

3D VIDEO: CONCEPTS, TOOLS AND SYSTEMS

Fernando Pereira



Light is Life ... Both are Complex !



The World and the Humans ...





Visual, Visual, Visual ...



- **It is believed that up to 50% of the human brain is involved in some way in processing visual information**
 - **This reflects the significance of vision for function and survival**
 - **And 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 a 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 !**

Cisco Visual Networking Index, "Forecast and Methodology, 2013-2018", 2014.

**Let's replicate the world to enjoy more realistic
and immersive experiences !**



Visual Representation: What and Why ?

- **Replicating the visual world**
- **Driven/conditioned by available sensors, transmission/storage channels, displays and devices**
- **.... and by Human Vision**
- **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 !**

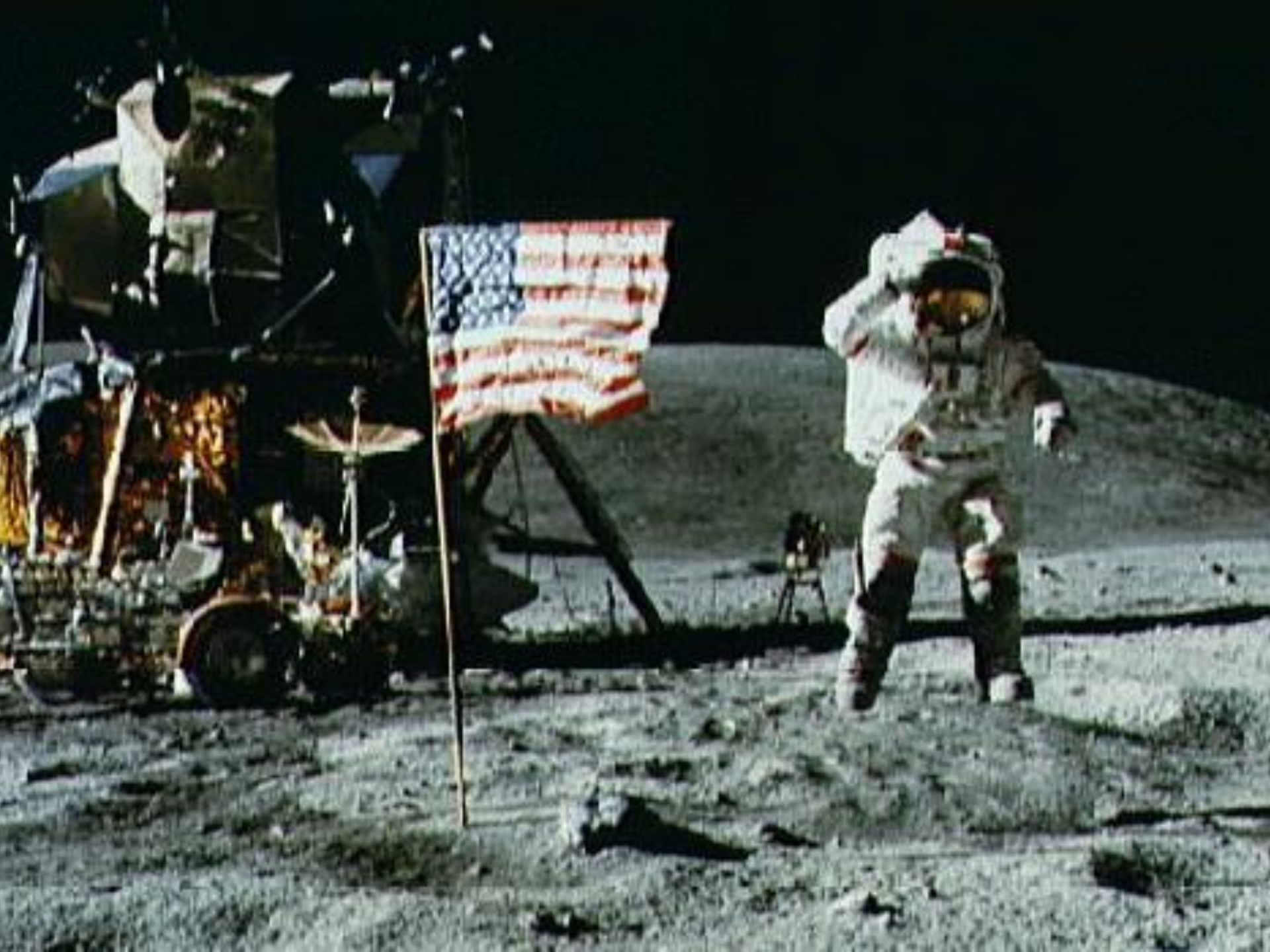


Today, Reality Becomes a Plane ...





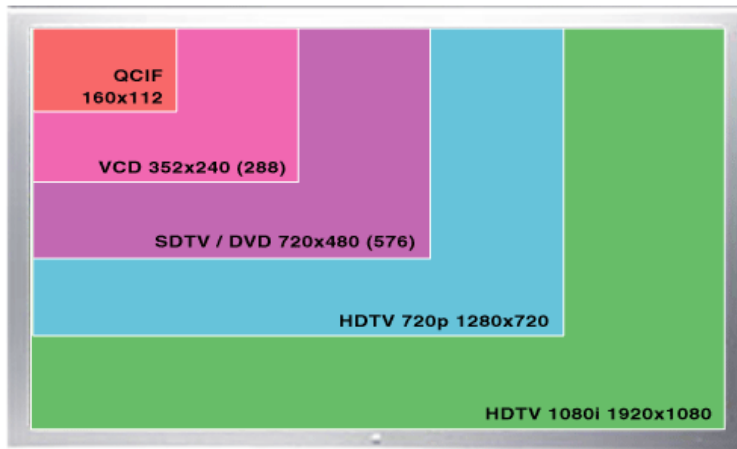




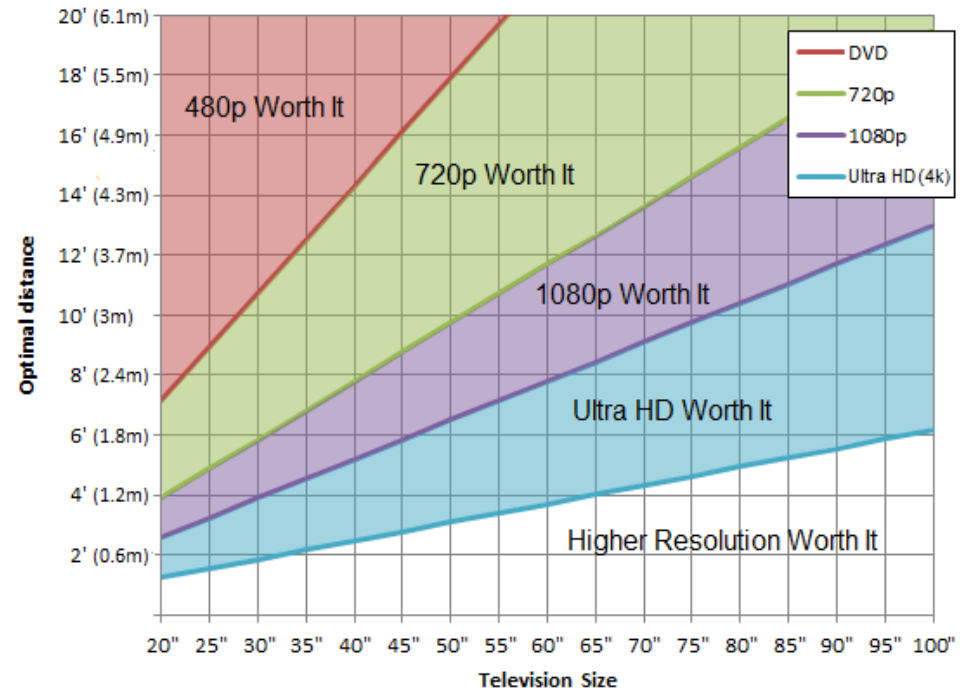




Spatial Resolution: Racing for Immersion



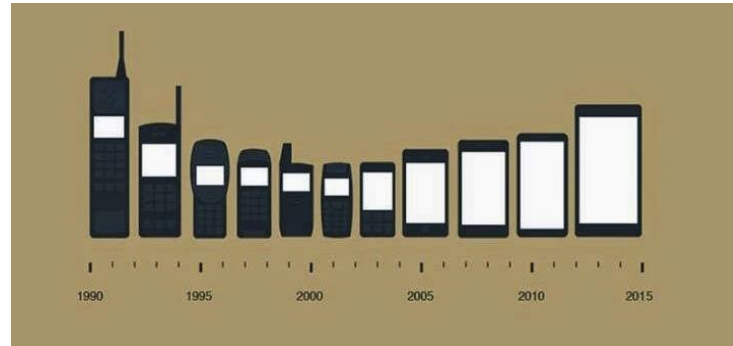
Optimal viewing distance by the size of the television and the resolution



- Higher resolutions are relevant for shorter viewing distances ...
- Shorter viewing distances and large screens increase the sense of immersion ...

The 'End of Times' Approach ...

- **Higher resolutions (at least above 4K) are useless**
 - **New generations just use handheld terminals**
 - **Human visual system does not see the difference anymore**
 - **HD to 4K receiver interpolators are good enough**
- **Sofa TV and big TVs have no future as only old people nowadays see TV ...**
- **3D is dead ...**
- ...



The 'Beginning of Times' Approach ...



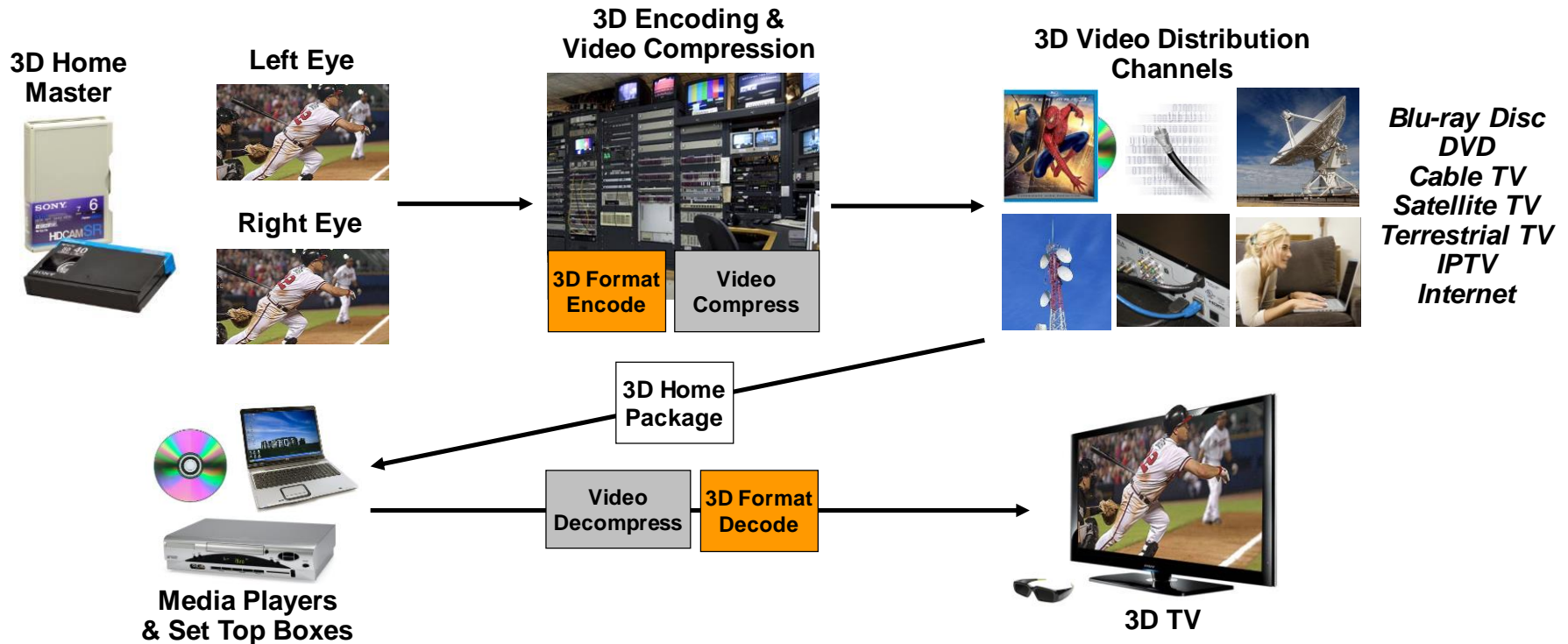
- **Until now was pre-history !**
- **Real fun is just starting ...**
- **Deep immersion is coming ...**

It's a 3D World !

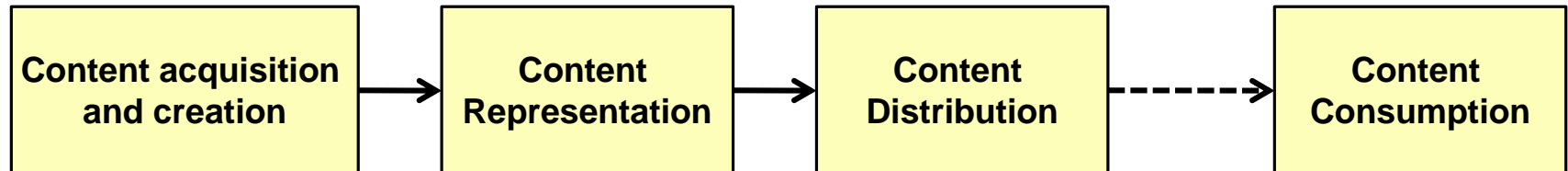


3D Video Applications ...

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.

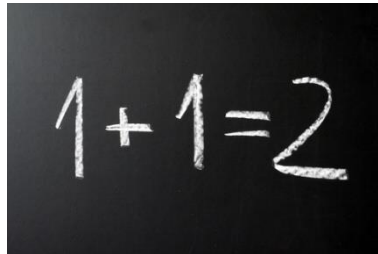


3D Video Content Chain ...

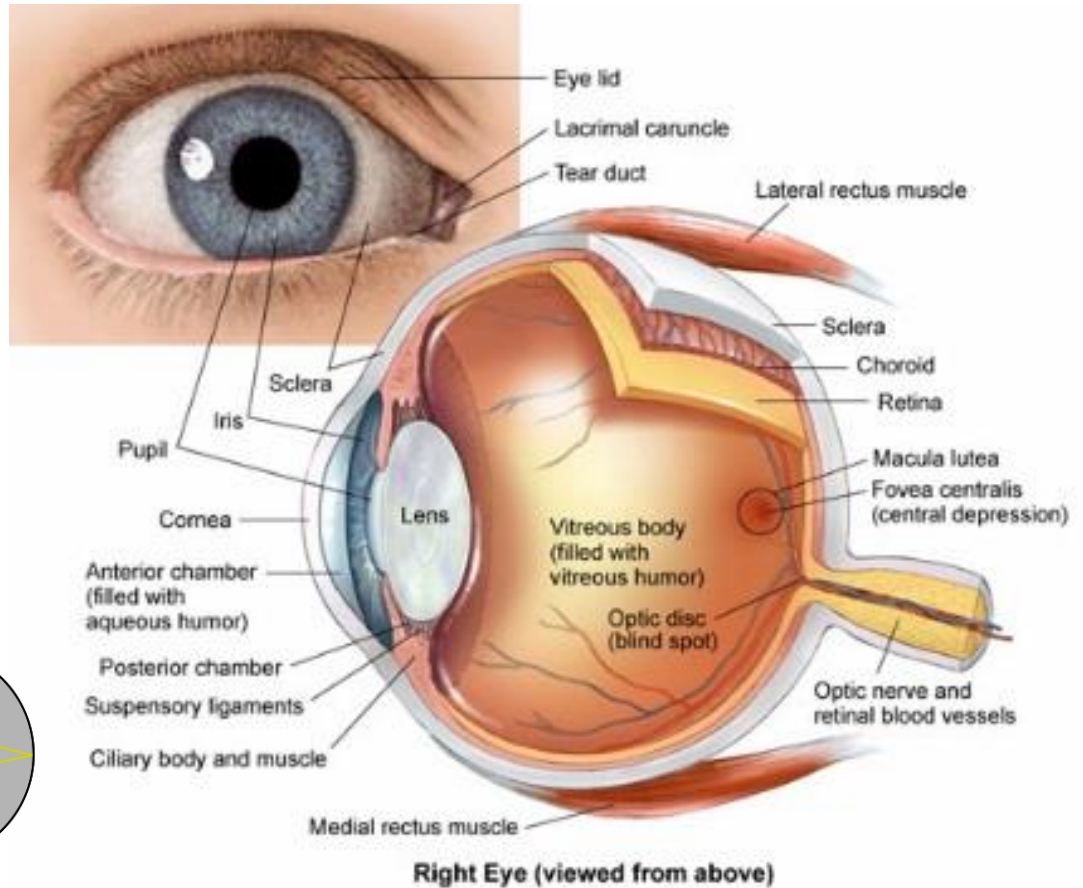
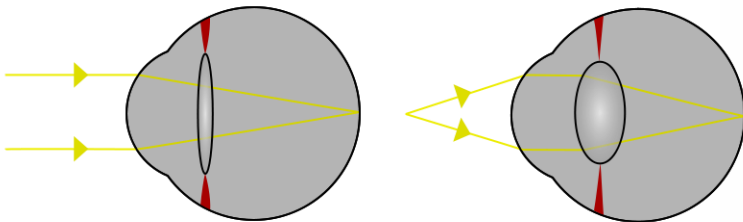


- **The 3D content chain includes a sequence of modules which closely mirror a conventional 2D system but are quite different; they have all to evolve towards 3D regarding the 2D available solutions.**
- **3D content creation involves special production “rules”, e.g. avoid fast pans and manage depth 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 Perception Basics

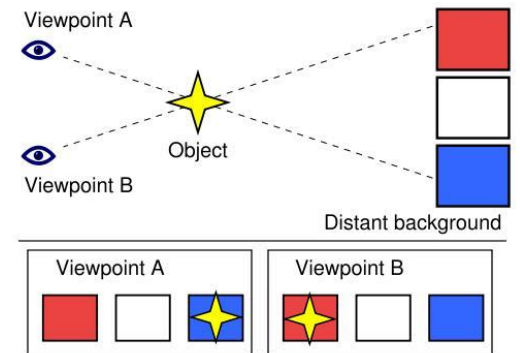
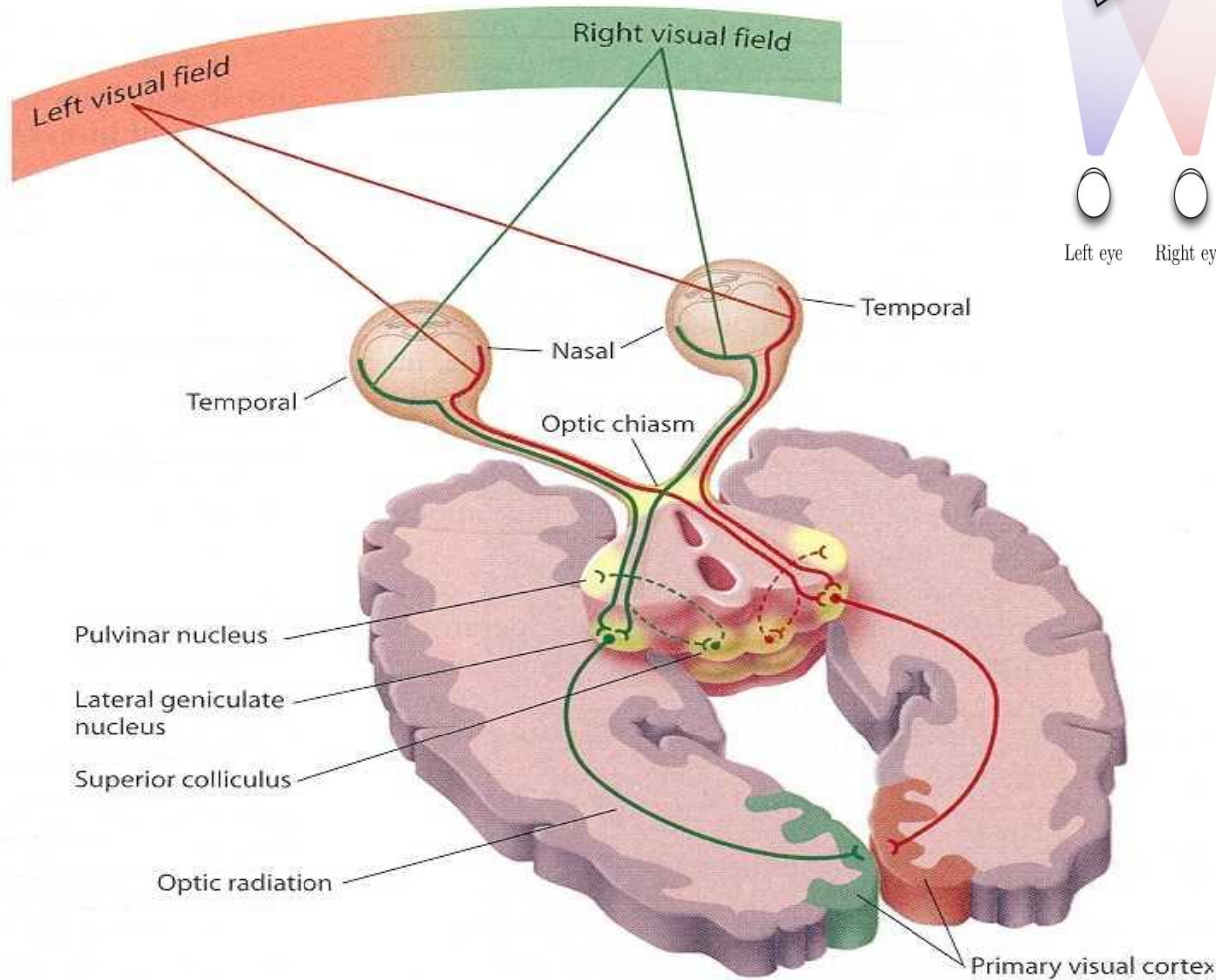
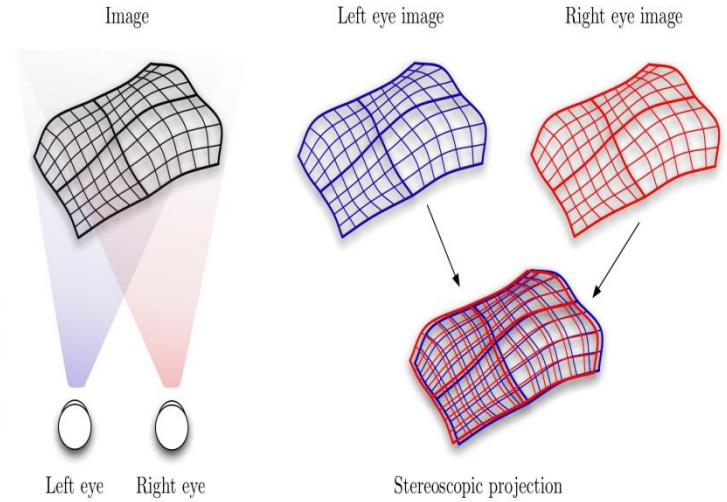


Rod and cone cells in the retina allow conscious light perception and vision including color differentiation and the perception of depth.

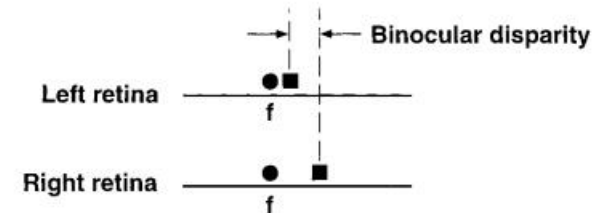
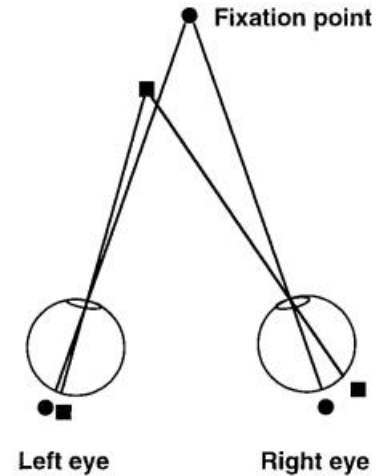
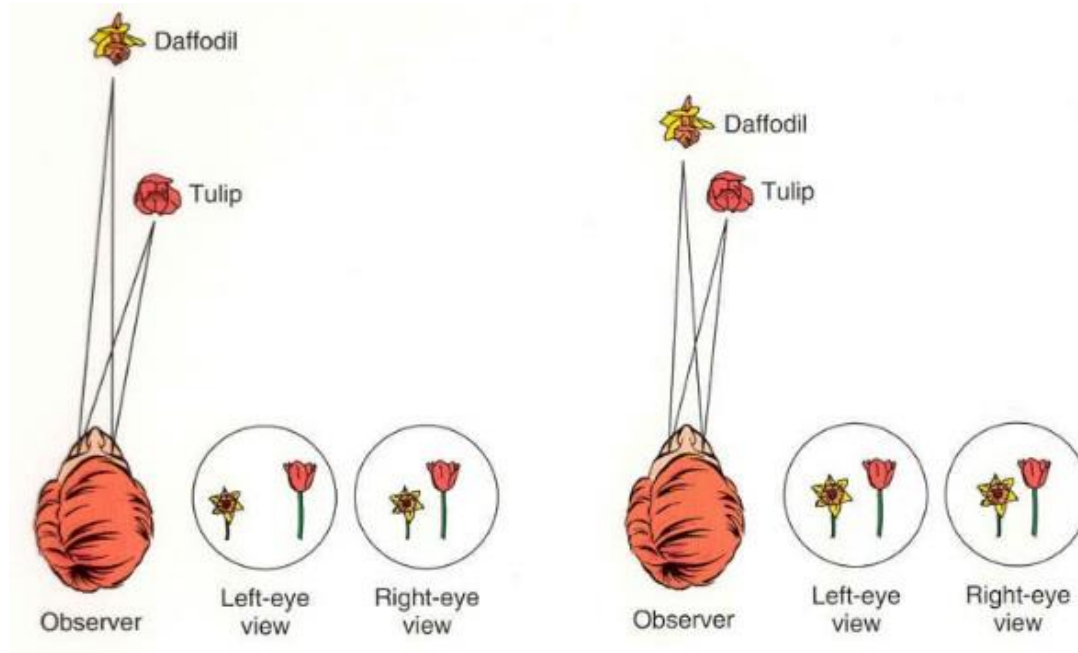


The crystalline lens changes/focus for the light to strike the retina

Human Visual System

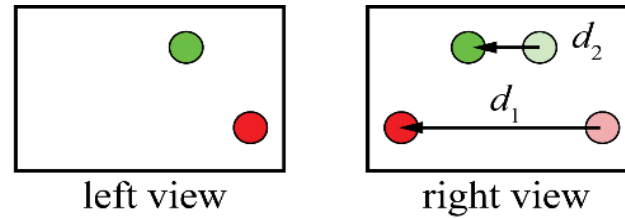


Retinal Disparity

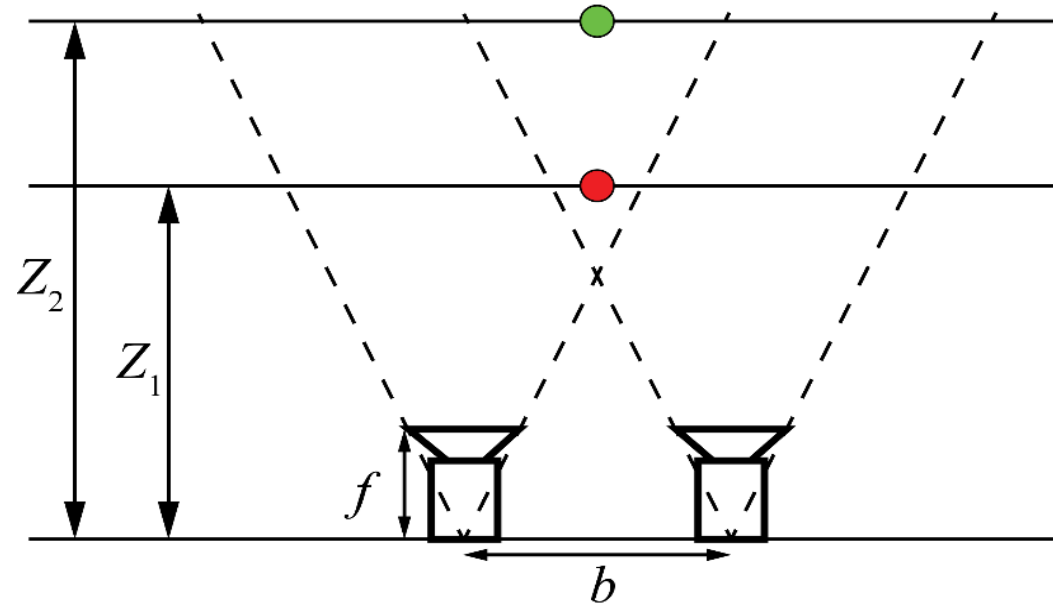


- The retinal images are different in the left and right eyes depending on the object distance and angle.
- The retinal disparity is the slight difference in the two retinal images due to the angle from which each eye views an object.
- The retinal disparity is an important depth cue.

Relation between Depth and Disparity for Two Parallel Cameras

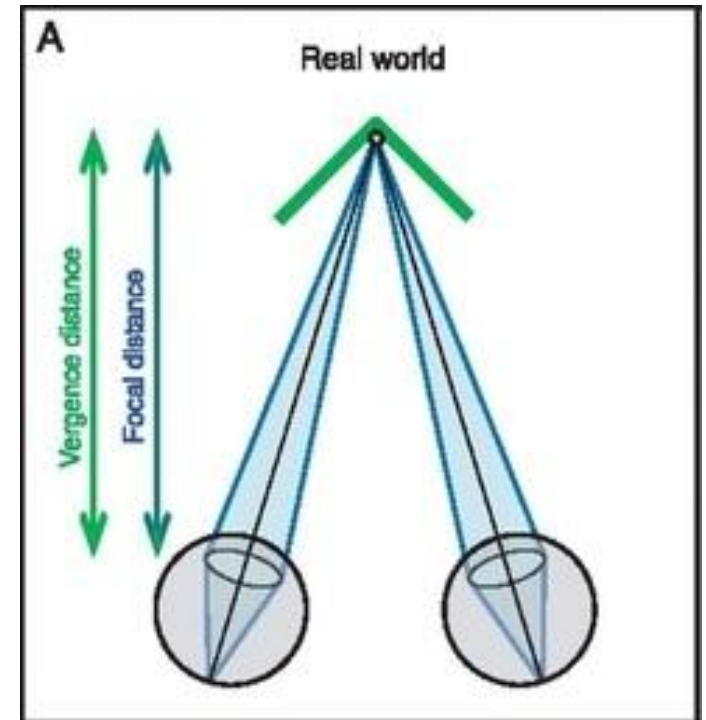


The larger the depth, the smaller the disparity ...

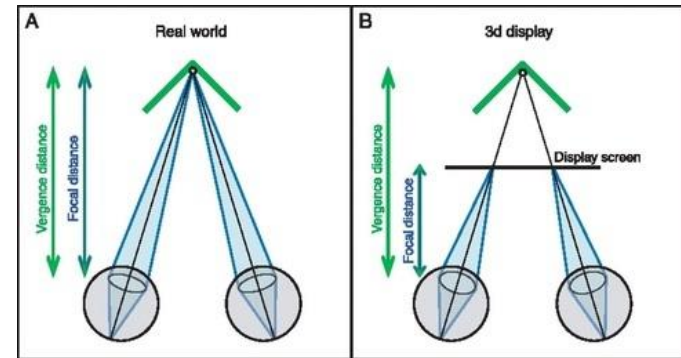
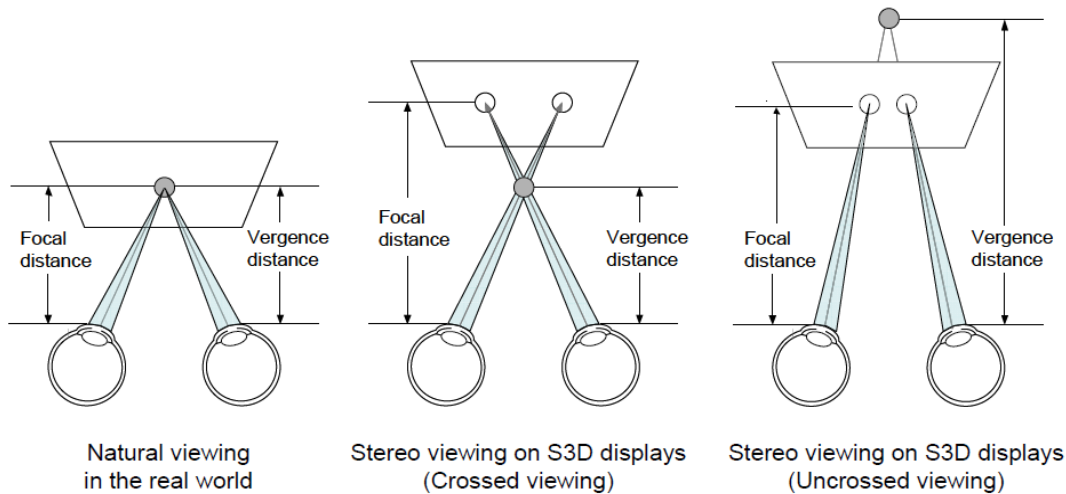


Stereoscopic Vision

- **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.**



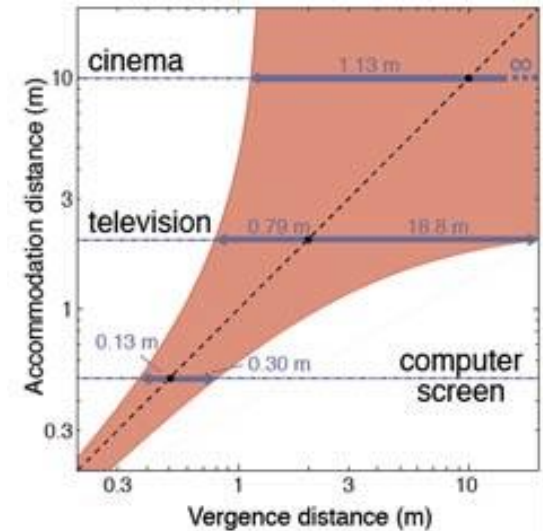
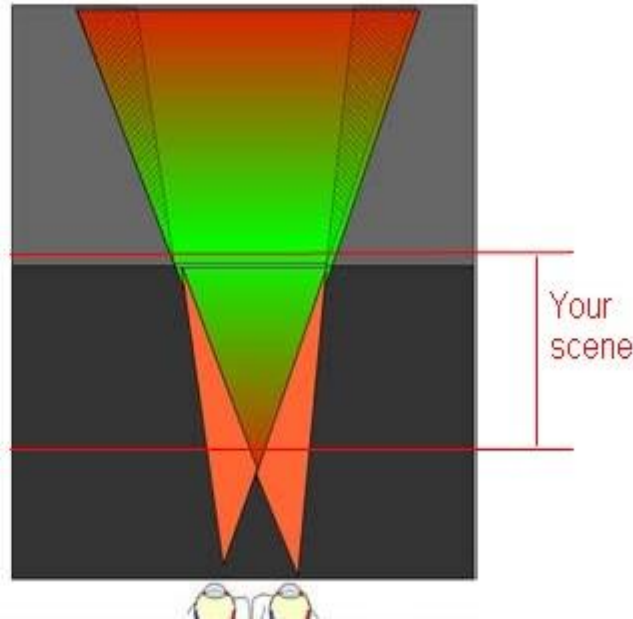
Accommodation-Vergence Conflict



- **In natural viewing, the vergence stimulus and focal stimulus are always at the same distance and, therefore, are consistent with one another.**
- **Stereo displaying creates (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 to have the sharpest image).**
- **The accommodation-vergence conflicts lead to problems, notably 3D structure distortions and visual fatigue.**

Depth Perception: the Comfort Zone

- Gray: Invisible to the audience
- Red: **Danger Zones**
 - Strong muscular activity
 - Convergence vs Accommodation
 - Do not stay too long
- Orange: **No Parking**
 - Retinal Rivalry Area
 - Move in, out and fast
- Green: **Rest Areas**
 - Close to the screen plane
 - Stripped: natural retinal rivalry zones



- **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.**

Camera Baseline



- **The *camera baseline or base* is the distance between the 2 stereoscopic lenses. This distance has a profound effect on stereo content.**
- **For most images, the baseline is close to the distance between the human eyes, which is around 65 mm.**
- **However, it is possible to use a shorter baseline for closeup photography or a longer baseline when shooting distant subjects such as the moon or mountains. This is critical to ‘put’ the real world in the comfort zone.**
- **Especially for 3D home entertainment, newer stereoscopic displays can vary the baseline between the views to adapt to different viewing distances.**

Depth Cues: Monocular and Binocular

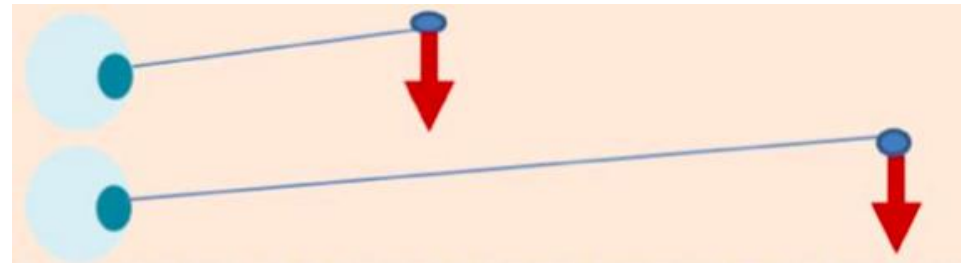
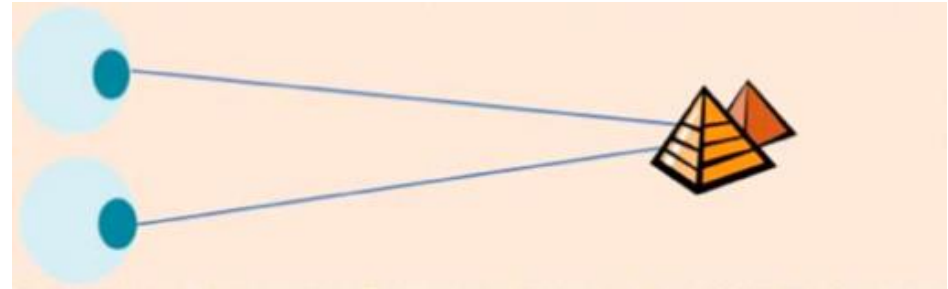
- Most of the depth cues used by humans to visualize the world's 3D structure are available in 2D projections; this is why images make sense on a (mono) TV/cinema screen.
- The depth cues can be classified into oculomotor cues coming from the eye muscles, and visual cues from the scene content itself. They can also be classified into monocular and binocular cues.
- Monocular cues for 3D perception include:
 - **Occlusion** - one object partially covering another
 - **Perspective** - point of view
 - **Familiar size** - we know the real-world sizes of many objects
 - **Atmospheric haze** - objects further away look more washed out
 - **Selective focus or Accommodation of the eyeball (eyeball focus)** - the eye changes optical power to maintain a clear image (focus) on an object as its distance changes.
 - ...



Main Binocular Depth Cues

Some main cues are missing from 2D media:

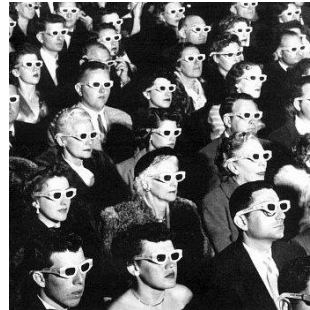
- **Stereo parallax** - seeing a different image with each eye, thus different aspects of the same object
- **Motion parallax** - seeing different perspective images when we move our heads; nearby objects appear to move faster across the view
- **Vergence** - muscular rotation of the eye balls, which is used to converge both eyes on the same object



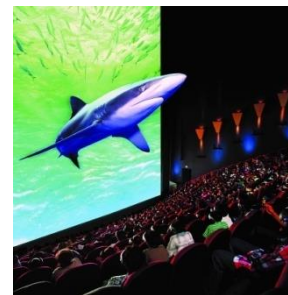
3D Video: Stereo



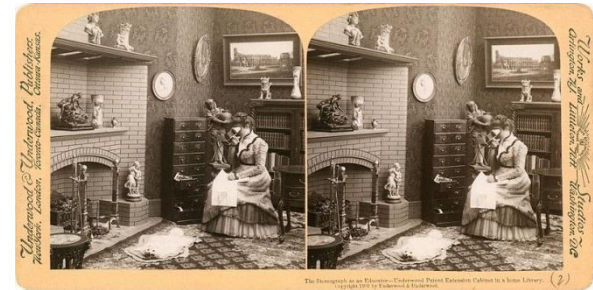
History of Stereo Video ...



<i>Wheatstone explains "stereopsis"</i>	<i>First stereo film camera</i>	<i>Boom year for 3D movies</i>	<i>First compression standard: MPEG-2 develops Multi-View Profile</i>	<i>MPEG-4 Multi-View Coding</i>				
1838	1851	1890	1915	1953				
1838	1851	1890	1915	1953	1990s	1995	2005-2009	2009
<i>Queen Victoria starts stereoscope rage</i>		<i>First red/blue 3D movies shown</i>		<i>3D starts to gain popularity with IMAX 3D</i>		<i>Expansion of 3D movies</i>		



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 is Still Mostly Stereo ...

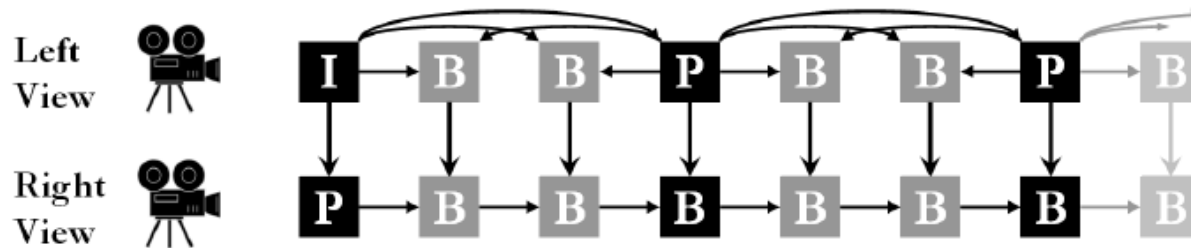


Stereo Cameras ...

- 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*. Stereo cameras may be used for making stereo views and 3D movies.
- The distance between the lenses 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.



Conventional Stereo Coding Format

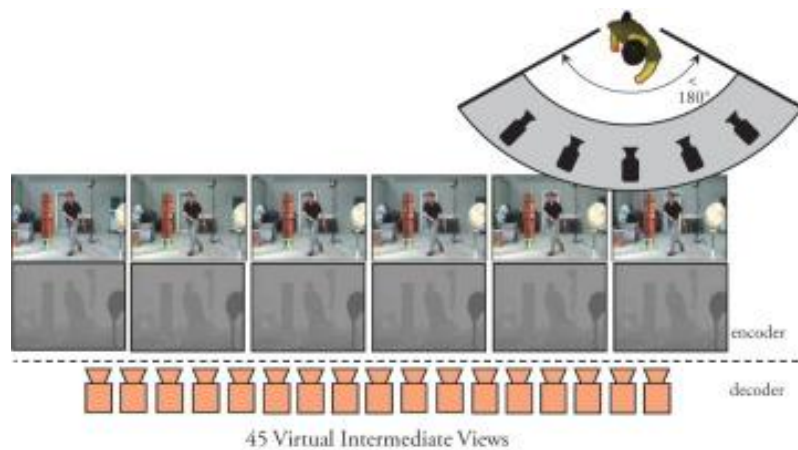


Blu-ray
3D

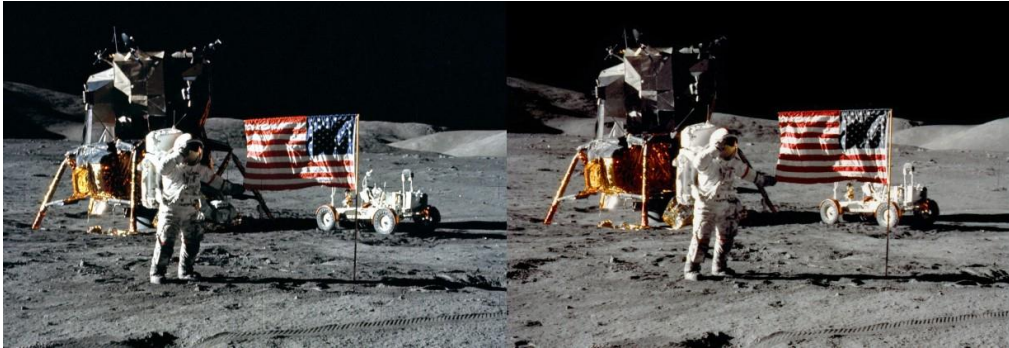
Combined temporal and interview prediction

- **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.**

3D Video: Multiview



Stereo and Multiview Video Data

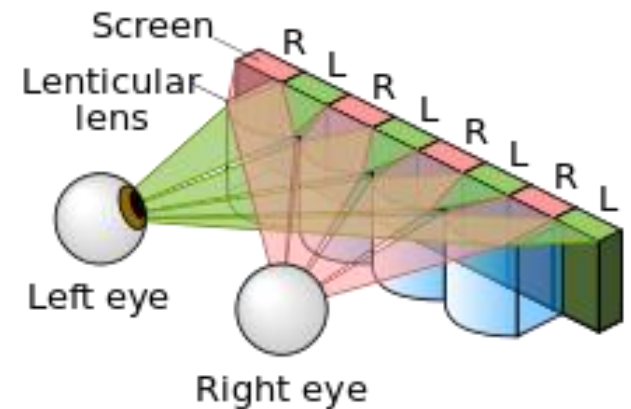
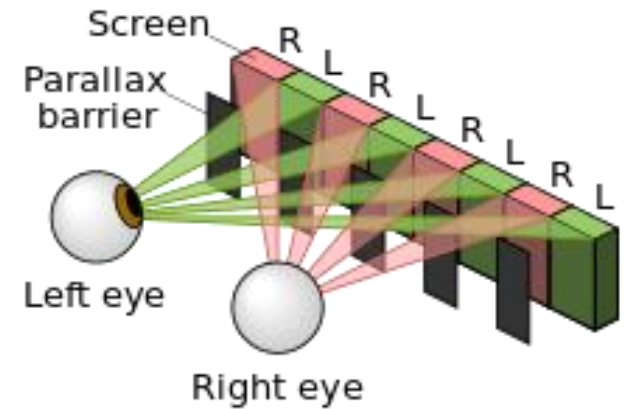


- **Redundancy reduction between camera views**
 - Need to cope with color/illumination mismatch problems
 - Alignment may not always be perfect either



Autostereoscopic Displays ...

- **Autostereoscopy is any method of displaying stereoscopic images (adding binocular perception of 3D depth) without the use of special headgear or glasses by the viewer.**
- **There are two main approaches to accommodate motion parallax and wider viewing angles:**
 - **eye-tracking**
 - **multiple views so that the display does not need to sense where the viewers' eyes are located**



Sensing More with Depth ...



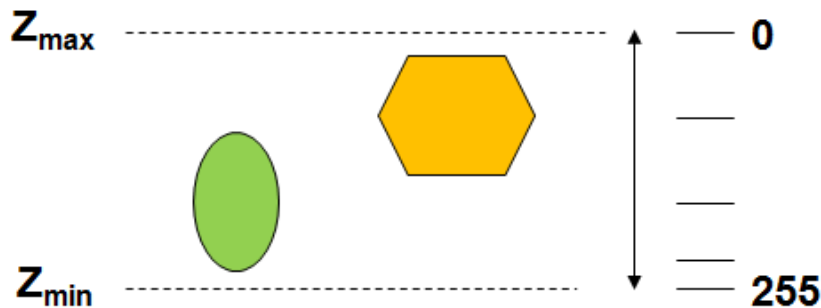
- A depth map is a ‘gray image’ containing information with the distance from the scene objects to the camera.
- Depth maps may be obtained by:
 - Special range cameras
 - Extraction from texture
 - Inherent to the content, e.g. computer-generated imagery
- *Depth maps provide important information about the scene geometry.*

Representing Depth ...

Store inverse depth

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

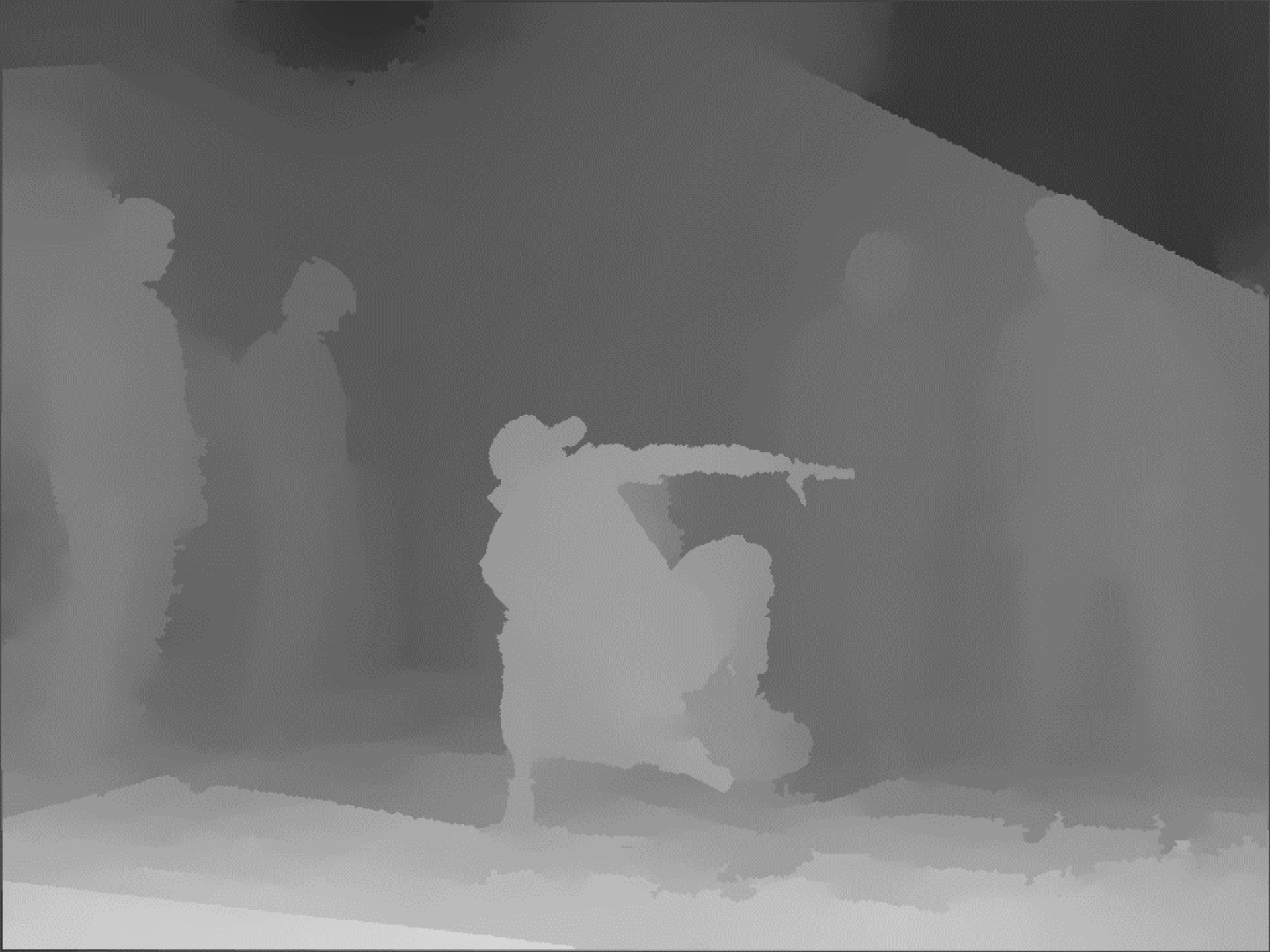
where z_{\min} and z_{\max} are the minimum and maximum depth of the scene, respectively



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 only based

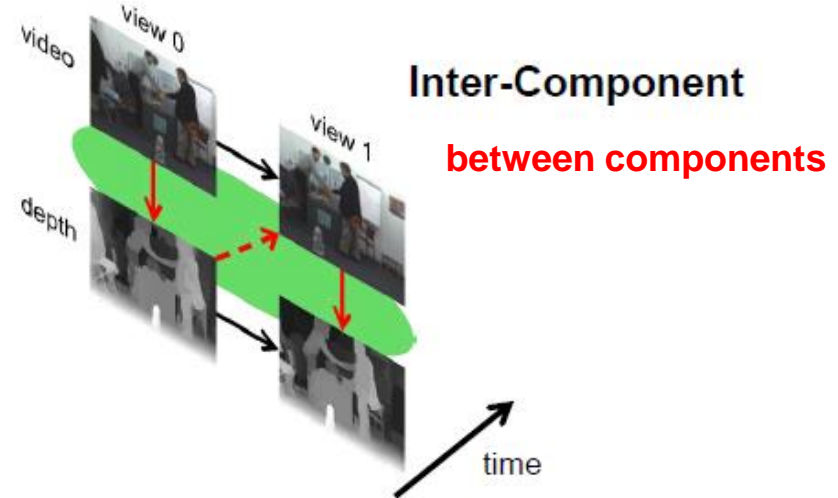
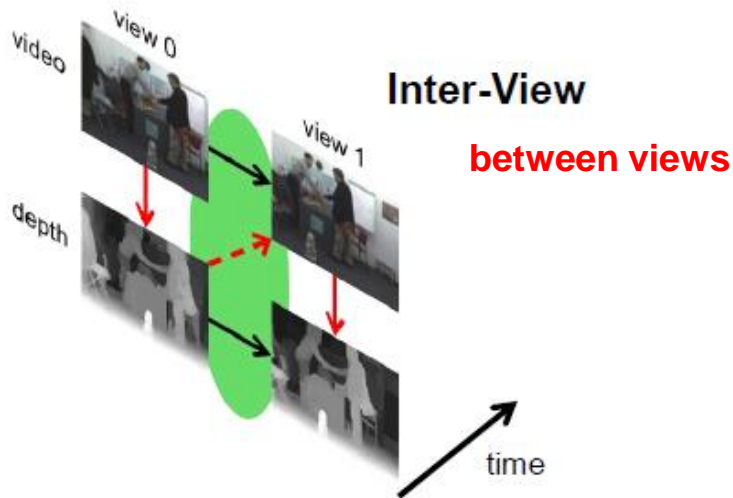
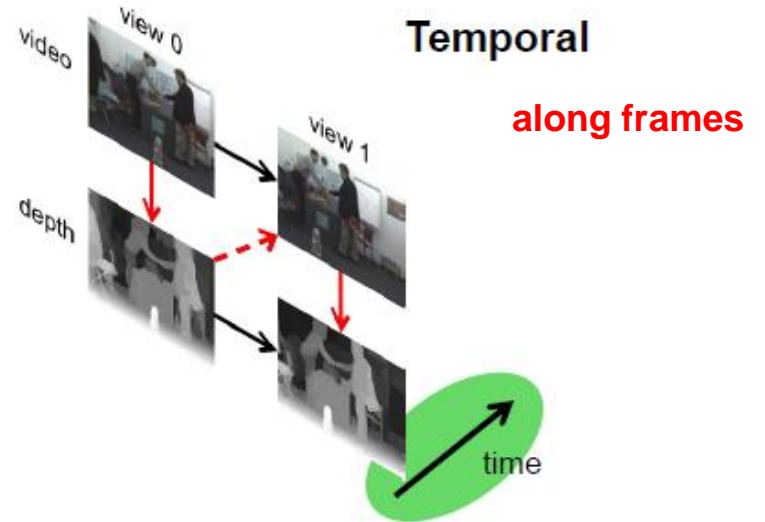
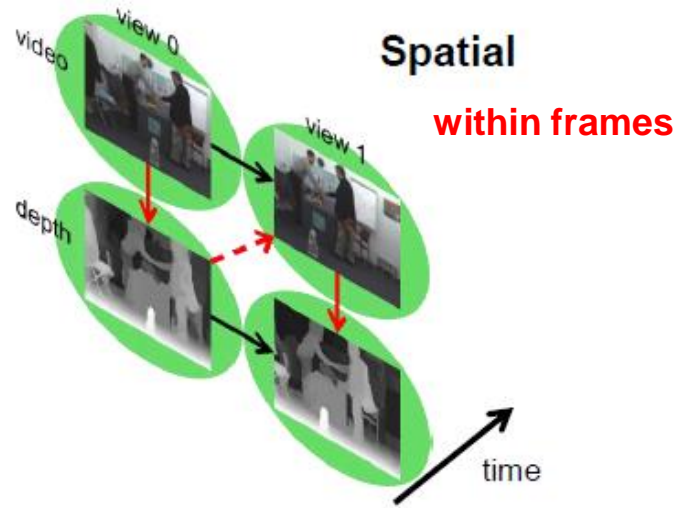
- Multiview Simulcasting
- Frame Compatible Stereo
- Conventional Stereo Video
- Multiview Video, MVC and MV-HEVC standards

Texture plus Depth based

- 2D (Texture)+Depth, MPEG-C standard
- Multiview+Depth (MVD), 3D-HEVC standard



Redundancies in 3D Video



The Texture Only Approach

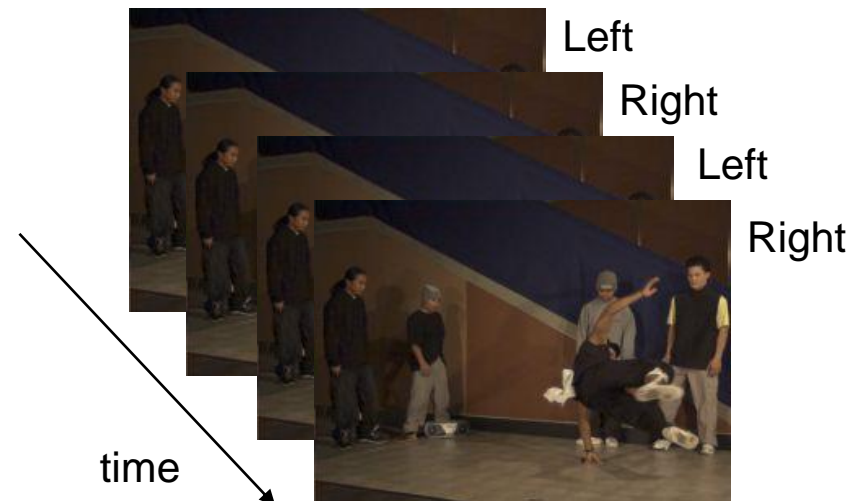
Multiview Simulcasting



- **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 H.264/AVC is more likely.**
- **This solution has been largely used in many countries, e. g. to broadcast the 2010 Football World Cup games.**

Frame Compatible Stereo Formats Examples

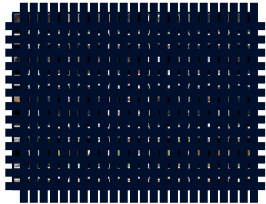
- **Basic concept: pack pixels from left and right views into a single frame to be coded ‘as usual’:**
 - **Spatial Multiplexing: side-by-side, top-bottom, checkerboard formats**
 - **Time Multiplexing: views interleaved as alternating frames or fields**
- **In such a spatial format, half of the coded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



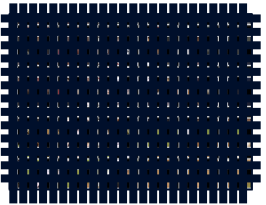
Frame Compatible Stereo: Spatial Multiplexing

Reduced picture resolution !

Left Eye



Right Eye



Top &
Bottom

Side-by-Side



Line
Interleave

Column
Interleave

Checkerboard

Left Eye



Right Eye





Frame Compatible Stereo: Temporal Multiplexing

Left Eye

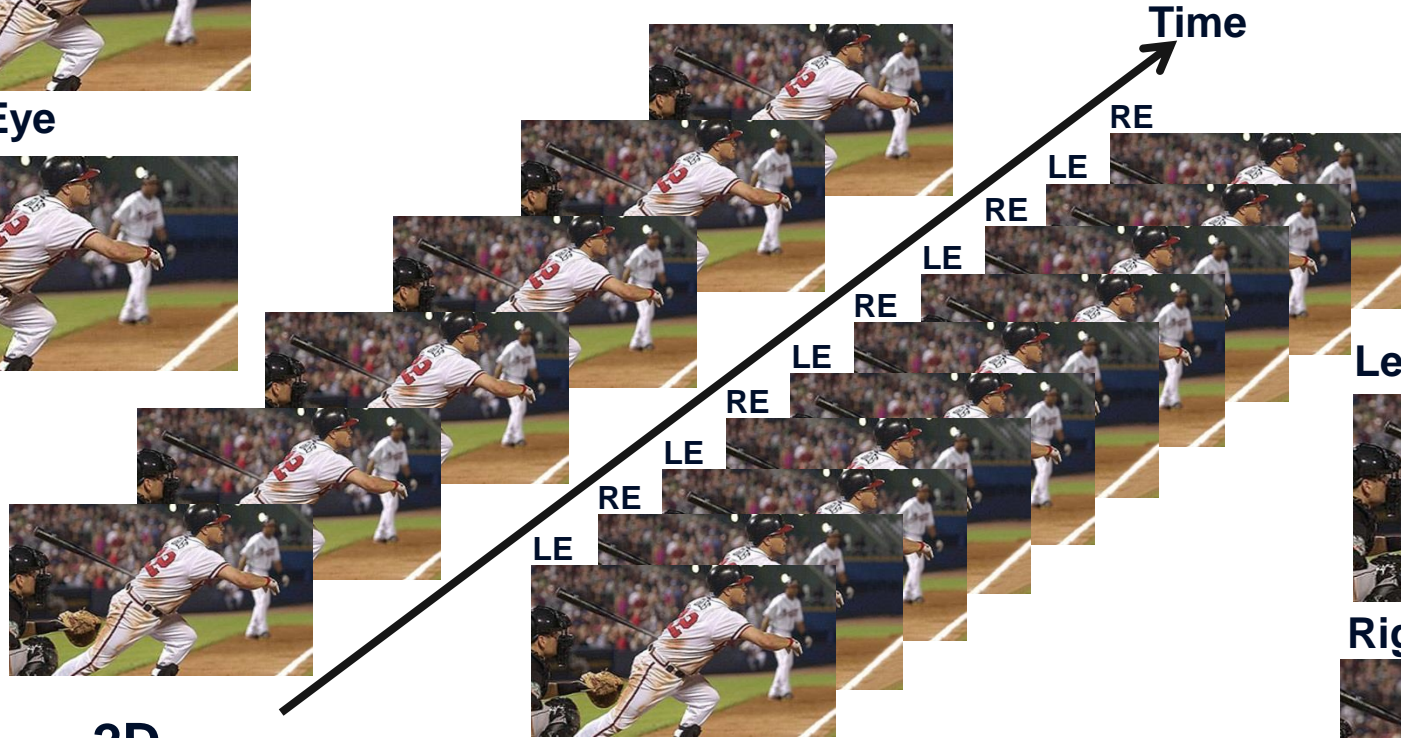


Right Eye



Provides full resolution quality but requires increased bandwidth and storage!

Time



Left Eye



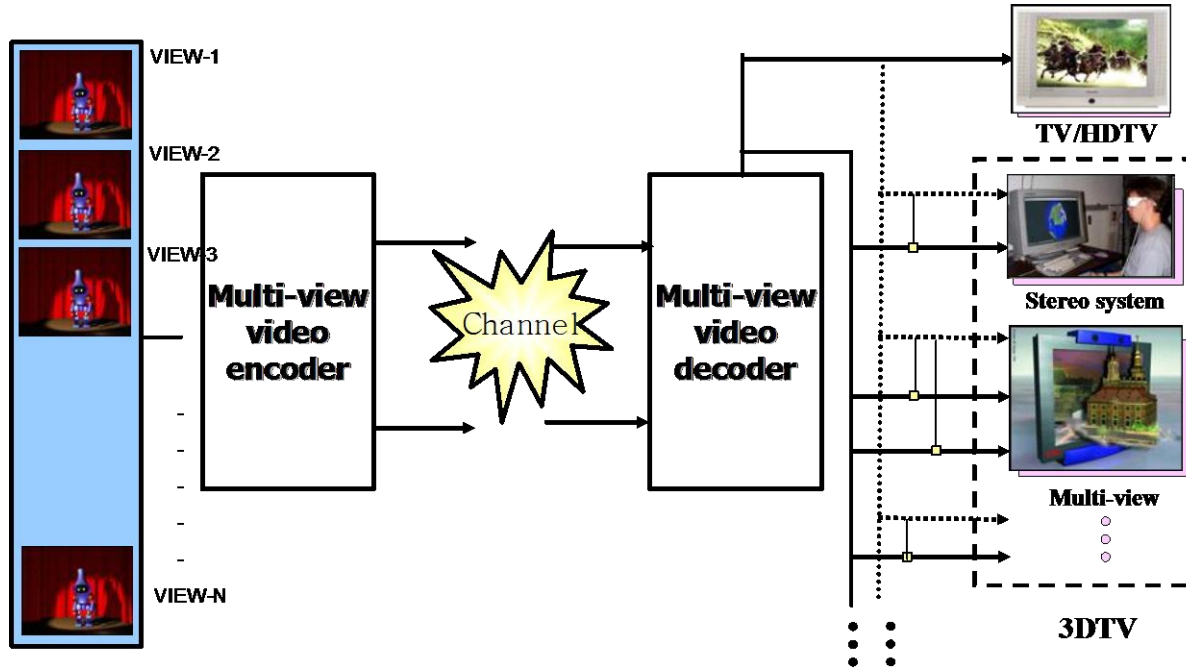
Right Eye



2D

3D Frame Sequential

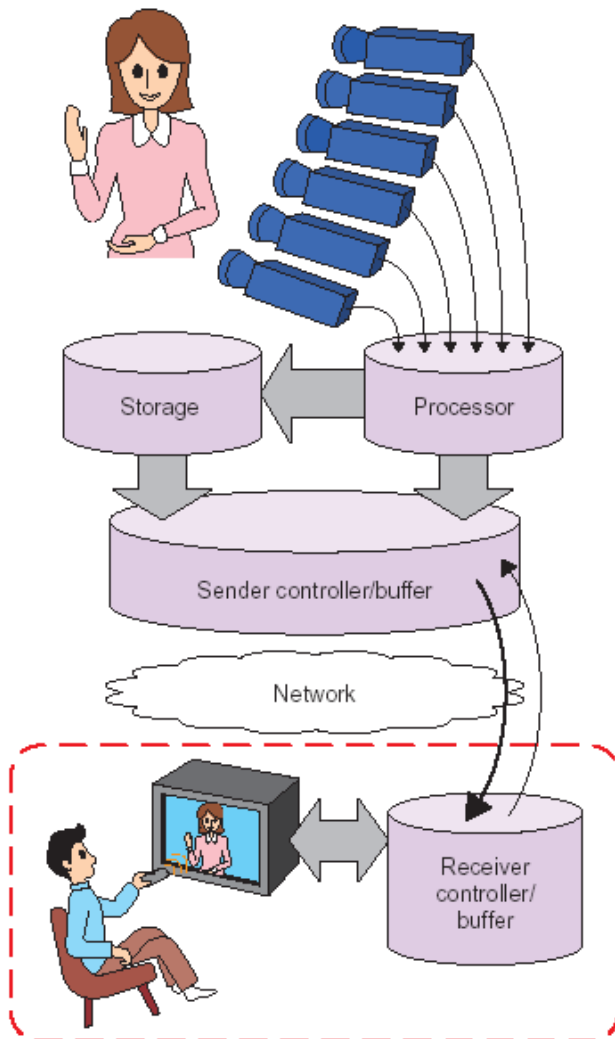
Multiview Video Coding Format



Multiview video (MVV) refers to a set of N temporally synchronized video streams coming from cameras capturing the same real scenery from different viewpoints.

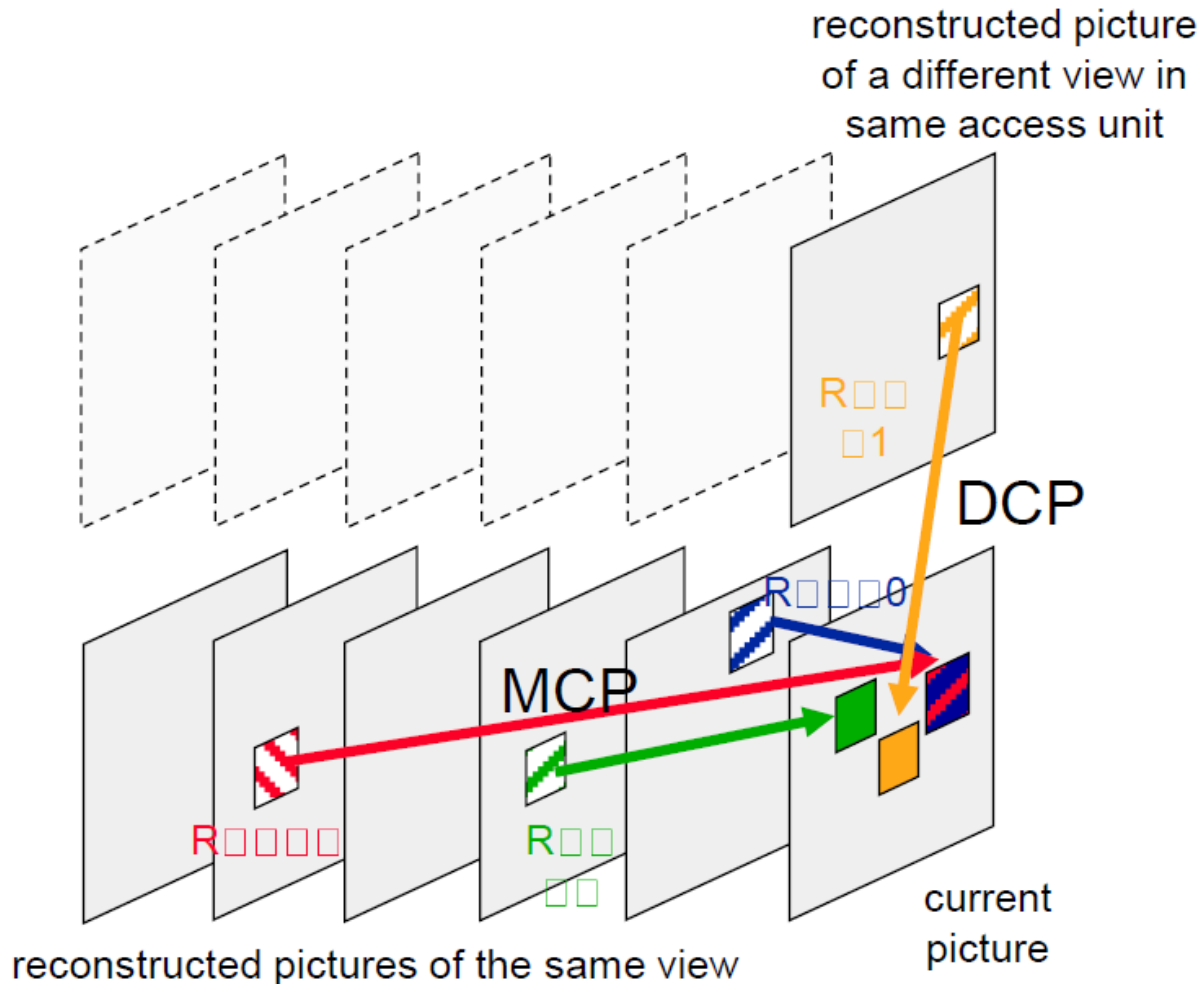
- Provides the ability to change viewpoint freely with multiple views available
- Renders one view (real or virtual) to legacy 2D display
- Most important case is stereo video ($N = 2$), generating a depth impression with each view derived for projection into one eye

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 H.264/AVC compliant 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%

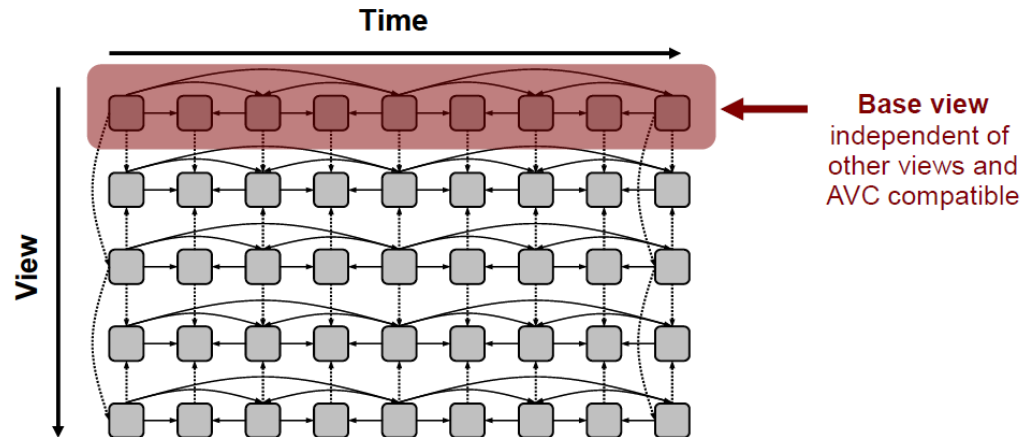
Disparity-Compensated Prediction



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

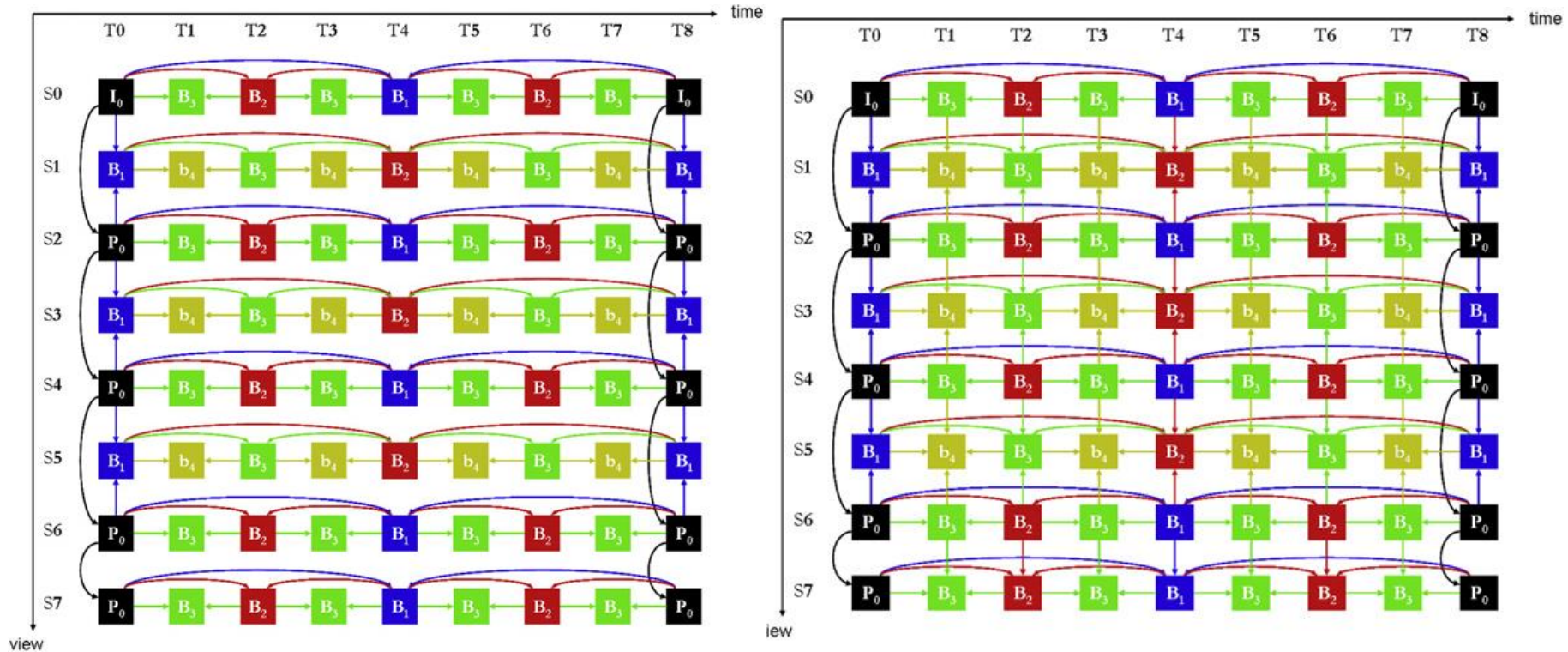
Interview Prediction: Basics

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



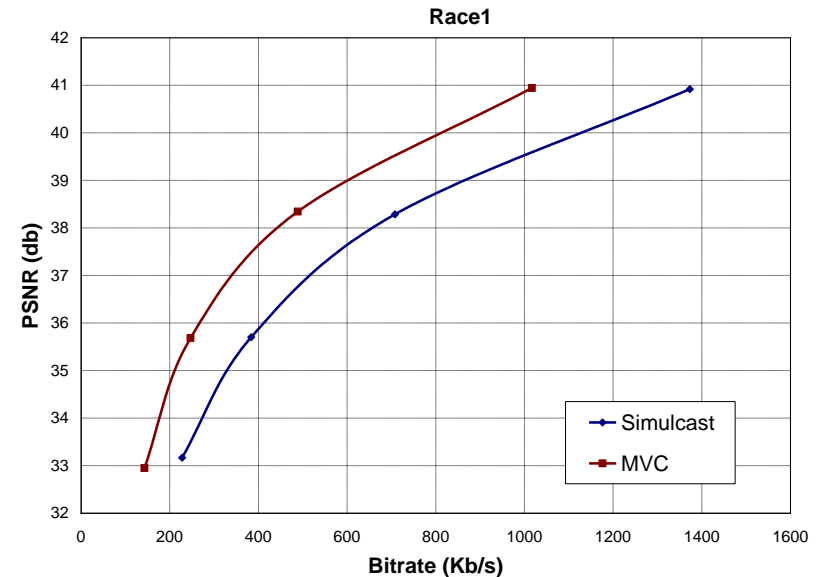
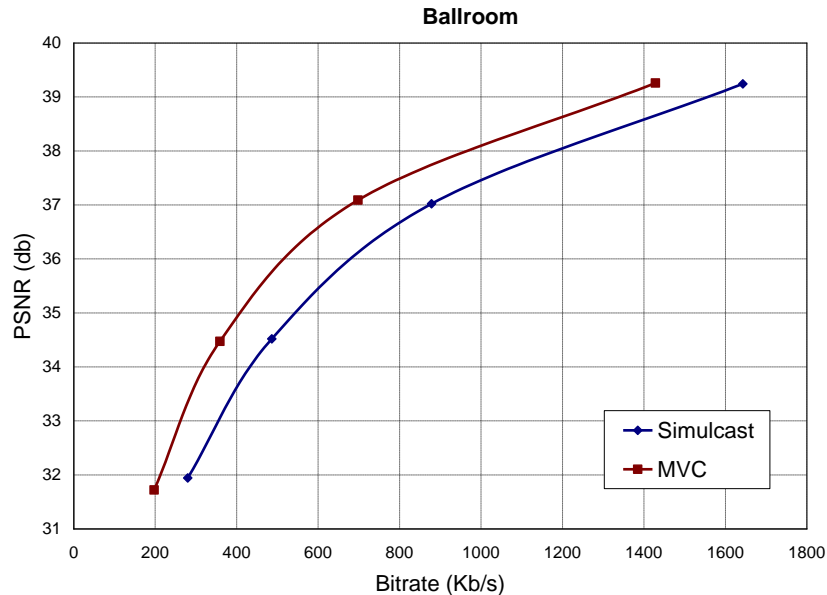
- Pictures in the non-base views are not only predicted from temporal references (*in the same view*), but also from interview references (*in the other views*).
- Limitations: i) inter-view prediction only from same time instance; ii) cannot exceed maximum number of stored reference pictures.
- The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

MVC Prediction 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

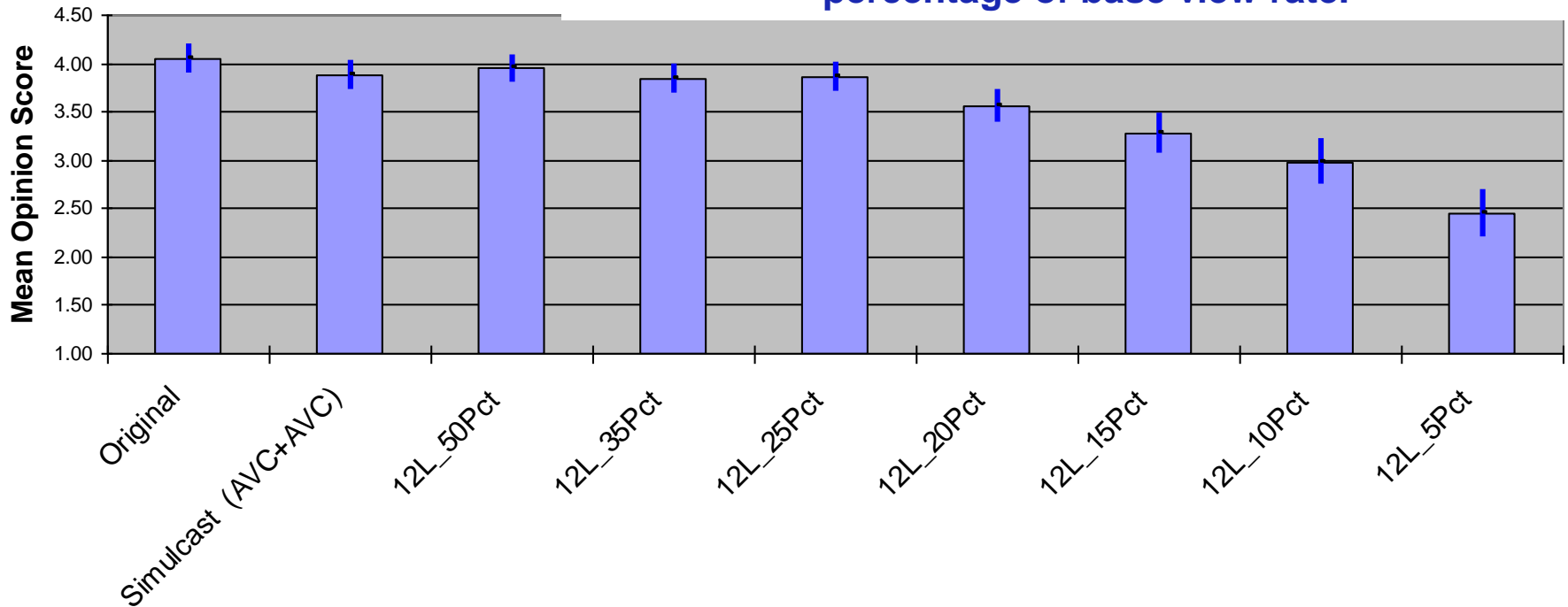
**8 views (with 640×480 resolution), and
considering the rate for all views**

**~25% bit rate savings over all views for
same PSNR**



MVC: Subjective Stereo Performance

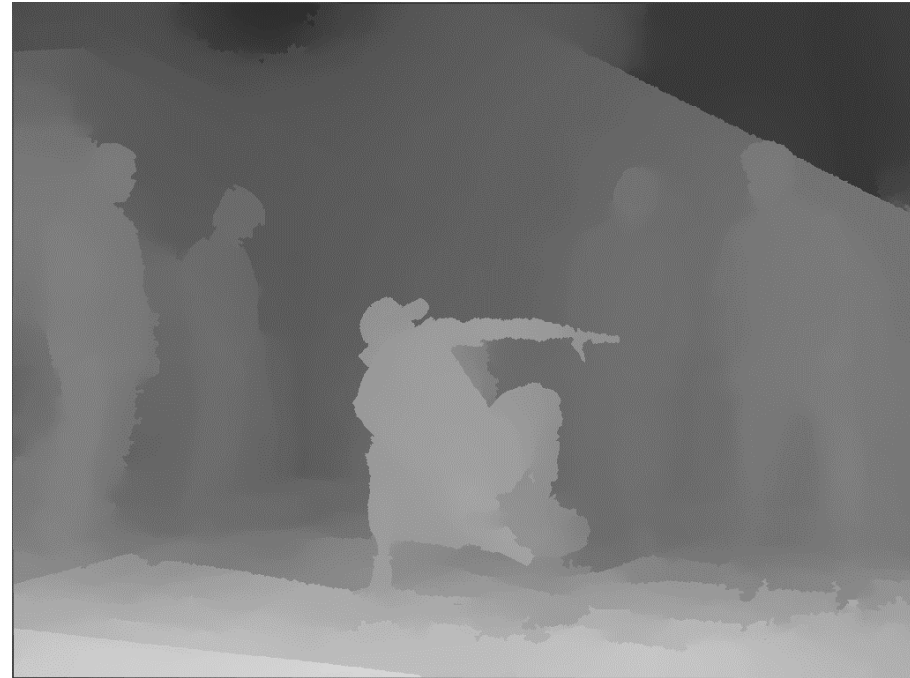
Base view fixed at 12 Mbit/s; dependent view at varying percentage of base view rate.



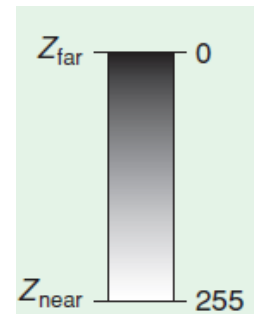
- MVC achieves comparable perceptual quality to simulcasting with as little as 25% rate for the dependent view (75% gain); this rate may have to be higher for lower rates than 12 Mbit/s for the main view.
- For similar PSNR, the gains are only about 30% for the dependent view.
- This experiment shows that the 2 views don't need to have the same PSNR quality.

The Texture+Depth Approach

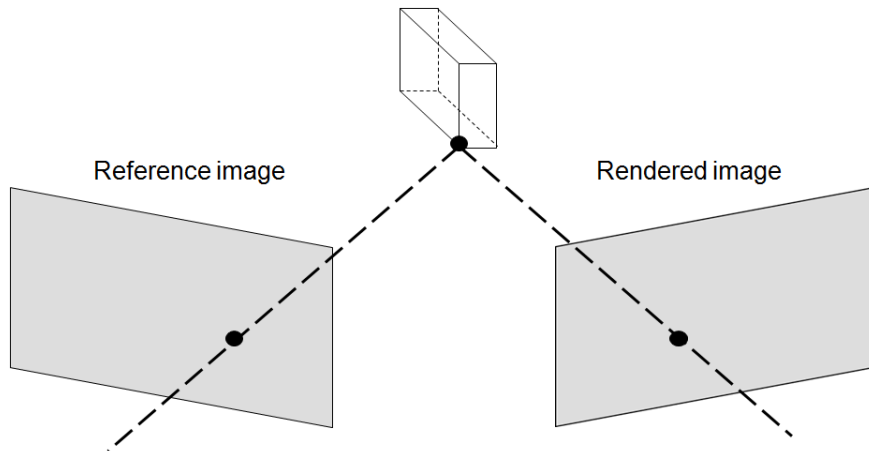
Texture and Depth ...



Depth-enhanced formats are suitable for generic 3D video solutions, where only one format is coded and transmitted while all necessary views for any 3D display are generated from the decoded data, e.g., by means of depth image based rendering (DIBR).

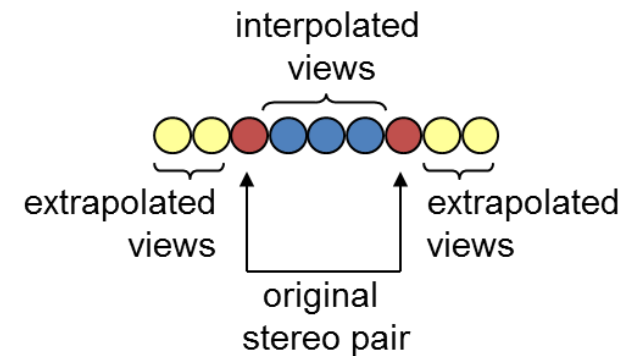
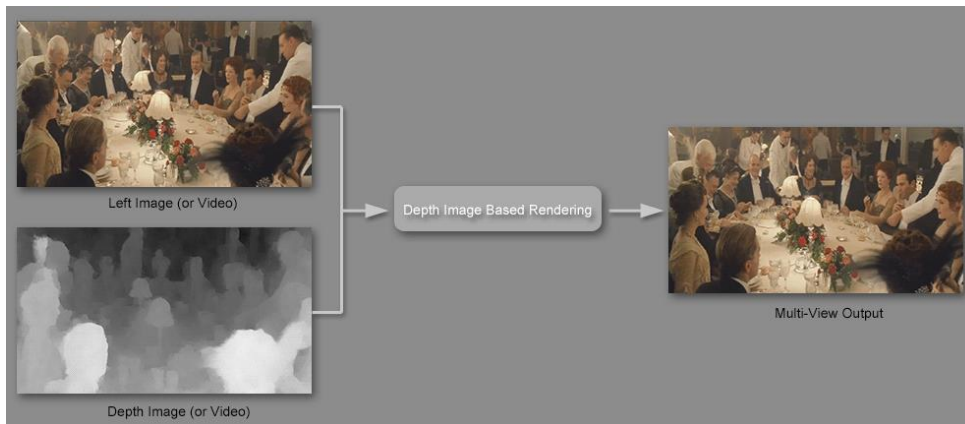


Depth-Image-Based Rendering (DIBR)



3D warping of pixel in reference image to rendered image based on geometry of the scene

- In general case, 3D warping is done using projective matrices and depth info
- When cameras are rectified, 3D warping amounts to a simple 1D shift
- Views may be either extrapolated or interpolated



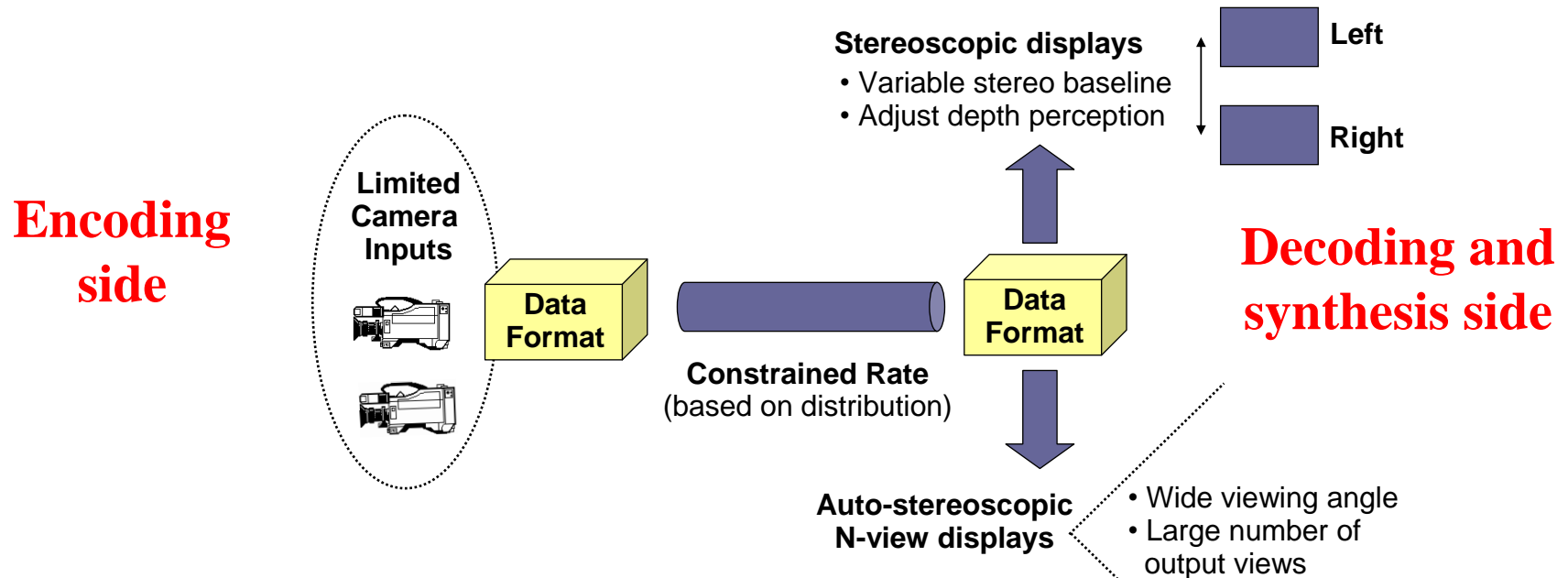
Multiview Video plus Depth (MVD)



- **The MVD format 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 independently encoded with MVC.**

Combining Coding with Synthesis

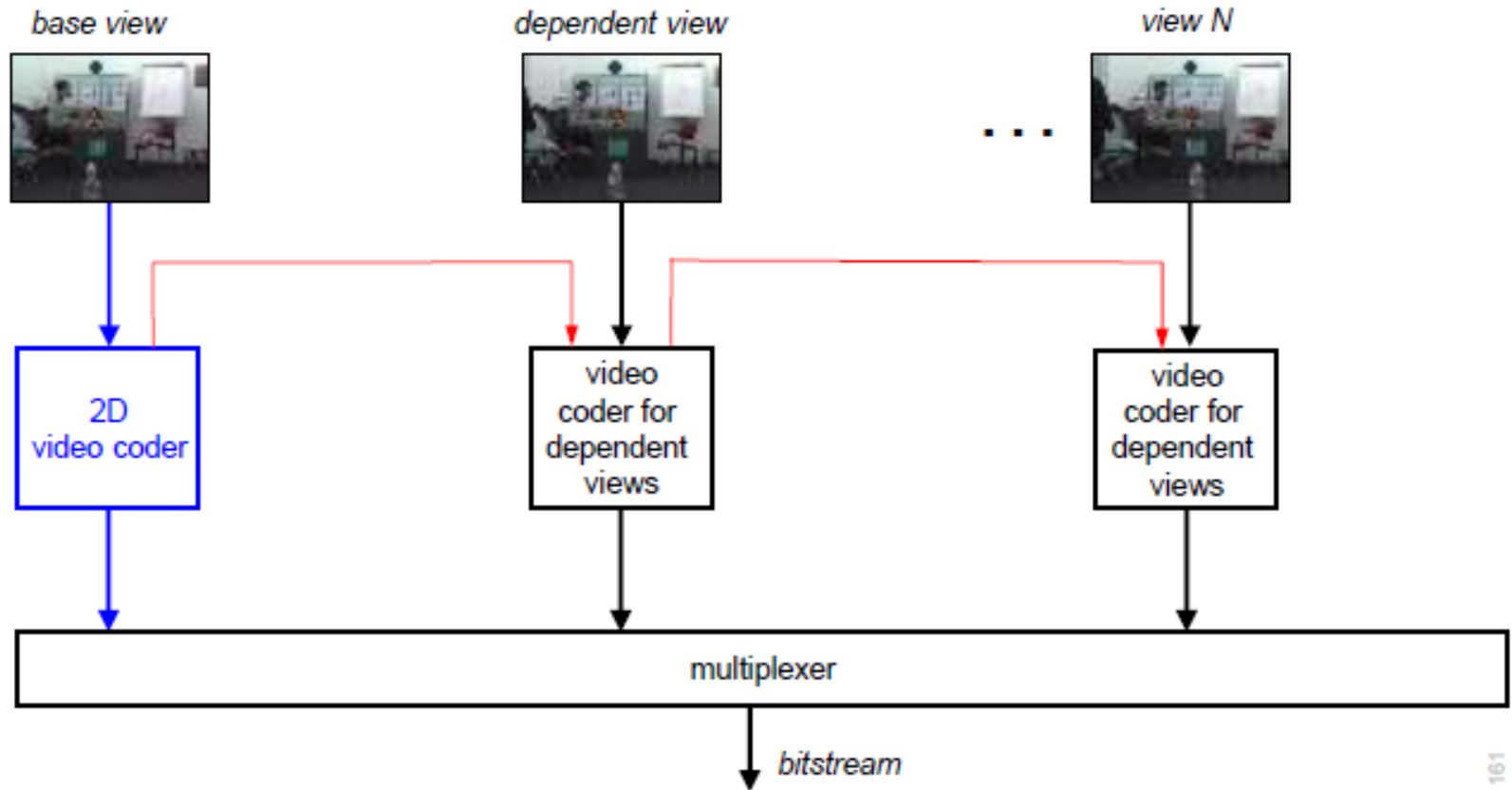
- As the transmission rate is limited, only a small number of texture and depth views may be coded.
- However, an arbitrarily large number of views may need to be rendered.
- Using depth-image-based rendering (DIBR) techniques, a continuum of views may be synthesized based on the limited set of decoded views.



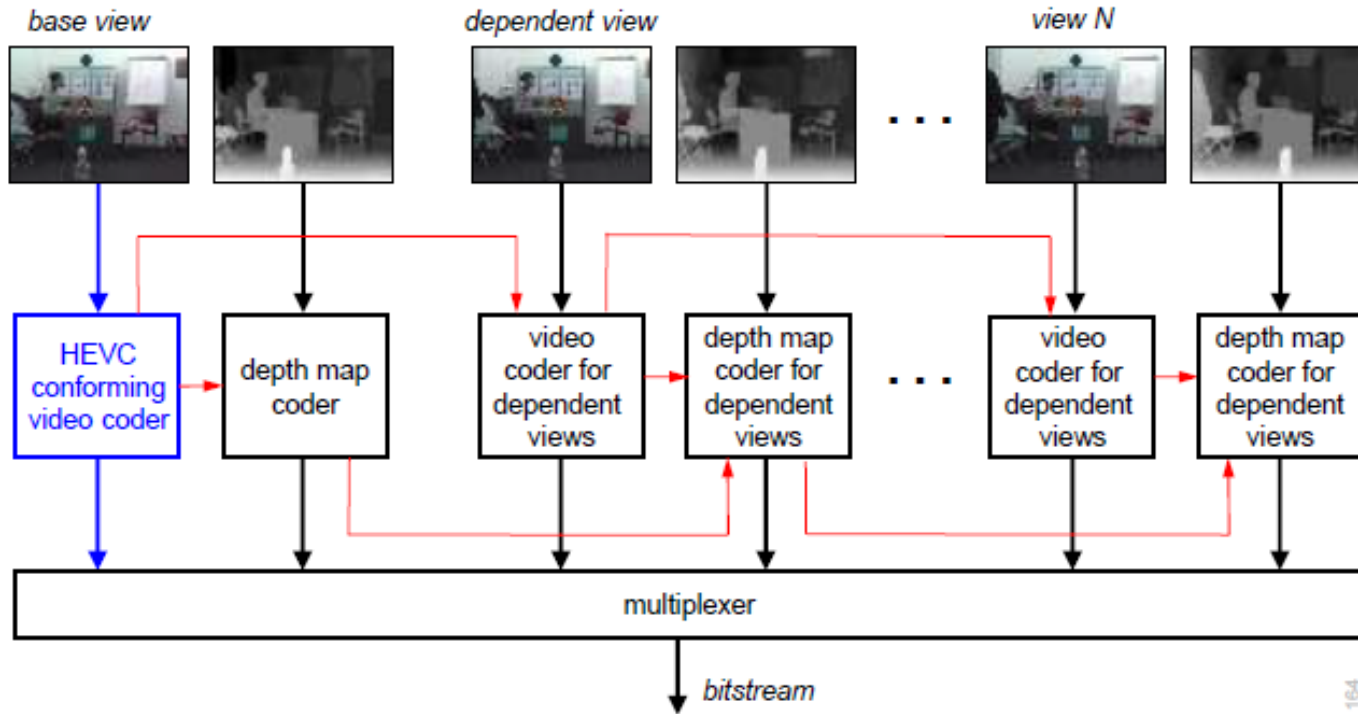
3D Related Video Coding Standard Extensions

- **MVC (Multiview Video Coding) - Simple stereo/multiview extension of the monoview H.264/AVC standard**
- **MV-HEVC - Simple stereo/multiview extension of the monoview HEVC standard**
- **3D-HEVC - More efficient extension of the monoview HEVC standard also considering 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



3D-HEVC Approach



Format scalable approach as sub-bitstreams representing a subset of the video views (with or without associated depth data) can be extracted by discarding NAL units from the 3D bitstream and be independently decoded.

K. Muller, Fraunhofer HHI

Why Didn't 3D Stereo Fly ?

It simply did not deliver the Quality of Experience that users expected ...

Immersion is really poor ... Only stereo parallax ... Glasses are cumbersome ...





Let's move forward

Wait ! Glasses, Again ?



Virtual Reality vs Augmented Reality

- **Virtual Reality (VR)** - Replicates an environment and simulates a physical presence in places in the real world or an imagined world, allowing the user to interact in that world.
- **Augmented Reality (AR)** - Live, direct or indirect view of a physical, real-world environment whose elements are augmented (or supplemented) by computer-generated sensory input such as sound, video, graphics or GPS data.







360° or Omnidirectional or Spherical Video ...

- **360-degree videos, also known as immersive videos or spherical videos, are video recordings where a view in every direction is recorded at the same time, shot using an omnidirectional camera or a collection of cameras (likely with stitching).**
- **During playback, the viewer has control of the viewing direction like a panorama.**
- **Spherical media enables a range of immersive viewing experiences and is currently an essential VR building block.**



360 Video Cameras ... For All Tastes ...



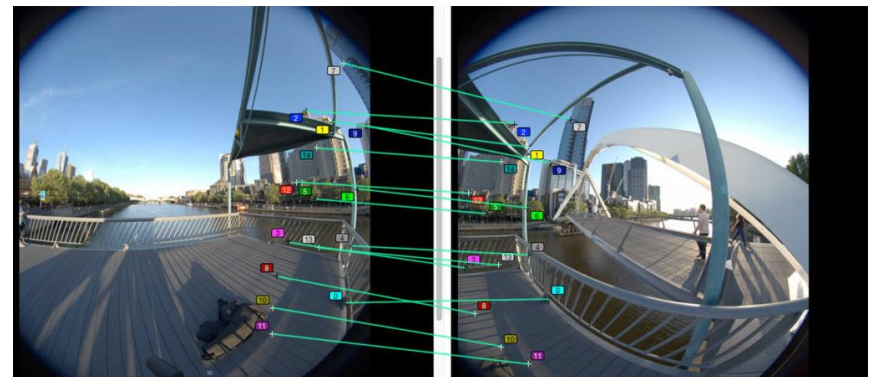
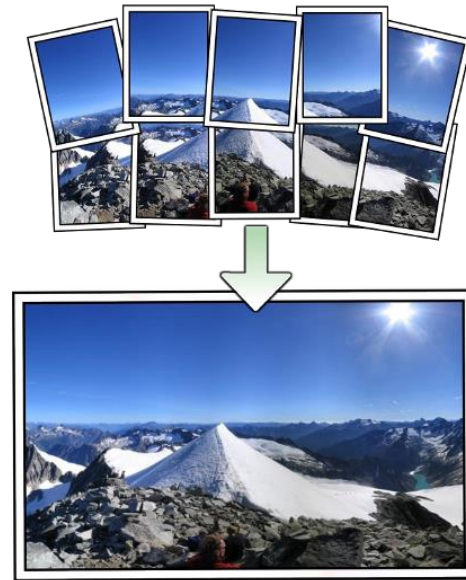
Acquisition: Facebook Surround 360



A production-ready, high-quality 3D-360 camera design with accompanying automated stitching technology that seamlessly marries the video from 17 cameras, reducing post-production effort and time.

- 14 wide angle cameras on a horizontal ring
- 1 fish eye lens on top and 2 on bottom for complete spherical coverage
- Global shutter ensures that each camera captures the pixels in sync

- **Image stitching or photo stitching is the process of combining multiple photographic images with overlapping fields of view to produce a segmented panorama or high-resolution image.**
- **Most approaches to image stitching require nearly exact overlaps between images and identical exposures to produce seamless results.**
- **Some digital cameras can stitch their photos internally.**

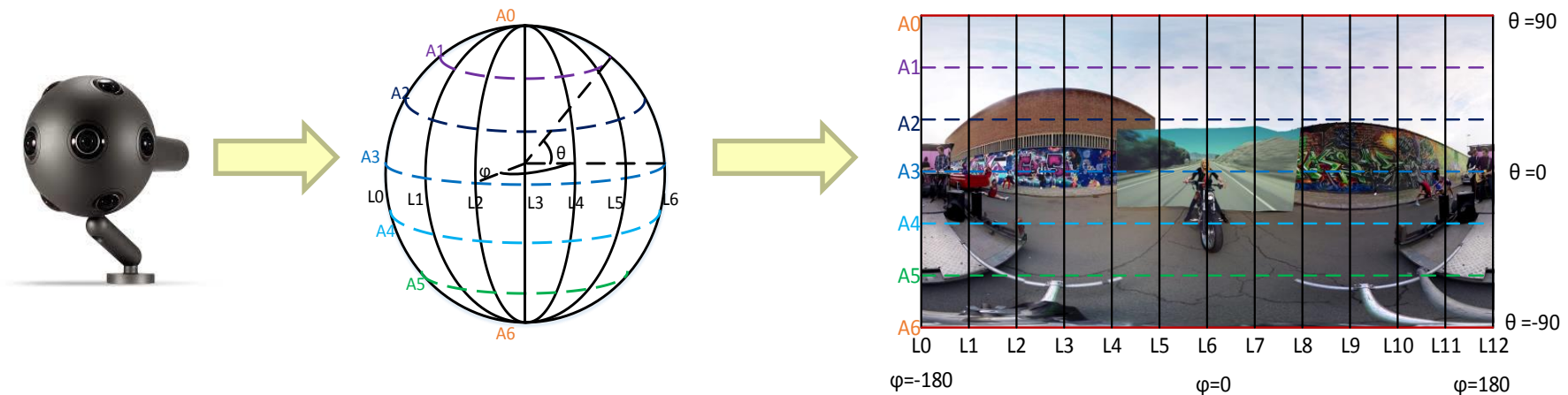




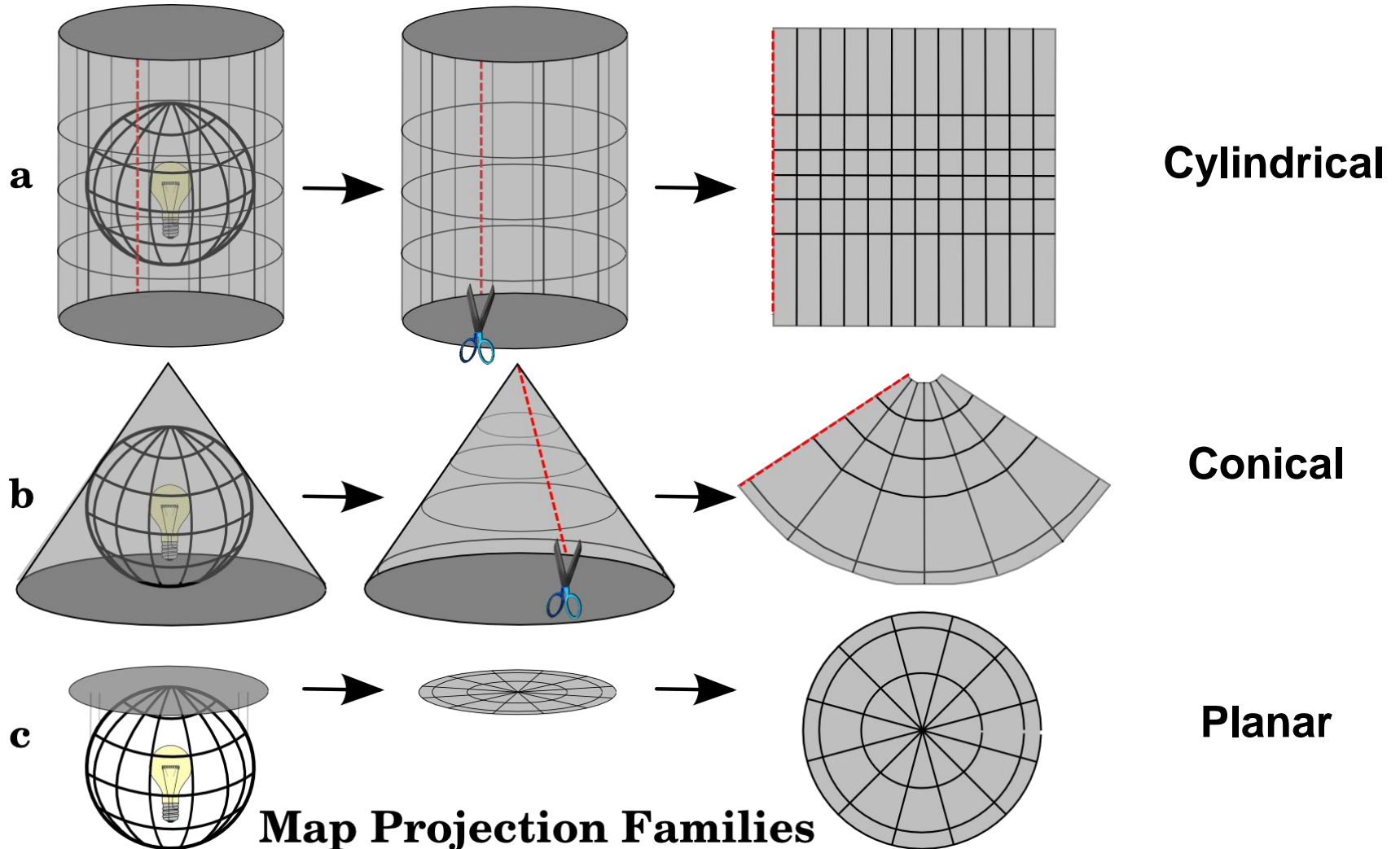
Projection



- A map projection is a systematic transformation of the latitudes and longitudes of locations from the surface of a sphere into locations on a plane.
- Maps cannot be created without map projections. All map projections necessarily distort the surface in some specific way.
- After the projection, which is basically a transformation of the stitched spherical image, the image is not spherical anymore and some areas may have been significantly stretched, thus creating distorted zones.



Map Projection Families



Map Projection Families

Before stitching



After stitching and projection



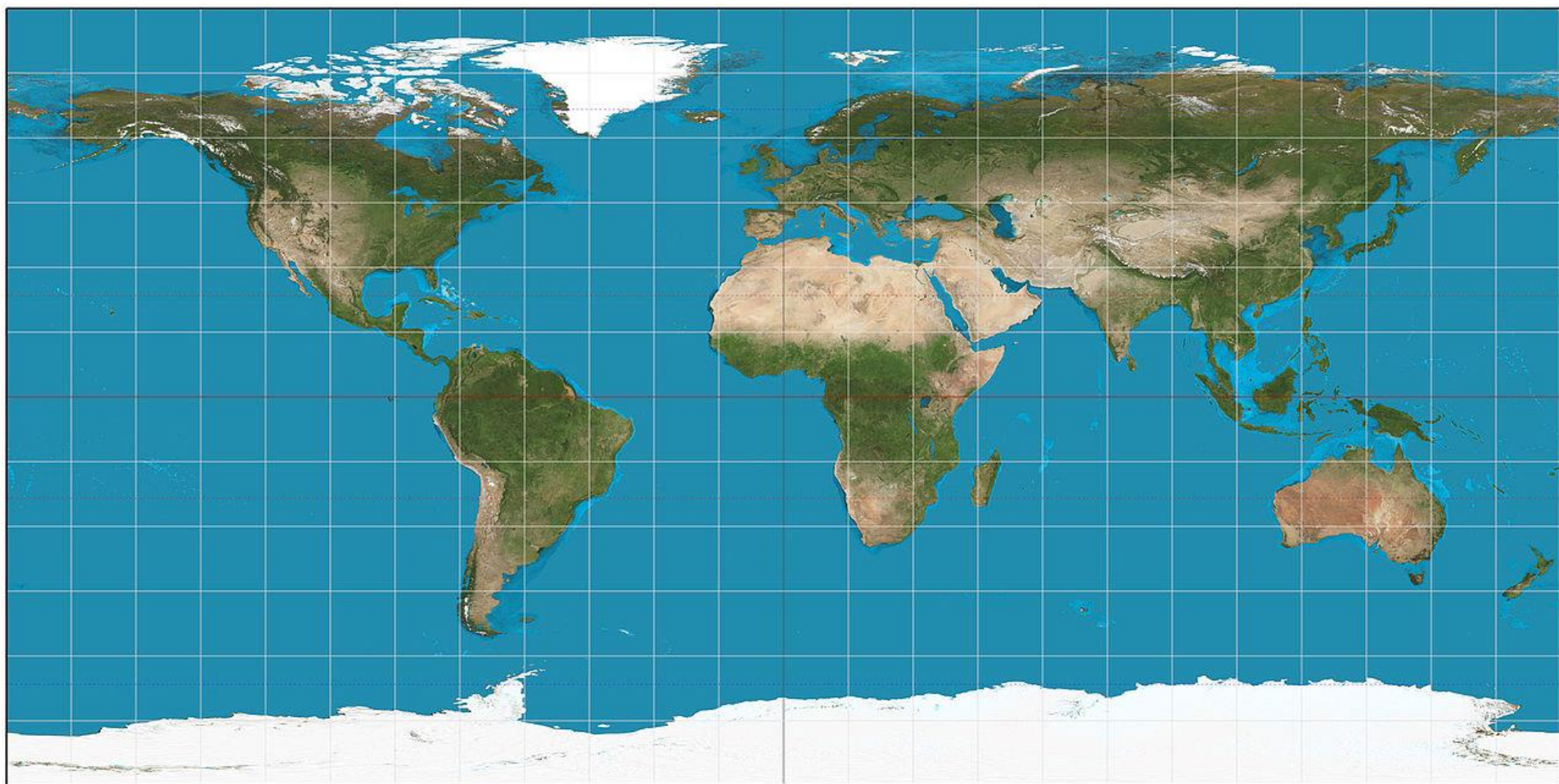
Equirectangular Projection



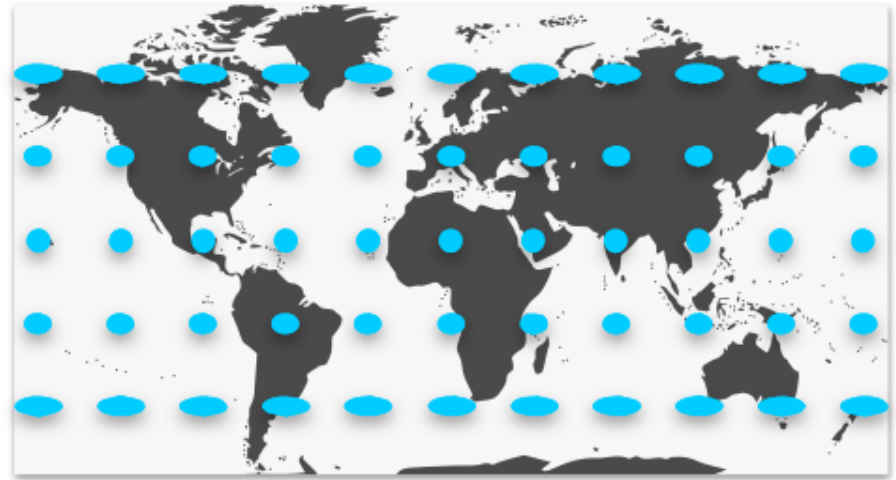
- The *equirectangular projection* (also called the equidistant cylindrical projection) maps meridians to vertical straight lines of constant spacing (for meridional intervals of constant spacing), and circles of latitude to horizontal straight lines of constant spacing (for constant intervals of parallels).
- The main advantages is that it has a particularly simple relationship between the position on the map and its corresponding position on the sphere, and it is a friendly format to preview video containing the whole Field Of View.
- The main drawback is that the associated distortion is considerable, notably for the polar areas; this affects the performance of standard coding methods, especially due to the larger number of redundant pixels representing small areas near the poles.



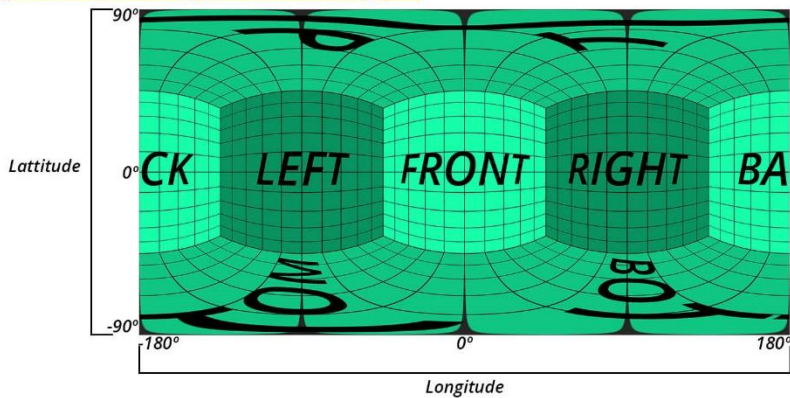
Equiarectangular Projection Example ...



Equirectangular Projection Deformations



Equirectangular image

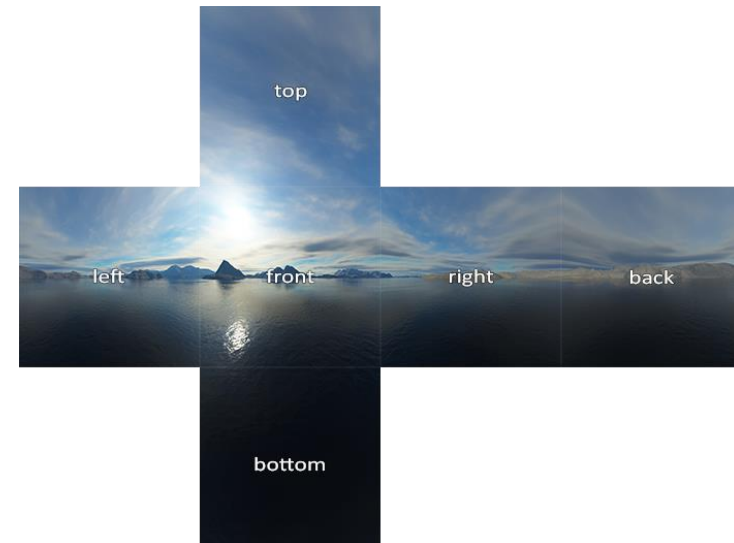
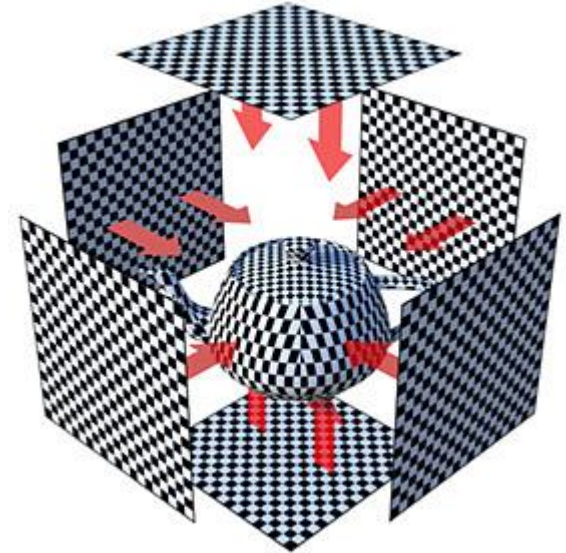




Cubic Projection



- **Cube mapping is a method of environment mapping that uses the six faces of a cube as the map shape.**
- **The cube map is generated by first rendering the scene six times from a viewpoint, with the views defined by a 90 degree view frustum representing each cube face.**
- **The faces may have different spatial resolutions.**
- **In the majority of cases, cube mapping is preferred over the older method of sphere mapping because it eliminates many of the problems that are inherent in sphere mapping such as image distortion, viewpoint dependency, and computational inefficiency.**



Cubic



**Cubic re-organized to
obtain a rectangular layout**



Equirectangular

Coding with Tiling ...

Spherical Video



8K ERP Video

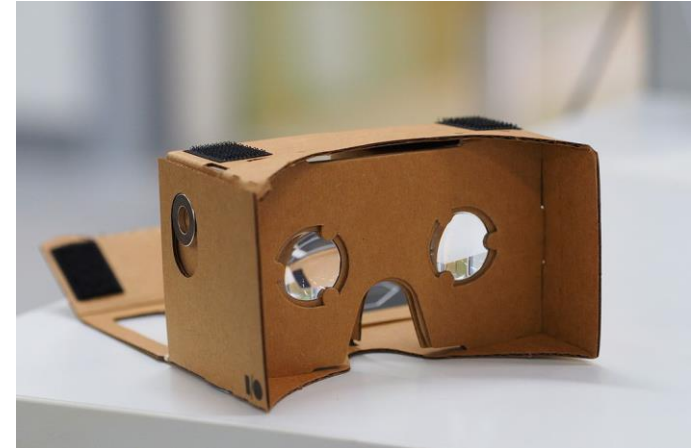


Tiled 8K ERP Video

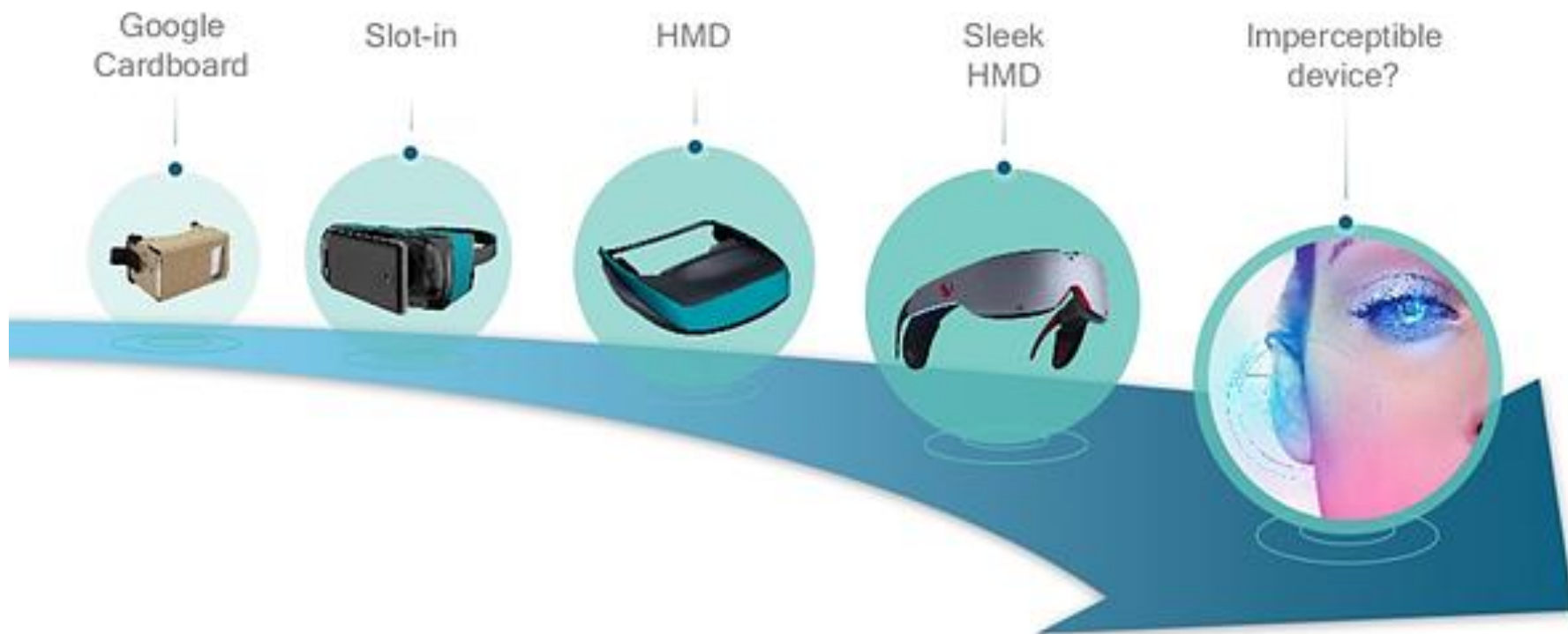


- **Coding of the rectangular projection may be performed with any of the usual image and video codecs, e.g. JPEG, H.264/AVC.**
- **Tiling is particularly relevant when streaming to avoid having to send the full projection, including the parts that are not being seen at all.**

- **Most 360-degree video is monoscopic (2D), meaning it is viewed as a one (360×180 degrees equirectangular) image directed to both eyes.**
- **Stereoscopic video (3D) is viewed as two distinct (360×180 equirectangular) images directed individually to each eye.**
- **360-degree videos are typically viewed via personal computers, mobile devices such as smartphones, or dedicated head-mounted displays.**
- **When viewed on PCs, the mouse is typically used to pan around the video by clicking and dragging. On smartphones, internal sensors such as the gyroscope are used to pan the video based on the orientation of the device.**



VR Display Evolution ...



What is a Viewport ...

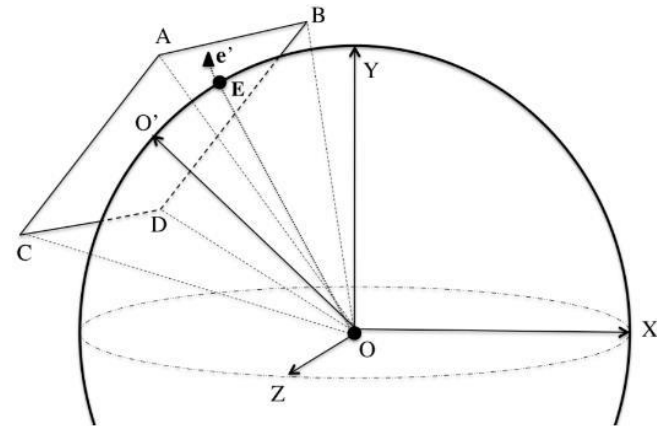
portion of video
rendered on a HMD

encoded video



Viewport Rendering ...

- It is necessary to select the pixels corresponding to the region that the user wants to visualize and send it to the display.
- The region that is presented to the display is often referred as the *viewport*. It is defined by horizontal and vertical Field-of-View (FOV), resolution and viewing direction corresponding to the center of the viewport that is controlled by the user.
- It is necessary to convert the omnidirectional video back to the spherical domain and after, knowing the viewport characteristics, project the spherical information into a tangent plane using a rectilinear projection (*corresponds to placing a flat piece of paper tangent to a sphere at a single point, and illuminating the surface from the spheres' center*).



Where is VR Nowadays ...

- *First steps: VR 360*
- **Video: Mono or Stereo**
- **Audio: Stereo or Spatial**
- **Very low resolution**
- **Limited (head) motion**
- **Large HMDs**



VR 360 Quality of Experience

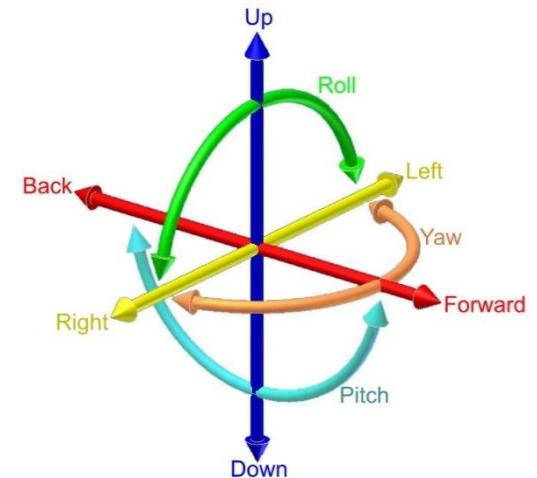
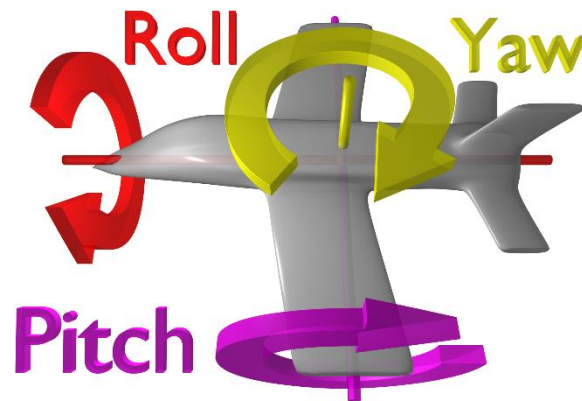
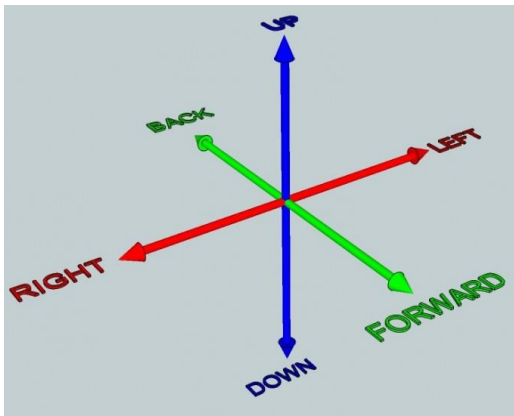


- Low spatial resolution
- Coding artifacts
- Motion-to-photon delay (time needed for a user movement to be fully reflected on a display screen)
- Amount of motion



The Visual Degrees of Freedom

- Degrees of Freedom (DoF) refer to the movement of a rigid body inside space, this means the “*different basic ways in which an object can move*”.
- There are only 6-DoF in total, essentially translations and rotations:
 - **Translations:** a body is free to translate in 3 degrees of freedom, forward/back, up/down, left/right.
 - **Rotations:** a body can also rotate with 3 degrees of freedom, pitch, yaw, and roll.



VR User Experience Quality ...

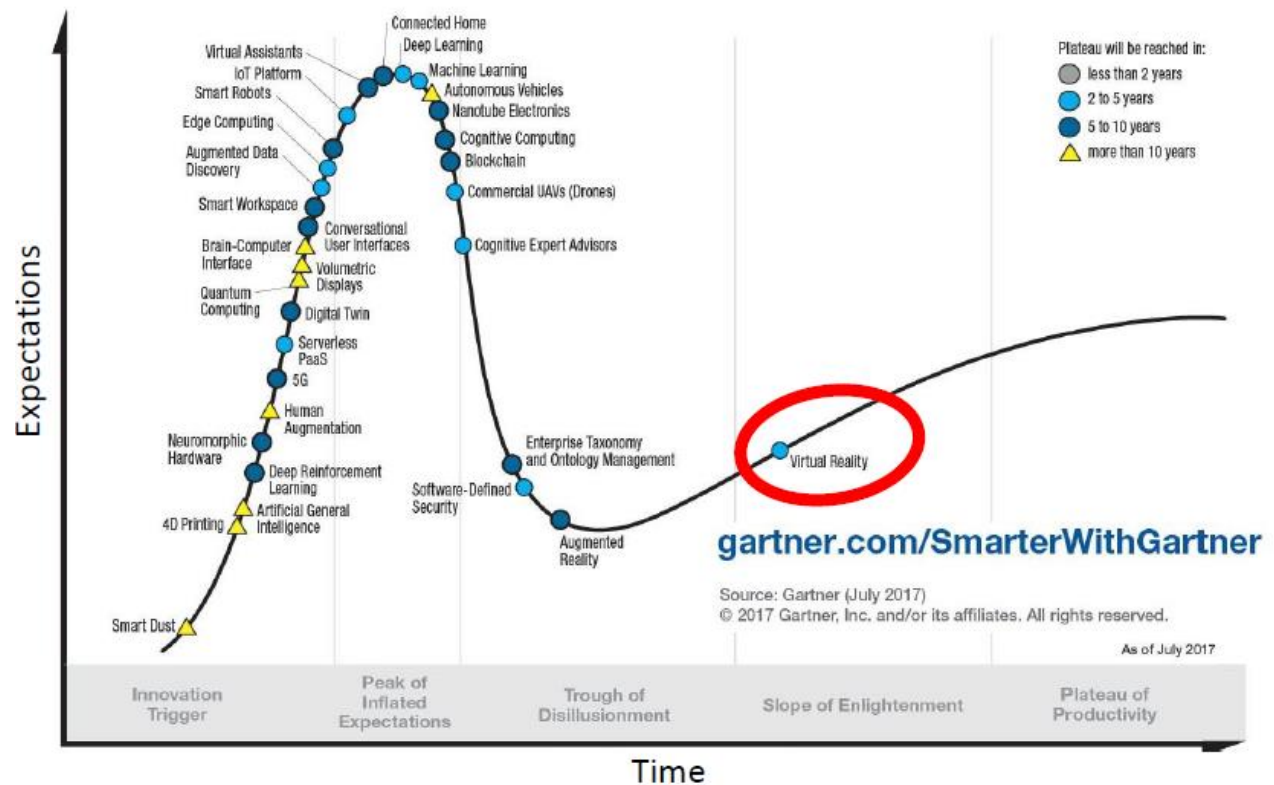
- **Current VR experiences have major cyber sickness issues, especially when moving cameras are used.**
- **6-DoF visual content (and not only 3-DoF) seems to be a critical step to achieve cyber sickness-free VR experiences.**
- **The 3-DoF to 6-DoF jump will significantly increase the amount of data.**



What Should be VR ...

- **Attractive user experience**
- **Great content**
- **Easy to use**
- **No side-effects**
- **Affordable for consumers and for providers**
- **Interoperable**

Gartner **Hype Cycle** for Emerging Technologies, 2017





**User experience has
to be great !**

**Technology alone is
not enough !**



Let's be Humble ... but Still Ambitious



**DREAM
BIG
AND
STAY
HUMBLE**

- **Visual representation WILL NOT be for ever what is today ...**
- **We have to keep trying opening new frontiers ...**
- **... with the courage to fail and keep trying ...**
- **And there are exciting things already happening ...**



**Whatever will be the
future, it has to be
researched today !**



**I hope you
have enjoyed
this
experience !**