Metadata for Creation and Distribution of Multi-view Video Content

Werner Bailer and Martin Höffernig

Institute of Information Systems
JOANNEUM RESEARCH Forschungsgesellschaft mbH
Steyrergasse 17, 8010 Graz, Austria
firstname.lastname@joanneum.at

Abstract. Recently the production of multi-view video content has attracted growing attention. The main driving force is stereoscopic cinema, but also 3D television is an upcoming application. In this paper we review the metadata requirements for multi-view video content and analyze how well these requirements are covered in existing metadata standards, both in terms of the coverage of metadata elements and the capabilities to structurally describe multi-view video content. The SMPTE metadata standards, MPEG-7 and EBU P_Meta are considered in this survey. We outline the issues that need to be addressed in future standardization activities.

1 Introduction

Media production and distribution workflows are increasingly shifting from a linear chain to flexible and dynamic processes. This is fostered by advanced tools for media creation and manipulation that blur the boundary between production and post-production and by the fact that productions are today often made for a broad range of target media and distribution channels. In addition, production workflows become increasingly distributed, involving many contributors located at different sites. Thus automation of workflows and metadata interoperability between different workflow steps is of growing importance. Previous work has analyzed the metadata needed in the audiovisual media production process (e.g. [1, 2]) and workflow automation based on workflow languages has been proposed, e.g. for movie production in the YAWL4Film[1] project [3].

Recently the production of multi-view video content has attracted growing attention. The main driving force is stereoscopic cinema, but also 3D television is an upcoming application. Multi-view production further increases the amount of material to be handled in the production process. Next to different language, subtitle, age, etc. versions, 3D adds another degree of freedom to the versions that need to be packaged and distributed. As with many emerging technologies, there are competing systems for stereoscopic exhibition that need to be supported, and of course 2D versions still need to be provided for the majority of theaters,

[1] http://www.yawl4film.com/

television and DVD viewers. Thus there is need for better asset management in distribution to support automatic packaging of the variety of versions.

In this paper we review the metadata requirements for multi-view video content and analyze how well these requirements are covered in existing metadata standards. Section 2 discusses the metadata that needs to be represented for multi-view video content. In Section 3 we analyze different relevant metadata standards w.r.t. these requirements. Section 4 summarizes the analysis and presents an outlook on possible future standardization activities.

2 Metadata Requirements

Various types of metadata exist throughout the digital cinema production workflow. These metadata are produced and consumed at different stages of the workflow. Typically the different devices and tools used in the chain also make use of different metadata representations. In some cases the same metadata properties are stored several times in different formats. Multi-source content adds additional requirements to the metadata representation, as the relations between different media elements need to be described (from the high-level fact that these are different views of the same scene down to precise measurements such as camera calibration parameters). We consider a wide range of visual, audio and several classes of descriptive metadata elements that are produced or used in the different stages of the 3D cinema production workflow. Our discussion does not include data derived from the essence that is in its structure similar to audiovisual essence, such as proxies, key frames, depth maps, maps of the scene geometry etc. Such data can be referenced from the description using relational descriptive metadata. The different properties can be related to three different granularities of the content: to the *production*, i.e. the entire set audiovisual content related to one movie production, to the *asset*, i.e. a single piece of audiovisual essence and to a *segment*, i.e. a (spatio)temporal part of audiovisual essence.

2.1 Technical Metadata

A wide range of technical metadata for video and audio is captured or created during the production process, mainly describing the sampling properties of the audiovisual essence and parameters of devices (e.g. cameras) and tools (e.g. encoders) used in the process. For multi-view video content the parameters describing the geometry of the scene and the recording process are of crucial importance. These include camera position and orientation, absolute lengths in scene needed for metric reconstruction and intrinsic camera parameters. As lenses introduce a number of distortions, precise parameters of the lens distortion model are also required. Another important kind of technical metadata is synchronization information between the different audiovisual streams.

2.2 Descriptive Metadata

The following types of descriptive metadata are created and used in the production and distribution process.

Identification Identification information contains IDs as well as the titles related to the content (working titles, titles used for publishing, etc.).

General content properties These are general description metadata items, not related to a specific modality, such as file size, checksums etc.

Production This describes metadata related to the creation of the content, such as location and time of capture as well as the persons and organizations contributing to the production.

Rights Basic rights information and references to more detailed description of rights and licenses.

Publication/distribution This describes metadata related to the creation of the content, such as location and time of capture as well as the persons and organizations contributing to the production.

Process-related Describes steps in the production workflow (e.g. applied tools, settings). Some processing steps may only apply to certain views (or use different parameters for each of the views), e.g. when performing color correction to adjust one view to another.

Content-related Content-related metadata is descriptive metadata in the narrowest sense. An important part is the description of the structure of the content (e.g. shots, scenes).

Relational/enrichment information Describes links between the content and external data sources, such as other multimedia content or related textual sources. For multi-view video content relational information is needed to link related views, calibration sequences for certain views and other captured data, such as e.g. depth maps.

Most of them are not specific to multi-view video content. However, some of these properties apply to all views, while others might differ. For example, the annotation might describe the objects present in the scene. In a certain setup, a background object could be placed in a corner of the scene so that it is not visible in one of the cameras.

3 Support in Standards

The following standards have been identified to be relevant in different stages of the digital cinema production process and have thus been considered in this study:

- SMPTE Metadata Dictionary [4],
- MXF Descriptive Metadata Scheme 1 [5],
- MPEG-7 Multimedia Content Description Interface [6], and
- EBU P_Meta metadata exchange format [7].

In the following, we analyze both the structural support for multi-view video content in these standards as well as the coverage of the metadata elements discussed in Section 2.

3.1 Structural Support for Multi-view Video Content

We have analyzed whether these metadata standards provide structural support for representing multi-view video content, i.e. allow to describe a set of audiovisual streams that capture the same scene from different positions in space and need to be synchronized, but may have different start times and durations, i.e. temporal offsets.

In most standards there is no explicit concept for representing different views of a scene, especially if they do not have the same temporal extent. Due to the longer tradition of multi-channel audio the support for it is typically much better. While it is in most standards possible to find a representation for multi-view video content, such a representation typically involves application defined semantics and several options might exist.

SMPTE MXF and DMS-1. The MXF container specification [8] provides means to represent several streams of the same modality. The MXF Generic Container [9] can have up to 127 visual or audio data items. However, the semantics of multi-view video content cannot be clearly represented. Depending on the semantics to be expressed two approaches can be chosen:

- Content play-list or edit item pattern for all streams, indicating the type of audiovisual stream (e.g. view from a certain camera) in the metadata. This approach is agnostic to the stream representation of the content, i.e. it could be multiplexed into a single item or be represented by several parallel items.
- Alternate packages representing the audiovisual content for a viewpoint. This requires that sources for different views are not multiplexed into one stream and allows accessing each stream separately. However, the semantics for playing several or all of the views is lost in the description and thus application defined.

MXF DMS-1 defines three frameworks for descriptive metadata: The production framework contains metadata related to all clips and all tracks, the clip framework contains metadata related to a single clip and the scene framework contains metadata for a set of related clips. Typically, the clips described as one scene are temporal segments of the same track. For multi-view video content the scene framework is the only one that could be used. However, a scene will then describe a set of temporal clips from a number of tracks that represent the different views. The semantics will be defined by the metadata of the individual tracks (e.g. camera ID) and their temporal relation. There are no means to describe metadata relating the different views (e.g. relative position information).

MPEG-7 provides flexible mechanisms for describing spatiotemporal decompositions of content and to attach metadata to each of the segments. However, as has been pointed out in other context (e.g. [10]), MPEG-7 allows to create descriptions that convey the same semantics but use different description tools and thus potentially cause interoperability problems. As there is no specific concept for multi-view video content the same problem applies here. Media source decomposition tools can be used to describe the decomposition of a content segment into constituent (subsequent) media of tracks (such as views). However, the semantics are not clear due to the following two issues:

- Structural composition: the decomposition of views could happen on any level, i.e. one could decompose the root segment representing the entire production into views and the describe the temporal decomposition (e.g. shots, scenes) separately for each view, or one could create a temporal structure of the content and then decompose each clip into views.
- Specification of decomposition criteria: unfortunately this is not a controlled property but free text, so that the semantics of a media source decomposition (e.g. whether into temporally subsequent media or views) are not well defined.

MPEG-7 provides no standard means to describe metadata relating the different views (e.g. relative position information).

EBU P_Meta The ItemGroup in P_Meta is intended the express the editorial relation of content items. It could be used to describe items representing different views of the content. An explanatory note element is provided to describe the relations informally. P_Meta provides no standard means to describe metadata relating the different views (e.g. relative position information).

3.2 Coverage of Required Metadata Elements

Traditional technical metadata, i.e. properties also needed for single-view content, is well covered by many standards, especially the SMPTE Metadata Dictionary and MPEG-7. P_Meta focuses on content exchange and thus mainly covers the technical properties needed there. The technical properties that are especially relevant for multi-view video content are not yet well supported by existing standards. Some camera calibration metadata elements are included in the SMPTE Metadata Dictionary while lens metadata is largely missing in all the standards investigated. Audio metadata are sufficiently covered in the SMPTE Metadata Dictionary, MPEG-7 and P_Meta.

The general descriptive metadata elements and identification metadata are well covered by all standards. The same holds for production metadata. Basic rights metadata is sufficiently supported by the standards coming from the motion picture and broadcast industries, while MPEG-7 lacks some elements[2], and the situation for publication and distribution metadata is similar. For process related metadata, the SMPTE metadata dictionary provides much better support than the other standards. Basic content description and relational metadata is available in all standards.

[2] Of course MPEG-21 could be used to complement this lack.

	structural	**calibration**	**lens**	**identif., prod.**	**process**	**rights**
SMPTE RP210	n/a	some	no	yes	yes	yes
MXF DMS-1	streams	→ RP210	no	yes	→ RP210	yes
MPEG-7	views (informal)	no	no	yes	no	limited
EBU P_Meta	views (informal)	no	no	yes	no	yes

Table 1. Summary of structural and metadata support for multi-view content in selected standards.

4 Summary and Outlook

We have analyzed the metadata requirements to describe multi-view video content and the coverage of these requirements in existing metadata standards. The analysis has shown that several metadata standards can be used for describing multi-view video content. As shown in Table 1, most of the required elements are covered by at least some of the standards. None of the standards provides supports for lens and some calibration metadata elements, so that one has to revert to proprietary or manufacturer specific solutions in this case. This is of course very unsatisfactory w.r.t. interoperability.

Concerning the structural description of multi-view video content we have identified possible solutions in all of the standards. However, in many cases several possible solutions exist, and the semantics are not defined in the standard. Application specific qualifiers and extensions are required in the structural description, leading to formally standard compliant descriptions, but with application defined semantics. Again, this leads to interoperability issues.

In order to improve the metadata workflow in multi-view content production, and establish interoperability between devices and tools, the following issues need to be addressed in standardization:

- Support the required calibration and lens metadata. These metadata elements are hardware related and need to be embedded with the captured essence. Thus SMPTE RP210 seems to be the appropriate standard for this kind of metadata.
- Structural description. Several standards are capable of describing multi-view content, but the semantics for using the standards' tools for representing multi-view content need to specified. This could for example achieved by defining MPEG-7 profiles.

Acknowledgements

The authors would like to thank their partners in the 2020 3D Media project for their contributions to requirements and information about currently used metadata in the production and distribution workflow. The research leading to this paper has been partially supported by the European Commission under the contract FP7-215475, "2020 3D Media – Spatial Sound and Vision" (http://www.20203dmedia.eu/).

References

1. Schinas, K., Schmidt, W., Höller, F., Zeiner, H., Bailer, W., Hausenblas, M.: D3.2.1 Metadata in the Digital Cinema Workflow and its Standards. Public deliverable, IP-RACINE (IST-2-511316-IP) (2005) `http://www.ipracine.org/documents/Del_3_2_1_metadata.pdf`.
2. Bailer, W., Schallauer, P.: Metadata in the audiovisual media production process. In Granitzer, M., Lux, M., Spaniol, M., eds.: Multimedia Semantics - The Role of Metadata. Volume 101 of Studies in Computational Intelligence. Springer (Jun. 2008) 65–84
3. Ouyang, C., Rosa, M.L., ter Hofstede, A.H., Dumas, M., Shortland, K.: Toward web-scale workflows for film production. IEEE Internet Computing **12**(5) (2008) 53–61
4. SMPTE: Metadata dictionary registry of metadata element descriptions. SMPTE RP210.11 (2004)
5. SMPTE: Material Exchange Format (MXF) - Descriptive Metadata Scheme-1. SMPTE 380M (2004)
6. ISO: Information Technology - Multimedia Content Description Interface (MPEG-7). ISO/IEC 15938 (2001)
7. EBU: EBU P_META 2.0 Metadata Library. EBU Tech 3295-v2 (Jul. 2007)
8. SMPTE: Material Exchange Format (MXF) - File Format Specification. SMPTE 377M (2004)
9. SMPTE: Material Exchange Format (MXF) - MXF Generic Container. SMPTE 379M (2004)
10. Troncy, R., Bailer, W., Hausenblas, M., Hofmair, P., Schlatte, R.: Enabling Multimedia Metadata Interoperability by Defining Formal Semantics of MPEG-7 Profiles. In: 1^{st} International Conference on Semantics And digital Media Technology (SAMT'06), Athens, Greece (2006) 41–55