International Telecommunication Union

ITU-T G.1036

(07/2022)

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

SERIES G: TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS

Multimedia Quality of Service and performance – Generic and user-related aspects

Quality of experience influencing factors for augmented reality services

Recommendation ITU-T G.1036

ITU-T G-SERIES RECOMMENDATIONS

TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS

INTERNATIONAL TELEPHONE CONNECTIONS AND CIRCUITS	G.100–G.199
GENERAL CHARACTERISTICS COMMON TO ALL ANALOGUE CARRIER-TRANSMISSION SYSTEMS	G.200–G.299
INDIVIDUAL CHARACTERISTICS OF INTERNATIONAL CARRIER TELEPHONE SYSTEMS ON METALLIC LINES	G.300–G.399
GENERAL CHARACTERISTICS OF INTERNATIONAL CARRIER TELEPHONE SYSTEMS ON RADIO-RELAY OR SATELLITE LINKS AND INTERCONNECTION WITH METALLIC LINES	G.400–G.449
COORDINATION OF RADIOTELEPHONY AND LINE TELEPHONY	G.450–G.499
TRANSMISSION MEDIA AND OPTICAL SYSTEMS CHARACTERISTICS	G.600–G.699
DIGITAL TERMINAL EQUIPMENTS	G.700–G.799
DIGITAL NETWORKS	G.800–G.899
DIGITAL SECTIONS AND DIGITAL LINE SYSTEM	G.900–G.999
MULTIMEDIA QUALITY OF SERVICE AND PERFORMANCE – GENERIC AND USER-RELATED ASPECTS	**G.1000–G.1999**
TRANSMISSION MEDIA CHARACTERISTICS	G.6000–G.6999
DATA OVER TRANSPORT – GENERIC ASPECTS	G.7000–G.7999
PACKET OVER TRANSPORT ASPECTS	G.8000–G.8999
ACCESS NETWORKS	G.9000–G.9999

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

Recommendation ITU-T G.1036

Quality of experience influencing factors for augmented reality services

Summary

Recommendation ITU-T G.1036 lists typical use cases of augmented reality (AR) services and identifies the key quality of experience (QoE) factors within them, and also gives a suggested scheme for AR QoE assessment in future work.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T G.1036	2022-07-29	12	11.1002/1000/15009

Keywords

Augmented reality, QoE, quality of experience.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

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

Recommendation ITU-T G.1036

Quality of experience influencing factors for augmented reality services

1 Scope

This Recommendation describes two main types of augmented reality (AR): face AR and space AR. Face AR refers to a service that superimposes the AR effect in and around the human face. Space AR refers to a service in which an AR effect is superimposed on a natural space, such as a tabletop, a wall surface, a common object, an indoor space or an outdoor open space.

In order to measure the quality of experience (QoE) that users perceive in face and space AR services, benchmarking is critical. Due to addition of new ways to locate the self-position of users and a new display mode of perceptual information, a set of new requirements for QoE assessment presents itself. The challenge is to characterize immersive video, spatial audio and interactivity of AR. It is also important to address the requirements and basic factors affecting virtual reality (VR) QoE before work is benchmarked. This Recommendation identifies the taxonomy of the two AR services and key factors affecting user-perceived experience.

This Recommendation includes:

– face and space AR service taxonomy and corresponding use cases;

– QoE-influencing factors for face and space AR services.

AR is an immersive experience of a real-world environment in which objects within it are "augmented" by computer-generated perceptual information, e.g., one or several visual, auditory, haptic, somatosensory and olfactory factors can be included. AR differs from VR, which aims to completely replace the user's real world environment with a simulated one; AR keeps the real world and adds information to the natural environment. The value of AR is to integrate immersive sensations into the natural environment, thus deepening perception when compared to actual reality. The first AR application systems were used for military training; subsequently, its functional scope has been expanded to gaming, education, navigation, etc.

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.

None.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 quality of experience (QoE) [b-ITU-T P.10]: The degree of delight or annoyance of the user of an application or service.

3.1.2 QoE influencing factors [b-ITU-T P.10]: Include the type and characteristics of the application or service, context of use, the user's expectations with respect to the application or service and their fulfilment, the user's cultural background, socio-economic issues, psychological profiles, emotional state of the user, and other factors whose number will likely expand with further research.

3.1.3 cybersickness or **simulator sickness** [b-ITU-T G.1035]: A physiological condition arising when exposed to a virtual reality environment.

NOTE – Definition based on combined information in [b-Kennedy] and [b-Stanney].

3.1.4 degree of freedom (DoF) [b.ITU-T G.1035]: Represents the ways an object can move within a space, which is a key element in helping create an immersive environment for a user.

3.1.5 immersion [b-ITU-T G.1035]: A psychological state characterized by perceiving oneself to be enveloped by, included in and interacting with an environment that provides a continuous stream of stimuli and experiences.

NOTE – Definition based on [b-Witmer].

3.1.6 motion-to-photon latency [b-ITU-T G.1035]: The time it takes between the user moving their head and this motion being reflected on the screen of the head-mounted display (HMD).

NOTE – Definition based on [b-Brandenburg].

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 geometric consistency: The degree of similarity of registration and tracking, to ensure that virtual content has correct locating relationship with the real environment as time changes.

3.2.2 illumination consistency: The degree of similarity of texture and lighting being simulated by augmented reality devices to those of the surrounding environment.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

3D three Dimensional

A/V Audio and Video

AR Augmented Reality

BCI Brain-Computer Interface

CG Computer Graphics

DoF Degree of Freedom

FoV Field of View

HCI Human-Computer Interaction

HDR High Dynamic Range

HMD Head-Mounted Display

ID Identifier

OST Optical See-Through

QR Quick Response

SAR Spatial Augmented Reality

SDR Standard Dynamic Range

VP*n* Video Payload type *n*

VR Virtual Reality

VST Video See-Through

5 Conventions

Within the scope of this recommendation, AR services are sorted into three kinds, namely: optical see-through (OST); video see-through (VST); and spatial augmented reality (SAR), mainly depending on their difference in display method. People using and experiencing these AR services are referred to as users or players.

6 Augmented reality system configuration and taxonomy

6.1 Overview

To achieve an ideal and realistic AR effect, an AR service application system is usually composed of the support of AR mobile device, network and platform(s). It should be noted that at the current stage, the AR service locally processed on a device occupies a large part of the form, and in such applications, network and platform support may not be needed [b-3GPP TR 26.998].

AR device: AR glasses, smart phone, model acquisition device (depth camera, camera group, infrared transceiver), interactive facilities.

Network: (low-delay) wireless local area network, fourth or fifth generation network.

AR platform: content authoring platform, audio and video (A/V) stream-forwarding platform, content recognition platform, A/V rendering platform, geographic model storage platform.

6.2 Typical augmented reality service use cases

There are two typical AR services: face and space based. A face AR service mainly processes and displays augmented information on the human face, while a space-based AR service is suitable for other objects and spaces remaining.

6.2.1 Face augmented reality service

There are two main types of face AR service: user local and remote call. User local services are designed for the user to experience AR effects in real time on the capture device itself, while the processing workflow is completed inside the device in most cases. Remote call services deliver

captured content to distant user device(s) through a network and platform, and the processing workflow can be executed either inside the device or on the platform.

6.2.2 Space-based augmented reality service

The development of a space-based AR service can be divided into three stages: marker-given or plane-based AR service; indoor room scale AR service: and outdoor environment scale AR service. The first requires the ability to detect a known marker like quick response (QR) code or a horizontal plane. The second requires the service to recognize all objects in a single room and obtain their surfaces. While in the most advanced last stage, the AR service works pretty well and puts virtual parts in outdoor space, which could be quite large.

6.3 Technical classification

AR service implementation technology can be divided into three categories, mainly by display method: VST AR, OST AR and SAR.

6.3.1 Video see-through augmented reality service

A VST AR service captures real world images via a camera system and displays both images and virtual parts integrated on one screen, e.g., those mounted on smartphones or tablets. Users usually handle the VST device manually and experience the service in a similar way to using common applications.

The most widely used applications of this type are: face AR services based on data captured by a front camera on mobile smart devices; and daily or gaming applications based on data captured by a rear camera on the same devices.

6.3.2 Optical see-through augmented reality service

An OST AR service adds virtual parts directly to a real environment without taking images of it, usually achieved by the use of an HMD. At the time of publication, this type of HMD mainly uses optical material produced by means of optical waveguide or free-form surface technology, and the experience of superimposing into real space is realized by displaying the picture in a transparent lens. There are also some schemes that use laser projection to achieve imaging on the retina directly.

This type of service is used for commercial purposes in most cases.

6.3.3 Spatial augmented reality service

SAR is a technology that utilizes displays built into projectors or real-world environments. It directly projects virtual content on to a real environment without the use of an imaging device on the user side.

This type of AR application allows multiple users to view the same AR effect at the same time. In recent years, projection mapping technology, which attracts a lot of attention, is one of its representative forms.

7 Augmented reality quality of experience influence factors

Some factors influencing the AR user experience are similar to those in other multimedia, such as video and VR services. Clauses 7.1 to 7.3 focus on specific factors influencing the experience of AR services. Clause 7.4 briefly introduces other general factors.

7.1 Human factors

Human factors can be understood as the interdisciplinary field covering many aspects of the study of human-computer interaction (HCI) technologies [b-Puig]. Human factors impact end user QoE in the aspects described in clauses 7.1.1 to 7.1.3.

7.1.1 Vision

Human visual functions and characteristics differ individually. Various visual impairments, e.g., myopia, hyperopia, astigmatism and amblyopia, may exist in a user's visual system to varying degrees, and the left and right eyes of a single individual may also exhibit symptoms of different types and degrees. OST AR display devices need to consider the type of visual impairment that can affect visual perception, by using their own hardware configuration to provide adjustment capabilities to compensate for visual impairment.

7.1.2 Hearing

Hearing impairment may affect one or both ears and can occur anywhere in the auditory pathway. As a result, the audio perceptibility of a certain frequency range or full frequency band is reduced. The user experience of AR service might then be degraded.

The head-related transfer function can simulate the spatial position of sound in AR service, but the function is highly individualized and requires restoration for each user to provide a better listening experience.

7.1.3 Tactile or force sensation

The initial stage AR system focuses on visual and auditory enhancements, while advanced AR experiences can include tactile or force and other types of perceptual enhancement.

There are also differences in tactile or force sensations for different individuals. Since the tactile nerves are insufficiently coordinated with the external environment, people may experience tactile disorders. Tactile dysfunction can manifest as hypersensitivity or slow tactile function, and the experience of tactile or force system feedback can be excessive or slow. AR somatosensory enhancement devices can provide corrections and compensation for such problems.

AR tactile or force enhancement is currently manifested in two forms of hand perception enhancement and body perception enhancement. Human hand and body disability may lead to the loss of some or all of the perception ability of a device.

7.2 System influence factors

7.2.1 Content related

AR content consists of two parts: the real environment; and the virtual content. In VST AR, the real environment is multimedia content acquired by sensors, e.g., video cameras and microphones; in OST and SAR, the real environment is directly seen by the user through the lens or without any obstruction, and real-world audio produced by the environment is perceived.

Virtual content is generally stored in a computer-specific format, and transmitted over a network. The virtual audio content is generated by computer simulation software and played using an AR HMD or dedicated sound field generation device.

Real environment

The VST AR service needs to use the camera and microphone to collect the environment. Recording quality, of the camera for video and of the microphone for audio, has a great influence on the experience. Real ambient light in OST AR passes directly through the lens to the user's eyes. Optical quality of the lens, such as transparency and distortion, affects the appearance of the real environment.

Depth information of a real environment plays an important role in AR service. When collecting environmental information, the device needs to obtain the depth value of the spatial point by using methods such as triangulation, structured light measurement or optical time measurement.

Virtual content

There are two ways to generate virtual content; one is drawing virtual objects and a virtual environment model by computer graphics (CG); and the other is to establish a model of real objects and real environment by using capture device.

The quality of virtual content drawn by CG is mainly determined by the parameter settings at the time of rendering and the performance of the rendering engine. The number of triangles and vertices, materials, and lighting reflection of the virtual model are used as the main user experience evaluation parameters.

Virtual content obtained through device detection can be acquired actively or passively. Active acquisition techniques include structured-light and time-of-flight measurement. Passive acquisition techniques include multiview stereo vision, motion structure restoration and beam adjustment. Different detection methods are applicable to different distances. The structured-light method is currently used for close range acquisition, such as face shape, the time-of-flight method is mostly used for environment detection at a meter level distance, while the passive acquisition method does not need to rely on additional devices except video capture camera and inertial measurement unit, which is more versatile. The quality differences in such virtual content collection is mainly reflected in the density of data points and the reproduction of the real scene. It is generally believed that the higher density and higher degree of reproduction for virtual content provides a higher level user experience.

Superimposition of real environment and virtual content

The primary value of AR is to promote the fusion of real-world environments and virtual content; however, these two layers of virtual and real scenes will influence each other and the fusion of these two components will influence the user's experience.

The primary theory underlying AR is human visual confusion, which allows users to perceive the real-world scenes and augmented content (virtual-world scenes) simultaneously by superimposing them together. The superimposition of real-world scenes and AR content may cause visual confusion, and may further affect the perceptual quality of both views. Thus, the QoE evaluation of AR services, especially OST AR services, is recommended to consider both AR content and real-world scenes together, since the superimposition of two views may cause visual confusion and various see-through environments may bring different QoE [b-Duan 2022a]. More specifically, the QoE of OST AR services should consider both the perceptual quality of superimposed scenes [b-Duan 2022a] and human visual attention behaviour when viewing such scenes [b-Duan 2022b].

7.2.2 Media and coding related

Codec

The A/V codecs are responsible for compressing and encoding the originally acquired data, removing the redundancy inherent in the data itself. Codecs commonly used in video compression in the industry include advanced video coding, high-efficiency video coding, video payload type 8 (VP8) and VP9. Moving Pictures Experts Group group I (MPEG-I) is developing a standard point cloud compression version 2, which can be used for depth information compression in AR scenarios in the future. The industry's effort on compression for three dimensional (3D) audio is currently underway, and MPEG-H is developing a 3D standard that can be applied to spatial audio compression in AR scenarios in the future. The metrics such as compression ratio and delay of different codec libraries will be different, which will ultimately affect the media quality experience perceived by users.

Bit rate

Bit rate is the number of binary bits processed per second of the media data, which is determined by the settings of the encoder. The bit rate of the media content can be changed by adjusting the parameter configuration of the encoder. If the original media content source is determined, the bit rate value is usually positively correlated with the media quality and can be used as a key indicator of user experience evaluation.

Resolution

The resolution of a video refers to how many pixels are in the video picture of each frame. At present, the resolution of the AR video is similar to that of the conventional video, and is about 720 px to 2 kpx in the case of a matching playback device. Increasing the video resolution can improve the clarity of the picture and give the user a better experience when watching AR video.

Frame rate

The frame rate refers to how many pictures are received by the human eye in 1 s. The frame rate of the AR video content should match the display frame rate of the AR display device. In VST AR, the frame rate matches the frame rate of the mobile device camera acquisition and screen display. In OST AR, hardware features allow a relatively higher frame rate, such as 60 frames/s, to resist display hysteresis and smear that may occur in virtual parts during head-on motion.

Coding delay

When the codec is configured to perform encoding processing via different profiles, different degrees of coding delay are introduced according to the complexity of the multimedia content. AR real-time services often require ultra-low overall latency. Perceptual thresholds exist for television broadcasting (e.g., [b-ITU-T J.248]), and delay perception thresholds for AR content are for further study.

Streaming

Augmented content is transmitted to end users by streaming media, whose quality can significantly influence the QoE of AR services. Specifically, the statistics of some buffering and bit rate-changing behaviours are usually useful for QoE measurement of streaming media. For example, initial buffering (time or percentage), stalling (count or percentage), bit rate switching (count or magnitude), as well as bit rate (average) itself, are widely recognized as major factors influencing streaming QoE [b- Seufert].

Storage and transport

For AR digital content, it is necessary to propose a more efficient way to reduce bandwidth and resource requirements compared with traditional storage and transmission methods. Research into point cloud coding and transmission may play a new role in processing AR content [b-Li].

7.2.3 Network and transmission related

Bandwidth

The bandwidth required by AR services is related to service content types. When the content transmitted is a pure media stream, the bandwidth required is the same as for traditional media content at the same quality level. When the content being transferred is a virtual component, the bandwidth depends on the size of the virtual content. When the transmitted content is that of the entire environment with 3D depth information to provide a 6 DoF experience, the bandwidth requirement increases dramatically.

Network congestion can lead to bandwidth fluctuations, packet loss and high latency. AR content that cannot be transmitted in a timely fashion can reduce immersion when users experience the system.

Sufficient network bandwidth guarantees the transmission of virtual content that improves user experience [b-Cao][b-Cui].

Latency

In AR services that need to transmit real-time content over a network, low latency is highly demanding. Real-time content may include that from captured video, audio, virtual and interactive device parameters. Network delays include those arising from queuing and over-the-air, which play a bottleneck role in real-time AR services.

High network latency may result in a matching time lag between the virtual content and the real environment or an increase in the response time of interactive operation on virtual content. This may result in the user experiencing dizziness, nausea and other discomforts.

Especially, for those AR applications on a mobile device, tracking performance decreases when network delay occurs [b-Cui].

Packet loss

The impact of network packet loss on AR service is related to the network transmission protocol used. In an AR content transmission service with a retransmission mechanism, packet loss increases the end-to-end transmission delay of the network. In services without these mechanisms, the video, audio, virtual content and interactive information carried by the network may be missing; furthermore, the presentation of multimedia and interactive content may be reduced. For AR applications on a mobile device, wireless channel conditions that change over time require consideration when evaluating the QoE [b-Cui].

Jitter

This term describes the variation in timing of a picture caused as packets are received, buffered and distributed to the screen as the available bandwidth changes. It is recommended that jitter be maintained under a few milliseconds for uninterrupted online AR service. An increase in jitter caused by an underpowered network connection can cause skipping or freezing of a picture, resulting in noticeable disruptions.

7.2.4 Recognition related

Marker attributes

Markers are used for registration between a virtual component and the real environment in AR service. They are used to obtain the identifier (ID) number to match the corresponding virtual component, and also to obtain the current pose of the AR device in real time.

Commonly used AR markers include rectangular logos, random dot markers, and lenticular sheet IDs.

Rectangular identification generally uses a template matching method or QR code-like ways to identify the ID, and calculates the pose by matching with the identification image stored in the library. It is simple and easy to use but the occlusion robustness is poor, and the front view accuracy may be low. Moreover, posture estimation and aesthetics quality can also be poor.

The random point marker adopts a randomly distributed small point as the identification subject, and the ID and the pose are calculated by extracting its position and matching with the template stored in the library. To some extent, this solves problems such as low aesthetic quality and occlusion sensitivity.

Lenticular sheet ID uses variable moiré pattern technology to further solve the problem of limited viewing angle, but the identification complexity and the cost are both high.

The choice of marker type affects user experience in AR service. A decision on which marker type needs to be adopted depends on service characteristics.

Recognition response

In another type of AR service, ID confirmation and pose calculation are implemented by recognizing real objects or real environment. Image feature point extraction methods, such as those designated scale-invariant feature transform, Harris and features from accelerated segment test, are generally used. The feature points of the real environment are then matched with those stored in the library, and the ID of the current environment or object and the pose of the AR device are obtained.

Feature extraction and matching methods vary in speed and accuracy, which will affect the recognition response. The characteristics of the service itself need to be taken into account in selection of the method to provide the ideal user experience.

7.2.5 Consistency

Geometric consistency

Geometric consistency is mainly the consistency of registration, tracking and object size, and whether there is a correct occlusion effect.

Registration consistency means that the correct initial position relative to the scene can be obtained when the AR device is in a static or dynamic situation. Tracking consistency means that the correct relative position after a change in angle of view can be continuously obtained at any time after the device starts moving.

Consistency of size means that in the AR environment, the geometric size of a virtual object in AR is consistent with its presented size. With movement of the device, the angle of view, and the distance between the user and the virtual object, change; the perceived size of objects should be consistent. Whether it is an object in the real world, such as a teacup, table or chair, or a purely virtual object, such as an elf in a game, the size should be consistent with what it is intended to represent, which

helps users understand the scene. In AR applications, such as AR-based shopping, size perception plays an important role, even a key factor.

The occlusion effect refers to whether the virtual content has a correct context with the real environment, i.e., the object in the foreground should block the overlapping part of the background object. The occlusion effect requires the system to process, based on computer vision (CV), the contours and positions of the environment and objects. When the virtual content is placed at a certain position, the mutual occlusion relationship with the surrounding objects is calculated and correctly displayed according to its shape, size and position.

Poor registration will result in different degrees of error in the initial appearance of the virtual content, affecting the user's understanding of the virtual effect, and poor tracking consistency will be reflected in the abnormal changes in the position and size of the virtual content during the user's movement, so that immersion is reduced. The appearance of the occlusion effect will affect the user's perception of the authenticity of the virtual content [b-Gutiérrez].

Illumination consistency

Illumination consistency in AR is the consistency of user retina effect that occurs after virtual content is placed in the real environment.

To achieve illumination consistency, it is necessary to measure: real-object characteristics; light source environment; and camera setting information. Consistency is then achieved through virtual content light source processing. Real-object characteristics include shape, position and reflection. A light source environment includes source type, position, intensity and colour. The first two types of information can be collected by CV technology. Camera settings include external parameters, lens parameters, and imaging settings, which can be obtained directly from the camera.

The experience of illumination consistency is mainly reflected in whether the real environment perceived by the user is consistent with the direction of the virtual content light source, whether the intensity is the same and whether the colour is matched and even.

For those AR scenarios that involve visualizing 3D objects, QoE evaluation is more complex. Currently, the research community is taking into account the new technical and perceptual factors offered by AR. Regarding QoE evaluation of 3D objects, lighting factors play an important role [b-Gutiérrez].

Time consistency

Time consistency is the synchronous change of virtual content and the real environment in the time dimension. The quality of user experience depends on whether the virtual content has the correct time setting and whether time synchronization with the real environment is accurate.

7.2.6 Hardware related

Comfortableness

In OST AR services, users generally need to wear HMDs. The weight and fixture method of devices affect user wearing comfortableness, which will affect the experience and duration of usage. The way in which lens are manufactured may cause eye discomfort and is a key point of comfort considerations.

In SAR services, users view the AR effect in a real scene with the naked eye. Light directly reaches the retina after environmental reflection, and the intensity, frequency band and duration of the illumination affect the comfort level of the user.

Field of view

Field of view (FoV) is the margin of image content that user can see at a certain moment. The larger the FoV, the more environmental content the user can perceive, and the easier it is to create an immersive experience in a scene. Increasing the value of FoV is currently a focus of improvements in AR display devices.

Stereo

The capability of devices to provide stereo vision determines whether the user can observe depth information of an object. VST AR service requires user equipment itself to provide stereoscopic video playback ability. OST AR service requires the device to have layered imaging capability on different focal planes. An SAR service needs to design the depth of the virtual content in combination with the projected target environment structure.

Stereo audio renderition capability determines whether the user can hear different intensities of sound from sources scattered at various locations in the environment.

The stereoscopic rendering function of hardware affects user experience of visual depth information and spatial audio perception and thus has influence on the degree of immersion.

Depth range

The value of the depth range of hardware devices determines the extent to which AR service with depth information effects can be provided. Larger depth range values provide a more realistic and more immersive experience.

In VST AR, the value of the depth range depends on the fineness of CV in environmental analysis. By adjusting the size and position of virtual content, the effect of depth range change can be achieved. In OST AR, in addition to CV fineness, it is also necessary to consider the furthest distance of the focal plane that the optical imaging system of the hardware itself can provide, which determines the upper limit of the depth range.

7.2.7 Interaction related

Hand gesture

Hand gesture recognition means that an AR device can recognize the position of a user's hand and position of the fingers. A hand gesture can be recorded by means of a camera shot, mobile phone mechanism sensing, etc.

The accuracy of hand gesture recognition, delay and types of posture supported have a direct impact on user experience. A poor gesture recognition system increases the probability of misoperation and increases the sense of delay in operation, while excessive types of gesture may increase the learning cost of the user, thereby reducing the willingness to use the system.

Speech

Speech recognition is the function by which an AR device collects a user voice signal through its own microphone and converts it into an operational command with a literal meaning. Speech recognition can replace manual system operation and functional operation of the device.

A low speech recognition success rate increases the probability of misoperation and reduces user willingness to use the system further.

Body posture

Body posture is the function of the AR device by which it identifies and digitizes the position and shape of the human torso and limbs. It can be recorded by camera shot, depth sensing hardware, wearable somatosensory device, etc. Digitized body data can be added to the real environment as part of the virtual content for fusion display.

A low body recognition rate reduces both the realism of the human body in the AR service and immersion in the user experience.

Tangible interfaces

Tangible AR interfaces allow users to use real, physical objects and tools in an intuitive way to interact with virtual content. Tangible AR interfaces are usually accurate and easy to use because the objects used have familiar physical properties that the users are familiar with. Representative tangible AR interfaces include rigid controllers like a paddle or handle and touch pad, as well as deformable surfaces like a haptic glove.

The comfortableness, usability, and accuracy of tangible AR interfaces are key factors influencing overall AR QoE.

Eye- or head-based interfaces

Eye and head movement can be leveraged to enrich user interaction with the augmented world. As representative HCI techniques, eye- and head-tracking have long been powerful ways of communication and control, especially for people with physical disabilities.

A low eye- or head-tracking and recognition accuracy can significantly reduce the experience of AR services.

Brain-computer interfaces

Brain-computer interfaces (BCIs) allow users to interact with an AR system by building a direct communication pathway between the electrical activity of the brain and the AR device. Depending on how close electrodes get to the brain, BCIs for AR can be categorized as: non-invasive; partially invasive; and invasive.

The reliability, safety and accuracy of BCIs are key factors influencing AR QoE.

7.3 Context influence factors

7.3.1 Physical context

Physical influencing factors relate to what physical environment conditions are when users experience AR services. Since mobility and environmental dependency are fundamental attributes of AR, the physical context plays a decisive role in its user experience.

The complexity of a physical environment affects the accuracy of AR systems to detect, identify and track a target object, i.e., the more complex the background, the more difficult it is to distinguish the target. A freely moveable range determines how close the AR effect is to the user. A close virtual object display may bring more information. Whether the background sound is noisy affects a user's aural perception of a virtual object. Contrast in ambient light challenges the optical latitude of an AR imaging system. The number and distribution of light sources affect the complexity and realism of AR ray estimation and shadow calculations.

In more complex olfactory and tactile-enhanced AR systems, factors such as temperature, humidity, odour type and intensity also have impacts on the user experience.

7.3.2 Collaboration context

The collaborative context factor relates to whether the user uses the AR service alone or with others. Multi-person AR services include games, teaching and office in the same space. On one hand, using AR services with family, friends and strangers increases the channels of interaction and information input, allowing users to have fun and satisfaction. On the other hand, cooperating with colleagues and classmates to complete a common AR operation task may increase requirements for accuracy and matching of the interaction; thus, the user may feel the experience to be more laborious. The impact of a collaborative environment under different preconditions is for further study.

7.3.3 Task context

The task context relates to what tasks and goals users are aiming at in their minds. When involved in some entertainment and social tasks, users may be concerned about the fidelity, complexity and ease of communication with others. When completing work and teaching tasks, the user may be more concerned about accuracy of the position of AR objects and accuracy in operation control.

7.3.4 Temporal context

The temporal context factor relates to how high the frequency and duration of use is in AR services. In VST AR service, users may need handheld devices to achieve AR effects, thus long-term use can cause fatigue and experience can degrade over time. In OST AR services, the display device worn by users generally adopts near-eye imaging technology, and high frequency and duration of use may cause eye fatigue and loss of visual acuity by the user. Therefore, for different usage modes and usage scenarios, it is necessary to determine which usage times are reasonable.

7.4 General factors

7.4.1 Human factor

Olfactory

Olfactory-based AR services are relatively rare. Unlike the types of perceptual enhancements discussed in clauses 7.1 to 7.3, they deal with not physical but chemical signals. The mechanism of olfactory perception is not fully understood, but studies have shown that its perceived experience changes with environment and subjective state [b- Kerruish].

The human olfactory experience is affected by the pathology of the respiratory system, with a lack of experience for a particular or all kinds of smell or taste.

Multisensory integration

Human sensory experiences are not independent, and two or more senses can interact and coordinate. For example, a multisensory illusion such as the McGurk effect indicates that two convergent perceptual bimodal stimuli can produce a new perception that is different not only in magnitude from the sum of its parts, but also in a quality. AR services need to take into account the impact of different sensory synergies and consider service design.

Simulator sickness

Discomfort relates to symptoms such as dizziness, nausea, vomiting, eye strain and cold sweat, which may occur when wearing an AR display device or watching an SAR video. The causes of the problem include a low refresh rate of the display hardware device, flickering of the screen, hysteresis of the virtual component during motion and mismatch between multi-sensory integration and the real world.

The discomfort caused by devices for VST AR and SAR may be weaker than that caused by one for OST AR. The occurrence of discomfort may be reduced by improving hardware parameters and system-processing performance.

Static and dynamic human factor

QoE is closely related to user characteristics [b-Reiter]. Some of these factors that can affect the user experience are static and do not change during use of AR services (such as age and gender), while others are more or less dynamic, e.g., emotions. In both cases, the role of characteristics in shaping user experience is important. Although these factors are not specific to AR services, they need to be taken into consideration.

7.4.2 Hardware-related factor

Colour space

The colour space of hardware determines the number of types of colours that can be rendered. Colour spaces such as in [b-ITU-R BT.709] and in [b-ITU-R BT.2020] can serve as reference values. A large display colour space can provide high colour consistency and image quality consistency in illumination consistency experience, making the colour performance of the virtual content closer to that of the real environment and enhancing user identity.

Dynamic range

The dynamic range is the ratio of maximum and minimum value of display light intensity and sound levels in AR hardware. The high dynamic range (HDR) provides a wider scope of visual and auditory experiences than the standard dynamic range (SDR) and approximates the realities of the real world. For example, in the video dynamic range, the evolution of SDR to HDR technology can bring greater contrast to AR hardware to adapt to the illumination changes in an actual physical environment.

Refresh rate

The refresh rate of hardware relates to how many pictures can be acquired from the image processing unit per second when displayed on a screen. A low refresh rate can cause smearing and hysteresis of virtual content during exercise, increasing user discomfort.

8 Augmented reality quality of experience assessment

Considering the factors affecting AR QoE discussed in clause 7, a scheme for AR QoE assessment is proposed. Consideration of a scheme based on the three dimensions shown in Figure 1 is recommended.

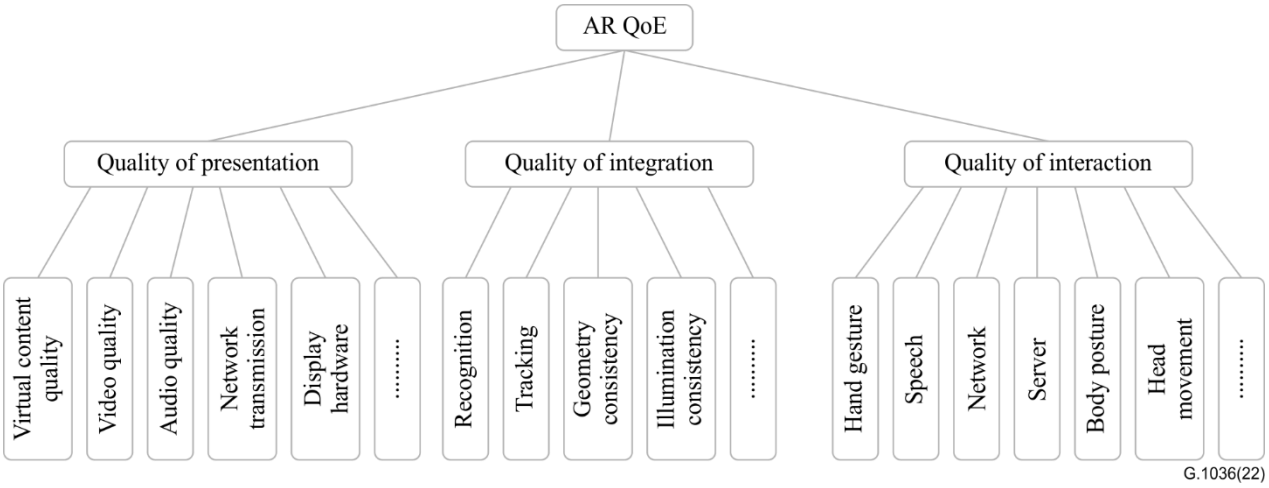

Figure 1 – Augmented reality quality of experience assessment

8.1 Presentation quality

Presentation quality includes the quality of each step related to content flow and display, including production and encoding quality of multimedia and virtual content, quality of network transmission, quality of display hardware processing and playback.

Presentation quality is recommended for assessment by traditional A/V media evaluation methods, for network bandwidth, delay, packet loss and device display quality, and finally to give a calculation method for this parameter.

Related metrics are listed in Table 1.

Table 1 – Related metrics for presentation quality assessment

Traditional audio video media quality metrics	video resolution
	video frame rate
	video bit rate
	audio parameters
	audio and video synchronization
Network transmission metrics	delay
	packet loss
	jitter
Device display quality metrics	screen refresh rate
	field of view
	screen size

8.2 Combination quality

Combination quality relates to the quality of merging virtual content with the real environment. This includes detection and recognition of target objects, tracking, as well as geometric and illumination consistency.

It is recommended that subjective evaluation criteria of combination quality be discussed first, and then assessment methods specified from the perspective of various influencing factors.

8.3 Quality of interaction

The quality of interaction relates to the quality of information input and response between the user and AR device and service.

Quality of interaction is recommended for assessment mainly on aspects including response speed, accuracy and degree of naturalness.

Appendix I

Augmented reality services use cases

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

I.1 Face augmented reality use scenario

Such applications are mainly for users employing the front camera of their smart device to present various augmented effects of the face locally or to remotely see another participant's facial AR effect through a video call.

I.1.1 Makeup trial

Users can change their lip gloss colour, hair colour, eyebrow shape, etc., just like makeup in the real world. This can help users save time and materials for makeup, and on other websites and applications where cosmetics are sold, users can more intuitively see the effect of using the product.

I.1.2 Masking

Users can wear a mask on their face and then record a video or make a video call. The effect of the mask varies depending on the degree of fineness of face tracking. A good 3D face tracking system can provide a mask effect that fits perfectly into the shape of the face, even more than a real world mask. Such applications are mostly used in the fields of entertainment or video calling.

I.1.3 Face shape control

Such applications allow users to change the shape of their facial organs, such as the size of the eyes, the length of the nose, the width of the mouth and the thickness of the lips. It is generally used to create a "distorting mirror" effect in entertainment applications, to allow users to incorporate their favourite famous person's facial features on their own faces or to let the user see the finished situation before plastic surgery.

I.1.4 Background substitution

The application of the background replacement class is complementary to where the effects of clauses I.1.1 to I.1.3 are superimposed in the foreground. It is an application that first recognizes the background content outside the user's face and then superimposes a virtual effect on these areas. It is possible to create a user effect in a particular environment or scene.

I.2 Space-based augmented reality use scenario

Such applications primarily use the rear camera of a user's smart device or the video capture camera of AR glasses. The result of a process of identification, virtual effect generation and effect overlay of actual objects and environments in space is then displayed. The application fields are very extensive, mainly as described in clauses I.2.1 to I.2.5.

I.2.1 Education

When space-based AR is applied in the field of education, the slide information that is currently fixedly displayed in the plane can be superimposed into the space, and more annotation, combination and dynamic information is generated than objects actually seen by the user. It makes the learning process easy to understand while introducing more additional information, and is mainly used in education and professional courses.

I.2.2 Gaming

In a gaming application, one or more users can see the game scenes in space, and the objects in those scenes remain at a fixed size and position. The game mode has changed from a fixed screen display to a user's deployment around real space, providing immersion and authenticity that has never been seen before. The devices used may include smart phones and AR glasses, and multiple people need to synchronize their devices when participating at the same time.

I.2.3 Navigation

When the user is navigating, the application can render a road sign, arrow, location name, etc. in real time to the current screen for display. The navigation mode thus changes from traditional "map-translation" to space exhibition, and becomes a viewing and experience in real space, making the choice of roads and the positioning of buildings easier. It can be used in navigation for walking, driving, indoors and other scenarios.

I.2.4 Training

In the fields of military driving, aircraft piloting and large-scale equipment operation, training and supervision are difficult and costly. An AR service can help these trainees see all kinds of information they need in real time, and automate and refine the training process and programme.

I.2.5 Remotely industrial maintenance/instruction

In industrial maintenance, actual on-site workers often encounter situations in which the equipment is too complicated for problem location and so machines cannot be operated and maintained. Using a common video calling service to provide remote guidance by professionals is an available solution. However, there are also contradictions in which the language description is inaccurate and the problem points cannot be located in the machine. The AR service can allow maintenance personnel to see various positions, shapes and treatments drawn by professionals on the equipment in real time during the video call, so that the feasibility of remote maintenance is enhanced.

Bibliography

[b-ITU-T G.1035] Recommendation ITU-T G.1035 (2021), *Influencing factors on quality of experience for virtual reality services*.

[b-ITU-T J.248] Recommendation ITU-T J.248 (2008), *Requirements for operational monitoring of video-to-audio delay in the distribution of television programs.*

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

[b-ITU-R BT.709] Recommendation ITU-R BT.709 (2015), *Parameter values for the HDTV standards for production and international programme exchange.*

[b-ITU-R BT.2020] Recommendation ITU-R BT.2020 (2015), *Parameter values for ultra-high definition television systems for production and international programme exchange.*

[b-3GPP TR 26.998] 3rd Generation Partnership Project Technical Report 26.998 V17.0.0 (2022), *Support of 5G glass-type augmented reality/mixed reality (AR/MR) devices – Release 17*.

[b-Brandenburg] van Brandenburg, R., Koenen, R, Sztykman, D. (2017). *CDN optimization for VR streaming*. Available [viewed 2022-08-24] at: https://www.ibc.org/download?ac=3831

[b-Cao] Cao, Y., Ritz, C., Raad, R. (2013). How much longer to go? The influence of waiting time and progress indicators on quality of experience for mobile visual search applied to print media. In: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 112-117. DOI: 10.1109/QoMEX.2013.6603220

[b-Cui] Cui, L., Chen, S., Jiang, Z. (2017). Poster Abstract: Quality of experience aware collaborative augment reality system (Q-CARS) – In:_ *2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*, 2017), pp. 345-346.

[b-Duan 2022a] Duan, H., Min, X., Zhu, Y., Zhai, G., Yang, X., Le Callet, P. (2022). *Confusing image quality assessment: Towards better augmented reality experience* [preprint]. DOI: 10.48550/ARXIV:2204.04900.

[b-Duan 2022b] Duan, H., Shen, W., Min, X., Tu, D., Li, J., Guangtao, Z. (2022). *Saliency in augmented reality* [preprint]. DOI: 10.48550/ARXIV.2204.08308.

[b-Gutiérrez] Gutiérrez, J., Vigier, T., Le Callet, P. (2020). Evaluation of 3D objects in mixed reality for different lighting conditions. In: *Proc. IS&T Int. Symp. Electronic Imaging: Human Vision and Electronic Imaging*, pp 128-1-128-7. DOI: 10.2352/ISSN.2470-1173.2020.11.HVEI-128

[b-Kennedy] Kennedy, R.S., Lane, N.E., Berbaum, K.S., Lilienthal, M.G. (1993). Simulator sickness questionnaire: An enhanced method of quantifying simulator sickness. *Int. J. Aviat. Psychol*. **3**(3), pp. 203–220. DOI: 10.1207/s15327108ijap0303_3

[b-Kerruish] Kerruish, E. (2019). Arranging sensations: Smell and taste in augmented and virtual reality. *Senses and Society*, **14**(1), pp. 31-45. DOI: 10.1080/17458927.2018.1556952

[b-Li] Li, J., Zhang, C., Liu, Z., Sun, W., Li, Q. (2020). Joint communication and computational resource allocation for QoE-driven point cloud video streaming. *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pp. 1-6. DOI: 10.1109/ICC40277.2020.9148922

[b-Puig] Puig, J., Perkis, A., Lindseth, F., Ebrahimi, T. (2012). Towards an efficient methodology for evaluation of quality of experience in augmented reality. In: *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 188-193, doi: 10.1109/QoMEX.2012.6263864

[b-Reiter] Reiter, U., Brunnström, K, De Moor, K., Larabi. M.-C., Pereira, M., Pinheiro, A., You, J., Zgank, A. (2014)., Factors influencing quality of experience. In: Möller, S., Raake, A., editors. *Quality of experience: Advanced concepts, applications and methods*, pp. 55-72. *T-Labs Series in Telecommunication Services*. Berlin: Springer.

[b-Seufert] Seufert, M., Egger, S., Slanina, M., Zinner, T., Hoßfeld, T., Tran-Gia, P. (2014). A survey on quality of experience of HTTP adaptive streaming. *IEEE Commun. Surv. Tut.* **17**(1), pp. 469-492. DOI: 10.1109/COMST.2014.2360940

[b-Stanney] Stanney, K.M., Kennedy, R.S., Drexler, J.M. (1997). Cybersickness is not simulator sickness. *Proc. Human Factors and Ergonomics Society Annual Meeting*. **41**(2). pp. 1138–1142. DOI: 10.1177/107118139704100292

[b-Witmer] Witmer, B.G., Singer, M.J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments.* **7**(3), pp. 225–240. DOI: 10.1162/105474698565686

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

Printed in Switzerland
Geneva, 2022