JPEG Pleno: Toward an Efficient Representation of Visual Reality

Ebrahimi, Touradj; Foessel, Siegfried; Pereira, Fernando; Schelkens, Peter

Published in:
IEEE Multimedia

DOI:
10.1109/MMUL.2016.64

Publication date:
2016

JPEG Pleno: Toward an Efficient Representation of Visual Reality

Touradj Ebrahimi
Ecole Polytechnique Fédérale de Lausanne

Siegfried Foessel
Fraunhofer Institute IIS

Fernando Pereira
Instituto Superior Técnico—Instituto de Telecomunicações

Peter Schelkens
Vrije Universiteit Brussel—iMinds

Recently, the JPEG standardization committee created an initiative called *JPEG Pleno*. "Pleno" is a reference to "plenoptic," a mathematical representation that not only provides information about any point within a scene but also about how it changes when observed from different positions. "Pleno" is also the Latin word for "complete," a reference to the JPEG committee's desire for future imaging to provide a more complete description of scenes, well beyond what's possible today. Here, we discuss the rationale behind the vision for the JPEG Pleno initiative and describe how it can potentially reinvent the future of imaging.

Plenoptic Representation

Plenoptic representation, which has been widely studied,[1] provides a veritable holographic characterization of the visual world. The following should give you an idea of what a representation describes without going into cumbersome mathematical details.

Imagine you're observing a scene with a pinhole camera (see Figure 1). The camera captures the set of light rays passing through the pinhole, known as the pencil P, at that point in space and time. The camera registers the intensity distribution of the light parameterized by spherical coordinates $P(\theta, \Phi)$. If we consider a color camera, where the color is characterized by a wavelength λ, we can extend the pencil definition to $P(\theta, \Phi, \lambda, t)$, with t representing time in a dynamic scene. The full plenoptic characterization of the visual world is obtained by positioning the pinhole camera at every viewpoint in spatial coordinates x, y, z, resulting in a seven-dimensional representation, depicted as $P(x, y, z, \theta, \Phi, \lambda, t)$, which is in fact the plenoptic representation of the scene in question.

Current representation models of visual information based on a 2D and 3D scan of a scene address only a subset of the information contained in the plenoptic representation just described. Such content is further discretized, windowed, subsampled, and quantized along different dimensions—including the number of pinhole camera positions, resolution of the intensity distribution, number of components representing the color information (such as red, green, blue), and number of frames captured per second to cope with the scene's dynamics. Consequently, the camera you used to take pictures at your last holiday gathering merely produces one approximation (among many) of the plenoptic representation of the places where you took photos.

Recently, a plethora of new sensor and camera systems have attempted to capture a more complete approximation of the plenoptic representation. These systems generate richer forms of visual content, including the following:

- *Omnidirectional* content is generated by a single camera or multiple cameras, enabling a wider field-of-view and larger viewing angles of the surroundings. Omnidirectional content is often captured as 360° panorama or complete sphere and mapped to a 2D image.

- *Depth-enhanced* content is produced by depth sensing cameras (such as the time-of-flight camera) or structured light systems, or by extracting depth from multiple images.
- *Point cloud* content is typically captured with a 3D scanner or Light Detection And Ranging (Lidar) scanner and can be subsequently used to generate 3D surfaces through operations such as mesh generation. Lidar scanners can measure a high-resolution structure and create a depth profile of the visual world, but they often must be combined with additional modalities—originating from classical cameras, for example—to characterize a richer plenoptic representation of the real world. Point clouds can also be computed from an array of cameras with different viewpoints.
- A *light field* is produced by capturing a scene either via an array of cameras or by a single and more compact sensor augmented by microlenses, to sample individual rays of light emanating from different directions. Light-field cameras based on microlenses are more compact but offer limited support for the viewing angle—that is, they have a narrower baseline for light-field data. In contrast, light-field cameras based on discrete optoelectronic sensors in linear or matrix configurations can support larger viewing angles—that is, they offer a wider baseline for light-field data—but they're bulkier and more expensive.
- *Holography* uses interferometric capturing or computer-generated holography (CGH) techniques to describe a 3D scene. Holography portrays the full plenoptic function—excluding the time dimension— to a 2D map. Although the technology still lacks maturity in macroscale applications, holography has already reached a sufficient level of maturity in microscale applications, such as in holographic microscopy.

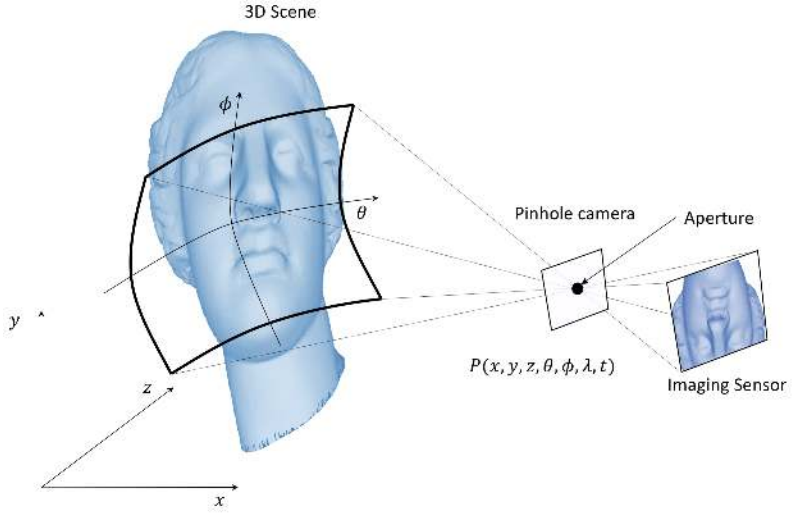

Figure 1. Plenoptic capture of a pinhole camera scanning a scene. The rays of light passing through the pinhole correspond to values of the plenoptic function along different directions.

A better understanding of plenoptic representation and its underlying practical implications as well as its challenges involves a better understanding of the human visual system and an efficient modeling of the end-to-end plenoptic processing flow.

Plenoptic Processing Flow

Although an infinite number of approaches can be used to capture, process, deliver, store, consume, and interact with plenoptic content, such systems can all be viewed as a chain composed of distinct components that interact to define a plenoptic processing flow. To better understand the end-to-end plenoptic dataflow, Figure 2 shows the main components in the process, starting with content acquisition/creation and ultimately leading to the display and user interaction.

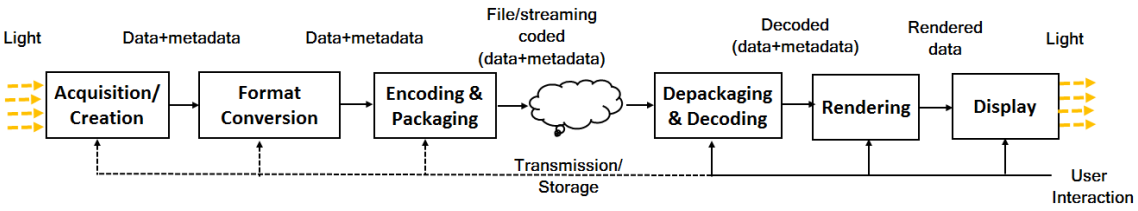

Figure 2. End-to-end plenoptic processing flow. The processing flow represents an architecture of key components in typical plenoptic content coding and decoding.

Acquisition/Creation

The first component in the plenoptic processing flow deals with the light-data acquisition by exploiting appropriate sensors. A wide range of sensors can be used, from 2D regular camera arrays and devices bearing microlenses to depth sensors. Each sensor relies on a specific data-acquisition model, which can use a regular sampling grid—for example, a regular array of cameras—or an irregular sampling grid (such as a point cloud). Although sensors capture natural and real-world light, plenoptic content can also be computer generated, meaning not acquired from a physical scene—at least not directly. A computerized creation might generate data using the same models as in natural acquisition, and data with different origins might be mixed in a single content in a seamless way.

Both the acquisition and creation processes can add relevant metadata to the visual data in some relevant format. This metadata might have very different purposes—it could describe the sensor, lenses, camera arrangements or support a semantic search, or it could be used to support privacy protection privacy, ownership rights, or security control.

Examples of sensors include conventional 2D cameras, regular and arbitrary arrays of cameras, light-field cameras with microlenses, and depth sensors.

Format Conversion

The acquisition/creation data model is largely determined by sensors and acquisition constraints. Consequently, it might not be the most appropriate way to represent the data when considering the overall application requirements—such as compression, display, or user-interaction requirements. A format conversion might be needed to convert the acquired (or even synthetically created) plenoptic data into a more suitable model.

For interoperability reasons, the number of representation models should be as few as possible, ideally corresponding to a single model. However, considering the variety of application scenarios and associated requirements, it's more reasonable to assume a rather limited number of models and expecting alternative models to be efficiently converted as needed. This conversion can happen both for the data and metadata, although the data conversion should involve higher complexity and have a greater impact on the final user experience, depending on the conversion assumptions and constraints. An example of a relevant conversion might be transforming a set of texture and depth views into a set of point clouds with RGB values or even meshes with associated textures.

Encoding and Packaging

The encoder targets a more efficient representation of the data in terms of compactness while fulfilling an additional set of relevant functionalities, such as random access, scalability, and error resilience. The encoding process can be either lossless or lossy, although it's most commonly lossy when considering human perception characteristics. While ideally a single coding format would maximize interoperability, the coding format will depend on the number of selected representation models. Different parts of a scene, such as background and foreground, might adopt different representation models, thus requiring different coding formats to coexist in the same scene.

Again, the encoding process applies both to data and metadata, eventually considering slightly different requirements. After encoding, the coded data and metadata are packaged in an appropriate file format able to provide the identified system-level functionalities.

Depackaging and Decoding

After transmission or storage, the packaged data is depackaged for decoding, thus recovering the data and metadata in the adopted representation model. If a lossy encoder is used, the decoded data will have a certain amount of distortions when compared to the corresponding original data, which will depend on the codec itself, the data characteristics, and the coding rate used.

Rendering

The rending must extract appropriate data from the decoded data to provide the best experience possible with the available display. In most conventional image and video applications, decoded content is directly given to the display without much rendering—although post-processing to improve quality and aesthetics, such as filtering or contrast enhancement, can be applied. However, for the most part, this won't be the case with plenoptic content, because the rich available scene representation will have to be "mined" to extract the relevant data—for example, to obtain a specific viewpoint or focus plane. Furthermore, the plenoptic content will eventually be under user control through explicit command interactions or indirect interactions, such as via the motion of a head-mounted display. Because rendering can now become rather complex and sophisticated, it will critically determine the quality of experience, eventually even more so than other components in the processing flow, which can introduce distortions before rendering.

Again, the metadata should also be processed, eventually in close synchronism with the data. It should also be possible to enrich the displayed scene with locally captured or stored natural and synthetic content, such as the user's hands and body within the scene presented on a head-mounted display.

Display

The final display will not only determine the user experience but also the technical solutions and requirements for the other components. In addition to the recent evolution of sensors, displays have also been evolving significantly—examples include autostereoscopic, light-field, and head-mounted displays. **//Okay?//** Naturally, if a light-field display is available, the user can benefit from richer plenoptic representation.

However, especially in the early days of applications using plenoptic processing, it will be common to have rich data—light fields and point clouds, for example—that must be displayed on conventional displays, such as 2D regular displays. This process gives the rendering module the critical role of extracting the available rich data as a subset to be displayed, targeting the best user experience. Because it's critical to assess the quality of the experience, both subjective assessment methodologies and new objective metrics will be required.

User Interaction

The availability of richer scene representations lets users experience interactions with plenoptic content similar to interactions that naturally happen in the real world. By interacting with the display, the user can not only perform simple actions, such as contrast enhancement, but also control the point of view, focal plane, lighting conditions, and so on.

This type of interaction is certainly relevant to the rendering component, which must extract the appropriate displayable data from fully or partially decoded data in reaction to the user's request. However, the interaction might go all the way back to the decoding module if only the data relevant for the requested user experience (such as a specific subset of the point cloud) was decoded as a way of preserving decoding resources. The interaction might even back to the sending side—defining the preferred points of view that should be captured or captured with more details. With time, we expect user interactions with plenoptic content to become as natural and powerful as interactions in the real world. **//Okay?//**

Although this processing flow opens new doors to visual content models and processing, it's well known that standard solutions for the depackaging and decoding, rendering, and display components (the right side of Figure 2)—and especially for the decoding component—play a paramount role in the wide adoption of applications and services. This is where JPEG Pleno comes to play a major role.

JPEG Pleno

JPEG Pleno intends to provide a standard framework to facilitate the capture, representation, and exchange

of omnidirectional, depth-enhanced, point-cloud, light-field, and holographic imaging modalities. It aims to define new tools for improved compression while providing advanced functionalities at the system level. It also aims to support data and metadata manipulation, editing, random access and interaction, protection of privacy, and ownership rights and other security mechanisms.

JPEG Pleno will provide an efficient compression format that will guarantee the highest quality content representation with reasonable resource requirements in terms of data rates, computational complexity, and power consumption. In addition to features described next, supported functionalities will include low latency, backward and forward compatibility with legacy JPEG formats, scalability, random access, error resilience, privacy protection, ownership rights and security, parallel and distributed processing, hierarchical data processing, and data sharing between displays or display elements. The associated file format will be compliant with JPEG systems specifications and will include signaling syntax for associated metadata for the capture, creation, calibration, processing, rendering, and editing of data and for user interactions with such data.

Features and Potential Applications

As mentioned earlier, plenoptic processing opens doors to new approaches for representing both real and synthetized worlds in a seamless and undistinguishable manner. This will enable a richer user experience for creating, manipulating, and interacting with visual content. Such new experiences will, in turn, enable an infinite number of novel applications that are either difficult or impossible to realize with existing image-coding formats.

JPEG Pleno will facilitate these by offering a collection of new features, in addition to those already offered by existing image-coding formats, such as scalability, random access, error resilience, and high compression efficiency. Some of the new features offered by plenoptic representation in JPEG Pleno include the ability to

- change the field depth after capture;
- change the focus and refocus on objects of interest after capture;
- change the lighting in an already captured or synthetized scene;
- change the perspective and observer's position within the scene, including free navigation; and
- enhance the analysis and manipulation of objects within a scene, such as segmentation, modification, and even removal or replacement of specific objects.

All these features enable new applications that were only possible with computer generated content, but now also possible to apply to real world scenes or used in virtual, augmented, and mixed-reality scenarios. Here, we briefly describe a few examples.

Depth-enriched photography. Most digital images today are stored in the well-known legacy JPEG file format. Wouldn't it be nice to be able to select two positions in an image and measure their distance? Then you could determine, for example, whether a new sofa you plan to purchase will fit in your living room. This is just one example application of depth-enriched photography. The JPEG committee intends to add new features to the existing JPEG format, such that your next smartphone can offer experiences such as this.

Enhanced virtual and augmented reality. Today, most visual content in form of images and video from 360-degree capture devices are stitched together based on a fixed geometric mapping. This leads to artifacts due to a missing common nodal point. With depth-based processing based on plenoptic representation, parallax compensated stitching will be possible.

In addition, JPEG Pleno can be used to view content from different viewpoints, as when a physical scene is observed in the real world. This means when viewing 360-degree panoramas, you can move around the scene as if you were physically there. Furthermore, real and virtual objects, such as the user's hands and body, can be included in the scene, and interactions will be possible.

Enhanced post production. All blockbusters today include special effects to make the movie more immersive or to allow actors to move beyond physical limits. For this reason, many parts of the movie are computer-generated. With plenoptic image processing, real scenes can be processed similar to synthetic scenes, accelerating the production process. Potential depth-based effects include depth-based color grading without manual masking of objects, virtual backlots without a green screen, and scene relighting.

Mixed reality, teleconferencing, and telepresence. Wouldn't it be nice to have your relatives in the same room, so you could interact with them even though they might be thousands of miles away? Similar to enhanced post production, plenoptic representation offered by JPEG Pleno will enable new immersive experiences as an extension of conventional teleconferencing and telepresence in real time and with ultralow latency.

Digital holographic microscopy (DHM). Microscopes that produce holograms from which intensity and phase images can be derived allow for example for cell refractive index tomography to facilitate 3D reconstruction of cells. Examples of life science applications also include monitoring of the viability of cell cultures in suspensions, automated multiwell plate screening devices measuring the cell density and the cell coverage of adherent cell cultures, or simultaneous fluorescence and holographic cell imaging.

Framework Vision

According to Aristotle, "The whole is greater than the sum of its parts." This is also the JPEG Pleno underlying vision, as it targets more than a collection of tools to seriously and effectively take into account that there is a unifying root entity: light and its plenoptic representation with various related modalities. This is the conceptual charter for JPEG Pleno, which aims to provide a standard framework for representing new imaging modalities, such as texture-plus-depth, light-field, point-cloud, and holographic imaging. Furthermore, such imaging should be understood as light representations inspired by the plenoptic function (see Figure 3), regardless of which model captured or created all or part of the content.

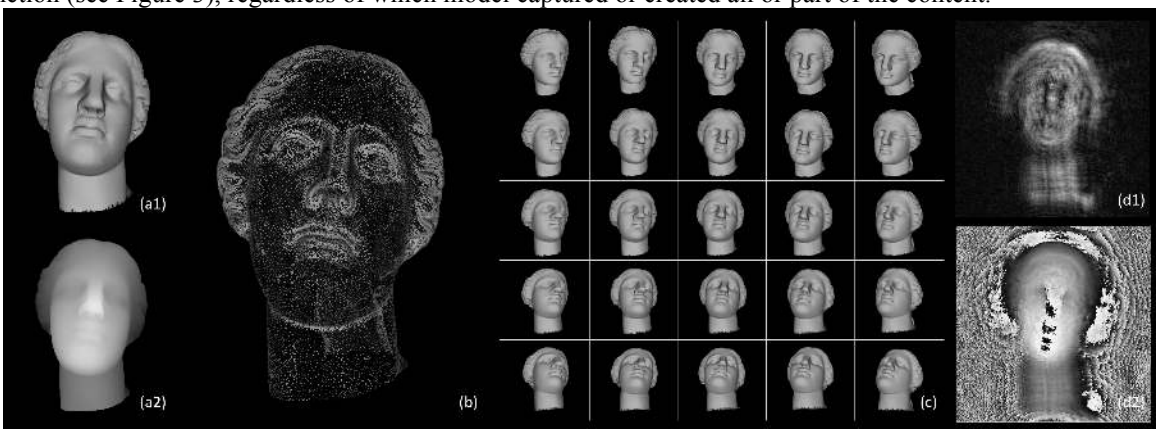

Figure 3. A conceptual illustration of the plenoptic function of a scene represented by different modalities: (a) a texture and (b) depth map, (c) a point cloud, (d) a light field, and finally (e) the amplitude and (f) phase components of a full parallax hologram.

In this context, the required standard tools must be designed together to consider their synergies and dependencies for the whole to be effectively greater than the sum of its parts. To fully exploit this holistic approach, JPEG Pleno isn't just a set of efficient coding tools addressing compression efficiency; it's a representation framework understood as a fully integrated system for providing advanced functionality support for image manipulation, metadata, random access and interaction, and various file formats. In addition, it should offer privacy protection, ownership rights, and security. The JPEG Pleno framework is end-to-end—from the real or synthetized world to the replicated world—in its focus on harmoniously integrating all necessary tools into a single system to represent the same visual reality while considering different modalities, requirements, and functionalities.

This highlights that, for example, light-field, point-cloud, and depth photography aren't fully independent representation models but rather "different sides of the same coin" (even if considering different constraints), which should then often share metadata and other system-level tools.[2] The framework philosophy might also be brought to the point of incorporating JPEG's legacy formats when needed, offering users a continuously evolving ecosystem.

The roadmap for JPEG Pleno follows a path that started in 2015 and will continue beyond 2020, with the objective of having the same type of impact on today's digital imaging applications that the original JPEG format had on earlier applications. **//Okay?//** After an initial period of exploration through organized discussion meetings and workshops, which invited contributions in terms of potential applications, use cases, requirements, and desired features, the JPEG standardization committee has officially launched a new work item in 2016, and a call for proposals is in progress to produce a first working draft for JPEG Pleno specifications in the second part of 2017 and a first international standard in 2018.

References

1. E.H. Adelson, and J.R. Bergen, The Plenoptic Function and the Elements of Early Vision, MIT Press, 1991.
2. *Technical Report of the Joint Ad Hoc Group for Digital Representations of Light/Sound Fields for Immersive Media Applications*, tech. report ISO/IEC JTC1/SC29/WG1N72033(1N16352), Int'l Standards Organization/Int'l Electrotechnical Commission, 2016; https://jpeg.org/items/20160603_pleno_report.html.

Touradj Ebrahimi *is a professor at Ecole Polytechnique Fédérale de Lausanne, where he heads the Multimedia Signal Processing group. He is also the convenor (chairman) of the JPEG Standardization Committee. Contact him at touradj.ebrahimi@epfl.ch.*

Siegfried Foessel *is head of the Moving Picture Technologies department at Fraunhofer Institute IIS in Erlangen, Germany. He is chairing the JPEG systems subgroup. Contact him at siegfried.foessel@iis.fraunhofer.de.*

Fernando Pereira *is an associate professor at Instituto Superior Técnico - Instituto de Telecomunicações (University of Lisboa) and is heading the Multimedia Signal Processing group of the Instituto de Telecomunicações in Lisbon. He is also the Requirements subgroup chair of the JPEG Standardization Committee. Contact him at fp@lx.it.pt.*

Peter Schelkens *is a professor at Vrije Universiteit Brussel and is the VP of Research of Data Science at iMinds. He is chairing the Coding and Analysis and the Public Relations and Liaisons subgroups of the JPEG Standardization Committee. Contact him at peter.schelkens@vub.ac.be.*

In discussing the rationale behind the vision for JPEG Pleno and how the new standardization initiative aims to reinvent the future of imaging, the authors review plenoptic representation and its underlying practical implications and challenges in implementing real-world applications with an enhanced quality of experience.

Keywords: multimedia, light field, point cloud, holography, virtual reality, augmented reality, mixed reality, graphics, visualization, virtualization, plenoptic representation, JPEG Pleno