# A review of methods for human motion prediction in video sequences with application to 3D telepresence and holoportation systems

Maria Yotova[1], Nikolay Neshov[2], Agata Manolova[3]

*Abstract* – **Nowadays, the new technologies especially 5G are paving the way of innovative solutions for communication between humans. One very new technology is holoportation, but it also presents a lot of challenges in terms of digital data gathering and transmission. Human motion prediction methods will offer one of the solutions for real time communication in the holoportation context. We present a conceptual model of holoportation architecture for real time communication based on highly accurate 3D modelling of the human face and body, recognition and prediction of human actions and facial expressions to achieve realistic communications. Designed this way the proposed conceptual model of the holoportation system addresses the challenges from the information transmission aspect where real time constraints and narrowband channels are imposed.**

*Keywords* – **holoportation, immersive telepresence, mixed reality, prediction of human actions.**

## I. INTRODUCTION

The way people see the world and interact between themselves is changing; 5G and other new technologies are paving the way of innovative solutions for real time life like long distance communication between humans. Without doubt video communication tools such as Skype and FaceTime are useful for many applications and tasks. However, text, 2D image and voice are not enough of a satisfying alternative to the personal interaction because the real eye contact is missing, touch is not possible between interlocutors and there is a limited feeling of presence (sharing the same space) [1].

These facts bring a sense of incompleteness and dissatisfaction from the communication process to the user. So how the ever evolving technology will help?

Currently in their infancy, augmented reality (AR), mixed reality (MR) and virtual reality (VR) technologies offer great potential to include all five human senses in the communication process to make it more meaningful for all participants. These technologies are already changing the way people interact between each other and with the machines and eventually will radically restructure the work process and the way of living. They will make it easier and more effective for executives, teachers, technologists, people with different professions to communicate in a more natural way anywhere around the globe, with all the advantages of physical presence but without the limitations of the current 2D video communication systems.

One of the new technologies that can enhance the current state of the 2D communication is holoportation. Holoportation is a complete human-computer-machine interface combining AR and VR telepresence thus creating real-life digital scans and realistic 3D avatars of subjects, displayed in MR environment, used for real time communication and interaction between remote users [2]. Holoportation can incorporate all five senses - optic, auditory, olfactory, gustatory, and tactile. This technology opens the door to new heights of interpersonal communication, but it also presents a lot of challenges in terms of digital data gathering and transmission.

In this paper we present a schematic model of a holoportation architecture addressing real time communication needs, based on highly accurate 3D modelling of the human face and body, recognition and prediction of human actions. We will concentrate our efforts to present and review the most current methods for human action prediction – a vital step for the real time holoportation process.

The rest of the paper is organized as follows: the next section presents the current state of the art of immersive telepresence systems. In section III the key features/building blocks of the proposed holoportation system will be described with accent to the human action prediction step. The final section draws the conclusion and suggest the scope of future work.

## II. CURRENT TELEPRESENCE SYSTEMS

In the last couple of years many research groups approached in different ways the above noted challenges with system developments for different use case scenarios and contributed to the advancing field of telepresence research. With the increased availability and sophistication of consumer RGB-D sensors and VR/AR glasses, there is currently an exponential emergence of 3D telepresence systems.

One of the most comprehensive reviews and classification of the current state of the immersive telepresence is done in [3]. The author presents an "Immersive Group-to-Group Telepresence" system by implementing a 3D capture and reconstruction pipeline which generates and distributes realistic 3D user representations in real time. One of the functionalities of the systems is the use of two large multi-user projection displays which offer individual stereoscopic perspectives for up to six co-located users. The distributed VR

[1]Maria Yotova, Agata Manolova and Nikolay Neshov are with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria, E-mail: E-mail: m.yotova08@gmail.com, nneshov@tu-sofia.bg amanolova@tu-sofia.bg.

framework delivers gaze tracking and eye contact between participants. It suffers from fairly coarse user's representation due to the low resolutions of the utilized RGB-D sensors. The framework was tested over a local area network so no data for more remote connection is yet available.

One of the very recent papers describes a low-cost, low-bandwidth telepresence system capable of rendering people and objects in 3D through data fusion and reconstruction without the use of head-mounted displays or any other wearable devices [4]. The visualization is done on a quadrangular acrylic pyramid by projecting the images using a video projector. This approach albeit requiring low bandwidth suffers from a serious resolution decrease from 1920 × 1080 to 640 × 480 with limited size holographic display and no high fidelity audio transmission and synchronization.

SLAMCast [5] is a very promising multi-client real time telepresence framework for remote collaboration which enables efficient remote exploration of quasi-static scenes by multiple independent users while at the same time being able to observe the other user's interactions with the environment. An oriented towards interpersonal communication cost-effective telepresence framework based on 3D data streaming and real time 3D reconstruction is described in [6]. The authors use skeleton data extracted from the depth sensors to animate a 3D human model including rigging the face to express facial expressions. The initial experimental setup show encouraging results to accomplish real time shared person space but no latency information or compensation of data loss is supplied; the sensors used do not allow transferring complex hand movements and multi-user real time telepresence is not assured.

AVATAREX [7] follows the principles of MR and connects users that are simultaneously occupying the same space in the real world and its virtual replica. This framework deals with the idea of how users experience the co-presence in AR and VR. VR users share the same gaming area, but they are physically in another room than the AR users. Even though AVATAREX covers both aspects of shared space it does not currently deal with distance communication so data transfer and latency are not considered in this research.

Many more research papers on the topic of immersive telepresence including different senses can be cited but there is no universal solution. The proposed approaches' success depends on the sensors used, the type of data gathered and way of data processing. The immersive telepresence systems are built on several complex software and hardware technologies, so it is very important to pinpoint what are the current challenges and requirements for real time holoportation.

## III. Schematic Model of the Holoportation System

### *A. Online and offline communication*

A holoportation system is composed of multiple consecutive steps in order to receive an avatar object.

The first step is the object scanning, where we aim to capture the body of the person in order to estimate its shape, skinning, and pose. There are multiple ways to implement a high quality 3D body scanner, but the most common one is by using high quality cameras in a controlled environment. The scanning process can be performed either online or offline, but in both cases there are shortcomings.

The online scanning system must have a very high performance in order to reconstruct the body. For this scenario a Kinect is used, which captures the raw data, then this data needs to be processed in order to be received correctly.

On the other hand –the offline scanning is performed only once and therefore it is more difficult to animate the avatar in real-time. However, the offline scanning creates more accurate and noiseless avatar, because the cameras used for the scan have very high quality.

### *B. Data processing*

The next logical step is the data processing. The captured images are used to create a 3D model of the human facial characteristics and body with real textures such as skin, facial expressions and clothing. Creating such model based on images is usually done in multiple stages [8]. Capturing and representing accurately the facial characteristics and expressions especially when the parties in the communication process wear large physical devices such as VR/AR glasses, which occlude the majority of the face, is very challenging. So in order to identify the user and capture in detail the facial expressions, the facial data will be gathered and transmitted separately in the communication channel. We must parametrize the facial characteristics of the participants during the teleconference [9]. The facial key points transmitted in real time are used to reanimate the avatar's face.

The second task for the proposed holoportation will be the real time avatar animation visualized at the remote site, based on the metadata captured at the home site. The created 3D model needs to be rigged with the captured skeleton hierarchy created by the Kinect sensors and appropriate texture maps. Skeleton data is transmitted separately in the communication channel. A skeleton based animation strategy is employed for robustly and accurately fitting the avatar to the skeleton and then large scale deformations and movements are applied in real time. Thanks to the multiple Kinect sensors employed, there are no occlusions of joints.

Additionally, we will employ a method for recognizing human activities, to perform short term prediction of the skeleton movements to compensate for network latency.

### *C. Motion prediction*

To provide a real time experience in a two way communication, between the participants, all the data, including body movement, speech and facial expressions, must be transferred without any lag, because it will lead to inconsistency of the speech-movement relation, for example. There is no need to transfer the static objects, because they can be created in both systems – such objects are furniture, walls, etc.

In a short distance, transferring a 3D human model a real time communication isn't an issue, but unfortunately, in a long distance communication there are restrictions due mainly to the speed of the data transfer, because a real-time transfer on a LAN network reaches a speed between 1-2Gbps, and to achieve a real time communication, we need to transfer hundreds of megabits. To avoid the network delay, a human motion prediction algorithm can be applied. This way, by scanning the avatar, we can send a metadata, which will be received by the recipient, containing the predicted position of the sender. Then the skeleton, presented by a number of joints that represent key body parts, animates the avatar, avoiding the long-distance data-transfer delay.

In this section we will make an overview of the recent methods that are used to make a short-term human motion prediction. To make a motion prediction in a video sequence there are several steps to be completed. First must be performed an action detection, then this action must be tracked and after that, we may say that this action is recognized. After the recognition of the action, comes the classification and then it is possible to make a short term prediction on what the next body position will be. Traditional approaches use Markovian assumptions [10, 11], but the latest works on this problem are based on different methods which use convolutional neural networks(CNN) or recurrent neural networks, or more specifically on Long short term memory (LSTM) and Gated Recurrent Unit (GRU). The LSTMs are widely used in action and speech recognition [12] thanks to their ability to learn long-term feature relationships by processing overlapping sequences of consecutive frames.

To solve the motion prediction task, Fragkiadaki et al. [13]. prose two architectures: LSTM-3LR (3 layers of Long Short-Term Memory cells) and ERD (Encoder-Recurrent-Decoder). Both are based on concatenated LSTM units, but the latter adds non-linear space encoders for data pre-processing. The authors also note that, during inference, the network is prone to accumulate errors, and quickly produces unrealistic human motion. Therefore, they propose to gradually add noise to the input during training which forces the network to be more robust to prediction errors. This noise scheduling makes the network able to generate plausible motion for longer time horizons, especially on cyclic walking sequences. Jabri et al. [14] have shown competitive performance on VQA with a simple baseline that does not take images into account, and state-of-the-art performance with a baseline that is trained to exploit the correlations between questions, images and answers.

In their paper Martinez et al. [15] propose a different approach than the LSTM networks. Their work is based on the GRU [16], with the help of which, they manage to drop the spatial encoding layer. This allows them to train their model on 3.6M Dataset for a few hours, and there is no need to train a different model for each different action. They analyse the motion continuity as velocity, rather than a set of poses, which allows them to model only one velocity, as opposed to presenting all possible human poses. To achieve this, they base their work on Sequence-to-Sequence architecture [13] with the help of which during training, the ground truth is fed to an encoder network, and the error is computed on a decoder network that feeds its own predictions. The decoder also has a residual connection, which effectively forces the RNN to internally model angle velocities.

Pavllo et al. [17] propose similar approach as Martinez et al. for short term predictions using a quaternion network, we consider predicting either relative rotation deltas (analogous to angular velocities) or absolute rotations. We take inspiration from residual connections applied to Euler angles, where the model does not predict absolute angles but angle deltas and integrates them over time. For quaternions, the predicted deltas are applied to the input quaternions through quaternion product. First they create a recurrent architecture, where they also use a GRU, because of its simplicity. The idea is that at each step, the model receives the previous state and features, in order to make a next pose estimation. After further work, they create a convolutional based architecture, where they replace the GRU and the linear layer with convolutional layers.

In Table 1 are summarized the most recent papers on motion prediction on the Human3.6M dataset (http://vision.imar.ro/human3.6m/description.php).

## IV. Conclusion

The main goal of this paper is to provide a comprehensive survey of the most recently published motion prediction methods that can be employed in immersive telepresence systems. And also to present a schematic model of a holoportation system and to outline some of its key concepts. The benefit of the proposed holoportation system is that it allows users to interact practically in real time - anywhere, anytime, with anybody – either in a virtual or integrated way offering the feeling of personal interactivity and the feeling of shared space. Holoportation offers the opportunity to address the current limitations in 2D communication permitting immediate, 3-D visual, auditory, tactile and emotional interaction between remote users. But a lot of work is still needed to make this system an everyday reality for the society.

## Acknowledgement

This work was supported in part by the contract № 182ПД0027-07 for research project: "Recognizing human activity from video sequences through recurrent and convolutive neural networks " of the Technical University of Sofia, Research Sector.

## References

[1] Cohen, Aviva, et al. "Sustaining a caring relationship at a distance: Can haptics and 3D technologies overcome the deficits in 2D direct synchronous video based communication?." Virtual System & Multimedia (VSMM), 23rd International Conference on. IEEE, 2017.

[2] Orts-Escolano, Sergio, et al. "Holoportation: Virtual 3d teleportation in real-time." Proceedings of the 29th Annual

Symposium on User Interface Software and Technology. ACM, 2016.
[3] Beck, Stephan. "Immersive Telepresence Systems and Technologies." (2019)., PhD thesis Fakultät Medien der Bauhaus-Universität Weimar.
[4] Córdova-Esparza, D.-M., Terven, J. R., Jiménez-Hernández, H., Herrera-Navarro, A., Vázquez-Cervantes, A., & García-Huerta, J.-M. (2019). Low-bandwidth 3D visual telepresence system. Multimedia Tools and Applications. doi:10.1007/s11042-019-7464-0.
[5] Stotko, P., Krumpen, S., Hullin, M. B., Weinmann, M., & Klein, R. (2019). SLAMCast: Large-Scale, Real-Time 3D Reconstruction and Streaming for Immersive Multi-Client Live Telepresence. IEEE transactions on visualization and computer graphics.
[6] Huynh, D., King, S. A., & Katangur, A. K. (2018). A framework for cost-effective communication system for 3D data streaming and real-time 3D reconstruction. Proceedings of the 3rd International Workshop on Interactive and Spatial Computing - IWISC '18. doi:10.1145/3191801.3191804
[7] Koskela, T., Mazouzi, M., Alavesa, P., Pakanen, M., Minyaev, I., Paavola, E., & Tuliniemi, J. (2018, February). AVATAREX: Telexistence System based on Virtual Avatars. In Proceedings of the 9th Augmented Human International Conference (p. 13). ACM.
[8] Zanfir, Mihai, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. "Human appearance transfer." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5391-5399. 2018.
[9] Manolova, Agata, Nikolay Neshov, Stanislav Panev, and Krasimir Tonchev. "Facial expression classification using Supervised Descent Method combined with PCA and SVM." In International Workshop on Biometric Authentication, pp. 165-175. Springer, Cham, 2014.
[10] McGhan, Catharine LR, Ali Nasir, and Ella M. Atkins. "Human intent prediction using markov decision processes." Journal of Aerospace Information Systems 12, no. 5 (2015): 393-397.
[11] Lehrmann, Andreas M., Peter V. Gehler, and Sebastian Nowozin. "A non-parametric bayesian network prior of human pose." In Proceedings of the IEEE International Conference on Computer Vision, pp. 1281-1288. 2013.
[12] Liu, Jun, Amir Shahroudy, Dong Xu, and Gang Wang. "Spatio-temporal lstm with trust gates for 3d human action recognition." In European Conference on Computer Vision, pp. 816-833. Springer, Cham, 2016.
[13] Fragkiadaki, Katerina, Sergey Levine, Panna Felsen, and Jitendra Malik. "Recurrent network models for human dynamics." In Proceedings of the IEEE International Conference on Computer Vision, pp. 4346-4354. 2015.
[14] Jabri, Allan, Armand Joulin, and Laurens Van Der Maaten. "Revisiting visual question answering baselines." In European conference on computer vision, pp. 727-739. Springer, Cham, 2016.
[15] Martinez, Julieta, Michael J. Black, and Javier Romero. "On human motion prediction using recurrent neural networks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2891-2900. 2017.
[16] Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).
[17] Pavllo, Dario, Christoph Feichtenhofer, Michael Auli, and David Grangier. "Modeling Human Motion with Quaternion-based Neural Networks." arXiv preprint arXiv:1901.07677 (2019).

| Modeling Human Motion | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Action | Walking | | | | Eating | | | | Smoking | | | | Discussing | | | | Base architecture and NN |
| milliseconds | 80 | 160 | 320 | 400 | 80 | 160 | 320 | 400 | 80 | 160 | 320 | 400 | 80 | 160 | 320 | 400 | |
| Zero-velocity (Martinez et al., CVPR 2017) | 0.39 | 0.68 | 0.99 | 1.15 | 0.27 | 0.48 | 0.73 | 0.86 | 0.26 | 0.48 | 0.97 | 0.95 | 0.31 | 0.67 | 0.94 | 1.04 | Sequence-to-sequence architecture, RNN, GRU |
| ERD (Fragkiadaki et al., CVPR 2015) | 0.93 | 1.18 | 1.59 | 1.78 | 1.27 | 1.45 | 1.66 | 1.80 | 1.66 | 1.95 | 2.35 | 2.42 | 2.27 | 2.47 | 2.68 | 2.76 | ERD architecture, RNN, LSTM (with non-linear space encoders) |
| LSTM-3LR (Fragkiadaki et al., CVPR 2015) | 0.77 | 1.00 | 1.29 | 1.47 | 0.89 | 1.09 | 1.35 | 1.46 | 1.34 | 1.65 | 2.04 | 2.16 | 1.88 | 2.12 | 2.25 | 2.23 | RNN, LSTM |
| QuaterNet abs. (Pavllo et al., BMVC 2018b) | 0.26 | 0.42 | 0.67 | 0.70 | 0.23 | 0.38 | 0.61 | 0.73 | 0.32 | 0.52 | 0.92 | 0.90 | 0.36 | 0.71 | 0.96 | 1.03 | Quaternion-based neural architectures, RNN, GRU,CNN |
| QuaterNet vel. (Pavllo et al., BMVC 2018b) | 0.21 | **0.34** | 0.56 | **0.62** | 0.20 | 0.35 | 0.58 | 0.70 | 0.25 | 0.47 | 0.93 | 0.90 | 0.26 | 0.60 | 0.85 | 0.93 | Quaternion-based neural architectures,RNN, GRU,CNN |
| QuaterNet vel. TF | **0.20** | 0.37 | 0.64 | 0.76 | 0.19 | 0.34 | 0.61 | 0.78 | **0.24** | 0.48 | 0.90 | 0.99 | 0.25 | 0.64 | 0.97 | 1.07 | Quaternion-based neural architectures,RNN, GRU,CNN |

Table 1: Results under the standard protocol (Fragkiadaki et al., 2015), with 4 samples per sequence. The mean angle error for short-term motion prediction on Human 3.6M for different actions is presented. Bold indicates the best result.