INTERNATIONAL STANDARD

ISO/IEC 23000-13

First edition
2014-05-15

Information technology — Multimedia application format (MPEG-A) —

Part 13: Augmented reality application format

Technologies de l'information — Format des applications multimedias —

Partie 13: Format pour les Applications de Realité Augmentée

Reference number
ISO/IEC 23000-13:2014(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC 23000-13 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

ISO/IEC 23000 consists of the following parts, under the general title *Information technology — Multimedia application format (MPEG-A)*:

— *Part 1: Purpose for multimedia application formats* [Technical Report]

— *Part 2: MPEG music player application format*

— *Part 3: MPEG photo player application format*

— *Part 4: Musical slide show application format*

— *Part 5: Media streaming application format*

— *Part 6: Professional archival application format*

— *Part 7: Open access application format*

— *Part 8: Portable video application format*

— *Part 9: Digital Multimedia Broadcasting application format*

— *Part 10: Surveillance application format*

— *Part 11: Stereoscopic video application format*

— *Part 12: Interactive music application format*

— *Part 13: Augmented reality application format*

Introduction

Augmented Reality (AR) applications refer to a view of a real-world environment (RWE) whose elements are augmented by content, such as graphics or sound, in a computer driven process. Augmented Reality Application Format (ARAF) is a collection of a subset of the ISO/IEC 14496-11 (MPEG-4 part 11) Scene Description and Application Engine standard, combined with other relevant MPEG standards (e.g. ISO/IEC 23005 - MPEG-V), designed to enable the consumption of 2D/3D multimedia content. Consequently, ISO/IEC 23000-13 focuses not on client or server procedures but on the data formats used to provide an augmented reality presentation.

Information technology — Multimedia application format (MPEG-A) —

Part 13: Augmented reality application format

1 Scope

This part of ISO/IEC 23000 specifies:

- Scene description elements for representing AR content
- Mechanisms to connect to local and remote sensors and actuators
- Mechanisms to integrated compressed medias (image, audio, video, graphics)
- Mechanisms to connect to remote resources such as maps and compressed medias

2 Normative references

The following documents, in whole or in part, are normatively referenced in this document and are indispensable for its application. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 14496-1, *Information technology — Coding of audio-visual objects — Part 1: Systems*

ISO/IEC 14496-3:2009, *Information technology — Coding of audio-visual objects — Part 3: Audio*

ISO/IEC 14496-11:2005, *Information technology — Coding of audio-visual objects — Part 11: Scene description and application engine*

ISO/IEC 14496-16:2011, *Information technology — Coding of audio-visual objects — Part 16: Animation Framework eXtension (AFX)*

ISO/IEC 23005-5:2013, *Information technology — Media context and control — Part 5: Data formats for interaction devices*

ISO/IEC 14772-1:1997, *Information technology — Computer graphics and image processing — The Virtual Reality Modeling Language — Part 1: Functional specification and UTF-8 encoding*

ISO/IEC 10646:2012, *Information technology — Universal Coded Character Set (UCS)*

ISO/IEC 8859-1:1998, *Information technology — 8-bit single-byte coded graphic character sets — Part 1: Latin alphabet No.1*

3 Abbreviated terms

For the purposes of this International Standard, the following abbreviated terms apply.

AR Augmented Reality

URI Uniform Resource Identifier

URL Uniform Resource Locator

URN Uniform Resource Name

4 ARAF Components

4.1 ARAF principle and context

Augmented Reality (AR) applications refer to a view of a real-world environment whose elements are augmented by content, such as graphics or sound, in a computer driven process. Figure 1 illustrates two real and virtual cameras and the composition of a real image and graphics objects.

Figure 1 — Simplified illustration of the AR principle

The Augmented Reality Application Format (ARAF) is an extension of a subset of the MPEG-4 part 11 Scene Description and Application Engine standard, combined with other relevant MPEG standards (MPEG-4, MPEG-V), designed to enable the consumption of 2D/3D multimedia content as depicted in Figure 2.

An ARAF, available as a file or stream, is interpreted by a device, called **ARAF device**. The nodes of the ARAF scene point to different sources of multimedia content such as 2D/3D image, 2D/3D audio, 2D/3D video, 2D/3D graphics and sensor/sensory information sources/sinks that are either remote or/and local.

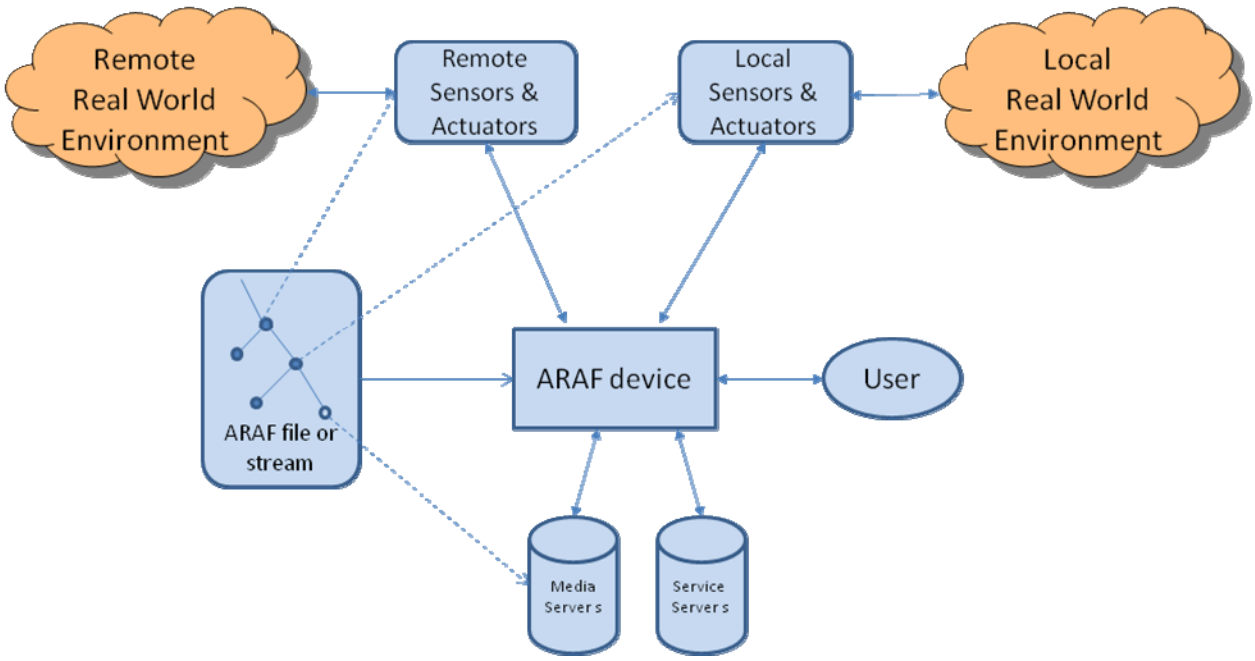

Figure 2 — The ARAF context

4.2 ARAF Scene Description

To describe the multimedia scene ARAF is based on ISO/IEC 14496-11 (MPEG-4 Part 11 BIFS) which at its turn is based on ISO/IEC 14772-1:1997 (VRML97). About two hundreds nodes are standardized in MPEG-4 BIFS and VRML, allowing various kinds of scenes to be constructed. ARAF is referring to a subset of MPEG-4 BIFS nodes for scene description as presented below.

Category	Sub-category	Node, Prototypes / Elements name in MPEG-4 BIFS / XMT
Elementary media	Audio	AudioSource
		Sound
		Sound2D
	Image and video	ImageTexture
		MovieTexture
	Textual information	FontStyle
		Text
	Graphics	Appearance
		Color
		LineProperties
		LinearGradient
		Material
		Material2D
		Rectangle
		Shape
		SBVCAnimationV2
		SBBone
		SBSegment
		SBSite
		SBSkinnedModel
		MorphShape
		Coordinate
		TextureCoordinate

		Normal
		IndexedFaceSet
		IndexedLineSet
Programming information		Script
User interactivity		InputSensor
		OutputActuator
		SphereSensor
		TimeSensor
		TouchSensor
		MediaSensor
		PlaneSensor
Scene related information (spatial and temporal relationships)		AugmentationRegion
		SimpleAugmentationRegion
		Background
		Background2D
		CameraCalibration
		Group
		Inline
		Layer2D
		Layer3D
		Layout
		NavigationInfo
		OrderedGroup
		ReferenceSignal
		ReferenceSignalLocation
		Switch
		Transform
		Transform2D
		Viewpoint
		Viewport
		Form
Dynamic and animated scene		OrientationInterpolator
		ScalarInterpolator
		CoordinateInterpolator
		ColorInterpolator
		PositionInterpolator
		Valuator
Communication and compression		BitWrapper
		MediaControl
	Maps	Map
		MapOverlay
		MapMarker
		MapPlayer
Terminal		TermCap

All the above listed elements are specified in MPEG-4 Part 11. However, to facilitate the implementation of ARAF content, the current document contains the XML syntax as well as the semantics and functionality.

MPEG-4 Part 11 describes a scene with a hierarchical structure that can be represented as a graph. Nodes of the graph build up various types of objects, such as audio video, image, graphic, text, etc. Furthermore, to ensure the flexibility, a new, user-defined type of node derived from a parent one can also be defined on demand by using the *Proto* method.

In general, nodes expose a set of parameters, through which aspects of their appearance and behavior can be controlled. By setting these values, scene designers have a tool to force a scene-reconstruction at clients' terminals to adhere to their intention in a predefined manner. In more complicated scenario, the structure of BIFS nodes is not necessarily static; nodes can be added or removed from the scene graph arbitrarily.

Certain types of nodes called *sensors*, such as TimeSensor, TouchSensor, can interact with users and generate appropriate triggers, which are transmitted to others nodes by routing mechanism, causing changes in state of these receiving nodes. They are bases for the dynamic behavior of a multimedia content supported by MPEG-4.

The maximum flexibility in the programmable feature of MPEG-4 scene is carried out with the *Script* node. By routing mechanism to Event In *valueIn* attribute of Script node, the associated function (defined in its URL attribute) with the same name Event In *valueIn* () will be triggered. The behavior of this function is user-defined, i.e. scene-designer can freely process some computations, and then sets the values for every Event Out *valueOut* attribute, which consecutively affect the states of other nodes linked to them.

Direct manipulation of nodes' states is also possible in MPEG-4 Part 11: the Field *field* attribute can refer to any node in the scene; through this link, all attributes of the contacted node will be exposed to direct setting and modifying operators within the *Script* node. The syntax of the language used to implement the function of Script node is ECMAScript [ISO/IEC DIS 16262 Information technology - ECMAScript: A general purpose, cross-platform programming language].

ARAF supports the definition and reusability of complex objects by using the MPEG-4 PROTO mechanism. The PROTO statement creates its own nodes by defining a configurable object prototype; it can integrate any other node from the scene graph.

The following table indicates the MPEG-4 Part 11 nodes that are included in ARAF. For each node, it is specified the version of the standard when it was published.

4.2.1 Elementary media

4.2.1.1 Audio

The following audio related nodes are used in ARAF: AudioSource, Sound, Sound2D.

4.2.1.1.1 AudioSource

4.2.1.1.1.1 XSD Description

```
      <complexType name="AudioSourceType">
       <all>
          <element ref="xmta:IS" minOccurs="0"/>
          <element name="children" form="qualified" minOccurs="0">
             <complexType>
                <group ref="xmta:SFAudioNodeType" minOccurs="0"
maxOccurs="unbounded"/>
             </complexType>
          </element>
       </all>
       <attribute name="url" type="xmta:MFUrl" use="optional"/>
       <attribute name="pitch" type="xmta:SFFloat" use="optional" default="1"/>
       <attribute name="speed" type="xmta:SFFloat" use="optional" default="1"/>
       <attribute name="startTime" type="xmta:SFTime" use="optional"
default="0"/>
       <attribute name="stopTime" type="xmta:SFTime" use="optional" default="0"/>
       <attribute name="numChan" type="xmta:SFInt32" use="optional" default="1"/>
       <attribute name="phaseGroup" type="xmta:MFInt32" use="optional"/>
       <attributeGroup ref="xmta:DefUseGroup"/>
   </complexType>
   <element name="AudioSource" type="xmta:AudioSourceType"/>
```

4.2.1.1.1.2 Functionality and semantics

As defined in ISO/IEC 14496-11 (BIFS), section 7.2.2.15.

This node is used to add sound to a BIFS scene. See ISO/IEC 14496-3 for information on the various audio tools available for coding sound.
The addChildren eventIn specifies a list of nodes that shall be added to the children field. The removeChildren eventIn specifies a list of nodes that shall be removed from the children field.

The children field allows buffered AudioBuffer or AdvancedAudioBuffer data to be used as sound samples within a structured audio decoding process. Only AudioBuffer and AdvancedAudioBuffer nodes shall be children to an AudioSource node, and only in the case where url indicates a structured audio bitstream. The pitch field controls the playback pitch for the structured audio, the parametric speech (HVXC) and the parametric audio (HILN) decoder. It is specified as a ratio, where 1 indicates the original bitstream pitch, values other than 1 indicate pitch-shifting by the given ratio. This field is available through the getttune() core opcode in the structured audio decoder (see ISO/IEC 14496-3, section 5). To adjust the pitch of other decoder types, use the AudioFX node with an appropriate effects orchestra.

The speed field controls the playback speed for the structured audio decoder (see ISO/IEC 14496-3, section 5), the parametric speech (HVXC) and the parametric audio (HILN) decoder. It is specified as a ratio, where 1 indicates the original speed; values other than 1 indicate multiplicative time-scaling by the given ratio (i.e. 0.5 specifies twice as fast). The value of this field shall be made available to the structured audio decoder indicated by the url field. ISO/IEC 14496-3, section 5.7.3.3.6, list item 8, describe the use of this field to control the structured audio decoder. To adjust the speed of other decoder types, use the AudioFX node with an appropriate effects orchestra (see ISO/IEC 14496-3, section 5.9.14.4).

The startTime and stopTime exposedFields and their effects on the AudioSource node are described in 7.1.1.1.6.2. The numChan field describes how many channels of audio are in the decoded bitstream.

4.2.1.1.2 Sound

4.2.1.1.2.1 XSD Description

```
      <complexType name="SoundType">
       <all>
          <element ref="xmta:IS" minOccurs="0"/>
          <element name="source" form="qualified" minOccurs="0">
             <complexType>
                <group ref="xmta:SFAudioNodeType" minOccurs="0"/>
             </complexType>
          </element>
       </all>
       <attribute name="direction" type="xmta:SFVec3f" use="optional" default="0
0 1"/>
       <attribute name="intensity" type="xmta:SFFloat" use="optional"
default="1"/>
       <attribute name="location" type="xmta:SFVec3f" use="optional" default="0 0
0"/>
       <attribute name="maxBack" type="xmta:SFFloat" use="optional"
default="10"/>
       <attribute name="maxFront" type="xmta:SFFloat" use="optional"
default="10"/>
       <attribute name="minBack" type="xmta:SFFloat" use="optional" default="1"/>
       <attribute name="minFront" type="xmta:SFFloat" use="optional"
default="1"/>
       <attribute name="priority" type="xmta:SFFloat" use="optional"
default="0"/>
```



```
        <attribute name="spatialize" type="xmta:SFBool" use="optional"
default="true"/>
        <attributeGroup ref="xmta:DefUseGroup"/>
    </complexType>
    <element name="Sound" type="xmta:SoundType"/>
```

4.2.1.1.2.2 Functionality and semantics

As defined in ISO/IEC 14496-11 (BIFS), section 7.2.2.116.

The Sound node is used to attach sound to a scene, thereby giving it spatial qualities and relating it to the visual content of the scene. The Sound node relates an audio BIFS sub-graph to the rest of an audio-visual scene. By using this node, sound may be attached to a group, and spatialized or moved around as appropriate for the spatial transforms above the node. By using the functionality of the audio BIFS nodes, sounds in an audio scene dscribed using ISO/IEC 14496-11 may be filtered and mixed before being spatially composited into the scene. The semantics of this node are as defined in ISO/IEC 14472-1:1997, section 6.42, with the following exceptions and additions.

The source field allows the connection of an audio sub-graph containing the sound. The spatialize field determines whether the Sound shall be spatialized. If this flag is set, the sound shall be presented spatially according to the local coordinate system and current listeningPoint, so that it apparently comes from a source located at the location point, facing in the direction given by direction. The exact manner of spatialization is implementation-dependant, but implementators are encouraged to provide the maximum sophistication possible depending on terminal resources. If there are multiple channels of sound output from the child sound, they may or may not be spatialized, according to the phaseGroup properties of the child, as follows. Any individual channels, that is, channels not phase-related to other channels, are summed linearly and then spatialized. Any phase-grouped channels are not spatialized, but passed through this node unchanged. The sound presented in the scene is thus a single spatialized sound, represented by the sum of the individual channels, plus an “ambient” sound represented by mapping all the remaining channels into the presentation system as described in ISO/IEC 14496-11, section 7.1.1.2.13.2.2. If the spatialize field is not set, the audio channels from the child are passed through unchanged, and the sound presented in the scene due to this node is an “ambient” sound represented by mapping all the audio channels output by the child into the presentation system as described in ISO/IEC 14496-11, section 7.1.1.2.13.2.2.

As with the visual objects in the scene, the Sound node may be included as a child or descendant of any of the grouping or transform nodes. For each of these nodes, the sound semantics are as follows. Affine transformations presented in the grouping and transform nodes affect the apparent spatialization position of spatialized sound. They have no effect on “ambient” sounds. If a particular grouping or transform node has multiple Sound nodes as descendants, then they are combined for presentation as follows. Each of the Sound nodes may be producing a spatialized sound, a multichannel ambient sound, or both. For all of the spatialized sounds in descendant nodes, the sounds are linearly combined through simple summation from presentation. For multichannel ambient sounds, the sounds are linearly combined channel-by-channel for presentation.

4.2.1.1.3 Sound2D

4.2.1.1.3.1 XSD Description

```
      <complexType name="Sound2DType">
       <all>
          <element ref="xmta:IS" minOccurs="0"/>
          <element name="source" form="qualified" minOccurs="0">
             <complexType>
                <group ref="xmta:SFAudioNodeType" minOccurs="0"/>
             </complexType>
          </element>
       </all>
       <attribute name="intensity" type="xmta:SFFloat" use="optional"
default="1"/>
```

```
      <attribute name="location" type="xmta:SFVec2f" use="optional" default="0
0"/>
      <attribute name="spatialize" type="xmta:SFBool" use="optional"
default="true"/>
      <attributeGroup ref="xmta:DefUseGroup"/>
   </complexType>
   <element name="Sound2D" type="xmta:Sound2DType"/>
```

4.2.1.1.3.2 Functionality and semantics

As defined in ISO/IEC 14496-11 (BIFS), section 7.2.2.117.

The Sound2D node relates an audio BIFS sub-graph to the other parts of a 2D audio-visual scene. It shall not be used in 3D contexts. By using this node, sound may be attached to a group of visual nodes. By using the functionality of the audio BIFS nodes, sounds in an audio scene may be filtered and mixed before being spatially composed into the scene.

The intensity field adjusts the loudness of the sound. Its value ranges from 0.0 to 1.0, and this value specifies a factor that is used during the playback of the sound. The location field specifies the location of the sound in the 2D scene. The source field connects the audio source to the Sound2D node. The spatialize field specifies whether the sound shall be spatialized on the 2D screen. If this flag is set, the sound shall be spatialized with the maximum sophistication possible. The 2D sound is spatialized assuming a distance of one meter between the user and a 2D scene of size 2m x 1.5m, giving the minimum and maximum azimuth angles of –45° and +45°, and the minimum and maximum elevation angles of -37° and +37 °. The same rules for multichannel audio spatialization apply to the Sound2D node as to the Sound (3D) node. Using the phaseGroup flag in the AudioSource node it is possible to determine whether the channels of the source sound contain important phase relations, and that spatialization at the terminal should not be performed.

As with the visual objects in the scene (and for the Sound node), the Sound2D node may be included as a child or descendant of any of the grouping or transform nodes. For each of these nodes, the sound semantics are as follows. Affine transformations presented in the grouping and transform nodes affect the apparent spatialization position of spatialized sound.

If a transform node has multiple Sound2D nodes as descendants, then they are combined for presentation. If Sound and Sound2D nodes are both used in a scene, all shall be treated the same way according to these semantics.

4.2.1.2 Image and video

The following image and video related nodes are used in ARAF: ImageTexture, MovieTexture.

4.2.1.2.1 ImageTexture

4.2.1.2.1.1 XSD Description

```
      <complexType name="ImageTextureType">
       <all>
          <element ref="xmta:IS" minOccurs="0"/>
       </all>
       <attribute name="url" type="xmta:MFUrl" use="optional"/>
       <attribute name="repeatS" type="xmta:SFBool" use="optional"
default="true"/>
       <attribute name="repeatT" type="xmta:SFBool" use="optional"
default="true"/>
       <attributeGroup ref="xmta:DefUseGroup"/>
   </complexType>
   <element name="ImageTexture" type="xmta:ImageTextureType"/>
```

4.2.1.2.1.2 Functionality and semantics

As defined in ISO/IEC 14772-1:1997, section 6.22.

The ImageTexture node defines a texture map by specifying an image file and general parameters for mapping to geometry. Texture maps are defined in a 2D coordinate system (s, t) that ranges from [0.0, 1.0] in both directions. The bottom edge of the image corresponds to the S-axis of the texture map, and left edge of the image corresponds to the T-axis of the texture map. The lower-left pixel of the image corresponds to s=0, t=0, and the top-right pixel of the image corresponds to s=1, t=1.

The texture is read from the URL specified by the url field. When the url field contains no values ([]), texturing is disabled. Browsers shall support the JPEG and PNG image file formats. In addition, browsers may support other image formats (e.g. CGM) which can be rendered into a 2D image. Support for the GIF format is also recommended (including transparency).

The repeatS and repeatT fields specify how the texture wraps in the S and T directions. If repeatS is TRUE (the default), the texture map is repeated outside the [0.0, 1.0] texture coordinate range in the S direction so that it fills the shape. If repeatS is FALSE, the texture coordinates are clamped in the S direction to lie within the [0.0, 1.0] range. The repeatT field is analogous to the repeatS field.

4.2.1.2.2 MovieTexture

4.2.1.2.2.1 XSD Description

```
      <complexType name="MovieTextureType">
       <all>
          <element ref="xmta:IS" minOccurs="0"/>
       </all>
       <attribute name="loop" type="xmta:SFBool" use="optional" default="false"/>
       <attribute name="speed" type="xmta:SFFloat" use="optional" default="1"/>
       <attribute name="startTime" type="xmta:SFTime" use="optional"
default="0"/>
       <attribute name="stopTime" type="xmta:SFTime" use="optional" default="0"/>
       <attribute name="url" type="xmta:MFUrl" use="optional"/>
       <attribute name="repeatS" type="xmta:SFBool" use="optional"
default="true"/>
       <attribute name="repeatT" type="xmta:SFBool" use="optional"
default="true"/>
       <attributeGroup ref="xmta:DefUseGroup"/>
   </complexType>
   <element name="MovieTexture" type="xmta:MovieTextureType"/>
```


4.2.1.2.2.2 Functionality and semantics

As defined in ISO/IEC 14496-11 (BIFS), section 7.2.2.86.

The loop, startTime, and stopTime exposedFields and the isActive eventOut, and their effects on the MovieTexture node, are described in ISO/IEC 14496-11, section 7.1.1.1.6.2. The speed exposedField controls playback speed. It does not affect the delivery of the stream attached to the MovieTexture node. For streaming media, value of speed other than 1 shall be ignored.

A MovieTexture shall display frame or VOP 0 if speed is 0. For positive values of speed, the frame or VOP that an active MovieTexture will display at time now corresponds to the frame or VOP at movie time (i.e., in the movie's local time base with frame or VOP 0 at time 0, at speed = 1): fmod (now - startTime, duration/speed) If speed is negative, then the frame or VOP to display is the frame or VOP at movie time: duration + fmod(now - startTime, duration/speed). A MovieTexture node is inactive before startTime is reached. If speed is non-negative, then the first VOP shall be used as texture, if it is already available. If speed is negative, then the last VOP shall be used as texture, if it is already available.