

A Comprehensive Infrastructure and Workflow for Acquisition of High-resolution Multi-view Content

Werner Bailer, Christian Schober

JOANNEUM RESEARCH, DIGITAL – Institute of
Information and Communication Technologies
8010 Graz, Austria
{firstname.lastname}@joanneum.at

Thomas Brune

TECHNICOLOR –
Research & Innovation
30625 Hannover, Germany
thomas.brune@technicolor.com

Abstract—The current success of immersive 3D experiences in feature films and the trend towards 3D TV require advanced tools and workflows for high-quality capture of multi-view live scenes including depth information. These requirements are not fulfilled by today’s capture workflows and infrastructures based on legacy technology from 2D capture solutions. We propose a comprehensive infrastructure for capture of multi-view content supporting different methods to obtain depth information. The paper describes a device for field capture and tools for online and offline content analysis and browsing of the captured content that have been integrated into the proposed infrastructure.

Keywords—10Gig; depth capture; tracking; browsing.

I. INTRODUCTION

The current success of immersive 3D experiences in feature films and the trend towards 3D TV call for advanced tools and workflows for high-quality capture of multi-view live scenes including depth information. This demand cannot be fulfilled by today’s stereoscopic high resolution capture devices/technology employing proprietary workflows based on legacy technology from 2D capture solutions.

Acquisition processes for multi-view content need to be flexible in order to cover different requirements depending on the type of production or scene. For example, no single depth capture approach is able to provide high-quality results for all types of scenes, thus different approaches need to be supported. Multi-view capture beyond stereo typically involves different types of sensors that provide different output formats. Each of them also has a range of static and dynamic metadata outputs that need to be captured together with the content. These requirements are also not fulfilled by today’s capture infrastructures.

This paper describes a comprehensive infrastructure for capture of multi-view content, supporting different methods to obtain depth information as well as tools for online and offline content analysis and processing of the captured content. Section II gives an overview of different capture methods for multi-view content that need to be supported. Section III then discusses the workflow requirements in more detail and Section IV proposes a comprehensive capture infrastructure capable to support these workflows. In Section V we discuss some selected components, which have

been integrated in this infrastructure, and Section VI concludes the paper.

II. CAPTURE METHODS FOR MULTI-VIEW CONTENT

Today, stereoscopic image capture and play-out produces depth impressions for the spectator. Disadvantageous of the stereoscopic approach is that play-out for multi-view devices is not addressed natively. Multi-view content can be created by means of a depth map of the captured scene, e.g., generated by using the triangulation method [4].

The method of ‘trifocal’ triangulation advances the basic triangulation approach and is subject of investigation within the EU funded project ‘2020 3D Media’¹. To gain high confidence depth maps, the ‘trifocal’ capture method is combined with a ‘structured light’ approach.

A. Trifocal Capture for Triangulation Processing

A trifocal system based on three cameras with common resolution is used to capture three synchronized video streams [4]. A post processing step performs a two-fold triangulation analysis based on the three captured video streams [5] following by a consistency check between the two analysis results. The output is a depth map accompanying the video of the central camera featuring the same resolution, with frame speed of 25 or 30 frames per second (fps) respectively

The data flow and annotated bandwidths of the current demonstrator setup is depicted in Figure 1. The setup contains one high end center camera delivering images in RGB 4:4:4, 10 bit log format as well as in addition two lower quality satellite cameras employing Bayer-patterns and 8 bit resolution. All three cameras feature HD resolution. The triangulation post processing takes place at the ‘consumer’

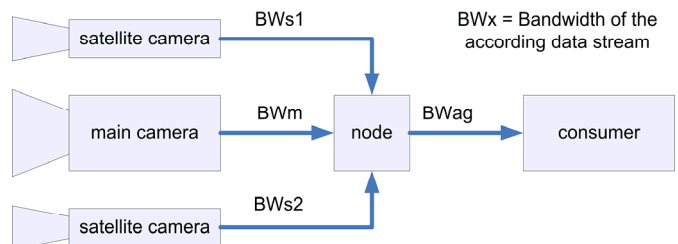


Figure 1. Example of a ONE-COLUMN figure caption.

¹ <http://www.20203dmedia.eu/>

side.

The rough data bandwidths generated by such a trifocal setup of HD cameras at 30 fps are:

- 2 Gbit/s driven by the center camera (BW_m)
- 0.6 Gbit/s driven by each satellite camera (BW_{s1}, BW_{s2})
- Therefore, the aggregated output bandwidth of all three cameras is 3.2 Gbit/s (BW_{ag}).

A possible trifocal set with 4K resolution at 24 fps can consist of a main camera with a 4096 x 2048, 16 bit Bayer sensor and satellite cameras with 8 bit depth. Rough data bandwidths are:

- 3.3 Gbit/s driven by the center camera (BW_m)
- 1.65 Gbit/s driven by each satellite camera (BW_{s1}, BW_{s2})
- Therefore, the aggregated output is 6.6 Gbit/s (BW_{ag}).

B. Structured Light Capture

A 2D deformation analysis of regular pattern structures projected upon voluminous objects can be used to calculate depths of a scene.

For capture of scenes applied with such patterns, setups with multiple HD camera arrangements are currently under investigation to detail the resulting depth maps [6]. The calculations presented in this paper with respect to the ‘structured light’ approach are exemplary based on a four camera setup (Figure 3), where each camera is pointing to the pattern applied scene from a different angle.

To capture a scene with the ‘structured light’ approach one has to separate between the images of the pure scene and the images of the pattern projected on the scene. To cope with movements of objects the capture of the pure scene and the scene with applied projection has to happen in a close series. One solution is to capture the pure scene and the pattern applied scene in a time multiplex manner with double frame rate according to Figure 2. Every even frame is captured pure and every odd frame is captured with applied pattern (grid). The depth map is processed in a post-processing step based upon the four video streams at the ‘consumer’ site.

Therefore the required data bandwidth of a structured light capture system based on e.g. four HD cameras, 8 bit Bayer pattern operating at 60 fps totals to approximately 4.4 Gbit/s.

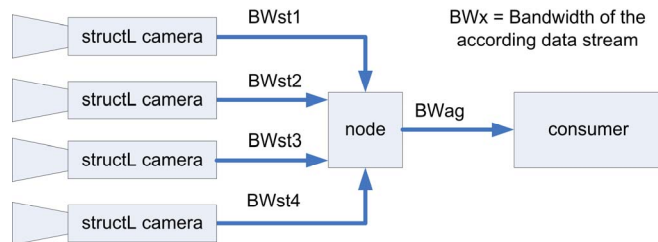


Figure 3. Data flow of an exemplary structured light capture.

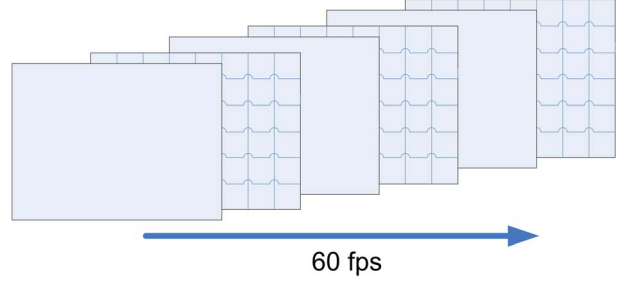


Figure 2. Pure and pattern applied image capture.

III. WORKFLOW REQUIREMENTS

Media production and distribution workflows are increasingly shifting from a linear chain to flexible and dynamic processes. This is caused by the diversity of capture methods and devices as well as 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 broader range of target and distribution formats. Multi-view capture further increases the diversity of sensors involved, each providing different data formats, image resolutions, colour representations, etc. For example, as is discussed in Section 2, there is a range of different methods for capturing depth information, each involving different types of sensors and producing different output data with different bandwidth. The capture infrastructure must be able to support these methods and also allow the usage of several methods in parallel as well as switching from one method to another between scenes.

In addition, many non-image data such as depth measurements, lens and camera parameters, tracking information, etc. need to be captured and transmitted to subsequent steps in the workflow together with and in sync with the essence.

On-set pre-visualization is becoming an increasingly important part of the capture infrastructure. This requires real-time or near-real-time analysis and processing of the captured essence and metadata. Thus it must be easily possible to integrate the equipment performing these tasks into the capture infrastructure in order to give it a low-latency access to the captured essence and metadata.

At the end of the acquisition process, the captured information must be ready for use in post-production. The capture infrastructure needs to support efficient viewing and selection of dailies and indexing of the captured content for efficient content organization and access for the editing stage.

IV. A PROPOSED INFRASTRUCTURE SUPPORTING THE CAPTURE OF MULTI-VIEW CONTENT

A. State of the Art of Video Capture Infrastructure

Today, in the field of professional video capture, most of the video data transmission and recording infrastructure is based on the single link HD-SDI [1], dual link HD-SDI [2]

and the recent 3G-SDI interface standard [3]. Drawbacks of these standards amongst others are the rigid space limits within the ancillary fields for additional data like metadata or device control or the limited amount of defined video standards not providing solutions for resolutions like 2K or 4K, colour depths like 14 bit or 16 bit, multiple frame rates or multiple streams. Also, there is no feedback channel defined in these standards.

To overcome parts of the mentioned drawbacks in the HD-SDI and 3G SDI standards circumventions are already defined, such as the ARRI-RAW T-Link solution [7] providing 12 bit raw Bayer sensor data for processing. Sony and GrassValley have introduced proprietary fiber optical interfaces that provide bandwidth for ‘triple speed’ capture [8][9].

B. Requirements for a New Capture Infrastructure

The preceding sections have discussed the basic needs for a versatile network when setting up capture sets for multi-view content using multiple camera arrangements. In the capture process the combination of the presented different 3D capture technologies has to be taken into account (e.g., as currently investigated in the 2020 3D media project). Aim of the combination is to gain more precise depth results. The mentioned basic needs are leading to the requirements as follows:

- The capture network has to provide headroom for 7.6 Gbit/s aggregated data bandwidth (6.4 Gbit/s, if

two structured light cameras can also be used as ‘satellite’ cameras)

- The network has to deal with multiple camera streams (Additional streams may be introduced dynamically, i.e., metadata and audio)
- Preferably, all camera streams have to be recorded with a single field-recorder.

These requirements are clearly beyond the scope of HD-SDI, 3G SDI, ARRI-RAW T-Link and the ‘triple speed’ fiber standards respectively solutions.

C. Design of a New Versatile Capture Infrastructure

The past few years have shown that the speed of innovation in capture technology is rapidly increasing because of the introduction of HD broadcast, 3D and 4K cinematography. To keep the pace, an appropriate capture infrastructure should behave accordingly in the future. Therefore the basic interface technology has to have a roadmap to avoid dead ends in terms of data bandwidth and flexibility.

Figure 4 shows the proposal for a versatile multi-view capture infrastructure based on 10G Ethernet, because Ethernet is a simple and universal, widely proven and adapted infrastructure. A multitude of logical connections respectively streams can be established between a single device or multiple different devices. They all can share the maximum available bandwidth dynamically. A roadmap in terms of data bandwidth is available with 10 Gbit/s -

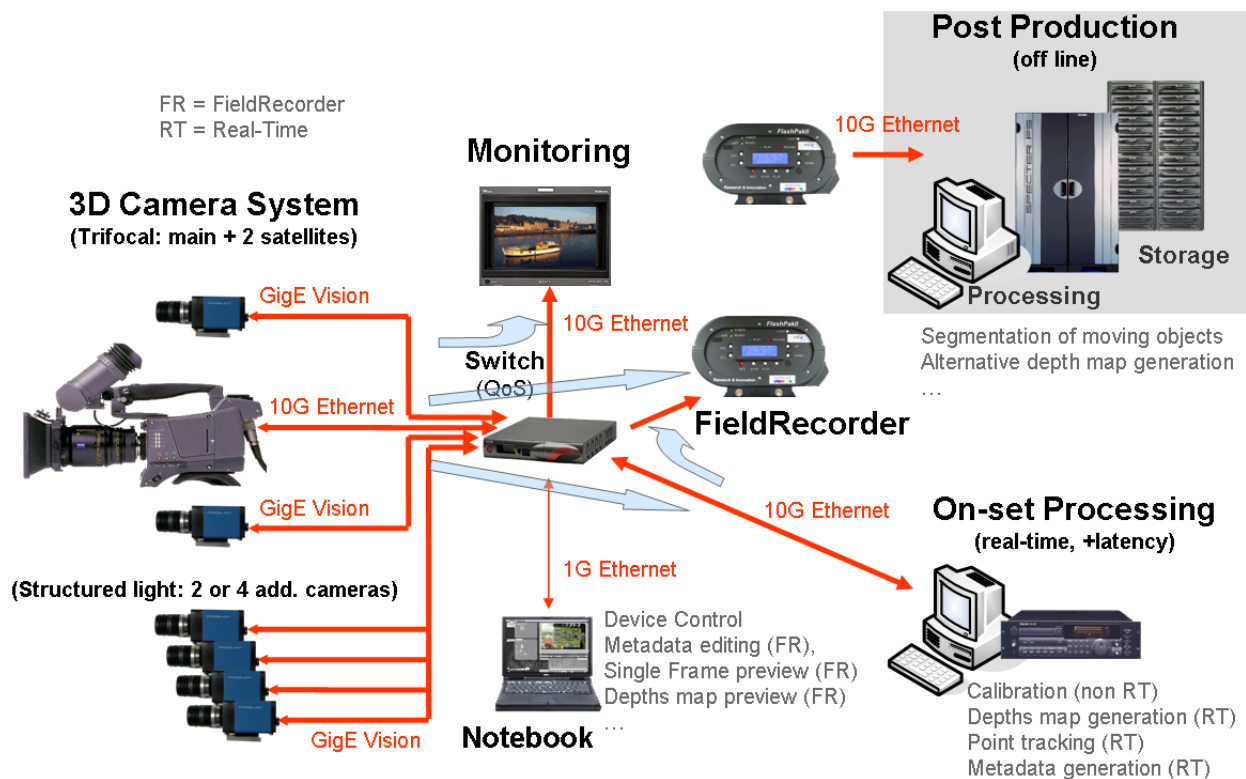


Figure 4. Proposed 3D capture infrastructure.



Figure 6. Capture infrastructure with FlashPakII recorder, preview monitor with 10G Ethernet converter box for the main camera, preview and control notebook for both two satellite cameras and notebook workstation for feature-point tracking (from left to write). The 10G Ethernet switch for stream routing is placed under the table.

currently of the shelf, 40 Gbit/s and 100 Gbit/s where first solutions already appear on the market [10].

In Figure 4 a trifocal 3D camera system features a cinematographic type lead camera and two industrial type satellite cameras. According to chapter 2.2, two or four industrial type cameras are added for a combined ‘structured light’ capture. In case of the satellite and ‘structured light’ cameras, industry standard HD cameras with the so called GigE Vision interface are proposed. The main camera features a 10G Ethernet interface. Therefore, all image capture devices are connected to an off the shelf 10G Ethernet network switch for aggregation purpose. The several resulting logical camera streams are routed via the switch to an on-set processing unit for real-time calculations like depth map and general metadata generation, or e.g. feature point tracking algorithms. All streams are recorded with a field recorder, since a 10G Ethernet switch supports a totally open routing matrix. Device control within the network and several previews taken from the field recorder are controlled by a connected notebook utilizing the feedback channel of the 10G Ethernet network. For moving image preview, a premium HD monitor is connected via 10G Ethernet.

For data transmission from capture to post, the field recorder can be directly connected to the post production facility via 10G Ethernet. It is also possible to either download the data from the field recorder to less expensive hard disk drives or tape for delivery, or to use a WAN link to the post location for large distance data offloads of e.g.,

several thousand kilometers [16]. Figure 6 and Figure 5 are displaying the evaluation of the components during a testbed.

The proposed 10G Ethernet based infrastructure is realized with dedicated 10G Ethernet interface modules in all devices. Regarding workstation and PC systems, there are off the shelf host bus adapter cards available. For all mobile and battery operated devices PCBs have been designed and



Figure 5. Camera rig for trifocal and structured light capture.

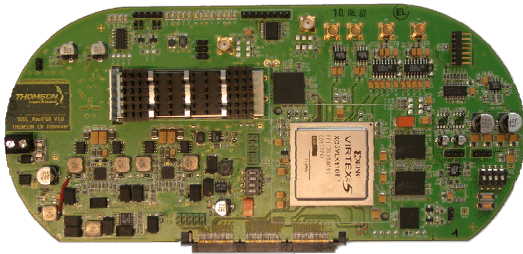


Figure 7. Embedded 10G Ethernet interface PCB.

manufactured that fit into the according device bodies and provide the required low power consumption. An optical 10 Gbit/s transceiver module (XFP), the ASIC for translation of four 3.125 Gbit/s streams to a single 10 Gbit/s stream and the Xilinx™ FPGA are the main building bricks of the ‘high speed’ data path operating with up to 10 Gbit/s net data rate. Layout of the PCB required intensive timing and field simulation to enable deterministic signal propagation for signal frequencies far beyond 5 GHz. Figure 7 displays a 10G Ethernet interface module for a field storage device. A PCB with another shaping is manufactured to fit into camera bodies.

V. SELECTED COMPONENTS

In this section we describe some components, which have been integrated into the proposed infrastructure, in more detail. In particular this is a device for field recording, a device for online feature point tracking as well as solutions for offline content analysis and browsing tools.

A. Field Recording

Since the start of uncompressed field recording with solid state FLASH technology for cinematography in 2005 (Venom FlashPak™ [17]), the demand for stereoscopic capture of 4:4:4 streams came up. This has been addressed recently e.g. by the S.two OB-1 [11] and Codex Onboard [12] recorders. To cope with versatile 3D capture



Figure 8. FlashPakII mit 10G Ethernet interface during playback.

scenarios like the trifocal capture, the ‘structured light’ capture or the combination of both, concepts of simple single stream storage, respectively ‘doubled’ single stream storage have to be enhanced significantly.

To realize a harmonized multi-view workflow from capture to post, the approach of an appropriate field storage has to fulfill the requirements according to chapter 4.2. Therefore, the architecture of the Venom FlashPak™ was significantly enhanced. Beside the introduction of the 10G Ethernet interface module, the internal write and read bandwidth was enhanced towards 1.1 GByte/s. In addition, the support of up to seven parallel data streams was implemented. The user interface provides stream control via IP address configurations. It is foreseen to provide a fast download capability towards post- and storage servers.

B. On-set Processing

One example for on-set processing is tracking of salient feature points, which can be used e.g. for reconstruction of the camera track in case that there is no tracker attached to the camera or for coarse segmentation of moving objects. We have integrated the real-time capable tracking algorithm described in [15] into the architecture described in Section 4.

The computer on which the tracker (cf.

Figure 9) is running is connected with a 1G Ethernet link to the system and receives the streams from the satellite cameras. Real-time tracking with up to several thousand feature points can be performed.

C. Offline Post-production Components

Some analysis and processing components are implemented as offline and non real-time tools as they perform computationally expensive operations. These tools also include an interactive browsing application.

1) *Multi-view content analysis.* Automatic content analysis can complement existing annotations and provide navigation aids or extract metadata from sparsely annotated



Figure 9. On-set processing: Feature point tracking.

content. Such metadata are crucial for search & retrieval of content or browsing applications as the one described in Section 5.3.2. However, in the case of multi-view content we cannot simply apply content analysis tools independently to each of the captured views. The experimental application of this naïve approach and discussions with users have revealed the following requirements: The workflow must be streamlined so that a single tool handling both single- and multi-view content is necessary. Such a tool shall be able to perform temporal alignment of the views if necessary and must produce a single metadata document containing both analysis results valid across views as well as specific to one view. The most relevant information to be described synchronously across views is temporal segmentation information as well as representations of these segments such as synchronous key frames. In addition, low-level visual features useful for retrieval and browsing, e.g. motion activity or color descriptors, shall be described jointly for all views.

Based on these requirements we have designed a multi-step process for performing content analysis on multi-view content (summarized in Figure 10). In the *first content analysis step* the mean visual activity of the scene is determined from the visual activities of the individual streams. Shot boundaries are detected and fused across the streams and shots are associated. This first analysis step also extracts a regularly sampled sequence of color, texture and visual activity descriptors. This information can be used to perform automatic temporal alignment of the video streams from different views in order to determine their temporal offset. We use the variant of the Longest Common Subsequence (LCSS) proposed in [13] for the identification of repeated takes of the same scene. For the view analysis problem we do not permit gaps in the matches and give 80% of the weight for similarity calculation to visual activity and 20% to color features. This makes the approach robust against the differences in visual content and color calibration of the different views.

The *second content analysis step* uses the fusion results as input. The key frame extraction is performed based on accumulated average activity of the scene. Frames are then extracted from each of the streams at synchronous time points. Color, texture and motion descriptors are extracted

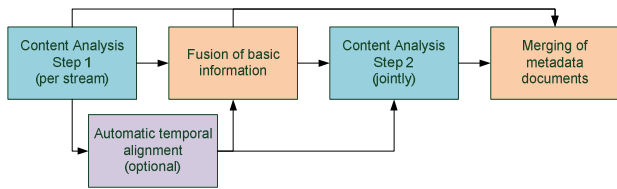


Figure 10: Multi-view content analysis workflow.

from the key frames. The final *merging step* combines all extracted information into one MPEG-7 document, containing stream specific and global descriptions.

The evaluation of the results on a data set from the 2020 3D Media project shows that 86% of the shot boundaries have a temporal offset between zero and five frames between the views. Only about 2% have an offset between 6 and 50 frames, for the others a shot boundary is simply not present in one of the views. Thus using a five frames time window for determining synchronous shot boundaries across multiple views is able to correctly associate nearly 93% of the shot boundaries that actually exist across all views. For key frames extracted based on the visual activity from single views, about 41% have been extracted at identical time points across all views and 65% within an 11 frame time window, but almost 20% of the key frame time points have been detected only in one or some views due to differences in visual activity.

2) *Video browsing.* The production of multi-view content further increases the volume of content that needs to be handled in the production process, especially in post-production, where selection of content takes place. The material is unedited and often very redundant, e.g. containing k takes of the same scene shot by n different cameras. Typically only few metadata annotations are available, mainly technical metadata that can be captured automatically, but only few descriptive metadata elements.

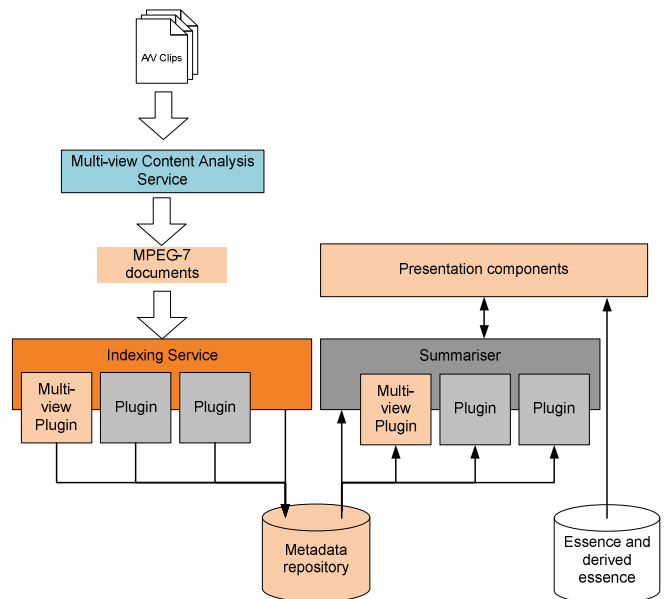


Figure 11: Multi-view content analysis and components of the content browsing framework.

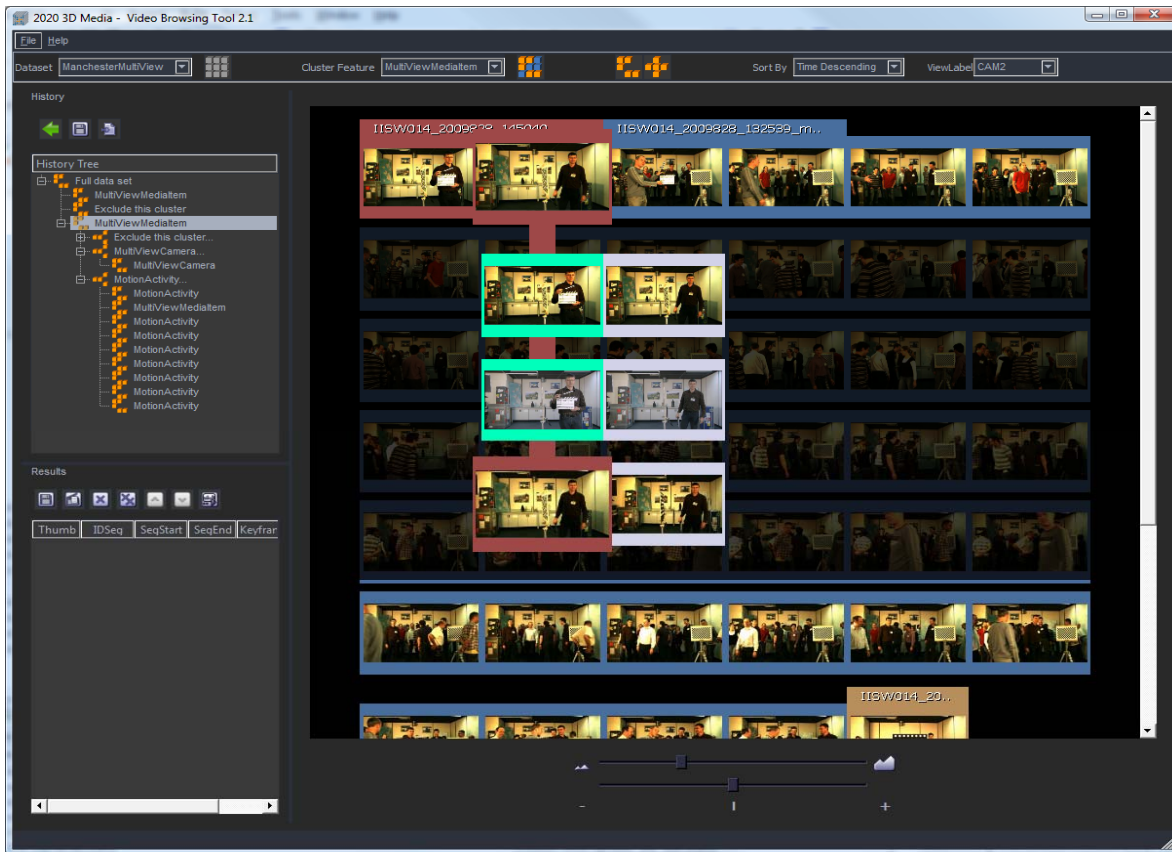


Figure 12: Screenshot of the video browsing tool.

The goal of the development of the video browsing tool is to support the user in navigating and organizing these material collections, so that unusable material can be discarded, yielding a reduced set of material from one scene or location available for selection in the post-production steps.

In order to enable the user to deal with such large amounts of content, it has to be presented in a form which facilitates the comprehension of the content and allows to quickly judging the relevance of segments of the content. Media summarization techniques employing strategies that aim at selecting parts of multimedia content which are expected to be relevant for the user are proposed for that task. The main challenge is therefore to decide which parts of a video are relevant and to select a visual representation for these parts. In the following we briefly summarize the features of the existing video browsing tool and focus on the proposed extensions for multi-view content. More details about the video browsing tool can be found in [14].

The central component of the tool's user interface is a light table view which shows the current content set and cluster structure using a number of representative frames for each of the clusters. The clusters are visualized by colored areas around the images, with the cluster label written above the first two images of the cluster. The size of the images in the light table view can be changed dynamically so that the user can choose between the level of detail and the number

of images visible without scrolling. The tool bar at the top of the screen contains the controls for selecting the feature for clustering and confirming the selection of a subset. The light table view allows selection of multiple clusters by clicking on one of their member images. By double clicking an image in the light table view, a small video player is opened and plays the segment of video that is represented by that image. A screen shot of the browsing tool is shown in Figure 12.

The size of the player adjusts relatively to the size of the representative frames. On the left of the application window the history and the result list are displayed. The history window automatically records all clustering and selection actions done by the user. By clicking on one of the entries in the history, the user can set the state of the summarizer (i.e. the content set) back to this point. The user can then choose to discard the subsequent steps and use other cluster/selection operations, or to branch the browsing path and explore the content using alternative cluster features. At any time the user can drag relevant representative frames into the result list, thus adding the corresponding segment of the content to the result set, which can then be saved as edit decision list (EDL).

a) *Multi-view ingest workflow.* Figure 11 shows the use of the multi-view content analysis service for ingesting content into the browsing tool. The video browsing tool uses the multi-view content analysis tools as a service and monitors a directory for the produced metadata document.

The components shown in orange in Figure 11 are those components of the content browsing framework that were adapted to support multi-view content. The component that is mainly affected is the indexing service. In addition plugins for handling multi-view specific metadata have been added to the indexing service and the summarizer and the structure of the metadata repository has been adapted to hold this kind of information. The indexing service is responsible for ingest of content into the abstraction system. The main change in the indexing service is the support of single metadata documents that contain both separate descriptions for different stream as well as descriptions that apply jointly to all streams. The indexing service must also add information about the relation of the streams to the database. Stream-specific metadata can be supported by the same indexing service plugins that handle single view content, while a new plugin has been developed handling cross-stream metadata.

b) User interface extensions. In order to support multi-view content in the user interface, clustering functionality based on the extracted multi-view features has been implemented, as well as clustering by views and multi-view clips. Also the preview functionality has been extended to show synchronised key frame time lines from the different views (cf. Figure 12). In case of a larger number of views, a subset of views is selected based on visual dissimilarity.

VI. CONCLUSION

This paper presents a comprehensive workflow and infrastructure for capture of high resolution multi-view content. The proposed infrastructure fulfils the requirements of today's capture workflows in terms of heterogeneity of supported devices and tools. Possible setups for capture of multi-view content and depth information according to the single capture paradigm are analyzed in terms of image generation, temporal, and spatial relationship of images, data bandwidth, and workflow integration. The proposed capture infrastructure is based on 10G Ethernet network technology. Implementation and the concept evaluation under real-time conditions are detailed.

Different devices and tools that have been integrated into this infrastructure are discussed in more detail. These include a new field recorder demonstrator, the FlashPakII, real-time feature point tracking and offline content analysis and browsing. The successful integration of these components demonstrates the ability of the proposed infrastructure to support diverse devices and tools that produce and consume a range of different essence and metadata. The main result of this investigation is a seamless data and metadata workflow from capture to post for universal 3D setups.

ACKNOWLEDGMENT

The authors would like to thank their partners in the "2020 3D Media" project for the collaboration. The research described here has been partially supported by the European Commission under the contract FP7-215475, "2020 3D Media" <http://www.20203dmedia.eu/>.

REFERENCES

- [1] HD-SDI: SMPTE 292M (SMPTE 274M)
- [2] Dual link HD-SDI: SMPTE 372M
- [3] 3G-SDI: SMPTE 424M
- [4] R. Tanger, 'Beyond Stereoscopic 3D'. Presentation at Dimension3 expo 2008, Chalon-sur-Saone, France. <http://www.20203dmedia.eu/materials/Dimension3-2020.pdf>
- [5] O. Schreer, N. Brandenburg, S. Askar, and P. Kauff. 'Hybrid Recursive Matching and Segmentation-Based Postprocessing in RealTime Immersive Video Conferencing'. In Proceedings of VMV2001, pp. 383-390, Stuttgart, Germany, November 2001
- [6] Y. Wang, K. Liu, Q. Hao, D. Lau, and L. G. Hassebrook. 'Multicamera Phase Measuring Profilometry for Accurate Depth Measurement', in Sensors and Systems for Space Applications: Camera Based Sensing. Proceedings of SPIE, B. E. Rogowitz, T. N. Pappas, and S. J. Daly; Orlando, Florida, Vol. 6555, Apr. 2007
- [7] <http://archiv.arri.de/arriraw/index.html>
- [8] Sony Super Motion Camera with 10 Gb/s optical fiber interface <http://pro.sony.com/bbsc/ssr/cat-broadcastcameras/cat-hdstudio/product-HDC3300R/>
- [9] GrassValley LDK8300 Live Super SloMo Camera with digital fiber transmission http://www.grassvalley.com/products/ldk_8300
- [10] <http://www.networkworld.com/news/2010/043010-interop-40g-ethernet.html>
- [11] S.two OB-1 field recorder, http://www.stwo-corp.com/Products/?item=OB_1
- [12] Codex Digital, Codex Onboard Recorder, <http://www.codexdigital.com>
- [13] W. Bailer, F. Lee, and G. Thallinger. "A distance measure for repeated takes of one scene," The Visual Computer, vol. 25, no. 1, pp. 53-68, Jan. 2009.
- [14] W. Bailer, W. Weiss, G. Kienast, G. Thallinger and W. Haas. "A Video Browsing Tool for Content Management in Post-production". International Journal of Digital Multimedia Broadcasting, Mar. 2010.
- [15] H. Fassold, J. Rosner, P. Schallauer, W. Bailer. "Realtime KLT Feature Point Tracking for High Definition Video". GravisMa workshop, Plzen, 2009.
- [16] http://www.asperasoft.com/en/industries/digital_media_10/Broadcast_Entertainment_Media_5
- [17] <http://www.grassvalley.com/docs/DataSheets/cameras/venom/CAM-1023D-1.pdf>