SIP (2015), vol. 4, e15, page 1 of 10 © The Authors, 2015.

doi:10.1017/ATSIP.2015.18

INDUSTRIAL TECHNOLOGY ADVANCES

Free viewpoint video (FVV) survey and future research direction

CHUEN-CHIEN LEE, ALI TABATABAI AND KENJI TASHIRO

Free viewpoint video (FVV) is one of the new trends in the development of advanced visual media type that aims to provide a new immersive user experience and interactivity that goes beyond higher image quality (HD/4K TV) and higher realism (3D TV). Potential applications include interactive personal visualization and free viewpoint navigation. The goal of this paper is to provide an overview of the FVV system and some target application scenarios. Associated standardization activities and technological barriers to overcome are also described. This paper is organized as follows: a general description of the FVV system and functionalities is given in Section I. Since an FVV system is composed of a chain of processing modules, an in-depth functional description of each module is provided in Section II. Examples of emerging FVV applications and use cases are given in Section III. A summary of technical challenges to overcome for wider usage and market penetration of FVV is given in Section IV.

Keywords: Multi-view 3D reconstruction, Shape, Texture, 4D video, Free viewpoint

Received 10 April 2015; Revised 19 September 2015

I. INTRODUCTION

Figure 1 illustrates the evolution of visual media types in terms of image quality (SD → HD → 4K), realism, depth sensation, and interactivity (2D → 3D → 4D). A key characteristic of earlier systems is that a desired two-dimensional/three-dimensional (2D/3D) scene can only be viewed from a fixed viewpoint and they usually lack the capability to manipulate interactively the viewpoints of captured 2D/3D scene.

For the past several years and because of advances made in computer graphics, computer vision and multi-view/3D multimedia technologies, a new type of interactive video navigation and visualization has been gaining popularity and is becoming available in the professional and consumer markets (e.g. bullet time movie and 360° panoramic video for VR head-mount (https://www.youtube.com/watch?v=82mXrQWlW38 https://www.youtube.com/watch?v=zKtAuflyc5w https://www.oculus.com/gear-vr/).) Navigation range offered by initial commercial products has been limited to a simple linear/angular change along the linear trajectory between capturing devices (e.g. camera array). Modern professional systems have however started to demonstrate a more flexible range of viewpoint navigation, independent of capturing device placements (e.g. sportscast, immersive concert streaming as shown in (https://www.youtube.com/watch?v=l_TxrOxCPSg&feature=youtu.be https://www.youtube.com/watch?v=h-7UPZg_qOM).) International Standardization Organizations, ISO-MPEG, and ITU-T VCEG, have also been working on various aspects of free viewpoint visual media technology, known as Free Viewpoint Television (FTV) standardization [1–5]. More specifically, Multi-view Video Coding (MVC) [6] is developed for efficient compression of multi-view cameras in phase 1 of FTV standardization activity. Phase 2 of FTV standardization, called 3DV, supports generation of virtual view(s) from small number of coded camera views together with their associated depth map(s) [7]. MPEG is currently considering the third phase of FTV standardization and is planning to issue CfE (Call for Evidence) in June 2015 [1].

A) Free viewpoint video (FVV)

As mentioned earlier, FVV is an advanced visual media type that offers flexible viewpoint navigation in 3D space and time (4D video) from multi-view captured video. The key benefit of FVV is interactivity allowing users, rather than broadcasters or content creators, to control the desired viewing angles and positions. A description of FVV system is given below.

B) FVV system

As described in [8] and shown in Fig. 2, a typical FVV system is composed of the following modules:

Sony US Research Center, San Jose, CA 95112, USA

Corresponding author:
A. Tabatabai
Email: ali.Tabatabai@am.sony.com

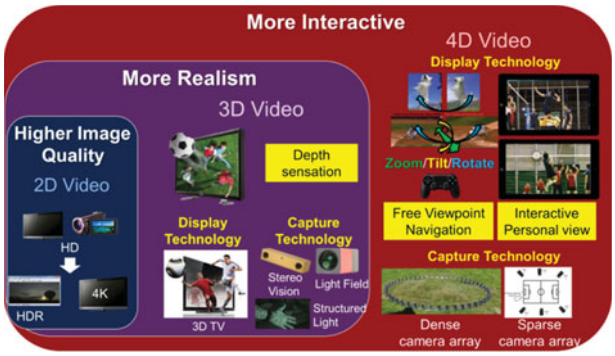

Fig. 1. Evolution of video capture and display system.

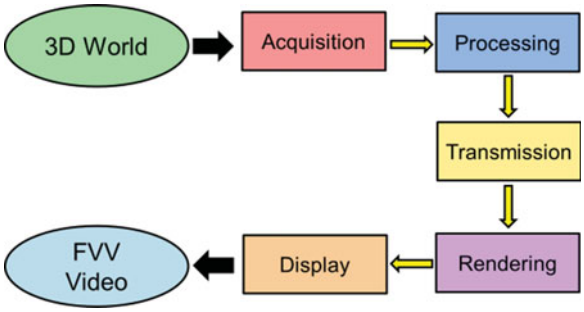

Fig. 2. FVV system overview (Smolic [8]).

- Acquisition module: Main functionality is to capture the 3D scene. It consists, in general, of multiple capture devices (e.g. sensors and camera arrays.)
- Processing module: It is used to convert captured data signals (e.g. image data) into a data format suitable for 3D scene representation (i.e. volumetric shape and texture) (Table 1).
- Transmission module: It provides a compact representation format of the 3D data for streaming and/or storage.
- Rendering module: It synthesizes the decoded 3D data for visualization and navigation at a desired scale in viewing and time.
- Display module: It presents rendered video to enhance user's scene immersive experience. It also provides user interface (UI) tools for viewpoint navigation.

A brief description of each module follows.

II. SYSTEM MODULE OVERVIEW

A) Acquisition module

Acquisition module consists of a number of capturing devices (e.g. camera array) characterized generally by three key configuration parameters: (1) topology, (2) synchronization, and (3) calibration.

1. Topology (inward versus outward)

The choice of topology depends largely on the goal of the target viewing content in scene production i.e. whether the main target and focus of immersive viewing is on the background (BG) or on the foreground (FG).

In case of BG viewing target, capturing devices are placed in an outward looking manner, as shown in Fig. 3. Scene capture in this case is typically done by a 360° spherical camera that consists of multiple sensors and optical modules.

For FG target viewing, capturing devices are instead placed in an inward looking manner, as shown in Fig. 4. Scene capture is typically done by either dense or sparse capturing devices such as camera array.

2. Synchronization

In order to capture dynamically moving or non-rigid targets, from multiple cameras, it is essential to maintain a degree of synchronization between cameras.

Wilburn *et al.* [9], for example, has shown that it is possible to maintain 1.2 μs synchronization accuracy between 100 cameras, by a single source of hardware trigger. Since synchronization is performed by wired connection between cameras, this approach may not be practical for consumer applications. An alternative and more cost-effective approach is based on GigE-Vision2 industry standard [59], which has an optional support for Precision Time Protocol (PTP, IEEE 1588). It has however some difficulties to guarantee predictable trigger latency and it is also necessary to have wired Ethernet connection between capturing devices. To eliminate the need for wired connections, Meyer *et al.* [10] has proposed a wireless and cost-effective approach that can achieve 0.39 μs synchronization accuracy by using GPS time sync via Wi-Fi. Use of audio information [11], is another way by which the need for wired trigger can be eliminated.

3. Calibration

With any multi-capturing devices, intrinsic and extrinsic camera calibrations are required, for 3D shape reconstructions and wider viewpoint navigation. Various calibration approaches are available in public domain and the detail survey of these approaches is out of the scope of this paper.

It should be noted that there are certain approaches for which it is possible to avoid calibration step. For instance, if view interpolation is expected along the physical boundary of camera arrays rather than 3D shape reconstruction, then calibration can be avoided. A second approach to avoid calibration is to apply structure from motion (SfM) (e.g. [12, 13]). SfM can estimate, over time, poses and positions of capturing device as well as 3D structure of the scene from captured trajectories. This approach is appropriate if the target object remains static/rigid. It cannot however be easily applied to dynamically moving or non-rigid objects that are typical of FVV applications.

It should also be emphasized that camera calibration error greatly affects the quality of 3D shape reconstruction. Furukawa and Ponse [14] have proposed a method to refine camera calibration from the multi-view stereo (MVS) system. A possible shortcoming of this approach is that it may have some difficulties dealing with sparse camera array.

https://doi.org/10.1017/ATSIP.2015.18 Published online by Cambridge University Press

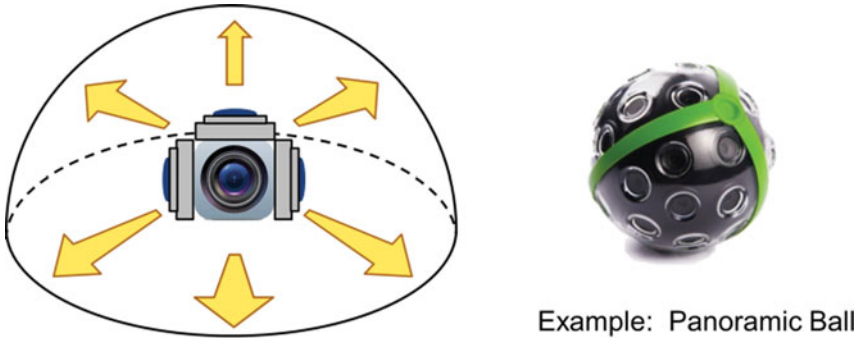

Fig. 3. Outward FVV.

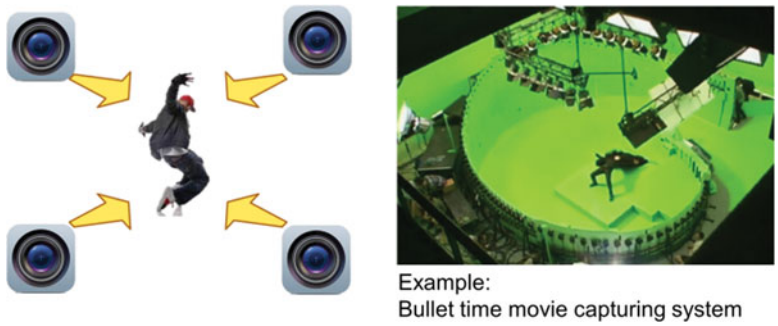

Fig. 4. Inward FVV.

In summary, camera calibration complexity due to high setup and installation cost is one of the main reasons that current applications of FVV have been limited to professional domain, only.

B) Processing module

1. Image-based versus model-based

In general, for an FVV system, with dense camera array, one can reduce computational complexity by applying image-based approaches (e.g. image stitching and view interpolation). In contrast, an FVV system with sparse camera array, tends to be model-based because of low correlation that exists between widely separated base-line of cameras [1, 8, 56, 57] (Table 1).

a) Image-based

In outward FVV systems, it is common practice to use image-based view interpolation methods to create 360° panoramic and photorealistic BG scene views. In such case, flexibility of viewpoint navigation will be limited to angular viewing only due to the lack of 3D scene geometry.

Table 1. Comparison of different 3D scene representations (Smolic [8]).

	Number of cameras	Viewpoint navigation range	Challenge
Model-based	Sparse	Wide	• Visual quality • Computational complexity
Depth-based	Medium	Medium	• Depth error prone
Image-based	Dense	Limited	• High bitrate

The inward FVV system, with dense camera array, can offer view interpolation with high image quality. However, if the target virtual viewpoint moves further, relative to capturing device, image quality could suffer due to limited 3D scene geometry information. Inward FVV systems incorporating depth (e.g. RGB_D or time-of-flight depth-camera) can reduce density of camera array and extends the potential range of viewpoint navigation [8, 15, 16]. Use of such depth cameras is limited to indoor scenes due to the use of near-infrared active sensing.

b) Model-based

In contrast with image-based approaches, model-based or geometry-based approaches offer a wider viewpoint navigation range.

For dense number of capturing devices, multi-view stereo (MVS) is a common technology that utilizes silhouette, texture, and shading (e.g. occluding contour) cues for 3D shape and texture reconstruction [17, 61]. However, effective application of the MVS system requires much shorter camera base-line for feature matching.

Whereas for sparse number of capturing devices one needs to exploit 3D reconstruction beyond stereo matching algorithm, due to low correlation between captured views.

We elaborate further about the 3D shape reconstruction and texture mapping.

2. 3D shape reconstruction

The general approach for 3D shape reconstruction is to use a number of visual cues, known as "Shape from X", in computer vision literature. It includes shape from silhouette (SfS) (silhouette cue), shape from texture (texture cue),

https://doi.org/10.1017/ATSIP.2015.18 Published online by Cambridge University Press

and shape from shading (shade cue). Moreover, if the target shape is known as *a priori*, 3D priori model can also be applied. We will discuss briefly about each approach.

a) Silhouette cue

Visual hull (VH) or SfS is a commonly used method for 3D shape reconstruction [18–21]. As mentioned earlier, this is mainly due to lack of correlation between widely separated views that could cause breakdown of feature-based matching methods. The first step in SfS is to extract silhouette(s) of the target FG object(s) from BG. This is usually done based on priori modeling of BG and by its subtraction from captured image. This process is known as background subtraction (BGS) and various algorithms have been proposed for both indoor and outdoor scenes [22]. Although extracted silhouette may be sufficient for FG detection, it may not be precise for 3D shape reconstruction due to the outliners of silhouette detection (e.g. excessive or missing body parts) and the low-pass filtering nature of BGS. To simplify the process of BGS, typical FVV capture systems place uniform color BG (e.g. green BG as shown in Fig. 4). This kind of simplification is not always possible in an outdoor scene. In [23, 24], a more robust SfS algorithms have been proposed. The goal is to improve silhouettes due to errors in camera calibration or inconsistent silhouettes.

Besides, silhouette-based VH produce low shape surface details due to its inability to recover concavity [19].

b) Texture cue

To address lack of surface details associated with SfS, photo consistency-based approach, as a complement to SfS, has been proposed [25–27] (Fig. 5). The general idea behind photo consistency is to enforce color consistency across views in addition to application of SfS or feature matching (e.g. SIFT [28] and SURF [29]) due to low feature correlation between sparse views. Surface detail refinement can be performed based on joint optimization of silhouette and photo consistency using Graph cut or Level-set [30, 31]. It should be pointed out that despite of its simple algorithm photo consistency has nevertheless high computational run-time due to many repetitive operations.

Fig. 5. VH(silhouette only) and improvement by joint-optimization of silhouette+photoconsistency (Esteban and Schmitt [31]). Left: silhouette based visual hull (low concavity). Middle: Silhouette and photoconsistency joint optimization by level-set. Right: Original input image.

Standard SfM tends to rely on spatially robust texture and corner feature points associated primarily with static or rigid objects [12, 13]. Torresani *et al.* [32] have proposed non-rigid SfM by learning temporal dynamics of object shape. In his approach, temporally robust feature points are manually assigned. For estimation of 3D shape deformation, automatic detection of sufficient number of such feature points becomes a challenging task, however.

c) Shading cue

Traditional shape from shading assumes the existence of a calibrated light source or a known surface material with no color or texture [33, 34]. Recently, Barron and Malik have proposed SIRFS model [35] with no priori assumption about shading, shape, light, and reflectance and has applied it to a single image. With the availability of RGB-D, the quality of shading and shape can be further refined for higher accuracy [36]. This state-of-the-art technique has nevertheless some difficulties in dealing with textured objects in natural light and outdoor conditions (Fig. 6).

d) 3D Priori Model

Presumably, with the target shape known as *a priori*, a more robust 3D shape reconstruction is possible. With the exception of human body model, there is however a limited number of priori 3D shape available. FVV application would typically require realistic surface-level reconstruction (e.g. skin and clothing) going beyond the skeleton model.

Various realistic and morphable 3D human body models have been proposed in the literature [37–39]. Anguelov *et al.* [37] have proposed a popular SCAPE model. It is based on the 3D scanned human body template model for estimation of the articulated pose and non-rigid shape deformation.

Using SCAPE or other 3D morphable models, various researchers have addressed non-linear 3D deformation and registration algorithm from 3D to multi-view 2D images [40–45]. Main advantage with these approaches is to provide temporally robust shape, particularly on non-rigid body parts (e.g. arms and legs) as shown in Figs 7 and 8 [41] and Fig. 9 [42]. Main remaining challenge include scalability of the model: a typical non-scalable system requires generating a new, non-trivial 3D template per each subject for the specific type of activities. Use of generic 3D template model tends to lose specifics of individuals with minimum shape surface details and they tend to look rather graphic-like than photorealistic.

3. Texture mapping

In addition to 3D shape, texture mapping is a critical factor affecting the visual quality of the reconstructed 3D scene and it is often the source of computational bottle-neck in the computer graphics (CG) processing flow [46]. In a typical CG processing chain, the texture is projected from multiple cameras' pixel data to the entire surface of the model using proximity or directional optimization between cameras and

https://doi.org/10.1017/ATSIP.2015.18 Published online by Cambridge University Press

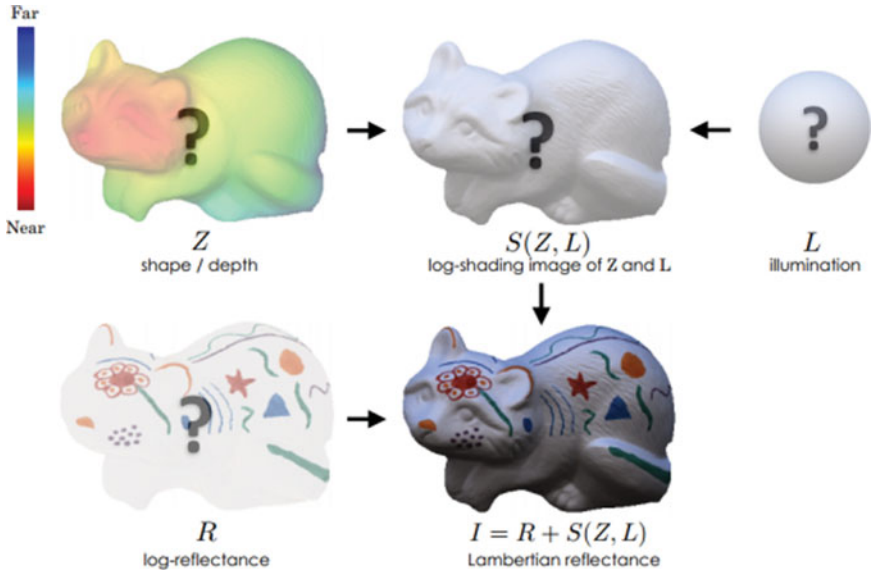

Fig. 6. SIRFS model for Shape from Shading (Barron and Malik [35]).

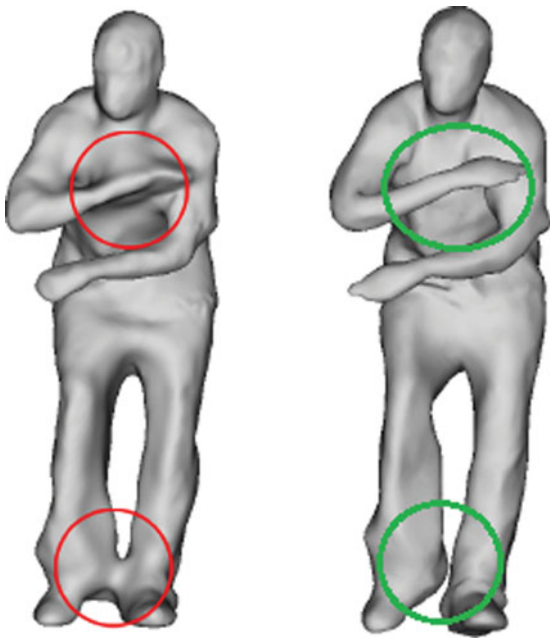

Fig. 7. 3D priori model improves non-rigid body parts. (Left: visual hull, Right: 3D priori model) (Vlasic *et al.* [41]).

surface normal. This view-independent texture mapping can generate universal dataset, independent of virtual viewpoint. For potentially higher texture quality optimized per particular viewpoint, texture map can be generated in a view-dependent context [18, 47–49]. Hisatomi [60] shows that view-dependent approach can suppress the impact of geometrical errors by processing images of cameras visible and close to the viewpoint. However, view-dependent texture mapping typically requires higher computational complexity due to the need for generating new virtual view every time viewpoint changes.

The key challenges with texture mappings are occlusion [8] and texture seam [47, 48, 50]. It is worth to note that unless proper texture synthesis approaches are applied, occlusion and texture seam could cause further degradation in visual quality, as the camera array becomes sparser (Table 2).

a) Occlusion

Some parts of the target object may be invisible by all capturing devices. In that case, the texture needs to be filled by in-painting or some other type of image completion [8] to allow 360° virtual viewpoint. Occlusion also needs to be handled in a temporally consistent manner for video, unless the time is stopped during the viewpoint navigation (e.g. bullet time movie).

b) Texture seam

Since texture on the target object is projected from multiple camera views and due to potential camera calibration and estimated shape errors, a globally mapped texture becomes blurry or duplicated near border lines (i.e. contribution of texture projection switches from one camera to another).

Various approaches have been proposed to minimize texture seam. Eisemann [50] and Casas *et al.* [47] (Fig. 10) have proposed local texture alignment using optical flow. Takai *et al.* [48] deform the multi-view images based on the virtual camera position to harmonize the texture mapping while maintaining the 3D shape and camera calibration.

C) Transmission

For smooth rendering of virtual views and scene navigation, a traditional multi-view system consists of a large number of sensor (camera) views. In [9], for example, a custom array of 100 cameras is constructed, where each three cameras are connected to PC for handling high bandwidth video data. From cost and efficiency point of view it is therefore necessary to develop appropriate visual coding and representation formats for efficient transmission and rendering of arbitrary views.

ISO/IEC MPEG and ITU-T VCEG International standard bodies have been engaged in the development of

https://doi.org/10.1017/ATSIP.2015.18 Published online by Cambridge University Press

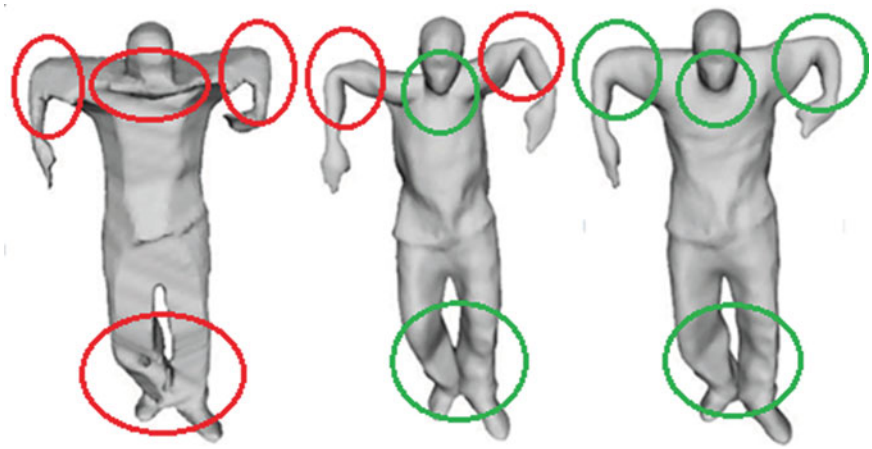

Fig. 8. Visual quality of shape reconstruction by Visual hull (Left), linear articulation of 3D model (Middle), and non-rigid deformation of 3D model (Right) (Vlasic *et al.* [41]).

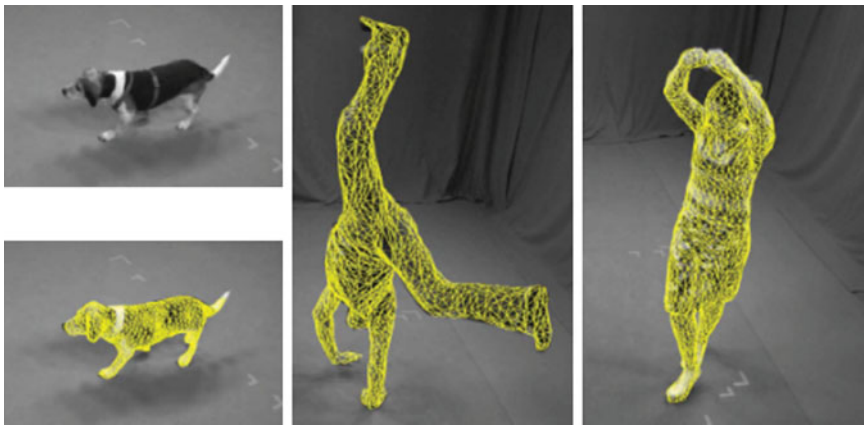

Fig. 9. Gall [42] captures the motion of animals and humans accurately by non-linear surface deformation of 3D priori model.

Table 2. Comparison of texture-mapping approach.

	Pros	Cons
View independent	• Universal data independent from viewpoint • Typically lower complexity	• Higher impact of geometrical errors
View dependent	• Optimized per particular viewpoint • Lower impact of geometrical errors	• Typically higher complexity due to generating virtual view every time viewpoint changes

efficient representation and compression of such data as early as 1996 [51]. More recent activities include phase 1 of FTV (MVC) which started in March of 2004 and was completed in May of 2009 and it is based on the extension of H.264/MPEG-4 AVC [6]. Subsequently, in phase 2 of FTV, multi-view extension of HEVC, known as MV-HEVC [7], was developed and completed in July of 2014, by JCT-3 V. Compared with MVC, MV-HEVC has a higher coding efficiency and provides means for optional coding of depth data associated with each view – see Fig. 12. MV-HEVC is in particular well suited for delivery of 3D content for autostereoscopic displays for which many views are needed for scene immersive visual experience and sensation.

(a) (b) (c) (d) (e)

Fig. 10. Texture local alignment by optical flow (Casas *et al.* [47]). (a, b) camera captured walk and run views; (c) naïve direct blend of textures to a geometric proxy at virtual view; (d) optical flow and (e) proposed alignment.

https://doi.org/10.1017/ATSIP.2015.18 Published online by Cambridge University Press

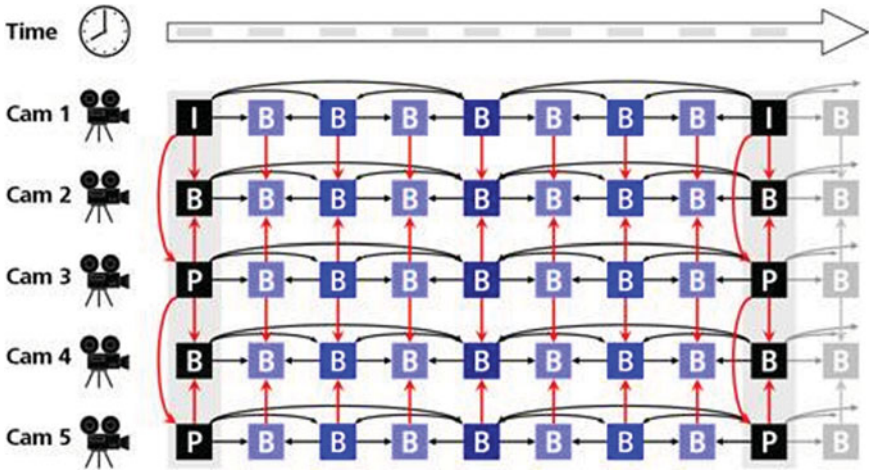

Fig. 11. Temporal/inter-view prediction for MV-HEVC. (Smolic [8]).

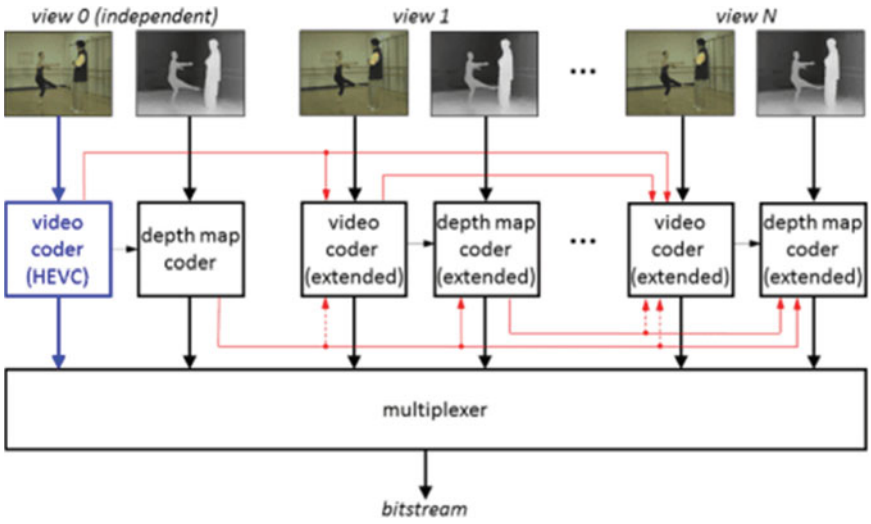

Fig. 12. Simplified block diagram of MV-HEVC [52].

As illustrated in Fig. 11, in either MVC or MV-HEVC correlation across views and time are used for better prediction; thus, providing an efficient compression of multi-view video.

For coding of 3D model data, MPEG-4 provides a process by which 3D (graphics) content is represented and coded. It is known as Animation Framework eXtension (AFX) [53]. Similarly, MPEG has also specified a model for interfacing AFX 3D graphics compression tools to graphics primitives defined in other standards referred to as 3DCG [54].

ISO/IEC MPEG is currently in phase 3 of FTV targeting free navigation (walk-through or flying through experience) and super multi-view (ultra-realistic 3D viewing) [58].

D) Display

It should be emphasized that a key feature of FVV is to gain control over virtual view point and direction, based on personal preferences. A side effect of increased viewing controllability is higher viewing complexity. This is because the FVV system can generate infinite number of views despite the fact that human eyes can only see one to two views at a given time instant [5]. A key challenge is therefore to provide users with natural and user friendly UI for viewpoint navigation by taking into account the above property of human visual system.

Various types of display devices can present FVV video content [2]. Display devices of first type include traditional 2D/3D monitors (e.g. TV, laptop, smartphone, and VR head-mount display), which could be equipped with UI for viewpoint navigation (e.g. joystick, mouse, head/eye-tracker, remote controller, and touch panel). This type of display devices provide single user interface. Second type of display devices provide all the 360° views and multiple users are able to see any views by changing their locations. Yendo [55] proposed 360° ray-producing display that allows multiple viewing of FVV videos.

III. APPLICATION EXAMPLES

As part of current MPEG FTV activities and discussions, two application scenarios are being considered: super multi-view display (SMV) and free navigation (FN) [1].

https://doi.org/10.1017/ATSIP.2015.18 Published online by Cambridge University Press

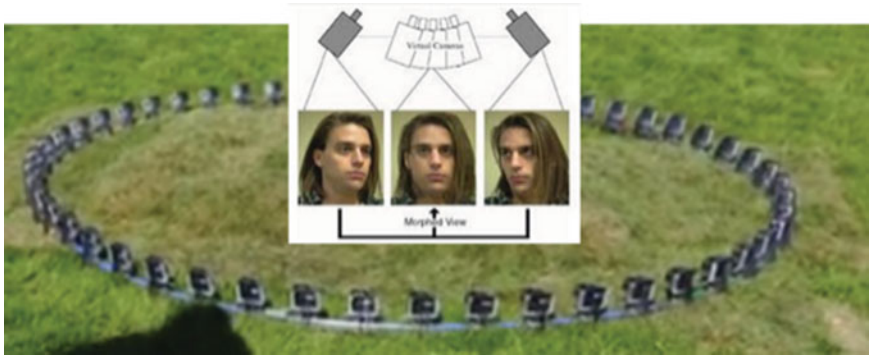

Fig. 13. 1D arc dense camera array for SMV (Lafruit *et al.* [1]).

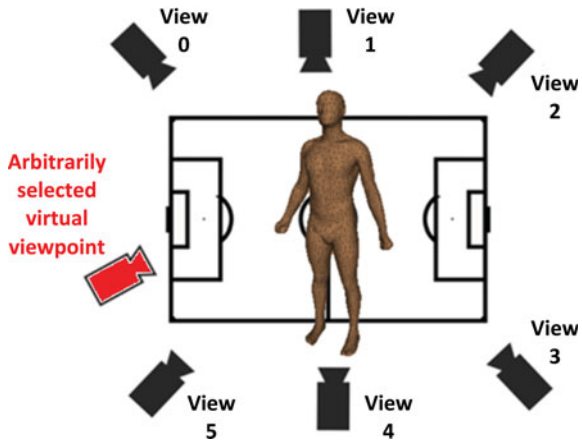

Fig. 14. Sparse camera array for FN (Lafruit *et al.* [1]).

A) Super multi-view display

The system is constructed by 1D or 2D dense camera array (typically 80 cameras.) The main challenge here is to address high bandwidth transmission. Video synthesis technology is an image-based approach (e.g. view interpolation), and therefore, the system requires a large number of cameras, and the viewpoint change is typically limited around physical camera array (e.g. angular viewpoint change, if the camera topology is 1D arc) (Fig. 13).

B) Free navigation

The system is constructed by sparse number of cameras (typically 5–10 cameras.) Due to a low correlation between widely separated views, the view synthesis technology is based on geometry or model-based approach (e.g. 3D reconstruction). Emphasis of FN is on visual quality of CG rendering and view synthesis (Fig. 14).

IV. DISCUSSION

Generally, the key issue that could cause barriers to wide market penetration and roll-out of FVV system can be summarized as high set-up cost and difficulty in system operation. In other words, it is the burden of installing and calibrating multiple cameras and performing non-trivial video operations at high bandwidth. Innovative solution(s) to this issue calls for collective efforts across industry (i.e. capture, authoring, and display), standardization (interoperability, transmission, and storage format), and academic research (processing and rendering). Herewith, we are listing some of the technical challenges that need to be overcome and for which new technologies and standards need to be introduced:

- *Automated or dynamic calibration of capture devices*: Time-consuming multi-view camera calibration is key bottle-neck for the consumer. For example, recalibration becomes necessary, if any camera is moved.
- *Photorealistic view synthesis by sparse camera array*: Focus of standard activities has been mostly on image-based view interpolation with dense camera array and point-cloud data representation, for higher visual quality. Due to a large number of capturing devices, the set-up cost is high. Accordingly, in an effort to reduce number of capturing devices and set-up cost, current academic trend of FVV research is model-based 3D reconstruction. In spite of these efforts, generation of photorealistic view synthesis of shape and texture that fits within a reasonable computational complexity is still a major challenge (e.g. challenges for texture mapping, namely occlusion and texture seam, become critical, as the camera array becomes sparser).
- *Establishing of a ground truth*: Due to the nature of arbitrary virtual viewpoint, it is also difficult to define the ground truth for quality measure that correlates well with user viewing experience. In the absence of a well-established quality measure it will be a major challenge to compare the quality performance of one system versus another.
- *Robust FG extraction from general indoor/outdoor BG*: Automatic and robust extraction of high-quality FG objects from general BG scene remains challenging in an uncontrolled indoor/outdoor environment. Multimodal sensing approach (e.g. silhouette, texture, motion, and depth) is needed with less reliance on BG priors (e.g. natural BG versus green screen).
- *Easy and intuitive mean of view point navigation*: User experience will suffer, if viewpoint navigation is not easy. TV and head-mount devices with eye-tracker have been

https://doi.org/10.1017/ATSIP.2015.18 Published online by Cambridge University Press

proposed for the direct UI. This is nevertheless not at practical level to create immersive user experience.

ACKNOWLEDGEMENT

This work is supported by Naofumi Yanagihara, Visual Technology Development Department of Sony, Corp. We also would like to thank Ken Tamayama who provided technical insight and expertise that have greatly assisted in the drafting of the paper.

REFERENCES

[1] Lafruit, G.; Wegner, K.; Tanimoto, M.: Draft Call for Evidence on FTV, in ISO/IEC JTC1/SC29/WG11 MPEG2015/N15095, Geneva, February 2015.

[2] Tanimoto, M.: FTV (Free-viewpoint television) for ray and sound reproducing in 3D space, in IEEE ICASSP, March 2012, 5441–5444.

[3] Tanimoto, M.; Panahpour Tehrani, M.; Fujii, T.; Yendo, T.: FTV for 3-D spatial communication. Proc. IEEE, 100 (4) (2012), 905–917.

[4] Tanimoto, M.: FTV (free-viewpoint television). APSIPA Trans. Signal Inf. Process., 1 (1) (2013).

[5] Tanimoto, M.; Tehrani, M.P.; Fujii, T.; Yendo, T.: Free-viewpoint TV. IEEE Signal Process. Mag., 28 (2011), 67–76.

[6] ITU-T and ISO/IEC JTC 1: Advanced video coding for generic audiovisual services. ITU-T Recommendation H.264 and ISO/IEC 14496–10 (MPEG-4 AVC), Annex. H.

[7] ITU-T and ISO/IEC JTC 1. High efficiency video coding. – ITU-T Recommendation H.265, and ISO/IEC 23008–2:2013, Annexes F & G.

[8] Smolic, A.: 3D video and free viewpoint video – from capture to display. Pattern Recognit., 44 (2011), 1958–1968.

[9] Wilburn, B. *et al.*: High performance imaging using large camera arrays, in ACM SIGGRAPH, 2005.

[10] Meyer, F. *et al.*: Wireless GPS-based phase-locked synchronization system for outdoor environment. J. Biomech., 45 (2012), 188–190.

[11] Lichtenauer, J. *et al.*: Cost-effective solution to synchronized autio-visual data capture using multiple sensors. Image Vis. Comput., 29 (2011), 666–680.

[12] Tomasi, C.; Kanade, T.: Shape and motion from image streams under orthography: a fractorization method. Int. J. Comp. Vis., (1992), 137–154.

[13] Snavely, N.; Seitz, S.M.; Szeliski, R.: Photo tourism: exploring photo collections in 3D. J. ACM Trans. Graph., 25 (3) (2006), 835–846.

[14] Furukawa, Y.; Ponse, J.: Accurate camera calibration from multi-view stereo and bundle adjustment. IEEE Comput. Vis. Pattern Recognit. (2008), 1–8.

[15] Zitnick, C.L. *et al.*: High-quality video view interpolation using a layered representation. J. ACM Trans. Graph., 23 (3) (2004), 600–608.

[16] Zollhofer, M. *et al.*: Real-time non-rigid reconstruction using and RGB-D camera. J. ACM Trans. Graph., 33 (4) (2014). doi: 10.1145/2601097.2601165.

[17] Shan, Q.; Curless, B.; Furukawa, Y.; Hernandez, C.; Seitz, S.M.: Occluding contours for multi-view stereo. IEEE Comput. Vis. Pattern Recognit. (2014), 4002–4009.

[18] Matusik, W.; Buehler, C.; Raskar, R.: Image-based visual hulls, in ACM SIGGRAPH, 2000, 369–374.

[19] Laurentini, A.: The visual hull concept from silhouette-based image understanding. IEEE Trans. Pattern Anal. Mach. Intell., 16 (1994), 150–162.

[20] Franco, J.S.; Boyer, E.: Efficient polyhedral modeling from silhouette. IEEE Trans. Pattern Anal. Mach. Intell., 31 (2009), 414–427.

[21] Furukawa, Y.; Ponce, J.: Carved visual hulls for image-based modeling. Int. J. Comput. Vis., 81 (1) (2008), 53–67.

[22] Sobral, A.: BGSLibrary: an opencv C++ background subtraction library, in WVC'2013, Rio de Janeiro, Brazil, June 2013.

[23] Landabaso, J.L.; Pardas, M.; Casas, J.R.: Shape from inconsistent silhouette. Comput. Vis. Image Underst., 112 (2008), 210–224.

[24] Haro, G.: Shape from silhouette consensus. Pattern Recognit., 45 (2012), 3231–3244.

[25] Seitz, S.M.; Dyer, C.R.: Photorealistic scene reconstruction by voxel coloring. IEEE Comput. Vis. Pattern Recognit., (1997), 1067–1073.

[26] Kutulakos, K.N.; Seitz, S.M.: A theory of shape by space carving. Int. J. Comput. Vis., 38 (3) (2000), 199–218.

[27] Yezzi, A. *et al.*: A surface evolution approach to probabilistic space carving, in IEEE 3DPVT, 2002, 618–621.

[28] Lowe, D.G.: Object recognition from local scale-invariant features, in Proc. Int. Conf. on Computer Vision 2. pp. 1150–1157, 1999.

[29] Bay, H.; Tuytelaars, T.; Gool, L.V.: SURF: speeded up robust features, in Proc. Ninth European Conf. on Computer Vision, 2006.

[30] Vogiatzis, G. *et al.*: Multi-view stereo via volumetric graph-cuts. IEEE Comput. Vis. Pattern Recognit., 2 (2005), 391–398.

[31] Esteban, C.H.; Schmitt, F.: Sihouette and stereo fusion for 3D object modeling. Comput. Vis. Image Underst., 96 (3) (2004), 367–392.

[32] Torresani, L.; Hertzmann, A.; Bregler, C.: Nonrigid structure-from-motion: estimating shape and motion with herarchical priors. IEEE Trans. Pattern Anal. Mach. Intell., 30 (2008), 878–892.

[33] Dovgard, R.; Basri, R.: Statistical symmetric shape from shading for 3D structure recovery of faces, in ECCV, 2004.

[34] Tankus, A.; Sochen, N.; Yeshurun, Y.: Perspective shape-from-shading by fast marching. IEEE Comput. Vis. Pattern Recognit., 1 (2004), 43–49.

[35] Barron, J.; Malik, J.: Shape, illumination, and reflectance from shading. IEEE Trans. Pattern Anal. Mach. Intell., 37 (8) (2015), 1670–1687.

[36] Barron, J.; Malik, J.: Intrinsic scene properties from a single RGB-D image. IEEE Comput. Vis. Pattern Recognit., (2013), 17–24.

[37] Anguelov, D.; Srinivasan, P.; Koller, D.; Thrun, S.; Rodgers, J.; Davis, J.: SCAPE: shape completion and animation of people, in ACM SIGGRAPH, 2005, 408–416.

[38] Hasler, N.; Stoll, C.; Sunkel, M.; Rosenhahn, B.; Seidel, H.P.: A statistical model of human pose and body shape. EUROGRAPHICS, 28 (2009), 337–346.

[39] Chen, Y.; Liu, Z.; Zhang, Z.: Tensor-based human body modeling. IEEE Comput. Vis. Pattern Recognit., (2013), 105–112.

[40] Balan, A.; Sigal, L.; Black, M.J.; Davis, J.; Haussecker, H.W.: Detailed human shape and pose from images. IEEE Comput. Vis. Pattern Recognit., (2007), 1–8.

[41] Vlasic, D.; Baran, I.; Matusik, W.; Popovic, J.: Articulated mesh animation from multi-view silhouettes. ACM Trans. Graph., 27 (2008). http://dl.acm.org/citation.cfm?doid=1399504.1360696

[42] Gall, J.; Stoll, C.; Aguiar, E.D.; Theobalt, C.; Rosenhahn, B.; Seidel, H.P.: Motion capture using joint skeleton tracking and surface estimation. IEEE Comput. Vis. Pattern Recognit., (2009), 1746–1753.

https://doi.org/10.1017/ATSIP.2015.18 Published online by Cambridge University Press

[43] Guan, P.; Weiss, A.; Balan, A.O.; Black, M.J.: Estimating human shape and pose from a single image, in ICCV, 2009, 1381–1388.

[44] Straka, M.; Hauswiesner, S.; Ruther, M.; Bischof, H.: Rapid skin: estimating the 3D human pose and shape in real-time, in 3D Imaging, Modeling, Processing, Visualization & Transmission (3DIM/3DPVT), 2012.

[45] Aguiar, E.D.; Stoll, C.; Theobalt, C.; Ahmed, N.; Seidel, H.P.; Thrun, S.: Performance capture from sparse multi-view video, in ACM SIGGRAPH, 2008.

[46] Akenine-Moller, T.; Haines, E.; Hoffman, N.: Real-Time Rendering. 2008.

[47] Casas, D. *et al.*: 4D video textures for interactive character appearance. J. Comput. Graph. Forum, 33 (2) (2014).

[48] Takai, T.; Hilton, A.; Matsuyama, T.: Harmonised texture mapping, in 3DPVT, 2010.

[49] Nobuhara, S.; Ning, W.; Matsuyama, T.: A real-time view-dependent shape optimization for high quality free-viewpoint rendering of 3D video. 3DV, 2014.

[50] Eisemann, M.: Floating textures. EUROGRAPHICS, 27, (2008), 409–418.

[51] ITU-T and ISO/IEC JTC 1: Generic coding of moving pictures and associated audio information. – ITU-T Recommendation H.262, and ISO/IEC 13818–2, 1994.

[52] ISO/IEC JTC1/SC29/WG11 N14425: Test Model 8 of 3D-HEVC and MV-HEVC. 2014.

[53] ISO/IEC 14496–16: Information technology – Coding of audio-visual objects – Part 16: Animation Framework eXtension (AFX), 2009.

[54] ISO/IEC 14496–16: Information technology – Coding of audio-visual objects – Part 25: 3D Graphics Compression Model, 2009.

[55] Yendo, T.; Fujii, T.; Tanimoto, M.; Tehrani, M.P.: The seelinder: cylindrical 3D display viewable from 360 degrees. J. Vis. Commun. Image Represent., 21 (5–6) (2010), 586–594.

[56] Kanade, T.; Narayanan, P.J.: Virtualized reality: perspectives on 4D digitization of dynamic events. IEEE Comput. Graph. Appl., May/June (2007), 32–40.

[57] Chan, S.C.; Shum, H.Y.; Ng, K.T.: Image-based rendering and synthesis. IEEE Signal Process. Mag., 24 (6) (2007) 22–33.

[58] Tanimoto, M.: FTV standardization in MPEG, in IEEE 3DTV_Conf., 2014, 1–4.

[59] GigE Vision Video Streaming and Device Control Over Ethernet Standard Version 2.0, AIA, 2011.

[60] Hisatomi, K.: Method of 3D reconstruction using graph cuts, and its application to preserving intangible cultural heritage, in ICCV, 2009, 923–930.

[61] Furukawa, Y.: Accurate, dense, and robust multi-view stereopsis. IEEE Trans. Pattern Anal. Mach. Intell., 32 (8) (2010), 1362–1376.

Chuen-Chien Lee is senior Vice President at Sony Electronics in charge of US Research Center based in San Jose, California. He is responsible for overseeing research in basic science as well as research and development of emerging technologies for the next-generation products and new businesses. He has accordingly led US Research Center in a wide range of research and development activities, including: advanced camera signal processing, computational intelligence, video codec standardization, 3D visualization and interactivity, medical imaging, media streaming & indexing, wireless technology and standardization, smart energy, and green nanotechnologies. These technologies have been deployed in many Sony consumer products including: digital cameras and camcorders, Bravia TVs, mobile phones, and image sensors. Dr. Lee received his Bachelor degree in Electrical Engineering from National Chiao-Tung University in Taiwan and he holds a Ph.D. & Master of Science in Electrical Engineering and Computer Sciences from the University of California, Berkeley.

Ali Tabatabai is a Director and IEEE Fellow at Sony US Research Center, San Jose, California. He started his professional career with Bell Laboratories and later Bell Communication Research where he worked on algorithmic research and on the application of sub-band techniques for still image coding for which he was the co-winner of the IEEE CSVT Best Paper Award. He later joined Tektronix as the manager of digital video research group where his responsibilities included algorithmic research and development of video compression techniques for studio and production quality applications. He chaired AdHoc group in MPEG-2 whose work resulted in the standardization of a highly successful MPEG-2 4:2:2 Profile & Main Level. In his current work at Sony, he is responsible for managing R&D activities in 3D visualization and next generation video codec. He obtained his bachelor degree from Tohoku University and Ph.D. degree from Purdue University both in Electrical Engineering.

Kenji Tashiro is a manager of 3D visualization group at Sony US Research Center, San Jose, California. He started his professional career at PULNiX (later JAI) as a research & development engineer for industrial and traffic capture & vision processing systems. He later joined Teledyne Scientific as a senior research scientist to research neuromorphic vision processing algorithms and hardware architecture for defense applications. In his current work at Sony, he is responsible for managing the 3D visualization group. His main research interests are 3D modeling, texture synthesis, and video processing & codec. He received his bachelor's and master's degrees in Precision Machinery from the University of Tokyo, Japan.

https://doi.org/10.1017/ATSIP.2015.18 Published online by Cambridge University Press