# 11-755— Spring 2021
# Large Scale Multimedia Processing

# Lecture 5/6

# Multimedia capture and storage

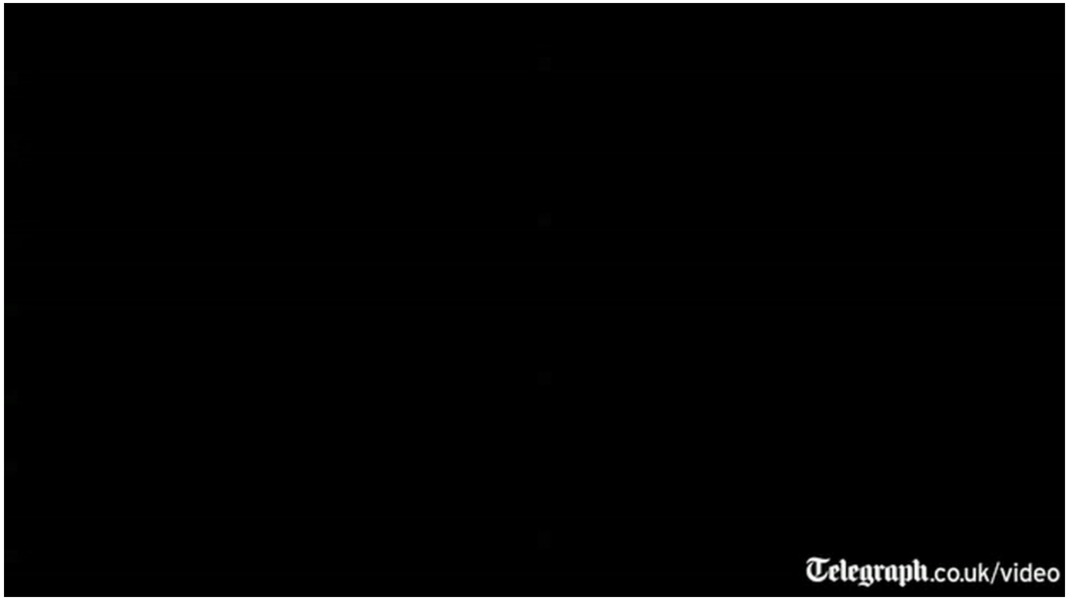
Rita Singh

Carnegie Mellon University

# In this lecture

- Digital multimedia: Recording and devices
  - Audio
  - Images
  - Video
  - Text
- Digital multimedia: Processing
  - Audio processing
  - Two generic processing techniques

# The first video: 1877



- https://www.telegraph.co.uk/culture/culturevideo/8035681/The-worlds-first-films.html
- The footage was shot from 1877, and projected by Muybridge's 'Zoopraxiscope' invented two years later in 1879. The dates mean that he was able to make and show films 16 years before the Lumiere Brothers used the more commercially successful cinematograph to popularise the art form.
- The dozen films are only a few seconds long and were commissioned to illustrate Victorian experiments in animal and human motion. Instead of using celluloid passed across a lens for a fraction of a second, Muybridge spaced out still cameras triggered by tripwires broken when the animals ran along, then projected the resulting sequence.

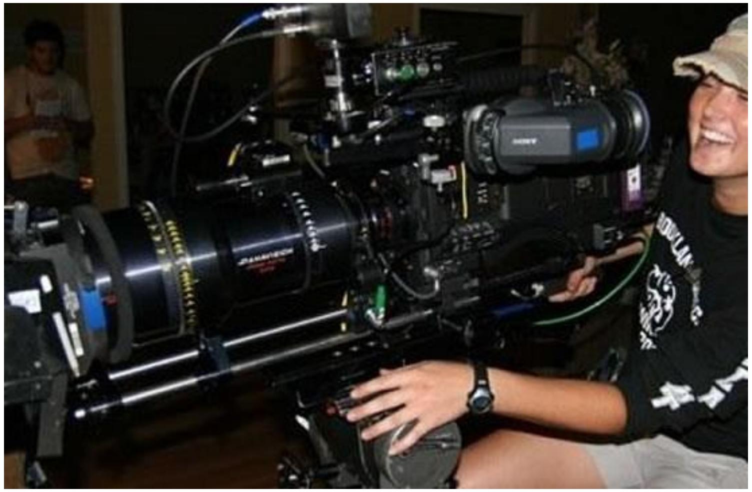
# Video cameras

- Analog video/movie cameras:
  - Mechanical devices that captured images on moving film (multiple photographs per second on a continuous film)

Movie Camera 35mm Mitchell
Film Original, circa 1940s

# Video cameras

- Digital video/movie cameras:
  - Just like still image cameras
  - Take multiple shots **automatically** at a fixed rate per second
  - The rate is decided based on persistence of vision
  - Usually 30 photos/second



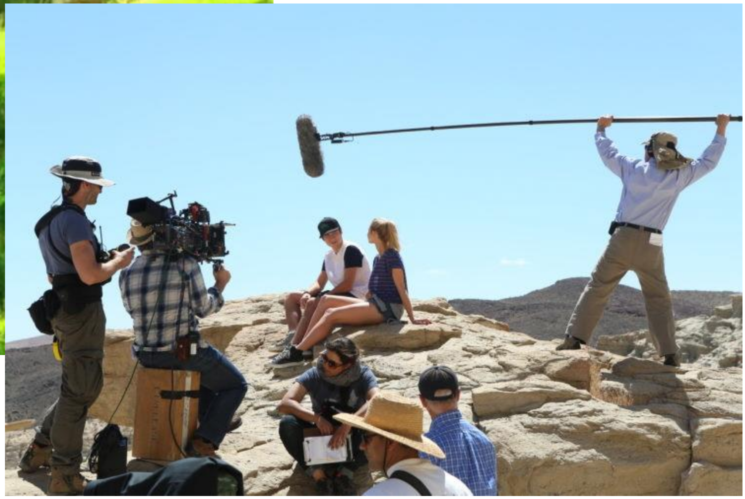
# Video cameras

- What happens to the sound?
  - Sound is **recorded alongside** and **time synchronized** with the photo frames

https://zaxcom.com/the-hobbit-a-production-sound-perspective/

# What happens to the sound?

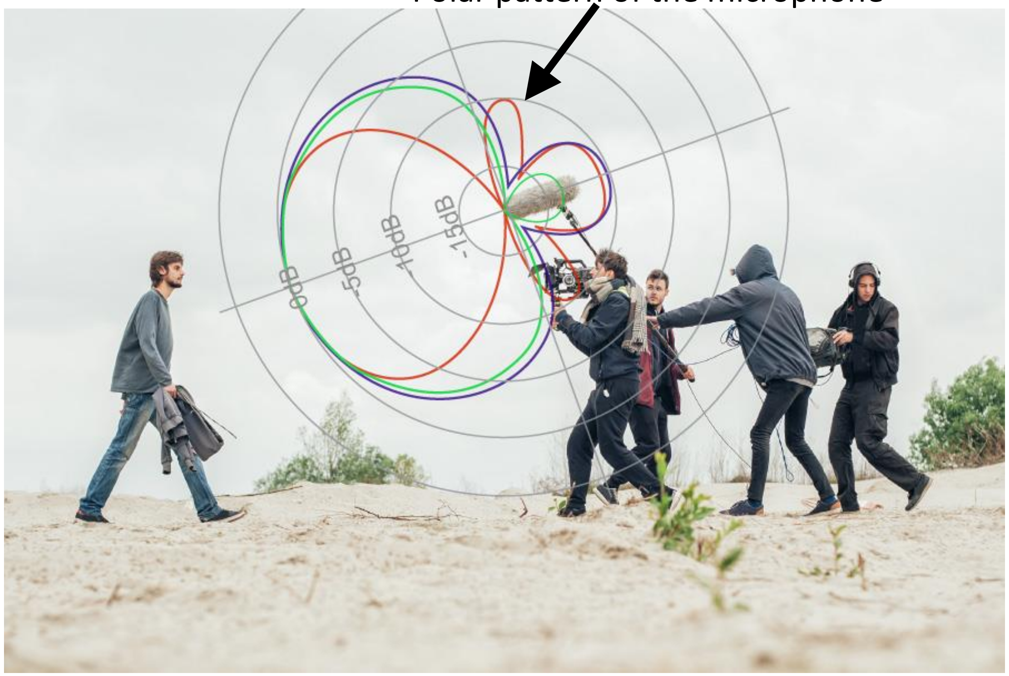
shotgun microphones - boom mics for movies



- Each film created its own set of challenges, but Johnson and his two boom operators, Corrin Ellingford and Steven Harris, consistently took the time to wire all the actors each day. The team used Zaxcom TRX900 transmitters in combination with Countryman B6s, DPA 4063s, and the occasional Sanken COS-11D for lavs depending on the application or actor. For the hobbit Bilbo (Martin Freeman), two lavs were needed at times. “Bilbo had this jacket he wore with lapels on it. What we did was put two wireless lavs, one on each side of his lapels, being Countrymen B6’s. They were placed at identical heights and levels so no matter which way he would swivel his head, the audio would match,” explains Johnson. “It was the first time I used two mics on an actor and it worked very well.” For the fourteen different dwarves, they used little hair clips to hide B6s into their beards. And in the case of Gandalf (Ian McKellen), sound hid a TRX900 with a B6 into the cone of his hat. “That saved so much production sound because it was like having a boom a few inches away from his mouth,” notes Johnson.

# Capturing audio

Polar pattern of the microphone

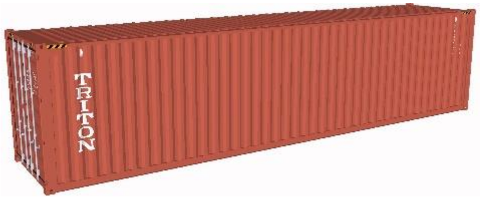


- **Captured** via one or more microphones, and then
  - **Digitized**
  - **Stored**

# A video is created

- A digital camera takes multiple shots **automatically** at a fixed rate
  - Usually 30 photos/second
  - Photos are stored in sequence
- Sound is **recorded alongside** and **time synchronized** with the photo frames
- Each type of data is a “stream”

- **AND THEN…..**
  - Everything is bundled up and stored together in a “container”
    - **Streams are still separate**
      - Image stream
      - Audio stream
      - Text stream (metadata!)



# Streams

- There can be many streams!
  - An image sequence: a visual data stream
  - multiple audio streams
    - E.g. a movie often contains multiple audio streams,  each in a different language, or with the director's commentary
  - multiple text streams
    - Can contain multiple subtitle tracks in different languages
    - A text stream is also called a metadata stream
    - **Metadata stream can contain:** information about the image compression **codec** used, audio codes used, bit rate, frames per second, closed captions, GPS information etc. etc.

- This is what allows us to watch the same movie in different languages from the same "file"...

# Video storage

- **Container types:** .avi, .wmv/.asf [Microsoft]; .mov [Apple]; .mkv (Matroska), .ogg [Opensource]; AVCHD [Panasonic/Sony], .mp4 (MPEG), .flv (Flash Video)

# Containers

Not every container can hold every type of codec

- **.AVI**
  - old and cannot hold the modern H.265 codec
- **.WMV, .ASF**
  - hold Microsoft codecs and cannot be played back on other operating systems
- **.MOV**
  - **supports all codecs**
  - can be played back using quicktime on all systems. Accepted by most  video editing software
- **.AVCHD**
  - developed for commercial camcorders,
  - only format that does not store in a single file, but in a folder structure. (Do not delete the folders to keep only the MTS files, many editing or converting software will not be able to work on it).

# Containers

- **.WEBM (.WEBP is for images)**
  - created by Google for efficient large-scale distribution.
  - small in size, not very high quality.
  - used for HTML5 video streaming sites, such as YouTub
- **.OGG**
  - **Can hold all codec formats**
  - Used for high-quality videos to be streamed via the internet.
  - Higher quality than WEBM files, takes longer to be delivered to the end-user.
  - Opensourced, used in a range applications, e.g. GPS receivers and media players (both desktop and portable).
- **.MKV**
  - **Can hold all codec formats**
  - Opensource, but not much integrated into editing and other processing software, so not in much use

# Containers

- **.MPG, .MP2, .MPEG, .MPE, .MPV**
  - Can contain audio/video media, or **simply audio**.
  - Small size, low quality (relatively),
  - use lossy compression, so best used when video will be recorded once and never edited.
- **.MP4, .M4P, .M4V**
  - Can contain audio/video media, or **simply audio**
  - does not store a wide variety of codecs but does support the most widely used codec, H.264
  - File formats are lossless, ideal for editing, saves don't iteratively lose quality.
  - used for streaming video via the internet. Higher in quality than WEBM files, also larger. Used ubiquitously, almost all devices, almost all video sharing websites.
  - **M4V**: proprietary iTunes files, share the same qualities of MP4 and M4P files,  DRM copy-protected.

# The culprits!!

The 94th ISO/IEC JTC1 SG29/WG11(MPEG) Meeting、The 53rd ISO/IEC JTC1 SG29/WG1(JPEG) Meeting、The 3rd JCT-VC Meeting of MPEG&VCEG

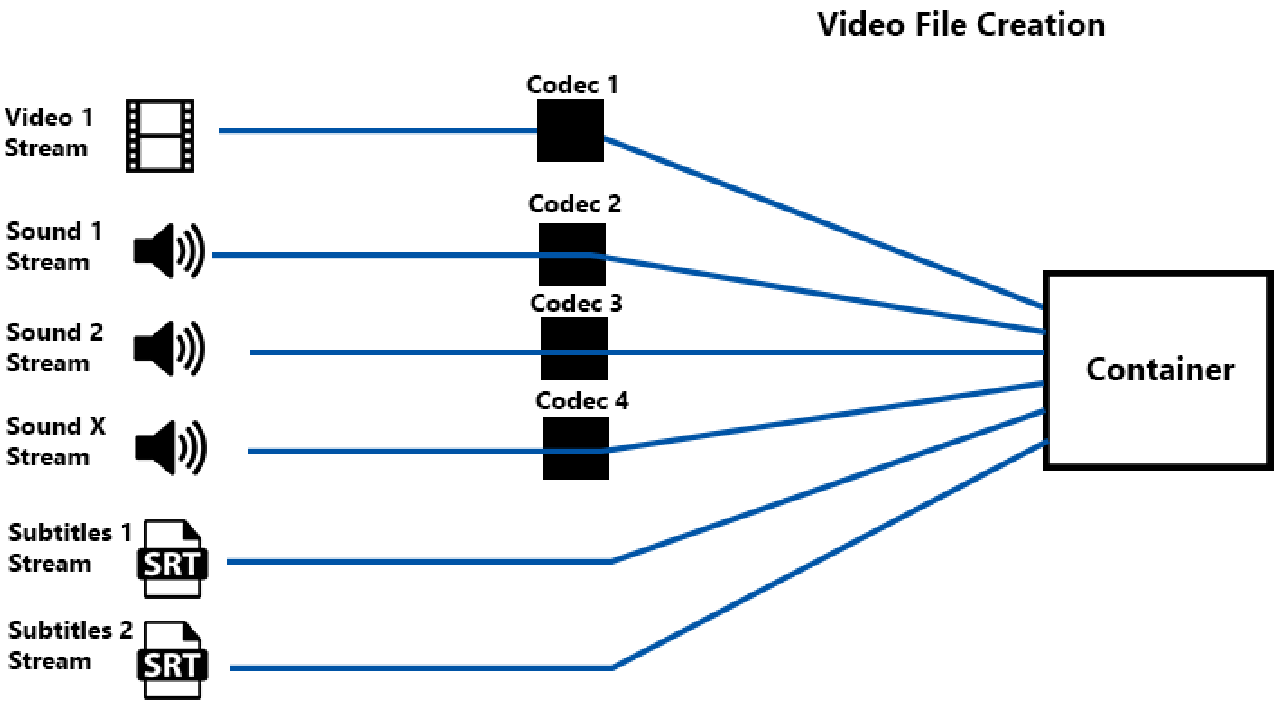
This is the power of MPEG

- The culprits: The group that created mp4
- MPEG: Moving Picture Experts Group

# Latest group…

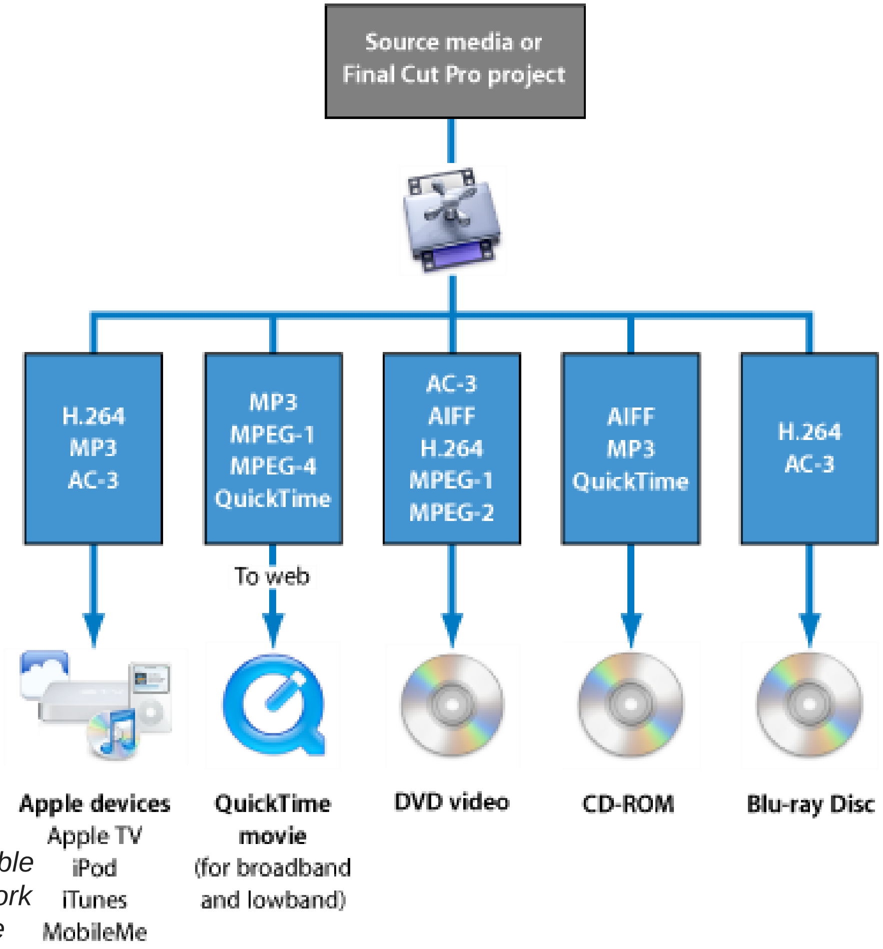
- MPEG 126 meeting, held on March 29, 2019 in Geneva, Switzerland. Some of the topics covered: 3Dof+, MPEG-G, and MMT.

# Creation of a video container



# Packaging

Codecs and streams are chosen according to the expected application mode of playback



*QuickTime: extensible multimedia framework developed by Apple*

# Video containers…

- MP4 format (.mp4)
- AVI format (.avi)
- FLV format (.flv)
- iPod/iPhone H.264 MP4 format (.mp4)
- raw MPEG-1 video (.mpg)
- PSP MP4 format (.mp4)
- Flash format (.swf)
- Windows Media Video (.wmv)
- Ogg (.ogg)
- DV video format (.dv)
- MOV format (.mov)
- 3GP/3G2 Video (.3gp)
- MJPEG (Motion JPEG) (.mjpeg)
- GIF Animation (.gif)
- MPEG-2 PS format (DVD VOB) (.dvd)

**Audio only**

- MPEG audio layer 3 (.mp3)
- WAV format (.wav)
- Ogg (.ogg)
- ASF Audio (.wma)

# Codecs

- Codec
  - (portmanteau of two words….)
  - **scheme** to **co**mpress images for storage and distribution… and to **dec**ompress the compressed images for display
  - Jpeg is a **codec**

# Image codecs (formats)… images

For raster images

- Joint Photographic Experts Group (.jpg)
- Windows bitmap (.bmp)
- Encapsulated PostScript (.eps)
- Graphics interchange format (.gif)
- Portable network graphics (.png),
- Portable Document Format (.pdf)
- Adobe Photoshop bitmap file (.psd)
- Tagged Image File Format (.tiff)
- Adobe PostScript (.ps)
- Apple Macintosh QuickDraw (.pict)
- AVS X image (.avs)
- CCIR 601 4:1:1 (.yuv)
- Flexible Image Transport System (.fits)
- Irix RGB image (.sgi)
- Kodak Cineon Image Format (.cin)
- Magick image file format (.miff)
- On-the-air Bitmap (.otb)
- Palm pixmap (.palm)
- Photo CD (.pcd)
- Photo CD - sRGB color (.pcds)
- Portable anymap (.pnm)
- Portable bitmap format black/white (.pbm)
- portable pixmap file format (.pgm)
- Portable pixmap format - color (.ppm)
- SMPTE digital moving picture exchange (.dpx)
- Truevision TGA (.tga)
- Utah RLE encoded image (.rle)
- X Windows system bitmap black/white (.xbm)
- X Windows system pixmap (.xpm)
- Xv's Visual Schnauzer thumbnail (.p7)
- ZSoft IBM PC multi-page Paintbrush (.dcx)
- ZSoft IBM PC Paintbrush file (.pcx)
- Magick Vector Graphics (.mvg)
- Windows Metafile (.wmf)

For vector images

- AI (ps based Adobe Illustrator 5.0) (.ai)
- Scalable Vector Graphics (.svg)
- SK (Sketch/Skencil format) (.sk)
- SK1 (sK1 format) (.sk1)
- CGM (Computer Graphics Metafile) (.cgm)
- WMF (Windows Metafile) (.wmf)
- PDF (Portable Document Format) (.pdf),
- PS (PostScript) (.ps)

# Audio codecs in the containers…

- MPEG-1/2 Audio Layer 3 (.mp3)
- WAVE form audio format (.wav)
- Ogg Vorbis (.ogg)
- 3GP Music (.3gp)
- ATSC A/52A (AC-3) (.ac3)
- Adaptive Multi-Rate (.amr)
- Advanced Audio Coding (.aac)
- ALAC (Apple Lossless Audio Codec) (.m4a)
- ASF Audio (.wma)
- Audio Interchange File Format (.aiff)
- Free Lossless Audio Codec (.flac)
- MPEG advanced audio coding (.mp4)
- MPEG audio layer 2 (.mp2)
- MPEG-4 advanced audio coding (.m4a)
- Soundblaster file (.voc)

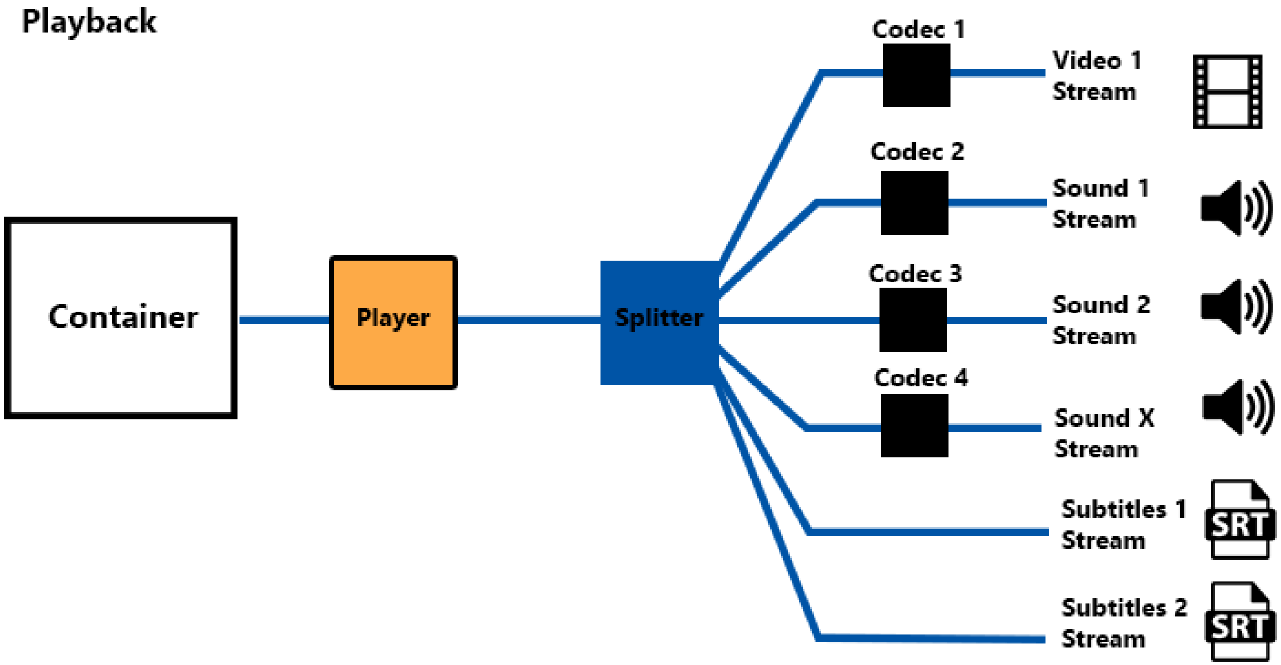
# Video codec??

- There are many kinds of codecs for video
  - E.g. H.265, MPEG-4, DivX Pro, MPEG-4 Part 2 codec such as Xvid, Ffmpeg etc.
  - Not every type of container can hold every type of codec
- More about codecs later…

# Extraction of streams from a container

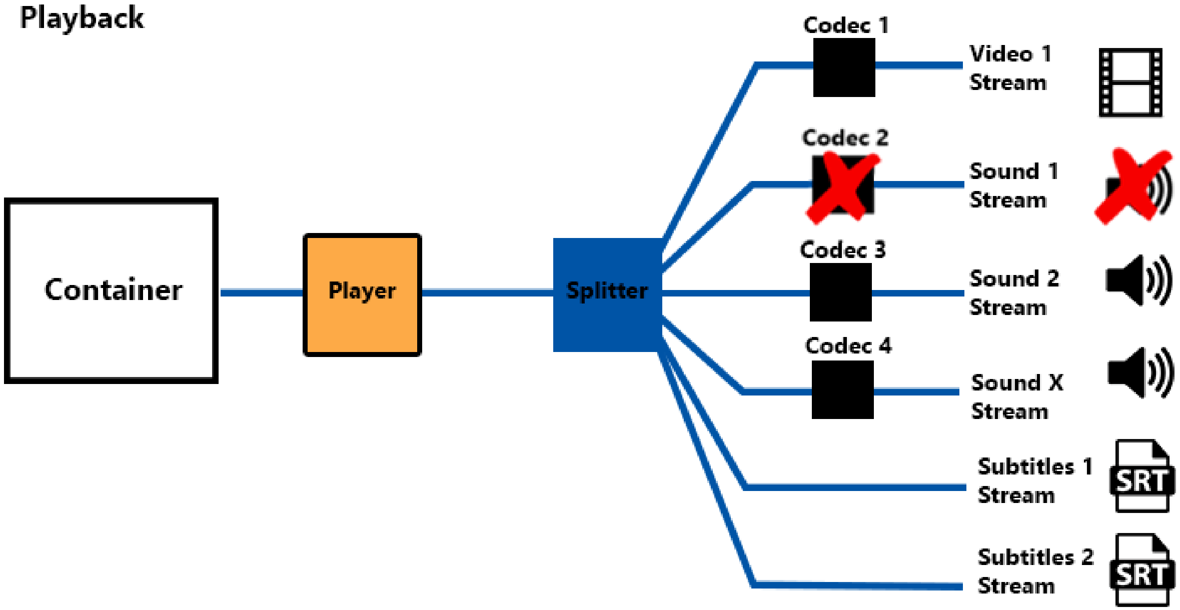
- To extract the streams from a container, we need a **splitter**
  - Process is called *decoding for/or playback*
  - Player detects container first
  - Then searches for and uses the appropriate splitter
  - Streams are compressed, they need to be decompressed
  - The *decoder of the codec* is used to decompresses them

# Decoding



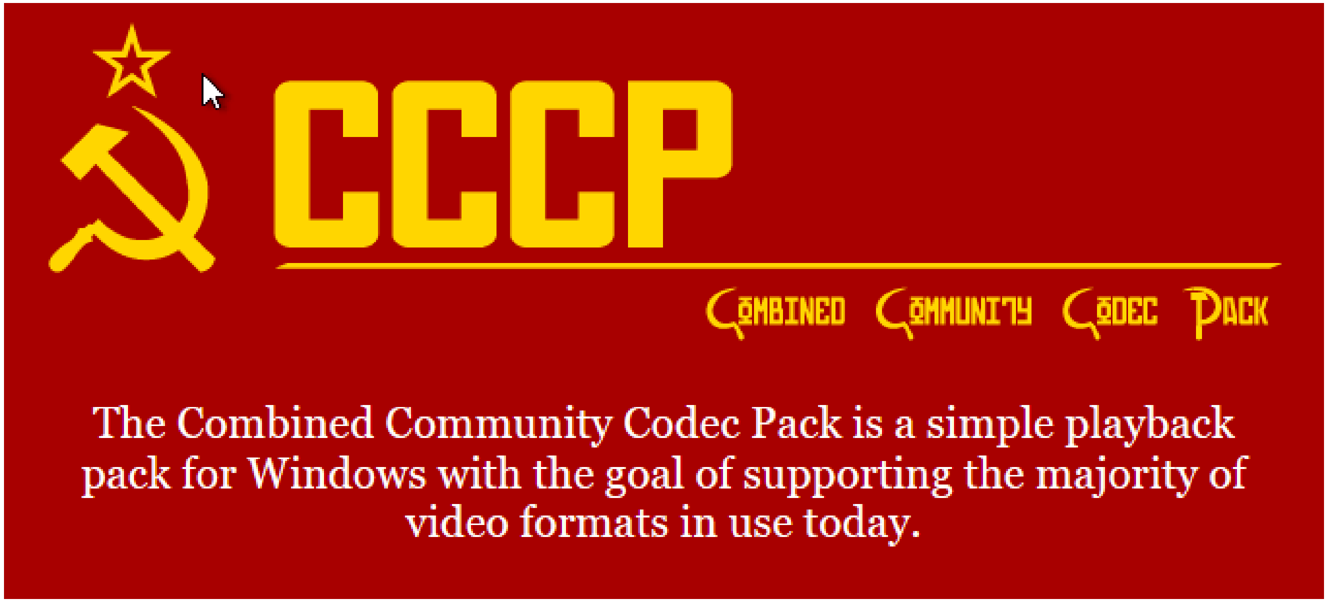
- Playback process

# Playback problems



- Video does not play: missing codec, wrong splitter
- E.g. VLC is well known for its ability to play any video file since it was first released. That’s because it already contains a variety of codecs and splitters.

# Codec packs



CCCP

Combined Community Codec Pack

The Combined Community Codec Pack is a simple playback pack for Windows with the goal of supporting the majority of video formats in use today.

- You can install a codec pack on your machine to enable video decoding and playback if there is a problem
- E.g. **K-Lite Codec Pack,** CCCP

# Back to video stream codecs

**Why do we need a codec? Why compress?**

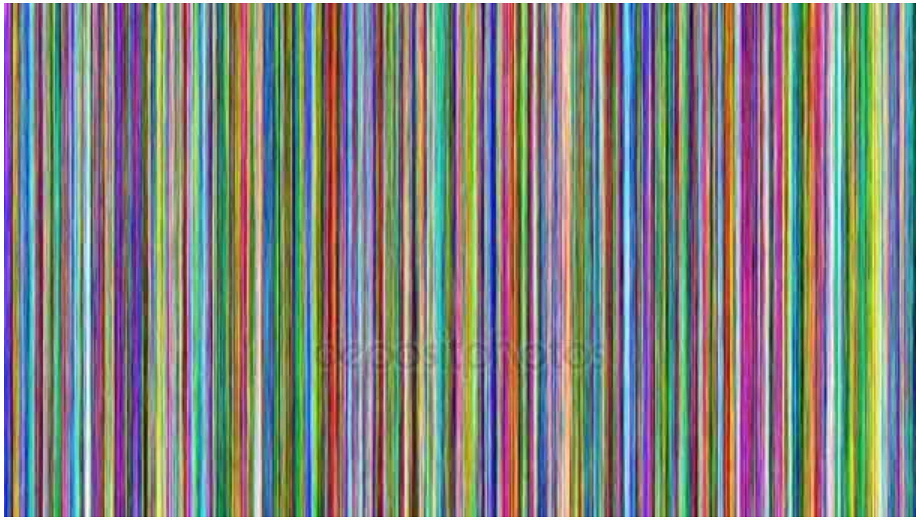
- **Videos take up a lot of storage**
  - E.g. HD video can take up to 400GB per hour of film!!
- **Why use HD then?**
  - We want to see the world in as much detail as we see in real life
  - For this we need high-resolution (or high definition) in the image stream

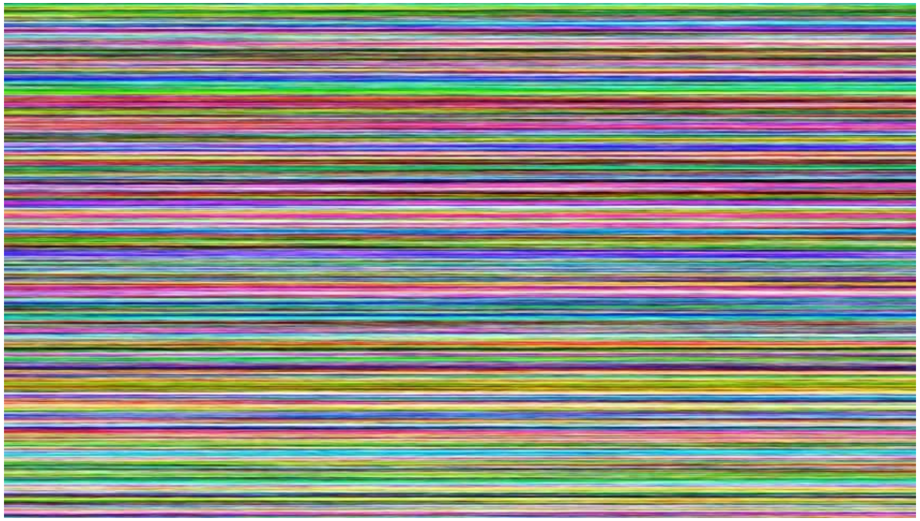
# Quality: HD video

- **High-definition video**
  - Video of higher resolution and quality than standard-definition
- No standardized meaning for *high-definition*
  - BASED ON RESOLUTION: Generally any video image with more than 480 vertical lines (North America) or 576 vertical lines (Europe) is considered HD
  - BASED ON SPEED: Images of standard resolution captured at rates faster than normal (60 frames/second North America, 50 fps Europe), by a high-speed camera may be considered high-definition in some contexts
    - E.g. Some television series shot on high-definition video are made to look as if they have been shot on film, a technique which is often known as **filmizing**.

# Scan lines

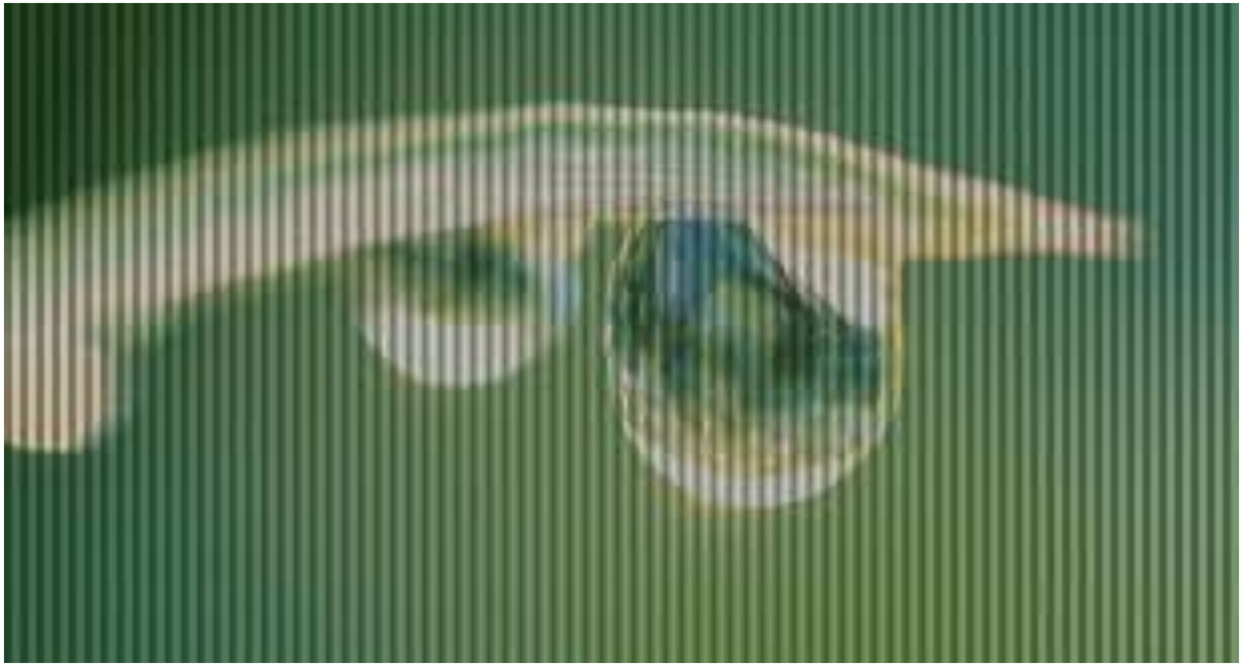
Vertical scan lines (HD has >= 480 vertical lines)





- Vertical and horizontal scan lines

# Interlacing



- Interlacing of vertical scan lines

# Video editing can cause artefacts

- **To study the artifacts we need the right information/features that can capture them, e.g.**
  - **Features:** HOG, SIFT, Mo-SIFT etc…
  - **Information (parameters):** Motion vectors, optical flow, gradient fields, noise residuals, etc… (many are used to encode video in different codecs)

# Editing artefacts

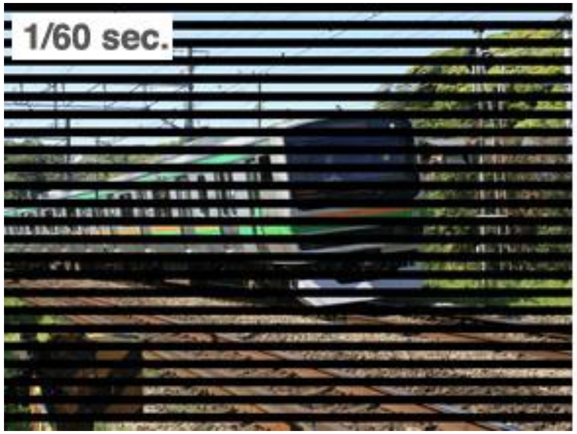
Example: Double/Multiple compression

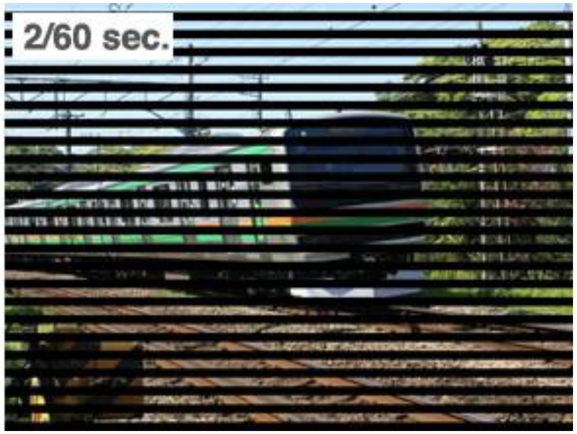
- In untampered video, the distribution of block Discrete Cosine Transform (DCT) coefficients (DC/AC) is usually Gaussian or Laplacian.
  - This is true even after the coefficients are quantized

- If recompressed with different encoding parameters, the distribution changes

- When I-frames undergo double JPEG compression
  - Relatively larger motion estimation errors occur when frames are moved from one GOP to another (in an attempt to perform frame deletion or insertion).
  - Periodic spikes are observed  in the Discrete Fourier Transform (DFT) of the P-frame prediction error sequence.

# Intra-frame editing: interlacing artefacts





- Interlacing:
  - 30 **IMAGES** per second = 30 **FRAMES** per second. If interlaces, each FRAME has two consecutive **FIELDS**
    - Fields are captured in succession at a rate twice that of the nominal frame rate.
    - One contains odd lines 1,3,… (**upper field**) and other contains even lines 2,4,…. (**lower field**)
    - PAL and SECAM systems have a rate of 25 frames/s or 50 fields/s
    - NTSC system delivers 29.97 frames/s or 59.94 fields/s
    - **Each FIELD is slightly motion displaced**
- To edit, you need to **de-interlace**!

# Intra-frame editing

- Terminology and de-interlacing: an example
  - This example is of Adobe After Effects (most commonly used tool to do this kind of tampering, deinterlaces well)
-

Watch full video at: https://www.youtube.com/watch?v=LA-Pc2jMn60

# Intra-frame editing

- De-interlacing is done by adding the two fields and smoothing
- Many algorithms for this
  - SmoothDeinterlacer
    GreedyHMA
    TomsMoComp
    FieldDeinterlace
    etc…
- Smoothing introduces spatial and temporal correlations

TAMPERING

  - De-interlacing correlations are destroyed
  - Distorts  motion across fields of neighboring frames in de-interlaced videos and the equality of the motion between fields of a single frame in interlaced video

# HD video

**High definition video:** 3 criteria are used to categorize as HD

1. ***The number of lines in the vertical* display resolution**
2. ***The scanning system***
3. ***The number of frames or fields per second (Hz)***

**Explained in the next slide….**

# HD video

- High-definition television (HDTV) resolution is 1080 or 720 lines
- Regular digital television (DTV)
  - 480 lines: NTSC is based on this, has 480 visible scanlines out of 525
  - 576 lines: PAL/SECAM are based on this, 576 visible scanlines out of 625
- **Most of the time HDTV defaults to DTV!!**
  - When HD is broadcast digitally within the limitations of DTV
- DVD quality is not HD, only disc systems such as Blu-ray Disc and the HD DVD are HD

- Read: https://en.wikipedia.org/wiki/High-definition_video for further details

# HD video

***The number of lines in the vertical* display resolution**

- Generally any video image with more than 480 vertical lines (North America) or 576 vertical lines (Europe) is considered HD

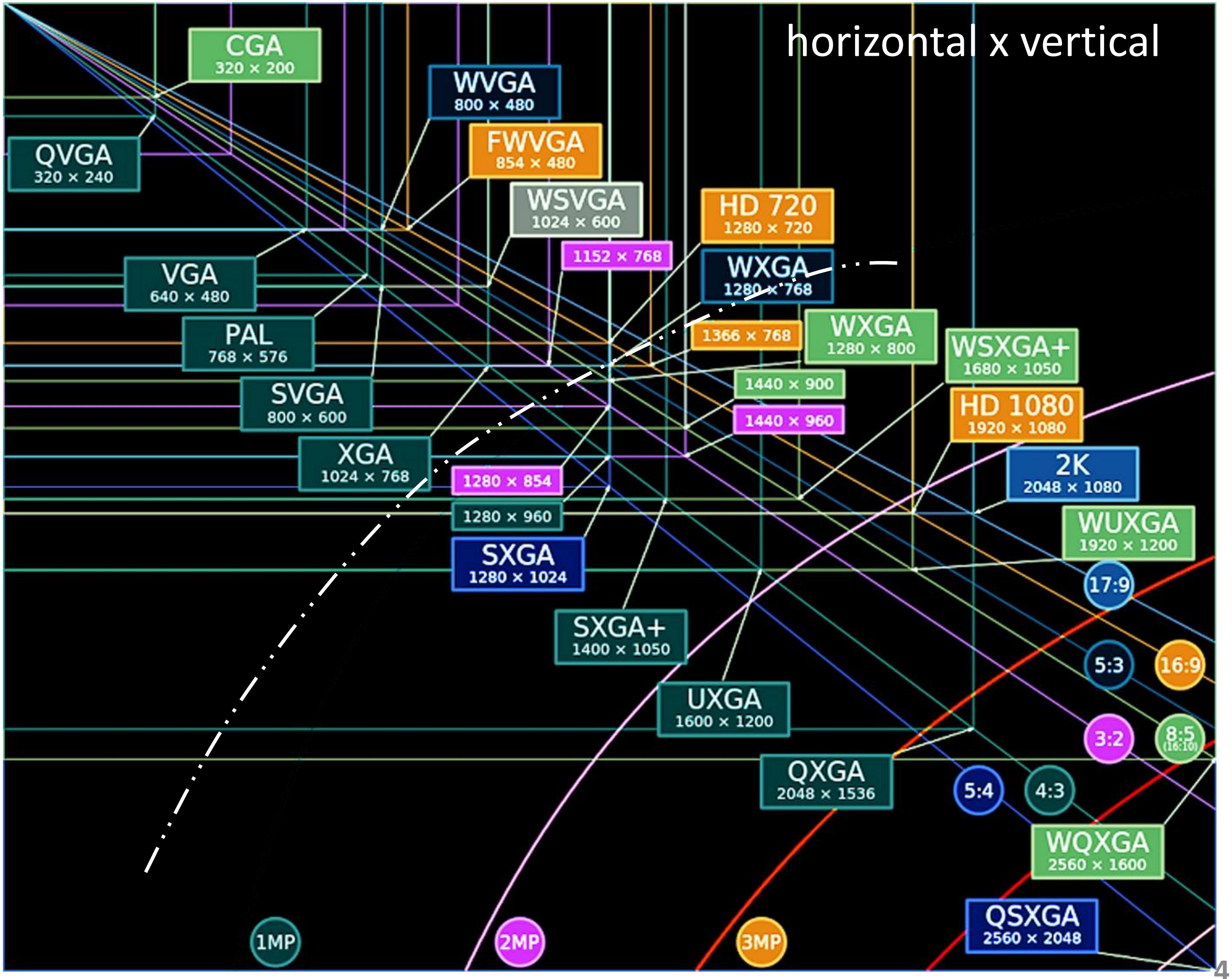
***The scanning system:***

- *2 types*
  - ***progressive scanning (p)****:* redraws an image frame (all of its lines) when refreshing each image
  - ***interlaced scanning (i):*** *r*edraws the every other line or "odd numbered" lines during the first image refresh operation, and then draws the remaining "even numbered" lines during a second refreshing,
    - E.g. 1080i refers to 1080 vertical scan lines interlaced system

***The number of frames or fields per second (Hz).***

- Europe: mostly 50 Hz television broadcasting; USA: 60 Hz
  - Example: The 1080i50/1080i60 format is 1920 × 1080 pixels, **interlaced encoding** with 50/60 fields, (50/60 Hz) per second.

horizontal x vertical



# HD video

- *interlaced scanning (i)*
  - *Loses some resolution in each frame but more than makes up in other advantages*
    - e.g 1080i. Is high res for still images, loses up to half of the resolution and suffers "combing" artifacts when subject is moving.
  - All HDTVs are **progressive-scan** displays — so even if the signal being sent to the HDTV is **interlaced**, the HDTV **will** convert it to **progressive scan** for display on the screen.

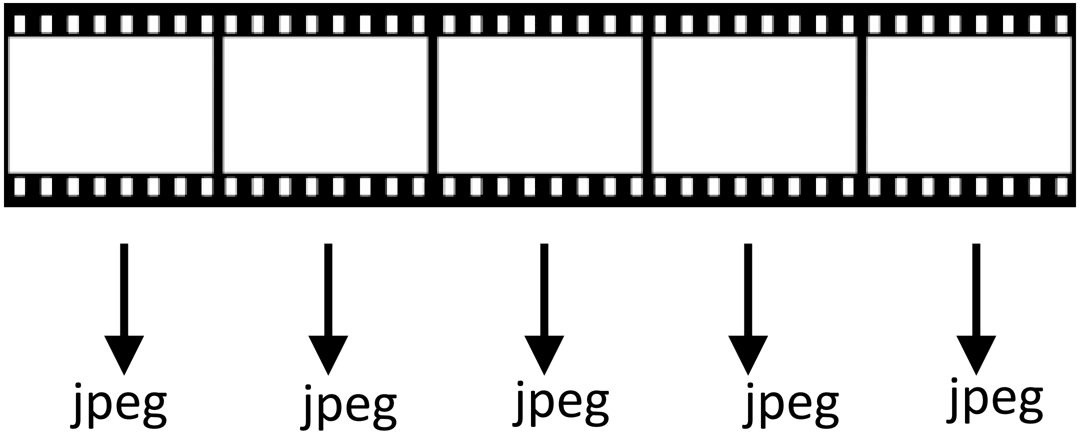
# Notes for later: HD video

- ***The number of frames or fields per second (Hz).***
- Europe: mostly 50 Hz television broadcasting system
- USA: 60 Hz
  - The 720p60 format, Also called HD Ready or standard HD, is 1280 × 720 pixels, **progressive encoding** with 60 frames per second (60 Hz).
  - 720 horizontal lines and an aspect ratio (AR) of 16:9, normally known as widescreen HDTV (1.78:1).
  - All major HDTV broadcasting standards (such as SMPTE 292M) include a 720p format which has a resolution of 1280×720; however, there are other formats, including HDV Playback and AVCHD for camcorders, which use 720p images with the standard HDTV resolution.
  - The frame rate is standards-dependent
    - for conventional broadcasting: 50 progressive frames per second in former PAL/SECAM countries (Europe, Australia, others), and 59.94 frames per second in former NTSC countries (North America, Japan, Brazil, others).

# Video stream codecs

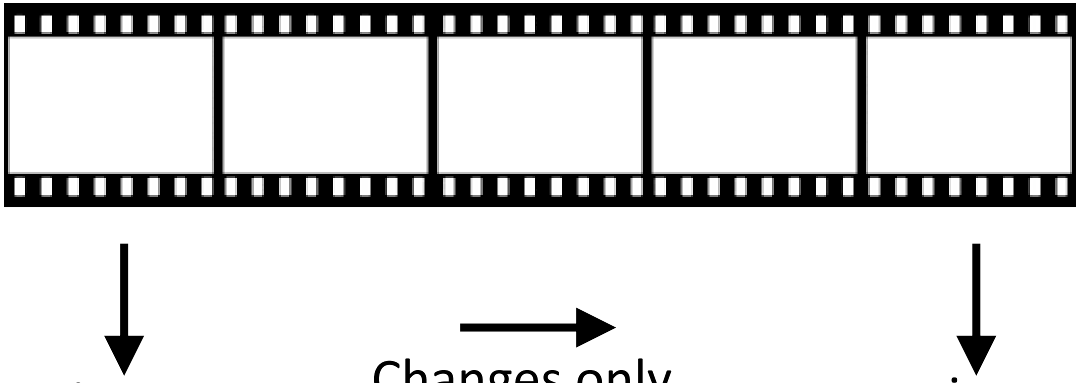
- A video codec Is used to compress the image sequence that comprises the video
  - Some codecs can be used to compress audio as well
  - Lossy codecs reduce quality but compress a lot more compared to lossless codecs.
  - 2 types: **interframe** and **intraframe** codecs

# Intraframe codec



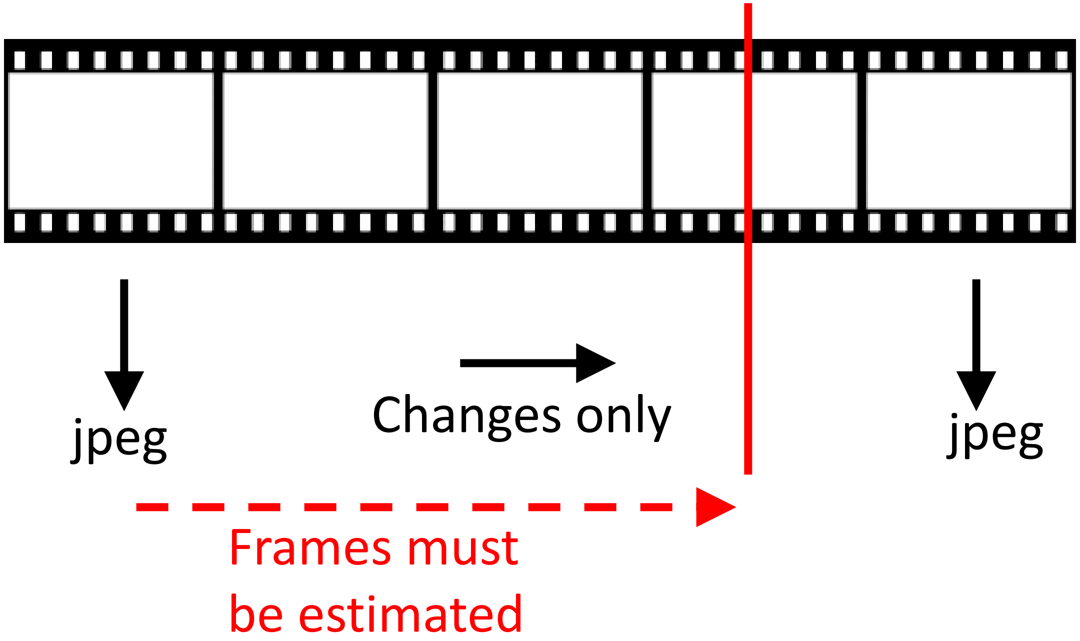
- Compress each frame separately, e.g. with jpeg compression

# Interframe codec



- Compress one frame every few seconds, and in between, save changes only:  areas that stay the same, moving objects or camera pans

# Interframe codec



- During editing if a cut is made at the red line, all frames from the last saved one must be calculated
  - More work for the computer

# Interframe and Intraframe codecs

Best intraframe codecs

- ProRes
- DNxHD

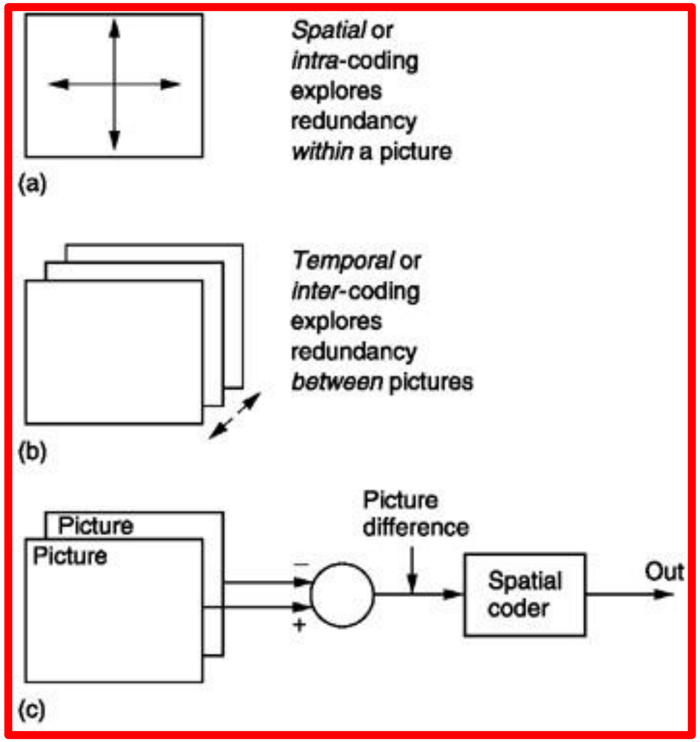
More at: https://wolfcrow.com/blog/intra-frame-vs-inter-frame-compression/

Best Interframe codecs

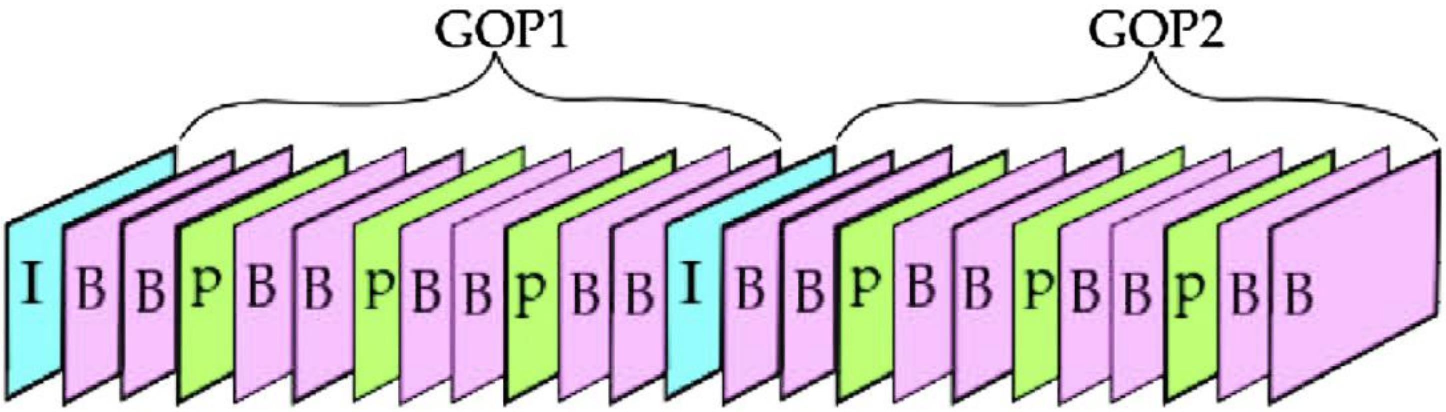
- mpeg2 (for DVD)
- divx
- mpeg4 (best for sharing)
- H.264 (H.265 is on its way)

# Video coding

- Videos: can be represented in 2 spatial and 1 temporal dimension
- Coding exploits redundancies in spatial and temporal domains
- **Spatial coding**
  - Also called intra-coding
  - Exploits redundancies within a frame
  - Involves *transform domain* coding
- **Temporal coding**
  - Also called inter-coding
  - Exploits redundancies  across frames
  - Involves *predictive* coding

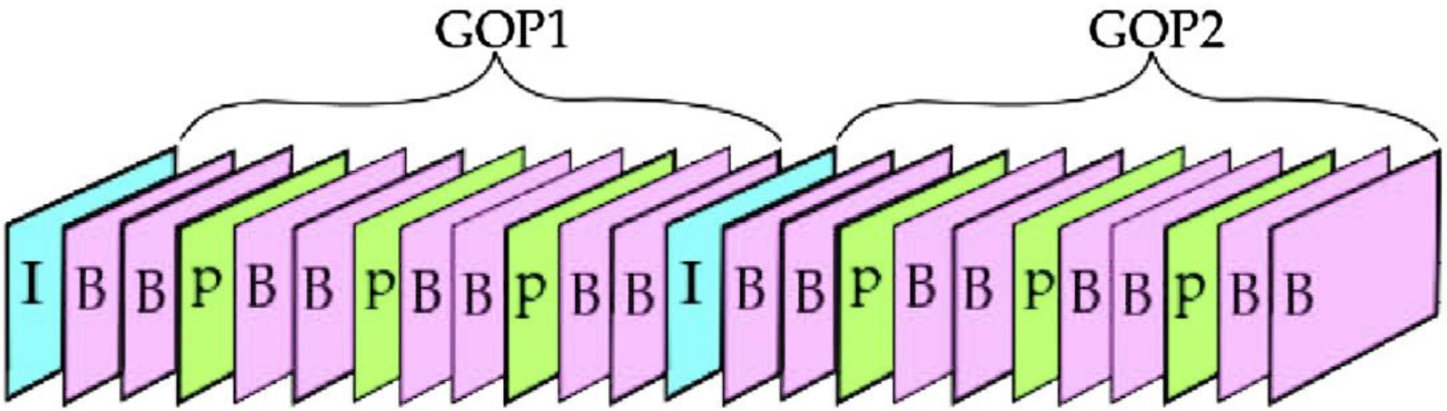


# Temporal (predictive) coding



- 3 three types of frames
- I-frame (Intra-coded), P-frame (Predicted) and B-frame (Bi-directionally predicted)

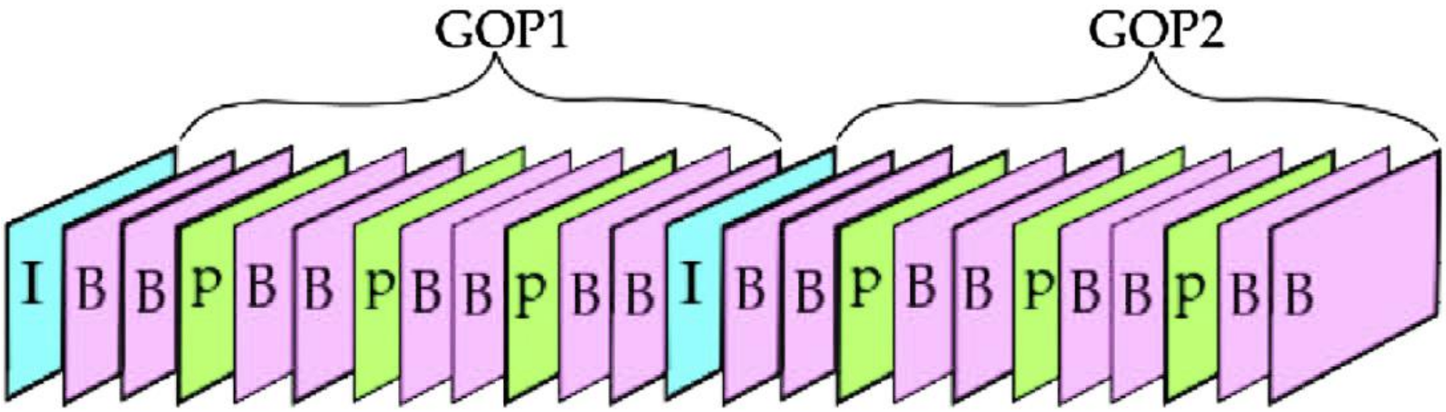
# Temporal (predictive) coding



- I-frames are coded using a JPEG like scheme. Exploits spatial redundancies ( as in image coding). The first frame of a video is an I-frame
- P-frames are predicted from previous I or P frames. Only changes from the reference are stored.
- B-frames are predicted from previous and future (backward) reference frames. More compression!

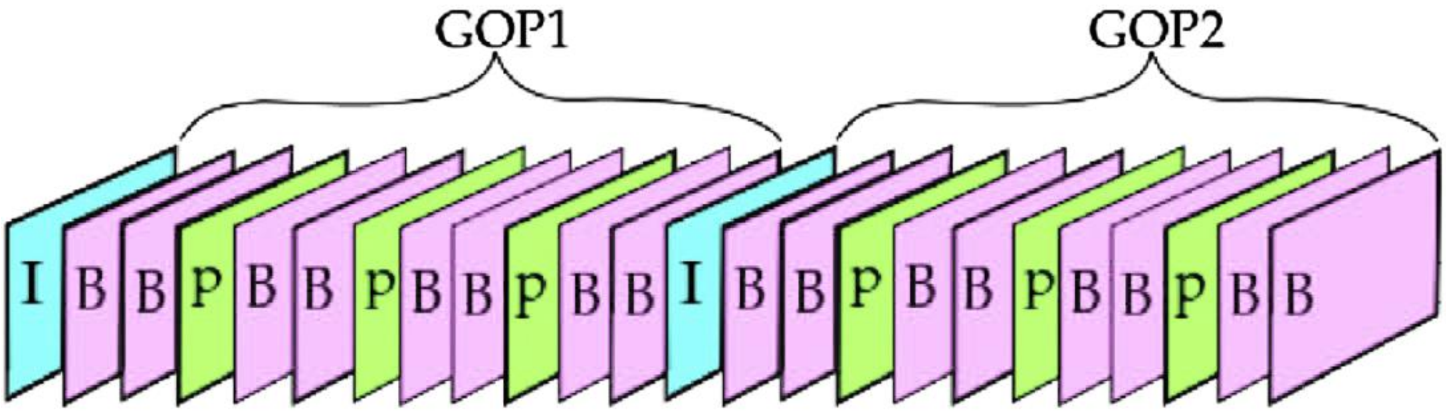
Often not possible to predict all frames in a video from the first frame. I-frames are inserted at regular intervals **or** inserted based on the motion in the video

# Temporal (predictive) coding



- Video sequence is divided into fragments, called Group of Pictures (GOP).
- In each GOP
  - I, B and P frames are ordered such that I-frame is first, followed by B and P frames.
  - Most encoders use fixed number of frames in a GOP
    - easy to implement, but affects coding efficiency and visual quality
    - Adaptive GOP (AGOP) is better

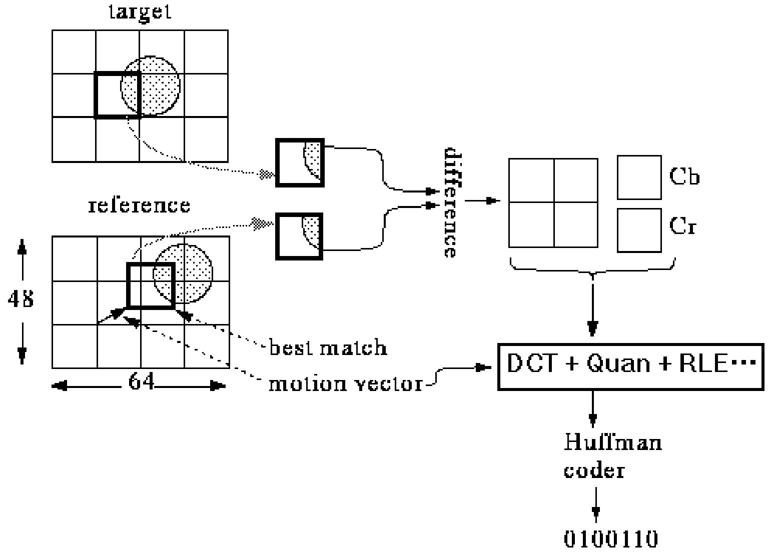
# Temporal (predictive) coding

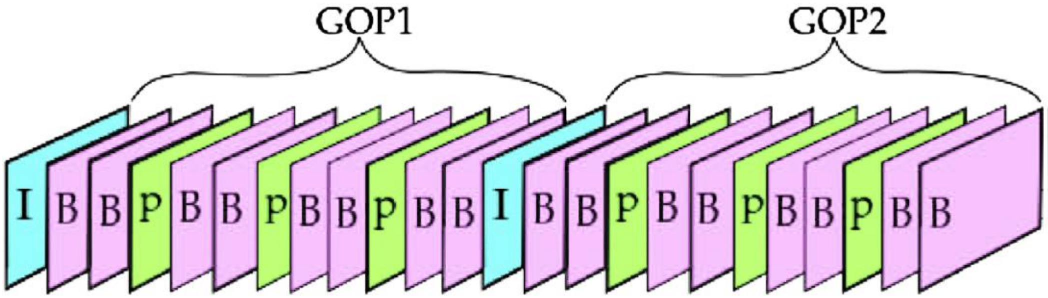


## Adaptive GOP (AGOP)

- length of GOP (number of frames) varies according to the video content
- Each frame in a GOP is segmented into blocks, called **macro-blocks (MB)**
- RGB frames areas converted to YCbCr or YUV
- One MB is divided into four 8 x 8 blocks from Y component (luminance component) and one 8 x 8 sample block from each chrominance components (sensitivity of human eyes is more for the luminance component)
  - = total of six 8 x 8 blocks in an MB
  - Allows video to be compressed in 4:2:0 format

# Video coding



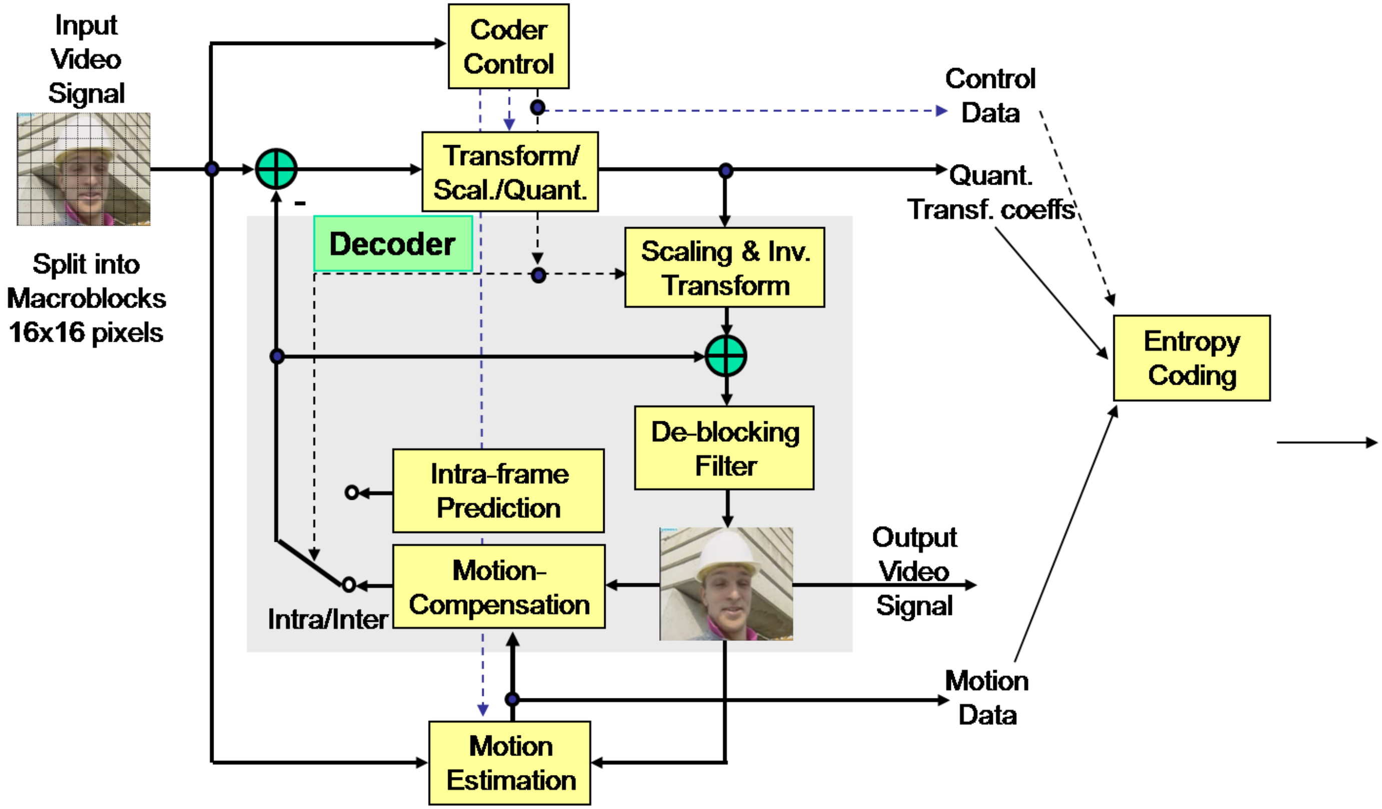


- **Encoding a macro block (MB) in P or B frames:**
  - Search for its best matched MB in the reference frames and store the location along with how far this matched MB has to be displaced (motion vectors) to create the predicted frame.
    - [Other features used: gradient fields, noise/error residuals etc.]
  - This information is coded using a JPEG-like process.
  - Intra-coded MB (IMB), Predicted MB (P-MB), Bi-predicted MB (B-MB) and skipped-MB (S-MB); no coding is required for skipped-MBs. A frame is coded as one or more slices, which contains one or more MBs.

# A closer look at video codecs: H.264

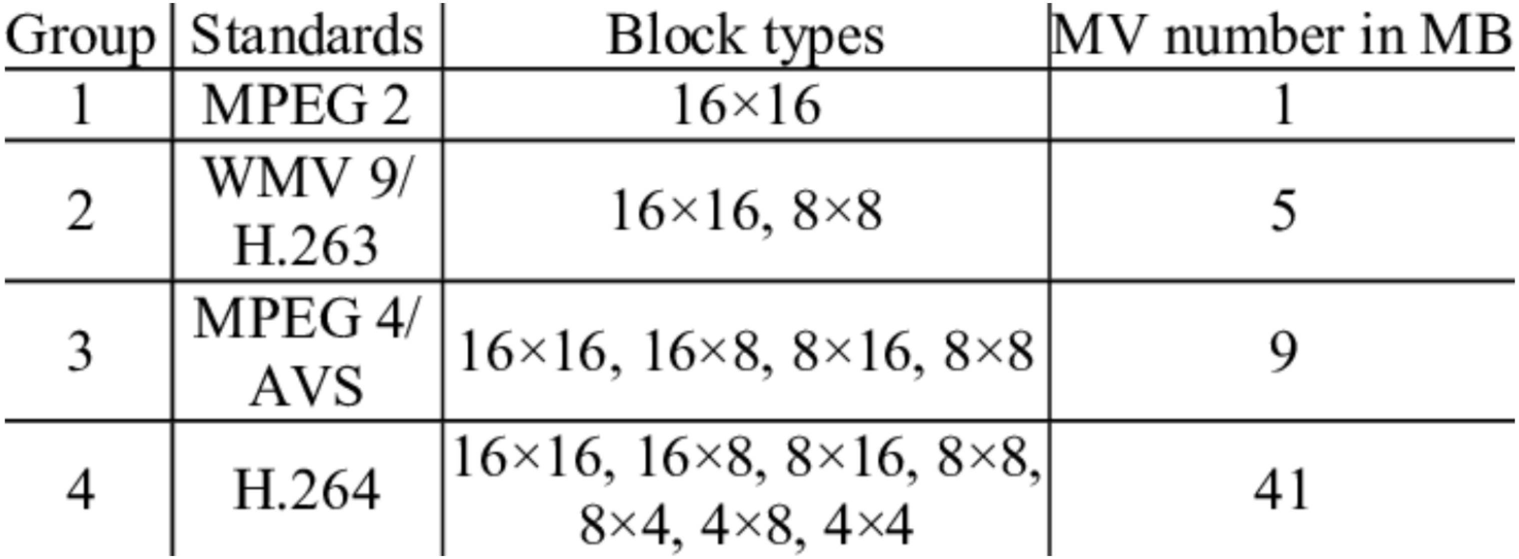
- **What is H.264 (**MPEG-4 AVC**)?**
  - A video coding format, nearly lossless coding
- Wide range of applications
  - low bit-rate streaming applications (YouTube, iTunes, Vimeo, Facebook, Instagram)
  - HDTV broadcasts over terrestrial, cable, and satellite
  - Blu-ray Discs
  - DVD storage
  - IP packet network
  - Digital cinema applications

# H.264



- The H.264 standard codec

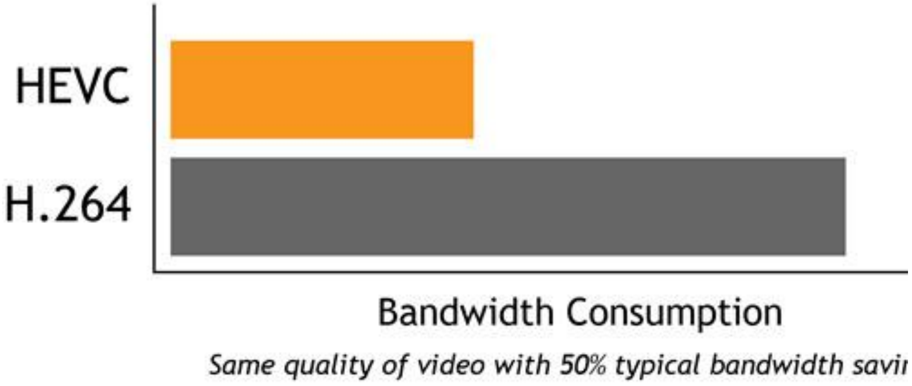
# Block choice differences between codecs



| Group | Standards | Block types | MV number in MB |
|---|---|---|---|
| 1 | MPEG 2 | 16×16 | 1 |
| 2 | WMV 9/ H.263 | 16×16, 8×8 | 5 |
| 3 | MPEG 4/ AVS | 16×16, 16×8, 8×16, 8×8 | 9 |
| 4 | H.264 | 16×16, 16×8, 8×16, 8×8, 8×4, 4×8, 4×4 | 41 |

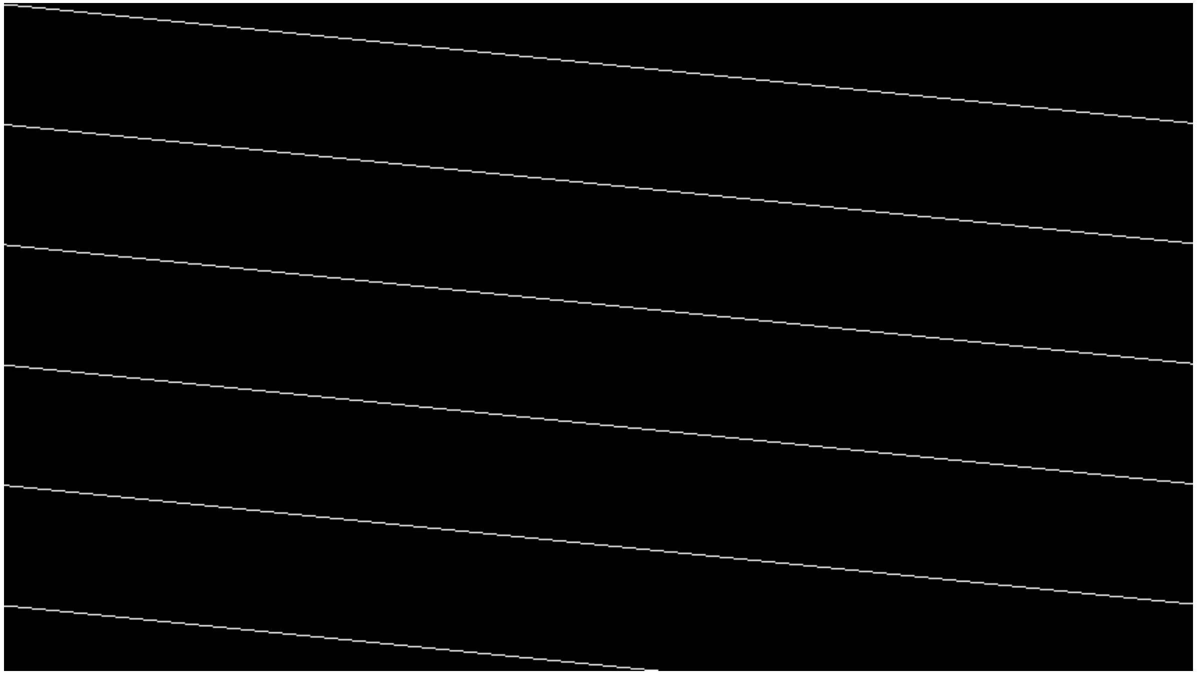
- Example… (MV stands for Motion Vector)

# Comprison: H.264 vs. H.265



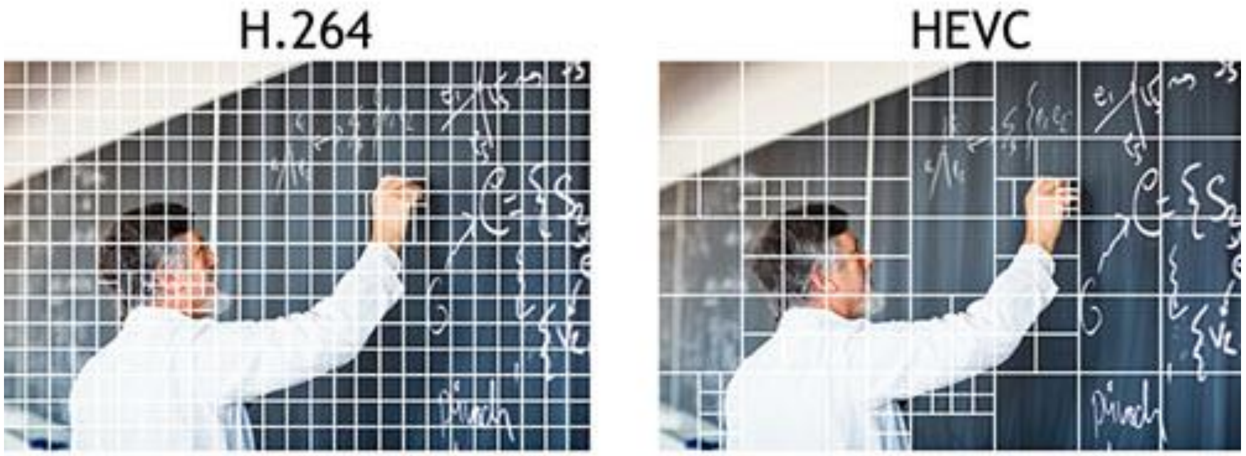
- H.264 itself saves 50% or more on bit rate compared with its predecessor MPEG-2 Part 2!!

# Comparisons



- H.264 vs. H.265

# H.264 vs H.265



- Blocking is content based
- H.265 is also called High Efficiency Video Coding (HEVC). As its name suggests, H.265 is twice high efficiency than H.264 when encoding video.

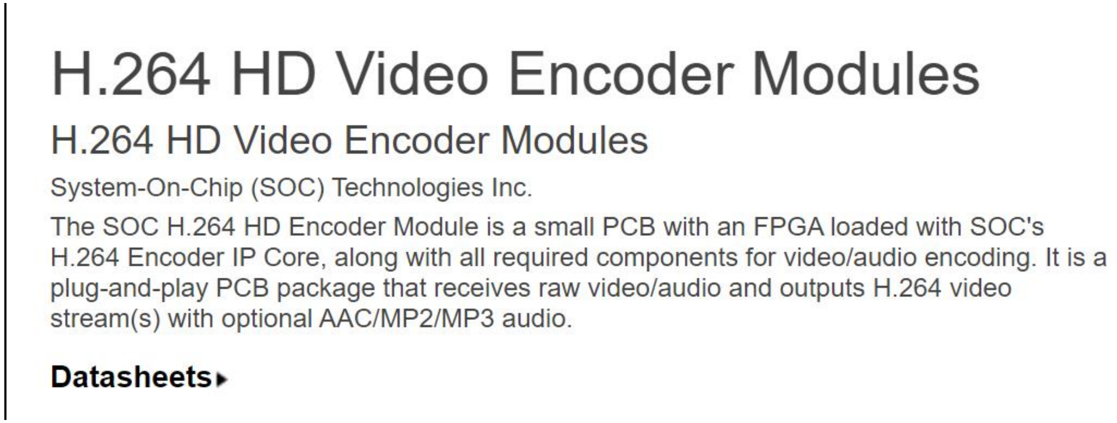
# H.264 codec chip in cameras

Complex enough to need a chip!

eASIC eDV9200 H.264 codec promises HD for all devices

- eASIC … is introducing a new H.264 codec aimed to bring high-def capabilities to all manners of devices, including (but certainly not limited to) toys, baby monitors, public transportation, wireless video surveillance and wireless webcams.
- The highly integrated eDV9200 is said to "dramatically lower the cost of entry into the high-definition video market..
- The chip -- which captures streaming data directly from a CMOS sensor, compresses it, and transfers it to a host system or to a variety of storage devices -- is priced at just $4.99 each in volume. HD oven timers, here we come!

# H.264



## H.264 HD Video Encoder Modules

### H.264 HD Video Encoder Modules

System-On-Chip (SOC) Technologies Inc.

The SOC H.264 HD Encoder Module is a small PCB with an FPGA loaded with SOC's H.264 Encoder IP Core, along with all required components for video/audio encoding. It is a plug-and-play PCB package that receives raw video/audio and outputs H.264 video stream(s) with optional AAC/MP2/MP3 audio.

**Datasheets ▸**

# Editing a video can cause artefacts

**We will consider**

1. Double/multiple compression

2. Video Intra-frame forgery detection techniques
   1. Also called spatial region tampering
   - **Region alteration** (Complex, resulting in content alteration over multiple frames without any frame displacement): example only
   - Interlacing artefact based techniques

3. Video Inter-frame forgery detection techniques
   - **Frame insertion** (also called region splicing)
   - **Frame deletion**
   - **Frame duplication** (also called region duplication or copy-paste)
   - **Frame shuffling**

# In the next lecture

- Digital multimedia: Recording and devices
  - Audio
  - Images
  - Video
  - Text
- Digital multimedia: Processing
  - Audio processing
  - Two generic processing techniques