Plenoptic Imaging Representation

Fernando Pereira
Instituto Superior Técnico - Instituto de Telecomunicações
Lisboa – PORTUGAL

Eduardo A. B. da Silva
Universidade Federal do Rio de Janeiro
Rio de Janeiro - BRAZIL

European Signal Processing Conference
Nice, France
31st August 2015
Note

- This set of slides builds a rather complete and sometimes detailed story …

- Due to its limited duration, some of the details/slides may have to be skipped at the live tutorial
About Me …

🌟 Associate Professor at University of Lisbon, Portugal
🌟 Senior Researcher at Instituto de Telecomunicações, Lisbon, Portugal
🌟 More than 250 publications in international journals and conferences
🌟 Chairman of the MPEG Requirements group for several years
🌟 One of the designers of the MPEG-4 and MPEG-7 standards
🌟 ICIP, PCS, VCIP, WIAMIS, QoMEX General or Technical Program Chair
🌟 Associate Editor of many journals
🌟 Editor-in-Chief of the IEEE Journal of Selected Topics in Signal Processing
🌟 ISO/IEC Award for contributions to the MPEG-4 Visual Standard
🌟 SPS Distinguished Lecturer
🌟 IEEE Fellow in 2008 for “contributions to object-based digital video representation technologies and standards”
🌟 EURASIP Fellow in 2013 for “contributions to digital video representation technologies and standards”
🌟 IEEE SPS Board of Governors and EURASIP Board of Directors
🌟 Several Excellence Teaching Awards
Visual, Visual, Visual …

★ It is believed that up to 50% of the human brain is involved in some way in processing visual information
  ● Reflects the significance of vision for function and survival
  ● Also explains its capacity to entertain, and inform

★ Visual experiences are important drivers:
  ● By 2018, the sum of all forms of video traffic will be in the range of 80-90%
  ● By 2018, over half of all traffic will originate from non-PC devices
  ● By 2020, the number of network-connected devices will reach 1000 times the world's population

★ New, more immersive and effective visual experiences are continuously asked for!

The Periodic Visual Coding Existential Crisis ... Striking Again ...

WHERE AM I GOING?
WHAT AM I DOING?
WHAT IS THE MEANING OF CODING?
Visual Coding: What and Why?

- Replicating the visual world
- Driven/conditioned by available sensors, transmission/storage channels, displays and devices
- ... and by the Human Visual System
- To offer in an efficient, effective, immersive, resilient, scalable, adaptive, simple, ... way
- The relevant set of functionalities
- For each target application/service
- To provide the best USER EXPERIENCE!
More, More, and … Even More Data …

- Higher spatial resolutions
- Higher temporal resolutions
- From interlaced to progressive
- Higher pixel depths
- Higher number of views
- Larger color gammut
- Less color subsampling
- …
- More content variety

While cameras and displays are many times ready for further ‘jumps up’, the transmission infrastructure is typically not prepared to accommodate the associated growing rates!
The ‘End of Times’ Approach ...

★ Higher resolutions (at least above 4K) are useless
  ● New generations just use handheld terminals
  ● Visual system does not see the difference anymore

★ Sofa TV and big TVs have no future as only old people nowadays see TV ...

★ 3D is dead ...

★ ...

Higher resolutions (at least above 4K) are useless

- New generations just use handheld terminals
- Visual system does not see the difference anymore

Sofa TV and big TVs have no future as only old people nowadays see TV ...

3D is dead ...

...
Let’s be Humble … but Still Ambitious

- Visual representation CANNOT be for ever what is already today …
- We have to keep trying opening new frontiers …
  … with the courage to fail and keep trying …

Whatever will be the future, it has to be researched today!
Outline of this Talk

1. Visual Coding: Context and Motivation
2. 2D Video Coding *Status Quo*
3. 3D Basics, Perception and Systems
4. 3D Video Coding *Status Quo*
   1. Overall Landscape
   2. 3D-HEVC Standard
5. 3D Visual Coding Evolution
   1. Plenoptic Function and Related Concepts
   2. Super Multiview Imaging
   3. Light Fields Imaging
6. Standards: Again and More
7. Summary and Trends
A 2D Window to the World …
Video Coding Requirements

★ Maximizing the quality for a target bitrate
★ Minimizing the bitrate for a target quality (or lossless)
★ Maximizing the rate-distortion performance
★ Maximizing the user experience for the available resources
★ Maximizing the error resilience/robustness
★ Minimizing the encoding and decoding complexities
★ Providing random access
★ Providing efficient scalability
★ Providing interactivity
★ …
Compression Efficiency versus Other Requirements
Predictive Coding: a Winning Cocktail ...

No significant architectural changes over the standards!

E Pur Si Muove ...
To Be Efficient, Be Adaptive …
To Be Adaptive, Add Complexity …

from Fabio Sonnati
Quantization exploits the visual perception characteristics and it is essential to reach manageable bitrates, eventually at no perceptual quality penalty.
Video Codec: a Toolbox Approach

- Pre-processing
- Filtering
- Spatial and temporal segmentation
- Motion estimation
- Spatial and temporal prediction
- Spatial transforms
- Entropy coding
- Deblocking filtering
- Error concealment
- Post-filtering
- ...

[Diagram of a toolbox with tools]
A Long Journey, Step by Step …

![Graph showing PSNR vs. Rate for different video coding standards.](image)

- Variable block size (16x16 – 4x4) + quarter-pel + multi-frame motion compensation (H.264/AVC, 2003)
- Variable block size (16x16 – 8x8) (H.263, 1996) + quarter-pel motion compensation (MPEG-4, 1998)
- Frame Difference coding (H.120 1988)

Bit-rate Reduction: 75%

- Conditional Replenishment (H.120)
- Integer-pel motion compensation (H.261, 1991)
- Intraframe DCT coding (JPEG, 1990)

For Foreman 10 Hz, QCIF 100 frames

Jens-Rainer Ohm, Gary Sullivan, Thomas Wiegand: Trends in Video Coding Standardization

ICIP 2012 Tutorial
Video Coding Standards Over Time …

High Efficiency Video Coding (HEVC) Standard: Why?

- Video is continuously increasing in resolution and views
  - HD existing, Ultra HD (4K×2K, 8K×4K) appearing
  - Mobile services going towards HD
  - Stereo, multi-view emerging

- Devices available to record and display Ultra HD resolutions
  - Becoming affordable for home and mobile consumers

- Video has multiple dimensions to grow the data rate
  - Spatial resolution, temporal resolution
  - Color resolution, bit depth
  - Multi-view

- Necessary video data rate grows faster than feasible network transport capacities
  - Better video compression (than current H.264/AVC) needed in next decade
Main HEVC Requirements

- **Compression** - **Substantially greater bitrate reduction over the H.264/AVC High profile** is required for the target application(s); **at no point of the entire bitrate range shall HEVC be worse than existing standard(s).** Subjective visually lossless compression shall be supported.

- **Complexity** - Shall allow for feasible implementation within the constraints of the available technology at the expected time of usage. **HEVC should be capable of trading-off complexity and compression efficiency** by having: i) an operating point with significant decrease in complexity compared to H.264/AVC but with better compression efficiency than H.264/AVC; ii) an operating point with increase complexity and commensurate increase in compression performance.

- **Picture Formats** - Focus on a set of rectangular picture formats that will include all commonly used picture formats, ranging at least from VGA to 4K×2K, and potentially extending to QVGA and **8K×4K**.

- **Color Spaces and Color Sampling** - a) The YCbCr color space 4:2:0, 8 bits per component shall be supported; b) YCbCr/RGB **4:4:4** should be supported; c) Higher bit depth **up to 14 bits** per component should be supported.
Sensors and Displays
Leading the Process

... since MPEG-1

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Aspect Ratio</th>
<th>Frame Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UHDTV</strong></td>
<td>16:9</td>
<td>60p</td>
</tr>
<tr>
<td>7680 x 4320</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HDTV</strong></td>
<td>16:9</td>
<td>24p, 30p, 60i</td>
</tr>
<tr>
<td>1920 x 1080</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1280 x 720</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SDTV</strong></td>
<td>16:9</td>
<td>24p, 30p, 60i, 60p</td>
</tr>
<tr>
<td>704 x 480</td>
<td></td>
<td></td>
</tr>
<tr>
<td>704 x 480</td>
<td></td>
<td></td>
</tr>
<tr>
<td>640 x 480</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HEVC: Still a Quiet Revolution …

2D Video HEVC Extensions

✿ Range extensions

- Larger bit-depth or higher dynamic ranges (HDR), up to 16 bits
- Color sampling beyond 4:2:0
- Screen content coding
- …

✿ Scalability (SHVC)

- HEVC base layer, spatial scalability, 1.5× and 2×
- HEVC base layer, spatial scalability, intra-only
- HEVC base layer, SNR scalability (enhancement QP-2/-4/-6/-8)
- AVC base layer, spatial scalability, 1.5× and 2×
For some given available resources, e.g. in terms of bandwidth and memory, it may be critical to find the right balance between

- Spatial resolution
- Temporal resolution
- Dynamic range
- Colour subsampling
- Colour gammut
- Scalability combination
- ...

to provide the best 2D visual user experience …

But this is expected to be content and display dependent …
3D Basics, Perception and Systems
It’s a 3D World!
Let’s Remind About 3D …

★ The world is not 2D …
★ 3D is more than stereo parallax …
★ Strong interest in 3D applications …
★ 3D applications go beyond TV broadcasting …
★ Increasing production of premium content, e.g. movies and sports …
★ Numerous devices supporting stereoscopic displaying available to the consumer including mobile …
★ Autostereoscopic displays without glasses emerging …
★ Substantial investments to upgrade digital cinema theaters with 3D capabilities …
★ Many new standards being developed, e.g. production, distribution, digital interfaces …
★ The visual future cannot be only 2D …
★ …
## History of 3D Video

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1838</td>
<td>Wheatstone explains &quot;stereopsis&quot;</td>
</tr>
<tr>
<td>1851</td>
<td>First stereo film camera</td>
</tr>
<tr>
<td>1890</td>
<td>Boom year for 3D movies</td>
</tr>
<tr>
<td>1915</td>
<td>First compression standard: MPEG-2 develops Multi-View Profile</td>
</tr>
<tr>
<td>1953</td>
<td>Almost 30 3D movies produced only in 1953!</td>
</tr>
<tr>
<td>1990s</td>
<td>MPEG-4 Multi-View Coding</td>
</tr>
<tr>
<td>1995</td>
<td>3D starts to gain popularity with IMAX 3D</td>
</tr>
<tr>
<td>2005-2009</td>
<td>Expansion of 3D movies</td>
</tr>
<tr>
<td>2009</td>
<td></td>
</tr>
</tbody>
</table>

*Queen Victoria starts stereocope rage*
*First red/blue 3D movies shown*
3D Content is not New …

★ 165 3D movies released since 1953
★ Almost 30 3D movies only in 1953
★ Much more to come …
Critical 3D Success Factors

- High quality experience not burdened with high transition costs or turned off by viewing discomfort or fatigue
- Usability and consumer acceptance of 3D viewing technology, e.g., glasses vs no glasses
- Availability of premium 3D content in the home
- Availability of an appropriate data format providing interoperability through the delivery chain and taking into consideration the constraints imposed by each delivery channel
3D Perception Basics

1 + 1 = 2
The Human Eye

Rod and cone cells in the retina allow conscious light perception and vision including color differentiation and the perception of depth.

The crystalline lens changes/focus for the light to strike the retina
Depth Cues: Monocular and Binocular

- Most of the depth cues used by humans to visualize the world’s 3D structure are available in 2D projections; this is why images make sense on a (mono) TV/cinema screen.

- The depth cues can be classified into oculomotor cues coming from the eye muscles, and visual cues from the scene content itself. They can also be classified into monocular and binocular cues.

- Monocular cues for 3D perception:
Some main cues are missing from 2D media:

- **Stereo parallax** - seeing a different image with each eye, thus different aspects of the same object

- **Motion parallax** - seeing different perspective images when we move our heads; nearby objects appear to move faster across the view

- **Vergence** - muscular rotation of the eye balls, which is used to converge both eyes on the same object
Range of Effectiveness of Depth Cues

<table>
<thead>
<tr>
<th>Depth Information</th>
<th>0–2 Meters</th>
<th>2–20 Meters</th>
<th>Above 30 Meters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occlusion</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Relative size</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Accommodation and convergence</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Movement</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Relative height</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Atmospheric perspective</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

Source: Based on Cutting & Vishton, 1995.

* Not all cues have the same importance in the visual system, and their relative importance depends on the viewing distance, among other factors

* Some depth cues are independent of distance, such as occlusion or relative size, whereas others are distance-dependent, such as disparity or vergence
Stereoscopic Vision

- **Accommodation**, a monocular cue, refers to the variation of the crystalline lens shape and thickness (and thus its focal length), to allow the eye to focus on an object as its distance varies to maintain a clear image or focus.

- **Vergence**, a binocular cue, refers to the muscular rotation of the eye balls, which is used to converge both eyes on the same object.

- Under normal conditions, changing the focus of the eyes to look at an object at a different distance will automatically cause vergence and accommodation, sometimes known as the *accommodation-convergence reflex*.

- In real life, the viewer eyes accommodate (focus) and converge (point) to the depth of the object.
Accommodation-Vergence Conflict

* In natural viewing, the vergence stimulus and focal stimulus are always at the same distance and, therefore, are consistent with one another.

* Stereo displaying create (varying) inconsistencies between vergence and focal distances because the vergence distance varies depending on the image contents while the focal distance remains constant (in the screen).

* The accommodation-vergence conflicts lead to problems, notably 3D structure distortions and visual fatigue.
Due to the accommodation-vergence conflict, there is a limited disparity range allowing proper stereo vision and depth perception. In content production, the admissible disparity range is called *comfort zone*.

3D video production has to map the arbitrary depth range of the real world into this comfort zone by carefully modifying the stereo camera baseline and convergence settings.
3D Systems
Early Stereoscopy

Stereoscopy regards the capability of recreating 3D visual information or creating the illusion of depth in an image based on two appropriate views.

These two slightly different images are presented to each eye. Both of these 2D offset images are then combined in the brain to give the perception of 3D depth.

The motion parallax cue is not satisfied with stereoscopy and, therefore, the illusion of depth is incomplete.
3D Video Experiences …

★ **Depth perception in stereoscopic displays** – Effect provided through *stereo video pairs*, targeting the left and right eyes, allowing the *perception of depth using stereo parallax*

★ **Depth perception in auto-stereoscopic displays** – Effect provided through *n video views*, targeting the left and right eyes in multiple positions, allowing the *perception of depth using stereo and motion parallaxes*

★ **Navigation** – Effect provided through *n video views*, allowing navigating the 3D scene by changing the viewpoint and view direction within certain ranges; the viewer may experience a *look around effect* as well as depth perception
Stereo Cameras …

- A stereo camera is a type of camera with two lenses with a separate image sensor for each lens.

- This allows simulating human binocular vision, and gives the ability to capture 3D images, a process known as *stereo photography*.

- The distance between the lenses, called *baseline*, in a typical stereo camera (the *intra-axial distance*) is about the distance between one's eyes (known as the *intra-ocular distance*)—this is about 6.35 cm, although a longer *baseline* (greater *inter-camera distance*) produces more extreme 3D content.
Current 3D content production methods can be classified into three main categories:

- **Direct acquisition by stereo cameras** - Precise calibration and temporal synchronization of the cameras is very important for capturing high-quality stereo video.

- **Active depth sensing** - Comprise time-of-flight (ToF) sensors and methods based on structured light such as Microsoft's Kinect. ToF sensors estimate the depth, this means the distance between the sensor and an object by extracting phase information from received light pulses. The structured-light approach usually recovers 3D shape from monocular images using a projector to illuminate objects with special patterns. *Depths brings information about the geometry of the scene.*

- **2D-to-3D conversion** – Existing 2D content can be converted to 3D video by considering several depth cues such as motion parallax, vanishing points/lines, or camera motion in a structure-from-motion framework.
The complete 3D video system is relevant for multiple applications such as broadcast TV, teleconference, surveillance, interactive video, cinema, gaming and other immersive video applications.
The 3D content chain includes a sequence of modules which closely mirror a conventional 2D system but are quite different; they have all to evolve towards 3D regarding the 2D available solutions.

3D content creation involves special production “rules”, e.g. avoid fast pans and manage depth transitions.

Content representation, distribution and display may be performed with many different formats; the best choice depends on distribution constraints, display capabilities, available equipment, target quality, etc.

New 3D display technology is an important driving force: no glasses, multi-persons displays, higher display resolutions, avoid uneasy feelings (headaches, nausea, eye strain, etc.).
3D Video Coding Status Quo
Stereo and Multiview Video Data

- Redundancy reduction between camera views
  - Need to cope with color/illumination mismatch problems
  - Alignment may not always be perfect either
Arc versus Linear Camera Arrangements

*Too long linear camera arrangements become less interesting as border cameras will very likely shoot less relevant parts of the scene*

*Wide viewing range requires the consideration of arc arrangements*

*MPEG FTV recently explicitly acknowledged the importance of arc arrangements for multi-view scenarios (after ignoring for a long time)*
3D Video Format Requirements

- **DISPLAY INDEPENDENCE** – Format should be independent of specific 3D displays.

- **HIGH COMPRESSION EFFICIENCY** - Significant compression gains compared to the independent compression of each view so-called *simulcasting*.

- **CONTINUOUS VIEWING RANGE** – Smooth and high quality motion parallax.

- **VIEW-SWITCHING RANDOM ACCESS** - Any view can be accessed, decoded and displayed in a relatively short time by starting the decoder at a random access point and decoding a relatively small amount of data on which that view may depend.

- **VIEW SCALABILITY** – Only a portion of the bitstream has to be accessed to output a limited number (subset) of the set of encoded views.

- **QUALITY/RESOLUTION SCALABILITY** – A decoder is able to generate effective video output – although reduced in quality to a degree commensurate with the quantity of data in the subset used for the decoding process – although accessing only a portion of a bitstream.

- **BACKWARD COMPATIBILITY** - A subset of the bitstream corresponding to one ‘base view’ is decodable by a standard video decoder, e.g. HEVC decoder.
3D Video Coding Status Quo: Overall Landscape
Redundancies in 3D Video

K. Müller, VCIP2014

Spatial within frames

Temporal along frames

Inter-View between views

Inter-Component between components
Multiview Video Formats: the Menu …

Texture only based

- Multiview Simulcasting

- Frame Compatible Stereo

- Conventional Stereo Video

- Multiview Video, MVC and MV-HEVC standards

Texture plus Depth based

- 2D (Texture)+Depth, MPEG-C standard

- Multiview+Depth (MVD), 3D-HEVC standard
The Texture Only Approach
Multiview simulcasting refers to the independent encoding of each view (ignoring they are like ‘peers’ due to the interview redundancy).

May use any coding technology, e.g. MPEG-2 Video, but an advanced codec such as HEVC is more likely.

This solution has been largely used in may countries due to its quick deployment.
Frame Compatible Stereo Format

- Basic concept: pack pixels from left and right views into a single frame to be coded ‘as usual’:
  - Spatial Multiplexing: side-by-side, top-bottom, checkerboard formats
  - Time Multiplexing: views interleaved as alternating frames or fields

- In such a spatial format, half of the coded samples represent the left view and the other half represent the right view; thus, each coded view has half the resolution of the full coded frame.
Conventional stereo refers to the case where two full resolution stereo views are coded exploiting their interview redundancy.

MPEG-2 Video, MPEG-4 Visual and the MVC standards offer full stereo coding solutions with increased compression efficiency.
Multiview video (MVV) refers to a set of $N$ temporally synchronized video streams coming from cameras capturing the same real scenery from different viewpoints.

- Provides the ability to change viewpoint freely with multiple views available
- Renders one view (real or virtual) to legacy 2D display
- Most important case is stereo video ($N = 2$), generating a depth impression with each view derived for projection into one eye
Multiview Video Coding (MVC) Standard

- MVC is a H.264/AVC extension without any changes of the slice layer syntax and below and of the decoding process.
- Provides coding of multiple views, stereo to multiview.
- Exploits redundancy between views using inter-camera prediction to reduce the required bitrate.
- It is mandatory for the multiview stream to include a base view, which is independently coded from other non-base views.
- For similar PSNR, the MVC coding gains are:
  - For stereo video, the rate of the dependent view is reduced around 30%
  - For multiview, rate savings over all views are about 25%
Disparity-Compensated Prediction

- Use previously decoded pictures in neighbor views as additional reference pictures
- Only construction of reference picture lists is modified from H.264/AVC
Interview Prediction: Basics

Many prediction structures are possible to exploit interview redundancy, trading-off differently memory, delay, computation and coding efficiency.

- Pictures in the non-base views are not only predicted from temporal references (in the same view), but also from interview references (in the other views).
- Limitations: i) inter-view prediction only from same time instance; ii) cannot exceed maximum number of stored reference pictures.
- The prediction is adaptive, so the best predictor among temporal and interview references can be selected on a block basis in terms of RD cost.
MVCPrediction Structures

- **View-progressive encoding** – View dependencies are exploited only for the first frame of each GOP

- **Fully hierarchical encoding** – Bidirectional predictions are allowed both in the time and view dimensions
MVC Compression Performance

Simulcasting versus MVC comparison

8 views (with 640×480 resolution), and considering the rate for all views

~25% bit rate savings over all views for same PSNR
MVC: Subjective Stereo Performance

Base view fixed at 12 Mbit/s; dependent view at varying percentage of base view rate.

- MVC achieves comparable perceptual quality to simulcasting with as little as 25% rate for the dependent view (75% gain); this rate may have to be higher for lower rates than 12 Mbit/s for the main view.

- For similar PSNR, the gains are only about 30% for the dependent view.

- This experiment shows that the 2 views don’t need to have the same PSNR quality.
MVC Limitations

- Acquisition and production of video with large camera arrays is hard, expensive and uncommon

- Only horizontal parallax and linear camera arrangements are considered

- MVC is more efficient than simulcast but the rate is still rather proportional to the number of views (varies with scene, camera arrangements, etc)
The Texture+Depth Approach
A depth map is a ‘gray image’ containing information with the distance from the scene objects to the camera.

Depth maps may be obtained by:
- Special range cameras
- Extraction from texture
- Inherent to the content, e.g. computer-generated imagery

Depth maps provide important information about the scene geometry.
Representing Depth ...

Store inverse depth

\[
I_d(z) = \text{round} \left( 255 \cdot \frac{1}{z} - \frac{1}{z_{\text{max}}} \right) / \left( \frac{1}{z_{\text{min}}} - \frac{1}{z_{\text{max}}} \right)
\]

where \( z_{\text{min}} \) and \( z_{\text{max}} \) are the minimum and maximum depth of the scene, respectively.
Depth Maps Properties

- Sharp edges at object borders
- Large areas of gradual variation in object areas
- Edges in depth maps are correlated with edges in video pictures
Texture and Depth ...

Depth-enhanced formats are suitable for generic 3D video solutions. While a single format is used, all necessary views for any 3D display are generated from the decoded data, e.g., by means of depth image based rendering (DIBR).
Depth-Image-Based Rendering (DIBR)

- In the general case, 3D warping is done using projective matrices and depth info.
- When cameras are rectified, 3D warping reduces to a simple 1D shift.
- Views may be either extrapolated or interpolated.
The MVD format (independently) encodes both the texture and the depth data for the same number of views.

MVD is the reference format for other MPEG 3D Video formats where the texture and depth views are not independently encoded.
Depth Coding vs Texture Coding

- Depth has unique signal properties relative to natural images
  - Larger homogeneous areas inside scene objects
  - Sharp transitions along object boundaries

- Depth maps are not reconstructed for display but rather for view synthesis of the video data (we never see depth maps!)

- Depth quality is indirectly assessed through the synthesized texture quality

- Depth represents a shift value (disparity) for color samples from original views
  - Coding errors in depth maps result in wrong pixel shifts in synthesized views
  - Errors (in the synthesized views) are especially visible around depth discontinuities at the borders of objects with different scene depth

- Depth compression algorithm needs to preserve depth edges much better than current texture coding methods such as H.264/AVC and HEVC
Combining Coding with Synthesis

* As the transmission rate is limited, typically only a small number of texture and depth views may be coded.

* However, an arbitrarily large number of views may need to be rendered.

* Using depth-image-based rendering (DIBR) techniques, a continuum of views may be synthesized based on the limited set of decoded views.

**Encoding side**

- Limited Camera Inputs
- Data Format
  - Constrained Rate (based on distribution)

**Decoding and synthesis side**

- Auto-stereoscopic N-view displays
  - Variable stereo baseline
  - Adjust depth perception
- Stereoscopic displays
  - Wide viewing angle
  - Large number of output views

- Left
- Right
Trading-off Bitrate with 3D Rendering Capability

3DV coding should be compatible with:
• existing standards
• mono and stereo devices
• existing or planned infrastructure

More for less!
HEVC 3D Related Extensions

- MV-HEVC - Simple stereo/multiview extension, potentially including (independent) encoding of depth maps as additional color plane

- 3D-HEVC - More efficient video-plus-depth coding
  - Scalable stereo/multiview
  - Combined coding of video and depth
  - Closer integration with view synthesis to save data rate by irrelevance criteria, particularly for larger view ranges which are costly in terms of data rate
MV-HEVC Approach

from K. Müller, VCIP2014
3D-HEVC Approach

from K. Müller, VCIP2014

HEVC-based codec with additional coding tools for dependent views and depth maps
3D Video Coding Status Quo: 3D-HEVC Standard
Coding Correlated Temporal Cubes ...

Linear arrangement, horizontal parallax only
**Coding of Views**

- **Base view** - Coded using a fully HEVC compliant codec

- **Dependent views and depth data** – Coded with modified HEVC codecs including additional coding tools and inter-component prediction techniques using data from already coded components at the same time instance, notably
  - Coding of dependent views using disparity-compensated prediction, inter-view motion prediction and inter-view residual prediction.
  - Depth map coding using new intra coding modes, modified motion compensation and motion vector coding, and motion parameter inheritance.
  - Encoder control for depth-enhanced formats using view synthesis optimization with block-wise synthesized view distortion change and encoder-side render model.
  - Decoder-side view synthesis based on DIBR for generating the required number of display views.
Coding of Texture Views

Coding of independent view:
- Unmodified HEVC

Coding of dependent views:
Inter-view correlations are exploited by prediction-based coding tools:
- Disparity-compensated prediction
- View synthesis prediction
- Depth-based block partitioning
- Inter-view prediction of motion parameters
- Inter-view prediction of residual data

from K. Müller, VCIP2014
Disparity-Compensated Prediction

Used for texture and depth views!

**Disparity-comp. Pred.**

- Apply motion prediction between different views
- Use already coded pictures inside current access as additional reference pictures
- Main source for coding gain in MVC and MV-HEVC
- No changes at CU level, only construction of reference lists is modified

*from K. Müller, VCIP2014*
View Synthesis Prediction

- Extended interview prediction
- Usage of 3D scene information (depth data and camera parameters in 3DVC)
- Warping or projection of video pixels, using associated depth pixels

(from K. Müller, VCIP2014)
Depth-based Block Partitioning

- Prediction of segmentation information from an already decoded depth map
- Arbitrarily-shaped binary block partitioning for texture block is derived from depth map
- Motion compensation for each partition

from K. Müller, VCIP2014
Inter-View Motion Parameter Prediction

- Motion is similar in different views
- Use disparity-compensated motion vector of base view for motion prediction in dependent views

Disparity vector is calculated from an estimated depth map.

from K. Müller, VCIP2014
Inter-View Residual Prediction

Advanced Residual prediction (ARP):

* Predict residual of a current block using a coded residual block in a reference view
* In temporal ARP, a disparity-compensated residual between different time instances is used
* In inter-view ARP, a motion-compensated residual between different views is used

from K. Müller, VCIP2014
Coding of Depth Maps

from K. Müller, VCIP2014

Coding of depth or disparity maps:

- Inter-view and additionally inter-component correlations are exploited by prediction-based coding

Tools:

- Disparity-compensated prediction for dependent view
- Depth modelling modes
- Segment-wise DC prediction
- Motion parameter inheritance
- Quadtree prediction
- Synthesized view distortion optimization
Depth Modeling Modes

Usual HEVC tools are good to code the smooth depth blocks!

New intra prediction modes

- Representation of depth edges
- Block partition into two regions with constant sample values
- Direct signaling of wedgelet separation line (E->S) through index
- Prediction of partitioning information based on co-located texture block
- Optional transform coding of residual

from K. Müller, VCIP2014
Depth Encoder Control: View Synthesis Optimization

- Coding artifacts in depth data are only indirectly perceivable in synthesized video data as decoded depth maps themselves are not visible.

- An exact mapping between the distortion of a block of the depth data and an associated distortion in the synthesized view is not possible regarding only the depth data within a currently processed block, e.g. due to occlusions, etc.

- Depth coding efficiency can be improved by including in the RDO Lagrangian cost function a modified distortion measure for depth coding (and not simply some usual depth distortion such as SSD or SAD).

- To assess the impact of the depth distortion on the synthesized views, the encoder needs to include some synthesis/rendering capabilities.

from P. Merkle, Fraunhofer HHI
Depth-based View Synthesis

- To consider the impact of depth coding requires including rendering in the encoding process; since complexity is a critical factor, a simplified rendering method is used.

- After decoding the 3D video content, a decoder-side synthesis algorithm generates the required number of dense views for a particular multiview display.

- Since the proposed 3D video codec produces a view- and component-scalable bitstream, two main synthesis approaches can be applied:
  - View synthesis from a video-only decoded bitstream – only operates on the decoded video data (depth may be generate from disparities)
  - View synthesis from a full MVD decoded bitstream - based on classical depth image based rendering (DIBR) solutions
How to measure the quality of the ‘synthetic’ views for which no ‘real’ references exist?

A common solution is to compute a PSNR comparing the decoded synthesized views with the synthesized views from original uncoded video and depth data.

Naturally, subjective testing is also largely used …
## Average Bitrate Savings (BD-Rate)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>MV-HEVC over Simulcast</th>
<th>3D-HEVC over Simulcast</th>
<th>3D-HEVC over MV-HEVC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>overall PSNR vs. total bitrate, (original and synthesized positions)</td>
<td>overall PSNR vs. total bitrate, (original and synthesized positions)</td>
<td>overall PSNR vs. total bitrate, (original and synthesized positions)</td>
</tr>
<tr>
<td>Balloons</td>
<td>-24%</td>
<td>-39%</td>
<td>-20%</td>
</tr>
<tr>
<td>Kendo</td>
<td>-21%</td>
<td>-38%</td>
<td>-22%</td>
</tr>
<tr>
<td>Newspaper</td>
<td>-26%</td>
<td>-39%</td>
<td>-19%</td>
</tr>
<tr>
<td>GT_Fly</td>
<td>-40%</td>
<td>-54%</td>
<td>-24%</td>
</tr>
<tr>
<td>Poznan_Hall</td>
<td>-24%</td>
<td>-38%</td>
<td>-19%</td>
</tr>
<tr>
<td>Poznan_Street</td>
<td>33%</td>
<td>41%</td>
<td>-13%</td>
</tr>
<tr>
<td>Undo_Dancer</td>
<td>-36%</td>
<td>-50%</td>
<td>-21%</td>
</tr>
<tr>
<td>Shark</td>
<td>-40%</td>
<td>-58%</td>
<td>-30%</td>
</tr>
<tr>
<td>Average</td>
<td>-40%</td>
<td>-45%</td>
<td>-21%</td>
</tr>
</tbody>
</table>

*from K. Müller, VCIP2014*
Coding Efficiency: Objective Evaluation

Average PSNR of Original and Synthesized Views versus Bitrate

from K. Müller, VCIP2014
Coding Efficiency: Objective Evaluation

Average PSNR of Original and Synthesized Views versus Bitrate

from K. Müller, VCIP2014
Coding Efficiency: Objective Evaluation

Average PSNR of Original and Synthesized Views versus Bitrate

from K. Müller, VCIP2014
Coding Efficiency: Objective Evaluation

Average PSNR of Original and Synthesized Views versus Bitrate

from K. Müller, VCIP2014
3D-HEVC Extensions: Limitations

The current visual representation *status quo* only provides standard efficient multiview video coding solutions for

* Linear and horizontal-only parallax camera arrangements
* Narrow baselines
* Reduced viewing ranges

Moreover

* 3D-HEVC reference software considers a limited number of horizontal-only parallax views (64)
* Some evidence that 3D-HEVC does not provide efficient performance for a scenario with many high density views
More advanced 3D displays are emerging, providing some or all of the following:

- Glass-free experience
- Render hundreds (dense set) of linearly or angularly arranged views
- Very wide viewing range
- Both horizontal and vertical parallaxes
- Smooth transition between adjacent views (motion parallax), with “walk-around” feeling – *no limited number of sweet spots*
- Effective image resolution NOT divided by the number of displayed views
- Reduced eye fatigue as reduced accommodation-vergence conflict
- Higher immersion, realism and comfortable viewing experience
Towards Interactive Reality …

★ Instead of passively undergoing the producer’s choices, the user now decides what he/she wants to see from the real world …

★ Users should be able to interact with reality in a very smooth and realistic way …

★ This may be named as free viewpoint, free navigation, interactive reality …

★ Interactive Reality is very much a symbiosis between video & gaming/interactivity …

★ Interactive Reality is not ‘virtual or augmented reality’ that gives a too strong feeling of ‘synthetic content’ …

★ Interactive Reality is about natural content …

inspired by G. Lafruit, July 2015
WHAT'S YOUR NEXT STEP?
5

3D Visual Coding Evolution
“The most perfect photograph currently shows only one aspect of reality; it reduces to a unique image fixed on a plane, as a drawing or a painting would be traced by hand.”

“Can we ask photography to render all the richness that the direct view of an object offers?”

Gabriel Lippmann (1845 – 1921)

Lippmann is remembered as the inventor of a method for reproducing colours by photography, based on the interference phenomenon.

Nobel Prize in Physics for 1908.
3.1 Plenoptic Function and Related Concepts
The Plenoptic Function and the Observer

- The world is made of 3D objects, but these objects do not communicate their properties directly to an observer.
- Rather, the objects fill the space around them with the pattern of light rays that constitutes the plenoptic function, and the observer takes samples from this function.
- The plenoptic function serves as the sole communication link between physical objects and their corresponding retinal images.
- The plenoptic function is the intermediary between the world and the eye/vision.

The 7D Plenoptic Function…

The Plenoptic function measures the intensity of light seen from:
- any viewpoint, camera centre 3D spatial position \((x,y,z)\)
- any angular viewing direction \((\theta,\phi)\)
- over time \((t)\)
- for each wavelength \((\lambda)\)

The Plenoptic function represents all the information available to an observer at any point in space and time.

The Plenoptic function can represent every possible view, from every position, at every moment, and at every wavelength.
Restricting the Plenoptic Function: Light Fields

- It is possible to reduce the dimensionality of the Plenoptic function adopting the following restrictions:
  - Radiance of a light ray remains constant along its path through empty space (one spatial dimension reduction)
  - Time is fixed (static scene)
  - Specific wavelength

The 4D light field is a plenoptic representation describing the amount of light faring through every point in space \((x,y)\) in every direction \((\theta,\phi)\).

- It is essential to measure/sample the Plenoptic function using appropriate sensor devices.

- Image-based rendering regards the sampling and reconstruction of the Plenoptic function, e.g. creating new views from sampled views.
The Challenge: Sampling the Plenoptic Function …

How do we Light Field sample this type of scene?

Again Sensors and Displays in the lead…

and New Representation Models …

and Increased Immersion!
A sensor is a transducer whose purpose is to sense some characteristic of its environment. It detects events or changes in quantities and provides a corresponding output, generally as an electrical or optical signal.

✶ Up to Now … Video sensors
  - With increased spatial resolution
  - With increased frame rate
  - With increased dynamic range
  - ...

✶ From now on …
  - Arrays of video sensors
  - Arrays of lens
  - 3D scanners
  - ...
Representation Models … we Need New Representation Models …

★ Up to now … Image and video represented as rectangular sets of regular grid positioned samples
  ● No need for sample position coding
  ● Adoption since JPEG, H.261, MPEG-1 Video … up to HEVC …

★ From now on …
  ● The one million dollars question!
Interactive Reality should involve:

- Moving laterally such as moving along the cameras in current MPEG multiview video solutions (linear, horizontal arrangements)
- Moving in the z-axial direction such as zooming
- Changing the focus position such as refocusing
- Moving freely in any direction like there were cameras anywhere
- Combination of the above to have a real Free Navigation in the volume enclosed by the set of cameras

...
Behind each microlens, a micro-image (MI) is formed...
Plenoptic Imaging: the Sampling Models

Regular Sampling (no explicit coding of sample positions)

- **Super Multi-View** - High density, wide range, array of conventional, monocular cameras with horizontal or both horizontal and vertical parallaxes, with linear or arc arrangement
  - *Video around Object* – Sequence of frames taken with regular camera moving around an object
  - *Rotating Object* – Sequence of frames taken with a fixed, regular camera when an object rotates around itself

- **Light Field Imaging** - Lenslet array into the optical path of a monocular camera; *multicamera in a box*

Irregular Sampling (sample positions need coding)

- **Labeled Point Cloud** - Set of data points in some coordinate system acquired with a 3D scanner with associated colour labels (or changing over directions)
Models … The Less, the Better …

Is there a single representation model that may accommodate all the previous acquisition models, notably for coding purposes?

Is there at least a very limited number of relevant representation models considering the relevant application scenarios and associated displays?
### Super Multi-View (SMV)
- Tens or hundreds of cameras
- Expensive, camera rig with many cameras
- Wider baseline
- Horizontal or full parallax
- Full resolution for each view
- Sparser sampling of the light field
- Linear, arc or sparse camera arrangements
- Outcome is several viewpoints with horizontal and vertical disparities

### Light Field
- Single camera, no need for camera synchronization
- Lenticular array composed of a large number of micro-lenses (ML)
- Baseline limited by size of ML array
- Full parallax
- Full resolution shared by ML
- Denser sampling of the light field
- Trade-off between spatial and angular information
- Outcome is array of Micro-images each associate to a ML with light coming from several view angles
Combining Regular and Light Field Cameras

Super multiview video

Light field camera

Hybrid cameras super multiview video

Light field super multiview video
More Information, Better Analysis ...

- Plenoptic imaging gathers significantly more light information, capturing a richer 4D/5D light field structure with textural and geometric information.

- Early vision extracts as much information as possible about the structure of the plenoptic function but clearly only a small portion of all information...

- It should be possible to reach better analysis performances, notably increasing robustness to difficult environmental conditions (e.g. unfocused, low light, rain, fog, snow, smoke, glare), unstructured scenes and unconstrained acquisition:
  - **Computer vision**
    - Mapping, modelling, segmentation, localization, tracking, classification, object recognition, …
  - **Biometrics**
    - Face, gait, palmprint, etc. recognition
  - …
5.2 Plenoptic Imaging: Super Multiview Imaging
Playing with Cubes … But What Cubes?
The 3D-HEVC Benchmark

Skipping input views in a dense camera array drastically reduces the quality of the view synthesis (around 5 to 10-13 dB quality loss), especially when coding only a few input camera views in large baseline settings.
# 3D-HEVC Strengths and Weaknesses

- Using depth maps improves the coding efficiency of 3D-HEVC in narrow baseline scenarios.
- There is evidence that dropping 5 to 9 intermediate views, and resynthesizing them may still achieve a good MOS (depending on the content, depth quality, synthesis method).

- **Baseline limitation** - Early experiments suggest that at a relatively moderate baseline the coding performance is not better than simulcast-HEVC, where all camera views are independently coded.

- **Quality limitation** - The view synthesis quality hardly reaches beyond 35dB with the current status of the MPEG Depth Estimation Reference Software (DERS) and View Synthesis Reference Software (VSRS).

- **Camera arrangement limitation** - The influence of non-linear camera arrangements and disparity/depth map errors on the view synthesis quality remains an open issue.

*inspired by G. Lafruit, Dec. 2014*
The Bigger Quality Problem …

- PSNR is highly sensitive to occlusions and object silhouette errors, and less tolerant to view synthesis artifacts than human viewers.

- PSNR hardly reflects the subjective quality
  - especially the Human Visual System (HVS) masking effects over adjacent views in SMV displays
  - especially in wide baseline applications with few coded views, where the rendered images should be experienced by the user as “plausible” rather than perfectly reflecting the scene’s reality

- The subjective Mean Observation Score (MOS) and PSNR metrics are weakly correlated, and only their monotonic relationship (both increase and decrease together) is not put in question. However, there is still no consensus on a better metric …

inspired by G. Lafruit, Dec. 2014
Super Multiview Imaging Coding
Super Multiview Imaging Coding

Just Extending the Conventional Way
Playing with Cubes … But What Cubes?
Full parallax SMV content can be coded with a multi-view video coding standard with an adaptation of the inter-view prediction structure.

- The views are first scanned in spiral and after realigned horizontally.
- The horizontal arrangement is then MVC coded, e.g. using a IBP prediction structure.
- There are some unsuitable and ineffective predictions.

from F. Dufaux, “Full parallax 3D video content compression”, ICIP2015
Full Parallax by Horizontal Realignment

2D Scan order: (a) spiral, (b) perpendicular, (c) diagonal and horizontal inter-view reference picture structures: (d) hierarchical, (e) IBP, (f) IPP

from F. Dufaux, “Full parallax 3D video content compression”, ICIP2015
The main drawback is the limited number of vertical inter-view predictions.

Or the limited number of views using both horizontal and vertical references.

from F. Dufaux, “Full parallax 3D video content compression”, ICIP2015
The central view (I) is coded first and cannot use inter-view references.

The N-1 (respectively M-1) views that are in the same horizontal (resp. vertical) axis as the central view are coded using only one inter-view reference, being the nearest view in the central direction.

All the other views are coded using one horizontal and one vertical inter-view references being the nearest views in the central direction.

- Allows the use of an horizontal and a vertical inter-view reference picture for a large number of views.
- Minimizes the distance between the coding views and their inter-view reference pictures and does not use diagonal references.
- Better RD performance than previous solutions.

from F. Dufaux, “Full parallax 3D video content compression”, ICIP2015
Super Multiview Imaging Coding

Going to the Essence
★ MV-HEVC and 3D-HEVC use essentially stereo disparity/depth estimation and view synthesis techniques.

★ Typically, only the two surrounding views of the requested virtual viewpoint are used, not taking benefit of all available camera views for achieving better compression performance.

★ Future solutions should overcome this approach to “extract the essence” of the plenoptic information in the observed scene, considering the information for all available views.

*inspired by G. Lafruit, Dec. 2014*
Playing with Cubes … But What Cubes?

For simplification, horizontal only parallax
Epipolar-Plane Images (EPI) Cube

An EPI corresponds to a \((u, s)\)-slice of the cube of views (fixed \(v\) coordinate, \(v^*\))

- A scene point (voxel) is mapped into a line segment in a EPI image.
- The denser the view sampling, the better defined the EPI line segments.
- The slope of the line segment is proportional to the scene point depth.
- Line segments with smaller slopes (points closer to the camera) occlude line segments with larger slopes.
**EPI Compact Representation**

- An EPI exhibits high coherence and redundancy
  - Positions along an EPI line segment represent the same scene point in various views
  - The slope of a voxel trajectory tells how that point in 3D space moves left-right in the projected images, when the camera moves left to right (by switching over different cameras).
  - All positions along an EPI line segment have the same disparity value $d$

- EPI redundancy can be exploited to obtain a more compact EPI representation
  - An EPI line segment can be compactly represented by a tuple $l = (d, u, s, r)$ where $r$ is the average radiance of the EPI position $(u, s)$ and $d$ its disparity
  - It is possible to fully reconstruct an EPI only knowing the EPI line segment tuples

*inspired by C. Kim et al., “Scene reconstruction from high spatio-angular resolution light fields”, SIGGRAPH2013*
EPI based 3D Video Coding

- Limited set of input views is conventionally coded, e.g. with 3D-HEVC.
- Conventionally decoded views are used to create the EPI cube for a specific time instant.
- Intermediate views are rendered at decoder based on the generated EPI cube.
EPI based View Rendering: Pixel Rendering

- Intermediate views are directly rendered from the EPI compact representation.
- Row $v^*$ of an intermediate view $s^*$ frame is rendered by intersecting the EPI$_v$ tuples $l = (d, u, s, r)$ with a virtual line at viewing position $s^*$.
- The radiance value $r$ of the line segment, $ac$, is assigned to the intersection point, $b$, in the new intermediate view $s^*$.
- Unfortunately, this works well only for linear camera arrangements.
EPI based View Rendering: Major Benefits

- Any view position (between the leftmost and the rightmost views) can be directly rendered from a single EPI compact representation
  - There is no need to obtain depth every time a new view position has to be rendered
- EPI based disparity estimation may handle occlusions more robustly as the EPI compact representation takes in account all views available (cube of views)
Are temporal cubes exploiting interview redundancy, such as in 3D-HEVC, better than EPI-based light field cubes exploiting the temporal redundancy?
5.3 Plenoptic Imaging: Light Field Imaging
Light Field Cameras

A light-field camera, also called a *plenoptic camera*, captures the available light in a scene coming from many directions.

Light field cameras offer an appealing alternative to conventional imagery by gathering significantly more light over a wider depth of field, and capturing a rich 4D light field structure that considers textural and geometric information.
Light Field Imaging Acquisition

- The light field camera breaks up the main image with an array of microlenses over an image sensor.
- Each microlens works as an individual low resolution camera, recording a different perspective of the same scene with slightly different angles.
- “Taking a conventional photograph is like recording all the musicians playing together, rather than recording each instrument on a separate audio track.”, Ren Ng, 2006

from C. Conti
What is Changing in the Acquisition?
Dealing with Angles ...
Micro-Images and Pixels ...

Sub-aperture image view of acquired light field
(major: UV ; minor: XY)
32x32x512x512

Microlens view of acquired light field
(major: XY ; minor: UV)
512x512x32x32

from Loïc Baboulaz,
EPFL, 2014
Behind each microlens, a micro-image (MI) is formed...
Light Field Video: Original (2880×1620)
Display Data: a New Stage between Sensor Data and Metadata

Light fields call for computational imaging!

As the light field camera gathers richer information than a conventional camera, computational processes have to be applied to render data to a regular 2D display.

“... images are computed rather than directly recorded”, Ren Ng, 2006
An IMAGE is increasingly much more than the output of a sensor!
Digital Refocusing Example

- Focus the image after the fact, no more blurry pictures!
- Control the focus of every pixel with a “focus brush”
- Move the camera after the fact
- Select objects automatically – based on depth information
- Insert objects into a scene – with proper occlusion based on depth

Refocusing is equivalent to (appropriately) sum captured data extracted from several microlenses.

“Computation is truly an integral component of modern photography”, Ren Ng, 2006
Light Field Displays

In the display, a light field identical to the one emanating from the original scene is recreated.

- The micro-lenses of the lenticular array allow the user to see only a particular part of each micro-image corresponding to the angle of view (allowing motion parallax).

- Light field visual data provides both horizontal and vertical parallaxes across a large range of viewing angles.

HoloVizio 128 WD, WLD

HoloVizio 80WLT
Light Field Display

- A full optical representation is recovered by the intersection of the rays refracted through each microlens.

- Allows a more natural visualization as providing:
  - Continuous motion parallax (in horizontal and vertical directions)
  - More natural vergence-accommodation cues

*from C. Conti*
Light Field Imaging Coding
Light Field Imaging Coding Approaches

- **Standard codecs** - Direct application of standard image codecs such as JPEG, JPEG 2000, H.264/AVC Intra, and HEVC Intra, to the full light field image.

- **DCT-based** – DCT is applied to the micro-images, followed by quantization and lossless coding; possibly, a differential coding between Micro-Images (MI) can be used.

- **Wavelet-based** – Discrete Wavelet Transform (DWT) is applied to the viewpoint images.

- **Elemental images-based** - Viewpoint images or the MIs of a light field image are coded as if they were a video sequence (called *pseudo video sequence*) and then the temporal compression tools of traditional video coders are used.

- **Multi-view based** - The light field image is taken as a group of viewpoint images that is encoded as a multi-view sequence (using interview prediction).

- **Self-similarity based** - Exploits the non-local spatial correlation between MIs.

*from F. Dufaux, “Full parallax 3D video content compression”, 2015*
Light Field Imaging Coding

The Conventional Way
Light field: Bikes (EPFL)
Light field: Bikes zoom (EPFL)
Light field: Books, zoom (EPFL)
HEVC RD Performance per View: Bikes

Results from G. Alves, et al., UFRJ, 2015

Rendering with
Light Field Toolbox, v0.4, D. Dansereau, 2015

Results from
G. Alves, et al., UFRJ, 2015
Various Codecs RD Performance: Bikes

from G. Alves, et al., UFRJ, 2015
HEVC RD Performance per View: Books

Rendering with
Light Field Toolbox, v0.4, D. Dansereau, 2015

Results from
G. Alves, et al., UFRJ, 2015
Various Codecs RD Performance: Books

from G. Alves, et al., UFRJ, 2015
Light Field Imaging Coding

Extending the Conventional Way
Self-Similarity (SS) Compensated Prediction

- The SS compensated prediction allows to exploit the inherent spatial correlation of the light field content.

- A SS reference picture is formed by the previously coded and reconstructed area from the current frame itself.

- The prediction blocks can have different sizes and do not need to be aligned with the micro-image structure; new, appropriate prediction modes may be included in an HEVC codec.

*from C. Conti et al., “New HEVC prediction modes for 3D holoscopic video coding”, ICIP2012*
Test Conditions and Benchmarks

- **Plane and Toy** – 1920×1088
  28×28 Micro-Image

- **Demichelis Cut** – 2880×1620
  38×38 Micro-Image

- **Laura** – 7240×5432
  75×75 Micro-Image

- **HEVC**
- **HEVC Rext 6.0** – Range Extension (Reference Software 6.0)
- **HEVC SCC 1.0** – Screen Content Coding (Reference Software 1.0)
- **HEVC + SS** – with Self-Similarity Compensated Prediction

from C. Conti et al., “New HEVC prediction modes for 3D holoscopic video coding”, ICIP2012
SS RD Performance: Full Plenoptic Image

Plane and Toy

Demichelis Cut

Laura

### HEVC + SS

<table>
<thead>
<tr>
<th></th>
<th>Plane and Toy</th>
<th>Demichelis Cut</th>
<th>Laura</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BD-BR [%]</strong></td>
<td><strong>BD-PSNR [dB]</strong></td>
<td><strong>BD-BR [%]</strong></td>
<td><strong>BD-PSNR [dB]</strong></td>
</tr>
<tr>
<td><strong>HEVC</strong></td>
<td>-20.49</td>
<td>1.53</td>
<td>-29.37</td>
</tr>
<tr>
<td><strong>HEVC RExt 6.0</strong></td>
<td>-5.30</td>
<td>0.36</td>
<td>-17.71</td>
</tr>
<tr>
<td><strong>HEVC SCC 1.0</strong></td>
<td>-3.82</td>
<td>0.26</td>
<td>-14.78</td>
</tr>
</tbody>
</table>

from C. Conti et al., “New HEVC prediction modes for 3D holoscopic video coding”, ICIP2012
Extracting 2D Views from 4D Data

from T. Georgiev et al., “Focused plenoptic camera and rendering”, Journal of Electronic Imaging, 2010
PSNR RD Performance: Rendered Images

- Average and standard deviation PSNR for:
  - Set of 9 views rendered from the holoscopic image
  - Equally-spaced angular positions, main object “in focus”

- HEVC+SS BD-PSNR gains are up to (for Laura):
  - 2.27 dB compared to HEVC
  - 1.57 dB compared to HEVC Rext 6.0
  - 0.73 dB compared to HEVC SCC 1.0

from C. Conti et al., “New HEVC prediction modes for 3D holoscopic video coding”, ICIP2012
SSIM RD Performance: Rendered Images

from C. Conti

* Average and standard deviation SSIM for:
  * Set of 9 views rendered from the holoscopic image
  * Equally-spaced angular positions
  * Depth plane is chosen to have the main object “in focus”

from C. Conti et al., “New HEVC prediction modes for 3D holoscopic video coding”, ICIP2012
Display Scalable Coding Architecture

- To provide backward compatibility with legacy 2D and 3D displays a three-layers hierarchical approach may be used.

- 2nd Enhancement Layer combines a Self-Similarity (SS) prediction with an Inter-Layer (IL) prediction (S3DHolo coding solution).

Standards: Again and More
Video Coding Standards Over Time ...

MPEG Future Video Compression Technology

Cítius, Altíus, Fortíus
Faster, Higher, Stronger
... and More Efficient?
MPEG: What Way Forward?

- Higher compression in video coding seems to remain a fundamental need.
- To define a new standard, improvements over HEVC should be more than incremental.
- HEVC in the market can be expected to make further progress within the next few years, so we are facing a moving target.
- It is important to understand what are the built-in limitations (in normative technology) of HEVC and its extensions.
- Improvement of compression can mean either getting (much) higher compression with increased complexity, or (slightly) higher compression with decreased complexity.
- Naturally, increasing subjective quality versus rate is more important than increasing PSNR versus rate.
- What are the expected operational ranges in terms of bit rates needed in the future?
MPEG Future Video Compression: Starting Another Cycle

- MPEG has received expressions of interest that further improvement of video compression is desirable in various existing and emerging application areas.
- MPEG plans to launch an investigation leading towards the next generation of video compression standards (2020 ?) and intends to establish an experimental environment to enable the investigation of future video compression technology.
- MPEG is interested in receiving reports about compression technology that performs better than the HEVC standard (objectively and perceptually), or would fulfill requirements on compression technology that HEVC might be unable to cover.
- MPEG intends to organize a workshop on future video compression and capturing technologies, to be held during the week of 19-23 October 2015 in Geneva, CH
MPEG Free-Viewpoint Television

Call for Evidence on FTV, Doc. ISO/IEC JTC1/SC29/WG11 MPEG2015/N15095, Warsaw, Poland, June 2015
MPEG FTV: Context

★ 4k/8k UHDTV offers viewing at the highest resolution in visual media. However, it transmits only a single view and users can’t change the viewpoints.

★ Prior stereo and multiview coding standards, such as MVC and MV-HEVC, have focused on the compression of camera views “as is”, all rendered without means to facilitate the generation of additional views.

★ 3D-HEVC assumes a linear, horizontal and narrow baseline arrangement of cameras.

★ Super-Multi-View (SMV) displays are emerging, which render hundreds of linearly or angularly arranged, horizontal parallax ultra-dense views, thereby providing a very pleasant glasses-free 3D viewing experience with wide viewing angle, smooth transition between adjacent views, and some “walk-around feeling” on foreground objects.
MPEG FTV: Objectives

There are substantial commercial interests for more capabilities, e.g. immersive experiences for sporting events, interactive tele-presence systems with realistic and natural interface, and various professional and scientific applications.

1. FTV should enable users to view a scene by freely changing the viewpoints as we do naturally in the real world. It should provide a very realistic glasses-free 3D viewing without eye fatigue.

2. FTV should provide a new data format along with associated compression and rendering technology to address these application scenarios and needs.

3. FTV should enable the generation of additional views from arbitrary and sparse camera arrangements, as well as an ultra-dense representations of a 3D scene. An efficient coded representation of this data format should also be realized.
MPEG FTV: Application Scenarios

Supermultiview

Free navigation
FTV considers Super Multi-View (SMV) and Free Navigation (FN) scenarios/applications.

- Though there exist commonalities between SMV and FN, these two categories are evaluated in a different way: SMV aims at high compression exploiting at best the essential information embedded in all camera views, while improved view synthesis is an additional cornerstone for FN in large baseline camera arrangements.

- **Super Multi-View Objective:** To substantially reduce the data rate required to reconstruct the full set of input views at the receiver compared to existing MPEG state-of-the-art compression standards.

- **Free Navigation Objective:** To substantially improve rendering quality at arbitrary virtual view positions in 3D space. This may be achieved through an alternative representation format (different from 3D-HEVC), in which case compression efficiency must also be considered. There is no intention to standardize post-processing tools.
SMV and FN systems may require technologies that are not currently available in MPEG.

Companies that have developed compression technologies performing better than 3D-HEVC are invited to bring such information to MPEG.

If proposed technology significantly outperforms currently available MPEG technology, MPEG plans to issue a Call for Proposals (CfP), subsequent to the CfE, to develop standards that allow increased compression performances beyond 3D-HEVC in SMV and FN application scenarios.

The timeline for this Call for Evidence has been fixed as follows:
- Test sequences and preliminary 3D-HEVC anchors are available: 2015-06-15
- Final 3D-HEVC anchors are available: 2015-07-10
- Submission of contributions (descriptive document): 2016-02-22
- Decoded sequences, bitstreams and binary decoders are made available by 2016-02-01 (three weeks prior to the February 2016 MPEG meeting)
- Evaluation of the responses at the 114th MPEG meeting (2016-02-22 – 2016-02-26)
JPEG PLENO

JPEG PLENO Abstract and Executive Summary, Doc. ISO/IEC JTC 1/SC 29/WG1 N6922, Sydney, Australia, Feb. 2015
Why Should Pictures be Flat?

Images should be represented and consumed as volumes instead of planar datasets!

from T. Ebrahimi, June 2015
JPEG PLENO: Emerging Imaging Modalities

- LIGHT-FIELD DATA (aka plenoptic data) records the amount of light (the “radiance”) at every point in space, in every direction.

- POINT-CLOUD DATA is a set of data points in a given coordinate system. Such dataset is usually acquired with a 3D scanner or LIDAR and subsequently used to generate and represent a 3D surface.

- HOLOGRAPHIC DATA records interference patterns between a reference (laser) wave and the scene/object wave (reference wave diffracted by the scene); such data may be physically or computer generated.

These new data types can then be processed to recover additional scene (3D) information and to render this information in novel ways.

JPEG PLENO standardization may consider several phases …
JPEG Pleno Imaging Modalities
JPEG PLENO: Functionalities and Applications

Functionalities

★ Ability to manipulate the content after it has been captured
★ Possibility for users to change, in real time, focus, field of depth and stereo baseline, as well as the viewer perspective
★ Relighting would allow users to change the mood of an image
★ Simplification of image compositing and other manipulations such as recoloring based on extracted depth
★ Accurate 3D scene information could be used to provide localization within a scene and enhanced capabilities to better detect/recognize objects or actions

Applications

★ Interactive content viewing, cultural environments exploration, medical imaging checking, more immersive browsing with novel special effects and nicer or more realistic images
JPEG PLENO targets a standard framework for the representation and exchange of new imaging modalities such as light-field, point-cloud and holographic imaging.

- It also targets to define new tools for improved compression while providing advanced functionality support for – but not limited to – image manipulation, metadata, image access and interaction, privacy and security.

- JPEG PLENO will investigate how the evolution to computational imaging approaches can be properly addressed while taking into account JPEG’s legacy formats.
JPEG PLENO: Action Plan

- The JPEG committee intends to interact closely with the actors in conventional and emerging imaging systems
- Organize focused workshops targeted to understand industry needs in terms of technology and supported functionalities
- Requirements have already started to be identified (see doc) …
- Calls for evidence and/or contributions will be issued to launch new standards or extend existing JPEG standards
- To stay posted on the action plan for JPEG PLENO, see the JPEG website (www.jpeg.org) and subscribe to the JPEG PLENO AhG (jpeginnovations-join@listserv.uni-stuttgart.de)
7 Summary & Trends
Visual Coding: the Trends

- Since the 90s, major video compression gains have been obtained in an almost continuous way. However, this pace has been reducing ...
- HEVC is the state-of-the-art pixel-based representation from low to ultra-high resolutions … but the coding paradigm is still the usual one …
- Plenoptic representation should provide a major step forward towards visual realism and immersion
- Plenoptic imaging acquisition and display are still at their infancy. Combined with ultra high spatial resolutions, high frame rates and high dynamic ranges, uncompressed plenoptic imaging rates are scaring …
- For new visual sensors/displays and new visual data, new coding models should be developed 😊
- First light field imaging coding solutions simply extend available standard coding solutions, e.g. JPEG, JPEG 2000, HEVC, 3D-HEVC
- New standards are coming …
New sensors and new displays are coming, sampling better the plenoptic function …

thus, new representation models and new coding solutions have to be designed to provide more powerful user experiences, eventually also with new displays.
The Visual Coding Standardization Path ...

- H.261
- H.262/MPEG-2 Video
- H.263
- H.264/AVC/SVC/MVC
- HEVC
- SHVC

- JPEG
- JPEG-LS
- JPEG 2000
- MJPEG 2000
- JPEG XR
- JPEG XT
- JPEG Pleno ?

- MPEG-1 Video
- MPEG-4 Visual
- MV-HEVC
- 3D-HEVC
- RVC

- ITU
- ISO/IEC 14496-10
- MPEG
3D Video: Success or Not so Much?
Thanks to …

João Ascenso  Catarina Brites  Caroline Conti

Eduardo Silva  Gustavo Alves
Thanks for your Attention!

And thanks to the colleagues who provided me many of the slides …