master thesis independently. I did not use any outside support except for the quoted literature and other sources mentioned in the paper.

I clearly marked and separately listed all of the literature and all of the other sources which I employed when producing this academic work, either literally or in content.

I am aware that the violation of this regulation will lead to failure of the thesis.

29.06.2022.. Oleksandra Baga

Acknowledgments

This thesis was created in cooperation with the Fraunhofer Heinrich Hertz Institute.

First and foremost I would like to thank Prof FU Berlin, who supervised my thesis. Thank you very much for the helpful suggestions and constructive criticism.

A special thanks goes to the researcher Fraunhofer Heinrich Hertz Institute, Serhan Gül, who suggested an exciting topic for a research, which I was allowed to choose for my master thesis. I would like to express my sincere thanks for the commitment and the consultation during the preparation of this thesis.

Contents

List of Figures

Listings

List of Abbreviations

ANN	Artificial Neural Networks
AR	Augmented Reality
CNN	Convolutional Neural Network
CPU	Central processing unit
DoF	Degree of freedome
DL	Deep Learning
FFN	Feed-forward Neural Network
GRU	Gated Recurrent Unit
HMD	Head-Mounted-Display
IEEE	Institute of Electrical and Electronics Engineers
KF	Kalman Filter
LAT	Look ahead time
LSTM	Long-Short-Term Memory
M2P	Motion-to-Photon
MAE	Mean Absolut Error
MEC	Mobile Edge Computing
ML	Machine Learning
MR	Mixed Reality
NLP	Natural Language Processing
ReLu	Rectified Linear Unit
RNN	Recurrent Neural Network
RTT)	Round-trip time
SDG	Stochastic Gradient Descent
VR	Virtual Reality
3-DoF	Three degree of freedom
6-DoF	Six degree of freedom

Introduction

1

This thesis is focusing on designing and evaluation of the approach for the prediction of human head position in a 6-dimensional degree of freedom (6-DoF) of Extended Reality (XR) applications for a given look-ahead time (LAT) in order to reduce the Motion-to-Photon (M2P) latency of the network and computational delays. At the beginning of the work the existing 3-DoF as well as 6-DoF methods were analyzed, and their similarities differences were taken into account when a proposed Recurrent Neural Network-based predictor was developed. The investigation of different neural network architectures and the improvement of head motion prediction is the main goal of this thesis. Proposed approach was evaluated and at the end of the work the obtained results were discussed and the suggestions for future work were done.

1.1 Problem statement

The correct and fast head movement prediction is a key to provide a smooth and comfortable user experience in VR environment during head-mounted display (HDM) usage. The recent improvements in computer graphics, connectivity and the computational power of mobile devices simplified the progress in Virtual Reality (VR) technology. The way users can interact with their devices changed dramatically. With new technologies of VR environment user becomes the main driving force in deciding which portion of media content is being displayed to them at any time of interaction with VR Applications [13]. Until recently the high-quality experiences with modern Augmented Reality (AR) and VR systems were not widely presented in home usage and were mainly used in research labs or commercial setups. The hardware for displaying the VR environment was once extremely expensive but recent years became more broadly accessible and the 6-DoF VR headset designed for the end-user were released[1]. It is possible now to experience virtual reality scenes and watch new type of volumetric media at home and the market interest for development VR and AR applications expected to be huge next years.

In fact, the existing on this moment virtual environments can be divided into two main groups depending on position of the user and their ability to move inside the VR environment. The user motion and prediction within a 3-DoF environment has been intensely researched for years. Extending such approaches to a 6-DoF

[1] https://medium.com/@DAQRI/motion-to-photon-latency-in-mobile-ar-and-vr-99f82c480926

environment is not straightforward, due to the change of the user's viewing point from inward to outward and additional three degrees of freedom [14].
Although all mentioned above improvements, rendering of volumetric content remains very demanding task for existing devices. Thus the improvement of a performance of existing methods, design and implementation of new approaches specially for the 6-DoF environment could be a promising research topic.

1.2 Motivation for the research

Research efforts to reduce the computational load are being already wide attempted. However, these approaches designed for the client side. Recently presented technique of the rendering on a cloud server makes possible to decrease the computational load on the client device by offloading of the task to a server infrastructure and than by sending the rendered 2D content instead of volumetric data [9]. The calculated 2D view must correspond the current position and orientation of a user. However, cloud-based streaming approach adds network latency and processing delays due uploading to a server the user position, rendering a new 2D picture from the 3D data and sending it back to a device. Thus, a rendered 2D image can appear even later on a display than with usage of local rendering system.
The promising research topic is reducing the Motion-to-Photon (M2P) latency by predicting the future user position and orientation for a look-ahead time (LAT) and sending the corresponding rendered view to a client. The LAT in this approach must be equal or larger to the M2P latency of the network including round-trip time (RTT) and time need for calculation and rendering of a future picture at remote server.

1.3 Structure of the thesis

The organization of this thesis is as follows. The thesis starts from introduction and problem statement, followed by theoretical background related to the research topic. Literature review chapter introduces different approaches and technologies of motion prediction algorithms. The chapters 4 and 5 show the implementation of presented model and evaluation of the results that were obtained during experiments. Last, the discussion regarding method limitations and suggestions for the future work are done.

Chapter 1 - Introduction.
The current chapter shortly introduces a state of development on scientific field achieved at a time of master thesis creation in the context on XR applications. The necessity of timely action to improve the situation with increasing computational

and network latency is shown in problem statement section 1.1. Due to the breadth of the research topic, the section 1.2 focuses the research topic and clearly motivates the implementation with neural network model.

Chapter 2 - Background.
The next chapter includes a review of the area being researched. It starts with a short introduction of the concept of MR applications and presents a 6-DoF environment. The presence and influence of a computational and network latency is covered, followed by discussion of possible solutions for its reduction. In section 2.5 the head pose estimation algorithms and the challenges faced in predicting of the viewer's position are discussed.

Chapter 3 - Related work.
The chapter provides an overview of previous research in the field of prediction of user's head position and orientation. The related works divided into section corresponding the computational approach of the prediction method. The chapter places a master thesis's topic in the context of the existing literature and the last section focuses in the methods using neural networks.

Chapter 4 - Data and Model.
Fourth chapter presents the design of the proposed method. The dataset including data collection from head mounted display (HDM) and data understanding and preprocessing are described in section 4.1, followed by a design of the algorithm including network architecture, functions of an input layers and the training methods.

Chapter 5 - Data and Model.
Fifth chapter gives an overview of the implementation of the described in previous chapter method. It describes the conducted experiments with a data obtained from HMD. The evaluation metrics and results are presented in the section 5.3.

Chapter 6 - Analysis.
The last chapter presents a discussion about the limitation of proposed method and provides a conclusion about the received results including suggestions for potential types of future research.

Background 2

This chapter introduces theoretical background of the presented research problem. First, the concept of mixed reality (MR) and the relation of this huge topic to the research field is presented, followed by an introduction of six degree of freedom (6-DoF) environment and the difference to the three degree of freedom (3-DoF). The term motion-to-photon latency (M2P) is covered, followed by a short discussion about an influence of M2P latency on the decreasing of user experience. The new developed cloud-based rendering and streaming approach is shortly discussed in this chapter. The last section of this chapter highlights challenges with the prediction of viewer's head pose that arises in modern XR applications in connection especially with the added network latency due the using of remote cloud server for computational offload.

2.1 Mixed reality

2.2 Six degrees of freedom

2.3 Motion-to-photon latency

2.4 Cloud-based volumetric video streaming

2.5 Challenges of head motion prediction

Related work 3

This chapter presents the overview of previous research in the field of the prediction of user position. It includes both approaches for 3-DoF and 6-DoF environment, focusing on time series methods, Kalman Filter and overview of Deep Learning Algorithms including methods using Recurrent Neuronal Network.

3.1 Time series methods

3.2 Kalman Filter

KF-based extrapolation are deemed to be robust against fast fluctuations, but suffer from susceptibility to noise sensory data [3].

3.3 Deep Learning Algorithms

Typical HMD computes user positions in 6-DoF by using its tracker module and data comes as time series with a sequential order. The structure of an input is crucial and needed to be followed in order to predict the next future step for a look-ahead time correctly. A sequence of inputs can be processed with Artificial Neural Network (ANN) called Recurrent Neural Network (RNN). Moreover, RNN can processes input with remembering its state while processing the next sequence of inputs. In the last decade, RNN algorithms have been adopted for motion prediction of 3D sequences. The authors *Aykut et al., 2018* claims their research to be first work that applies deep learning for head motion prediction. The current three rotations in three dimensions the so-called Euler angles as well as past values thereof within a certain time window (W) used as input for the network [3]. The best results delivered when 20 last values for each orientation direction (W = 250 ms) were used [3]. Instead of the absolute orientation values *Aykut et al., 2018* suggest for a goal of generalization for other datasets to subdivide the inputs into their respective orientation groups and compute their normalized differences [3]. The loss function MAE performed better compared to the MSE in experiments done by researchers. The authors experimentally confirmed that Feed-forward Neural Network (FFN) indeed had difficulties to learn for different delays even after an architecture was extended with the present delay and injected into all NN nodes. Next *Aykut et al., 2018* reasoned using of LSTM-based architectures with feedback loop and ability to establish a way of memory and share weights over time [3]. Researchers used Adam optimization algorithm, the maximum number of epochs was set to 1000, early stopping technique (patience = 2, min. delta = 0) was used to avoid overfitting. Additionally, the learning rate was decreased by 70% from initial value of 0.001 every 30 epochs. The batch size was set to 2^{11}. Rectified Linear Unit (ReLU) as activation function for the FFN layers used with LSTM [3].

Three different architecture were tried. In all variants an input for NN was subdivided into dimensions and normalized within time window W set to 250 ms. Thus all variants received 3-dimensional vector of size $[[pan, tilt, roll], W, batchsize]$ as input. First interleaved variant receives at first step input into stacked LSTM architecture comprised of nine LSTM layers. The output of stacked LSTM is fed together with initial input into nine input layers of FFN. The next step is to feed an output of FNN with previously computed output from stacked LSTM into new stacked nine-layered LSTM and the result of FNN separately goes into similar FNN again. This cooperation from LSTM and FNN loops calculations four times. The last loop's output from stacked LSTM and FNN goes through nine-layered LSTM once again and finally thirty-layered FNN computes final result for orientation with 1s LAT [3]. Second variant called LSTM with subdivided inputs in which each orientation dimension with input vector of size $[[dimension, W, batchsize]$ goes through 10-layered stacked LSTM and the result of all three LSTMs forms new input vector of size $[[pan, tilt, roll], W, batchsize]$ that is feed into 30-layered stacked LSTM. The output than goes through 30-layered FNN and thus final result for orientation with 1s LAT will be computed [3].

Third variant looks much simplified. The input vector $[[pan, tilt, roll], W, batchsize]$ goes through input layer with 30 nodes containing each LSTM. Two additional intermediate layers with 30 LSTM nodes each perform computations and extract the features from the data. Final layer has 30 nodes of FNN and thus als computes final result with 1s LAT [3].

4 Data and Model

4.1 6-DoF Dataset

4.1.1 Data collection from HMD

4.1.2 Data Exploration

!! Data analysis AVG linear velocity position, plots

4.1.3 Data preprocessing

4.2 Neural Network

4.2.1 Network architecture

4.2.2 Network input

4.3 Training methods

5 Implementation and experiments

Here some code for my super neural network. The artificial neural networks discussed in this text are only remotely related to their biological counterparts. In this section we will briefly describe those characteristics of brain function that have inspired the development of artificial neural networks.

```python
class StudentFactory(DjangoModelFactory):
        class Meta:
                model = Student
        student_card = factory.SubFactory(StudentCardFactory)
        first_name = factory.Faker('first_name')
        second_name = factory.Faker('last_name')
```

Listing 5.1: StudentFactory

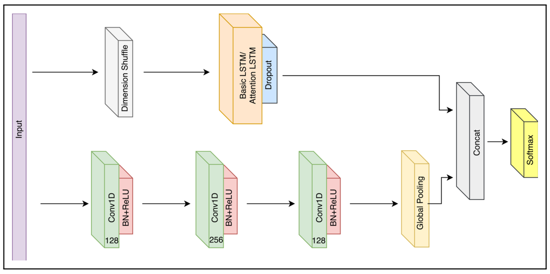

Figure 1: LSTM Fully Convolutional Networks for Time Series Classification

You might already know that you want to apply an established theory or set of theories to a specific context (for example, reading a literary text through the lens of critical race theory, or using social impact theory in a market research project).

5.1 Implementation

5.2 Experiments

5.3 Evaluation metrics

5.4 Results

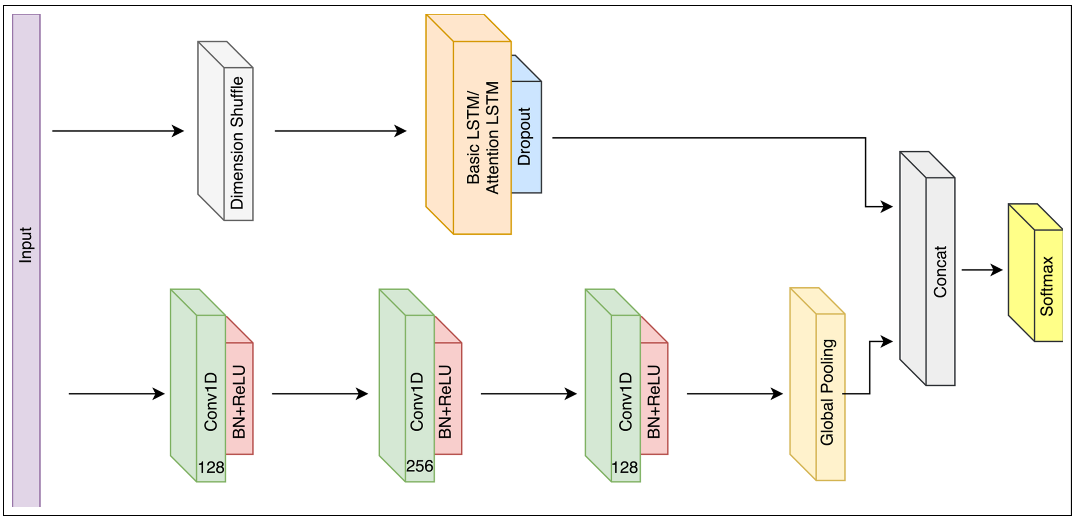

Figure 2: LSTM FCN WRAP

6 Analysis

The introduction text to analysis chapter

6.1 Limitations

6.2 Conclusion

6.3 Suggestions for future work

Glossary

AJAX

AJAX (asynchrones Javascript und XML) ist der allgemeine Name für Technologien, mit denen asynchrone Anforderungen (ohne erneutes Laden von Seiten) an den Server gestellt und Daten ausgetauscht werden können. Da die Client- und Serverteile der Webanwendung in verschiedenen Programmiersprachen geschrieben sind, müssen zum Austausch von Informationen die Datenstrukturen (z. B. Listen und Wörterbücher), in denen sie gespeichert sind, in das JSON-Format konvertiert werden.

Bibliography

[1] A. Deniz Aladagli, Erhan Ekmekcioglu, Dmitri Jarnikov, and Ahmet Kondoz. *Predicting head trajectories in 360° virtual reality videos*. https://ieeexplore.ieee.org/document/8251913. date access on 29.03.22. 2017. DOI: 10.1109/IC3D.2017.8251913.

[2] R.S. Allison, L.R. Harris, M. Jenkin, U. Jasiobedzka, and J.E. Zacher. *Tolerance of temporal delay in virtual environments*. https://www.researchgate.net/publication/2945506. date access on 17.03.22. 2001. DOI: 10.1109/VR.2001.913793.

[3] Tamay Aykut, Christoph Burgmair, Mojtaba Leox Karimi, and Eckehard Steinbach. *Delay Compensation for a Telepresence System With 3D 360 Degree Vision Based on Deep Head Motion Prediction and Dynamic FoV Adaptation*. https://arxiv.org/abs/2007.14084. date access on 23.02.22. 2018. DOI: 10.1109/WACV.2018.00222.

[4] Yun-Kai Chang, Mai-Keh Chen, Yun-Lun Li, Hao-Ting Li, and Chen-Kuo Chiang. *6DoF Tracking in Virtual Reality by Deep RNN Model*. https://ieeexplore.ieee.org/document/9394069. date access on 02.04.22. 2020. DOI: 10.1109/IS3C50286.2020.00057.

[5] Howie Choset et al. *Principles of Robot Motion: Theory, Algorithms, and Implementations*. the MIT Press, 2006, p. 630. ISBN: 978-0262033275.

[6] Alessandro Crivellari and Euro Beinat. *LSTM-Based Deep Learning Model for Predicting Individual Mobility Traces of Short-Term Foreign Tourists*. https://www.researchgate.net/publication/338377314. date access on 08.04.22. 2020. DOI: 10.3390/su12010349.

[7] Richard O. Duda, Peter E. Hart, and David G. Stork. *Principles of Robot Motion: Theory, Algorithms, and Implementations*. Wiley-Interscience; 2. edition, 2000, p. 688. ISBN: 978-0471056690.

[8] Serhan Guel, Sebastian Bosse, Dimitri Podborski, Thomas Schierl, and Cornelius Hellge. *Kalman Filter-based Head Motion Prediction for Cloud-based Mixed Reality*. https://arxiv.org/abs/2007.14084. date access on 19.02.22. 2020. DOI: 0.1145/3394171.3413699.

[9] Serhan Gül, Dimitri Podborski, Thomas Buchholz, Thomas Schierl, and Cornelius Hellge. *Low-latency Cloud-based Volumetric Video Streaming Using Head Motion Prediction*. https://arxiv.org/abs/2001.06466. date access on 19.02.22. 2020. DOI: 0.1145/3394171.3413699.

[10] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. *LSTM Fully Convolutional Networks for Time Series Classification*. https://arxiv.org/abs/1709.05206. date access on 14.04.22. 2017. DOI: 10.48550/arXiv.1709.05206.

[11] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. *Your Attention is Unique: Detecting 360-Degree Video Saliency in Head-Mounted Display for Head Movement Prediction*. https://www.researchgate.net/publication/328370817. date access on 15.03.22. 2018. DOI: 10.1145/3240508.3240669.

[12] Feng Qian, Lusheng Ji, Bo Han, and Vijay Gopalakrishnan. *Optimizing 360 video delivery over cellular networks*. https://dl.acm.org/doi/10.1145/2980055.2980056. date access on 12.03.22. 2016.

[13] Silvia Rossi, Irene Viola, Laura Toni, and Pablo Cesar. *A New Challenge: Behavioural Analysis Of 6-DOF User When Consuming Immersive Media*. https://ieeexplore.ieee.org/document/9506525. date access on 03.04.22. 2021. DOI: 10.1109/ICIP42928.2021.9506525.

[14] Silvia Rossi, Irene Viola, Laura Toni, and Pablo Cesar. *From 3-DoF to 6-DoF: New Metrics to Analyse Users Behaviour in Immersive Applications*. https://www.researchgate.net/publication/357172010. 1-7. 2021. date access on 13.04.22.

[15] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. the MIT Press, 2006, p. 672. ISBN: 978-0262201629.

[16] Dingqi Yang, Benjamin Fankhauser, Paolo Rosso, and Philippe Cudre-Mauroux. *Location Prediction over Sparse User Mobility Traces Using RNNs: Flashback in Hidden States*. https://www.researchgate.net/publication/338377314. date access on 07.04.22. 2020. DOI: 10.24963/ijcai.2020/302.

[17] Emin Zerman, Radhika Kulkarni, and Aljosa Smolic. *User Behaviour Analysis of Volumetric Video in Augmented Reality*. https://ieeexplore.ieee.org/document/9465456. date access on 13.04.22. 2021. DOI: 10.1109/QoMEX51781.2021.9465456.