ITUPublications

International Telecommunication Union

Standardization Sector

ITU Focus Group Technical Specification

(06/2024)

ITU Focus Group on metaverse
(FG-MV)

FGMV-38

Framework and requirements for the construction of human-driven 3D digital human application system for metaverse

Working Group 2: Applications & Services

Technical Specification ITU FGMV-38

Framework and requirements for the construction of human-driven 3D digital human application system for metaverse

Summary

In the future, the idea of 3D digital human will become familiar to people as "super agents". Three-dimensional digital humans can display human characteristics such as facial appearance, gestures, and even a biological brain. Anthropomorphic behaviour of 3D digital humans can be generated through different driving technologies, which can be divided into intelligent-driven technology and human-driven technology.

With the popularization of human-driven 3D digital human applications and the advancement of image recognition technology such as posture and facial expressions recognition algorithm, inertial or optical motion capture devices are no longer essential tools for driving the 3D digital human. Instead, ordinary cameras, combined with ideal recognition algorithms, can achieve accurate driving of the 3D digital human. This approach not only benefits from the inherent flexibility and interactive capabilities imparted by human operators but also substantially lowers the barriers to entry and cost associated with generating virtual content. Consequently, it facilitates the intelligent transformation of the content creation industry.

This technical specification provides the framework and requirements of the 3D human-driven digital human application system for metaverse.

Keywords

Metaverse; digital human; human-driven; framework and requirements

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Change Log

This document contains Version 1.0 of the ITU Technical Specification on "*Framework and requirements for the construction of human-driven 3D digital human application system for metaverse*" approved at the 7th meeting of the ITU Focus Group on metaverse (FG-MV) held on 12-13 June 2024.

Acknowledgements

This Technical Specification was researched and written by Qiuhong Zheng (China Telecommunications Corporation, China) and Liang Wang (ZTE Corporation, China) as a contribution to the ITU Focus Group on metaverse (FG-MV), with assistance with Hideo Imanaka (NICT, Japan). The development of this document was coordinated by Yuntao Wang (CAICT, China) and Yuan Zhang (China Telecom), as FG-MV Working Group 2 Co-Chairs, and by Qiuhong Zheng (China Telecom) as Chair of the Task Group on Generative Artificial Intelligence in the metaverse.

Additional information and materials relating to this report can be found at: https://www.itu.int/go/fgmv. If you would like to provide any additional information, please contact Cristina Bueti at tsbfgmv@itu.int.

Editor & Task Group Chair:	Qiuhong Zheng China Telecommunications Corporation China	Tel: +86 17300137101 E-mail: zhengqh@chinatelecom.cn
Editor:	Liang Wang ZTE Corporation China	Tel: +86 25 88014641 E-mail: wang.liang12@zte.com.cn
WG2 Co-Chair	Yuntao Wang CAICT China	E-mail: wangyuntao@caict.ac.cn
WG2 Co-Chair	Yuan Zhang China Telecom China	E-mail: zhangy666@chinatelecom.cn

Table of contents

Page

Table of figures

Technical Specification ITU FGMV-38

Framework and requirements for the construction of human-driven 3D digital human application system for metaverse

1 Scope

Human-driven 3D digital human can be utilized for the avatar in metaverse. The application system based on the image recognition algorithm can enable users to interact with others in the virtual world of metaverse through voice conversation, intuitive and natural gestures, movements, and facial expressions by controlling their avatars.

This Technical Specification specifies the framework and requirements for the construction of human-driven 3D digital human application system for metaverse.

The scope of this Technical Specification includes:

– Overview of human-driven 3D digital human application system for metaverse.

– Framework of human-driven 3D digital human application system for metaverse.

– Requirements of human-driven 3D digital human application system for metaverse.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T F.748.15] Recommendation ITU-T F.748.15 (2022), *Framework and metrics for digital human application system*.

[ITU-T F.748.27] Recommendation ITU-T F.748.27 (2023), *Framework and requirements for the construction of 3D intelligent driven digital human application systems*.

3 Definitions

3.1 Terms defined elsewhere

This Technical Specification uses the following terms defined elsewhere:

3.1.1 digital human [ITU-T F.748.15]: A computer application that integrates the technologies of computer graphics, computer vision, intelligent speech and natural language processing. It can be used for digital content generation and human-computer interaction to help improve content production efficiency and user experience.

3.2 Terms defined in this Technical Specification

This Technical Specification defines the following terms:

3.2.1 Human-driven 3D digital human: A 3D digital human, which is driven by a human to represent a series of actions by technical means in 3D manner.

4 Abbreviations and acronyms

This Technical Specification uses the following abbreviations and acronyms:

3D Three Dimensions

CG Computer Graphics

UHD Ultra High Definition

5 Conventions

In this Recommendation:

– The keywords "**is required**" indicate a requirement which must be followed strictly and from which no deviation is permitted if conformance to this Recommendation is to be claimed.

– The keywords "**is recommended**" indicate a requirement which is recommended but is not absolutely required. Thus, this requirement need not be present to claim conformance.

– The keywords "**can optionally**" indicate an optional requirement which is permissible, without implying any sense of being recommended. These terms are not intended to imply that the vendor's implementation must provide the option and that the feature can be enabled optionally by the network operator/service provider. Rather, it means that the vendor may provide the feature optionally and still claim conformance with the specification.

6 Overview of human-driven 3D digital human application system for metaverse

In the future, the idea of 3D digital human will become familiar as "super agents". Three-dimensional digital humans display human characteristics like voice conversation, facial appearance, gestures, and even a biological brain. Anthropomorphic behavior of 3D digital humans can be generated through different driving technologies, which can be divided into intelligent driven technology and human driven technology. Employing intelligent driven technology, the digital human can partially replace the human workforce, often exceeding human capabilities, thereby cutting human resource costs and boosting productivity. This technology finds applications in various roles, including intelligent assistants, customer service agents, content presenters, and more. With human driven technology, the digital human can be the avatar of the user in metaverse, for attending virtual conferences, virtual games, and other scenarios in the metaverse. It has also been used widely in live streaming scenarios, including e-commerce live streaming and personal media live streaming, and so on.

Human-driven 3D digital human applications have become increasingly prevalent. In fact, this technological approach represents an extension of CG technology within the realm of media content production. In recent years, the main technical breakthrough has been in the motion capture part. With the advancement of image recognition technology such as the posture and facial expressions recognition algorithm, inertial or optical motion capture devices – which are expensive and usually not easily obtainable – are no longer essential tools for driving the 3D digital human. Instead, ordinary

cameras, combined with ideal recognition algorithms, can achieve accurate driving of 3D digital human. This approach not only benefits from the inherent flexibility and interactive capabilities imparted by human operators but also substantially lowers the barriers to entry and cost associated with generating virtual content. Consequently, it facilitates the intelligent transformation of the content creation industry.

7 Framework of human-driven 3D digital human application system for metaverse

The high-level framework of the 3D intelligent driven digital human application system based on image recognition algorithm is shown in Figure 1, based on Figure 1 in ITU-T F.748.15.

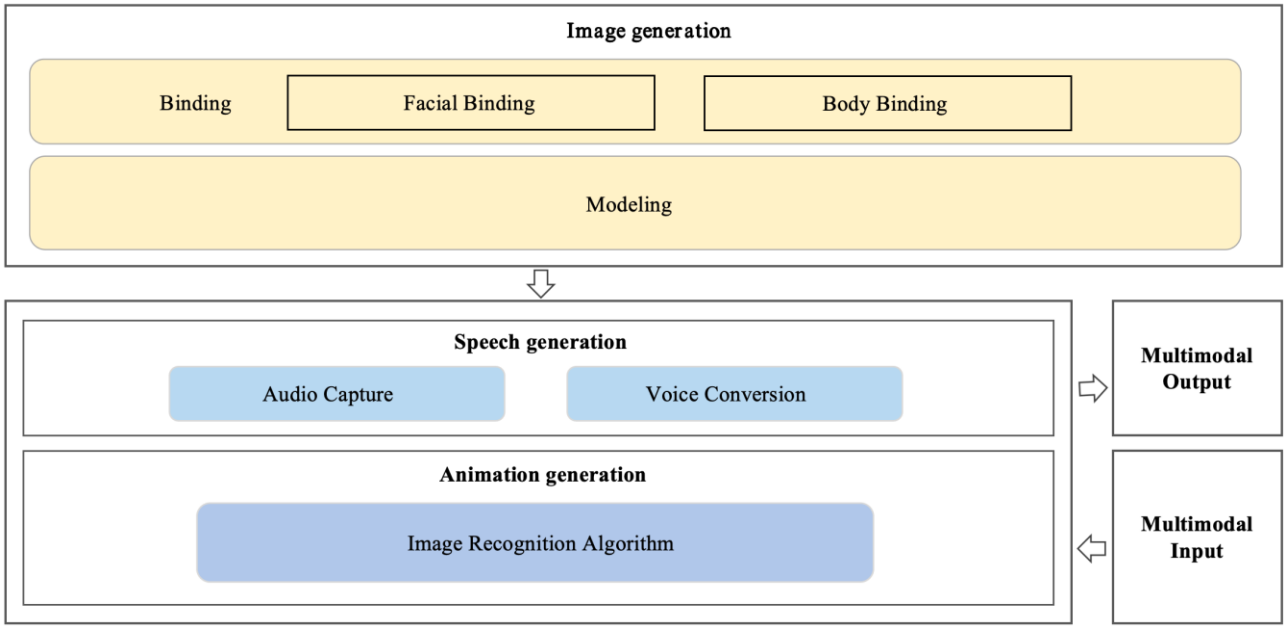

Figure 1 – Framework of human-driven 3D digital human application system for metaverse

– **Image generation module**: Generates the 3D model based on modelling and binding. Modelling can use a variety of modelling methods to generate the 3D model of the human-driven 3D digital human. Binding the facial features and body of the 3D model can achieve subsequent control of it.

– **Speech generation module**: A module that can generate the speech of the 3D digital human. One method is based on audio capture of the human voice, and the other is to perform voice conversion on the captured audio of the human (which can be used for voice cloning, speech enhancement, and voice transformation for hiding personal voice features or entertainment purposes).

– **Animation generation module**: A module that uses an image recognition algorithm to get the body movements and facial expression data of human from the input video, and applies it to 3D digital human, which can replicate the corresponding actions of the human.

– **Multimodal input module**: A module that is used to receive input, mainly the video captured by camera.

– **Multimodal output module**: A module that synthesizes animation and speech, and presents the output to the user.

8 Requirements of human-driven 3D digital human application system for metaverse

8.1 Image generation

The image generation module includes the modelling and binding of the 3D model of the digital human. Due to the consistent requirements for the 3D model of the digital human in human driven and intelligent driven 3D digital human application systems, image generation is required to meet the requirements in clause 8.1 of ITU-T F.748.27.

To enhance the representation of the diversity of human poses and expressions, and ensure high consistency between the movements and expressions of the 3D digital human and those of the human, it is recommended to achieve more precise skeletal binding for both facial (for example, no fewer than 687 blendshapes [b-Metahuman]) and body (for example, no fewer than 887 bones [b-Metahuman]).

8.2 Animation generation

In order to generate the movements and expressions of the human-driven 3D digital human, the animation generation component includes capturing fine facial expressions and body poses with minimal latency; the specific requirements are as follows.

- It is required to have the algorithm which recognizes human movements from the input image data provided by the multimodal input module. This includes major movements (such as walking, running, and jumping) and subtle gestures (such as facial expressions and hand movements).

- Due to the uncertainty of pose and expression changes when driven by a human, the 3D digital human's movements and facial expressions can vary significantly. Smoothing abrupt changes that may occur in raw motion capture data must be included, thus ensuring that the resulting animations appear natural and fluid.

- When the human who is driving the 3D digital human is replaced, the animation of the 3D digital human must be coherent and without any noticeable transition, interruptions or discrepancies.

NOTE - In certain scenarios, such as live streaming, there may be instances where the human operator (the anchor) who is driving the 3D digital human needs to be changed during the broadcast.

- It is recommended to have a real-time processing capability to process the captured data and generate the corresponding animations in real time or with minimal latency, thus providing an interactive experience for users.

- It is recommended to provide customization options for movement styles, offering users the ability to customize the style or characteristics of the 3D digital human's movements (e.g., more exaggerated or restrained), which can enhance the system's versatility and appeal.

- It is recommended to adapt to different lighting conditions, clothing types, and other environmental factors that might affect the quality of the captured data.

- It is recommended to include error correction mechanisms and to be robust to occasional errors in input data such as missing frames or incorrect tracking, and it should be possible to correct these errors seamlessly.

- Different applications may require different levels of detail in the skeletal structure of the 3D digital human. It is recommended to support multi-accuracy skeletal binding systems.

- When utilizing the 3D digital human for applications that require lip or body movement in sync with speech, it is recommended that it be possible to align animations with the timing of the speech.

8.3 Speech generation

8.3.1 Audio capture

The requirements for audio capture are the follows:

- To ensure the clarity and integrity of captured audio, it is recommended to have a dynamic range to capture sounds without clipping or distortion.

- It is recommended to be compatible with audio sources, such as microphones, headphones and speakers, to allow for flexible integration with various devices and environments.

- It is recommended to support real-time performance, avoiding delay between the spoken words and their capture by the system.

- It is recommended to support noise reduction and echo cancellation techniques to reduce background noise and environmental interference, so enhancing the intelligibility of the captured audio.

8.3.2 Voice conversion

The requirements for voice conversion are follows:

- It is required that the voice conversion process preserves the original content and meaning of the input audio. The converted voice is expected to accurately convey the same information and intent as the original audio.

- It is required to produce speech that sounds natural and resembles human-like speech, including the prosody, intonation, intonation contours, stress patterns, timing, and other cues that convey emotional meaning.

- To guarantee real-time performance, it is recommended to enable applications that support real-time functionality, minimize latency and ensure smooth, uninterrupted speech delivery.

- It is recommended to support multiple languages, enabling users to select from various options.

- It is recommended to support multiple voice styles, enabling users to select from various options.

NOTE - Multiple voice styles refer to the use of varied tones, moods, rhythms, and pitches to convey information or emotion.

- It can optionally support customization, allowing for the ability to fine-tune voice characteristics based on specific requirements or user preferences.

8.4 Multimodal input

The requirements for multimodal input are follows:

- It is required to support at least one type of input video, such as live video streams and offline video. For offline videos, it is recommended to support video storage and management.

- It is required to support one or multiple video data formats such as MP4, MOV, FLV and AVI.

- To widen the scope of usability, it is recommended to ensure compatibility with webcams, as these are devices commonly owned by users.

- It is recommended to utilize UHD cameras to capture subtle movements and expressions of the human.

8.5 Multimodal output

For the human-driven 3D digital human, the multimodal output is generated by real-time rendering or offline rendering of generated animation. In addition to the requirements in clause 8.6 of ITU-T F.748.27, the requirements for multimodal output are follows.

- It is required to support the integration of the 3D digital human with virtual backgrounds, ensuring consistent rendering and smooth transitions between the 3D digital human and the virtual environment.

- It is recommended to support switching between different virtual environments, allowing for dynamic changes in the background based on the requirements of the live streaming or virtual event.

- It is recommended to synchronize the 3D digital human's movements and facial expressions with audio, text, and other modalities to create a coherent experience.

- It can optionally support viewpoint switching, allowing for different angles of interaction with user and the 3D digital avatar.

Appendix I

Use Cases of Human-driven 3D Digital Human Application System Based on Image Recognition Algorithm

I.1 Webcaster in the metaverse

The human-driven 3D digital human can be used in a live streaming scenario, acting as the webcaster in the metaverse. The webcaster's movements and facial expressions are captured, driving the 3D digital human to replicate them in real time. Furthermore, it supports voice conversion, enhancing the overall live streaming experience. This meets the requirements of webcasters by utilizing virtual avatars in their live streaming, so optimizing the interactive experience, and safeguarding their PII. This allows them to act as hosts and engage in real-time interactions with the viewers.

I.1.1. Description

In live streaming scenarios in the metaverse, as shown in Figure I.1, the user (webcaster) can utilize the human-driven 3D digital human system to transform their own video stream into 3D digital human animation, completing their webcast tasks. This system can offers a variety of digital human options, allowing them to select their preferred 3D digital human and voice style from the system's 3D digital human library and voice library for personalized customization, and interact with the viewers.

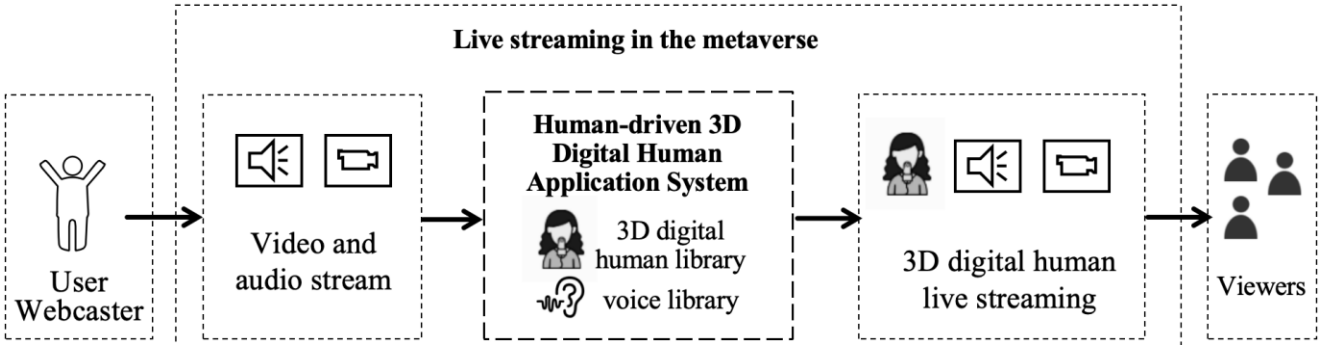

Figure I.1 – Human-driven 3D digital human for live streaming

During the live streaming in the metaverse, the system can utilize image recognition algorithms to capture the user's movements and expressions in real-time, driving the 3D digital human's actions accordingly. At the same time, it captures the user's voice and converts it into the specified voice style. This completes the transformation of the webcaster from a real person into a 3D digital human. Viewers in the live streaming can interact with the 3D digital human in real-time, which provides an immersive interactive experience.

By leveraging the capabilities of the human-driven 3D digital human, webcasters are able to overcome the limitations of their physical appearance. This enables them to showcase their talents, share valuable knowledge, promote products, and engage more effectively with their viewers. Additionally, the employment of the 3D digital human guarantees the protection of the webcaster's personal image privacy.

I.1.2. Assumptions

The assumptions related to this use case include the following;

– It is assumed that the user has video and audio input devices, such as webcams and microphone.

– It is assumed that the user has a terminal to interact with the system, such as a smartphone or computer.
– It is assumed that the user has power supply and network connection environment to ensure the normal operation of the system and the reliability of data transmission.

I.1.3. Service scenario

Figure I.2 shows the service flow for human-driven 3D digital human application system for live streaming.

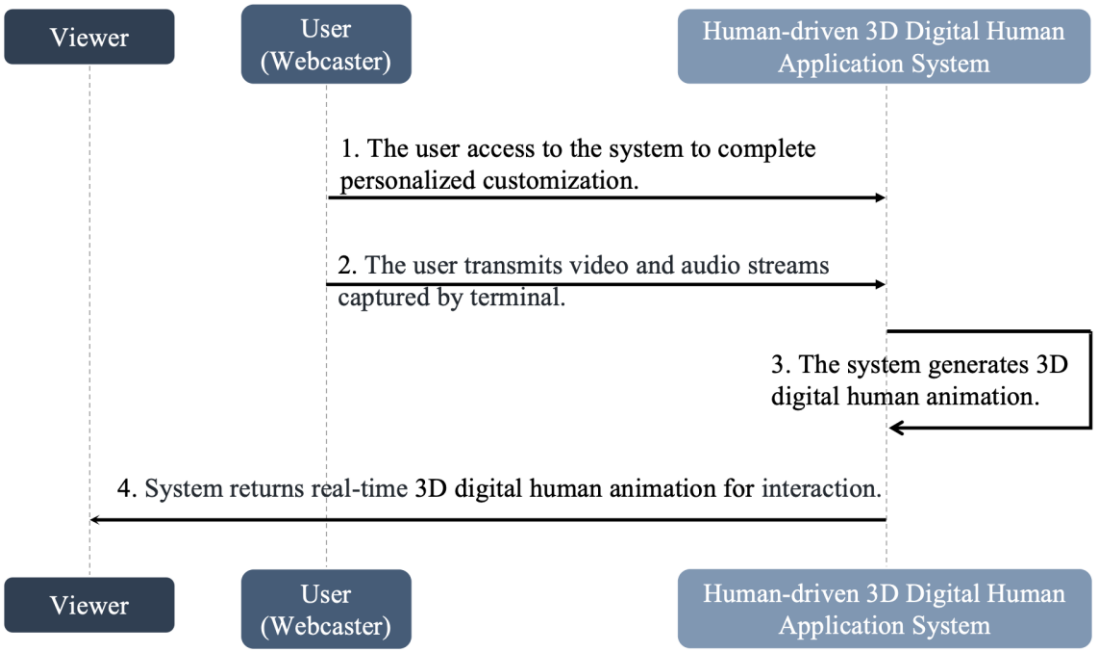

Figure I.2 – Service flow for human-driven 3D digital human for live streaming

1. The user accesses the system to complete the personalized customization. Through the smart terminal, the user (webcaster) browses the 3D digital human library provided by the system, selects their preferred 3D digital human, and chooses a voice style that matches the live streaming style, or just uses their own voice.
2. The user transmits video and audio streams captured by terminal. When the user starts the live streaming, the terminal captures the video and audio of the user by webcams and microphone, transmitting them to the system.
3. The system generates 3D digital human animation. The system receives the user's movements, facial expressions in real-time from the video stream data, as well as the user's voice from the audio stream data. Subsequently, it drives the corresponding 3D digital human based on the captured data, achieving real-time action synchronization of the 3D digital human and user. If the user chooses voice conversion, the captured audio is also converted into the corresponding voice style to maintain synchronization with the webcaster's voice.

NOTE - 2 and 3 are occurring at the same time.

4. The system returns real-time 3D digital human animation for interaction. The system transmits the generated animation to viewers. When viewers interact with the webcaster, the webcaster can give a reaction with another round of the driven process.

Bibliography

[b-Metahuman] MetaHuman documentation. Available at: https://dev.epicgames.com/documentation/zh-CN/metahuman/metahuman-documentation