ETSI GS ARF 003 V1.1.1 (2020-03)

Augmented Reality Framework (ARF); AR framework architecture

Disclaimer

The present document has been produced and approved by the Augmented Reality Framework (ARF) ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference

DGS/ARF-003

Keywords

API, architecture, augmented reality, context capturing and analysis, framework, model, real time

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Important notice

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

Foreword

This Group Specification (GS) has been produced by ETSI Industry Specification Group (ISG) Augmented Reality Framework (ARF).

The ISG ARF shares the following understanding for Augmented Reality: Augmented Reality (AR) is the ability to mix in real-time spatially-registered digital content with the real world. The present document specifies a functional reference architecture for AR solutions.

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1 Scope

The present document specifies a functional reference architecture for AR components, systems and services. The structure of this architecture and the functionalities of its components have been derived from a collection of use cases ETSI GR ARF 002 [i.3] and an overview of the current landscape of AR standards ETSI GR ARF 001 [i.4].

The present document introduces the characteristics of an AR system and describes the functional building blocks of the AR reference architecture and their mutual relationships. The generic nature of the architecture is validated by mapping the workflow of several use cases to the components of this framework architecture.

2 References

2.1 Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found at https://docbox.etsi.org/Reference/.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are necessary for the application of the present document.

Not applicable.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] Ronald T. Azuma, "A survey of augmented reality", in Presence: Teleoperators & Virtual Environments, 1997.

[i.2] Paul Milgram, H. Takemura, A. Utsumi, F. Kishino, "Augmented Reality: A class of displays on the reality-virtuality continuum", in Proceedings of Telemanipulator and Telepresence Technologies, 1994.

[i.3] ETSI GR ARF 002 (V1.1.1): "Augmented Reality Framework (ARF) Industrial use cases for AR applications and services".

[i.4] ETSI GR ARF 001 (V1.1.1): "Augmented Reality Framework (ARF); AR standards landscape".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the following terms apply:

Augmented Reality (AR): ability to mix in real-time spatially-registered digital content with the real world

NOTE: This definition is based on the work of Azuma [i.1] and Milgram [i.2].

AR anchor: coordinate system related to an element of the real world on which virtual content stays spatially-registered

NOTE: AR anchors help users to maintain the perception that static virtual content appears to stay at the same position and orientation in the real world.

AR application: software designed by using several **AR components** to perform a group of coordinated functions, tasks, or activities for the benefit of the user who is experiencing augmented reality

AR component: hardware or software that provides application-oriented computing functions and supports interoperability when connected with other components of the **AR system**

AR device: hardware that provides one or more functions offering an augmented reality experience to one or several users

AR experience: the real time perception of the mixture of the real world and spatially-registered digital content by user senses

AR system: combination of hardware and software that delivers an AR experience

AR scene: information describing the interactive content contributing to an augmented reality experience

Building Information Modeling (BIM): process supported by various tools and technologies involving the generation and management of digital representations of physical and functional characteristics of places

descriptor extraction: task consisting in extracting differentiating characteristics of a detected feature

feature detection: task consisting in detecting specific information from a given signal

function: collection of functionalities

Product Lifecycle Management (PLM): process of managing the entire lifecycle of a product from inception through engineering, design, and manufacture to service and disposal of manufactured products

pose: position and orientation of an object, defined in a given coordinate system

EXAMPLE: The camera pose defined in a world coordinate system.

pose estimation: task of determining the pose of an object

object recognition: task consisting in finding and identifying objects

EXAMPLE: Recognition may be performed on an image, a video sequence, or an audio stream.

object tracking: task consisting in locating an object over time

EXAMPLE 1: A 2D tracking consists in locating an object in a sequence of images.

EXAMPLE 2: A 3D tracking consists in locating an object in a 3D space from a sequence of images or an audio signal.

point cloud: set of data points in space defined in a common coordinate system

EXAMPLE: A 3D point cloud is a set of data points in a 3D space.

random forest: learning method based on a multitude of decision trees used for classification or regression tasks

reference point: point located at the interface of two non-overlapping functions and representing interrelated interactions between those functions

visual bag of words: simplified representation using image features as words for image retrieval task

visual descriptor: characteristics of a visual feature

NOTE: A descriptor is based on elementary characteristics such as the shape, the colour, the texture or the motion of the feature itself and its neighbourhood in the image.

EXAMPLE: SIFT, SURF, BRIEF, ORB, BRISK, FAST, etc.

visual feature: information representing an element of an image

NOTE: The feature are generally primitive geometric elements (points, edges, lines, polygons, colors, textures, or any shapes) used to characterize an image.

EXAMPLE: Keypoints, edges, blobs.

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

AR	Augmented Reality
AV	AudioVisual
BIM	Building Information Modeling
BRIEF	Binary Robust Independent Elementary Features
CAD	Computer-Aided Design
DoF	Degree of Freedom
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GPU	Graphic processing Unit
GUI	Graphical User Interface
IMU	Inertial Measurement Unit
IP	Internet Protocol
Li-Fi™	Light Fidelity
ORB	Oriented FAST and Rotated Brief
PLM	Product Lifecycle Management
RGB	Red, Green, Blue
RGB-D	Red, Green, Blue and Depth
SIFT	Scale-Invariant Feature Transform
SURF	Speeded Up Robust Features
TPU	Tensor Processing Unit
UWB	Ultra-WideBand
VPU	Vision Processing Unit
Wi-Fi™	Wireless Fidelity

4 Architecture Overview

4.1 Global Architecture

An AR system is based on a set of hardware and software components as well as data describing the real world and virtual content. Figure 1 presents a global overview of an AR system architecture. The architecture diagram is structured in three layers, in the upper part the hardware, in the middle the software, and in the lower part the data:

- Hardware layer:
  - **Tracking Sensors:** These sensors aim to localize (position and orientation) the AR system in real-time in order to register virtual contents with the real environment. Most of AR systems such as smartphones, tablets or see-through glasses embed at least one or several vision sensors (generally monochrome or RGB cameras) as well as an inertial measurement unit and a GPS™. However, specific and/or recent systems use complementary sensors such as dedicated vision sensors (e.g. depth sensors and event cameras), or exteroceptive sensors (e.g. Infrared/laser tracking, Li-Fi™ and Wi-Fi™).
  - **Processing Units:** Computer vision, machine learning-based inference as well as 3D rendering are processing operations requiring significant computing resources optimized thanks to dedicated processor architectures (e.g. GPU, VPU and TPU). These processing units can be embedded in the device, can be remote and/or distributed.
  - **Rendering Interfaces:** Virtual content require interfaces to be rendered to the user so that he or she can perceive them as part of the real world. As each rendering device has its own characteristics, the signals generated by the rendering software generally need to be transformed in order to adapt them to each specific rendering hardware.
- Software layer:
  - **Vision Engine:** This software aims to mix the virtual content with the real world. It consists of localizing (position and orientation) the AR device relative to the real world reference, localizing specific real objects relatively to the AR device, reconstructing a 3D representation of the real world or analysing the real world (e.g. objects detection, segmentation, classification and tracking). This software component essentially uses vision sensors signals as input, but not only (e.g. fusion of visual information with inertial measurements or initialization with a GPS), it benefits from the hardware optimization offered by the various dedicated processors embedded in the device or remote, and will deliver to the rendering engine all information required to adapt the rendering for a consistent combination of virtual content with the real world.
  - **3D Rendering Engine:** This software maintains an up-to-date internal 3D representation of the virtual scene augmenting the real world. This internal representation is updated in real-time according to various inputs such as user's interactions, virtual objects behaviour, the last user viewpoint estimated by the Vision Engine, an update of the World Knowledge to manage for example occlusions between real and virtual elements, etc. This internal representation of the virtual content is accessible by the renderer (e.g. video, audio or haptic) which produces thanks to dedicated hardware (e.g. Graphic Processing unit) data (e.g. 2D images, sounds or forces) ready to be played by the Rendering Interfaces (e.g. screens, headphones or a force-feedback arm).
- Data layer:
  - **World Knowledge:** This World Knowledge represents the information either generated by the Vision Engine or imported from external tools to provide information about the real world or a part of this world (CAD model, markers, etc.). This World Knowledge corresponds to the digital representation of the real space used for different usages such as localization, world analysis, 3D reconstruction, etc.
  - **Interactive Contents:** These Interactive Contents represent the virtual content mixed to the perception of the real world. These contents can be interactive or dynamic, meaning that they include both 3D contents, their animations, their behaviour regarding input events such as user's interactions. These Interactive Contents could be extracted from external authoring tools requiring to adapt original content to AR application (e.g. 3D model simplification, fusion, and instruction guidelines conversion).

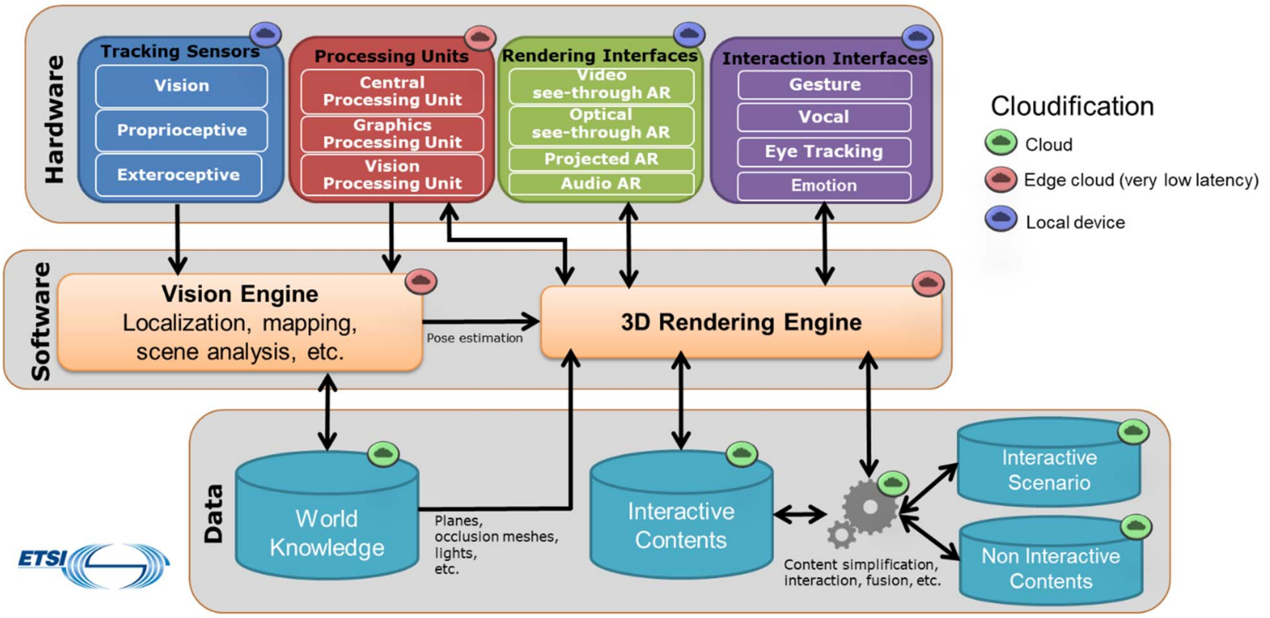

Figure 1: Global overview of the architecture of an AR system

4.2 Functional Architecture

Figure 2 shows the functional architecture specified by the present document addressing both fully embedded AR systems and implementations spread over IP networks in a scalable manner. Logical functions are shown as named boxes that may be nested in cases where a high-level function is composed of several subfunctions. The logical functions are connected by reference points. A reference point in a functional architecture is located at the conjunction of two non-overlapping functions and represents the interrelated interactions between those functions. A reference point allows a framework to aggregate those abilities that one function provides towards another function. In a practical deployment each of these reference points can be realized by a physical interface that conveys information between the connected subfunctions in a unidirectional or bidirectional way using a specified protocol. Depending on the deployment scenario and the applications that needs to be supported, multiple logical subfunctions can also be combined in one deployable unit. All of these subfunctions can either be deployed on the device that also presents the AR implementation or they can be provided via cloud technology.

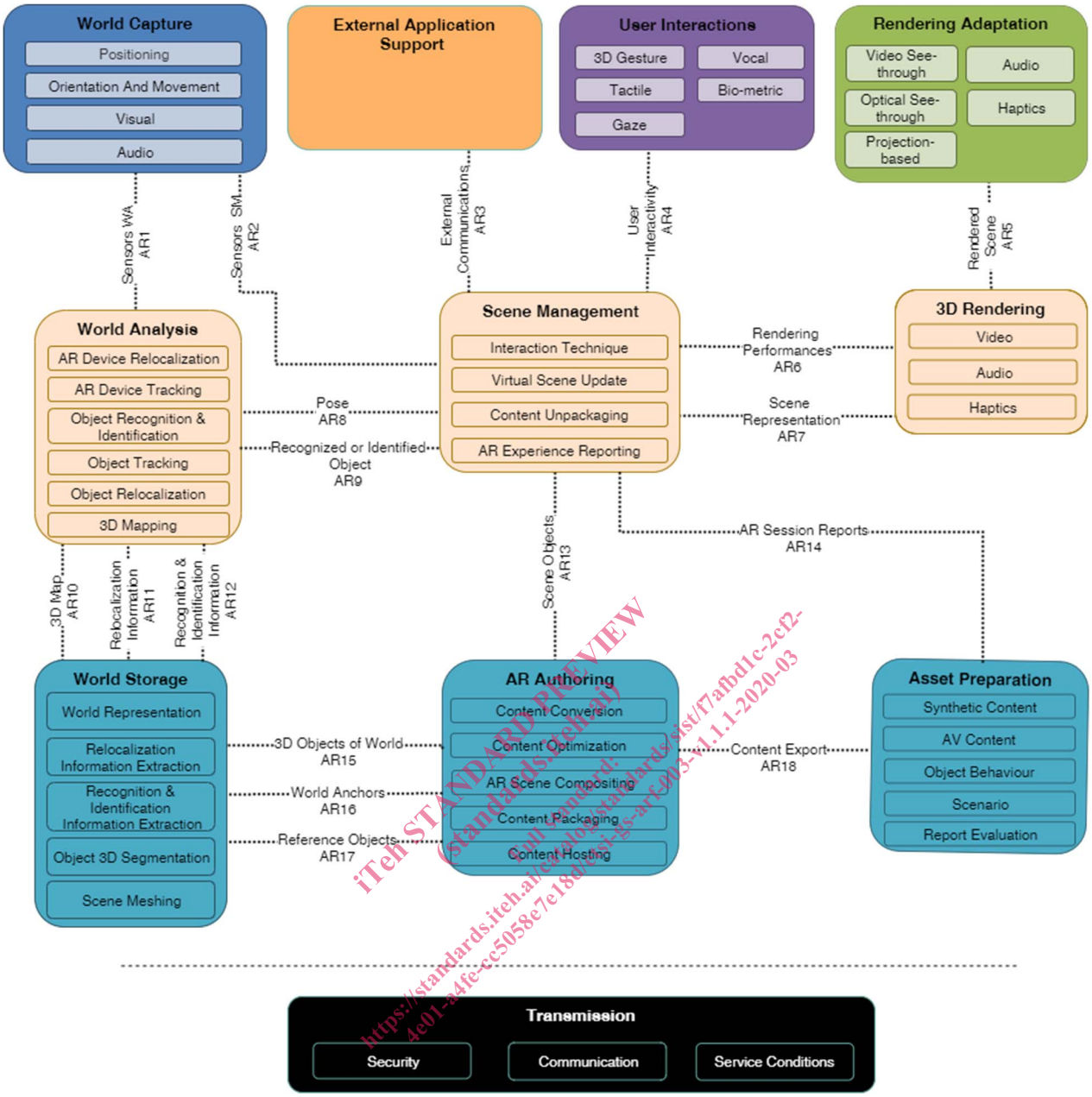

Figure 2: Diagram of the functional reference architecture

5 Functions and subfunctions of the functional architecture

5.1 World Capture

5.1.1 Introduction

This function deliver information relevant for the localization of the AR device or real objects, or for analysing the environment of the application. AR systems can embed various sensors aiming at better understanding the real environment as well as the pose (position and orientation) of the AR system or of real objects in this environment required to provide an accurate registration of virtual objects on the real world. The following subfunctions can address different kinds of sensors.

5.1.2 Positioning

This subfunction shall deliver the location of an AR device and may also deliver its orientation relatively to a coordinate reference system. The coordinate reference system shall be defined in relation to a reference in the real world and there shall be a mechanism to temporally synchronize this information with other **World Capture** subfunctions.

NOTE 1: The coordinate reference system can be related to the earth (global geo-positioning system), a factory, a room, an object, the positioning system itself, etc.

NOTE 2: GNSS are the most commonly used means to provide a position, but other solution such as beacons, Li-Fi™, UWB or other radio technologies can also provide a position or even an orientation of an AR system device.

5.1.3 Orientation and Movement

This subfunction shall deliver information about the movement or the orientation of the device that is providing the AR experience. This information shall be defined in a given coordinate system and there shall be a mechanism to temporally synchronize this information with other **World Capture** subfunctions.

NOTE 1: This subfunction can make use of the information provided by the subfunction **Positioning.**

NOTE 2: Inertial measuring can provide information about the movement and orientation of the device with a high frequency (~1 000 captures per seconds) which is useful to interpolate intermediate device poses between two vision-based pose estimations or when vision-based pose estimation fails.

5.1.4 Visual

This subfunction applies to AR systems making use of cameras (e.g. RGB, depth or event-based) which support the device providing AR experience. The cameras can be built into the device (interoceptive capture) or positioned outside the device (exteroceptive capture).

This subfunction shall deliver streaming video/event or still pictures to be used by the **World Analysis** function.

This subfunction should also deliver information about camera parameters relevant for the application (e.g. focal lengths, pixel size, principal point, image size, global or rolling shutter, distortion parameters, timing information and camera range) and in the case where the application uses several vision sensors, their relative positions and orientations.

Different kind of visual sensors can be addressed by this subfunction:

- RGB cameras make use of a photo-sensitive sensors by which coloured images are acquired. Colours are represented by an additive combination of the three primary colours red, green and blue.
- Depth cameras produce two-dimensional images that contain information about the distance to points of a scene from a given specific point. Several technologies can be used to achieve such information (e.g. stereoscopic triangulation, light patterns). In many cases, a depth camera can also provide a cloud of points defined in the coordinate system of the sensor.
- RGB-D cameras produce both two-dimensional images captured by a RGB camera and two-dimensional images captured by a depth sensor. For this reason, RGB-D cameras provide both camera and depth sensor interfaces. But, since an RGB-D camera is usually composed of two separate sensors, an RGB camera and a depth sensor (time of flight, structured-lights-based, stereoscopic, etc.), the raw images from the two sensors are not aligned. For this reason, RGB-D sensors offer complementary interfaces to match RGB and depth images.
- Event-based cameras measure the changes in brightness and colour in a scene over a given time period and allows the detection of movement of objects within the scene.
- Infrared cameras measure the infrared radiation of an object and maps the measured wavelength range into a picture using pseudo-colours.
- Laser trackers are used to measure the distance between such devices and objects that reflect laser pulses sent out by the tracker. Often, these measurements are mapped into a local coordinate system.