ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Recommendation

ITU-T F.748.27 (01/2024)

SERIES F: Non-telephone telecommunication services

Multimedia services

Framework and requirements for the construction of 3D intelligent driven digital human application systems

ITU-T F-SERIES RECOMMENDATIONS

Non-telephone telecommunication services

TELEGRAPH SERVICE	F.1-F.109
Operating methods for the international public telegram service	F.1-F.19
The gentex network	F.20-F.29
Message switching	F.30-F.39
The international telemessage service	F.40-F.58
The international telex service	F.59-F.89
Statistics and publications on international telegraph services	F.90-F.99
Scheduled and leased communication services	F.100-F.104
Phototelegraph service	F.105-F.109
MOBILE SERVICE	F.110-F.159
Mobile services and multidestination satellite services	F.110-F.159
TELEMATIC SERVICES	F.160-F.399
Public facsimile service	F.160-F.199
Teletex service	F.200-F.299
Videotex service	F.300-F.349
General provisions for telematic services	F.350-F.399
MESSAGE HANDLING SERVICES	F.400-F.499
DIRECTORY SERVICES	F.500-F.549
DOCUMENT COMMUNICATION	F.550-F.599
Document communication	F.550-F.579
Programming communication interfaces	F.580-F.599
DATA TRANSMISSION SERVICES	F.600-F.699
MULTIMEDIA SERVICES	**F.700-F.799**
ISDN SERVICES	F.800-F.849
UNIVERSAL PERSONAL TELECOMMUNICATION	F.850-F.899
ACCESSIBILITY AND HUMAN FACTORS	F.900-F.999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T F.748.27

Framework and requirements for the construction of 3D intelligent driven digital human application systems

Summary

Recommendation ITU-T F.748.27 outlines the framework and requirements for the construction of three dimensional (3D) intelligent driven digital human application systems. With the advancement of modelling, driving, rendering and interactive technologies, an increasing number of new services and applications involving 3D intelligent driven digital humans are emerging. It defines the concept, related terms, and fundamental functions of 3D intelligent driven digital human to specify the framework of 3D intelligent driven digital human application systems, including image generation, speech generation, animation generation, interaction processing, multimodal input and output modules with its specified functions and construction requirements. In addition, the Appendix presents some use cases of the workflow of 3D intelligent driven digital human.

History[*]

Edition	Recommendation	Approval	Study Group	Unique ID
1.0	ITU-T F.748.27	2024-01-13	16	11.1002/1000/15761

Keywords

3D intelligent driven digital human, animation generation module, image generation, interactive processing module, multimodal input module, multimodal output module, speech generation module.

[*] To access the Recommendation, type the URL https://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Page

Recommendation ITU-T F.748.27

Framework and requirements for the construction of 3D intelligent driven digital human application systems

1 Scope

This Recommendation specifies the framework and requirements for the construction of three dimensional (3D) intelligent driven digital human application systems.

The scope of this Recommendation includes:

– Overview of 3D intelligent driven digital human application systems;

– Framework of 3D intelligent driven digital human application systems;

– Requirements of 3D intelligent driven digital human application systems.

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 systems*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 application system [b-ISO/IEC TR 10032]: A collection of application processes which utilizes the services provided by the human-computer interface, communications facility, and data management system to perform the processing necessary to meet the requirements of the information system.

3.1.2 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.

3.1.3 intelligent driven digital human [ITU-T F.748.15]: The digital human is driven by a computer system to automatically complete a series of actions by technical means, including text-driven, audio-driven and video-driven.

3.1.4 text-to-speech synthesis (TTS) [b-ITU-T P.10]: A TTS process generates a speech signal from text codes. It is usually composed of two parts:

– a language-dependent text processing part (the high-level processing part), which generates from the character string (by reading rules, vocabulary and semantic analysis) a set of phonetic, prosodic, etc., parameters which are used by:

– an acoustical signal generating part, the synthesizer itself which generates the audible speech.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 3D intelligent driven digital human: An intelligent driven digital human whose graphic content contains information about x, y and z dimensions.

3.2.2 3D manually modelling: A modelling method based on specialized three-dimensional (3D) modelling software that manually constructs 3D geometric, material and texture information.

3.2.3 3D scanning modelling: A modelling method based on three-dimensional (3D) measurements technology that collects geometric, material, and texture information to reconstruct the object.

3.2.4 3D intelligent modelling: A modelling method based on computer vision technology for automatically reconstructing 3D geometric, material, and texture of objects based on two-dimensional images.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

3D	Three Dimensional
AAC	Advanced Audio Coding
ALAC	Apple Lossless Audio Codec
AO	Ambient Occlusion
APE	Monkey's Audio (format)
AR	Augmented Reality
CNN	Convolutional neural network
CPU	Central Processing Unit
DCC	Digital Content Creation
FBX	FilmBoX
FLAC	Free Lossless Audio Codec
GPU	Graphics Processing Unit
IK	Inverse Kinematics
LSTM	Long Short-Term Memory
MA	Maya ASCII
MP3	MPEG-1 Audio Layer 3
OBJ	Object
PBR	Physically Based Rendering
RNN	Recurrent Neural Network
VR	Virtual Reality
WAV	Waveform Audio File Format

5 Conventions

In this Recommendation:

- The keywords "**is required**" indicate a requirement which must be strictly followed 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 which is not absolutely required. Thus, this requirement needs not be present to claim conformance.

6 Overview of 3D intelligent driven digital human application systems

The 3D intelligent driven digital human is a stereoscopic anthropomorphic image driven by digital technology. It visualizes the digital identity and can replace real people for content production and simple interaction automatically. This technology has found extensive applications in various sectors, including finance, education, media, and social networking. For instance, 3D intelligent driven digital human for news broadcasting can automatically complete the production and output of news video according to a simple manuscript, which can meet the needs of video release on different platforms, greatly improving the production efficiency of media news.

As the 3D intelligent driven digital human evolves into an intelligent assistant, it can provide answers and product recommendations to customers. In addition to one-way content delivery, it can engage in two-way interactions. Therefore, 3D intelligent driven digital human is an important part of the interactive system in the future and will become one of the main forces of value production in the virtual world.

7 Framework of 3D intelligent driven digital human application systems

Based on the framework proposed in Figure 1 of [ITU-T F.748.15]. For the construction of 3D intelligent driven digital human application systems, this Recommendation proposes a framework of 3D intelligent driven digital human application systems in Figure 1. It clarifies the requirements for the construction of 3D intelligent driven digital human in terms of image generation, speech generation, animation generation, interactive processing, multimodal input and multimodal output. The coloured boxes in it indicate the contents concerned in this Recommendation. And the dashed box indicates that this section is optional.

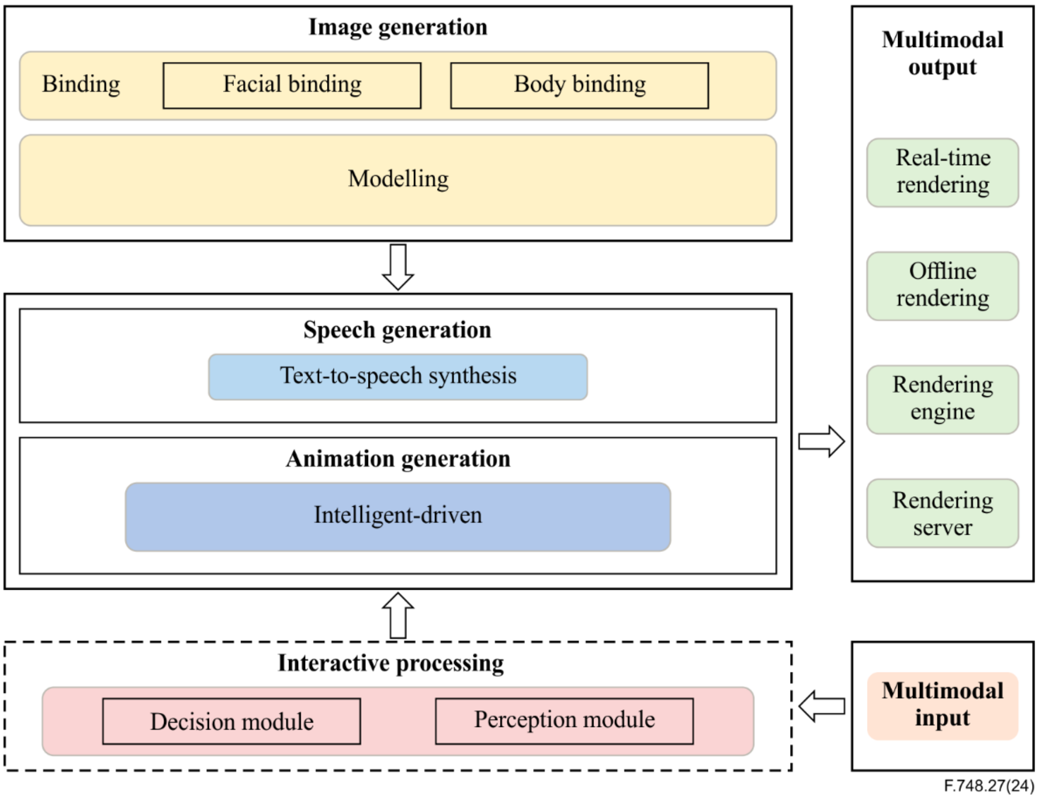

Figure 1 – Framework of 3D intelligent driven digital human application systems

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

– **Speech generation module**: A module that can enable 3D intelligent driven digital human to speak. Speech synthesis can generate speech based on the algorithms.

– **Animation generation module**: A module that generates the facial expression and body movement of a 3D intelligent driven digital human. Based on the different driving methods, animation generation modules can be divided into voice-driven, image-driven, and text-driven.

– **Interactive processing module**: Realizing the interactive ability of the 3D intelligent driven digital human to analyse multimodal input. The perception module realizes the recognition of data from the multimodal input. The decision module realizes the comprehension of user intention and generates decision-making data of 3D intelligent driven digital human. This module is optional.

– **Multimodal input module**: A module that is used to receive input, including text, voice, image, etc.

– **Multimodal output module**: A module that presents the output to the user. Including real-time rendering and offline rendering. The rendering capability of this module relies on the rendering engine and the rendering server.

8 Requirements for the construction of 3D intelligent driven digital human application systems

8.1 Image generation

8.1.1 Modelling

The modelling of 3D intelligent driven digital human should meet the following requirements:

GM-01: It is required to generate a 3D model of a digital human.

GM-02: It is required that the 3D model supports 3D visualization.

GM-03: It is required to enable the creation of various styles of 3D models, such as 3D cartoons, 3D realistic, 3D ultra-realistic, etc.

GM-04: It is required that the 3D model includes basic information of the digital human, including the head, limbs, hair, clothing, etc.

GM-05: It is required that the 3D model includes geometric, material and texture information.

GM-06: It is required to support multiple physically based rendering (PBR) materials for the 3D model, such as albedo, roughness, normal, displacement, ambient occlusion (AO), cavity, curvature, wrinkle and so on.

GM-07: It is required that the transition between the key features of the 3D model should be natural and the deformation should be realistic.

NOTE – The transition between the key features of the 3D model refers to changes between each component of a digital human, such as the head and body, the foot and leg, etc.

GM-08: It is required to provide support for generating 3D models in multiple formats, such as FilmBoX (FBX), object (OBJ), Maya ASCII (MA), etc.

GM-09: It is required for the 3D model to be compatible with various rendering engines.

GM-10: It is required for the 3D model to be compatible with various digital content creation (DCC) software.

GM-11: It is required to support different modelling methods to build 3D models, including 3D manual modelling, 3D scanning modelling, 3D intelligent modelling, etc.

8.1.2 Binding

The generated 3D model of a digital human is static, in order to drive its motion it is required to perform facial binding and body binding of it. Binding for a 3D digital human needs to follow one specification, and when a possible restriction on the binding of a 3D digital human is defined, the animations designed for one 3D digital human are deployable to another 3D digital human, see clause 4.8 of [b-ISO/IEC 19774-1].

The facial binding should meet the following requirements:

FB-01: It is required to support the binding of the facial bones and facial key feature points of the 3D digital human model.

FB-02: It is required to support the splitting and production of the corresponding expressions of the 3D digital human model.

FB-03: It is required to support the corresponding construction of the facial control system in the later stage.

FB-04: It is required to support real-time driving and manual correction.

Body binding should meet the following requirements:

BB-01: It is required to support the construction of the human skeleton of the 3D digital human model.

BB-02: It is required to support the binding of 12 joint points of the human body and 5 key points of the head of the 3D digital human model.

BB-03: It is required to support the creation of the control system of each part of the human body of the 3D digital human model.

BB-04: It is required to support the skin painting of the 3D digital human model.

BB-05: It is required to support the muscle modification or muscle and skeleton system construction of the 3D digital human model.

BB-06: It is recommended to support the loading of the full-body human inverse kinematics (IK) skeleton.

8.2 Speech generation

Speech generation is based on text-to-speech synthesis and it should meet the following requirements:

SG-01: It is required to support the synthesis of audio of different genders, such as male or female voices.

SG-02: It is required to support the synthesis of audio of different ages, such as children, youth, and elderly voices, etc.

SG-03: It is required to support the synthesis of audio with different emotions, such as neutral, sad, happy voices, etc.

SG-04: It is required to support the synthesis of the audio in multiple languages, such as Chinese, English, Japanese, Korean, Spanish, etc.

SG-05: It is required to support multiple voice styles for different scenarios, such as news, entertainment, emotional, information, and promotional scenarios.

SG-06: It is required to support voice-changing that can convert the voice of one person into a specific voice.

SG-07: It is required to support the synthesis of singing audios.

SG-08: It is recommended to support audio editing capabilities, including speed, tone, volume, and emotional adjustment.

SG-09: It is recommended to support text labels for correcting errors and improving the synthesis effect, such as supporting the editing of homophones, symbols, and other elements in the text.

SG-10: It is required to support the synthesis of audio in multiple formats to be compatible with different rendering engines, such as waveform audio file format (WAV), free lossless audio codec (FLAC), Monkey's audio (format) (APE), Apple lossless audio codec (ALAC), MPEG-1 audio layer 3 (MP3), advanced audio coding (AAC), Ogg Vorbis, etc.

NOTE – See [b-Audio formats] for a general description of common audio file formats.

SG-11: It is required to support the generation of audio time series information and segmented speech fragments while synthesizing the audio.

8.3 Animation generation

Animation generation can occur based on the user input text or in response to the text analysed from the user's input. In the latter case, when the user's input is an audio, a typical animation generation method involves the following steps: it is required to obtain the response text and feature information corresponding to the response text according to the audio input by the user. The feature information is blend shape and skeleton parameters used to indicate facial expressions and body motions of 3D digital human. Obtain the feature data corresponding to the response information based on the feature information. Then, it is required to generate dynamic images of 3D digital human corresponding to the response text according to the feature data.

NOTE – Response information is obtained from the user's audio by interactive processing in clause 8.4.

AG-01: It is required that the feature information includes at least one of the emotion information or pronunciation information.

AG-02: It is required that the feature data includes at least one of the expression data or lip-shape data. The expression data is determined according to the first blend shape data and the first skeleton data corresponding to the emotion information; the lip-shape data is obtained according to the second blend shape data and the second skeleton data corresponding to the pronunciation information.

AG-03: It is required that the skeleton data is determined according to the weighted sum of the initial skeleton data and multiple skeleton data components.

AG-04: It is required to obtain the initial weight of the feature data according to the feature information.

AG-05: It is required to randomly generate the actual weights of multiple key frames at the corresponding time within the initial weight and the value range defined according to the threshold. The value range includes values that are greater than the difference between the initial weight and the threshold and less than the sum of the initial weight and the threshold.

AG-06: It is required that generating the dynamic images of 3D digital human is based on the multiple key frames: smoothing the weighted feature data corresponding to the adjacent key frames to generate non-key frames between the adjacent key frames; the dynamic images are generated in time order according to the key frames and non-key frames and their timestamps.

AG-07: It is required to determine the timestamp of the feature data according to the feature information and to generate the actual weight of the corresponding time of the timestamp to generate the actual weight of the corresponding time of multiple key frames.

8.4 Interactive processing

8.4.1 Perception module

Perception module needs to realize the recognition of multimodal input data by 3D digital human based on the algorithm models. It supports 3D intelligent driven digital human's perception of the user's input.

PM-01: It is required to support the recognition of at least one data type, such as text, audio, video, touch/sensing data, etc.

PM-02: It is required to support at least one type of recognition capability, such as text recognition, speech recognition, image recognition, etc.

PM-03: It is recommended to support text recognition, such as text intention recognition, semantic recognition, etc.

PM-04: It is recommended to support speech recognition, such as speech-to-text recognition, voiceprint recognition, age and gender recognition, etc. In some cases, it is required to support specific person speech recognition, non-specific person speech recognition, multi-person speech recognition, etc.

PM-05: It is recommended to support image recognition, such as face recognition, gesture recognition, posture recognition, facial emotion recognition, scene recognition, place recognition, etc. In some cases, it is required to support specific person image recognition, non-specific person image recognition, multi-person image recognition, etc.

PM-06: It is required to support specific types of 3D digital human image editing. On a limited set, according to the text commands, 3D digital human support text commands are to be editable in dimensions such as hair colour, skin colour and hairstyle.

PM-07: It is required to support specific types of 3D digital human action and expression driving. On a limited set, according to the text instructions, audio instructions or video instructions, the 3D digital human can respond to actions and expressions that conform to the instructions.

PM-08: It is required to support algorithm models and model structures for various tasks, such as convolutional neural network (CNN), long short-term memory (LSTM), recurrent neural network (RNN), etc.

PM-09: It is required a high recognition accuracy rate.

PM-10: It is required that the algorithm models can update iteratively.

PM-11: It is required for the algorithm models to be robust against abnormal data.

PM-12: It is required for the algorithm models to be adapted to different devices, such as servers, hosts, terminal devices, etc.

8.4.2 Decision module

Decision-making module needs to realize the analysis of perception data by 3D digital human based on the AI model. It includes the following requirements:

DM-01: It is required to support the knowledge matching ability of the 3D digital human, such as natural language processing, knowledge graph, dialogue management, big data search, recommendation engine, etc.

DM-02: It is required to support the generation of response text based on intent understanding and business knowledge matching.

DM-03: It is required to support algorithm models and model structures for various tasks, such as CNN, LSTM, RNN, etc.

DM-04: It is required to support multi-decision algorithms that integrate intent understanding and business knowledge matching to realize linkage decision-making.

DM-05: It is required a high model recognition accuracy rate.

DM-06: It is required that the model can update iteratively.

DM-07: It is required for the model to be robust against erroneous data.

DM-08: It is required for the model to be adapted to different devices, such as servers, hosts, terminal devices, etc.

8.5 Multimodal input

The input data should meet the following requirements:

MI-01: It is required to support at least one of the input data types, such as text, audio, image, etc.

MI-02: It is required to support multiple types of input text, such as entered text, imported text files, text converted by third-party systems, etc.

MI-03: It is required to support multiple types of input audio, such as audio recorded by real people, synthetic audio, audio that has been processed by voice change, etc.

MI-04: It is required to support multiple types of input video, such as recorded video, live video streams, etc.

MI-05: It is required to support multiple data formats, such as structured data, semi-structured data, and unstructured data.

MI-06: It is recommended to support touch/sensing data input and other types of data input.

8.6 Multimodal output

8.6.1 Real-time rendering

Real-time rendering should meet the following requirements:

RR-01: It is required to support the rendering of the images while performing the 3D image calculation.

RR-02: It is required to support multi-resolution rendering of 3D digital human.

RR-03: It is required to support the rendering of fusion materials, lighting and physical characteristics.

RR-04: It is required to support the transmission of rendering results in the form of video streams.

RR-05: It is required to support the rendering quality output of multiple resolutions.

RR-06: It is recommended to support real-time adjustment of resolution.

RR-07: It is required to have real-time control and real-time interaction capabilities, and low real-time feedback delay of the screen.

RR-08: It is required to support real-time face control, real-time body control, real-time hair control and other functions.

RR-09: It is recommended to support multi-channel concurrent rendering and can control multiple users in real-time.

RR-10: It is recommended to support compatibility with a variety of terminal devices, such as flat panel display devices such as mobile phones, tablets, and computers, virtual reality (VR) / augmented reality (AR) devices and holographic projection devices, etc.

8.6.2 Offline rendering

Offline rendering should meet the following requirements:

OR-01: It is required to support the ability to edit 3D digital human images and voices, such as designing and adjusting their appearance, clothing, and actions according to the application scene to achieve better visual effects.

OR-02: It is required to support ultra-high-precision rendering of 3D digital humans.

OR-03: It is required to support fine rendering that integrates materials, lighting and physical characteristics.

OR-04: It is required to support the rendering quality output of multiple resolutions.

OR-05: It is required to support compatibility with a variety of terminal devices, such as mobile phones, tablets, computers and other flat panel display devices, VR devices, and holographic projection devices.

8.6.3 Rendering engine

For the intelligent driving of 3D digital human, the rendering engine should meet the following requirements:

RE-01: It is required to support the import of commonly used 3D digital human model file formats, such as object (OBJ), FilmBoX (FBX), Maya ASCII (MA), etc.

NOTE – See [b-3D formats] for a general description of common 3D file formats.

RE-02: It is required to support the adjustment of 3D digital human models, such as face, skin, hair, body, etc.

RE-03: It is required to support not less than one kind of operating system, such as Windows, Linux, etc.

RE-04: It is required to support offline and real-time rendering.

RE-05: It is required to support central processing unit (CPU) / graphics processing unit (GPU) rendering.

RE-06: It is required to support cloud rendering.

RE-07: It is required to support end-to-end rendering.

RE-08: It is recommended to support GPU rendering acceleration.

RE-09: It is recommended to support new rendering technologies such as pixel streaming and ray tracing.

8.6.4 Rendering server

Rendering server is utilized to deliver the rendering engine ensuring high-performance services to meet the rendering requirements of high-precision image quality in 3D intelligent driven digital human.

RS-01: It is recommended to use a high-performance server with 64-bit system architecture and multiple CPUs/GPUs.

RS-02: It is recommended to adopt the cluster deployment method and cooperate with the load balancing mechanism to achieve high availability of the rendering service.

RS-03: It is recommended to have scalability, and scale performance smoothly by adding server instances.

Appendix I

Use cases of workflow of 3D intelligent driven digital human application systems

(This appendix does not form an integral part of this Recommendation.)

I.1 Use case of image generation process for 3D intelligent driven digital human application systems

Figure I.1 depicts the image generation process for 3D intelligent driven digital human application systems.

– The first step is to generate a 3D model of an intelligent driven digital human. The modelling methods include 3D manually modelling, 3D scanning modelling and 3D intelligent modelling, etc. The 3D manually modelling uses modelling software to generate the 3D model. For 3D scanning modelling, the generation of the 3D model relies on a scanning modelling system to capture facial and body data. For 3D intelligent modelling, it realizes the construction of 3D models based on single frame images or video sequences.

– The second step is to perform character binding on the built 3D model, including the body binding and facial binding.

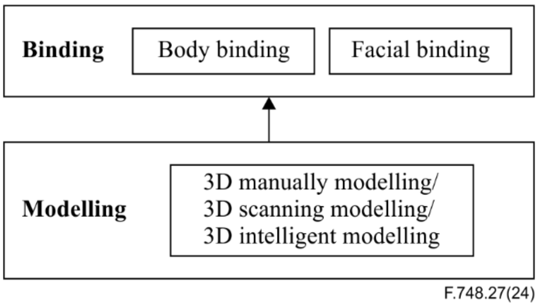

Figure I.1 – Image generation process for 3D intelligent driven digital human application systems

I.2 Use case of 3D intelligent driven digital human application systems for news broadcasting

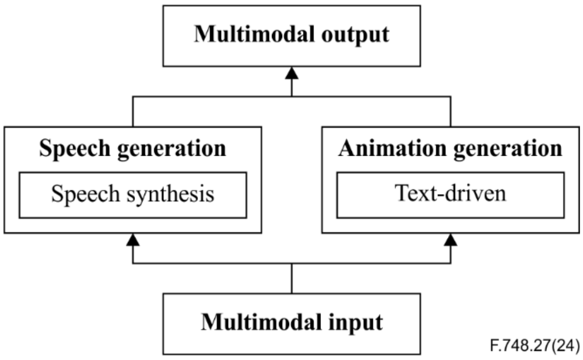

Figure I.2 – Workflow of 3D intelligent driven digital human application systems for news broadcasting

3D intelligent driven digital human application systems for news broadcasting need to automatically complete content production based on the simple text input.

Figure I.2 depicts the workflow of 3D intelligent driven digital human application systems for news broadcasting.

- First, input a text to the multimodal input module.
- Second, the speech generation module and animation module works based on the text data from the multimodal input module to generate corresponding speech and animation of 3D intelligent driven digital human.
- Finally, the multimodal output module does the synthesis and output of the speech and animation.

I.3 Use case of 3D intelligent driven digital human application systems for intelligent assistant

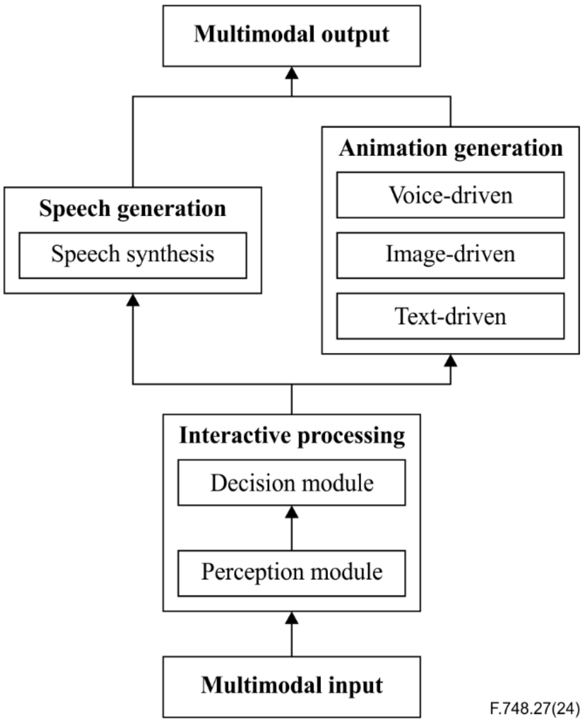

Figure I.3 – Workflow of 3D intelligent driven digital human application systems for intelligent assistant

3D intelligent driven digital human application systems for intelligent assistant should have the ability to automatically interact with users. For example, providing consulting, data query, business handling and other services.

Figure I.3 depicts the workflow of 3D intelligent driven digital human application systems for intelligent assistant.

- First, the multimodal input module allows multiple data types, such as voice, image, text, somatosensory data, touch data, etc.
- Second, the intelligent driven model of the interactive processing module enables the 3D intelligent driven digital human to analyse external input and generate decision data for it.
- Third, the animation generation module and speech generation module work according to the decision data. Among them, according to the different input types, animation generation can be divided into speech driven, image driven, text driven, and speech generation needs to rely on the speech synthesis technology. At the same time, the consistency of the speech and lips of the 3D intelligent driven digital human should be maintained.
- Finally, speech and animation are synthesized and rendered in real time.

Bibliography

[b-ITU-T P.10] Recommendation ITU-T P.10/G.100 (2017), *Vocabulary for performance, quality of service and quality of experience*.

[b-ISO/IEC 19774-1] ISO/IEC 19774-1:2019, *Information technology – Computer graphics, image processing and environmental data representation – Part 1: Humanoid animation (HAnim) architecture*. <https://www.iso.org/standard/64788.html>

[b-ISO/IEC TR 10032] ISO/IEC TR 10032:2003, *Information technology – Reference model of data management*. <https://www.iso.org/standard/38607.html#:~:text=ISO%2FIEC%20TR%2010032%3A2003%20defines%20the%20ISO%20Reference%20Model,persistent%20data%20in%20information%20systems.>

[b-3D formats] Wikipedia, *List of file formats, 3D graphics*. <https://en.wikipedia.org/wiki/List_of_file_formats#3D_graphics>

[b-Audio formats] Wikipedia, *Audio file format*. <https://en.wikipedia.org/wiki/Audio_file_format>

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	**Non-telephone telecommunication services**
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Published in Switzerland
Geneva, 2024