Technical Disclosure Commons

Defensive Publications Series

December 20, 2018

Lightfield Compression Using Commodity Hardware Video Codecs

Sai Deng

Pavel Krajcevski

Matthew Pharr

Stuart Abercrombie

Lightfield Compression Using Commodity Hardware Video Codecs

Abstract:

A lightfield video is a series of lightfield stills, which include a number of reference views and residual data describing differences between the reference views and various viewpoints. The set of reference views or images, though different enough to provide the required basis for predicting intermediate views in the dataset, are correlated enough to approximate a traditional video stream. Additionally, reference images from a single time point may be compressed as a collage of smaller images that make up a single frame of a traditional video. Temporal correlation between reference images at successive time points is exploited, similar to standard two-dimensional video, and reference images for a given frame are tied together by virtue of being part of a same two-dimensional video frame. The sequence of reference images can be treated as frames in a video, allowing the use of non-proprietary codecs on commodity hardware at times of decoding. Thus, at runtime, a video display system relying on the lightfield system can access the original data in a time and power efficient manner using common hardware and compression techniques.

Keywords:

Lightfield, compression, codec, hardware, GPU, graphics processing unit, CPU, central processing unit, video, MP4, MPEG-4, ASTC, HVEC, VP9, virtual reality, augmented reality, VR, AR, artificial environment.

Background:

Virtual reality (VR) environments rely on display, tracking, and VR-content systems. Through these systems, realistic images, sounds, and sometimes other sensations simulate a user's physical presence in an artificial environment. Each of these three systems are illustrated below in Figure 1.

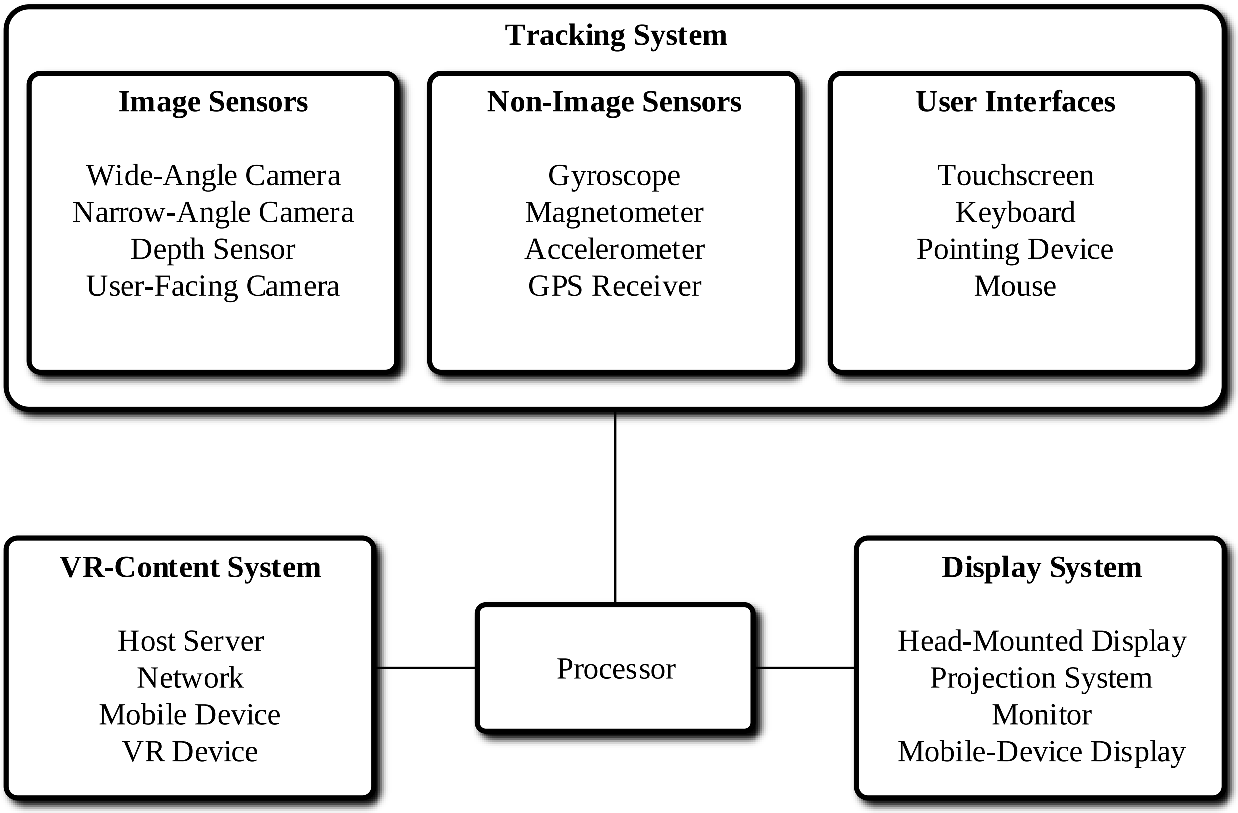

Figure 1

The systems described in Figure 1 may be implemented in one or more of various computing devices that can support VR applications, such as servers, desktop computers, VR goggles, computing spectacles, laptops, or mobile devices. These devices include a processor that can manage, control, and coordinate operations of the display, tracking, and VR-content systems. The devices also include memory and interfaces. These interfaces connect the memory with the systems using various buses and other connection methods as appropriate.

The display system enables a user to “look around” within the virtual world. The display system can include a head-mounted display, a projection system within a virtual-reality room, a monitor, or a mobile device’s display, either held by a user or placed in a head-mounted device.

The VR-content system provides content that defines the VR environment, such as images and sounds. The VR-content system provides the content using a host server, a network-based device, a mobile device, or a dedicated virtual reality device, to name a few.

The tracking system enables the user to interact with and navigate through the VR environment, using sensors and user interfaces. The sensors may include image sensors such as a wide-angle camera, a narrow-angle camera, a user-facing camera, and a depth sensor. Non-image sensors may also be used, including gyroscopes, magnetometers, accelerometers, GPS sensors, retina/pupil detectors, pressure sensors, biometric sensors, temperature sensors, humidity sensors, optical or radio-frequency sensors that track the user’s location or movement (*e.g.*, user’s fingers, arms, or body), and ambient light sensors. The sensors can be used to create and maintain virtual environments, integrate “real world” features into the virtual environment, properly orient virtual objects (including those that represent real objects, such as a mouse or pointing device) in the virtual environment, and account for the user’s body position and motion.

The user interfaces may be integrated with or connected to the computing device and enable the user to interact with the VR environment. The user interfaces may include a touchscreen, a keyboard, a pointing device, a mouse or trackball device, a joystick or other game controller, a camera, a microphone, or an audio device with user controls. The user interfaces allow a user to interact with the virtual environment by performing an action, which causes a corresponding action in the VR environment (*e.g.*, raising an arm, walking, or speaking).

The tracking system may also include output devices that provide visual, audio, or tactile feedback to the user (*e.g.*, vibration motors or coils, piezoelectric devices, electrostatic devices, LEDs, strobes, and speakers). For example, output devices may provide feedback in the form of blinking and/or flashing lights or strobes, audible alarms or other sounds, songs or other audio files, increased or decreased resistance of a control on a user interface device, or vibration of a physical component, such as a head-mounted display, a pointing device, or another user interface device.

Figure 1 illustrates the display, tracking, and VR-content systems as disparate entities in part to show the communications between them, though they may be integrated, *e.g.*, a smartphone mounted in VR goggles, or operate separately in communication with other systems. These communications can be internal, wireless, or wired. Through these illustrated systems, a user can be immersed in a VR environment. While these illustrated systems are described in the VR context, they can be used, in whole or in part, to augment the physical world. This augmentation, called "augmented reality" or AR, includes audio, video, or images that overlay or are presented in combination with the real world or images of the real world. Examples include visual or audio overlays to computing spectacles (*e.g.,* some real world-VR world video games or information overlays to a real-time image on a mobile device) or an automobile's windshield (*e.g.,* a heads-up display) to name just a few possibilities.

Description:

A lightfield video is a series of lightfield stills, which include a number of reference views and residual data describing differences between the reference views and various viewpoints. The set of reference views or images, though different enough to provide the required basis for predicting intermediate views in the dataset, are correlated enough to approximate a traditional video stream. Additionally, reference images from a single time point may be compressed as a collage of smaller images that make up a single frame of a traditional video. Temporal correlation between reference images at successive time points is exploited, similar to standard two-dimensional video, and reference images for a given frame are tied together by virtue of being part of a same two-dimensional video frame. The sequence of reference images can be treated as frames in a video, allowing the use of non-proprietary codecs on commodity hardware at times of decoding. Thus, at runtime, a video display system relying on the lightfield system can access the original data in a time and power efficient manner using common hardware and compression techniques.

The lightfield stills represent a correlated collection of reference views captured by several cameras at a single instance in time, which may or may not be the same wide-angle, narrow-angle, or depth cameras used in the viewing of virtual reality scenes discussed above. In some circumstances, dedicated cameras separate from a virtual reality viewing device may capture a scene prior to viewing by a user. In some augmented reality circumstances, the scene may be captured at a similar time as viewing by the user. Although the cameras capture several different views of a particular scene, uncaptured intermediate views between captured views remain. The lightfield system can construct these new intermediate views based on the captured views, such that a previously uncaptured view becomes available to the system. Any number of views can be combined into a single reference collage image, and the reference collage images can be combined and ordered in time to create a lightfield video.

In a virtual reality context, a lightfield video can be assembled based on individual decisions of a user or user-interface inputs to represent portions of a virtual environment not initially recorded or generated as the user looks around the virtual environment using goggles, computing spectacles, or a mobile device. Thus, a lightfield system capable of constructing new, intermediate views can produce a more-detailed user experience than initially captured by the various cameras.

Although the lightfield stills and videos can represent or generate nearly limitless uncaptured views as captured views, the generation of intermediate views not part of the originally captured views comes at non-trivial hardware and processing cost. The datasets can quickly become difficult to manage, often requiring tens of thousands of images to be stored at once, which may be impractical in many mobile devices, virtual/augmented reality goggles, or computing spectacles. One method of reducing the number of images to manage involves storing a limited number of reference images and differences between the reference images and the original dataset. Hopefully, the reference images have been captured by the cameras in a manner that contains little redundant content that minimizes overall dataset storage and management.

At runtime, the generated or reconstructed intermediate views represent a combination of data from many densely located views. In the context of virtual reality, the video selected and seen by a user can change at any given time. Thus, the lightfield system must be able to access all the data necessary to construct any portion of the lightfield at any given time. In order to do so, some techniques involve using the references views or reference images as a basis for compressing a series of views into a video for user viewing and consumption. However, these techniques often require the use of customized hardware or video compression formats instead of relying on non-proprietary video formats, such as VP9, MPEG-4, or HVEC, and existing decoding hardware on common computing platforms and devices. Indeed, many traditional techniques focused more on compression ratio than on decode speed, because lightfields have significantly more redundancy than videos, and could not take advantage of other hardware and decoding codecs or techniques on common platforms and devices.

A commodity or non-proprietary video codec is designed to record or present a view of a scene from a single point of view over time. A lightfield video records or presents multiple points of view of a scene over time. The system described herein can select frames that fit several points of view and assemble them into one or more traditional video streams processed by a single piece of commodity hardware. For example, consider the camera setup of Figure 2. Here, a single camera records a dinner scene of a couple in a restaurant. The dinner scene is recorded with a single, commodity camera using a non-proprietary video codec.

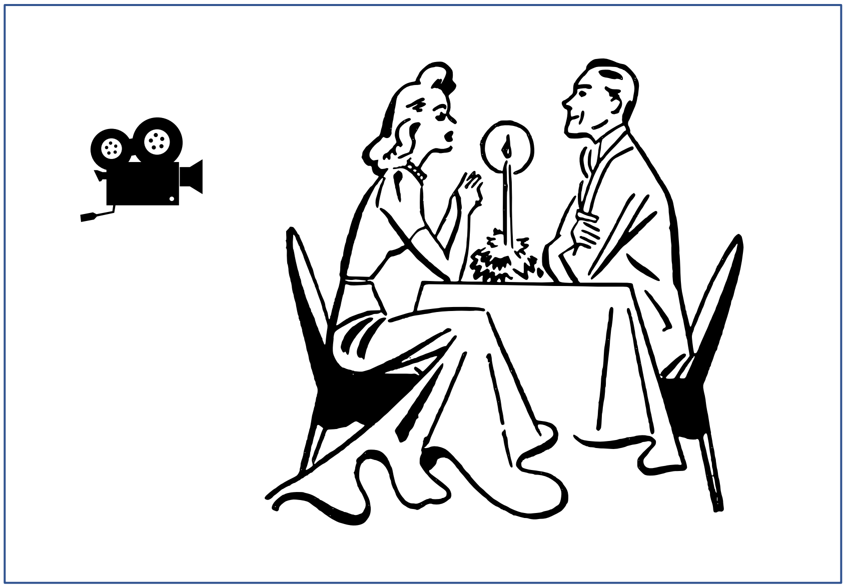

Figure 2

In a virtual reality context, a user may want to pan around the dinner scene and look at the face of the woman, the face of the man, the food on the table, the surrounding environment, or any number of different facets of the scene. By using a lightfield system, the scene would be recorded from many different viewpoints. Consider the dinner scene, as illustrated in Figure 3, recorded from any number of different viewpoints. Six different cameras are illustrated. In some situations, array cameras, array camera fields, or any other camera arrangement could be used. In some circumstances, the system could record hundreds or even thousands of viewpoints in order to assemble a view of the scene in a traditional lightfield system.

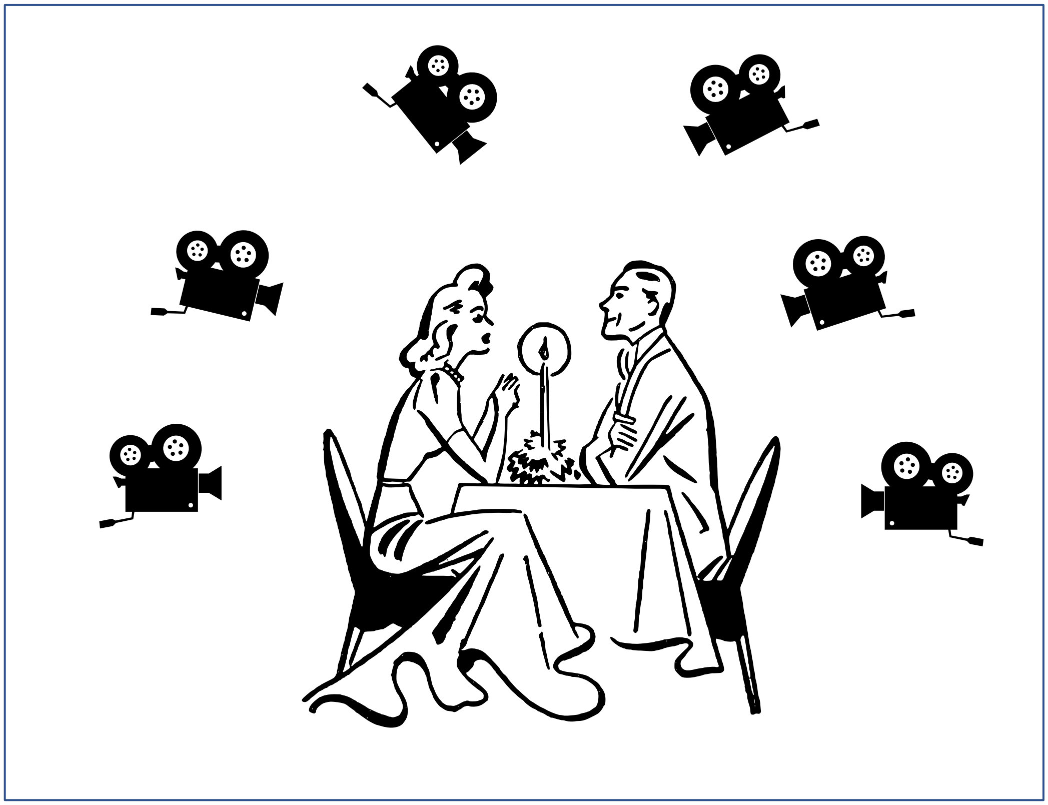

Figure 3

The set of reference views or images collected in a lightfield system, though different enough to provide the required basis for predicting intermediate views in the dataset, are correlated enough to approximate a traditional video stream. However, the sheer number of potential views would overwhelm a traditional video codec intended to operate on a commodity device. In order to use a traditional video codec, the system selects a subset of views to present using a standard video codec. The subset of views could be selected in many different ways. For example, the system could select a view from every third camera to reduce the overall number of viewpoints managed by the system. The system can strike a balance between enough viewpoints for a smooth virtual-reality presentation and data storage or computing limitations of a particular device.

The spatial correlation of the reference views allows the lightfield system to infer a temporal correlation and to order the reference views in a manner in which they remain consistent relative to one another. The sequence of reference images can be treated as frames in a video, allowing the use of non-proprietary codecs on commodity hardware at times of decoding. Thus, at runtime, a video display system incorporating or using a lightfield-based system can access the original data in a time and power efficient manner using common hardware and compression techniques. In fact, the lightfield system described herein can achieve fast and low-power access to views in the dataset during the construction of intermediate views, which enables the use of lightfield representations in low-latency displays like the head-mounted display in the virtual or augmented reality contexts discussed above.

Because the reference views do not change in time, the reference view can be encoded as a sequence of video frames in a spatial dimension. The data can be delivered to an encoder prior to any rendering such that all the reference views are decoded and available as a lightfield video plays back. The hardware decoder can be separate from other processing components, which can allow an application to decode the additional lightfield data needed in parallel to other processing operations. The lightfield system can select a subset of views in both space and time to encode as a non-proprietary video such that at runtime, the subset of views provide the necessary information to reconstruct other views, especially in virtual reality and augmented reality applications.

In some circumstances, such as times when a video is to be played on or viewed on a device without a hardware video decoder, the temporal coherence of the reference views can be exploited by using three-dimensional textures. Textures represent data used by a graphics processing unit correlating to certain properties related to memory access. Although common to store textures as two-dimensional images in memory, textures can be stored as volumetric data by storing a sequence of two-dimensional slices. Similar to some video codecs, certain texture compression formats, like ASTC, can encode a small sequence of images as a volumetric dataset. In like fashion, the lightfield system could use a sequence of three-dimensional textures to represent a video instead of using a full video codec, which requires a hardware video decoder at runtime, to compress the reference views. A three-dimensional texture could represent a small time segment containing a sequence of reference views, which other hardware could decode in a fast, low-power manner. In some virtual or augmented reality presentation devices, the ability to present the artificial environment without specialized hardware can be advantageous.

As described above, a lightfield assisted presentation system can take advantage of the viewpoint flexibility of traditional lightfield photography and still present captured data to a user in a time and power efficient manner using common hardware and compression techniques.