Unit 5
Digital Video

CS 3570

Shang-Hong Lai

Video

- Video is made up of a series of still images (*frames*) played one after another at high speed.
- This fools the eye into believing that it is observing a continuous stream.

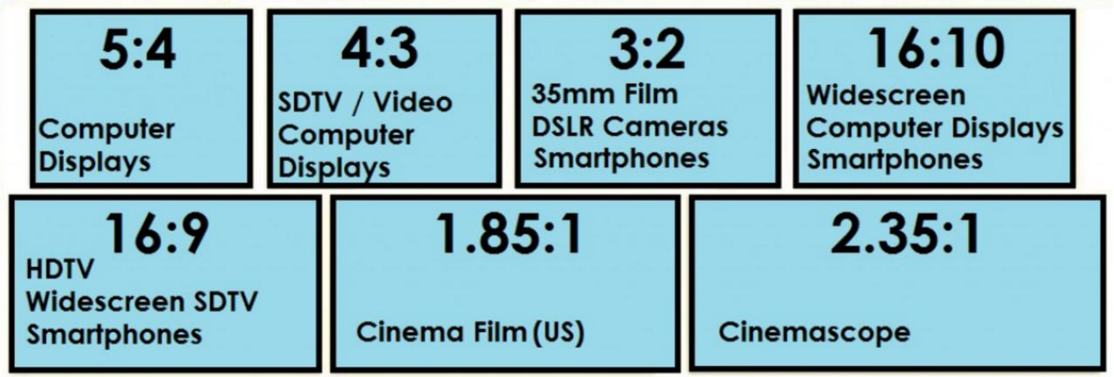
- Video (real-world pictures)
- Animation (Computer generated)

Video and Film

- ***Movie*** : a story told with moving images and sound.
- The word ***film*** seems to imply a movie that is shot and/or stored on cellulose film.
- Film and video both rest on the same phenomenon of human perception, called ***persistence of vision*** – the tendency of human vision to continue to “see” something for a short time after it is gone.
- A related physiological phenomenon is ***flicker fusion***—the human visual system’s ability to fuse successive images into one fluid moving image.

Standard film aspect ratios

- ***Aspect ratio*** is the ratio of the width to the height of a frame, expressed as *width : height*.
- Two common video aspect ratios are **4 : 3** (1.33 : 1), the universal video format in the 20th century, and **16 : 9** (1.77 : 1), universal for high-definition television.
- The most common aspect ratios used today in the presentation of films in cinemas are **1.85 : 1** and **2.35 : 1**



5:4
Computer Displays

4:3
SDTV / Video Computer Displays

3:2
35mm Film
DSLR Cameras
Smartphones

16:10
Widescreen Computer Displays
Smartphones

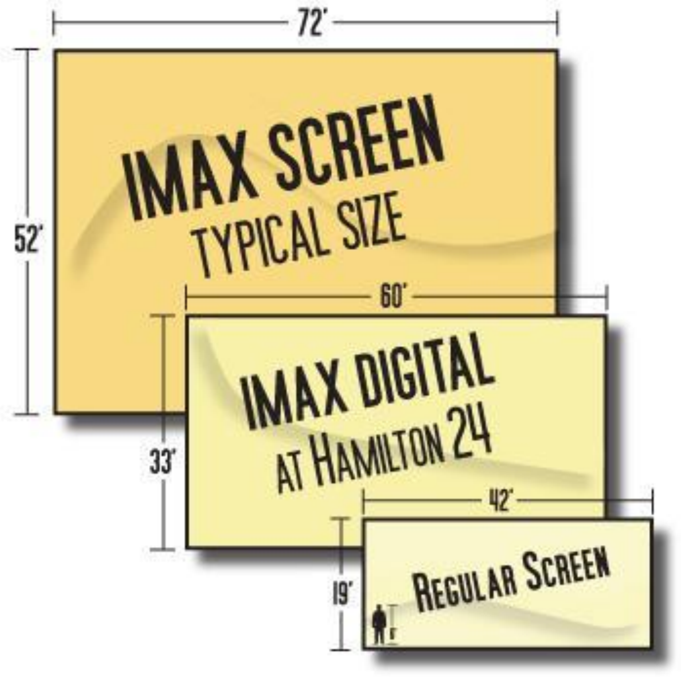
16:9
HDTV
Widescreen SDTV
Smartphones

1.85:1
Cinema Film (US)

2.35:1
Cinemascope

Film Development

- Silent movies and early sound movies were shot mostly on 16 mm film, introduced by Eastman Kodak in 1923.
- **IMAX**(Image MAXimum) is a system that displays images of greater size and resolution than conventional film systems.
    - IMAX movies are shot on 70 mm film with aspect ratio of 1.43:1.
    - IMAX movies are displayed on very large screens, e.g. a standard IMAX screen is 22 m × 16.1 m (72 ft × 52 ft).

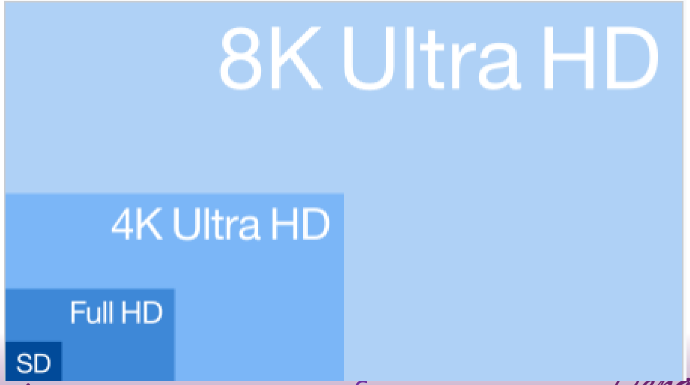


High-definition television

- In 1981, NHK began broadcasting what came to be known as ***high-definition television, HDTV***.
- The current definition of HDTV is television that has an aspect ratio of 16:9, surround sound, and one of three resolutions:
  1. 1920 × 1080 using interlaced scanning(**1080i**)
  2. 1920 × 1080 using progressive scanning(**1080p**)
  3. 1280 × 720 using progressive scanning (**720p**)
- Digital encoding is not part of this definition, and, historically, HDTV was not always digital.

Ultra High-Definition Television

- 4K UHD and 8K UHD, which are two digital video formats with an aspect ratio of 16:9.
- They were first proposed by NHK and later defined and approved by the International Telecommunication Union (ITU).
- Two resolutions are defined as UHDTV:
  - UHDTV-1 is 3840 pixels wide by 2160 pixels tall (8.3 megapixels), also known as 4K UHD
  - UHDTV-2 is 7680 pixels wide by 4320 pixels tall (33.18 megapixels), also referred to as 4320p and *8K UHD*.



Video and film displays

- Like film, video is created by a sequence of discrete images, called **frames**, shown in quick succession.
- Film is displayed at 24 frames/s. The standard frame rate for NTSC video is about 30 frames/s. The frame rate for PAL and SECAM video is 25 frames/s.
  - NTSC, PAL and SECAM are 3 main analog television systems
- A film frame is a continuous image. Video frames, in contrast, are divided into lines. Television has to be transmitted as a signal, line-by-line.
- Video is displayed (and recorded) by a process called ***raster scanning.*** The raster refers to a single frame.

Raster scanning

- For many years, the dominant video display technology was the *cathode ray tube* (*CRT*). Most television sets were built from CRTs, as were the computer monitors.

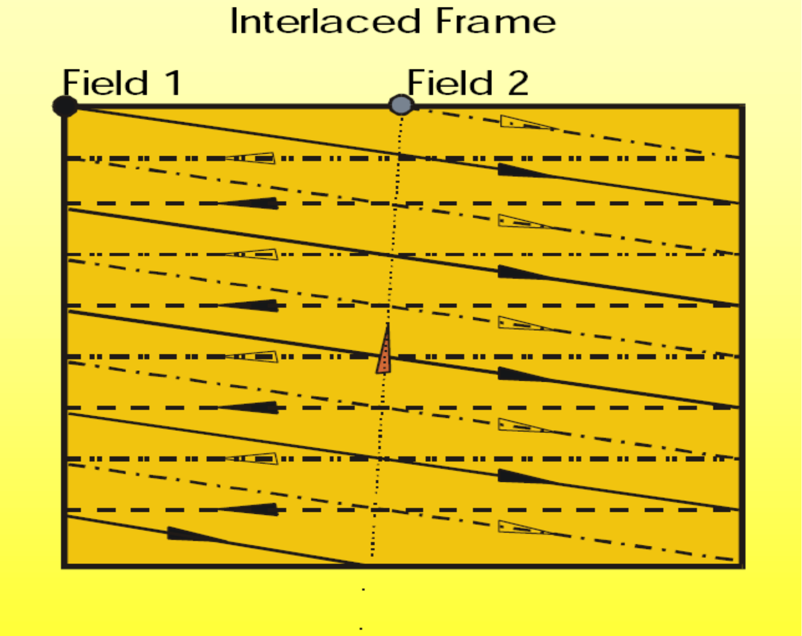
- Scanning can be done by one of two methods: either **interlaced** or **progressive scanning.**

- In ***interlaced scanning***, the lines of a frame are divided into two ***fields***: The odd-numbered lines, called the ***upper field (odd field)***, and the even-numbered lines, called the ***lower field (even field).***

- Video standards are sometimes described in terms of ***field rate*** rather than frame rate.
  - For PAL analog video, 50 fields/s = 25 frames/s

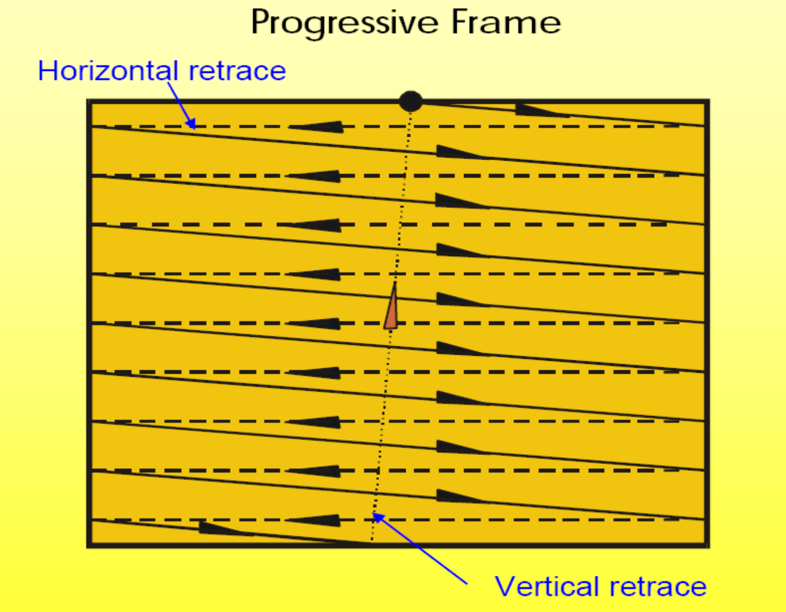
Interlaced Scanning

- In ***interlaced scanning***, the lines of a frame are divided into **two** *fields*:
  - The odd-numbered lines, called the ***upper field (odd field)***
  - The even-numbered lines, called the ***lower field (even field)***

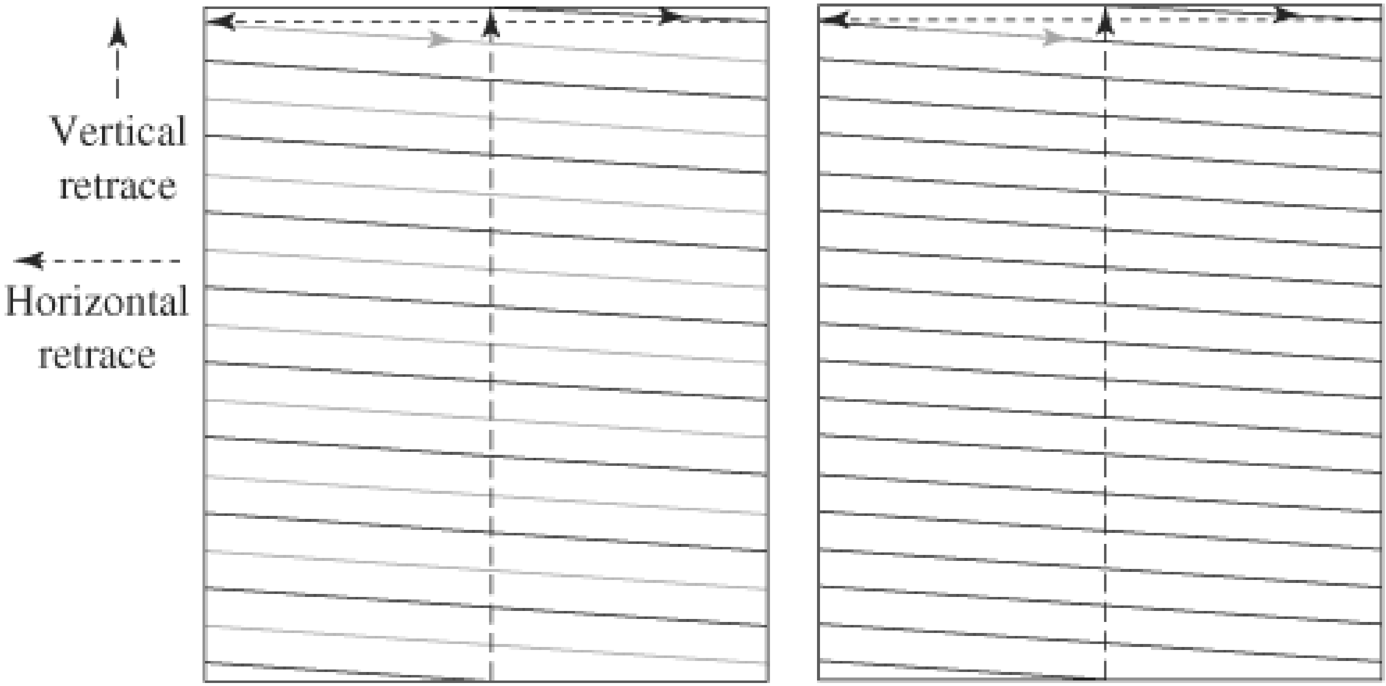


Progressive Scanning

- In ***progressive scanning***, each frame is scanned line-by-line from top to bottom.
- For progressive scanning, the frame rate and field rate are the same because a frame has only one field.
- Computer monitors and many digital televisions use progressive scanning.



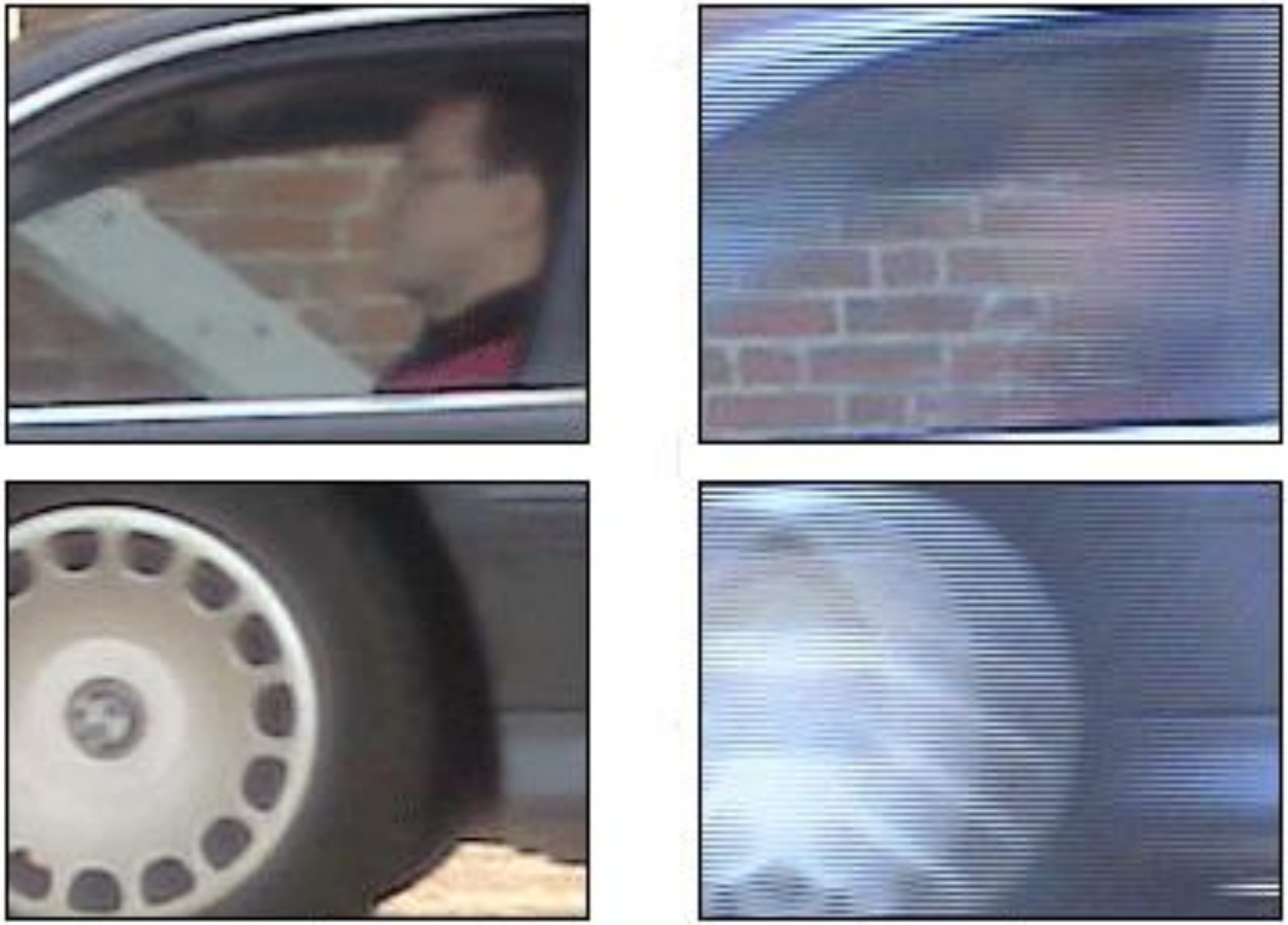
Interlaced and progressive scanning



Interlaced scanning:
Lower field (shown in gray) displayed first, one line at a time from top to bottom; then upper field (shown in black) displayed

Progressive scanning:
Lines displayed in order from top to bottom

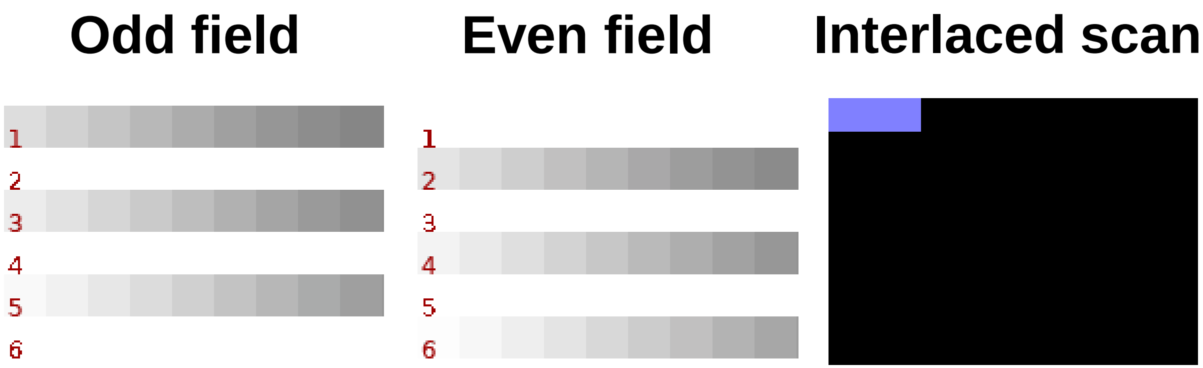
Interlaced and progressive scanning

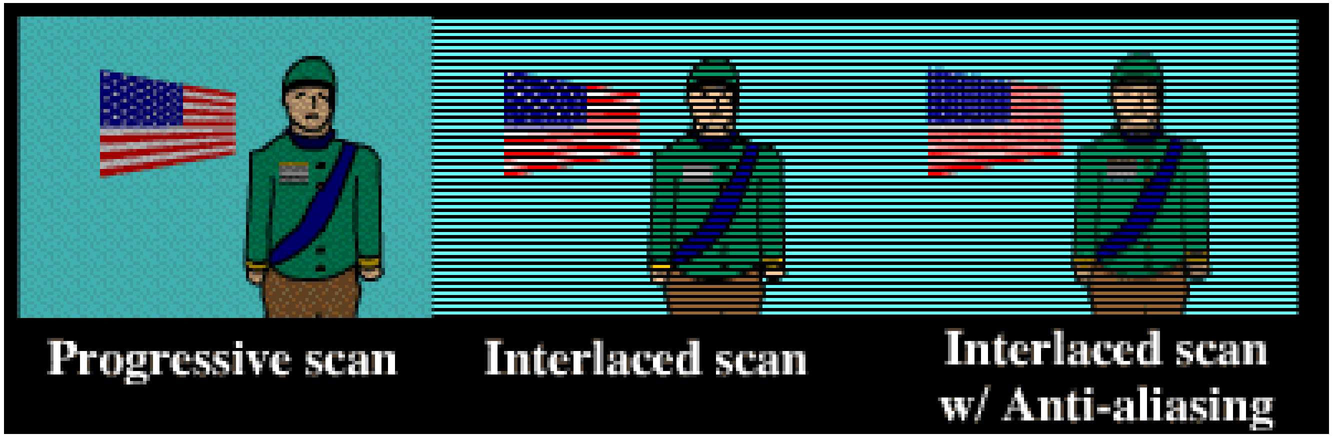


progressive scan

interlace

Interlaced scanning





Progressive vs. Interlaced Scan

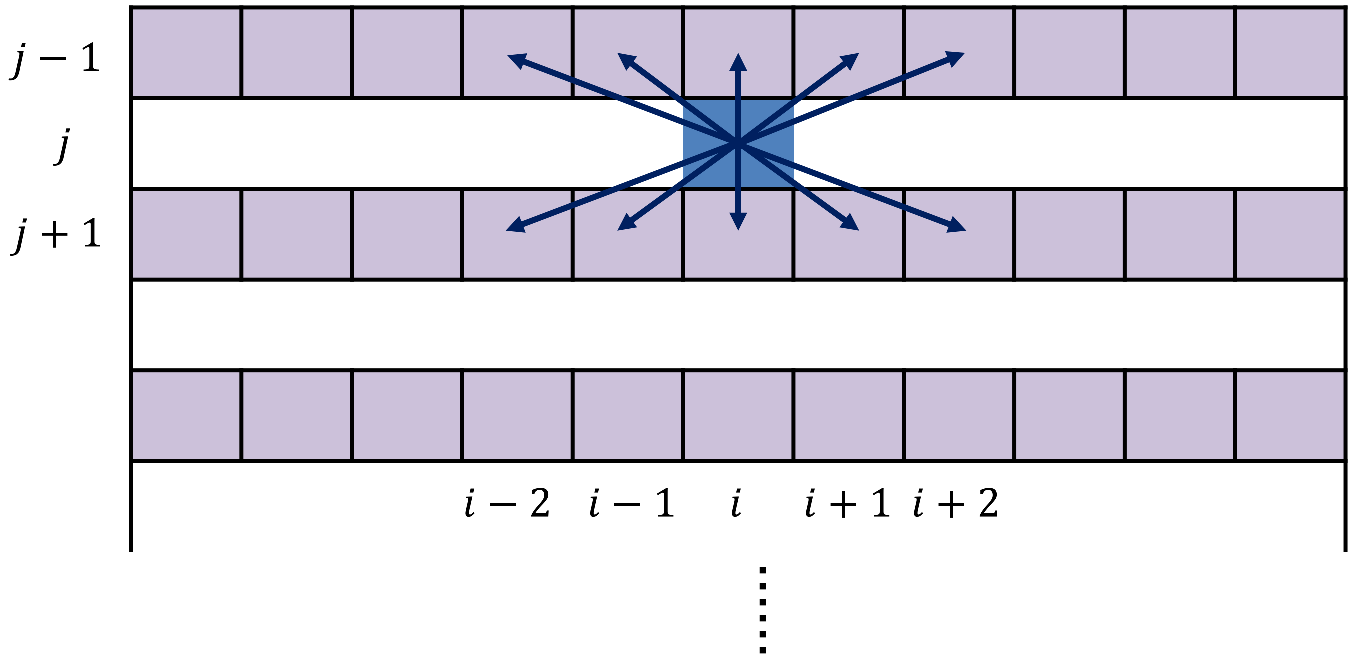
- Progressive
  - Computer monitors
  - Scans entire picture line by line
  - Eliminate flicker seen in interlaced
- Interlaced
  - Developed for **CRT** (Cathode Ray Tube) technology
  - Divides scans into odd and even lines
  - Alternately refreshes odd lines, then even lines
  - Slight delay between refreshes causes "jaggedness" or **interlace artifacts**
  - **Deinterlacing** can compensate somewhat

Deinterlacing

- Most modern displays only support progressive scanning.
- **Deinterlacing** is the process of converting interlaced video into progressive form.
- Deinterlacing methods:
  1. Intra-field interpolation deinterlacing
  2. Inter-field interpolation deinterlacing
  3. Motion adaptive deinterlacing
  4. Motion compensated deinterlacing

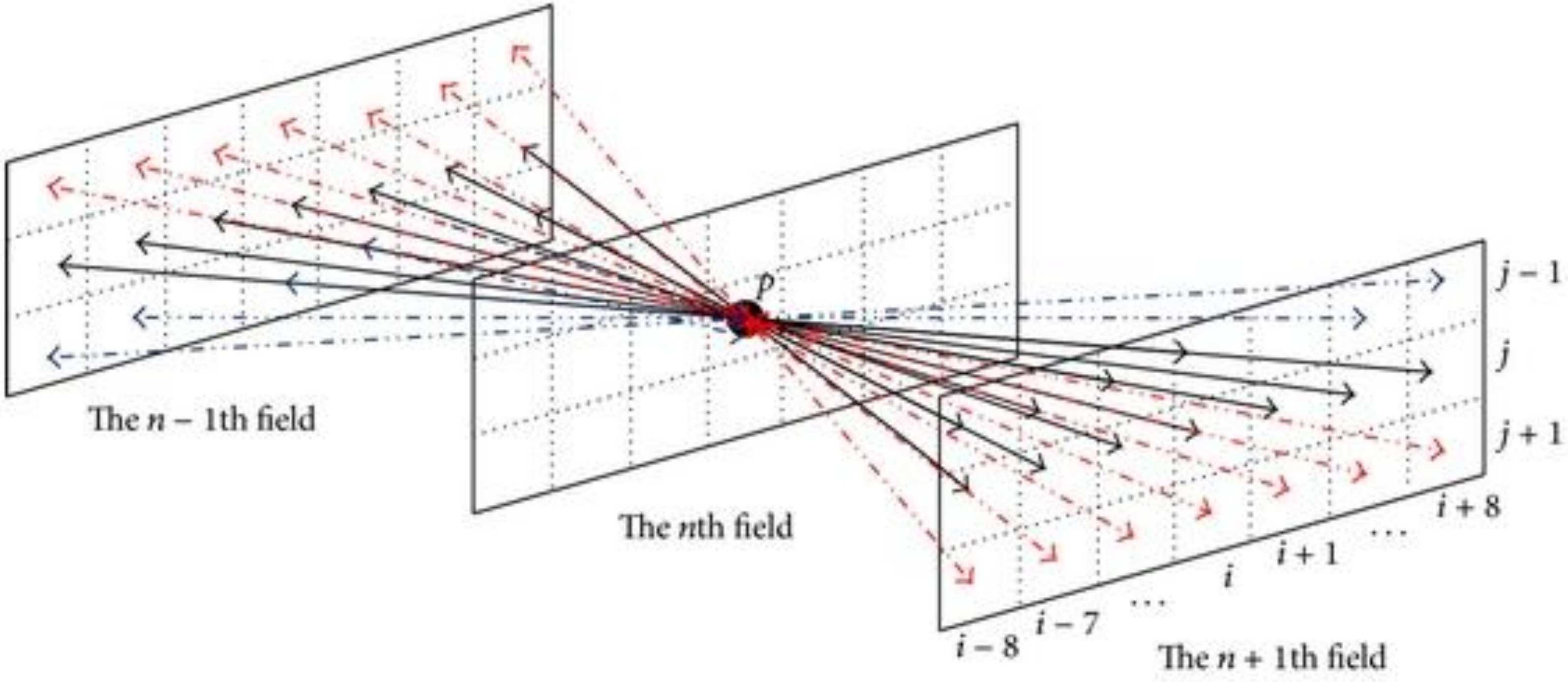
Deinterlacing

1. **Intra-field** interpolation deinterlacing



Deinterlacing

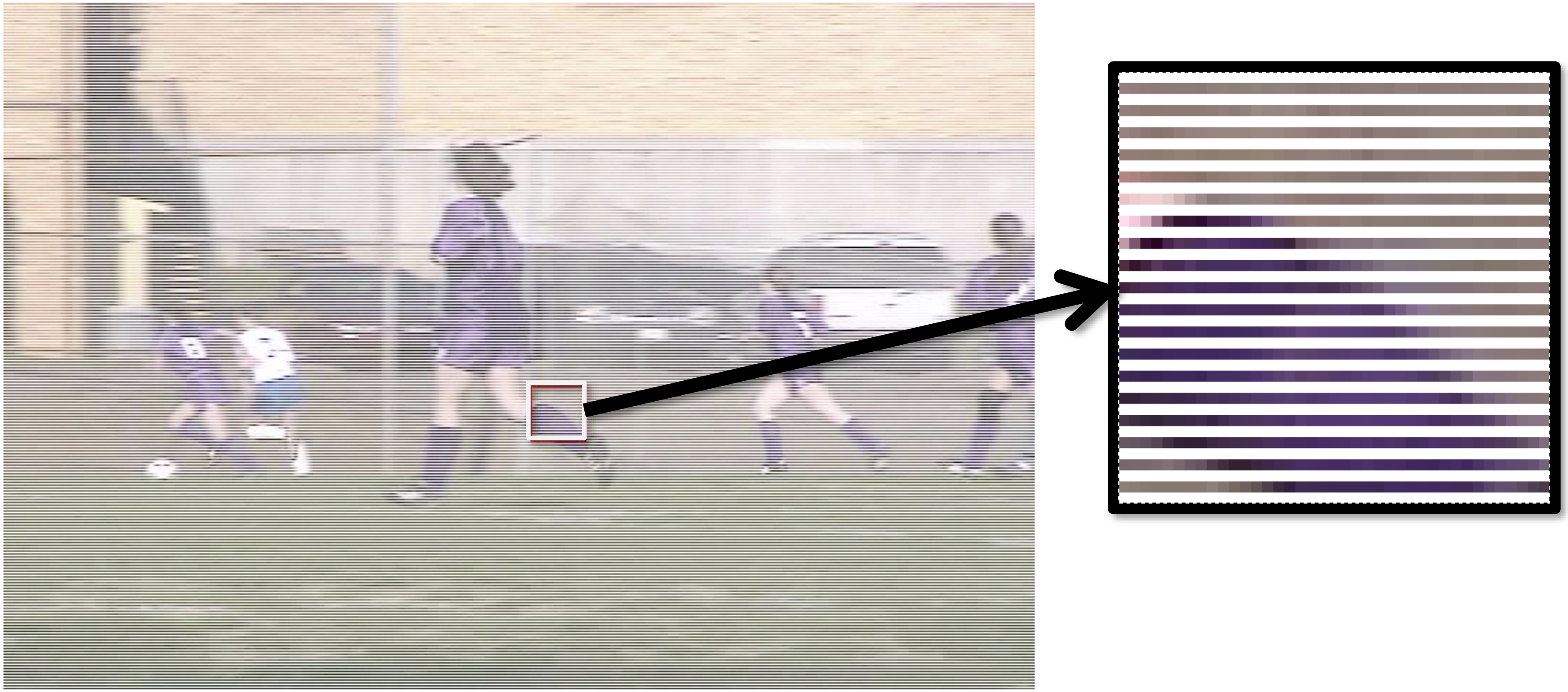
2. **Inter-field** interpolation deinterlacing



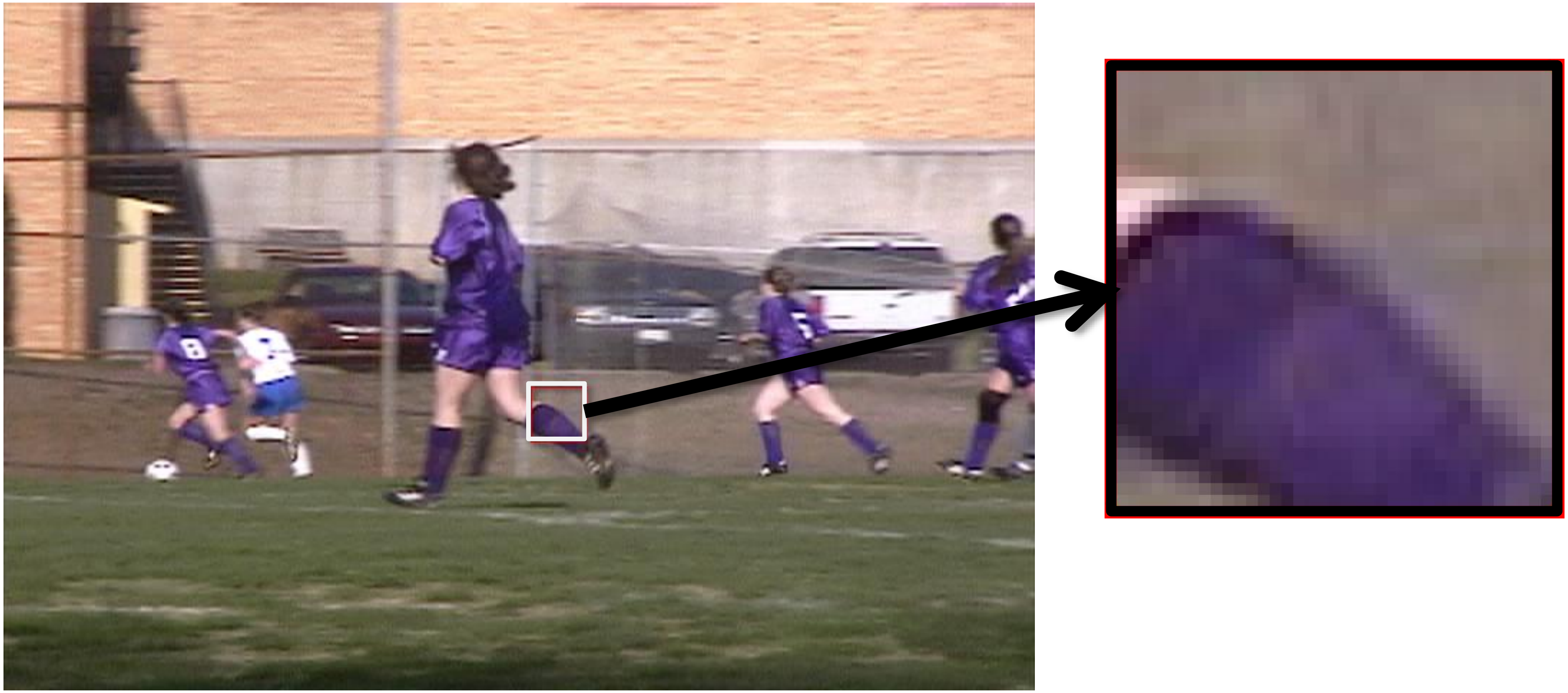
Deinterlacing

3. **Motion adaptive deinterlacing:** as areas that haven't changed from field to field don't need any processing, the regions apply intra-field interpolation. The other areas use inter-field interpolation.
4. **Motion compensated deinterlacing:** the exact original frames can be recovered by copying the missing field from a matching previous/next frame.

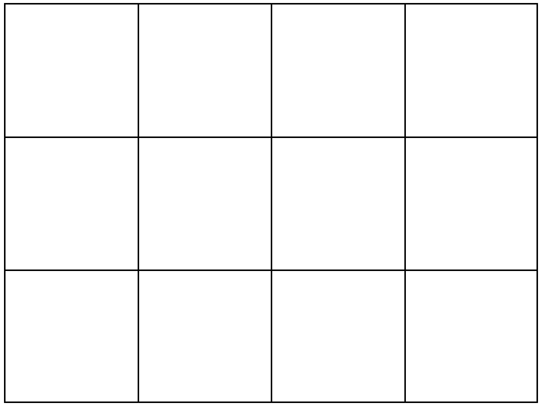
Interlacing



Deinterlaced Result



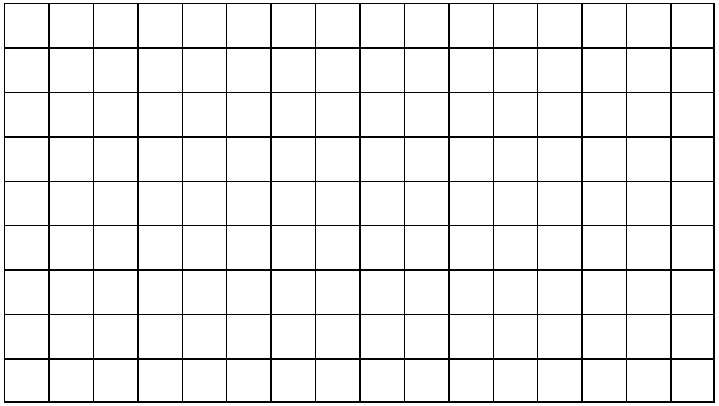
Frame Aspect Ratio Examples



4:3

Example:

- Standard definition NTSC standard format



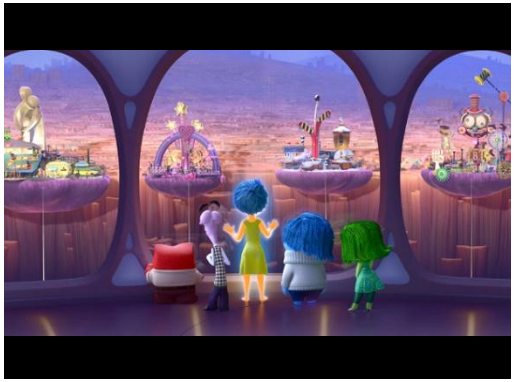
16:9

Examples:

- Standard definition NTSC wide-screen format
- High definition digital video
- High definition TV

Comparison of 4: 3 and 16: 9 image aspect ratio

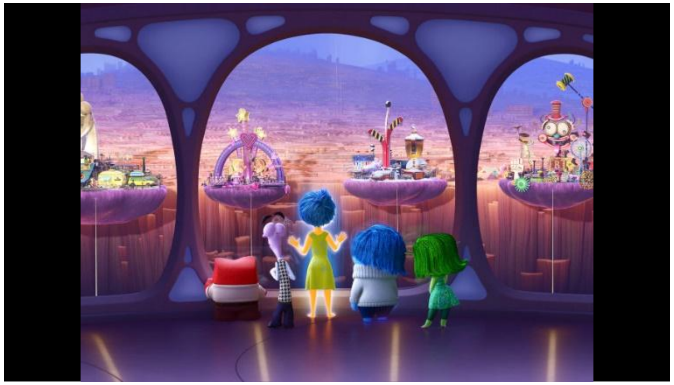
16:9 Aspect ratio



16:9 Aspect ratio displayed
on a 4:3 screen (letter box)

→ Letter box

4:3 Aspect ratio

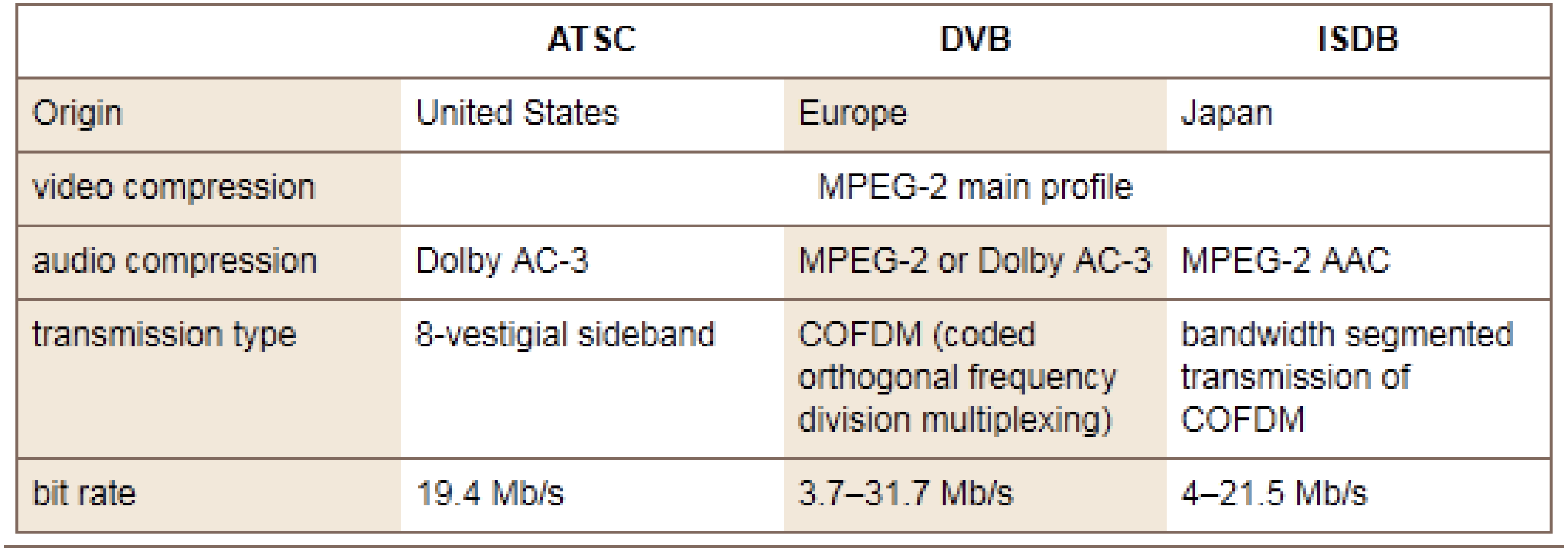


4:3 Aspect ratio displayed
on a 16:9 screen (pillar box)

→ Pillar box

Digital television (DTV)

- In the 1990s, the development of international standards for the transmission of ***digital television*** (***DTV***) became a hot topic.
- Three main standards organizations for DTV:



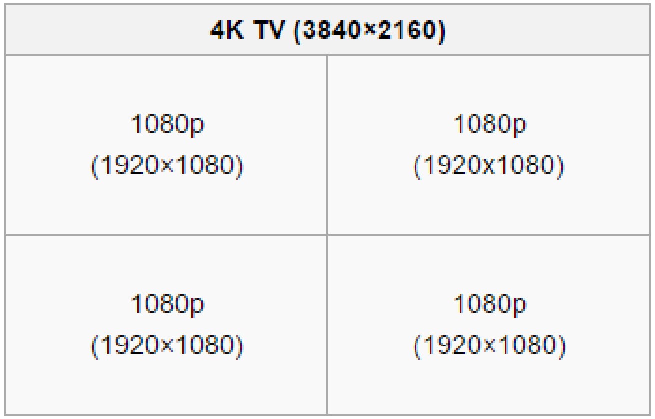
	ATSC	DVB	ISDB
Origin	United States	Europe	Japan
video compression	MPEG-2 main profile		
audio compression	Dolby AC-3	MPEG-2 or Dolby AC-3	MPEG-2 AAC
transmission type	8-vestigial sideband	COFDM (coded orthogonal frequency division multiplexing)	bandwidth segmented transmission of COFDM
bit rate	19.4 Mb/s	3.7–31.7 Mb/s	4–21.5 Mb/s

Standards for DTV

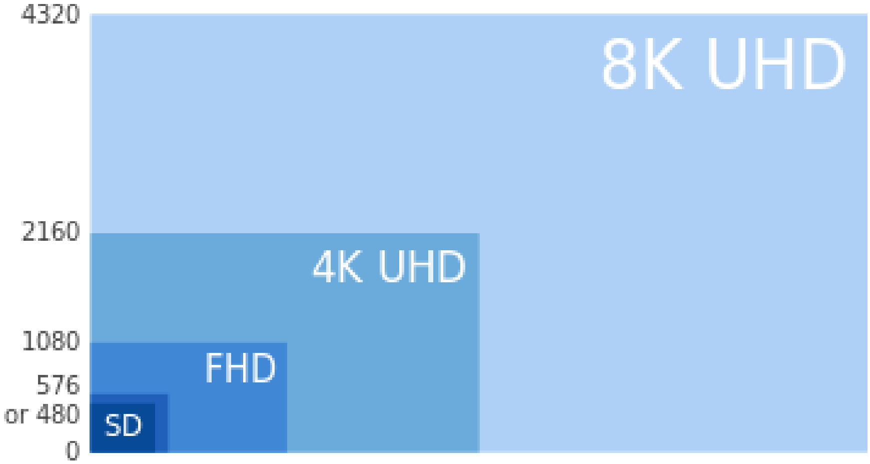
- ***ATSC*** (***Advanced Television Systems Committee***) is an international nonprofit organization that develops standards for digital television.
- ATSC developed DTV standards for the United States and Canada (**Taiwan** and south Korea have been adopted the standards).
- In Europe, standards for digital television were developed by ***DVB*** (***Digital Video Broadcasting Project***).
- DVB standards are divided into terrestrial (DVB-T), satellite (DVB-S), and handheld (DVB-H).
- Standards for digital video in Japan go by the name of ***ISDB*** (***Integrated Services Digital Broadcasting***).

Higher Resolution Nowadays

- There are 2 common standards on 4K videos:
  - A 4K resolution, as defined by Digital Cinema Initiatives, is 4096 x 2160 pixels(1.9:1 aspect ratio)
  - 4K Ultra HD(UHD) is 3840 x 2160 (1.78:1 aspect ratio)
- 8K UHD or FUHD (Full Ultra HD) is 7680 x 4230.



4K TV (3840×2160)	
1080p (1920×1080)	1080p (1920x1080)
1080p (1920×1080)	1080p (1920×1080)



Digital Video Distribution Media

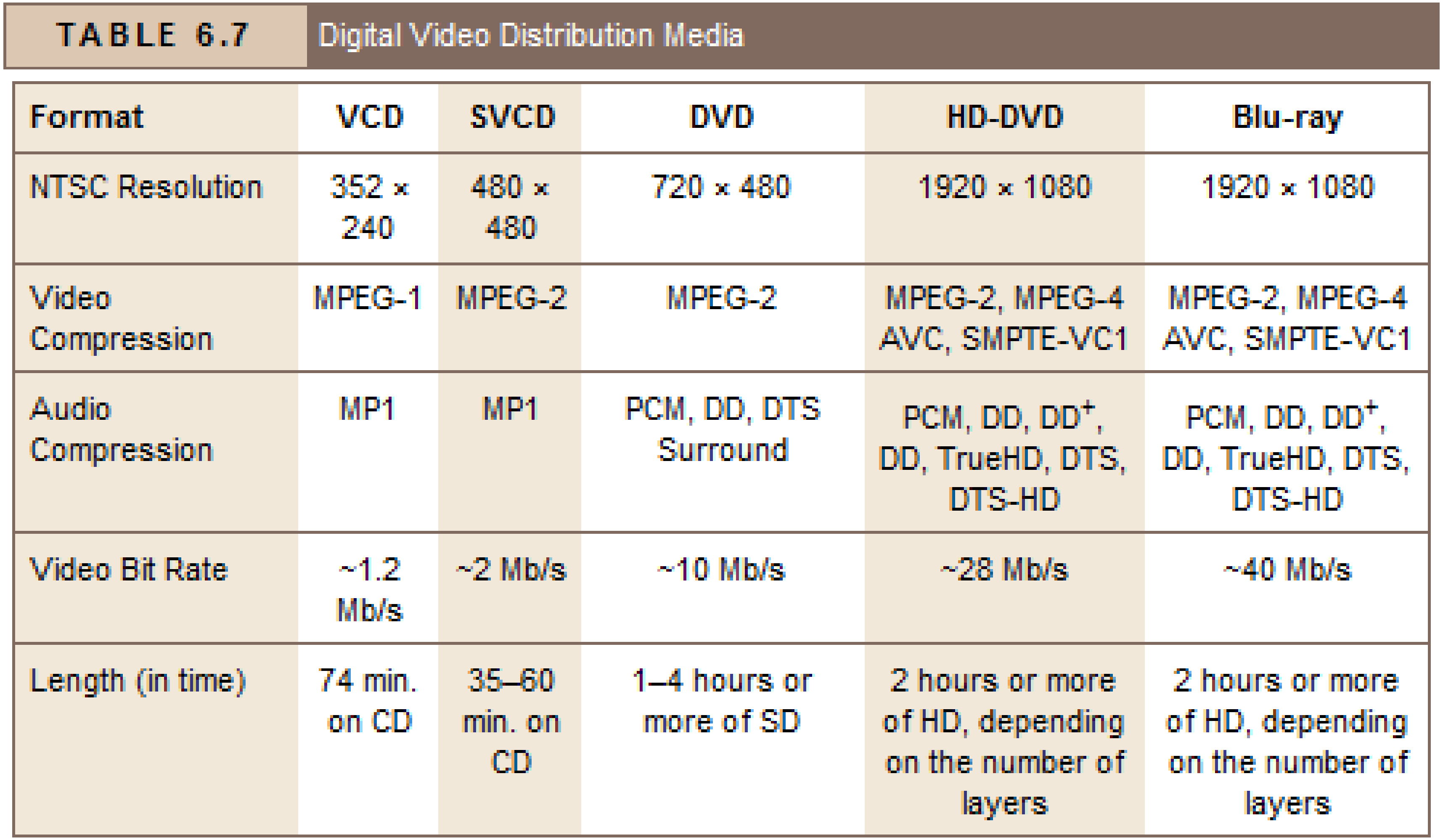


TABLE 6.7 Digital Video Distribution Media

Format	VCD	SVCD	DVD	HD-DVD	Blu-ray
NTSC Resolution	352 × 240	480 × 480	720 × 480	1920 × 1080	1920 × 1080
Video Compression	MPEG-1	MPEG-2	MPEG-2	MPEG-2, MPEG-4 AVC, SMPTE-VC1	MPEG-2, MPEG-4 AVC, SMPTE-VC1
Audio Compression	MP1	MP1	PCM, DD, DTS Surround	PCM, DD, DD$^+$, DD, TrueHD, DTS, DTS-HD	PCM, DD, DD$^+$, DD, TrueHD, DTS, DTS-HD
Video Bit Rate	~1.2 Mb/s	~2 Mb/s	~10 Mb/s	~28 Mb/s	~40 Mb/s
Length (in time)	74 min. on CD	35–60 min. on CD	1–4 hours or more of SD	2 hours or more of HD, depending on the number of layers	2 hours or more of HD, depending on the number of layers

Video Codecs

- Digital video files are very large. With no compression or subsampling, NTSC standard video would have a data rate of over 240 Mb/s; HD would have a data rate of about 1 Gb/s.
- Remove redundancies and extraneous information within one frame is called ***intraframe compression***. It also can be referred to as ***spatial compression.***
- There are two commonly used methods for accomplishing ***spatial compression***: transform encoding and vector quantization.
- ***Temporal compression*** is a matter of eliminating redundant or unnecessary information by considering how images change over time. it is also called ***interframe compression***.

Video Codecs

- The basic method for compressing between frames is to detect how objects move from one frame to another, represent this as a vector.
- Determining the motion vector is done by a method called ***motion estimation.***
- Some codecs allow you to select either ***constant*** or ***variable bit rate encoding*** (***CBR*** and ***VBR***, respectively). Variable bit rate varies the bit rate according to how much motion is in a scene.
- Codecs are mostly ***asymmetrical***. This means that the time needed for compression is not the same as the time needed for decompression.

Compression Standards

Year	Standard	Publisher	Applications
1984	H.120	ITU-T	
1990	H.261	ITU-T	Videoconferencing Videotelephony
1993	MPEG-1 Part 2	ISO, IEC	Video-CD
1995	H.262/MPEG-2 Part 2	ISO, IEC, ITU-T	DVD Video, Blu-ray, Digital Video Broadcasting, SVCD
1996	H.263	ITU-T	Videoconferencing, videotelephony, video on mobile phones (3GP)
1999	MPEG-4 Part 2	ISO, IEC	Video on Internet (DivX, Xvid ...)
2003	H.264/MPEG-4 AVC	Sony, Panasonic, Samsung, ISO, IEC, ITU-T	Blu-ray, HD DVD, Digital Video Broadcasting, iPod Video, Apple TV, videoconferencing
2013	H.265	ISO, IEC, ITU-T	4K Blu-ray Disc, Nvidia Graphic Cards, Windows 10, iOS 11, …

Bitrate

- Bit rate often refers to the **number of bits used per unit of playback time** to represent a continuous medium such as audio or video after data compression.
- The encoding bit rate of a multimedia file is the size of a multimedia file in bytes divided by the playback time of the recording (in seconds), multiplied by eight.

編碼率期望越小越好

Peak Signal-to-Noise Ratio

- Decibels - a dimensionless unit
  - that is used to describe the relative power or intensity of two phenomena.
- The definition of decibel (dB) is:
  - $1\,dB = 10\log_{10}\left(\frac{I}{I_0}\right)$, *I* and *I₀* are the intensities (power) of two signals
- Peak Signal-to-Noise Ratio

$$MSE = \frac{1}{m\,n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2$$

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

$$= 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)$$

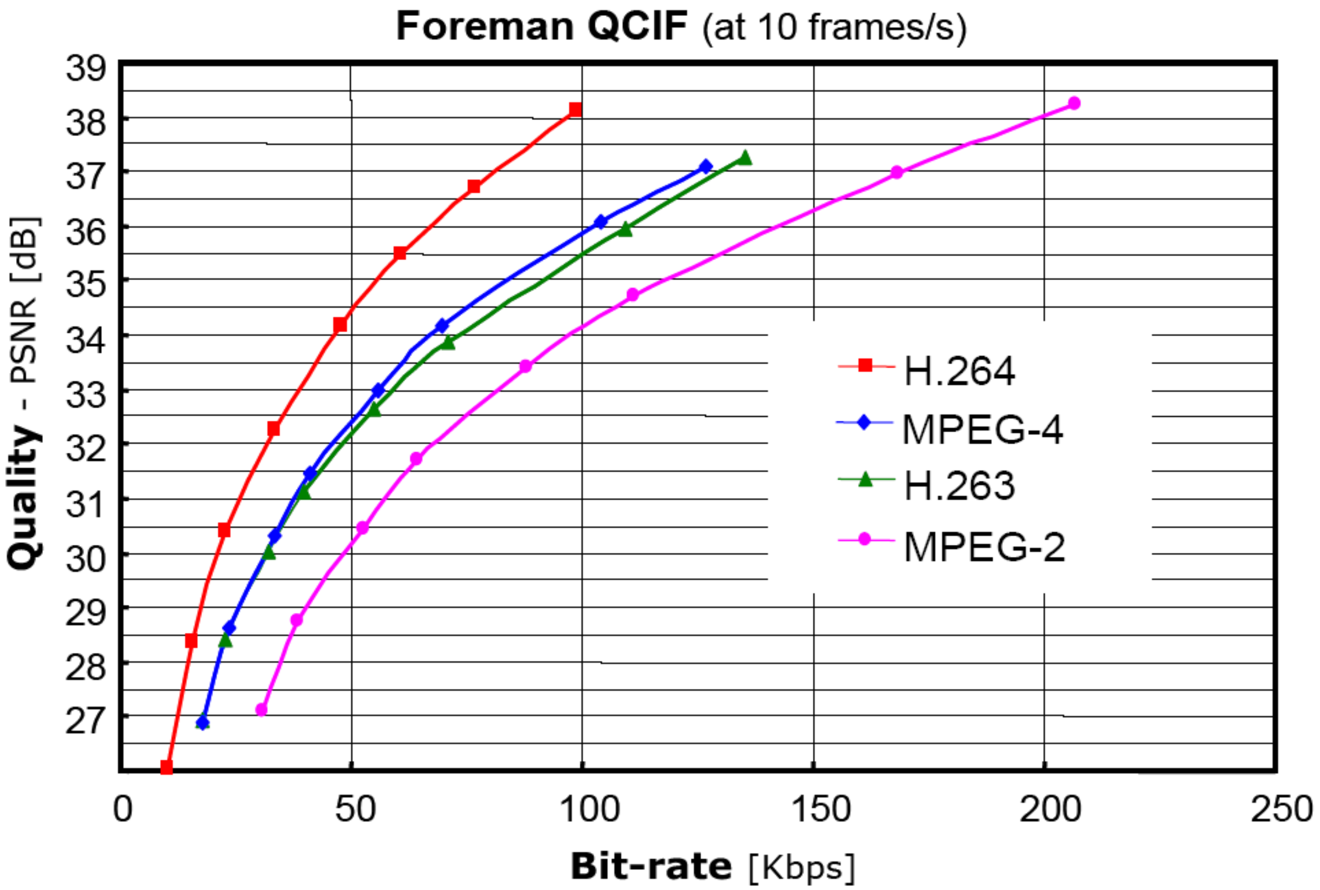
$$= 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE)$$

In data compression,
I(i,j) : the original image
K(i,j) : the compressed image

MAX is set to 255 if image is represented by 8 bit

Comparison of Different Standards

Rate-Distortion Curve



MPEG compression

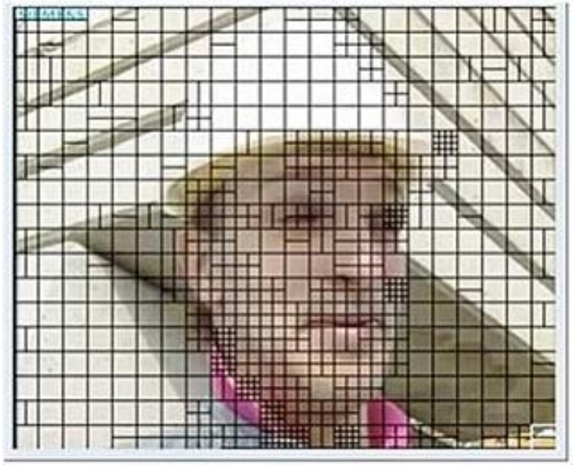
- MPEG compression was developed in two lines.
  - The first was the work of ITU-T and their subcommittee, the **Video Coding Experts Group**. We know this line of codecs as the H.26* series
  - The second line emerged from the **Motion Picture Experts Group**, from which we get the name MPEG
- The revolutionary advance in MPEG-4 compression is the use of object-based coding.
- ***MPEG-4 AVC*** (Advanced Video Coding) and equivalent to ***H.264***, is an improved MPEG-4 version introduced in 2003 that quickly achieved wide adoption for DVD; videoconferencing; videophone…

MPEG General Information

- Goal: data compression 1.5 Mbps
- MPEG defines video, audio coding and system data streams with synchronization.
- MPEG information
  - Aspect ratios: 1:1 (CRT), 4:3 (NTSC), 16:9 (HDTV)
  - Refresh frequencies: 23.975, 24, 25, 29.97, 50, 59.94, 60 Hz

MPEG Image Preparation - Blocks

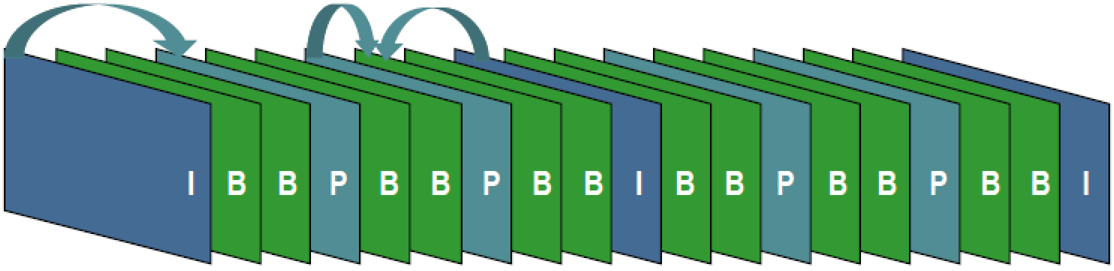
- Each image is divided into **macro-blocks.**
- Macro-block : 16x16 pixels for luminance; 8x8 for each chrominance component.
- Macro-blocks are useful for Motion Estimation.
- No MCUs which implies sequential non-interleaving order of pixels values.



H.264

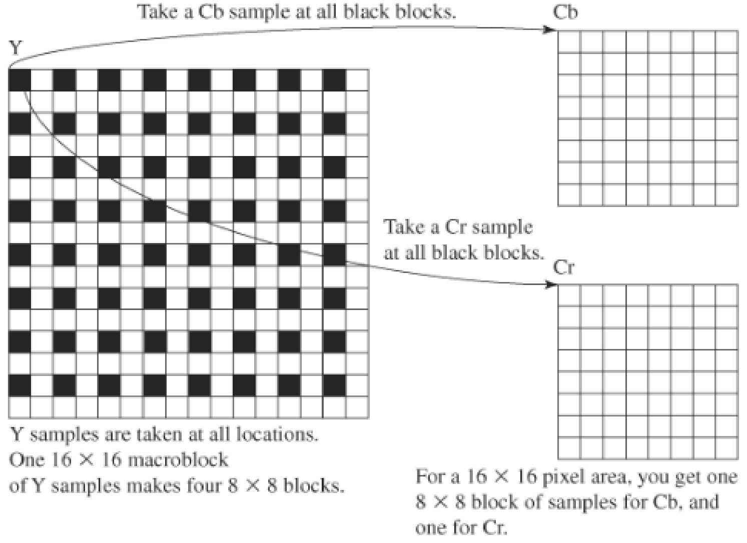
Main steps of MPEG compression

- **Step 1.** Divide the sequence of frames into GOPs, identifying I, P, and B frames.
  - A ***GOP*** is a ***group of pictures***, that is, a group of *n* sequential video frames
  - ***I frames***, or ***intraframes***, are compressed independently, as if they were isolated still images, using JPEG compression.
  - I frames serve as reference points for the ***P frames*** (***interframes***, also called ***forward prediction frames***) and ***B frames*** (***bidirectional frames***), which are compressed both spatially and temporally.



Main steps of MPEG compression

- **Step 2.** Divide each frame into macroblocks.
  - A ***macroblock*** is a 16 × 16 pixel area.
  - A 16 × 16 macroblock can be divided into 8 × 8 blocks. The way that macroblocks are divided depends on the particular compression standard, which can apply different types of chrominance subsampling.



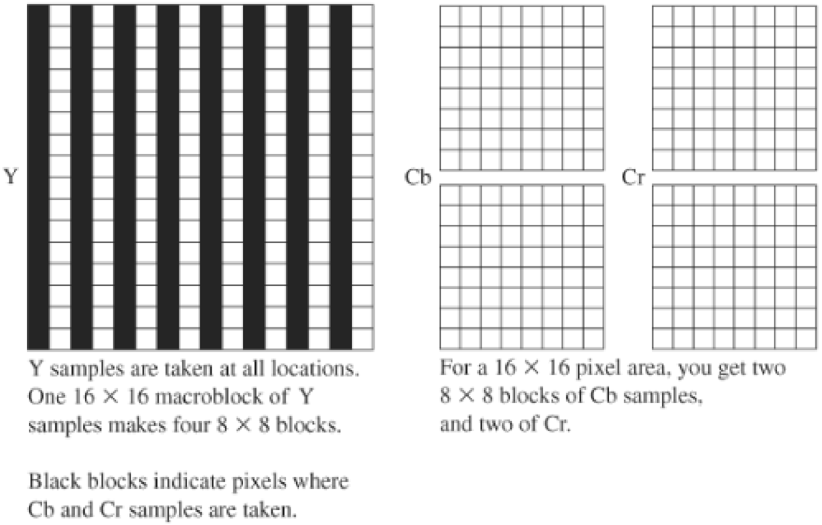
Take a Cb sample at all black blocks.

Take a Cr sample at all black blocks.

Y samples are taken at all locations. One 16 × 16 macroblock of Y samples makes four 8 × 8 blocks.

For a 16 × 16 pixel area, you get one 8 × 8 block of samples for Cb, and one for Cr.

Figure 7.22 Macroblock for 4: 2: 0 subsampling



Y samples are taken at all locations. One 16 × 16 macroblock of Y samples makes four 8 × 8 blocks.

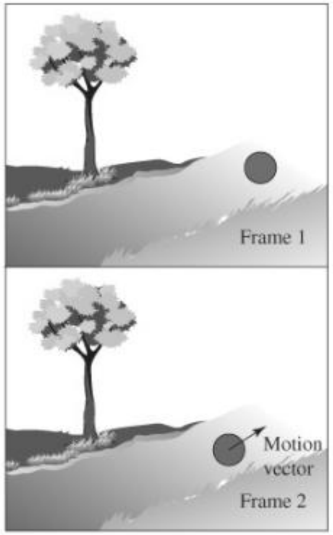
For a 16 × 16 pixel area, you get two 8 × 8 blocks of Cb samples, and two of Cr.

Black blocks indicate pixels where Cb and Cr samples are taken.

Figure 7.23 Macroblock for 4: 2: 2 subsampling

Main steps of MPEG compression

- **Steps 3 and 4.** For each P and B frame, compare the frame to the related I frame to determine a motion vector. Record differential values for P and B frames.
  - This step is called ***motion estimation.***
  - It's more economical to convey the difference between one frame and the next, a method called ***differential encoding.***
  - Motion estimation determines how much a frame has "moved" since the previous frame.
  - The difference between the macroblock in frame 2 and the matching macroblock in frame 1 is called the **prediction error**.

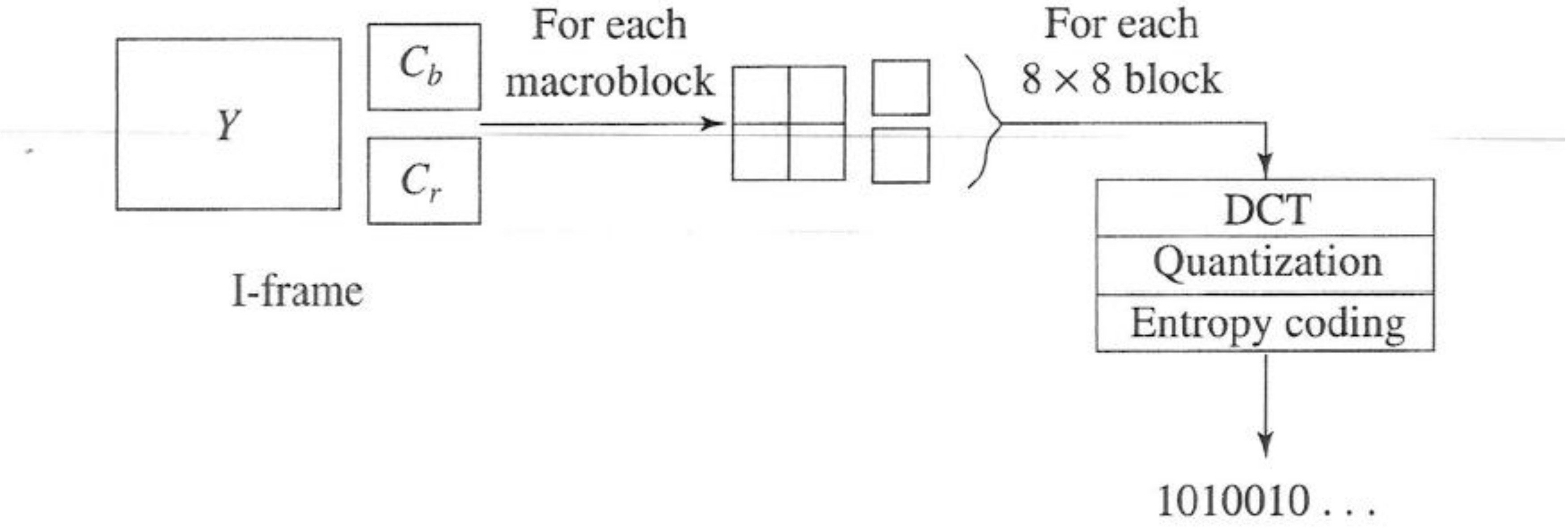


Main steps of MPEG compression

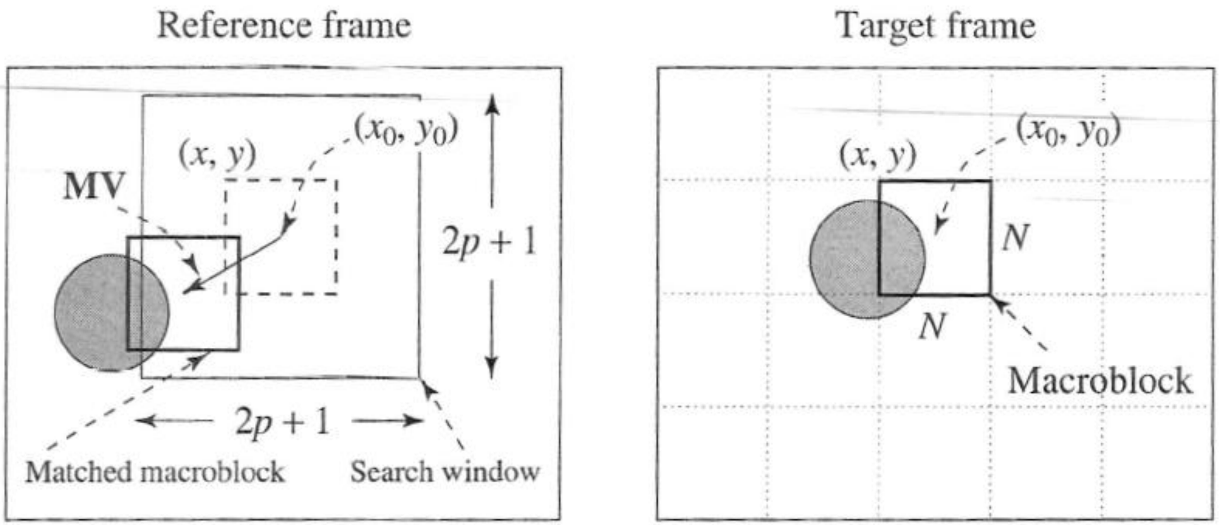
- The P or B frame being compressed is called the ***target frame***, and the I frame is named ***reference frame***.
- The reference frame to which a P frame is compared is called its ***forward prediction frame***. The reference frame to which a B frame is compared is called its ***backward prediction frame.***
- Assume we have a macroblock in the target frame T. We will search for a matching macroblock in reference frame R. We want to look in the vicinity of $\boldsymbol{R}_{x,y}$ for the macroblock that most closely matches $\boldsymbol{T}_{x,y}$.

Intra-frame coding

- Various lossless and lossy compression techniques use – like JPEG

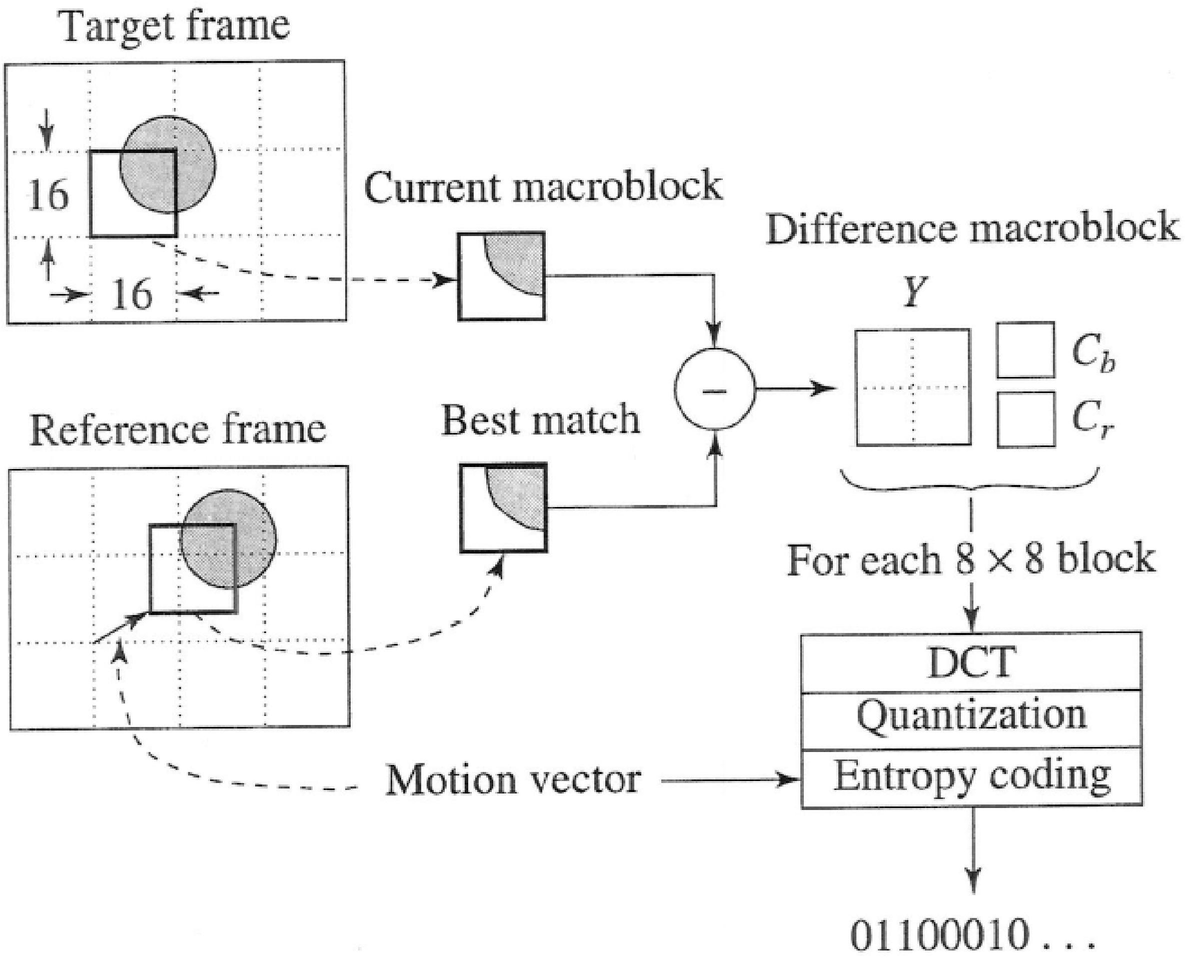


Motion Compensation



In H.261, motion vectors are in the range [-15,15]x[-15,15], e.g, p = 15.

MPEG Video for P-Frames



Estimating Motion Vectors

- Basic idea to find motion vectors is to search macroblocks
    - Within a $\pm p \times p$ pixel search window
    - Calculate **Sum of Absolute Difference** (SAD) of each macroblock(or **Mean Absolute Error** (MAE))
    - Choose macroblock which SAD/MAE is a minimum
- If the encoder decides that no acceptable match exists then
    - Coding that macroblock as an intra macroblock
    - In this manner, high quality video is maintained at a slight cost to coding efficiency

Sum of Absolute Difference (SAD)

- SAD is computed by

$$SAD(\boldsymbol{i},\boldsymbol{j}) = \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}|T(x+k,y+l) - R(x+k+\boldsymbol{i},y+l+\boldsymbol{j})|$$

- N is size of macroblock window typically (16 or 32 pixels)
- (x,y) the position of the **target** macroblock T, and R is the **reference** region to compute the SAD.
- $T(x+k,y+l)$ — pixels in the macroblock with upper left corner (x,y) in the target
- $R(x+k+i,y+l+j)$ — pixels in the macroblock with upper left corner $(x+i,y+j)$ in the reference

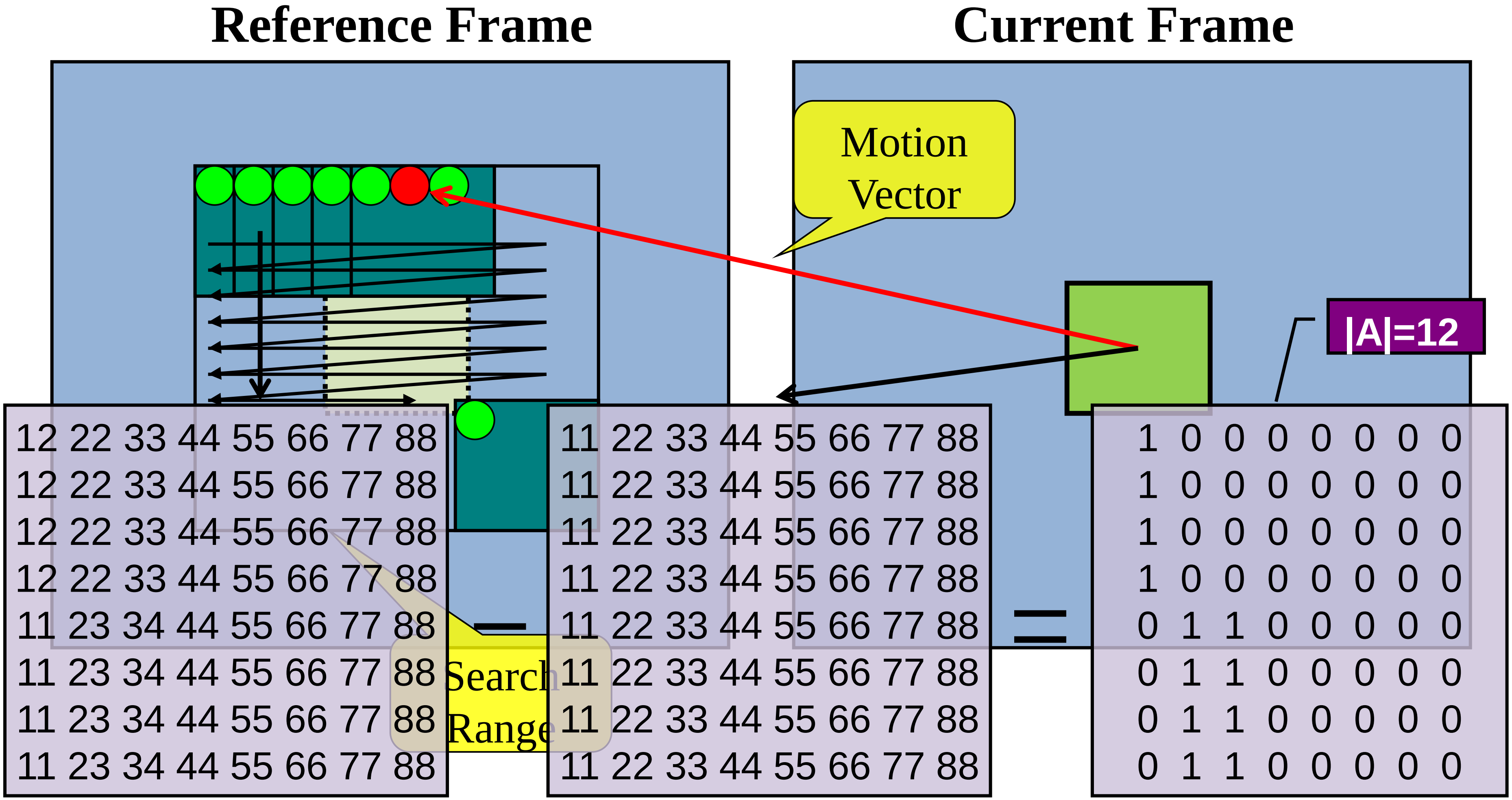
Sum Square Differences (SSD)

- Alternatively: sum of squared differences (SSD)

$$SSD(\boldsymbol{i},\boldsymbol{j}) = \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\Big(\big(T(x+k,y+l) - R(x+k+\boldsymbol{i},y+l+\boldsymbol{j})\big)\Big)^2$$

- Goal is to find a vector (i,j) such that SAD/SSD (i,j) is minimum

Exhaustive Block-Matching Algorithm



$$\begin{matrix}12 & 22 & 33 & 44 & 55 & 66 & 77 & 88\\ 12 & 22 & 33 & 44 & 55 & 66 & 77 & 88\\ 12 & 22 & 33 & 44 & 55 & 66 & 77 & 88\\ 12 & 22 & 33 & 44 & 55 & 66 & 77 & 88\\ 11 & 23 & 34 & 44 & 55 & 66 & 77 & 88\\ 11 & 23 & 34 & 44 & 55 & 66 & 77 & 88\\ 11 & 23 & 34 & 44 & 55 & 66 & 77 & 88\\ 11 & 23 & 34 & 44 & 55 & 66 & 77 & 88\end{matrix} - \begin{matrix}11 & 22 & 33 & 44 & 55 & 66 & 77 & 88\\ 11 & 22 & 33 & 44 & 55 & 66 & 77 & 88\\ 11 & 22 & 33 & 44 & 55 & 66 & 77 & 88\\ 11 & 22 & 33 & 44 & 55 & 66 & 77 & 88\\ 11 & 22 & 33 & 44 & 55 & 66 & 77 & 88\\ 11 & 22 & 33 & 44 & 55 & 66 & 77 & 88\\ 11 & 22 & 33 & 44 & 55 & 66 & 77 & 88\\ 11 & 22 & 33 & 44 & 55 & 66 & 77 & 88\end{matrix} = \begin{matrix}1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0\end{matrix}$$

Fast Block-Matching Algorithms

- **The characteristics of fast algorithm**
  - Not as accurate as exhaustive search algorithm
  - Save large amount of computation
- **fast motion estimation algorithm**
  - 2-D logarithm search method
  - 3-step search algorithm

2-D logarithm Search Method

- The search is accomplished by successively **reducing the area of search**.
- Each step consists of searching five locations which contain the center of the area, and the midpoints between the center and the four boundaries of the area along the axes passing through the center.
- In the final step all the nine locations are searched and the location corresponding to the minimum is the *direction of minimum distortion*.

J.R. Jain and A.K. Jain, “Displacement measurement and its application in interframe image coding,” *IEEE Trans. Commun.*, Vol. COM-29, pp. 1799–1808, Dec. 1981

The algorithm is given below.

For any integer $m > 0$, we define

$$N(m) = \{(i, j); \quad -m \leqslant i, j \leqslant m\}$$

$$M(m) = \{(0, 0), (m, 0), (0, m), (-m, 0), (0, -m)\}.$$

A 2-D Logarithmic Search Procedure for DMD:

Step 1: (initialization)

Error function

$$D(i, j) = \infty \qquad (i, j) \notin N(p)$$

P is search range

$$n' = \lfloor \log_2 p \rfloor$$

$$n = \max \cdot \{2, 2^{n'-1}\}$$

$$q = l = 0 \text{ (or an initial guess for DMD)}$$

where $\lfloor \cdot \rfloor$ is a lower integer truncation function.

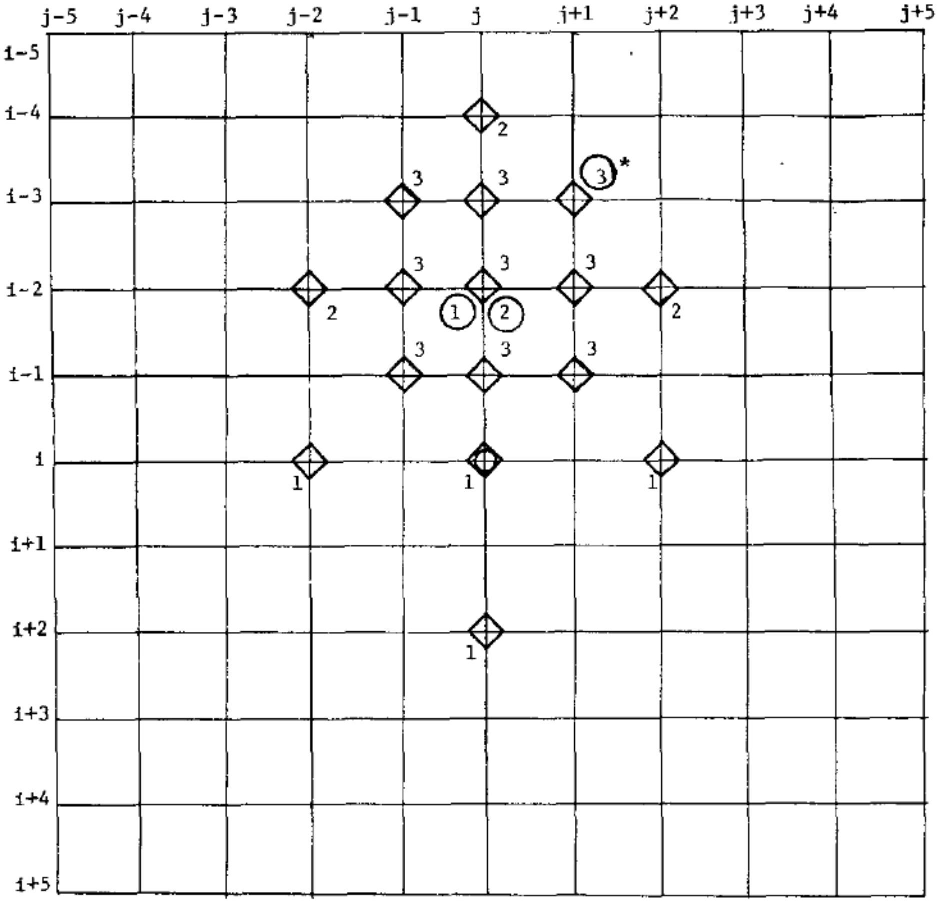
Step 2: $M'(n) \leftarrow M(n)$.

Step 3: Find $(i, j) \in M'(n)$ such that $D(i + q, j + l)$ is minimum. If $i = 0$ and $j = 0$, go to Step 5; otherwise go to Step 4.

Step 4: $q \leftarrow q + i, l \leftarrow l + j$; $M'(n) \leftarrow M'(n) - (-i, -j)$; go to Step 3.

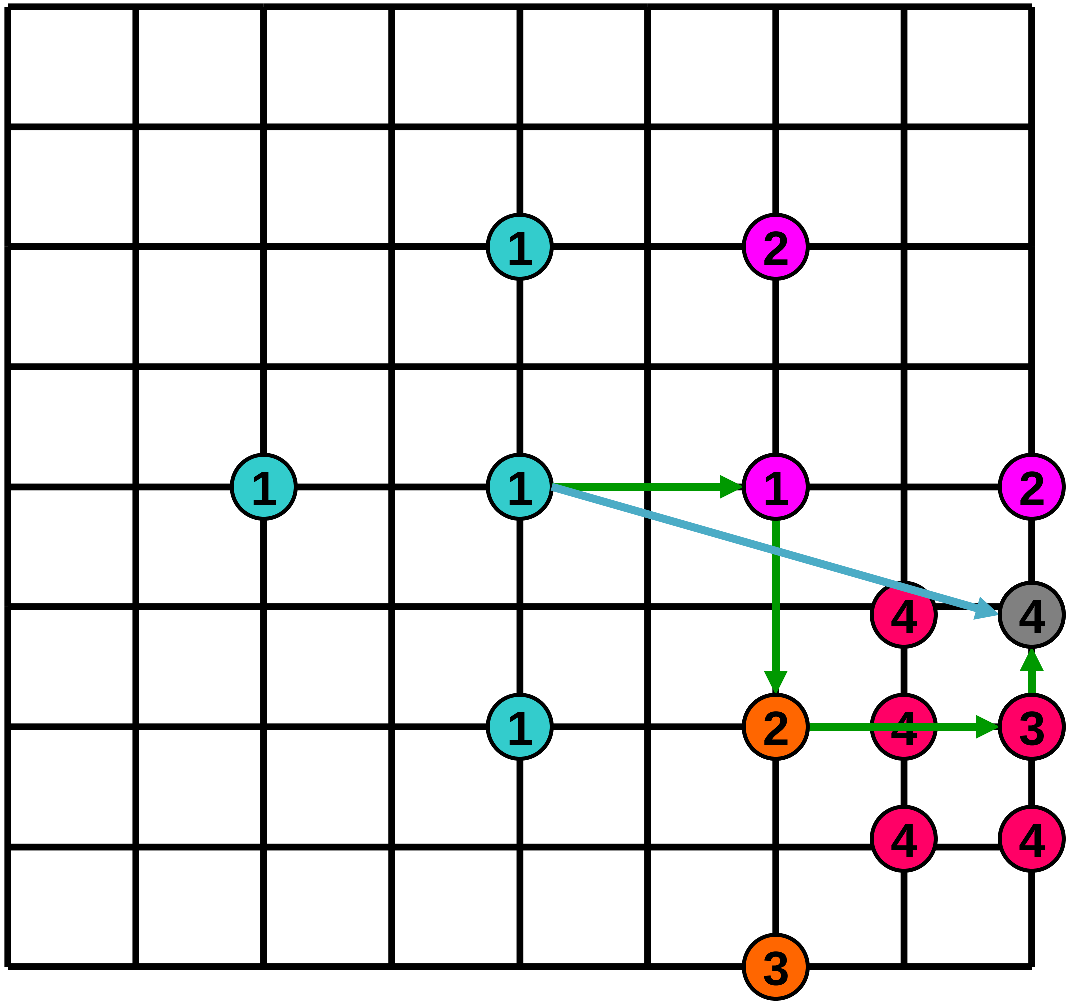
Step 5: $n \leftarrow n/2$. If $n = 1$, go to Step 6; otherwise, go to Step 2.

Step 6: Find $(i, j) \in N(1)$ such that $D(i + q, j + l)$ is minimum: $q \leftarrow i + q, l \leftarrow l + j$. (q, l) then gives the DMD.



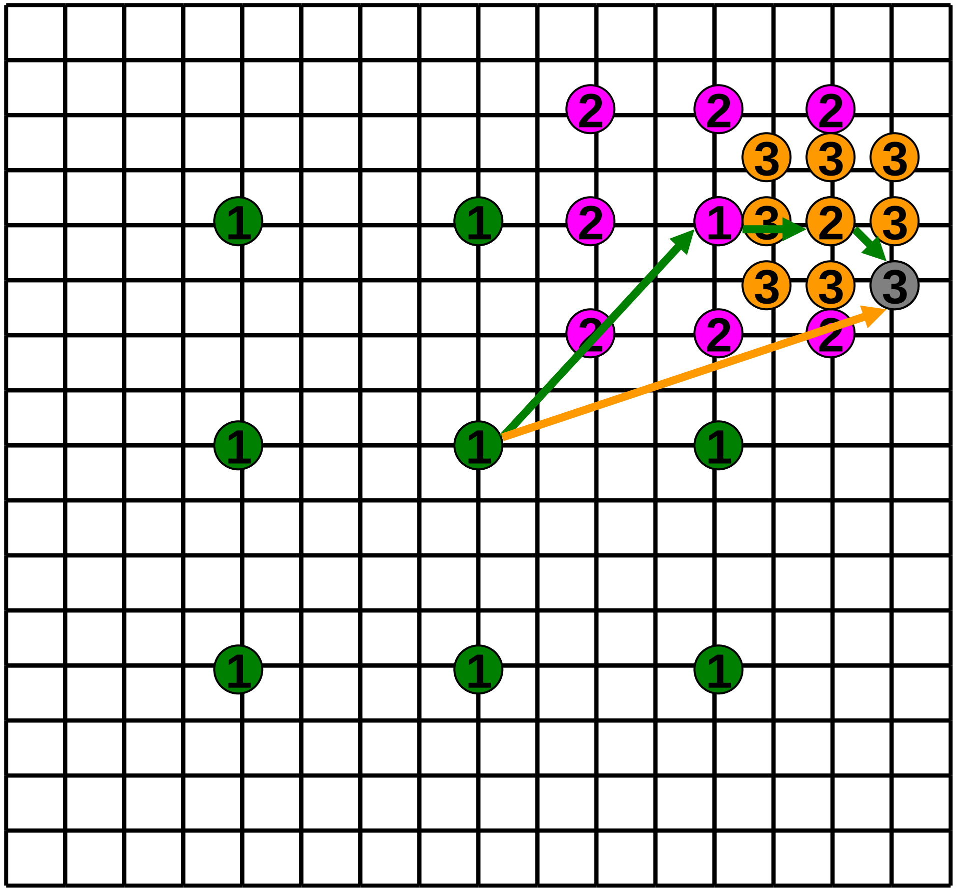
P=5

2-D logarithm Search Method



Three-Step Search Method

- Three-Step Search (TSS)
  - Koga et al. (1981)
  - 9 Points are searched in each step: 3X3 regular grid with equal space w
  - Find the best match
  - Use the previous best as the new center point
  - Reduce the spacing w by half, select 8 new candidates from the reduced 3X3 grid
  - Repeat the search 3 times
  - Examine 25 points in total



Motion Compensation

- The selected ‘best’ matching region in the reference frame is subtracted from the current macroblock to produce a residual macroblock
  - that is encoded and transmitted together with a motion vector describing the position of the best matching region.

Motion Compensation

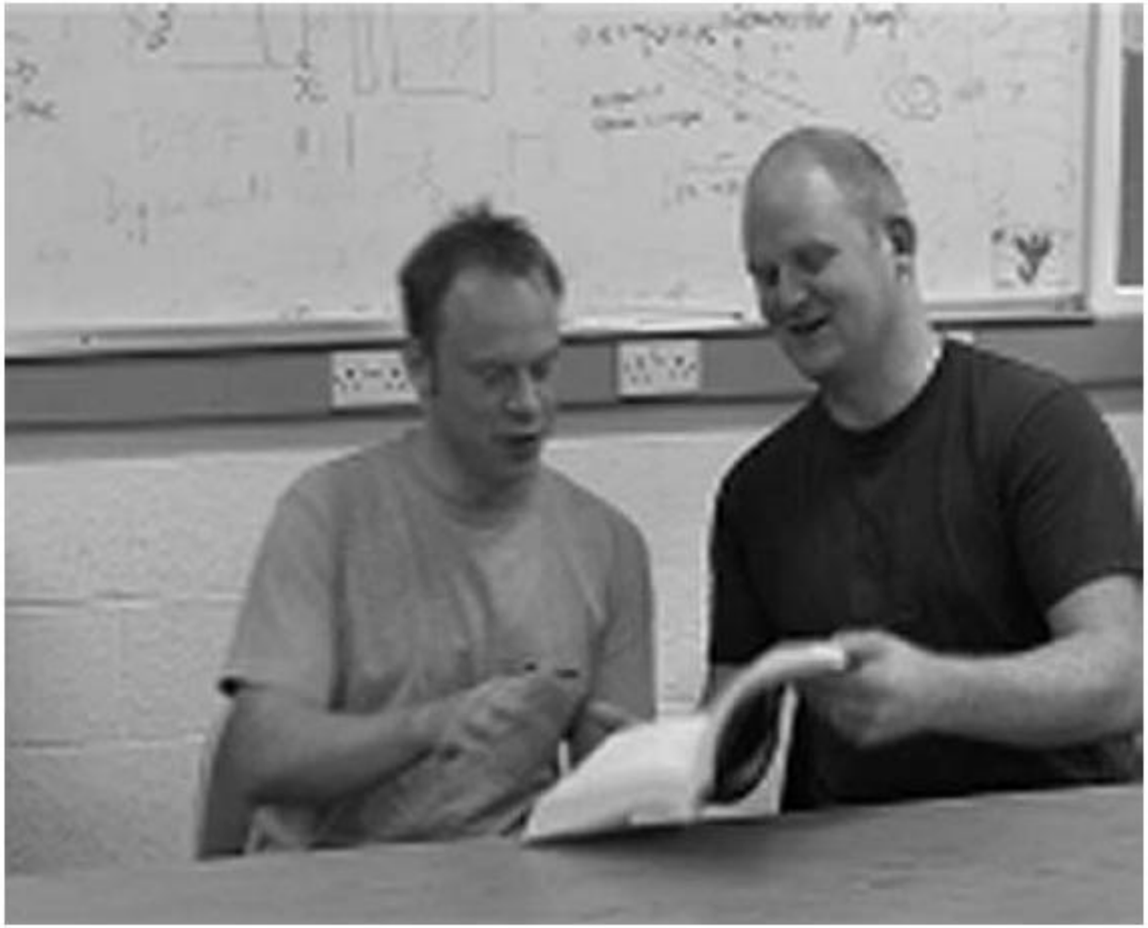


Figure 3.10 Frame 1

Motion Compensation

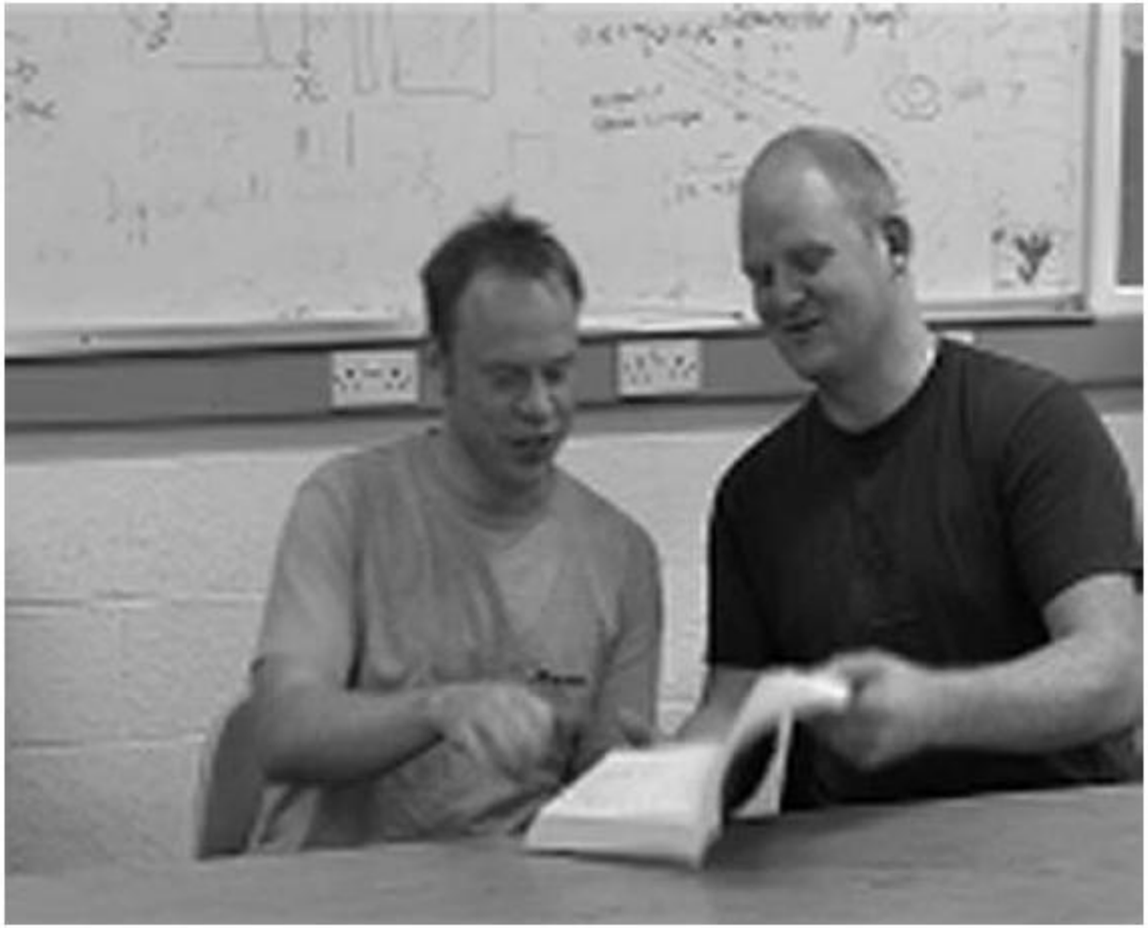


Figure 3.11 Frame 2

Motion Compensation

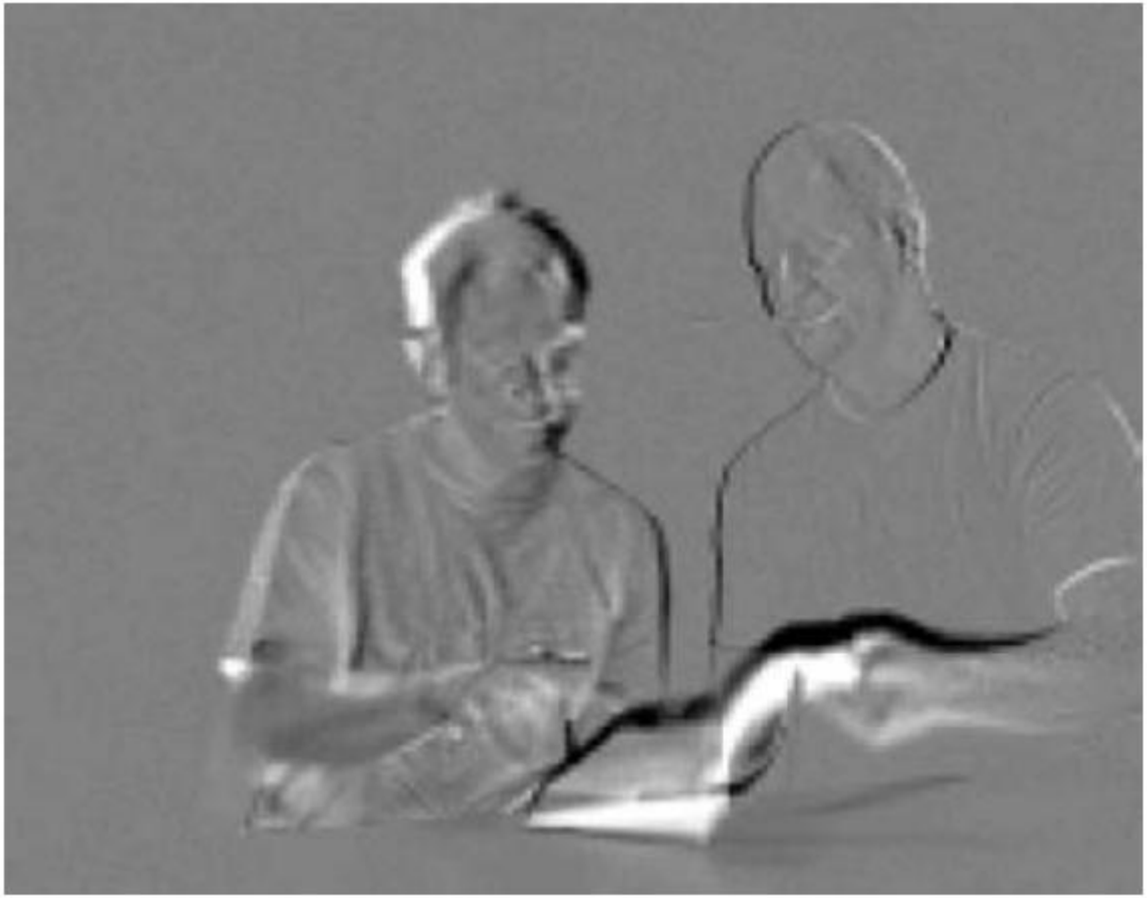


Figure 3.12 Residual (no motion compensation)

Motion Compensation

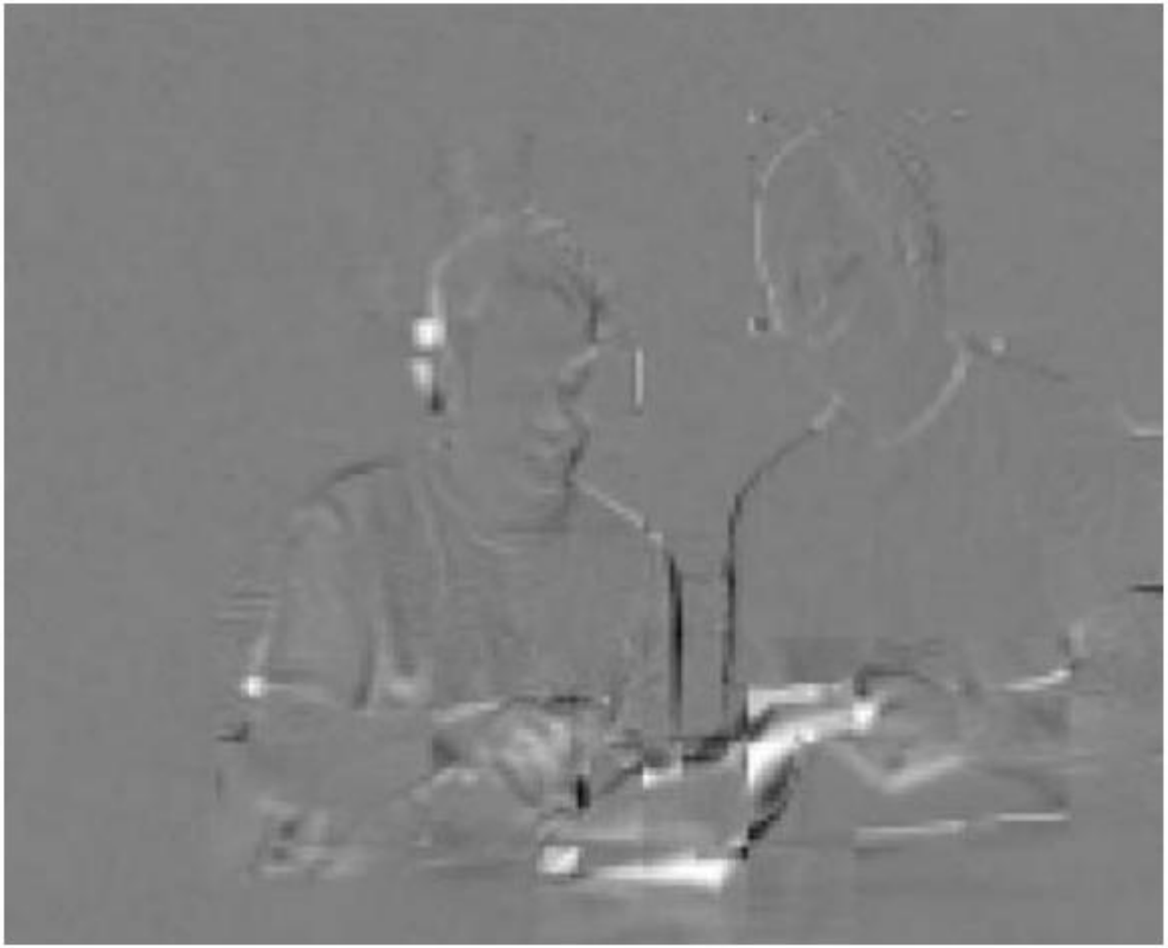


Figure 3.13 Residual (16 × 16 block size)

Motion Compensation

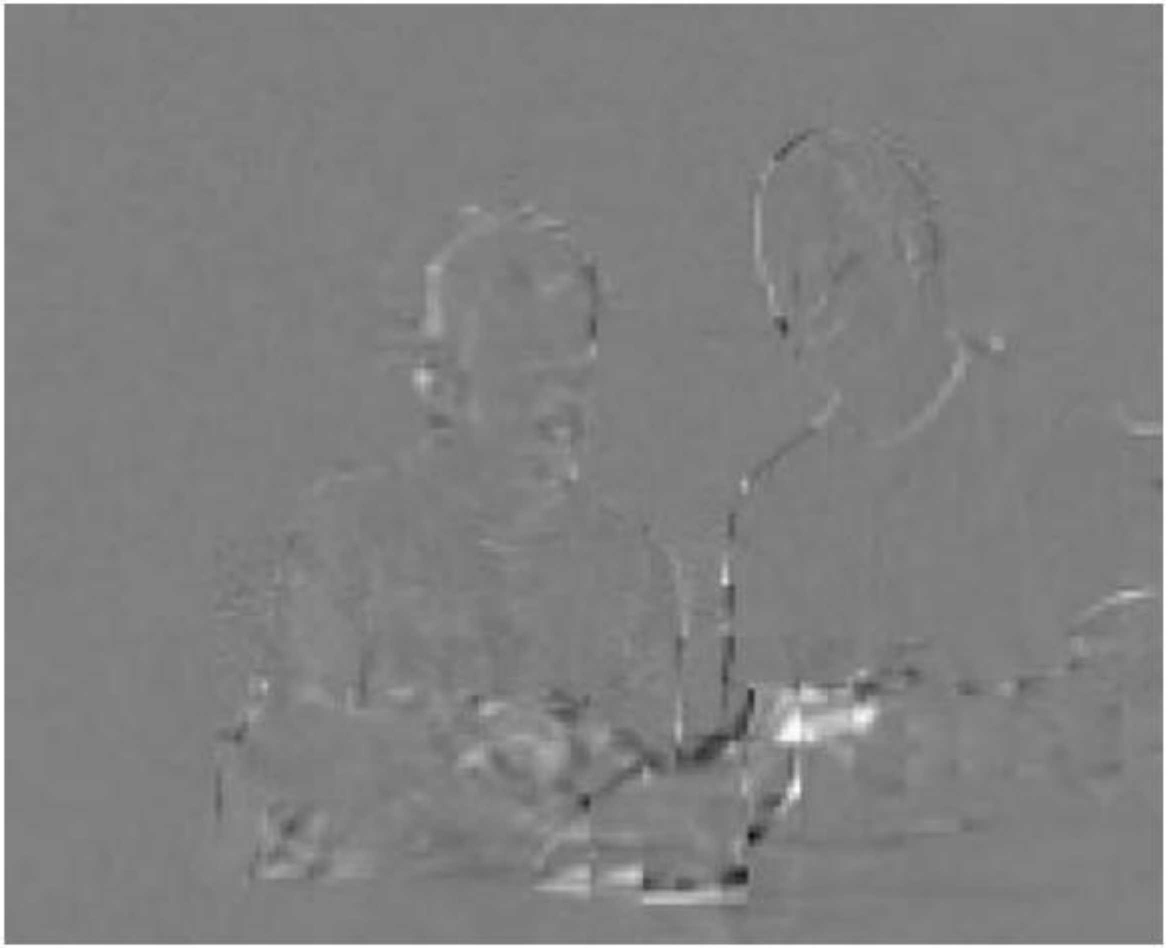


Figure 3.14 Residual (8 × 8 block size)

Motion Compensation

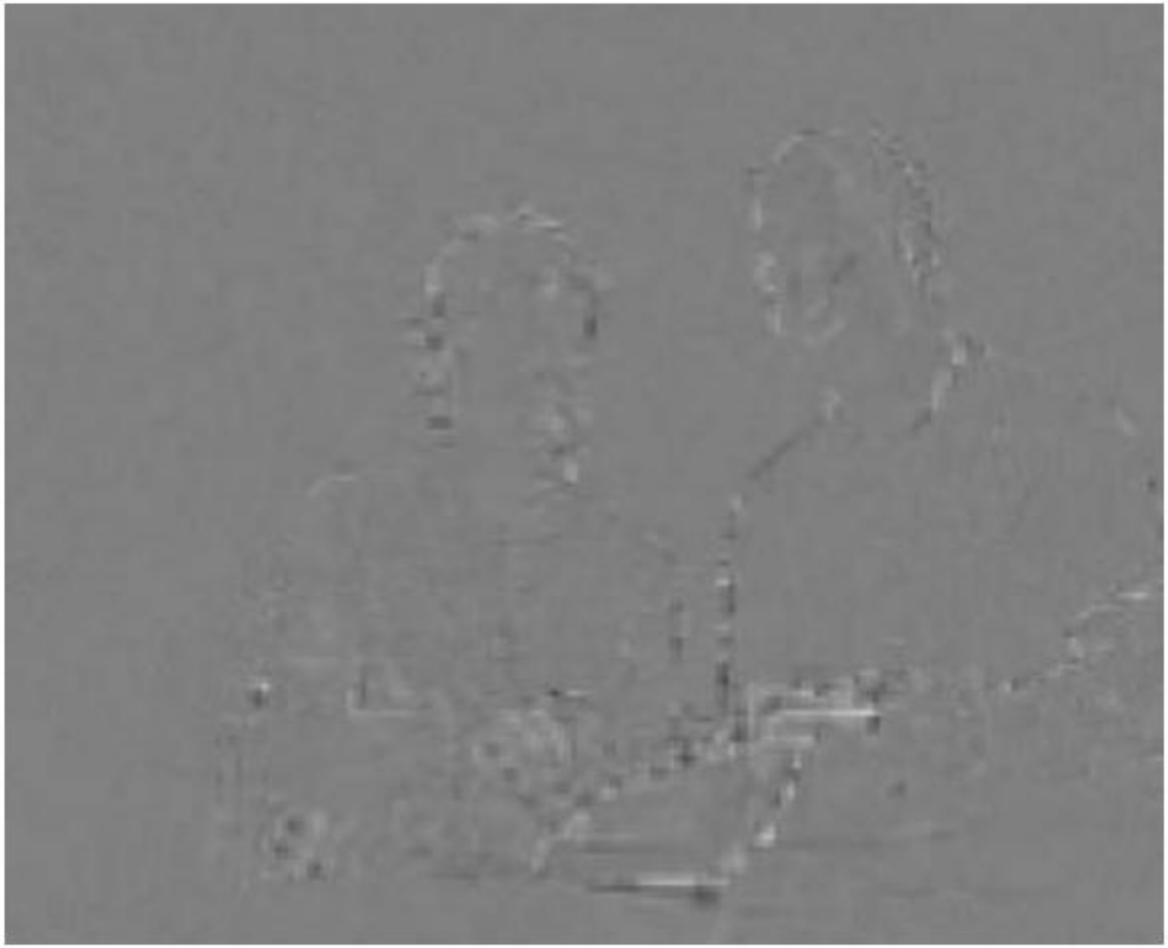


Figure 3.15 Residual (4 × 4 block size)

Main steps of MPEG compression

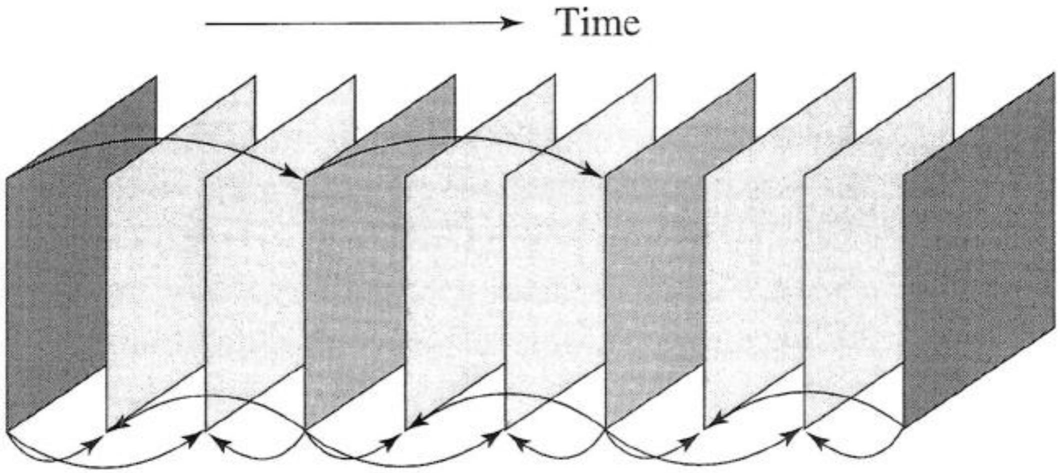
- **Step 5:** For all frames, compress with JPEG compression.
  - Compressing a frame of video is just like compressing a still image, and thus JPEG compression can be applied
  - **I frames** undergo intraframe compression only, without reference to any other image.
  - **P and B frames** first undergo motion prediction. Then the difference between the expected value of a pixel and its actual value is encoded.

MPEG-1 Video

- MPEG-1 was approved by ISO and IEC in 1991 for “Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5Mbps”.
- MPEG-1 standard is composed of
  - System
  - Video
  - Audio
  - Conformance
  - Software
- MPEG-1’s video format is called SIF(Source Input Format)
  - 352x240 for NTSC at 30f/s
  - 352x288 for PAL at 25f/s

MPEG-1 Motion Compensation

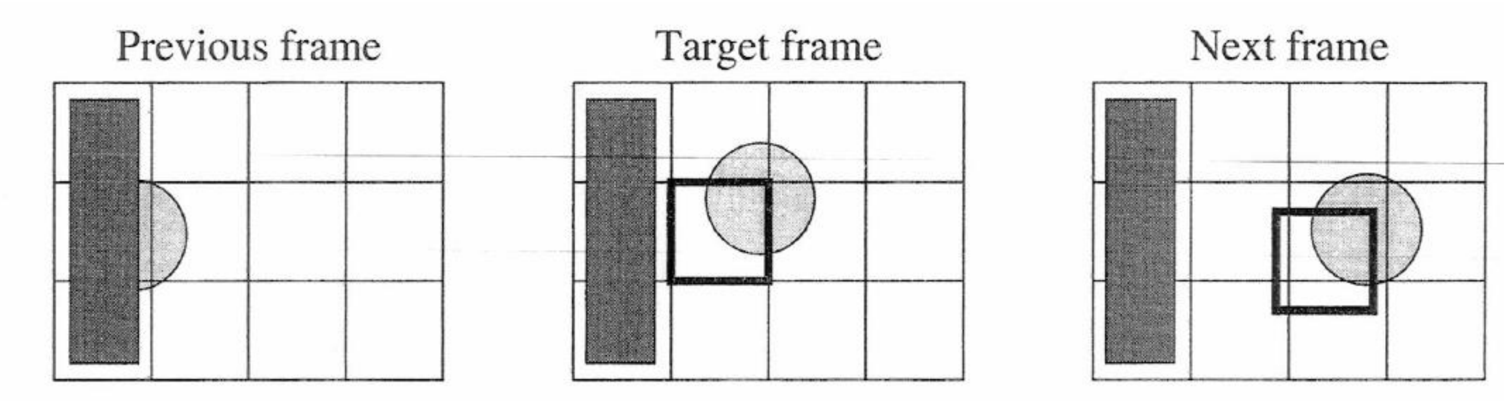
- MPEG-1 introduces a new type of compressed frame: the B-frame.



Display order	I	B	B	P	B	B	P	B	B	I
Coding and transmission order	I	P	B	B	P	B	B	I	B	B

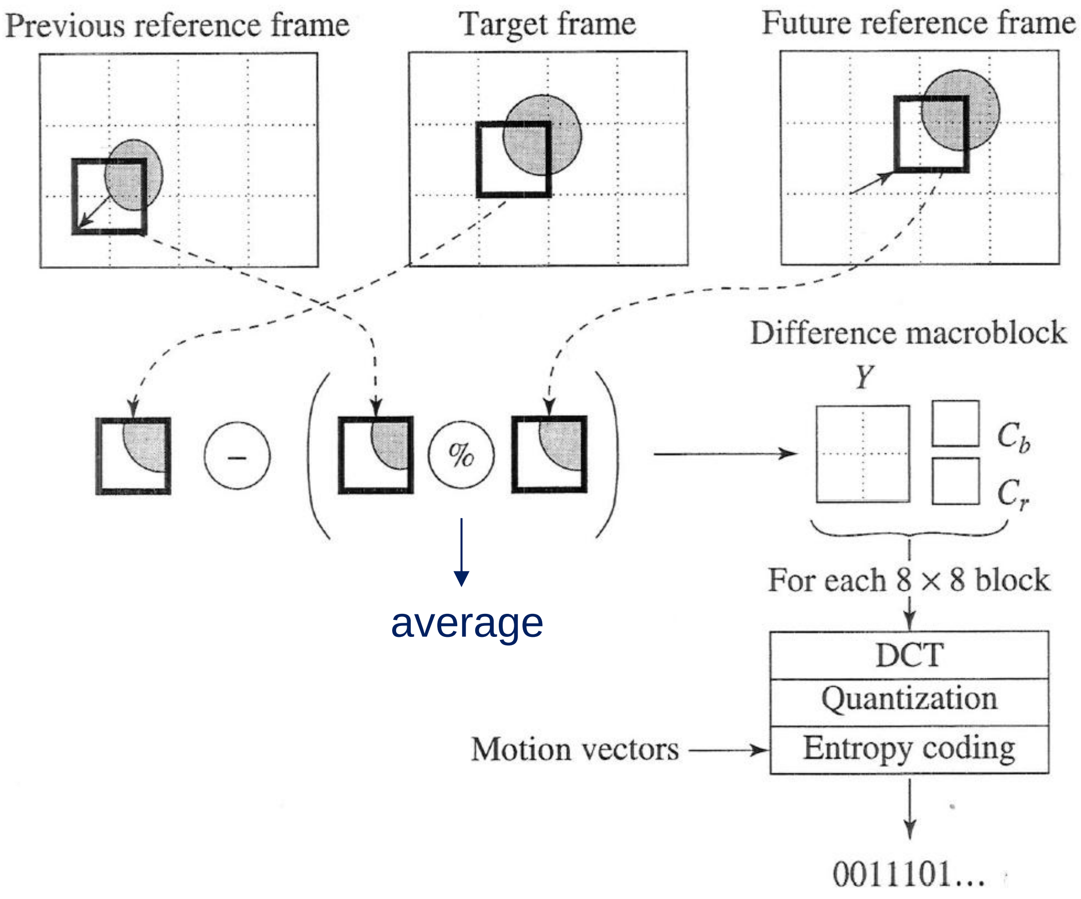
Why do we need B-frames?

- Bi-directional prediction works better than only using previous frames when occlusion occurs.



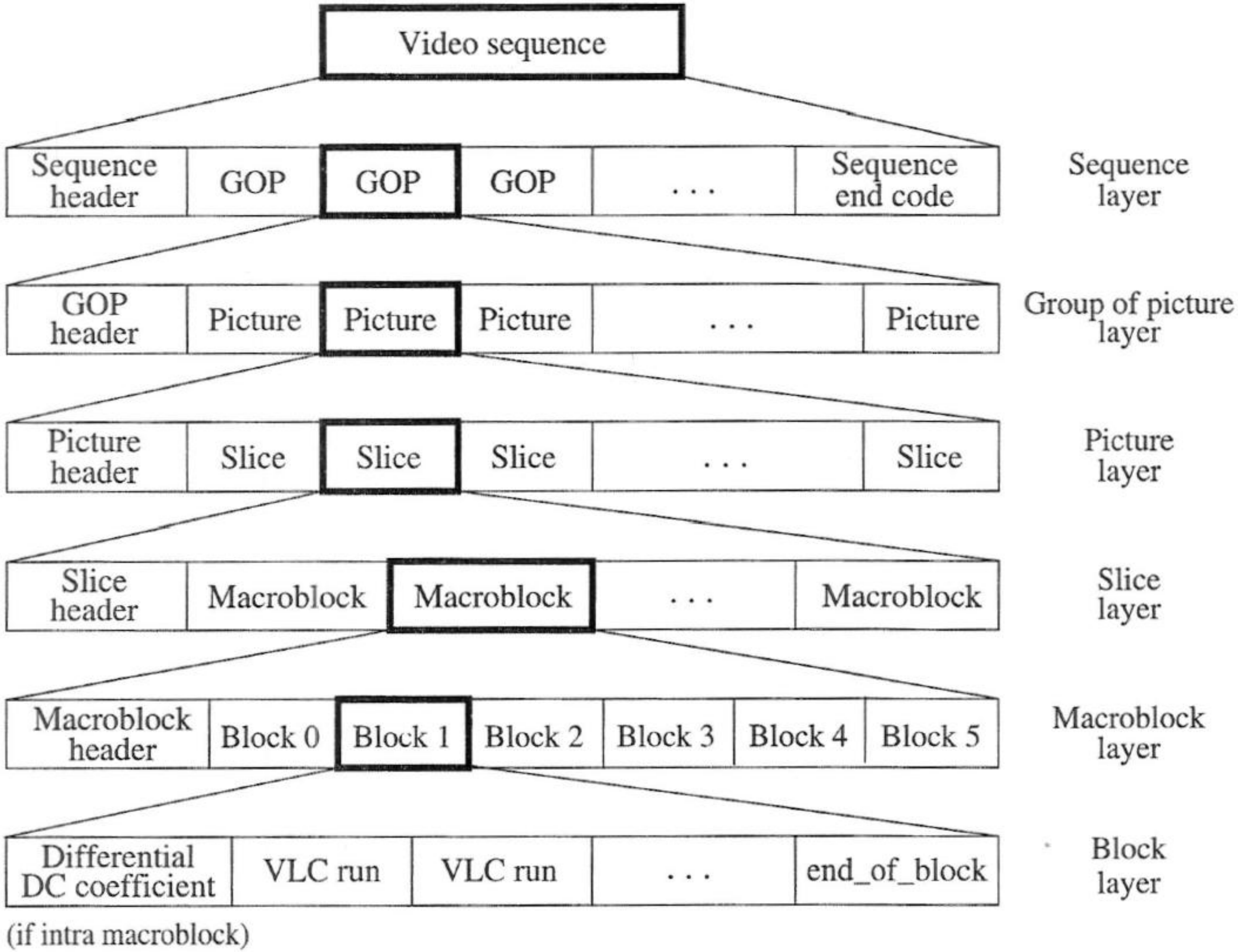
For this example, the prediction from next frame is used and the prediction from previous frame is not considered.

Compression of B-frames



average

MPEG-1 Video Stream

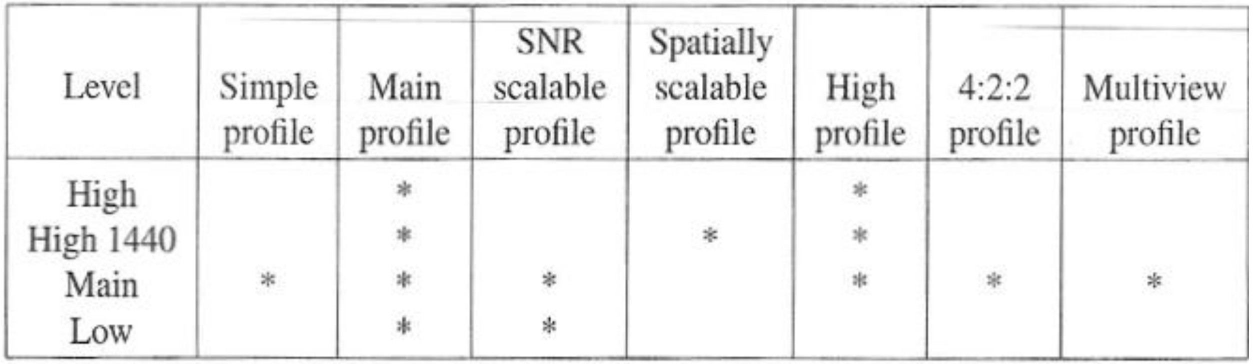


MPEG-2

- **MPEG-2** (aka H.222/H.262 as defined by the ITU) is a standard for "the generic coding of moving pictures and associated audio information“
- Backwards compatibility with existing hardware and software means it is still widely used, for example in the DVD-Video standard
- MPEG-2 evolved out of the shortcomings of MPEG-1

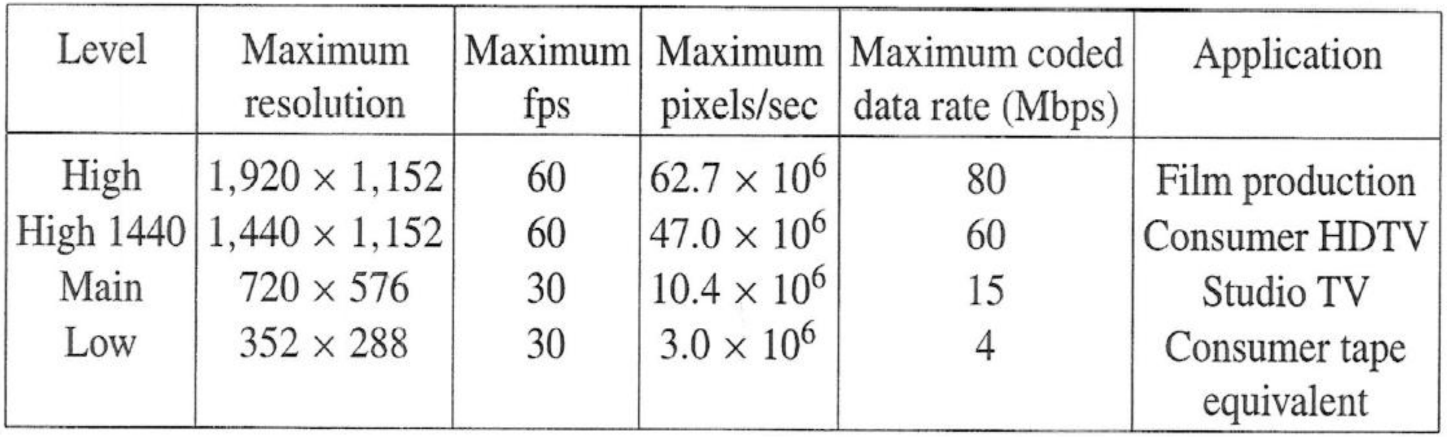
MPEG-2

- MPEG-2 profiles and levels:



Level	Simple profile	Main profile	SNR scalable profile	Spatially scalable profile	High profile	4:2:2 profile	Multiview profile
High		*			*		
High 1440		*		*	*		
Main	*	*	*		*	*	*
Low		*	*				

TABLE 11.6: Four levels in the main profile of MPEG-2.



Level	Maximum resolution	Maximum fps	Maximum pixels/sec	Maximum coded data rate (Mbps)	Application
High	1,920 × 1,152	60	62.7×10^6	80	Film production
High 1440	1,440 × 1,152	60	47.0×10^6	60	Consumer HDTV
Main	720 × 576	30	10.4×10^6	15	Studio TV
Low	352 × 288	30	3.0×10^6	4	Consumer tape equivalent

Scalability

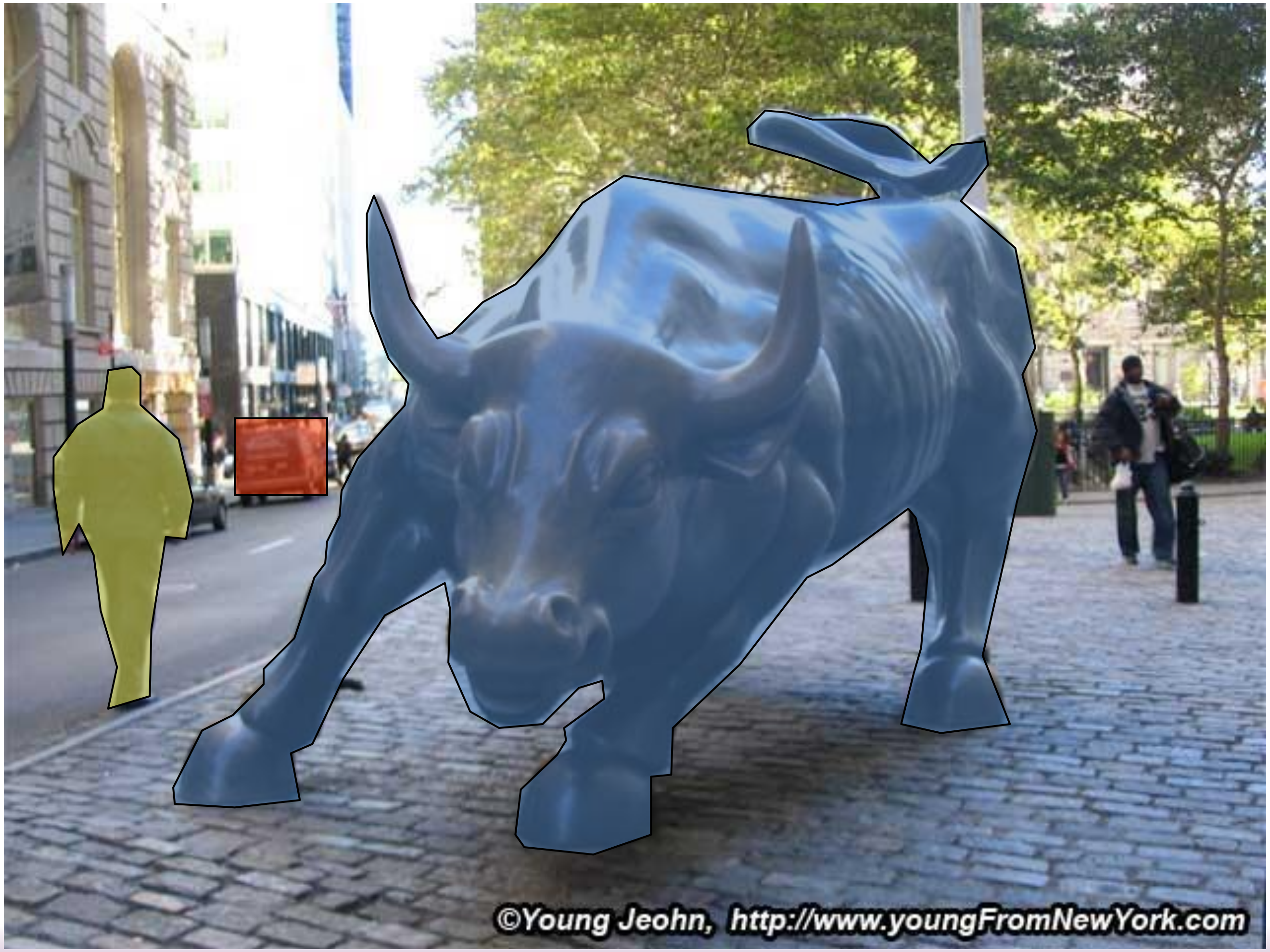
- SNR scalability
  - Base layer uses rough quantization, while enhancement layers encode the residue errors.
- Spatial scalability
  - Base layer encodes a small resolution video; enhancement layers encode the difference of bigger resolution video with the “un-sampled” lower resolution one.
- Temporal scalability
  - Base layer down-samples the video in time; enhancement layers include the rest of the frames.
- Hybrid scalability
- Data partitioning

MPEG-4

- Officially up to 10 Mbits/sec.
- Improved encoding efficiency.
- Content-based interactivity.
- Content-based and temporal random access.
- Integration of both natural and synthetic objects.
- Temporal, spatial, quality and object-based scalability.
- Improved error resilience.
- Support object-based features for content
- Enable dynamic rendering of content
  - defer composition until decoding
- Support convergence among digital video, synthetic environments, and the Internet

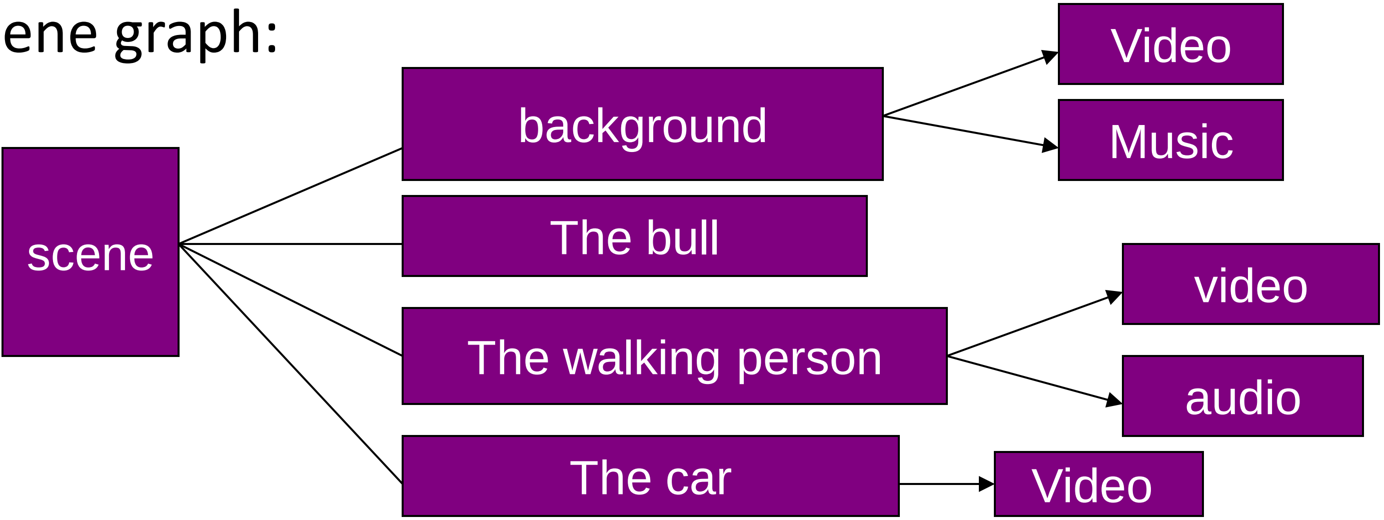
Audio-Video Object

- MPEG4 is based on the concept of media objects.



Audio Video Objects

- A media object in MPEG4 could be
  - A video of an object with “shape”.
  - The speech of a person.
  - A piece of music.
  - A static picture.
  - A synthetic 3D cartoon figure.
- In MPEG4, a scene is composed of media objects based on a scene graph:



MPEG-4 Standard

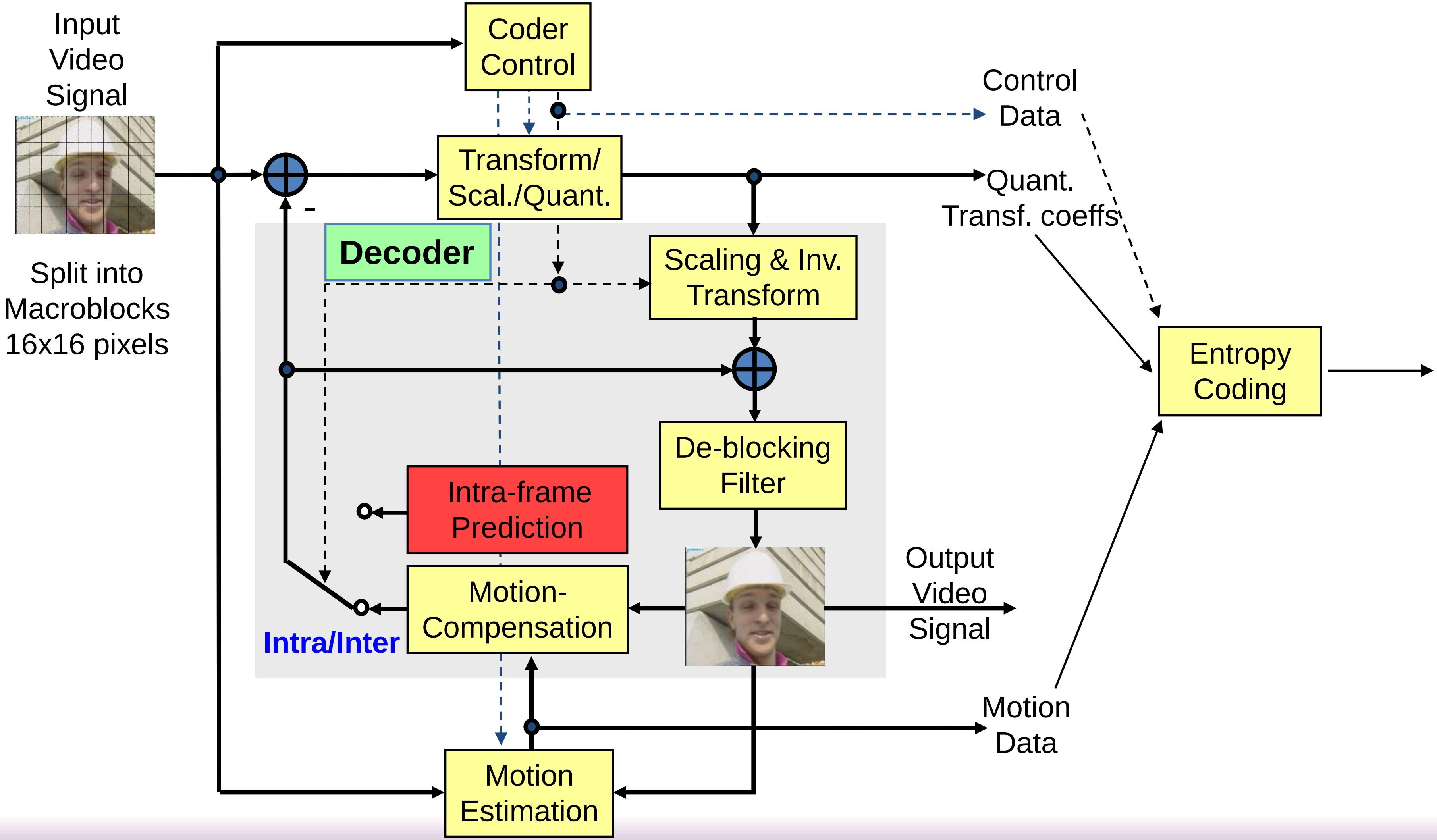
- Defines the scheme of encoding audio and video objects
  - Encoding of shaped video objects.
  - Sprite encoding.
  - Encoding of synthesized 2D and 3D objects.
- Defines the scheme of decoding media objects.
- Defines the composition and synchronization scheme.
- Defines how media objects interact with users.

Video Coding in MPEG-4

Support 4 types of video coding:

- Video Object Coding
  - For coding of natural and /or synthetic originated, rectangular or arbitrary shaped video objects.
- Mesh Object Coding
  - For visual objects represented with a mesh structure.
- Model-based Coding
  - For coding of a synthetic representation and animation of a human face and body.
- Still Texture Coding
  - For wavelet coding of still textures.

Basic coding structure of H.264/AVC for a macroblock

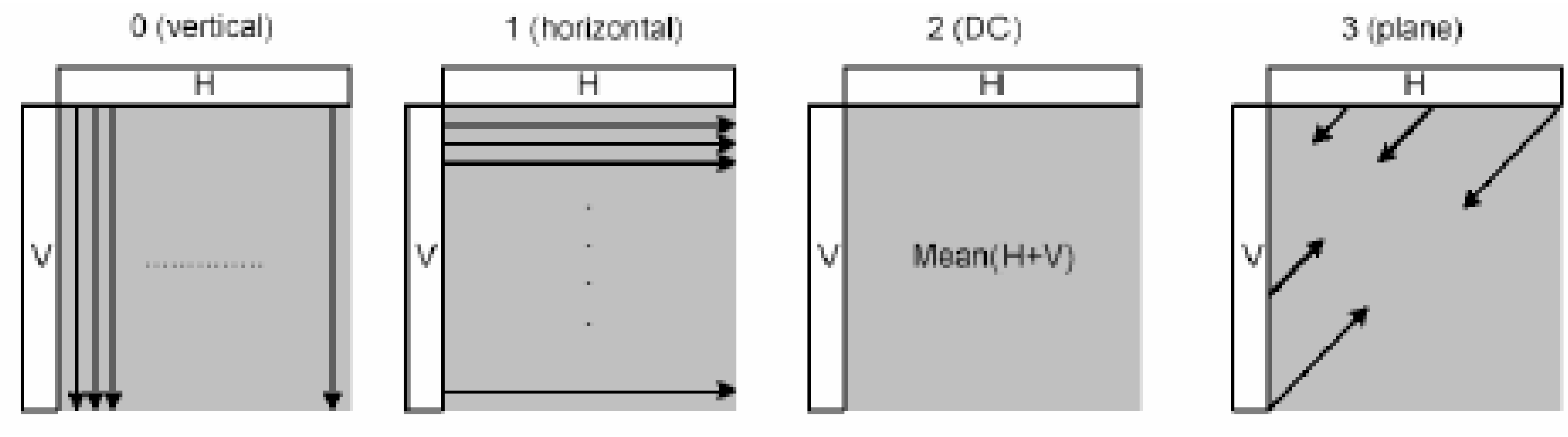


Intra-frame Prediction

- Intra-frame encoding of H.264 supports Intra_4×4, Intra_16×16 and I_PCM.
  - I_PCM bypass prediction and transform coding and, send the values of the encoded samples directly.
  - Intra_4 ×4 and Intra_16 ×16 allows the *intra prediction*.
    - Intra 4×4
      - ✓9 modes
      - ✓Used in texture area
    - Intra 16×16
      - ✓4 modes
      - ✓Used in flat area

Four modes of Intra_16×16

- Mode 0 (vertical) : extrapolation from upper samples(H)
- Mode 1 (horizontal): extrapolation from left samples(V)
- Mode 2 (DC): mean of upper and left-hand samples (H+V)
- Mode 3 (Plane) : a linear “plane” function is fitted to the upper and left-hand samples H and V. This works well in areas of smoothly-varying luminance



Example

Original image

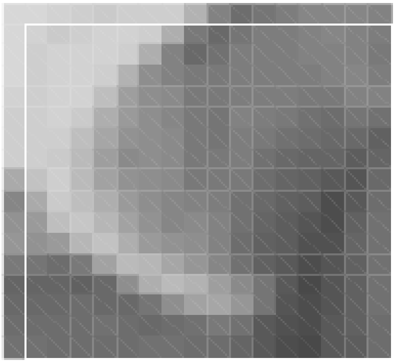
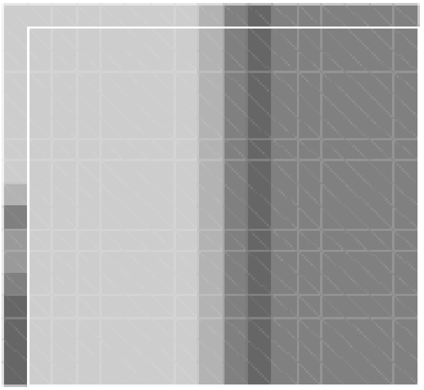
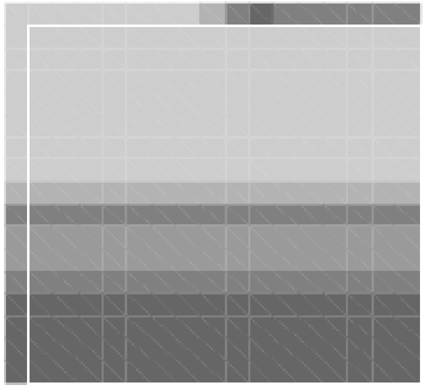


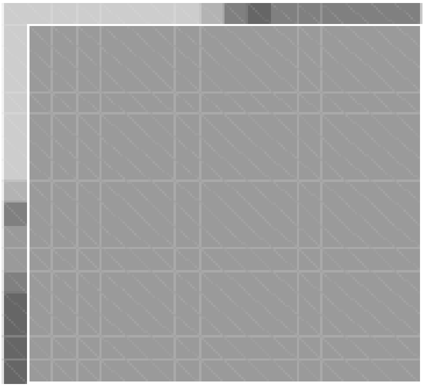
Figure 6 16x16 macroblock



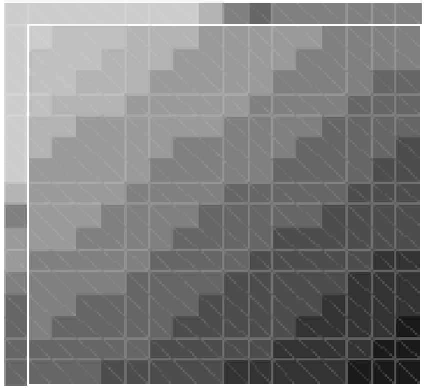
0 (vertical), SAE=8990



1 (horizontal), SAE=10898

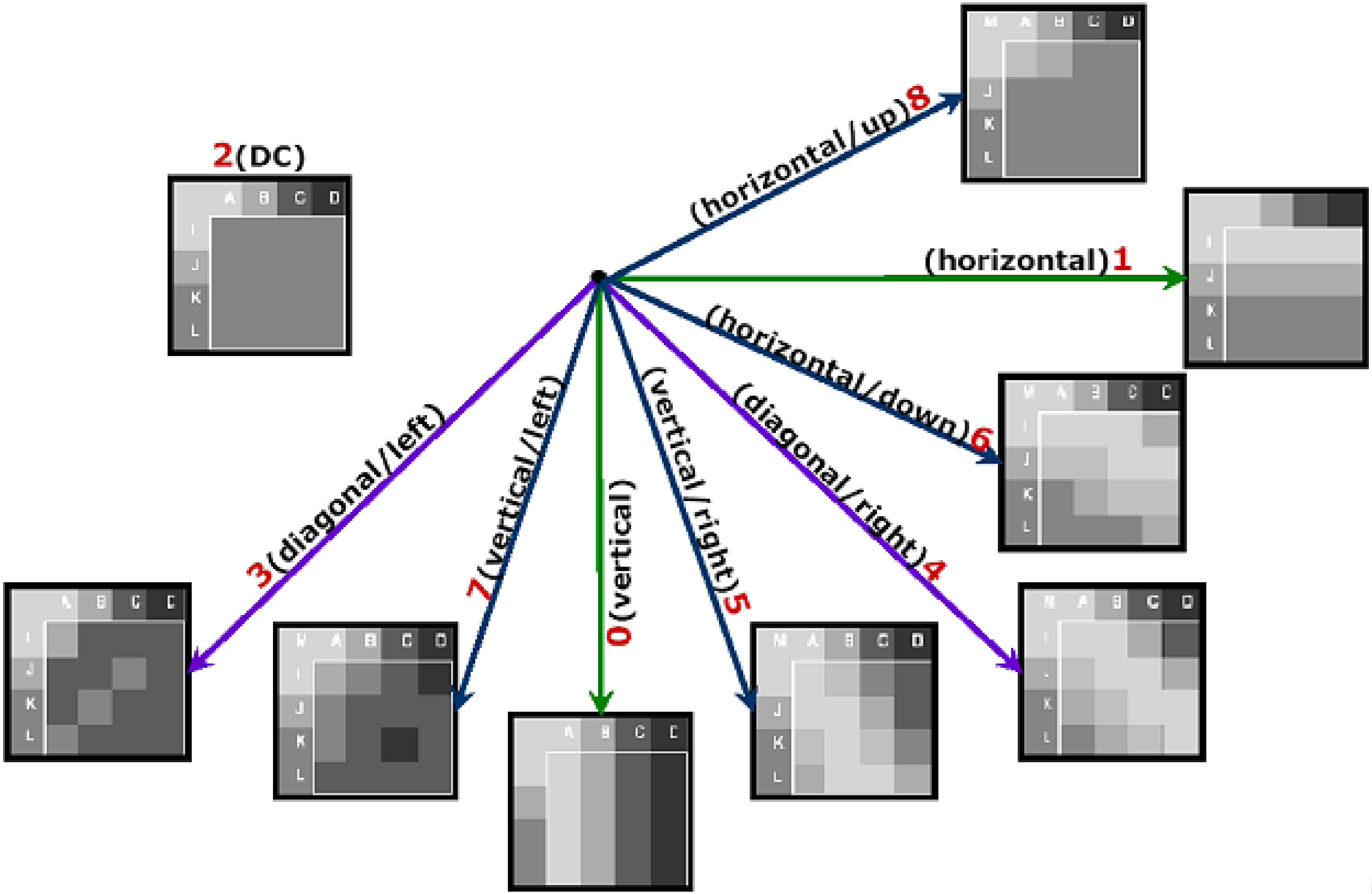


2 (DC), SAE=11210



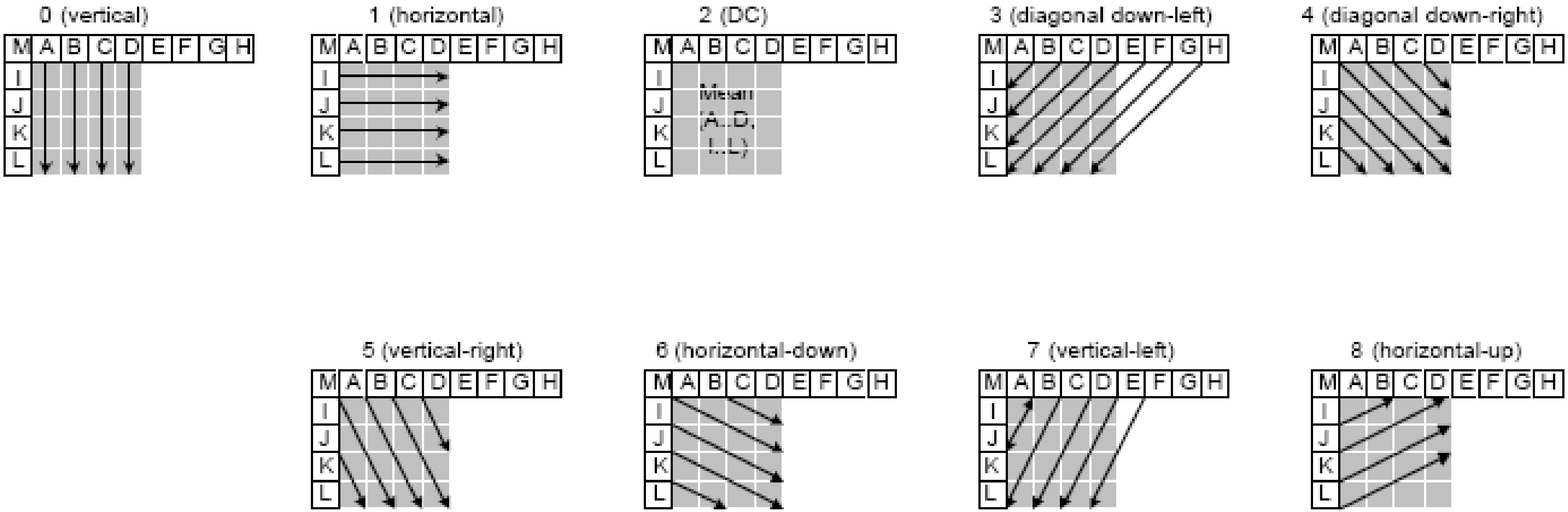
3 (plane), SAE=6264

Nine modes of Intra_4×4



Nine modes of Intra_4×4

- The prediction block P is calculated based on the samples labeled A-M.
- The encoder may select the prediction mode for each block that minimizes the residual between P and the block to be encoded



0 (vertical)

1 (horizontal)

2 (DC)

3 (diagonal down-left)

4 (diagonal down-right)

5 (vertical-right)

6 (horizontal-down)

7 (vertical-left)

8 (horizontal-up)

Intra Prediction Example

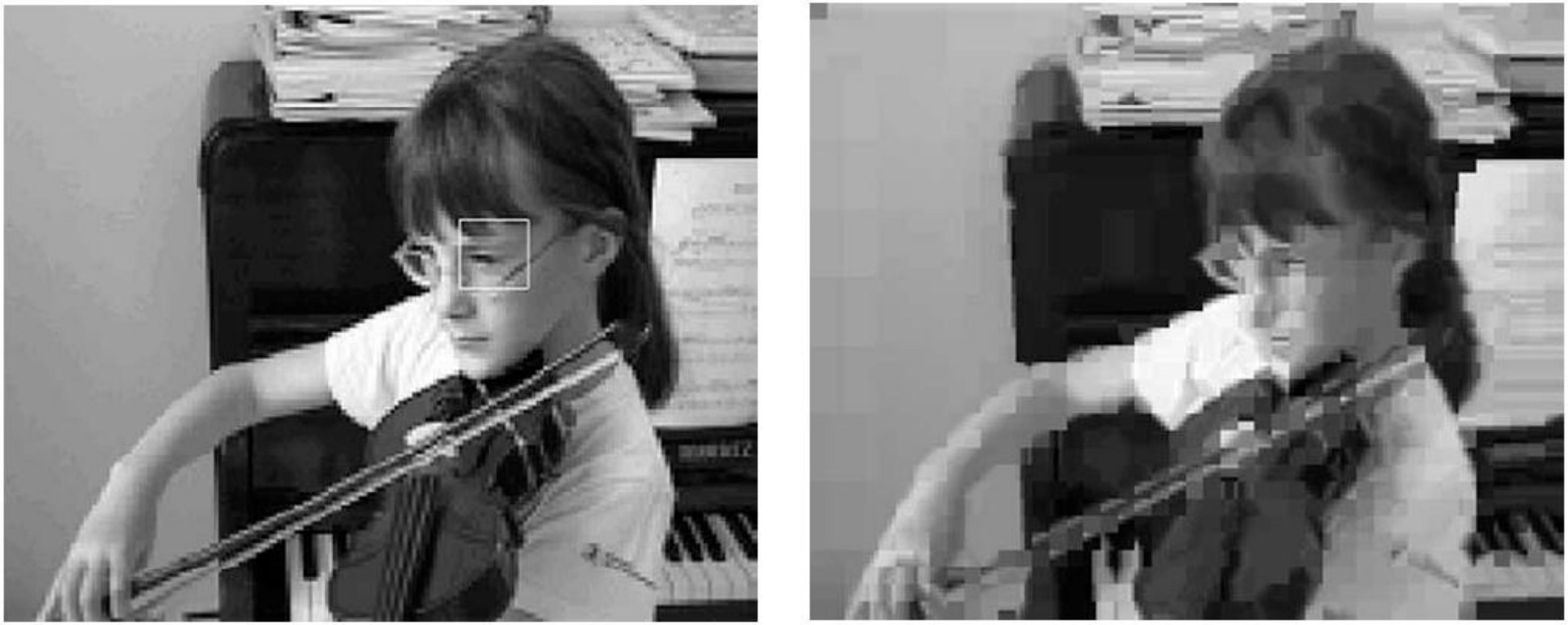
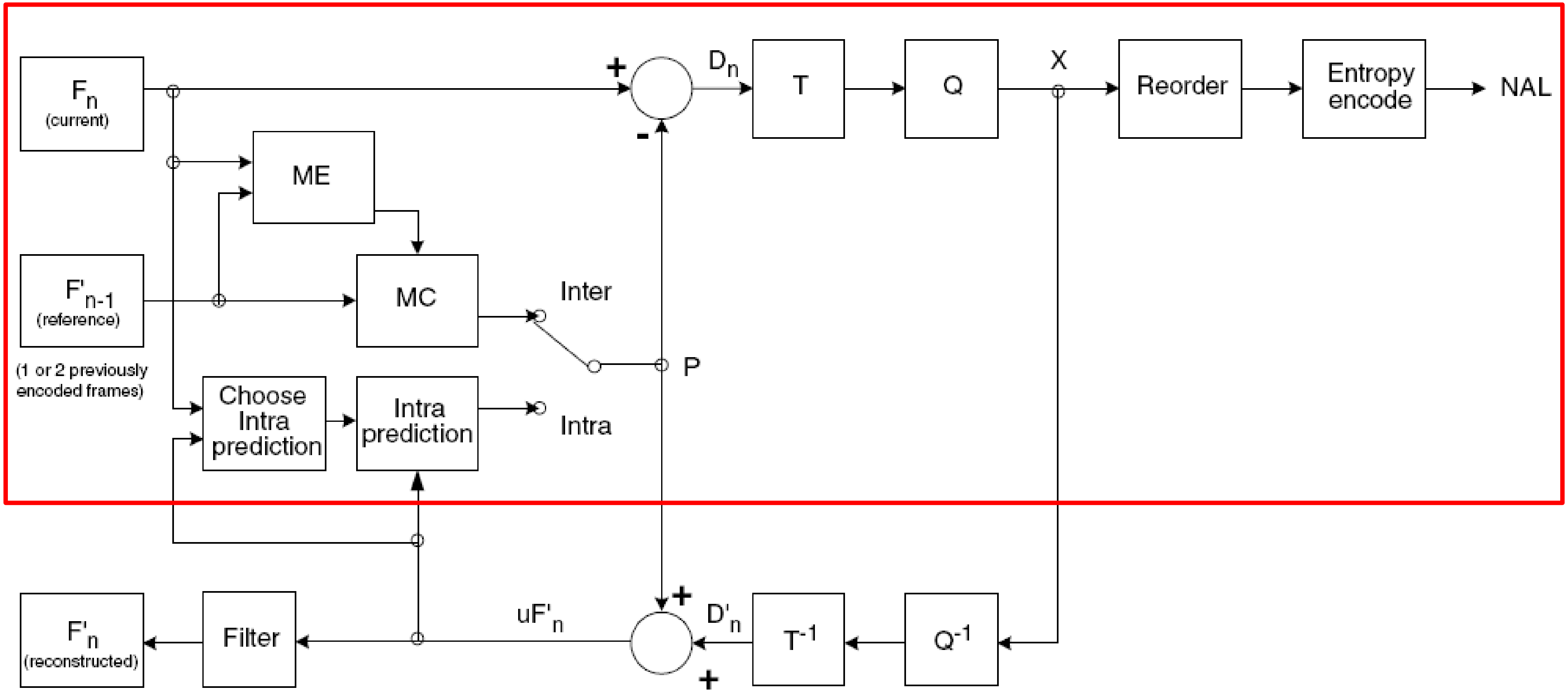


Figure 6.21 Predicted luma frame formed using H.264 intra prediction

H.264 Encoder

Forward Path



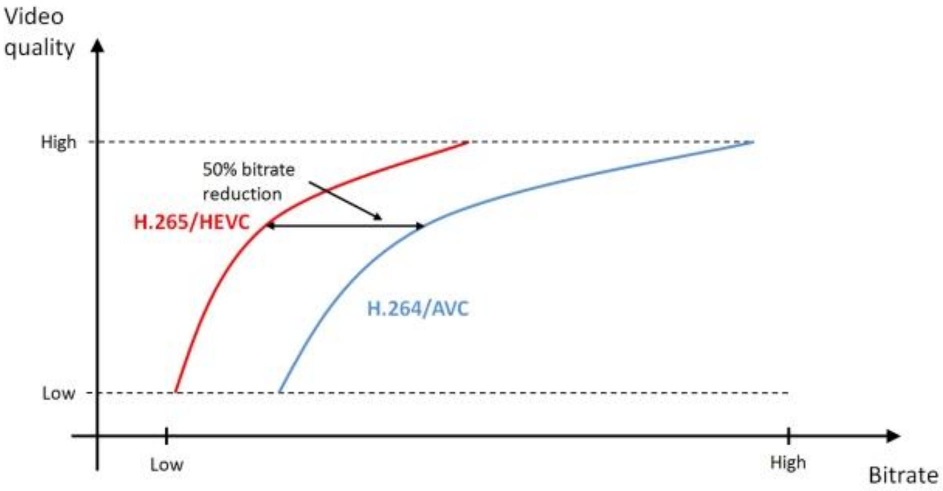
MPEG-4 Summary

- A lot of MPEG-4 examples with interactive capabilities
- Content-based Interactivity
  - Scalability
    - Spatial Scalability
    - Temporal Scalability
  - Sprite Coding
- Improved Compression Efficiency
- Universal Accessibility
  - re-synchronization
  - data recovery
  - error concealment

High Efficiency Video Coding (HEVC)

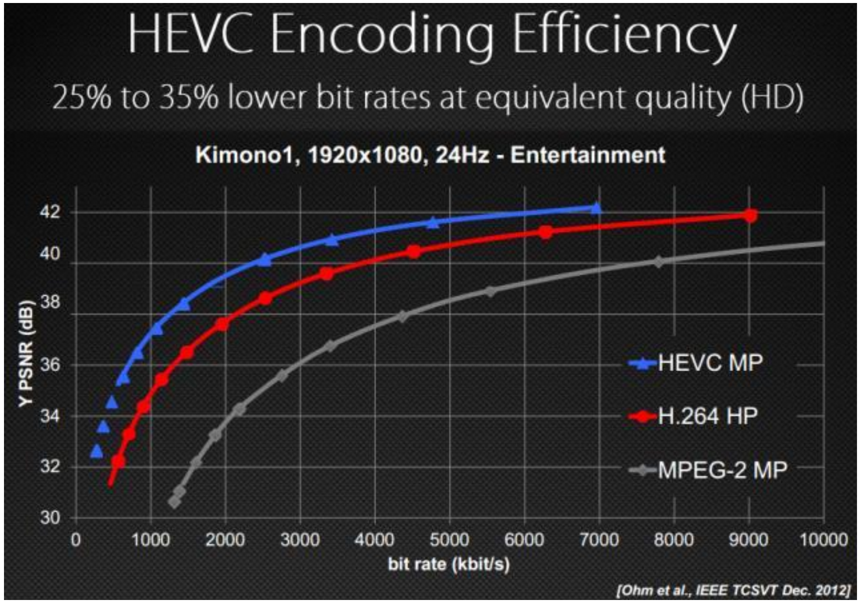
- High Efficiency Video Coding (HEVC) is the latest generation video compression standard.
- This standard was developed by the **ISO/IEC** MPEG and **ITU-T** VCEG, through their JCT-VC
- HEVC is also known as ISO/IEC 23008-2 MPEG-H Part 2 and ITU-T **H.265**

- Have a bit rate reduction of 50% at the same subjective image quality compared to the H.264/MPEG-4 AVC High profile



High Efficiency Video Coding (HEVC)

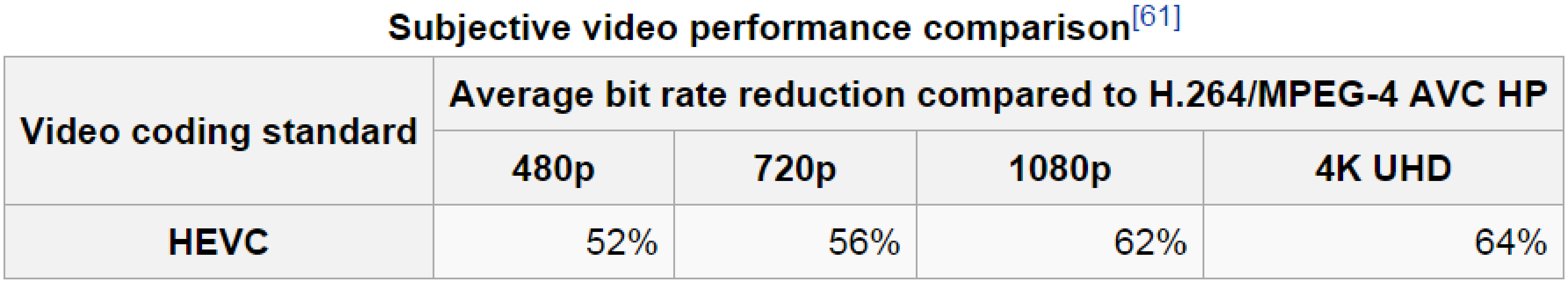
- It can support 8K UHD and resolutions up to 8192×4320.
- HEVC is said to double the data compression ratio compared to H.264/MPEG-4 AVC at the same level of video quality.



HEVC Encoding Efficiency

25% to 35% lower bit rates at equivalent quality (HD)

[Ohm et al., IEEE TCSVT Dec. 2012]



Subjective video performance comparison[61]

Video coding standard	Average bit rate reduction compared to H.264/MPEG-4 AVC HP			
	480p	720p	1080p	4K UHD
HEVC	52%	56%	62%	64%

H.265/HEVC/MPEG-H Part 2

- Main drivers
  - Get **low bitrate target – target 2:1** over H.264
  - Cheat your eyes – how much can you cut bits and still see the same quality
  - Improve **resolutions (8K by 4K and 4K by 2K) and frame rates**
  - Launch **1080p50/60 services** to compete against BluRay
  - Expect up to 10x more computational complexity (encode) and 2x-3x (decode)

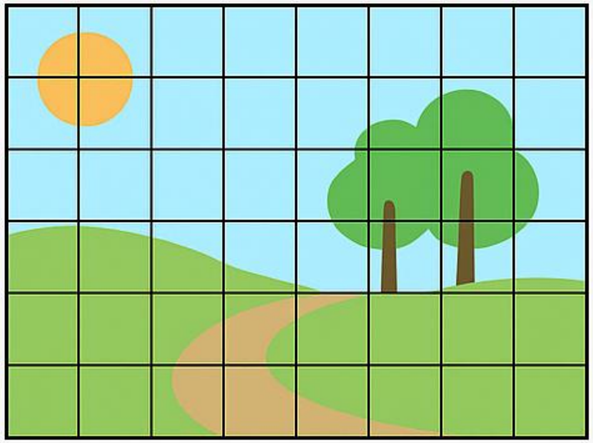
H.265

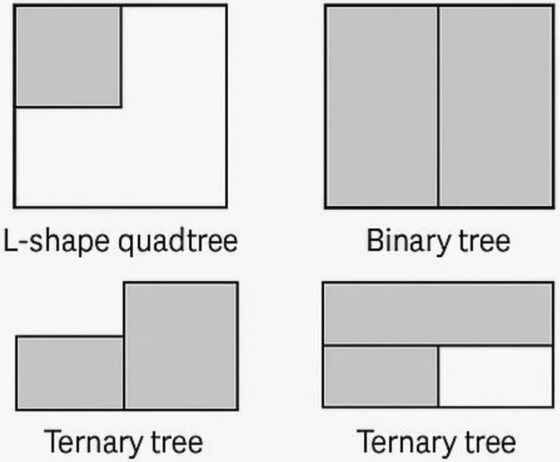
- Derived from H.264
  - More modes, tools and more interdependencies
  - More efficient search algorithms
  - More complex intra-prediction
  - Macroblocks vs Partitions

• AVC • 16x16 macro-blocks • 8x8 and 4x4 transform sizes	• HEVC • Coding unit size 64x64 to 8x8 • 32x32, 16x16, 8x8 and 4x4 transform sizes

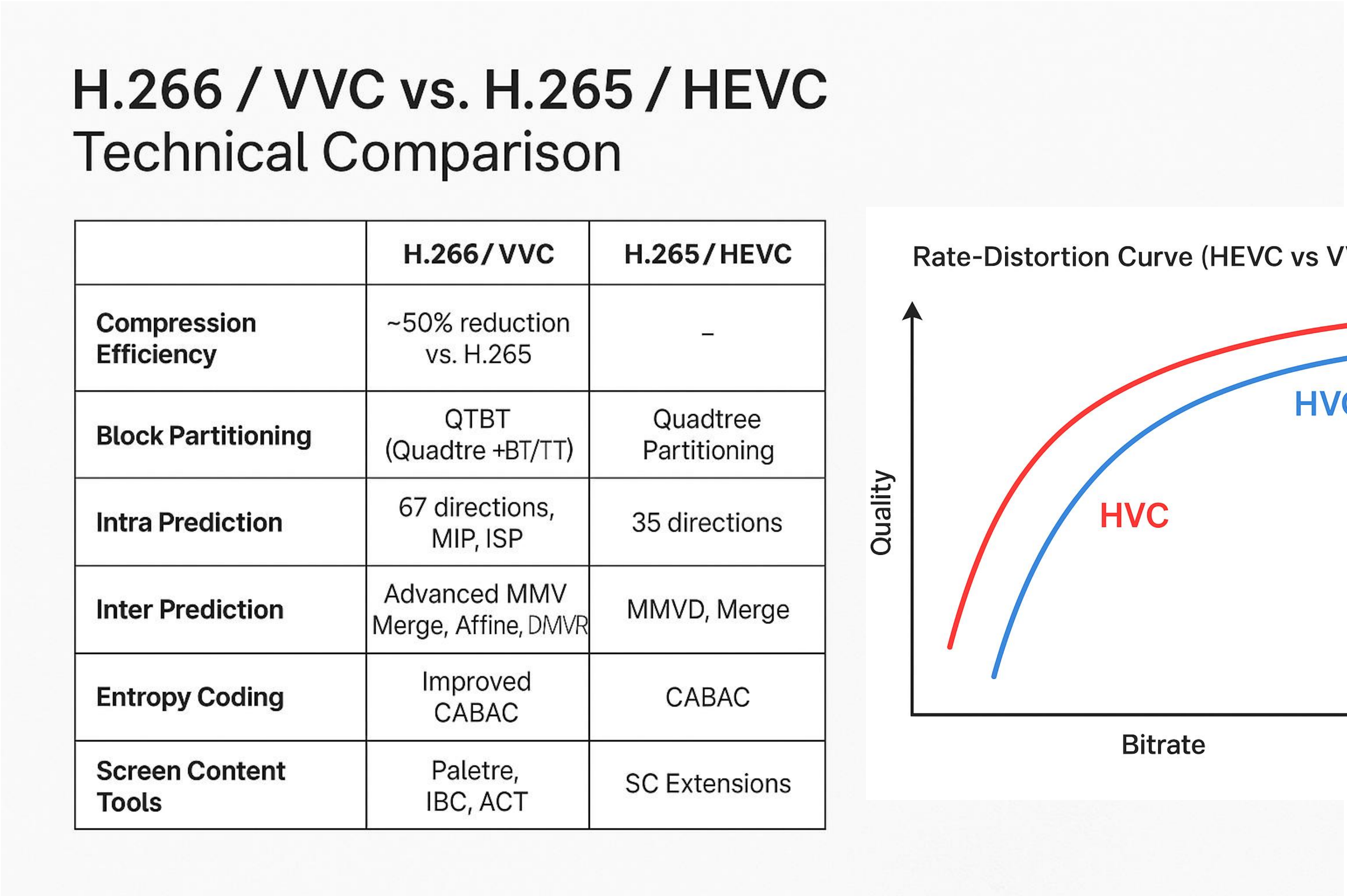
VVC (H.266) – Versatile Video Coding

- Main new technologies introduced compared to H.265 (HEVC
  - Larger set of block sizes for coding decisions
  - More advanced prediction and transform methods
  - Improved in-loop filtering (e.g. adaptive loop filter)
  - Enhanced partitioning of prediction blocks





H.266/VVC vs. H.265/HEVC



H.266 / VVC vs. H.265 / HEVC Technical Comparison

	H.266/VVC	H.265/HEVC
Compression Efficiency	~50% reduction vs. H.265	–
Block Partitioning	QTBT (Quadtre +BT/TT)	Quadtree Partitioning
Intra Prediction	67 directions, MIP, ISP	35 directions
Inter Prediction	Advanced MMV Merge, Affine, DMVR	MMVD, Merge
Entropy Coding	Improved CABAC	CABAC
Screen Content Tools	Paletre, IBC, ACT	SC Extensions

Rate-Distortion Curve (HEVC vs VVC)

High Definition TV (HDTV)

- The standard supports video scanning formats shown in Table 5.4. In the table, “I” mean interlaced scan and “P” means progressive (non-interlaced) scan.

- **Table 5.4:** Advanced Digital TV formats supported by ATSC

# of Active Pixels per line	# of Active Lines	Aspect Ratio	Frame Rate
1,920	1,080	16:9	60P 60I 30P 24P
1,280	720	16:9	60P 30P 24P
704	480	16:9 or 4:3	60P 60I 30P 24P
640	480	4:3	60P 60I 30P 24P

HDTV

- For video, MPEG-2 is chosen as the compression standard. For audio, AC-3 is the standard. It supports the so-called 5.1 channel Dolby surround sound, i.e., five surround channels plus a subwoofer channel.
- The salient difference between conventional TV and HDTV:
  a) HDTV has a much wider aspect ratio of 16:9 instead of 4:3.
  b) HDTV moves toward progressive (non-interlaced) scan. The rationale is that interlacing introduces serrated edges to moving objects and flickers along horizontal edges.
- **HDTV** : 720 active lines or higher. Popular choices are:
  - 720P (1,280 × 720, progressive scan, 30 fps)
  - 1080I (1,920 × 1,080, interlaced, 30 fps)
  - 1080P (1,920 × 1,080, progressive scan, 30 or 60 fps).

Ultra High Definition TV (UHDTV)

- UHDTV is a new generation of HDTV. The standards initiated in 2012 support 4K UHDTV: 2160P (3,840×2,160, progressive scan) and 8K UHDTV: 4320P (7,680×4,320, progressive scan).
- The aspect ratio is 16:9. The bit-depth is 10 or 12 bits per sample, and the chroma subsampling can be 4:2:0, 4:2:2, or 4:4:4.
- The supported frame rate has been gradually increased to 120 fps.
- The UHDTV will provide superior picture quality, comparable to IMAX movies, but it will require a much higher bandwidth and bitrate.

UHDTV

- 16K UHDTV has been demonstrated in 2018, targeting applications such as Virtual Reality with true immersion. Its resolution is 15,360 × 8,640 for a total of 132.7 megapixels.

- **Table 5.5:** A Summary of UHDTV

Type of UHDTV	Resolution	Bit Depth	Aspect Ratio	Frame Rate
4K UHD (2160P)	3840 x 2160	10 or 12 bits	16:9	Up to 120P
8K UHD (4320P)	7680 x 4320	10 or 12 bits	16:9	Up to 120P
16K UHD (8640P)	15360 x 8640	10 or 12 bits	16:9	Up to 240P

Video Display Interfaces

- **Analog Display Interfaces**

- Analog video signals are often transmitted in one of three different interfaces: *Component video*, *Composite video*, and *S-video*.

Connectors for typical analog display interfaces. From left to right: Component video, Composite video, S-video, and VGA.

Digital Display Interfaces

- Digital interfaces emerged in 1980s (e.g., Color Graphics Adapter (CGA)), and evolved rapidly. Today, the most widely used digital video interfaces include Digital Visual Interface (DVI), High-Definition Multimedia Interface (HDMI), and DisplayPort.
- Connectors of different digital display interfaces. From left to right: DVI, HDMI, DisplayPort.

High-Definition Multimedia Interface (HDMI)

- HDMI is a newer digital audio/video interface developed to be backward compatible with DVI.
  - HDMI doesn’t carry analog signal and hence is not compatible with VGA.
  - HDMI supports both RGB and YCbCr 4:4:4 or 4:2:2. [DVI is limited to the RGB color range (0-255).]
  - HDMI supports digital audio, in addition to digital video.
- The maximum pixel clock rate for HDMI 1.0 is 165 MHz, which is sufficient to support 1080P (1920×1200) at 60 Hz.
- HDMI 2.0 was released in 2013, which supports 4K resolution at 60 frames per second.

360^{o} Video

- 360^{o} Video is also known as *Omnidirectional Video, Spherical Video* or *Immersive Video*.
- 360^{o} Video can span 360^{o} horizontally and 180^{o} vertically.
- It is captured by cameras at (almost) all possible viewing angles. In actual implementations, the video is usually captured by an omnidirectional camera or a collection of cameras with a wide field of view.
- 360^{o} Video can be monoscopic or stereoscopic. The latter is especially suitable for VR applications.

3D Video and TV

- Three-dimensional (3D) pictures and movies have been in existence for decades.
- The rapid progress in R&D of 3D technology and the success of the 2009 film Avatar have pushed 3D video to its peak.
- Increasingly, it is in movie theaters, broadcast TV (e.g., sporting events), PCs, and various hand-held devices.
- The main advantage of the 3D video is that it enables the experience of immersion — be there, and really Be there!

Cues for 3D Perception

Monocular Cues:

- Shading, Perspective scaling, Relative size, Texture gradient, Blur gradient, Haze, Occlusion, Motion parallax.
- Among the above monocular cues, it is said that Occlusion and Motion parallax are more effective.

Binocular Cues

- The human vision system uses binocular vision, i.e., stereo vision, aka. Stereopsis.
- Our left and right eyes are separated by a small distance, on average approximately two and half inches, or 65 mm. This is known as the interocular distance.
- The left and right eyes have slightly different views. The amount of the shift, or disparity, is dependent on the object’s distance from the eyes, i.e., its depth, thus providing the binocular cue for the 3D percept.
- Current 3D video/TV systems are almost all based on stereopsis because it is believed to be the most effective cue.

3D Camera Models

- **Simple Stereo Camera Model:**
- The left and right cameras are identical (same lens, same focal length, etc.). The cameras' optical axes are in parallel, pointing at the Z-direction, the scene depth.

- Disparity: $d = fb/Z$

where f is the focal length, b is the length of the baseline (camera separation), $d = x_l - x_r$ is the disparity or horizontal parallax.

- Zero disparity for objects at the infinity.

3D Movie and TV Based on Stereo Vision

- 3D Movie Using Colored Glasses: glasses tinted with complementary colors, usually red on the left and cyan on the right – Anaglyph 3D.

- 3D Movie Using Circularly Polarized Glasses: polarized glasses that the audience wears allows one of the two polarized pictures to pass through while blocking the other, e.g., in the RealD cinemas.

- 3D TV with Shutter Glasses: the liquid crystal layer on the glasses becomes opaque (behaving like a shutter) when some voltage is applied. The glasses are actively (e.g., via Infra-Red) synchronized with the TV set that alternately shows left and right images in a Time Sequential manner.

Vergence-Accommodation Conflict

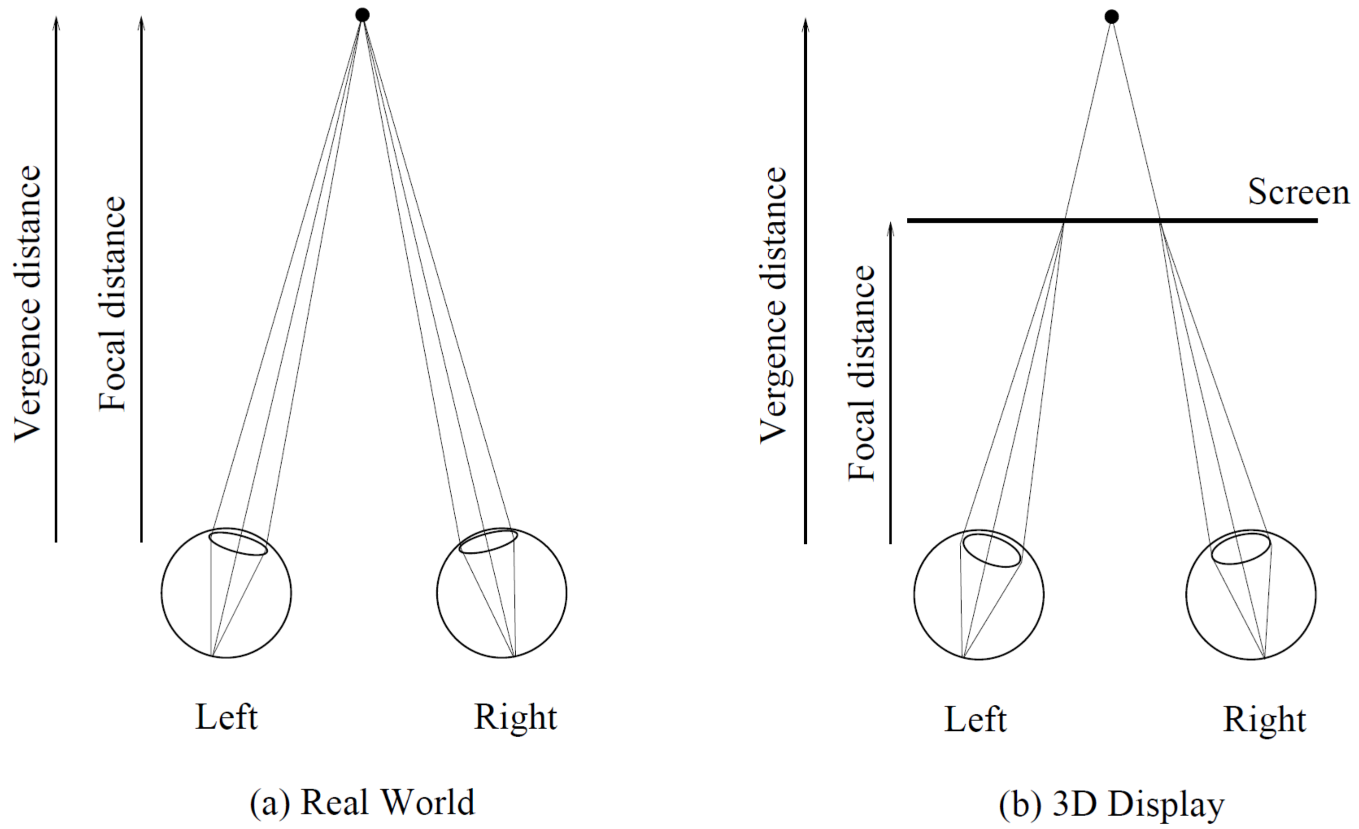
- ***Accommodation*** – to maintain a clear (focused) image on an object when its distance changes.
- As in Figure 5.9(a), in human vision, normally,

$$\textit{Focal distance} = \textit{Vergence distance}.$$

- As in Figure 5.9(b), most 3D video/movie/TV viewing requires

$$\textit{Focal distance} \neq \textit{Vergence distance}.$$

- This creates **the Vergence-Accommodation Conflict** – one of the main reasons for eye fatigue and strain while viewing 3D video/TV/movie.



(a) Real World

(b) 3D Display

- **Fig. 5.10:** The Vergence-Accommodation Conflict.

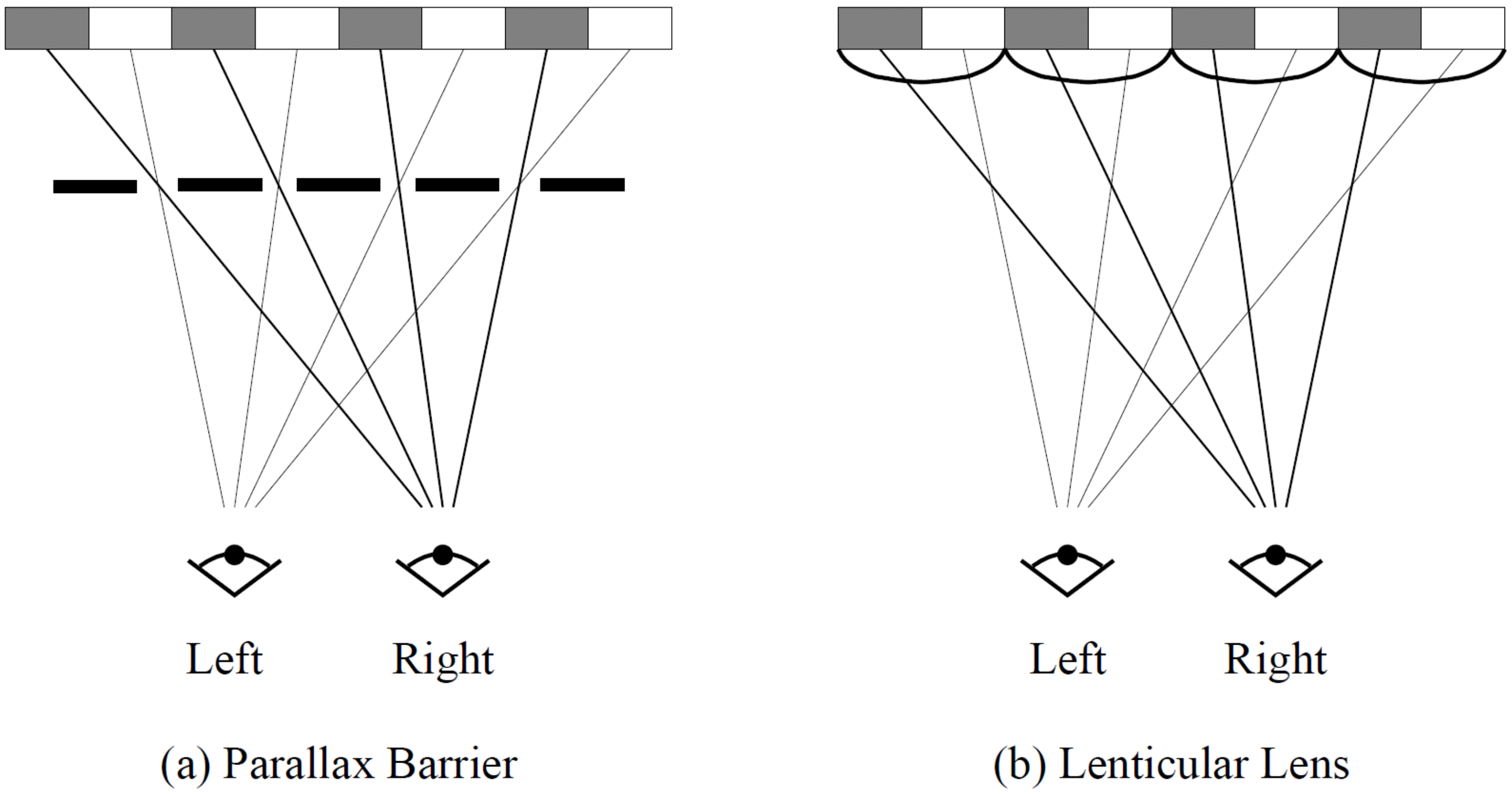
Autostereoscopic (Glasses-Free) Display

- **Parallax Barrier** – a layer of opaque material with slits is placed in front of the normal display device, e.g., an LCD. See Fig. 5.10(a). Used in e.g., Nintendo 3DS, Fujifilm 3D camera, and Toshiba's glasses-free 3D TV.

- **Lenticular Lens** – Instead of barriers, columns of magnifying lenses can be placed in front of the display to direct lights properly to the left and right eyes. See Fig. 5.10(b).

- **Integral Imaging** – Instead of cylindrical lenses as shown above, an array of spherical convex microlenses can be used to generate a large number of distinct micro-images. It enables the rendering of multiple views from any directions.

Autostereoscopic Display



Autostereoscopic Display Devices.

Disparity Manipulation in 3D Content Creation

- Disparity Range — map (often suppress) the original disparities into the range that will fit in the comfort zone of most viewers.
- Disparity Sensitivity — human vision is more capable of discriminating different depths when they are nearby, so do nonlinear disparity mapping (suppress the far range).
- Disparity Gradient — human vision has a limit of disparity gradient in binocular fusion, so avoid it in depth editing.
- Disparity Velocity — we cannot rapidly process large accommodation and vergence changes (i.e., disparity changes), so slow it down!

Video Quality Assessment (VQA)

- When comparing the coding efficiency of different video compression methods, a common practice is to compare the bitrates of the coded video bitstreams at the same quality.
- The video quality assessment approaches can be objective or subjective: the former is done automatically by computers and the latter requires human judgment.
- The most common criterion used for the **objective assessment** is *peak signal-to-noise ratio (PSNR)*, where I_{max} is the maximum intensity value, e.g., 255 for 8-bit images, *MSE* is the *Mean Squared Error* between the original image I and the compressed image I'.

$$PSNR = 10 \log_{10} \frac{I_{max}^2}{MSE}$$

Subjective Assessment

- The main advantage of PSNR is that it is easy to calculate. However, it does not necessarily reflect the quality as perceived by humans, i.e., visual quality.
- For subjective assessment, the original and compressed video clips are shown in succession to the human subjects. The subjects are asked to grade the compressed videos by their quality, 0 — lowest, 10 —highest. The Mean Opinion Score (MOS) is used as the measure for the subjective quality.
- Extensively employed in the development of the video compression standards such as H.265 and H.266, especially for HDR videos.

Other VQA Metrics

- The main efforts are to find better metrics than the simple measures such as PSNR, so that the assessments can be conducted objectively (by computers) and their results will be comparable to those of human subjects.
- Many VQA algorithms are derived from their counterparts in Image Quality Assessment (IQA).
- Wang et al. presented the Structural Similarity (SSIM) index that captures some simple image structural information (e.g., luminance and contrast). It has become very popular as a metric in IQA and VQA.

Summary

- Standard formats for video
- Main properties of video
- Deinterlacing
- Different kinds of codecs
- MPEG compression
- Motion estimation
- Main features of MPEG-4
- 3D video
- VQA