Unit 4
Digital Video

CS 3570
Chen-Kuo Chiang
CS Dept., NTHU

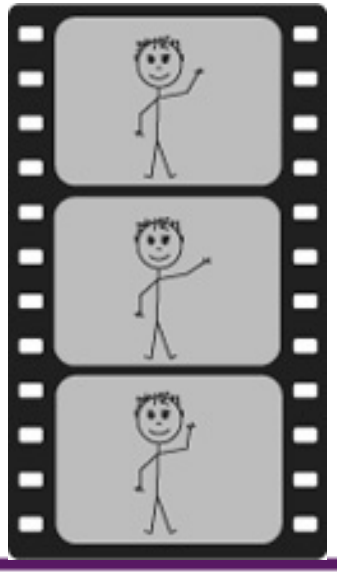
Video, Film, and Television Compared

- 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.

Frame rate

- Film and television create moving pictures—by a fast sequence of images, called **frames.**
- The speed at which images are shown is the ***frame rate.***
- A frame rate of about 30 frames per second is needed in order for successive images to be perceived as smooth motion with no flicker.
- **Sprocket holes**—also called ***perforations***—are holes on the sides of the film used to pull the film through the projector.

4-perf 35mm film



Standard film aspect ratios

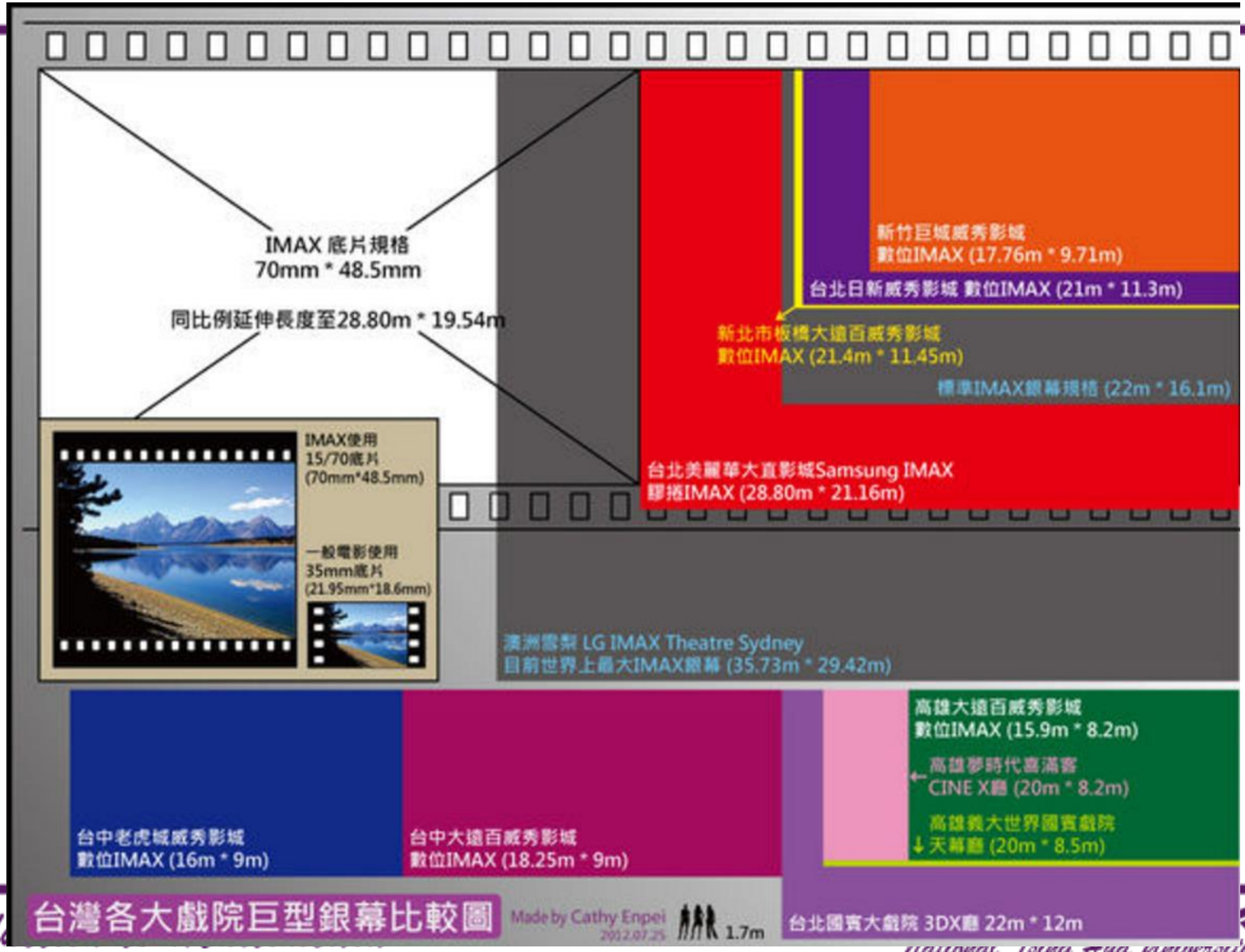
- Silent movies and early sound movies were shot mostly on 16 mm film, introduced by Eastman Kodak in 1923.
- ***Aspect ratio*** is the ratio of the width to the height of a frame, expressed as *width:height*.
- IMAX movies are shot on 70 mm film with aspect ratio of 1.44:1.
- **IMAX** movies are on very large screens, so the frames have to be enlarged more than they are in standard movie projection.

Academy aspect ratio 1.33:1

Widescreen, 1.85:1

Anamorphic widescreen, 2.39:1

台灣各影城IMAX銀幕比較



台灣各影城IMAX銀幕比較

- 台灣前四大IMAX
  - 大直-美麗華戲院 IMAX
    - ✓寬28.8 x 高21.16 (公尺) 約八層樓高
  - 西門町-日新威秀 IMAX
    - ✓寬21 x 高11.3 (公尺)
  - 西門町樂聲-樂聲廳
    - ✓寬22 x 高8.5 (公尺)
  - 西門町國賓數位大螢幕
    - ✓寬17.4 x 高7.8 (公尺)
- 台灣最小IMAX
  - 高雄大遠百,寬15.9 x 高8.2 (公尺)

Aspect Ratio 的影響

- IMAX比例為1.44:1，若在數位IMAX 1.78:1(紅框)或寬螢幕戲院 2.35:1(黃框)，畫面必須裁切。小丑面具沒了...

Standard-definition television

- In the beginning, television was transmitted as an analog signal. In comparison to the newer HDTV, we now sometimes refer to this as ***SDTV*** (***standard-definition television***).
- SDTV was broadcast through radio waves by land-based broadcast stations .
- Direct Broadcast Satellite (DBS) is received directly in the home, which must be equipped with a satellite dish.

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: 1920 × 1080 using interlaced scanning(**1080i**), 1920 × 1080 using progressive scanning(**1080p**), or 1280 × 720 using progressive scanning (**720p**).
- Digital encoding is not part of this definition, and, historically, HDTV was not always digital.

Frame Size (Resolution) Comparison between Standard Definition and High Definition

Same frame as 720p

Frame Size (Resolution) Comparison between Standard Definition and High Definition

A frame from a 1080i video

Frame Size (Resolution) Comparison between Standard Definition and High Definition

Same frame as standard definition DV wide-screen (16:9)

Same frame as standard definition DV standard 4:3

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***).

Three Video Standards

- **NTSC** governs standards in North America, Japan, **Taiwan**, and parts of the Caribbean and South America.
  - NTSC was instrumental in helping the television industry move from monochrome transmission to color.
- In 1967 **PAL** was adopted for color television broadcasts in the United Kingdom and Germany.
  - PAL has a number of variants that are now used in Europe, Australia
- **SECAM** was developed in France and accepted for color broadcasting in 1967.
  - It was later adopted by other countries in Eastern Europe.

Video and Film displays

- Like film, video is created by a sequence of discrete images, called **frames**, shown in quick succession.
- A film frame is a continuous image, enlarged and projected as a whole onto the movie screen.
- 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.***

Raster scanning

- The scanning process is a movement from left to right and top to bottom.
- When the scanner has finished with one line, it moves back to the left to start another in a motion called ***horizontal retrace***.
- ***Vertical retrace*** takes the scanner from the bottom of the monitor to the top again.
- In the case of video camera, the purpose of the scanning is to record the data that will be saved and/or transmitted as the video signal.

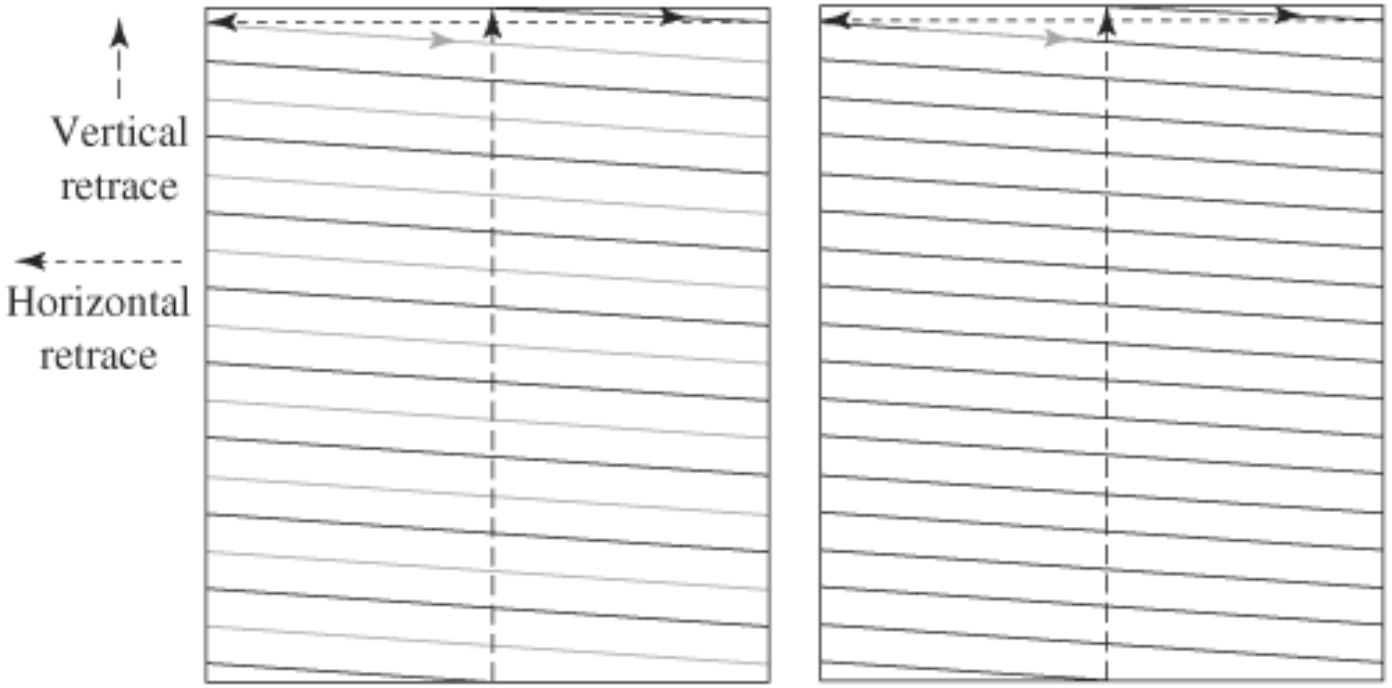
Raster scanning

- 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

Raster 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:

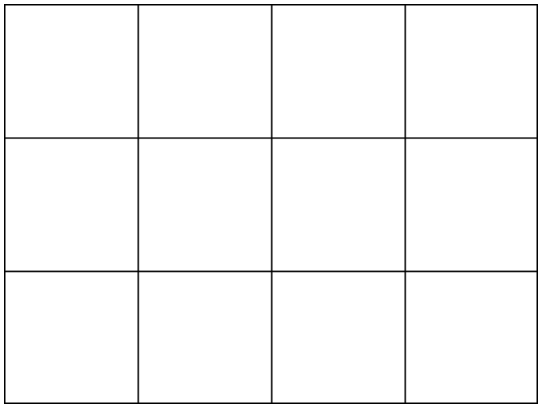
Lines displayed in order from top to bottom

Interlaced and progressive scanning

Native resolution

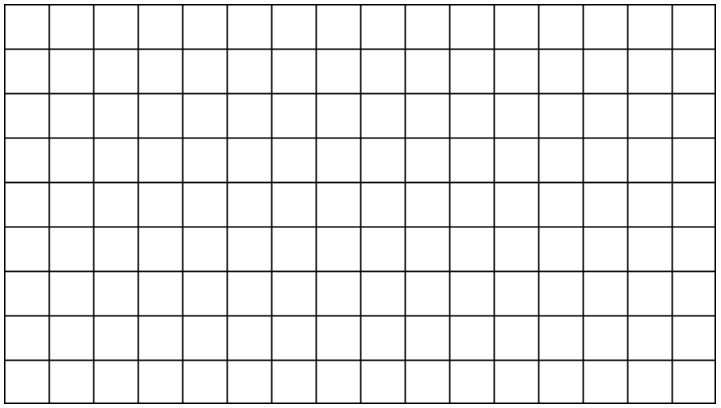
- A 720p television accepts a signal with 1280×720 pixels per frame and displays them with progressive scanning.
- However, it may display a different resolution that doesn't have 1280×720 pixels.
  - **1080i**: 1920×1080 using interlaced scanning
  - **1080p**: 1920×1080 using progressive scanning
  - **720p**: 1280×720 using progressive scanning
- For each frame, the ***logical pixels***—pieces of information saved and transmitted in a video signal—have to be mapped to the ***physical pixels***—points of light on the video display.

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

Seam Carving

- A smart way to change aspect ratio

Seam Carving

(a) Original (b) Crop (c) Column (d) Seam (e) Pixel (f) Optimal

Figure 2: Results of 5 different strategies for reducing the width of an image. (a) the original image and its e_1 energy function, (b) best cropping, (c) removing columns with minimal energy, (d) seam removal, (e) removal of the pixel with the least amount of energy in each row, and finally, (f) global removal of pixels with the lowest energy, regardless of their position. Figure 3 shows the energy preservation of each strategy.

The Operator

- Given an energy function, let **I** be an n×m image

$$e_1(\mathbf{I}) = |\frac{\partial}{\partial x}\mathbf{I}| + |\frac{\partial}{\partial y}\mathbf{I}|$$

- define a vertical seam

$$\mathbf{s}^{\mathbf{x}} = \{s_i^x\}_{i=1}^n = \{(x(i), i)\}_{i=1}^n, \text{ s.t. } \forall i, |x(i) - x(i-1)| \leq 1,$$

where ***x*** is a mapping ***x*** : [1, . . . ,n] → [1, . . . ,m].

- a vertical seam
  - is an 8-connected path of pixels in the image from top to bottom,
  - containing one, and only one, pixel in each row of the image
- The pixels of the path of seam s (e.g. vertical seam {si}) will therefore be

$$\mathbf{I}_\mathbf{s} = \{\mathbf{I}(s_i)\}_{i=1}^n = \{\mathbf{I}(x(i), i)\}_{i=1}^n$$

Vertical & Horizontal Seams

The Operator

- Given an energy function e, we can define the cost of a seam as

$$E(\mathbf{s}) = E(\mathbf{I}_\mathbf{s}) = \sum_{i=1}^{n} e(\mathbf{I}(s_i))$$

- look for the optimal seam s^* that minimizes this seam cost :

$$s^* = \min_{\mathbf{s}} E(\mathbf{s}) = \min_{\mathbf{s}} \sum_{i=1}^{n} e(\mathbf{I}(s_i))$$

- The optimal seam can be found using dynamic programming.

The Operator

- The first step is
  - to traverse the image from the second row to the last row
  - and compute the cumulative minimum energy M for all possible connected seams for each entry (i, j):

$$M(i,j) = e(i,j) + \min(M(i-1,j-1), M(i-1,j), M(i-1,j+1))$$

- At the end of this process,
  - the minimum value of the last row in M will indicate the end of the minimal connected vertical seam.
- In the second step
  - backtrack from this minimum entry on M to find the path of the optimal seam .
- The definition of M for horizontal seams is similar.

Example

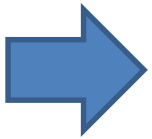
- Remove two vertical seams

5	10	25	87	93
66	11	20	3	45
0	47	88	35	57
3	5	64	31	44
10	8	39	22	39

Image

10	1	5	3	4
2	8	6	7	12
7	5	3	1	2
0	12	9	8	1
8	1	4	3	7

Energy map of image



Resized Image

Energy Preservation Measure

- examine the average energy of all of pixels in an image

$$\frac{1}{|\mathbf{I}|}\sum_{p\in I} e(p)$$

- removing the low energy pixels in ascending order
  - gives the optimal result.
  - This is closely followed by pixel removal.
  - But both methods destroy the visual coherence of the image.
- Cropping
  - shows the worst energy preservation.
- Column removal
  - does a better job at preserving energy,
  - but still introduce visual artifacts.
- Seam carving
  - strikes the best balance between the demands
    - ✓ for energy preservation
    - ✓ And visual coherency.

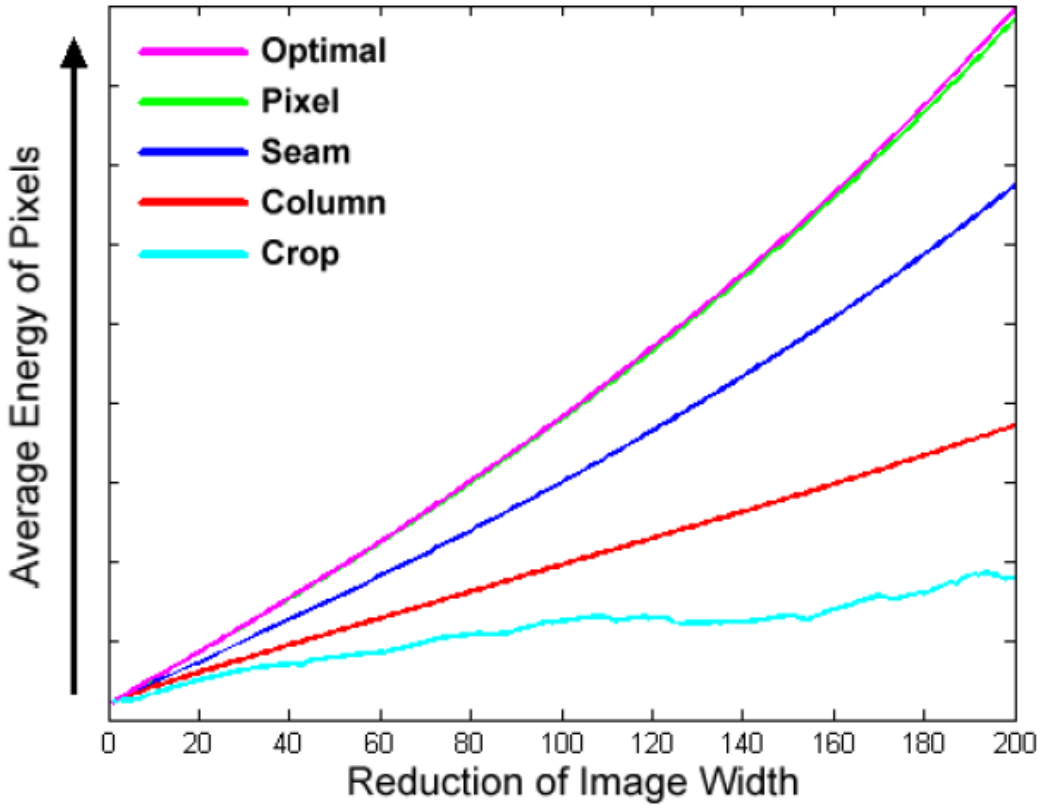


Figure 3: Image energy preservation. A comparison of the preservation of content measured by the average pixel energy using five different strategies of resizing. The actual images can be seen in Figure 2.

(a) Original

(b) e_1

(c) $e_{Entropy}$

(d) e_{HoG}

(e) Segmentation and L_1

Figure 4: Comparing different energy functions for content aware resizing.

Finding the missing shoes……

Figure 12: Object removal: find the missing shoe! (original image is top left). In this example, in addition to removing the object (one shoe), the image was enlarged back to its original size. Note that this example would be difficult to accomplish using in-painting or texture synthesis.

Find the Missing Person…

Original Image

Removing Ex-girl Friend

Input

Weights

Object Removed

Video connections

- In analog video, the color information can be sent in one of three types of *analog* ***video transmission formats***—component, S-video, or composite form.

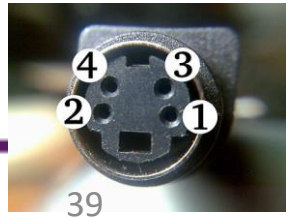
TABLE 6.3 Standards for Analog Video Recording Equipment

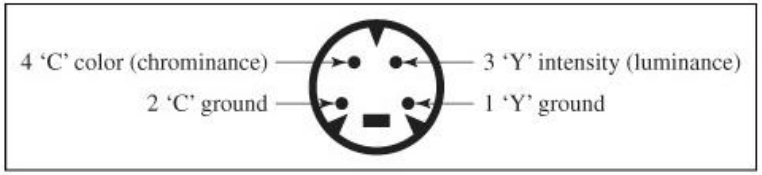
Video Format	Year Introduced	Color Transmission Format	Horizontal Resolution	Tape Width	Quality
VHS	1976	composite	~240	½" (12.5 mm)	consumer
Betamax	1976	composite	~240	½" (12.5 mm)	consumer
8mm (Video 8)	1984	composite	~240–300	8 mm	consumer
S-VHS	1987	S-video	~400–425	½" (12.5 mm)	high-end consumer
Hi-8	1998	S-video	~400–425	8 mm	high-end consumer
U-Matic	1971	composite	~250–340	¾" (18.75 mm)	professional
M-II	1986	component	~400–440	½" (12.5 mm)	professional
Betacam	1982	component	~300–320	½" (12.5 mm)	consumer
Betacam SP	1986	component	~340–360	½" (12.5 mm)	professional

Video transmission formats

- In ***component video*** a separate signal is sent for each part of the three luminance/chrominance components.
- Component video has three separate paths for the information and three connectors at the end.

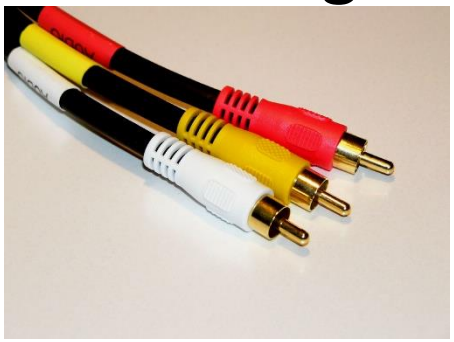
- ***S-video*** uses two data paths: one for the luminance and one for the two chrominance.
- An S-video jack has one connection at the display end, with two channels of information carried through the connection.





Video transmission formats

- ***Composite video*** is a video signal that is sent on just one channel.

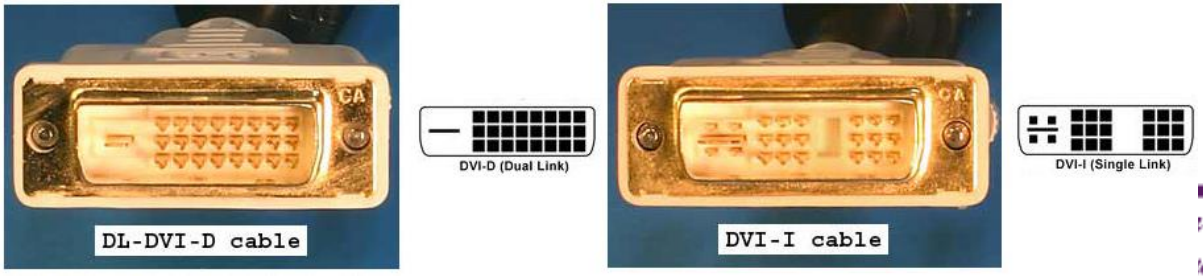


Composite video with two audio jacks

- Compositing the signal makes it possible to use just one broadcast channel through one physical connection from device to device.
  - Disadvantage to this technology is that crosstalk can occur between the color and luminance components, making composite video the lowest quality of all the alternatives.

Digital video transmission format

- There are two main types of digital video transmission format: ***DVI (digital video interface),*** and ***HDMI (high definition multimedia interface).***
- *DVI* connects an uncompressed digital video source (e.g., from a video card) to a digital display device.
- There are three basic DVI formats:
  - DVI-D, for a true digital-to-digital connection
  - DVI-A, connects and convert a digital signal to an analog display
  - DVI-I, have the flexibility to carry either DVI-D or DVI-A signals

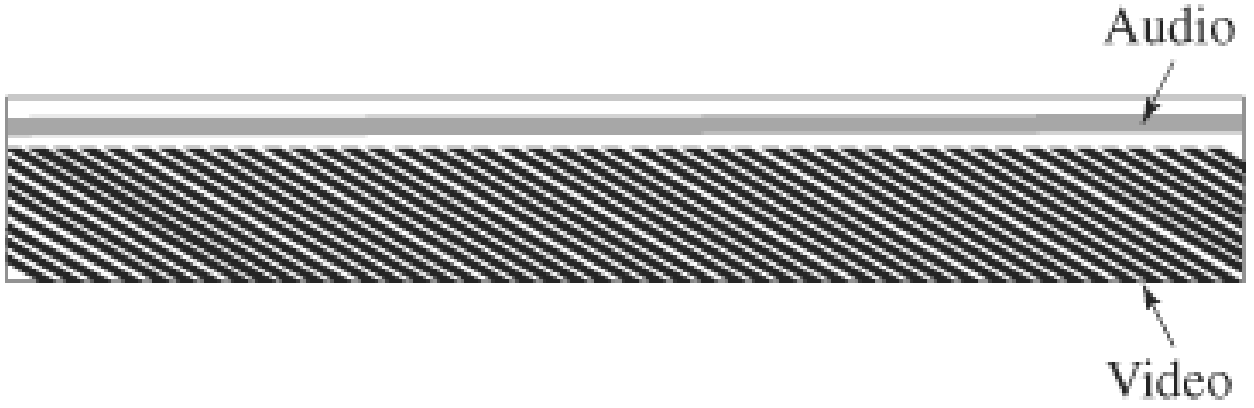


Digital video transmission format

- *HDMI* is an audio/video connection for transmitting uncompressed digital data.
- It is backward compatible with DVI and accommodates audio data on the signal.
- HDMI connections can apply ***HDCP (high bandwidth digital content protection)*** to signal.
  - HDCP is a digital copy protection protocol that prevents unauthorized copying.

Videotape

- Videotape is different from film. Instead of recording a whole frame in a rectangle, a video camera records an image line-by-line, on a magnetized piece of plastic.
- The audio track lies in a straight line along the edge, and the video information is written diagonally on the tape.



Digital video file

File Extension	Container File Type	Characteristics
.avi	Audio/Video Interleaved	A type of RIFF file designed for Windows Media. Created for the PC platform, now used on Mac and Linux also. Can be uncompressed, or compressed with a variety of codecs, including DivX, Cinepak, Indeo, MJPEG, or DV.
.mov (sometimes *.qt*)	QuickTime Movie	A multimedia container file framework created by Apple; it stores different media—including video, sound, and text—in different tracks. Cross-platform, for Mac, PC, and Linux. Accommodates a variety of audio and video codecs, including Sorensen, MPEG, Cinepak, and DivX.
.mpg (also *.mpeg, .m1v, .m2v, .m2t, .mp4, .mpv2*)	MPEG	A file that has been compressed with some version of the MPEG codec. MPEG-1 is greatly compressed with small resolution, for use on CD or web. MPEG-2 is the standard for video on DVD. MPEG-4 has highest compression rate and serves as a container file, modeled after QuickTime.
.flv, .f4v, .f4p, .f4a, .f4b	Flash video	Encoded audio and video streams playable by the Flash player, which can exchange audio, video, and data over RTMP (real-time messaging protocol) connections with Adobe Flash Media Server. Now used widely on the web.
.ogg	Ogg (by Xiph Foundation)	An open-source format, good for internet streaming.
.rm and *.rmvb*	Real Media	A proprietary file format with accompanying codecs, developed by RealNetworks; it works on multiple platforms, including Windows, Mac, Linux, and Unix.
.wmv (sometimes *.asf*)	Microsoft Windows Media	Originally Microsoft proprietary codec for Windows Media Player, standardized by SMPTE; uses Advanced Systems format (a container format) sometimes with .asf suffix. Uses its own codec.

Properties of 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***.

Properties of 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.

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…

History and Naming

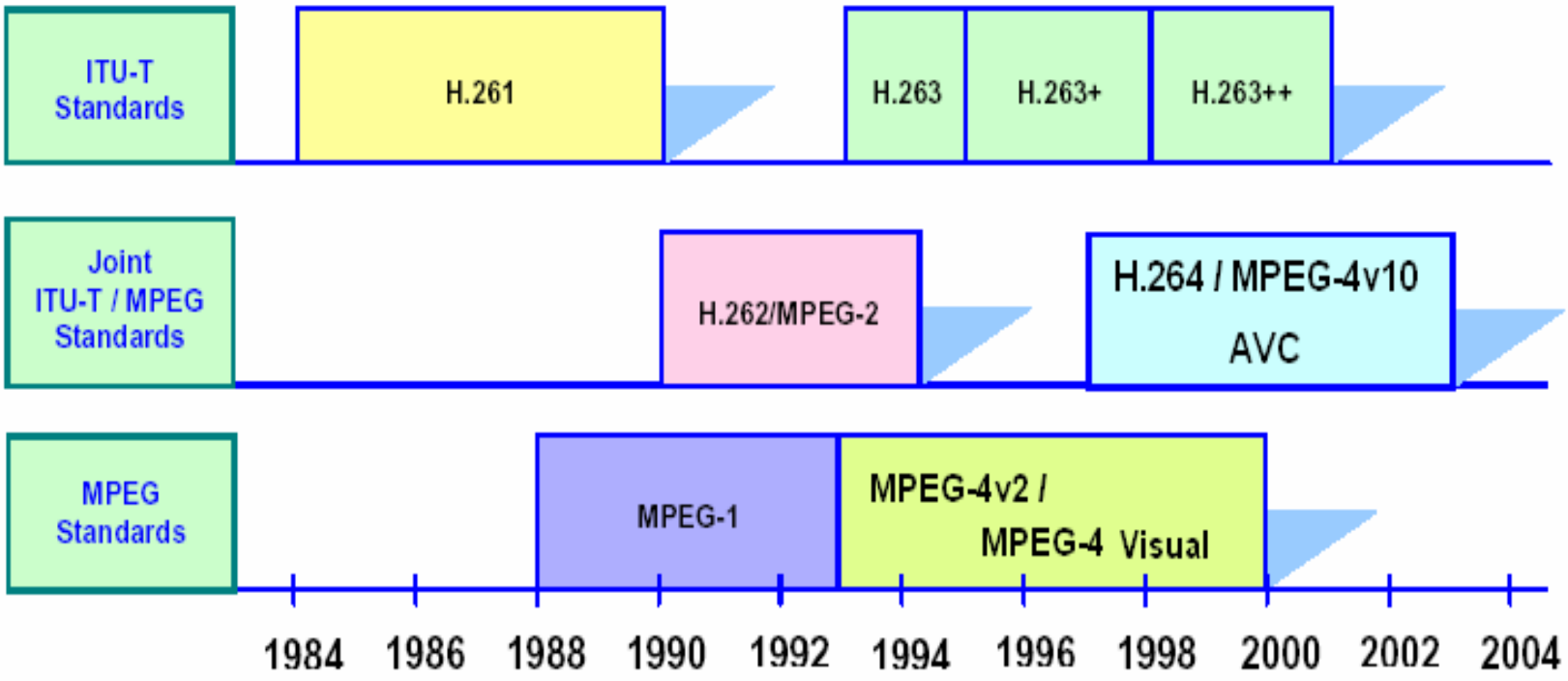
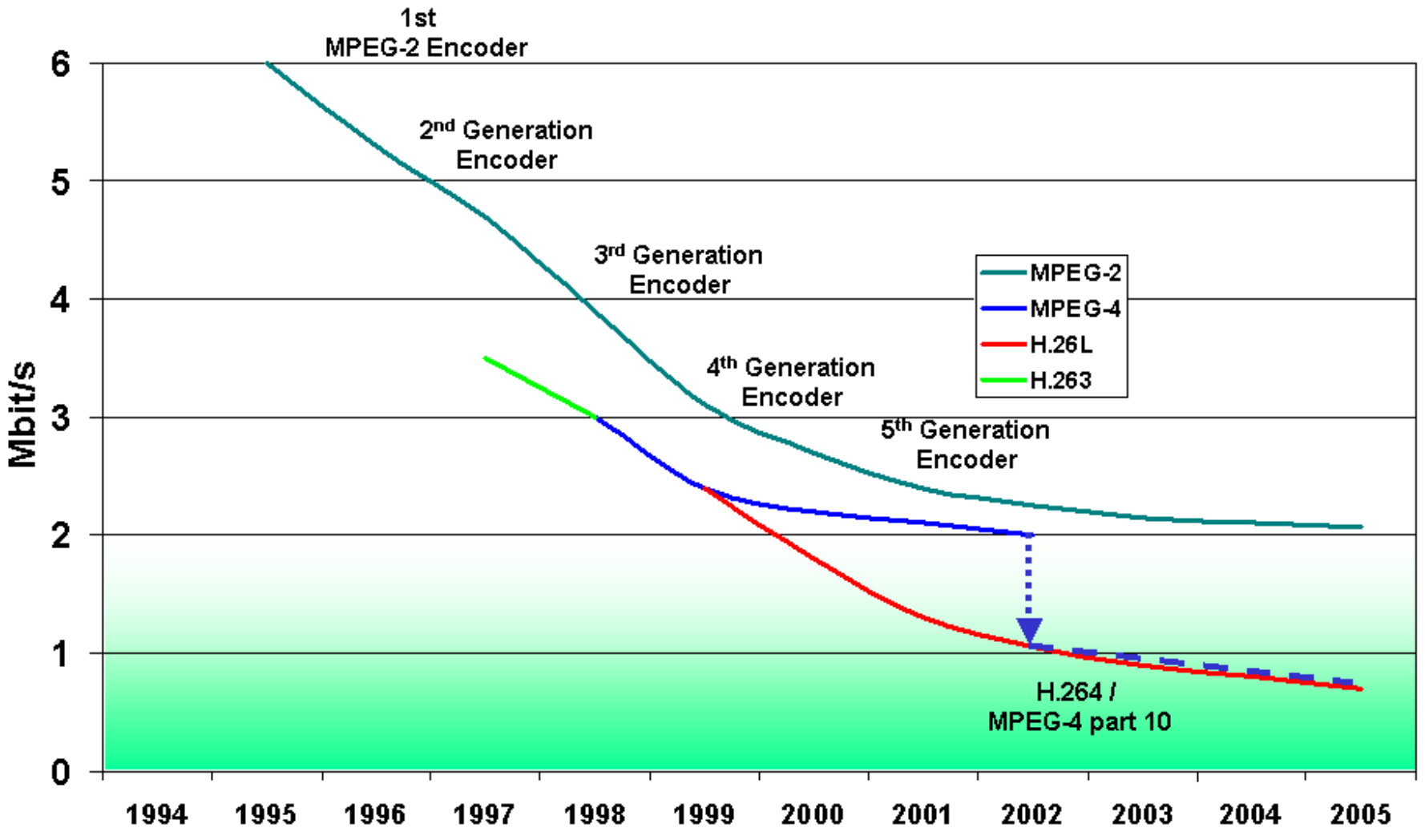


Figure 1. Progression of the ITU-T Recommendations and MPEG standards.

Goals of the H.264/AVC

- Video Coding Experts Group (VCEG), **ITU-T** SG16 Q.6
  - H.26L project (early 1998)
  - Target – double the coding efficiency in comparison to **any other existing video coding standards** for a broad variety applications.

Performance of video standard



Comparison to Previous Standards

Table 1.
Average bit rate savings for video streaming applications (from [10]).

Coder	Average Bit Rate Savings Relative To: MPEG-4 ASP	H.263 HLP	MPEG-2
H.264/AVC MP	37.44%	47.58%	63.57%
MPEG-4 ASP	–	16.65%	42.95%
H.263 HLP	–	–	30.61%

Table 2.
Average bit rate savings for video conferencing applications (from [10]).

Coder	Average Bit Rate Savings Relative To: H.263 CHC	MPEG-4 SP	H.263 Base
H.264/AVC BP	27.69%	29.37%	40.59%
H.263 CHC	–	2.04%	17.63%
MPEG-4 SP	–	–	15.69%

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 I_0 are the intensities (power) of two signals
- Peak Signal-to-Noise Ratio

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

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

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

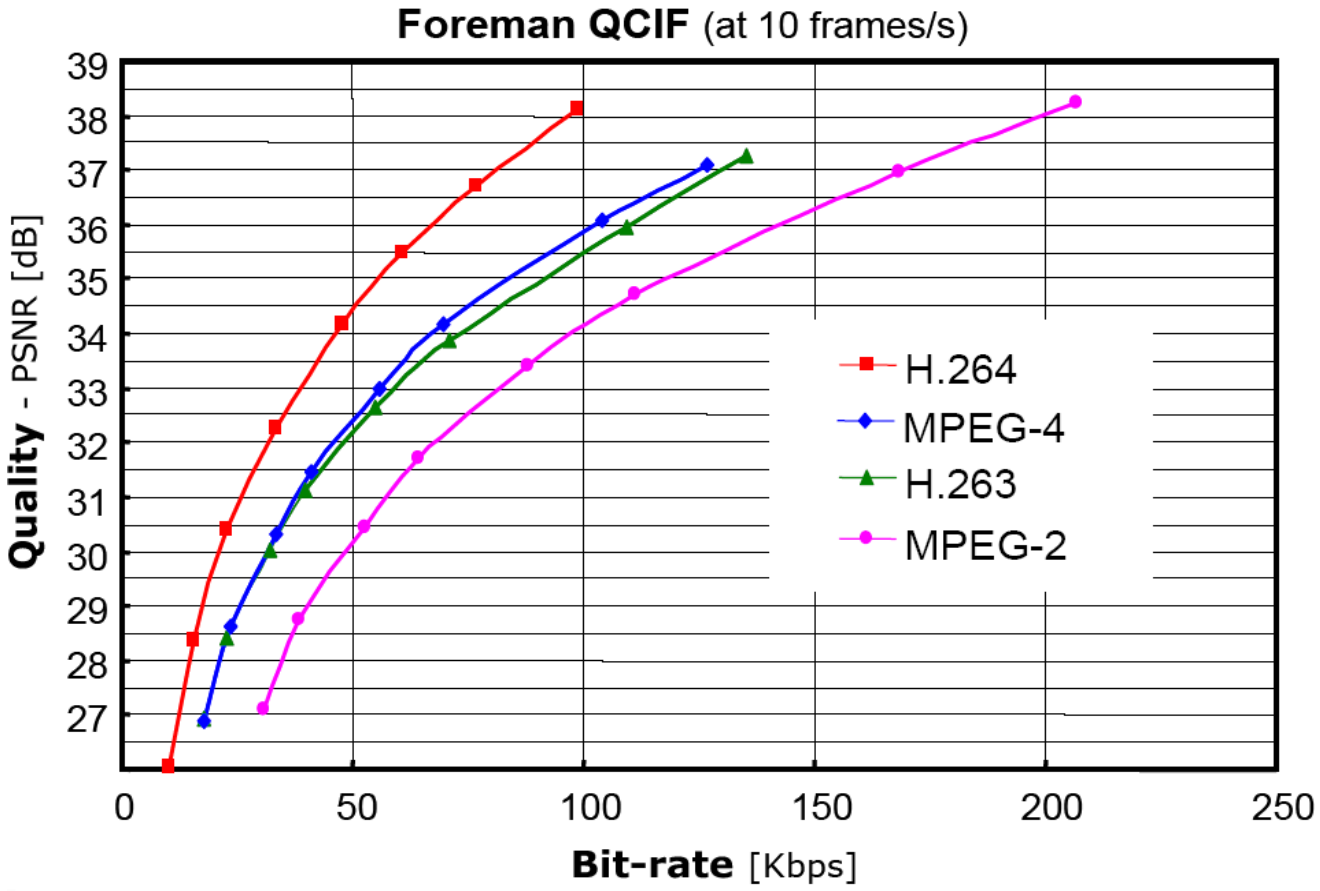
$$= 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE)$$

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

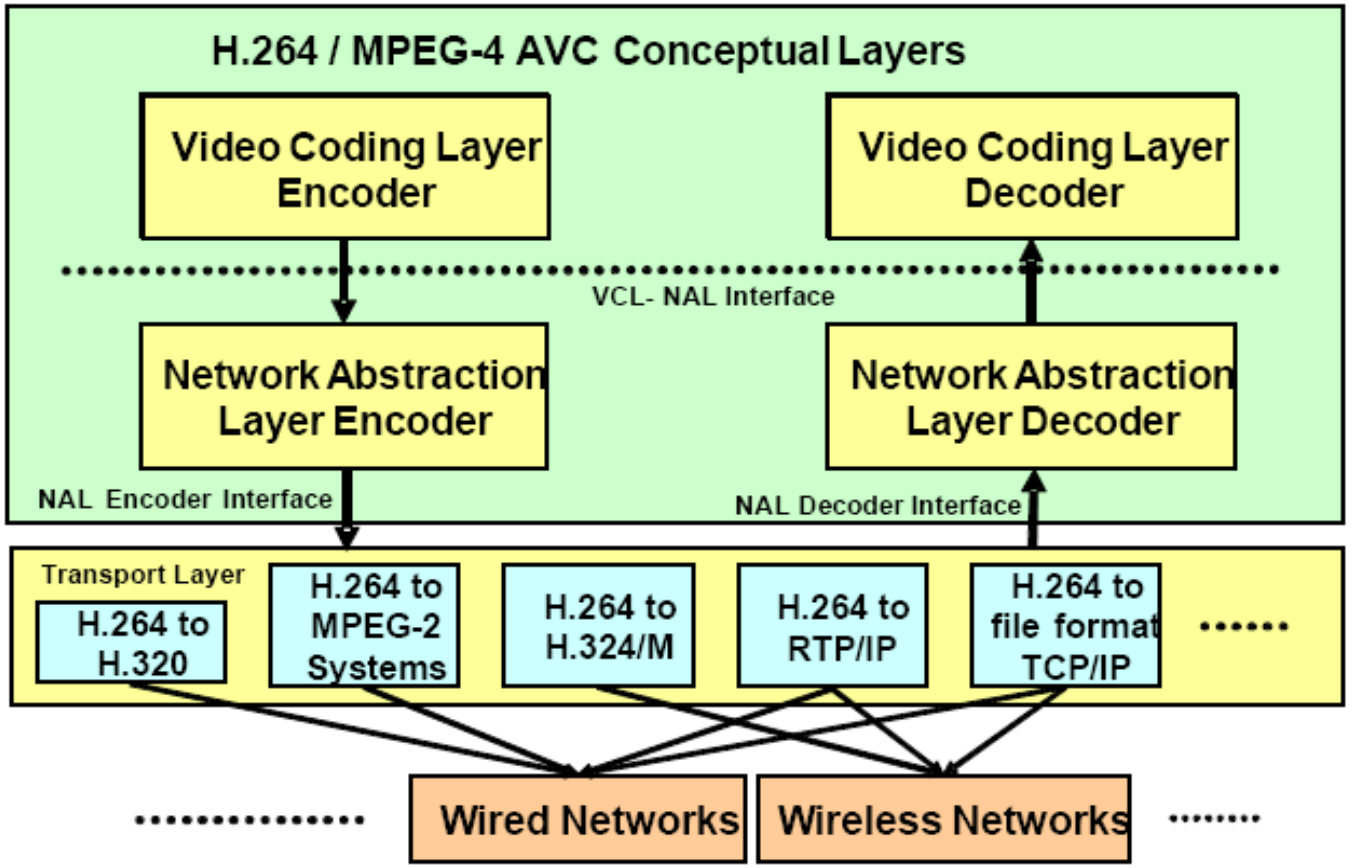
MAX is set to 255 if image is represented by 8 bit

Comparison to Other Standards

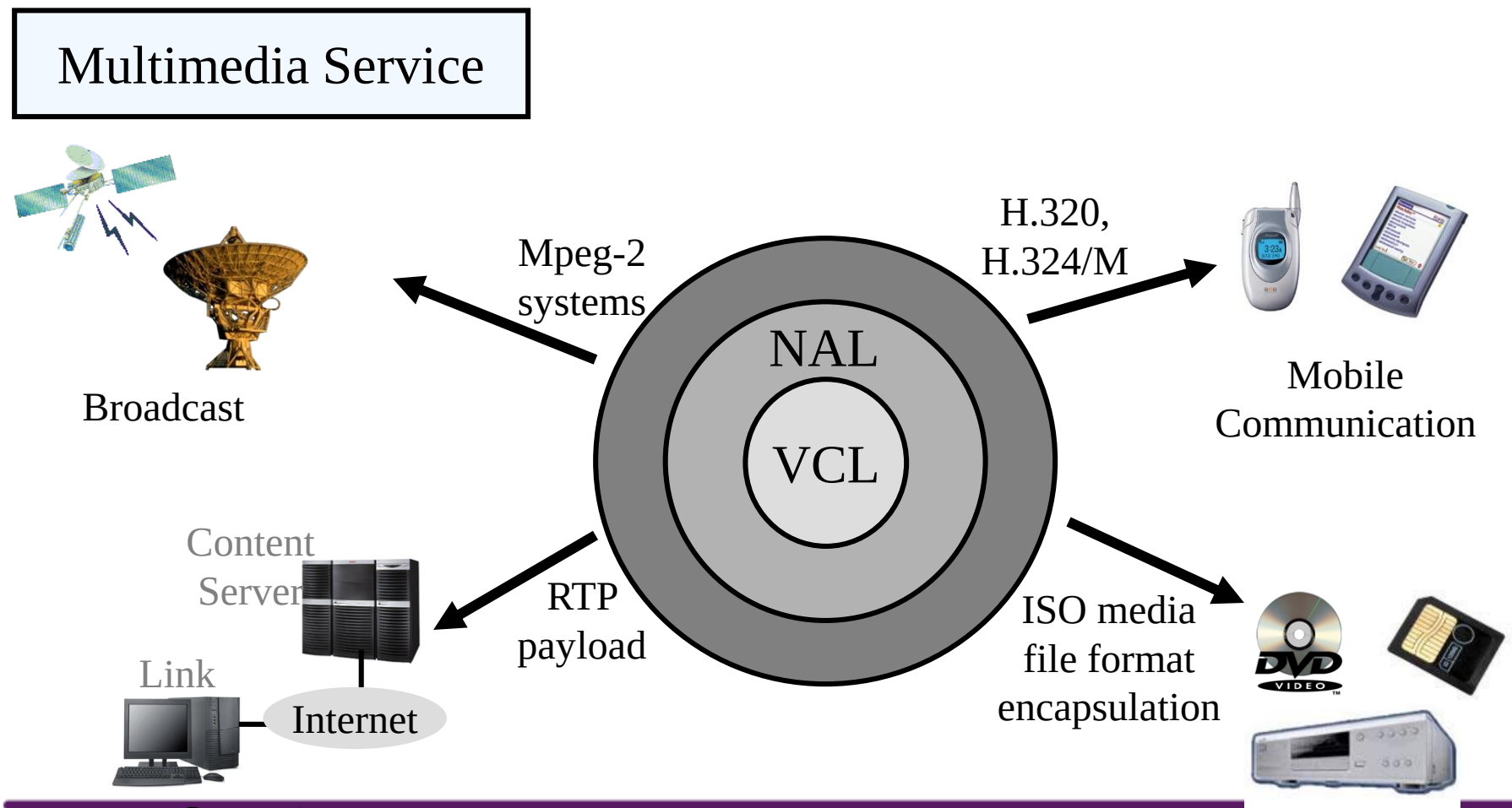
Rate-Distortion Curve



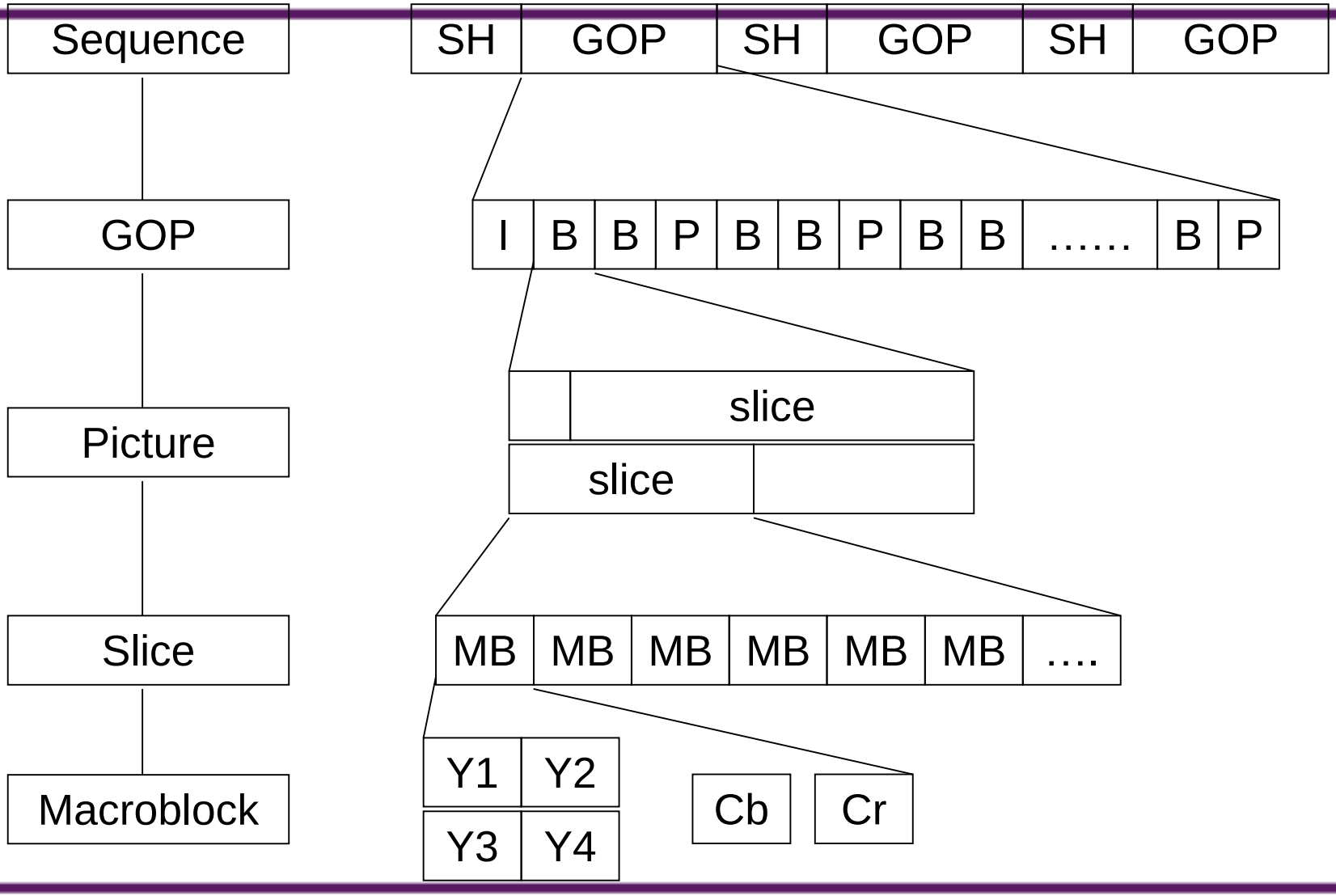
The Overall Conceptual Structure



Applications

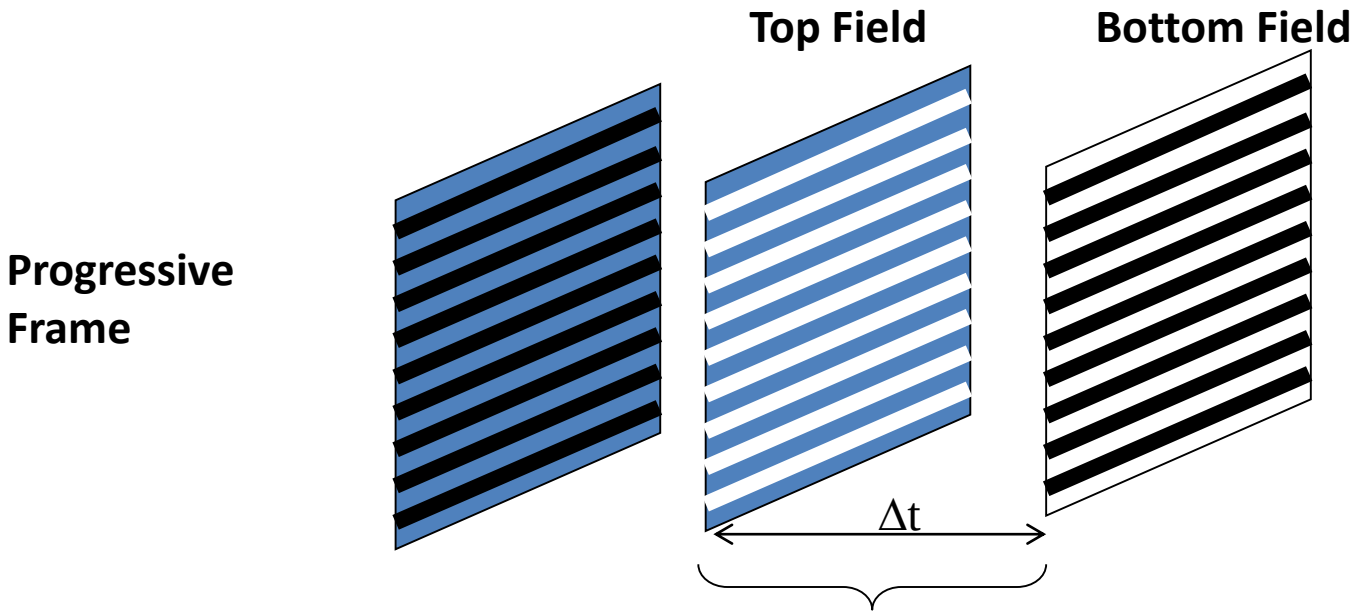


Data Structure



Video Coding Layer (VCL)

- The VCL follows the block-based hybrid video coding approach
- A sequence of coded pictures which can represent either an entire frame or a single field



Adaptive Frame/Field Coding Operation

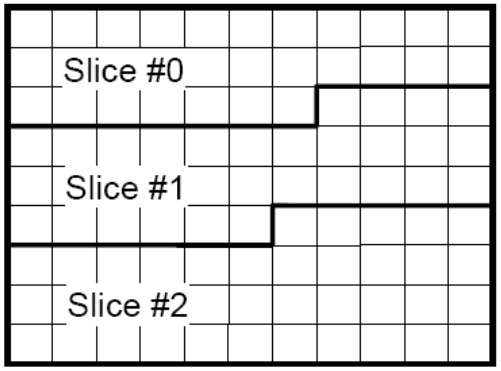
- Three modes can be chosen adaptively for each frame in a sequence.
  - Frame mode
  - Field mode

  Picture-adaptive frame/field (PAFF)
  16% ~ 20% save over frame-only
  for ITU-R 601 "Canoa", "Rugby", etc.

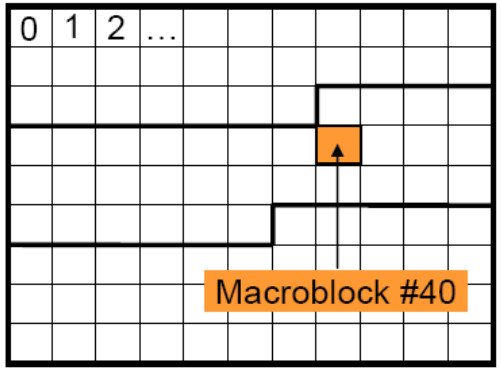
  - Frame mode / Field coded
- For a frames consists of mixed moving regions
  - Macroblock-adaptive frame/field (MBAFF)

Partitioning of a picture

- Slice
  - A picture is split into 1 or several slices
  - a sequence of macroblocks
  - minimal self-decodable unit



- Macroblocks
  - Basic syntax
  - Contains 16x16 luma samples and two 8x8 chroma samples



RGB → YC_bC_r

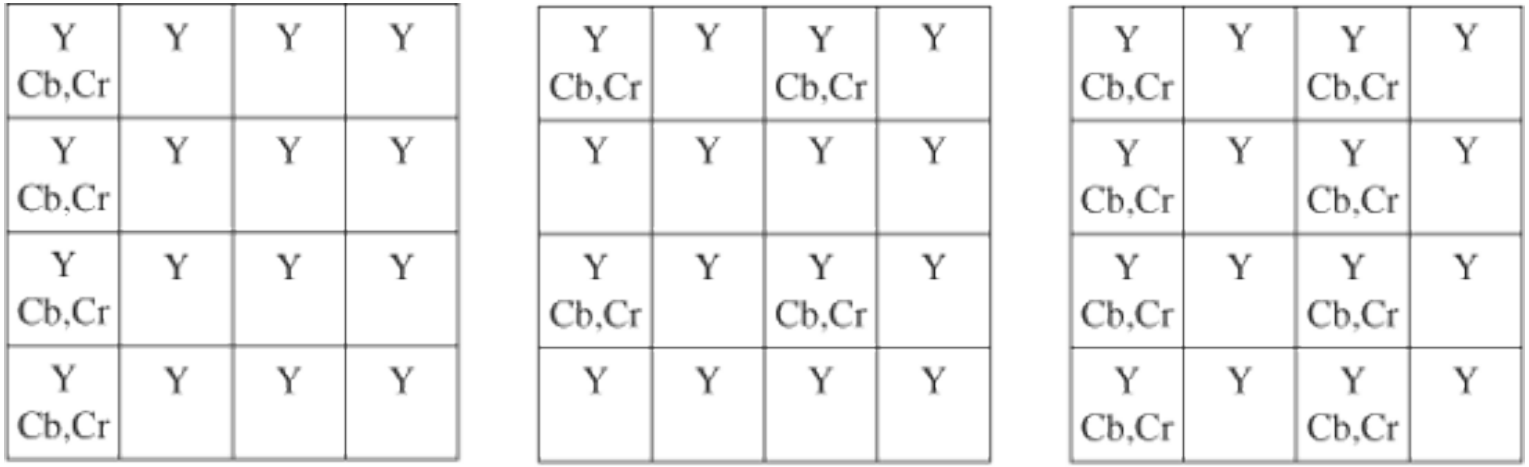
- YCbCr color model represents color in terms of one luminance component, Y, and two chrominance components, Cb and Cr.
- The human eye is more sensitive to changes in light (*i.e.*, luminance) than in color (*i.e.*, chrominance).

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1.164 & 0 & 1.596 \\ 1.164 & -0.392 & 0.439 \\ 1.164 & 2.017 & 0 \end{bmatrix} \left[\begin{bmatrix} Y - 16 \\ C_b - 128 \\ C_r - 128 \end{bmatrix} \right]$$

Chrominance subsampling

- luminance/chrominance subsampling is represented in the form *a*:*b*:*c*
- For each pair of four-pixel-wide rows, *a* is the number of Y samples in each rows, *b* and c are the numbers of C_b (C_r) samples in the 1st and 2nd rows, respectively.



Example

4:4:4

Y C_b,C_r	Y C_b,C_r	Y C_b,C_r	Y C_b,C_r
Y C_b,C_r	Y C_b,C_r	Y C_b,C_r	Y C_b,C_r
Y C_b,C_r	Y C_b,C_r	Y C_b,C_r	Y C_b,C_r
Y C_b,C_r	Y C_b,C_r	Y C_b,C_r	Y C_b,C_r

4:2:0

Y C_b,C_r	Y	Y C_b,C_r	Y
Y	Y	Y	Y
Y C_b,C_r	Y	Y C_b,C_r	Y
Y	Y	Y	Y

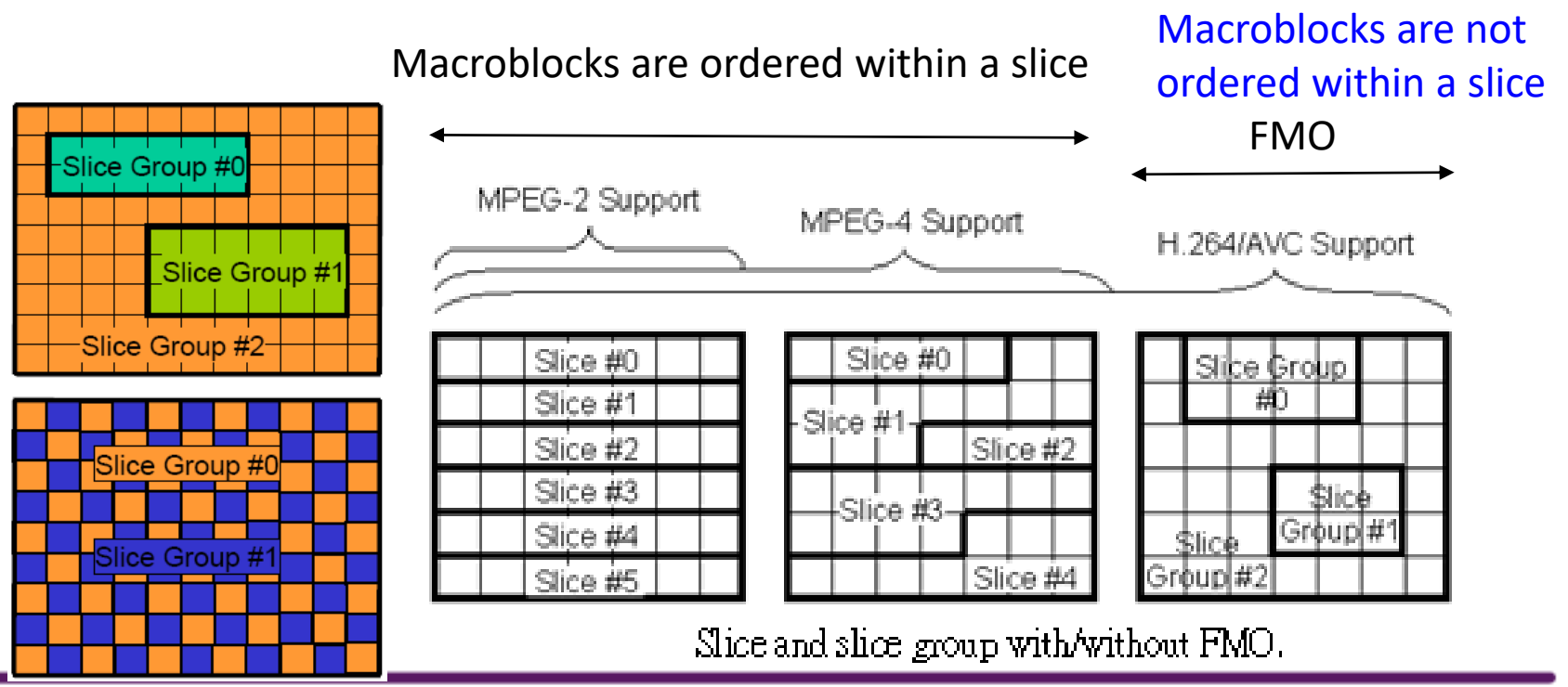
- Image resolution: 720 × 576 pixels
- Y resolution: 720 × 576 samples, each represented with 8 bits
  - 4:4:4 C_b, C_r resolution: 720 × 576 samples, each 8 bits
  - Total number of bits: 720 × 576 × 8 × 3 = 9953280 bits
  - 4:2:0 C_b, C_r resolution: 360 × 288 samples, each 8 bits
  - Total number of bits: (720 × 576 × 8) + (360 × 288 × 8 × 2) = 4976640 bits
- The 4:2:0 version requires half as many bits as the 4:4:4 version.

Field & Subsampling

- Allocaton of 4:2:0 samples to top and bottom fields

Flexible Macroblock Ordering (FMO)

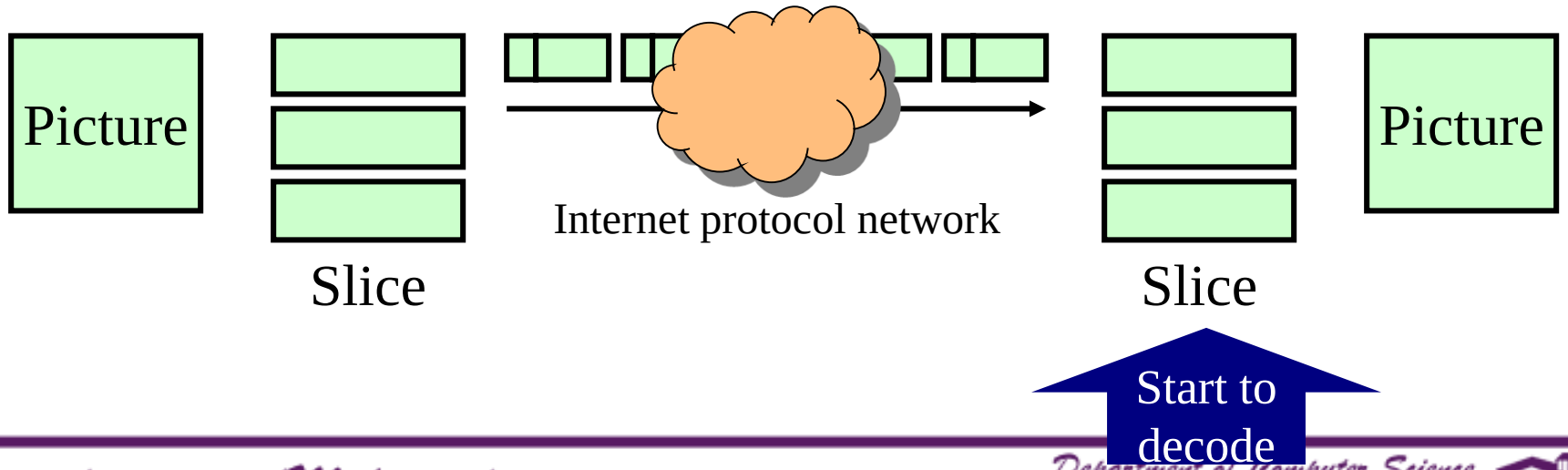
- Slice Group
  - Pattern of macroblocks defined by a **Macroblock Allocation Map**
  - A slice group may contain one to several slices



Slice and slice group with/without FMO.

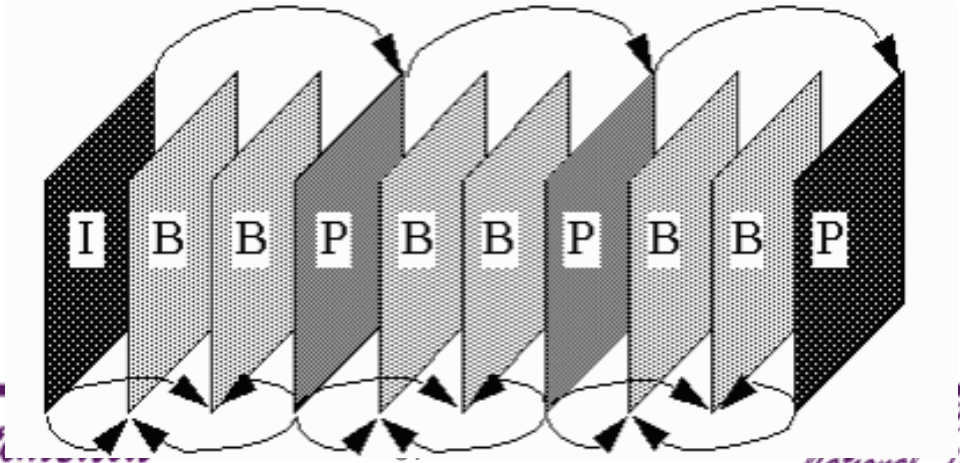
Arbitrary Slice Ordering

- ASO (Arbitrary Slice Ordering)
  - Independently-decoded Slice
    - ✓Enables sending and receiving the slice in any order
    - ✓Improve end-to-end delay in real-time application

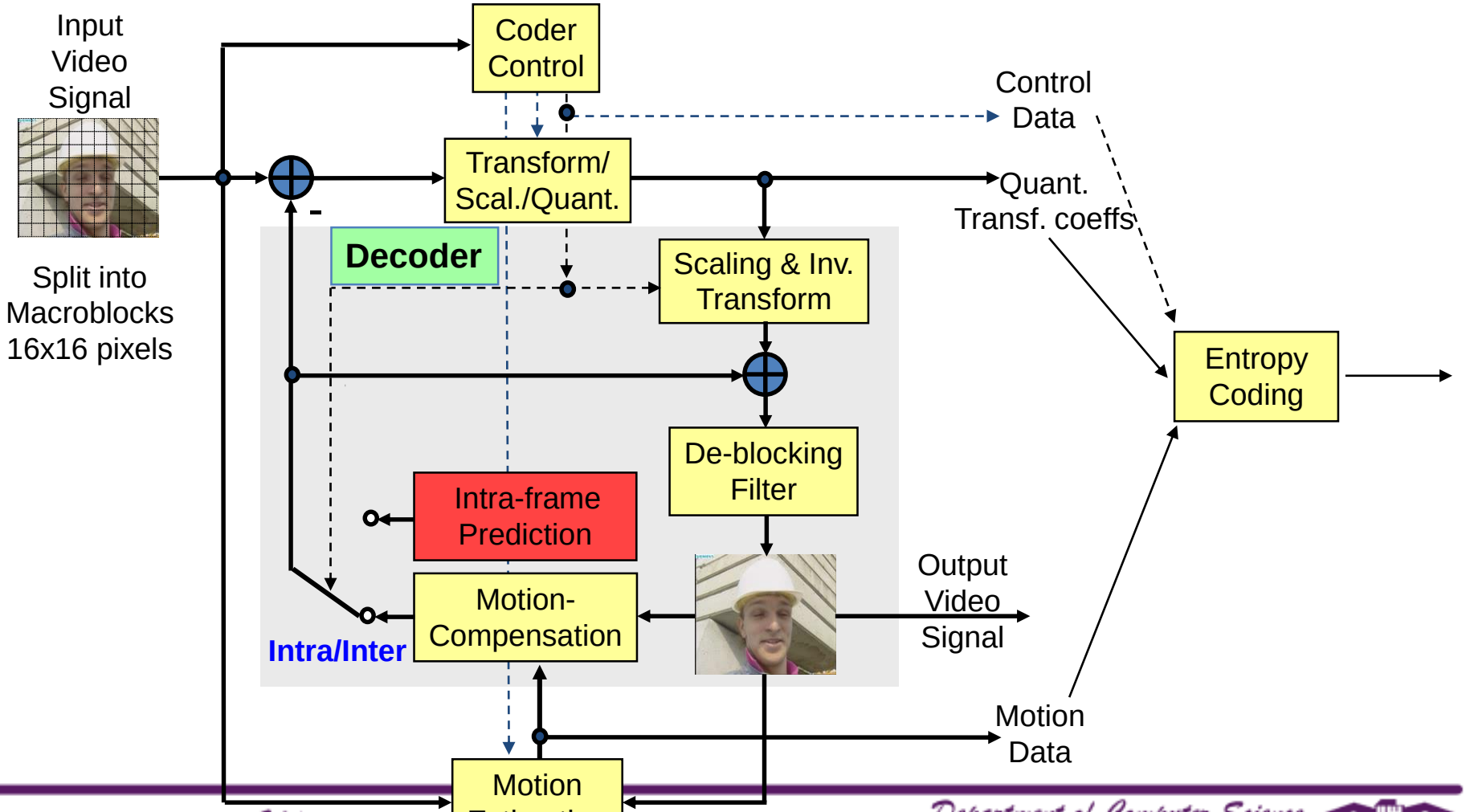


Coding the Pictures

- Each picture can be coded using different coding types
  - I frame**:** A picture in which all macroblocks of the picture are coded using **intra prediction**.
  - P frame**:** A picture which can be coded using **inter prediction** with at most **one** motion-compensated prediction signal per prediction block.
  - B frame**:** A picture which is coded using inter prediction with **two motion-compensated prediction** signals per prediction block.



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

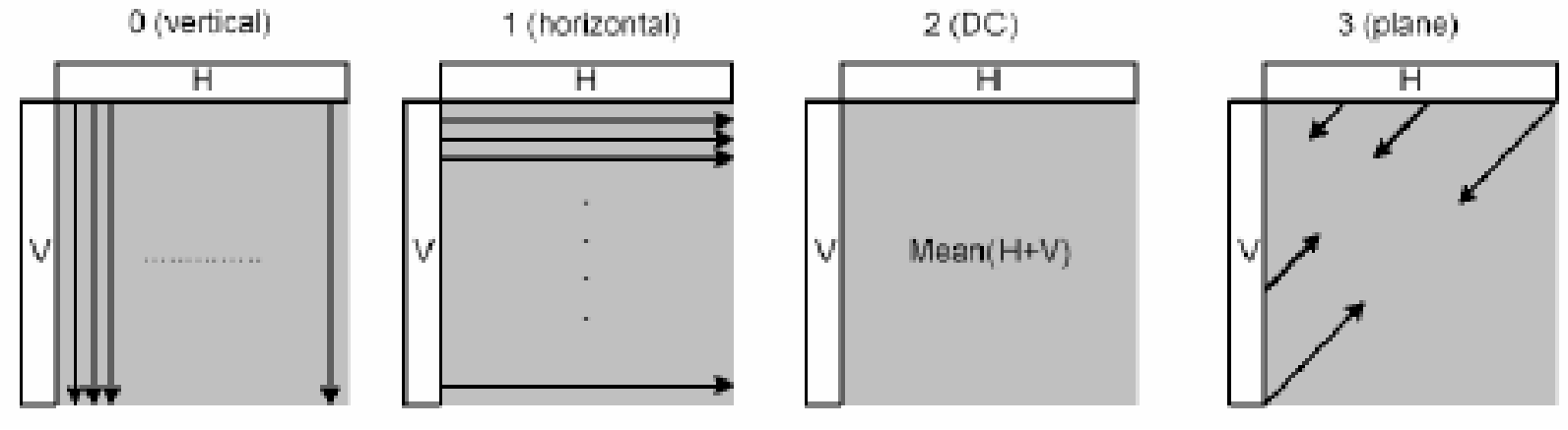


Intra-frame Prediction

- 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

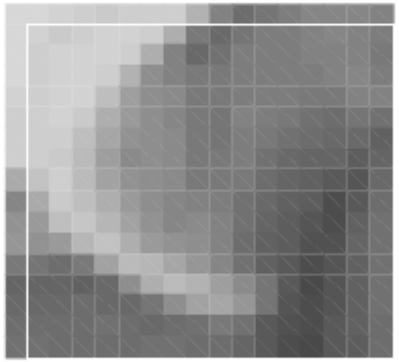
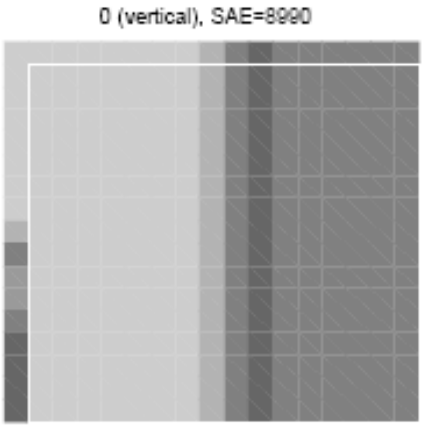
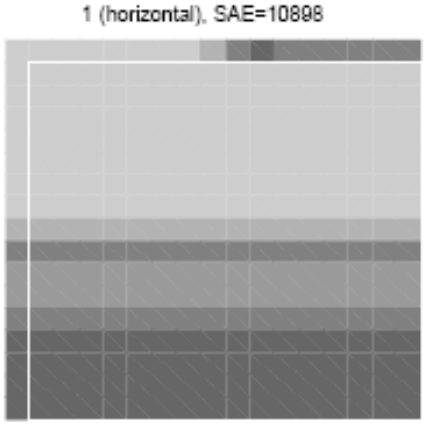
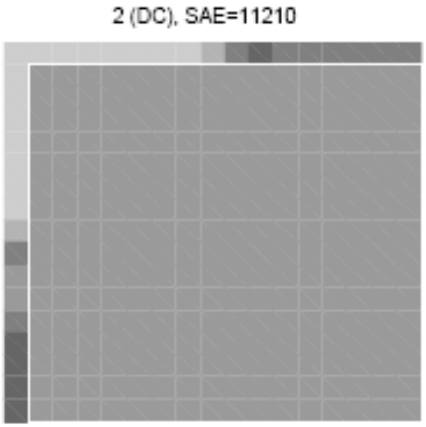
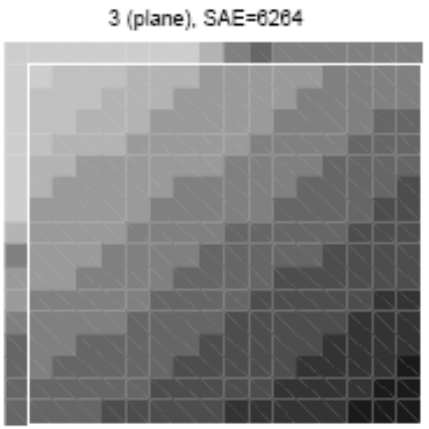


Figure 6 16x16 macroblock

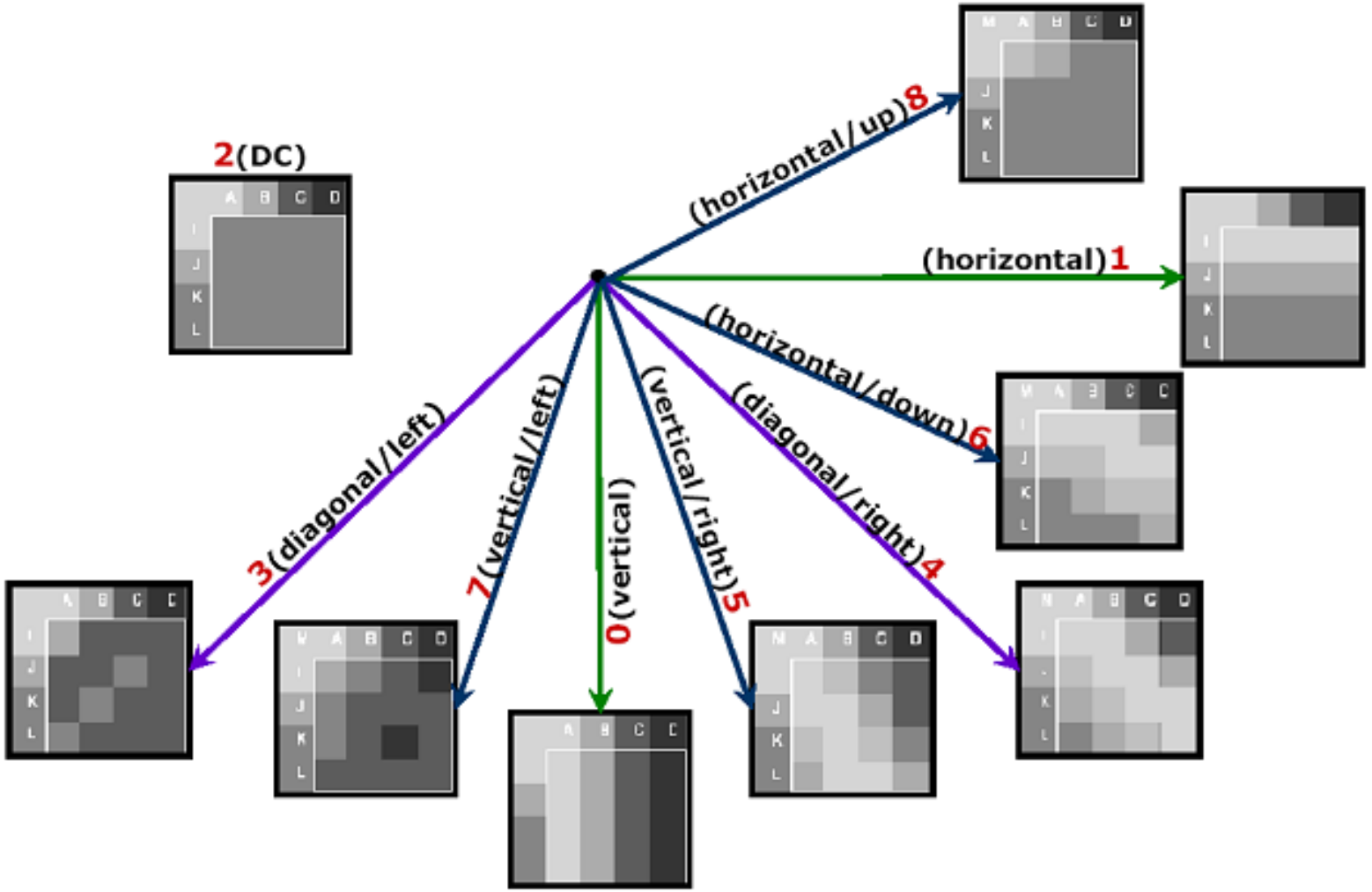






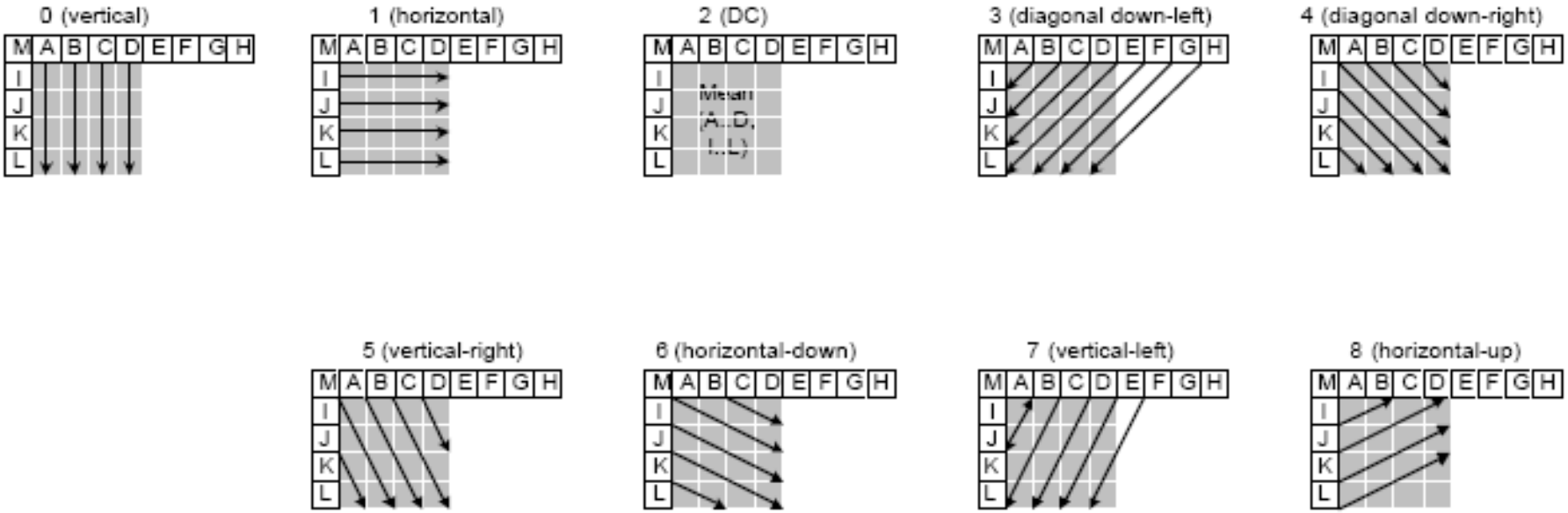


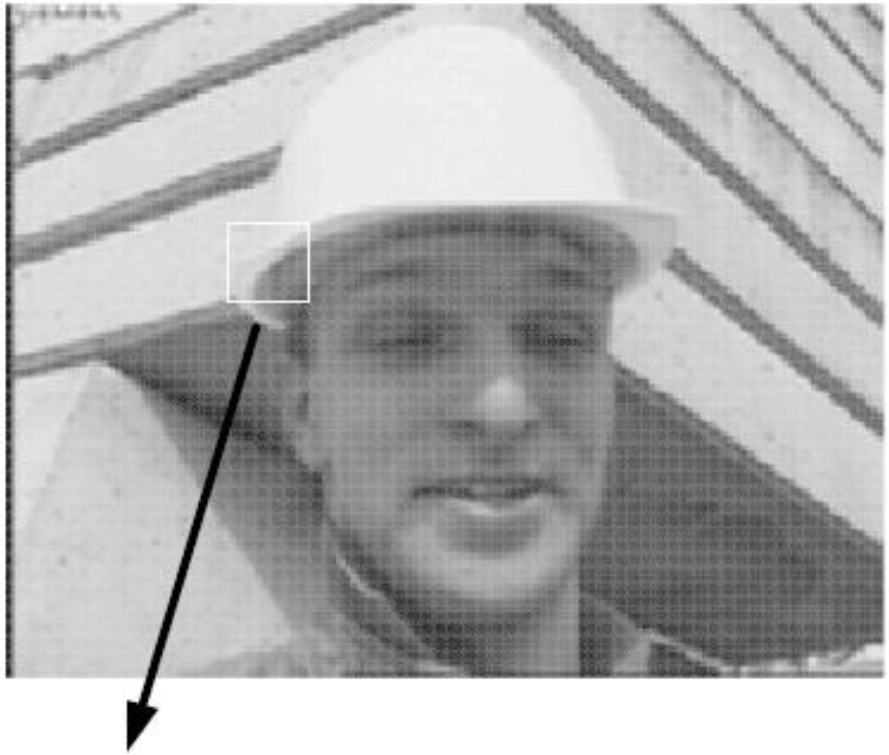
Nine modes of Intra_4×4



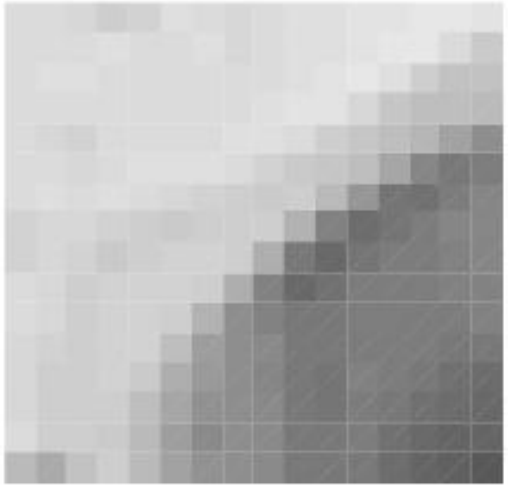
Nine modes of Intra_4×4

- 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

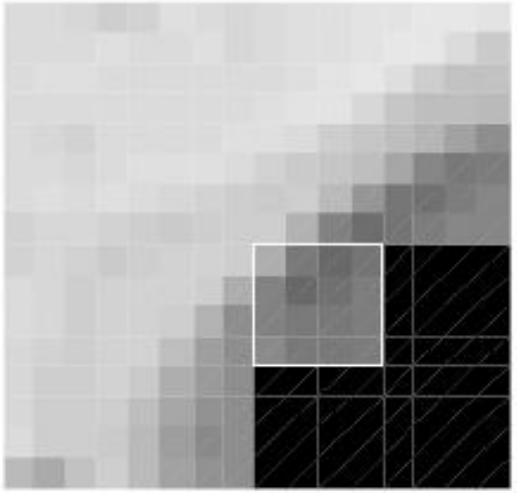


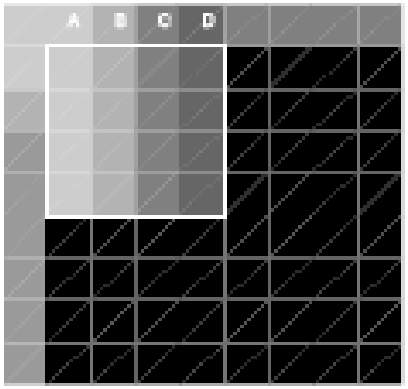


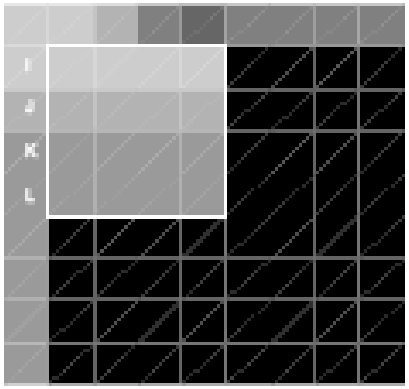
Original macroblock

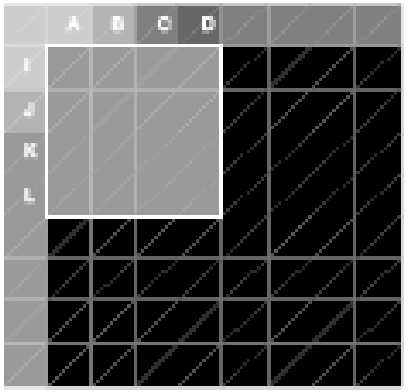


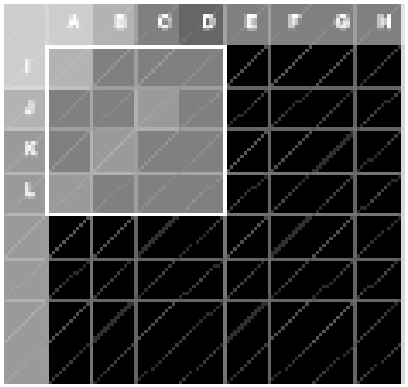
4x4 luma block to be predicted

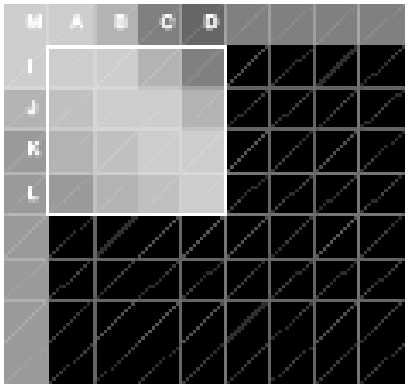


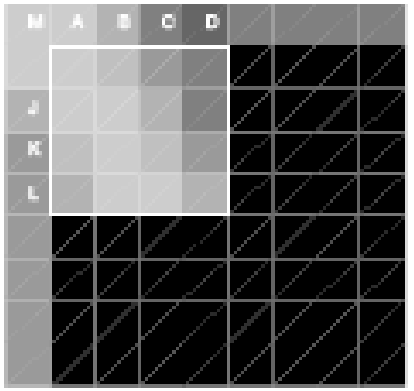

0 (vertical), SAE=619
A B C D

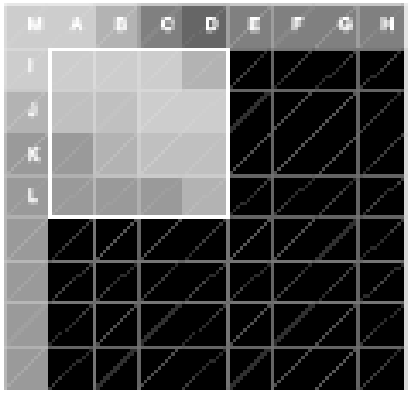

1 (horizontal), SAE=657
I
J
K
L

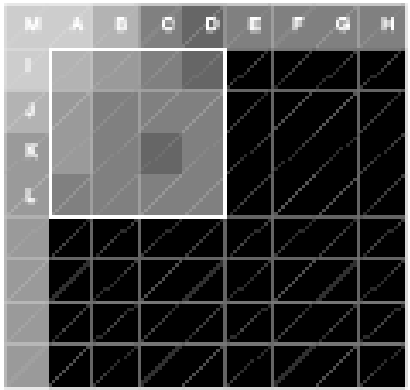

2 (DC), SAE=607
A B C D
I
J
K
L

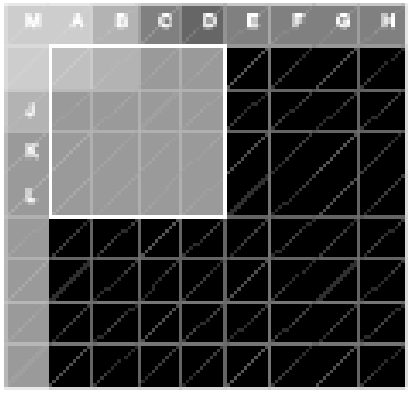

3 (diag down/left), SAE=200
A B C D E F G H
I
J
K
L


4 (diag down/right), SAE=1032
M A B C D
I
J
K
L


5 (vertical/right), SAE=608
M A B C D
J
K
L


6 (horizontal/down), SAE=839
M A B C D E F G H
I
J
K
L

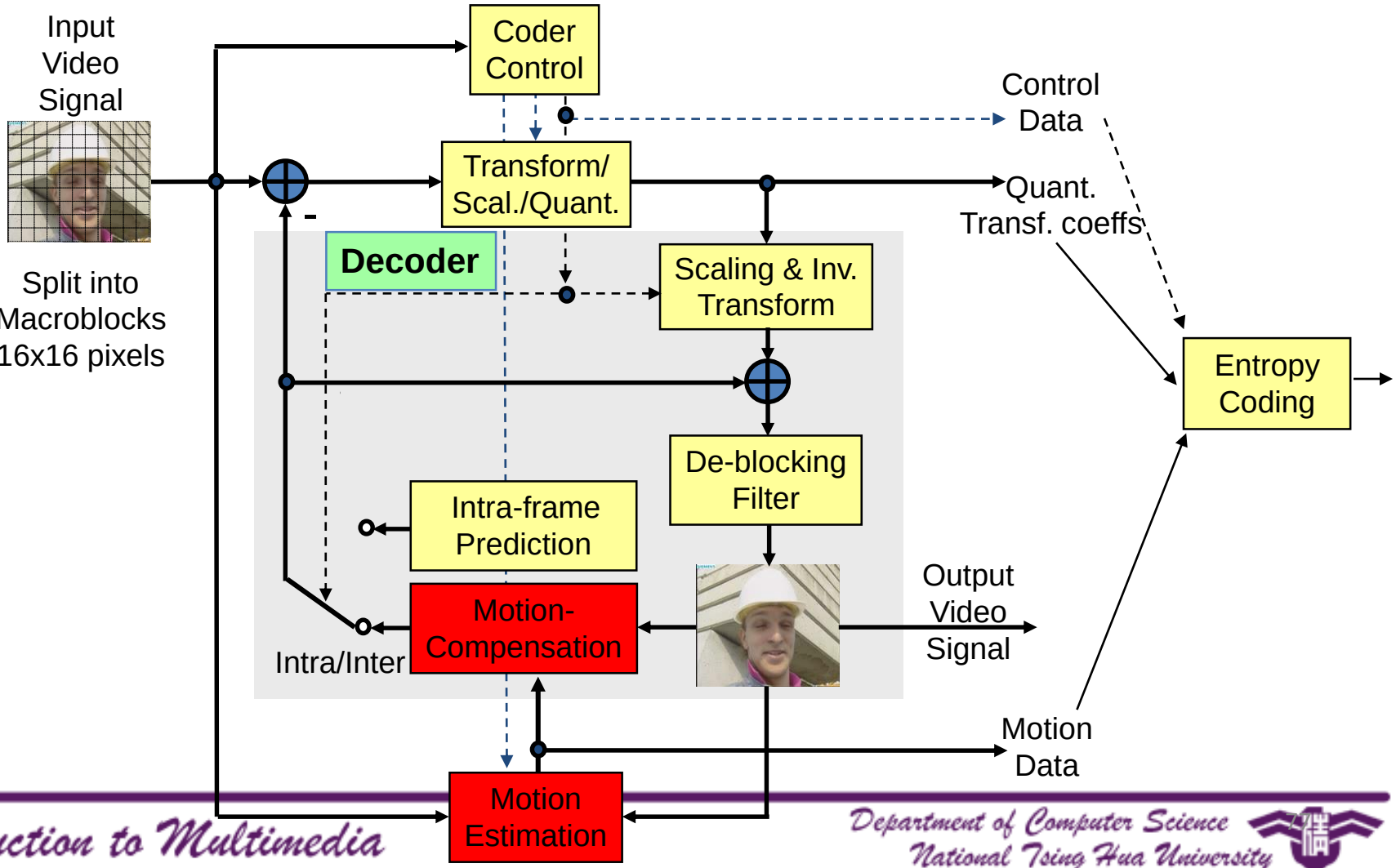

7 (vertical/left), SAE=187
M A B C D E F G H
I
J
K
L


8 (horizontal/up), SAE=399
M A B C D E F G H
J
K
L

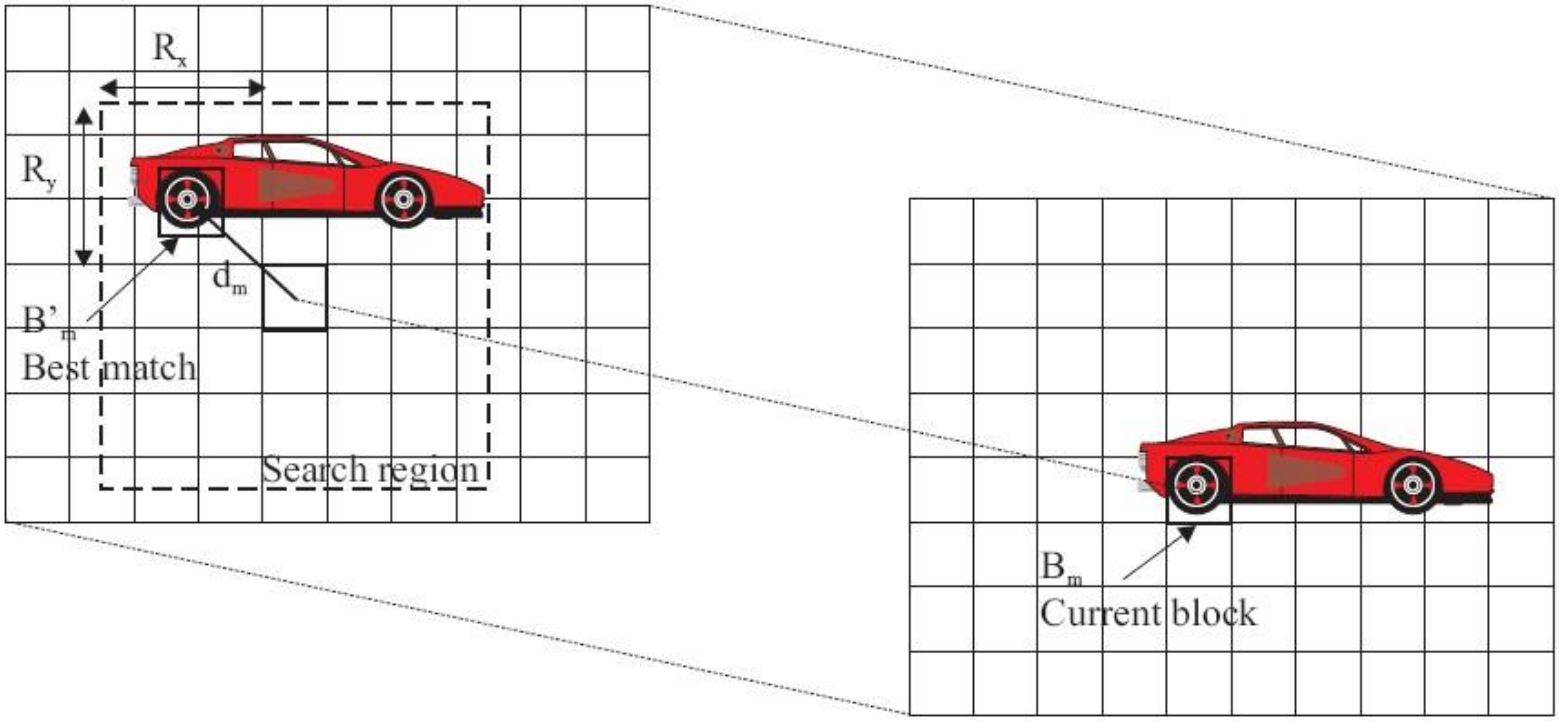
Intra Prediction Example

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

Motion Estimation/Compensation



The Block-Matching Algorithm



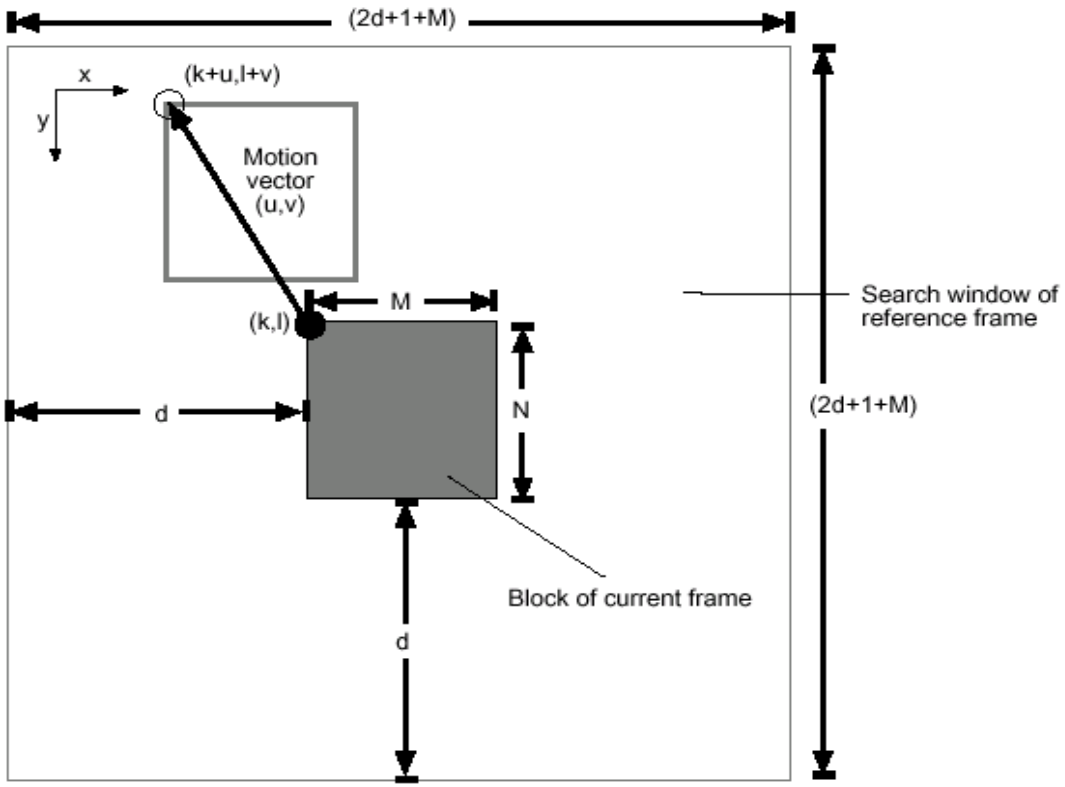
Matching Function

- The dissimilarity $D(s,t)$ between two blocks Ψ_n and

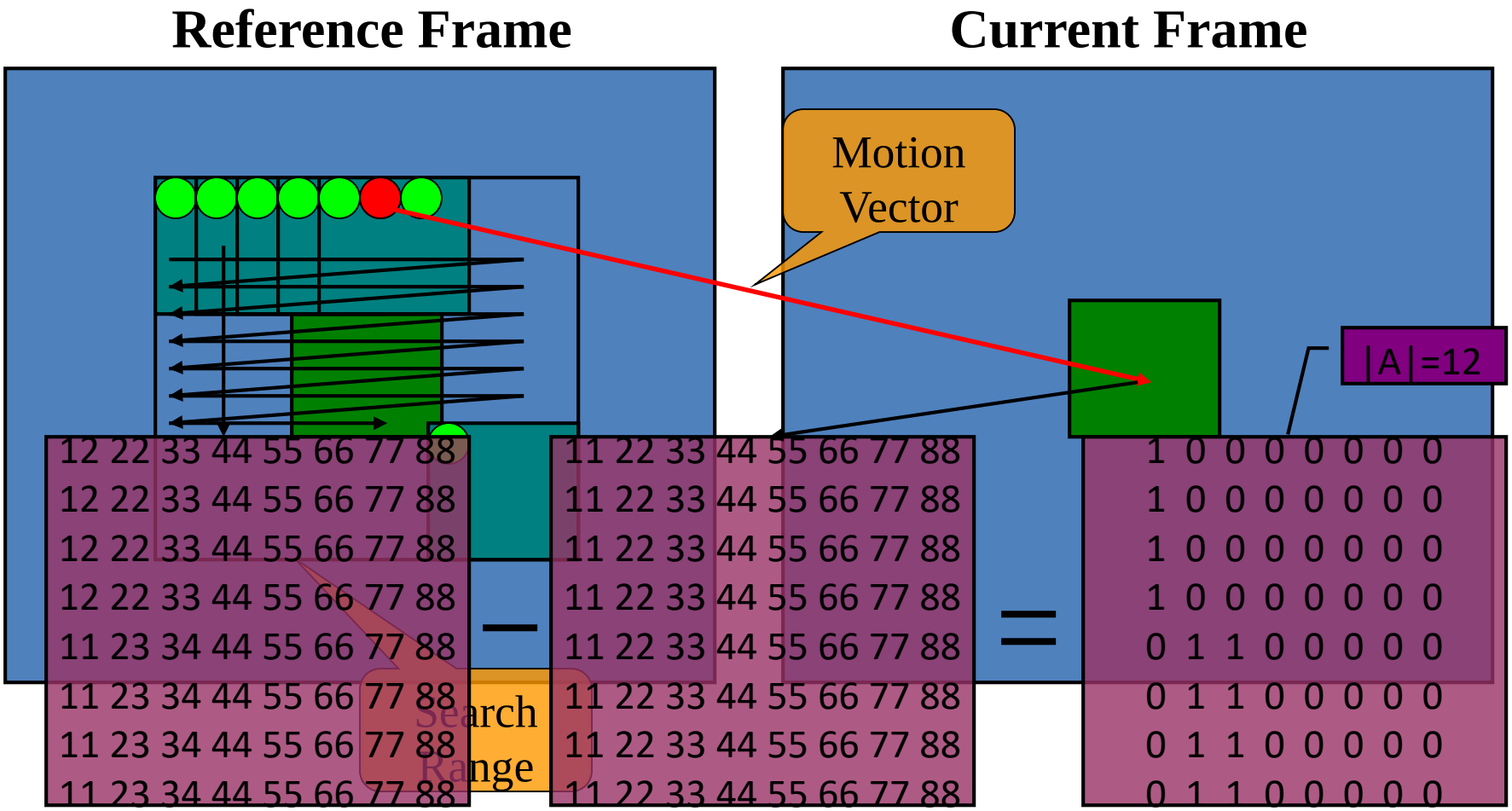
$$D(s,t) = \sum_{x=1}^{N}\sum_{y=1}^{N} M[\Psi_n(x,y), \Psi_{n-1}(x+s, y+t)]$$

- The matching criteria
  - Mean square error (MSE)
    - $M(u,v) = (u-v)^2$
    - High precision is needed
  - Mean absolute difference (MAD)
    - $M(u,v) = |u-v|$
    - Low precision is enough

Full Search


(2d+1+M)
x
y
(k+u,l+v)
Motion vector (u,v)
M
(k,l)
d
N
Search window of reference frame
(2d+1+M)
Block of current frame
d

The Exhaustive Block-Matching Algorithm



Fast Block-Matching Algorithms

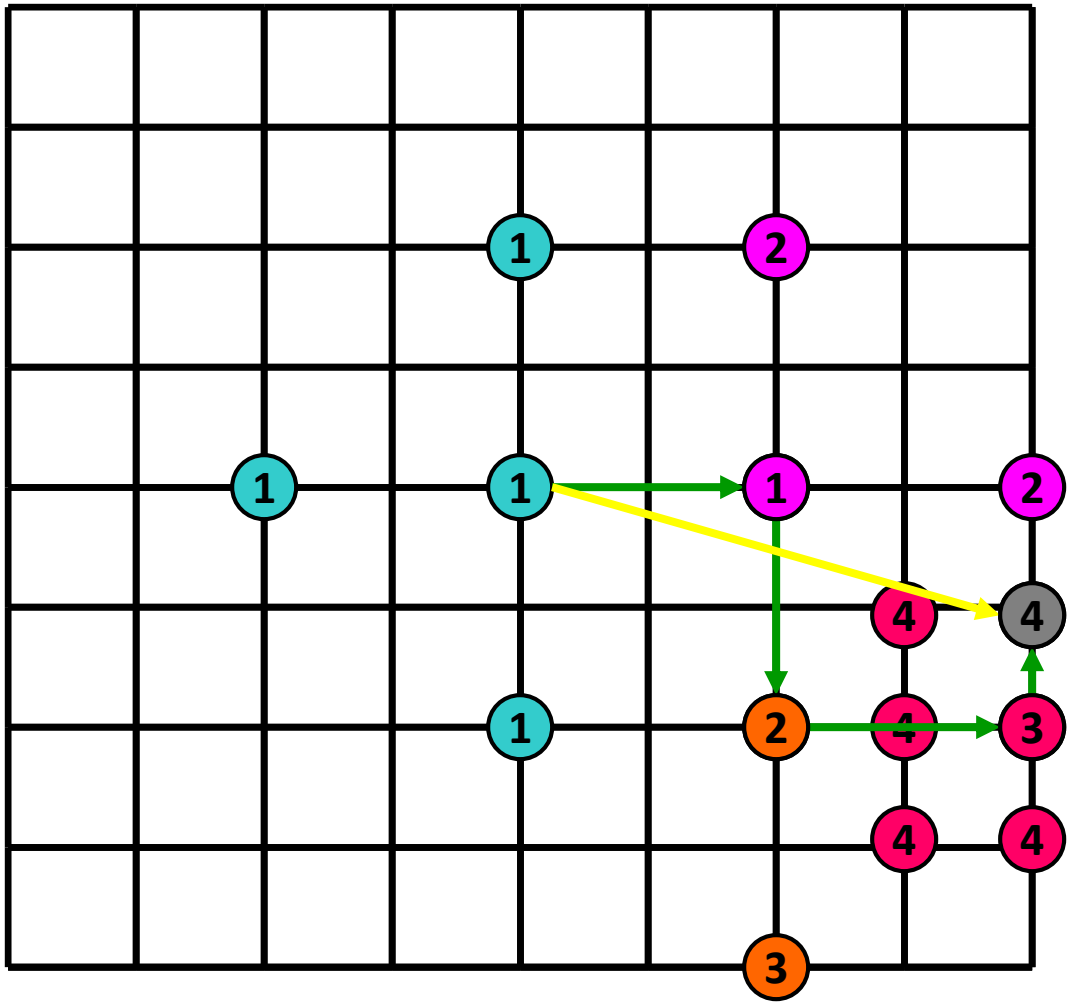
- **The characteristics of fast algorithm**
  - Not accurate as EBMA
  - Save large computation
- **Two famous fast algorithm**
  - 2-D logarithm Search Method
  - Three-Step Search Method

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 (DMD)*.

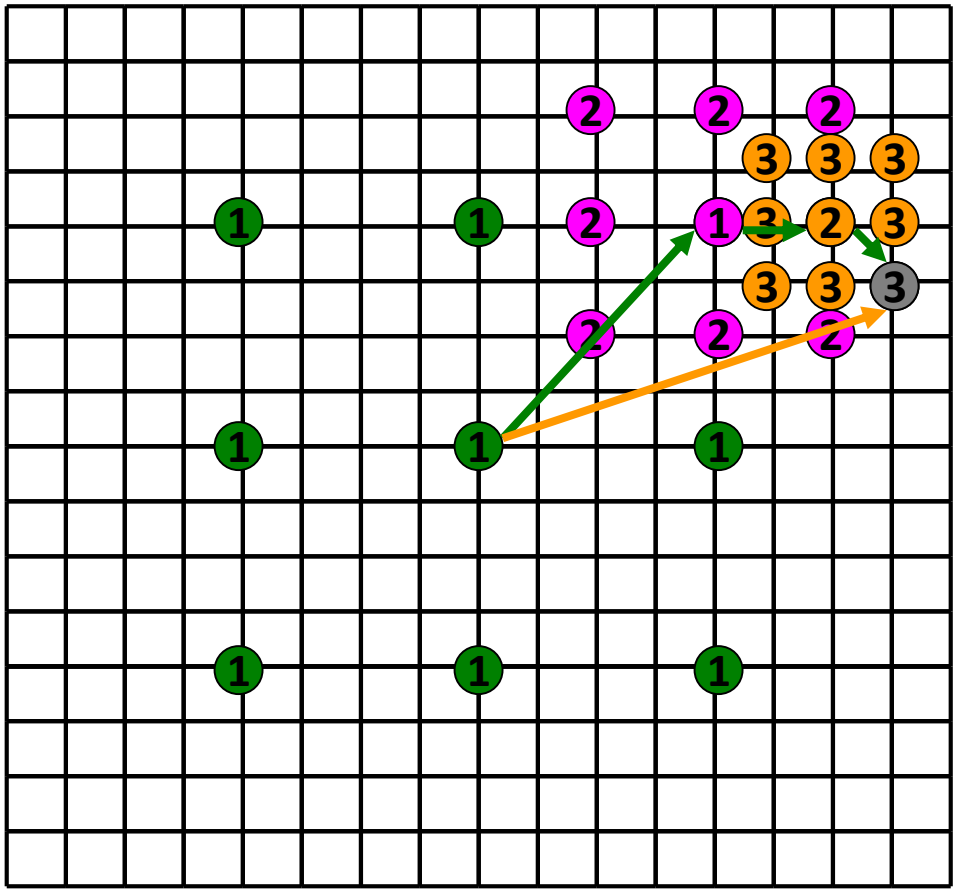
J.R. Jain and A.K. Jain, "Displacement measurement and its application in interframe image coding," *IEEE Trans. Commun.*, Vol. COM-29, pp. 1799–1808, Dec. 1981

2-D logarithm Search Method

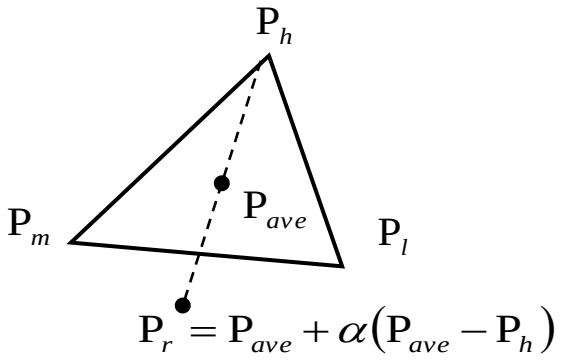


Three-Step Search Method

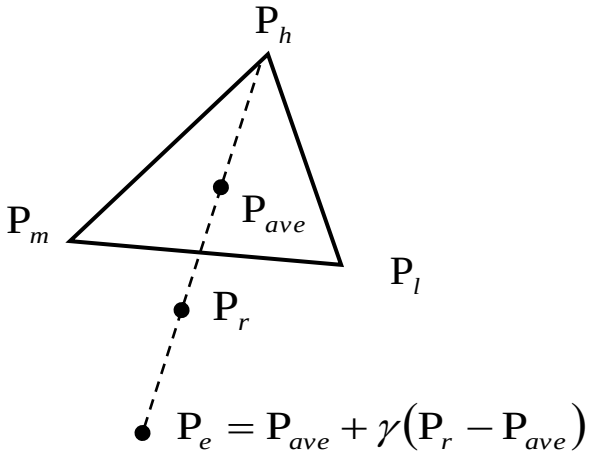
- Three-Step Search (TSS)
  - Koga et al. (1981)
  - 9 Points are searched in each step: 3X3 regular grid with equal space w
  - Find the best match
  - Use the previous best as the new center point
  - Reduce the spacing w by half, select 8 new candidates from the reduced 3X3 grid
  - Repeat the search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



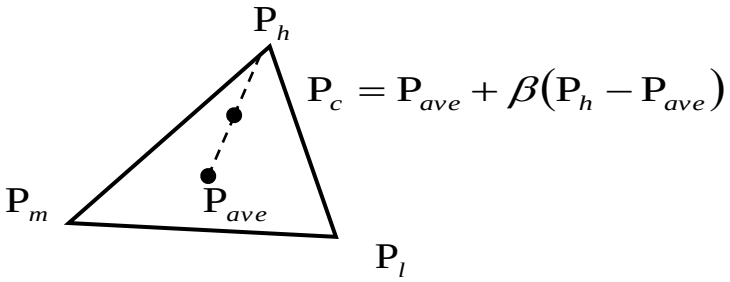
Enhanced Downhill Simplex Search



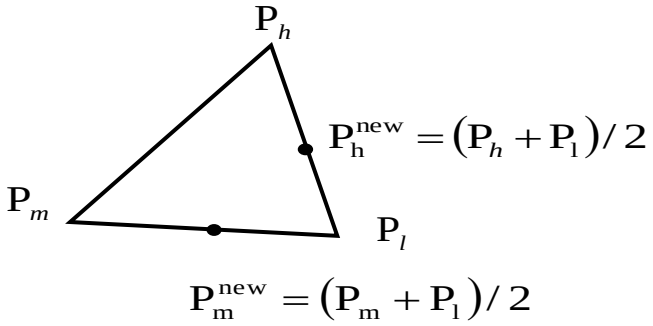
(1) Reflection



(2) Expansion



(3) Contraction



(4) Shrinkage

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

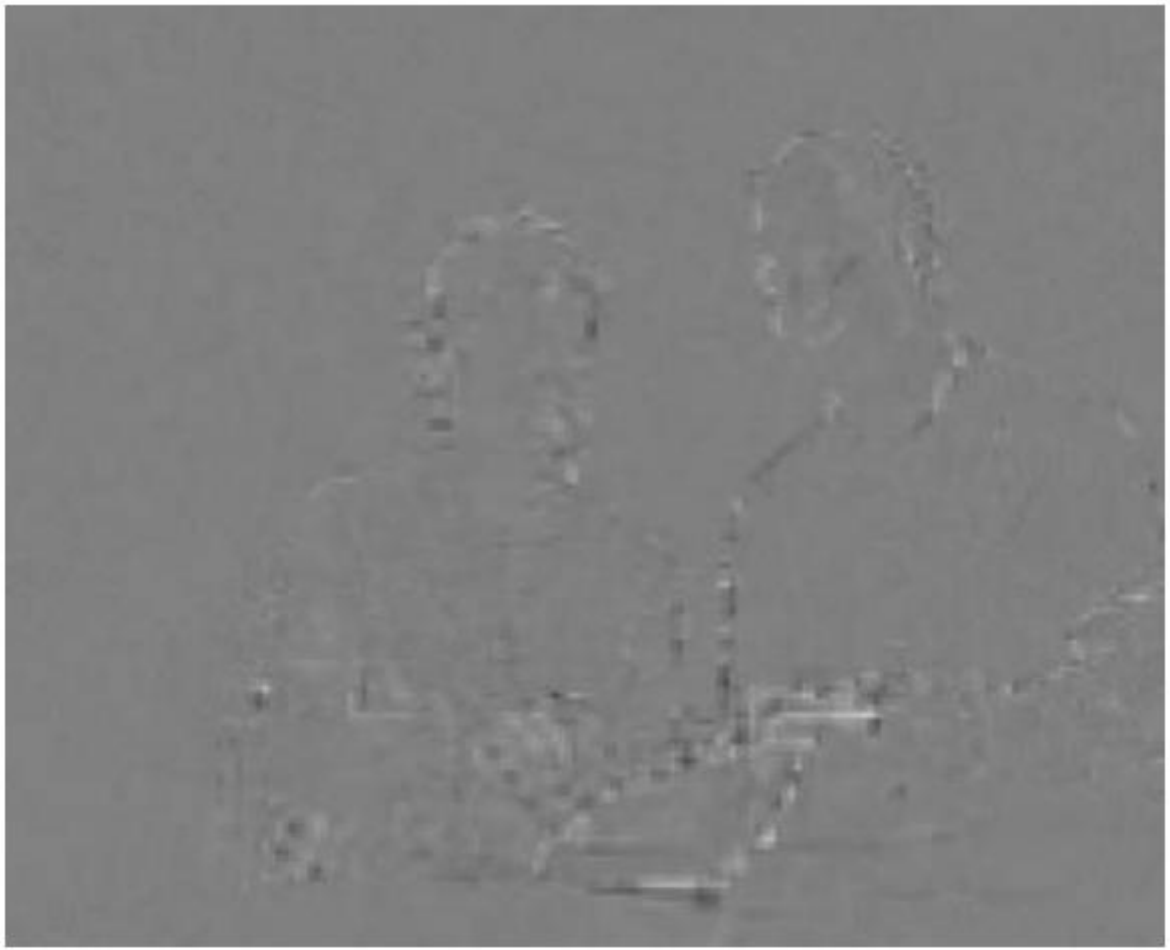


Figure 3.15 Residual (4 × 4 block size)

Example

10	1	5	3
2	8	6	7
7	5	3	1
0	12	9	8

Reference Frame

4	3	9	5
8	0	7	1
2	5	1	10
6	9	7	3

Current Frame

6	2	4	2
6	8	1	6
5	0	2	9
6	3	2	5

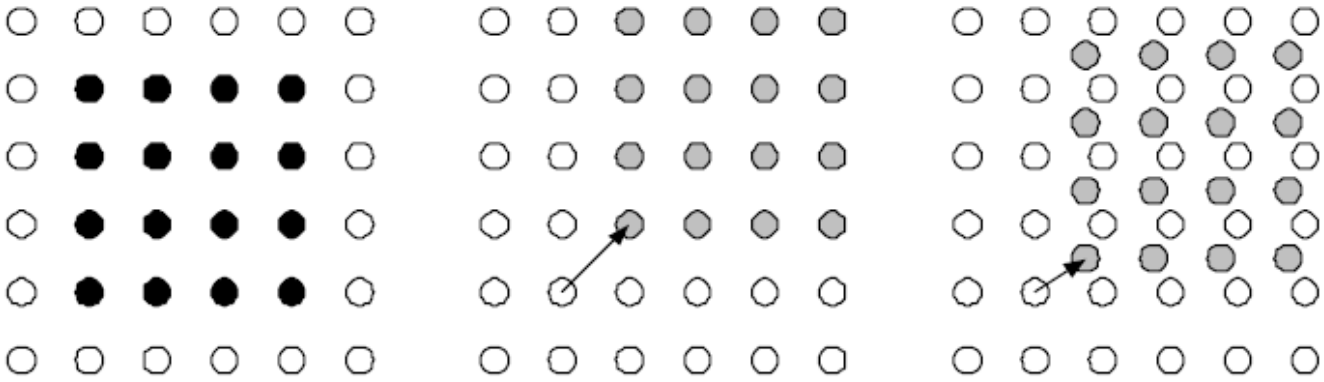
Residual = 52

(2,0)	

Motion Vector

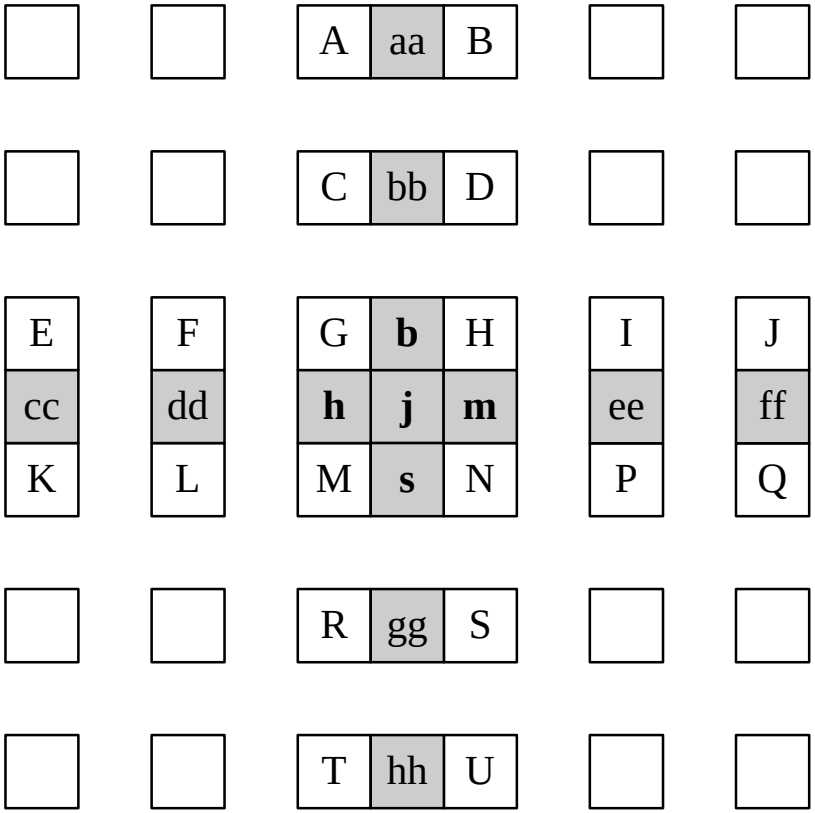
Fractional Motion Estimation

- In H.264, the motion vectors between current block and candidate block has ¼-pel resolution.
- The samples at sub-pel positions do not exist in the reference frame and so it is necessary to create them using interpolation from nearby image samples.



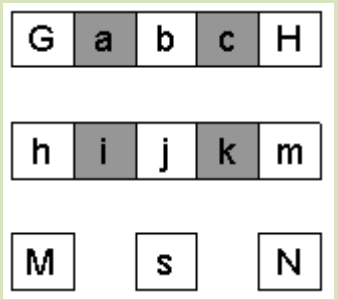
(a) 4x4 block in current frame (b) Reference block: vector (1, -1) (c) Reference block: vector (0.75, -0.5)

Interpolation of ½-pel Samples

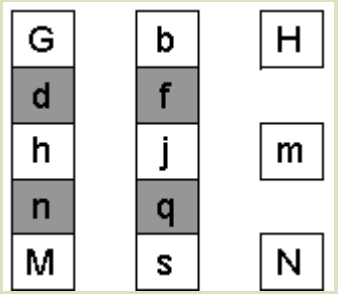


b=round((E-5F+20G+20H-5I+J)/32)
h=round((A-5C+20G+20M-5R+T)/32)
j=round((aa-5bb+20b+20s-5gg+hh)/32)

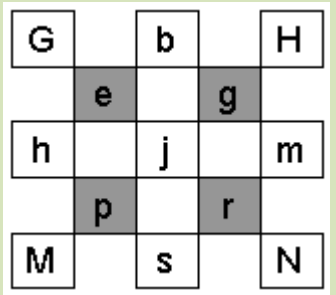
Interpolation of ¼-pel Samples



a=round((G+b)/2)



d=round((G+h)/2)



e=round((b+h)/2)

Half-pixel Compensation

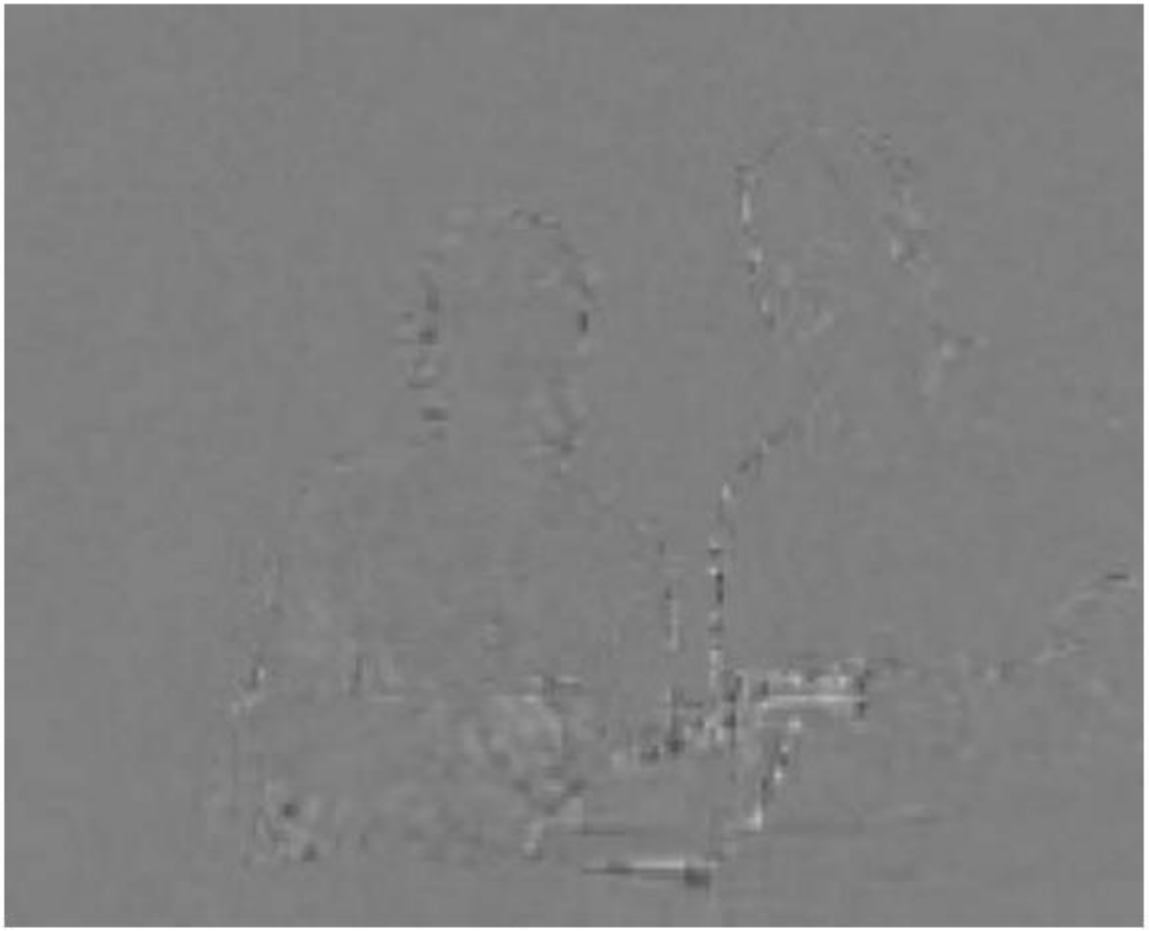


Figure 3.19 Residual (4 × 4 blocks, half-pixel compensation)

Quarter-pixel Compensation

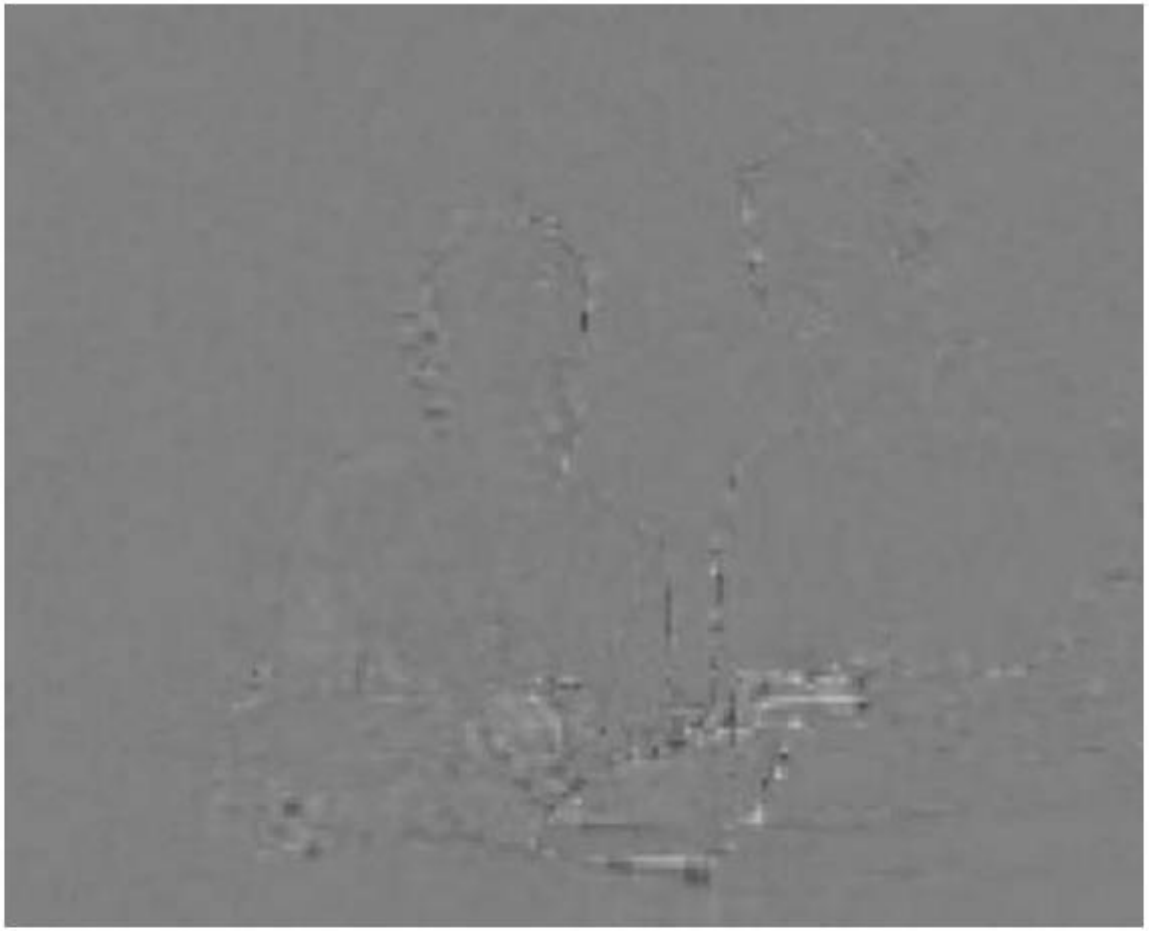
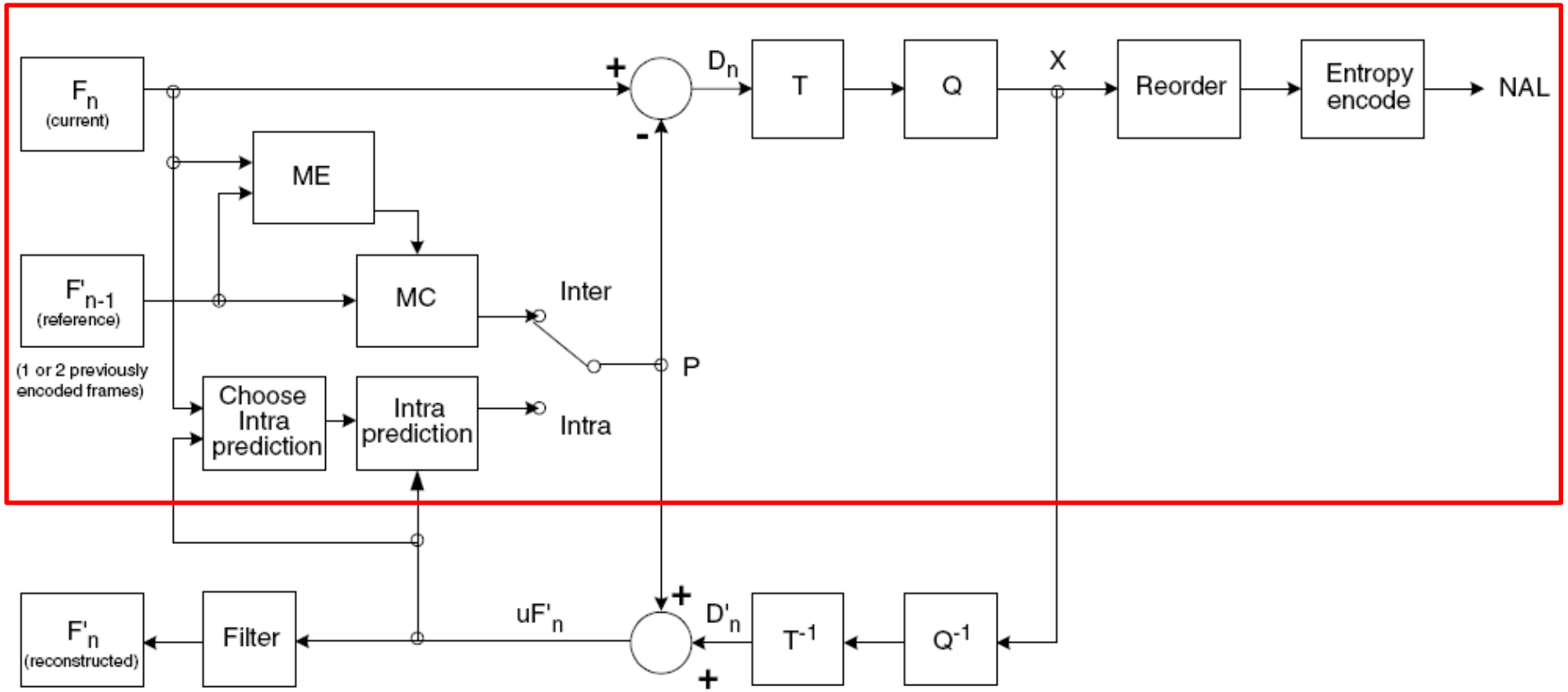


Figure 3.20 Residual (4 × 4 blocks, quarter-pixel compensation)

H.264 Encoder

Forward Path

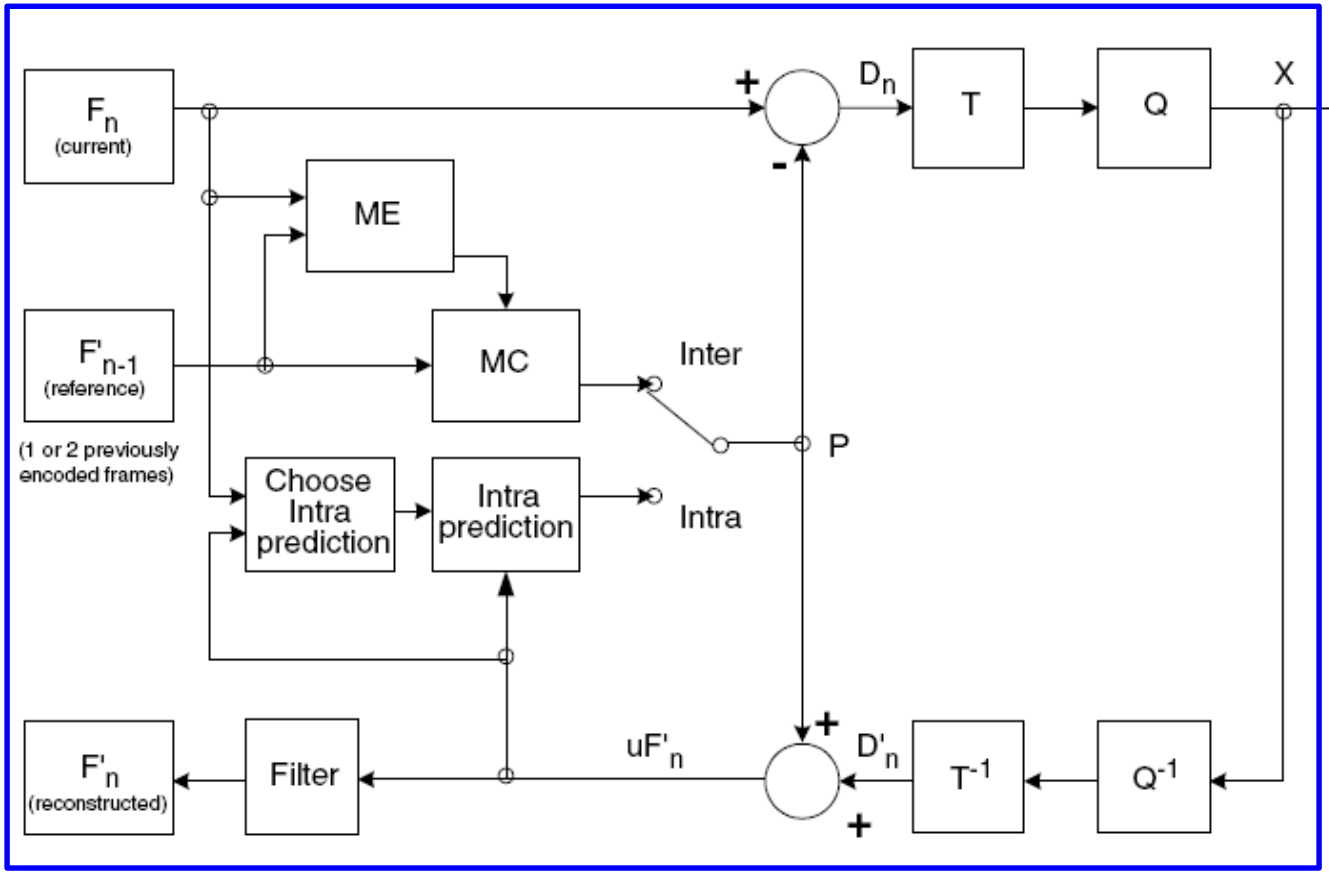


H.264 Encoder

- Encoder (Forward Path)
  - An input frame F_n is processed in units of a macroblock.
  - Each macroblock is encoded in intra or inter mode.
  - A prediction P is formed based on reconstructed picture samples.
    - ✓ In Intra mode, P is formed from samples in the current slice that have previously encoded, decoded and reconstructed.
    - ✓ In Inter mode, P is formed by motion-compensated prediction.
  - In the figures, the reference picture is shown as the previous encoded picture F'_{n-1}.
  - The prediction P is subtracted from the current block to produce a residual block D_n that is transformed and quantized to give X, a set of quantized transform coefficients which are reordered and entropy encoded.
  - The entropy-encoded coefficients, together with side information required, are passed to a Network Abstraction Layer (NAL) for transmission or storage.

H.264 Encoder

Reconstruction Path



H.264 Encoder

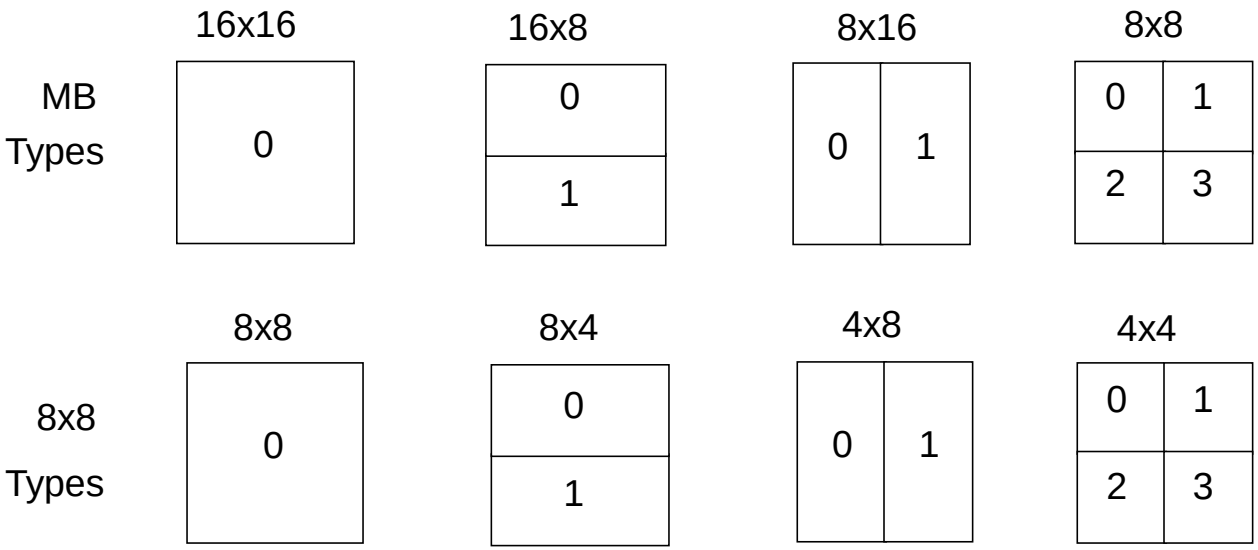
- Encoder (Reconstruction Path)
  - As well as encoding and transmitting each block in a macroblock, the encoder decodes (reconstructs) it to provide a reference for further predictions.
  - The coefficients X are scaled (Q^{-1}) and inverse transformed (T^{-1}) to produce a difference block D'_n.
  - The prediction block P is added to D'_n to create a reconstructed block uF'_n (a decoded version of the original block; u indicates that it is unfiltered).
  - A filter is applied to reduce the effects of blocking distortion and the reconstructed reference picture is created from a series of blocks F'n.

Features of H.264 Motion Estimation

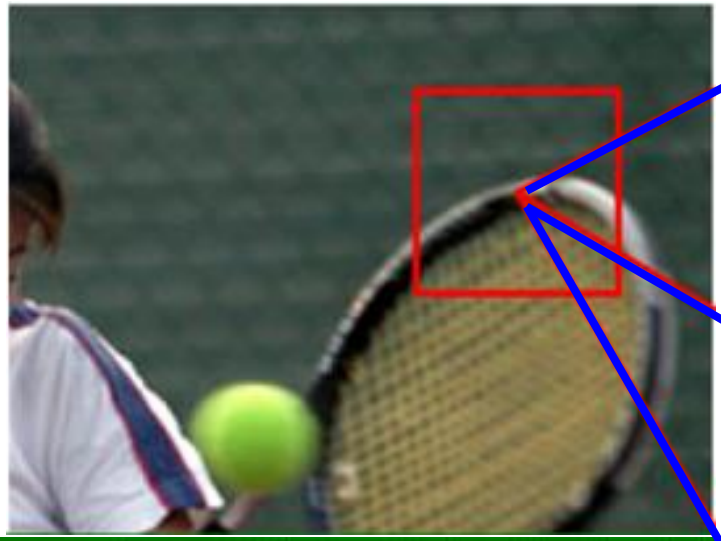
- Various block sizes motion compensation
- Motion Vector Prediction
- ¼ sample accuracy
- Multiple reference pictures
- Generalized B-Frames

Variable Block Size Block-Matching

- In the H.264, a video frame is first splitted using fixed size macroblocks.
- Each macroblock may then be segmented into subblocks with different block sizes.
- A macroblock has a dimension of 16 × 16 pixels. The size of the smallest subblock is 4 × 4

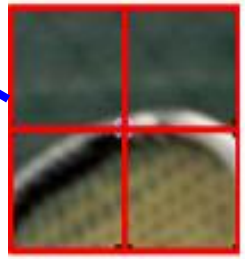


Example: H.264 MC



MPEG-2
- 16x16 block size
- Square shape
- ½ pel motion vector

➔ Weak Motion Isolation !

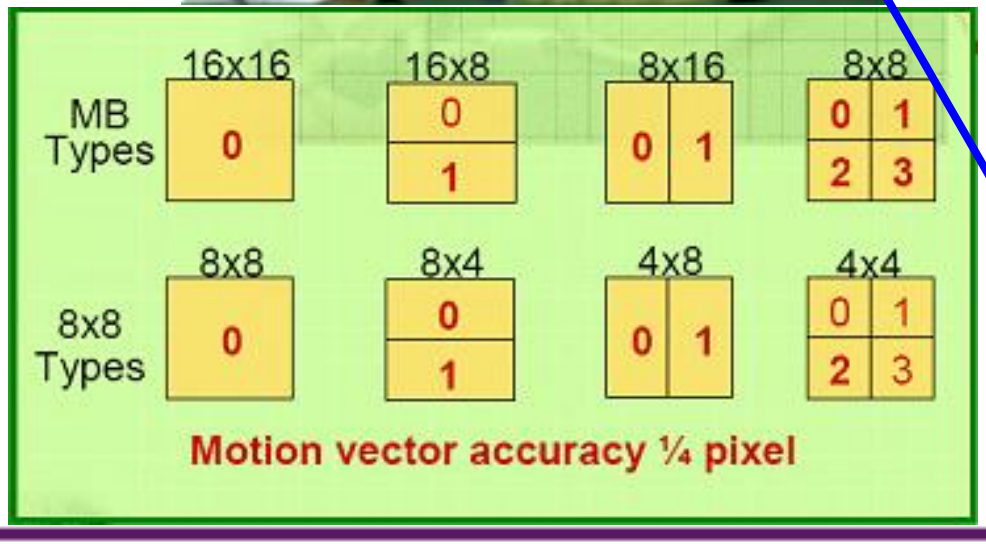


MPEG-4
- 8x8 block size
- Square shapes
- ½ pel motion vector

➔ Moderate Motion Isolation !!

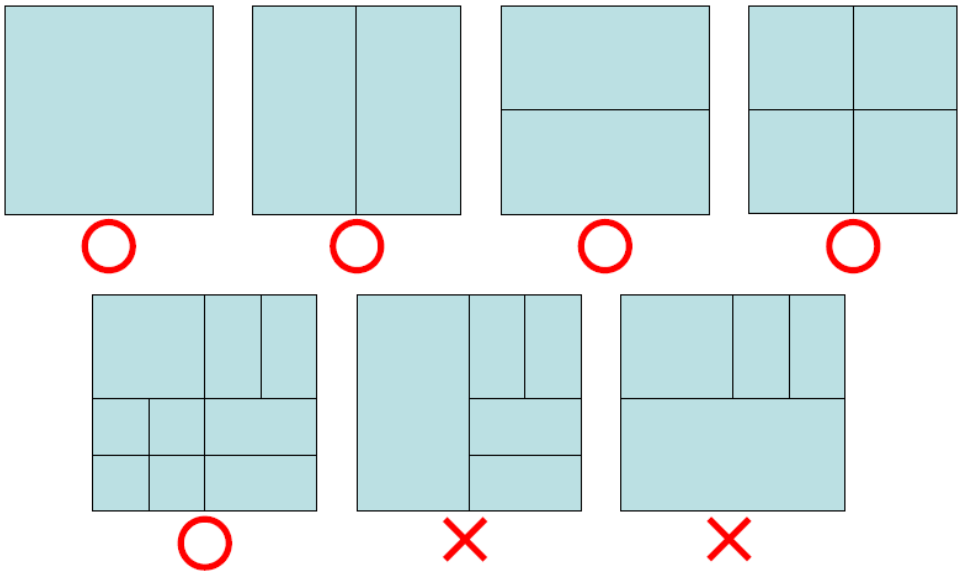
H.264
- 4x4 block size
- Arbitrary shapes
- ¼ pel motion vector

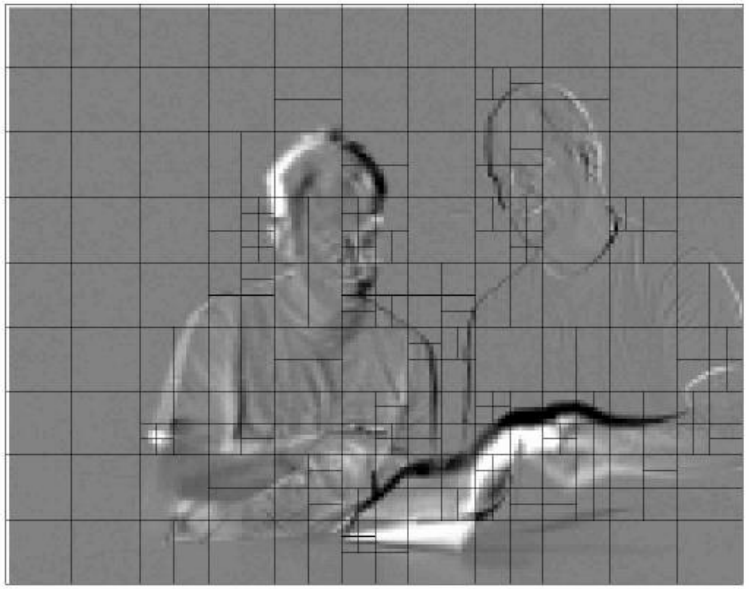
➔ Strong Motion Isolation !!!



Constraint

- To use a subblock with size less than 8×8, it is necessary to first split the macroblock into four 8×8 subblocks.





Motion Vector Prediction

- Encoding a motion vector for each subblock can cost a significant number of bits, especially if small block sizes are chosen.
- Motion vectors for neighboring subblocks are often highly correlated and so each motion vector is predicted from vectors of nearby, previously coded subblocks.
- The difference between the motion vector of the current block and its prediction is encoded and transmitted.

Neighboring Blocks

- Let E be the current block, let A be the subblock immediately to the left of E, let B be the subblock immediately above E, and let C be the subblock above and to the right of E.

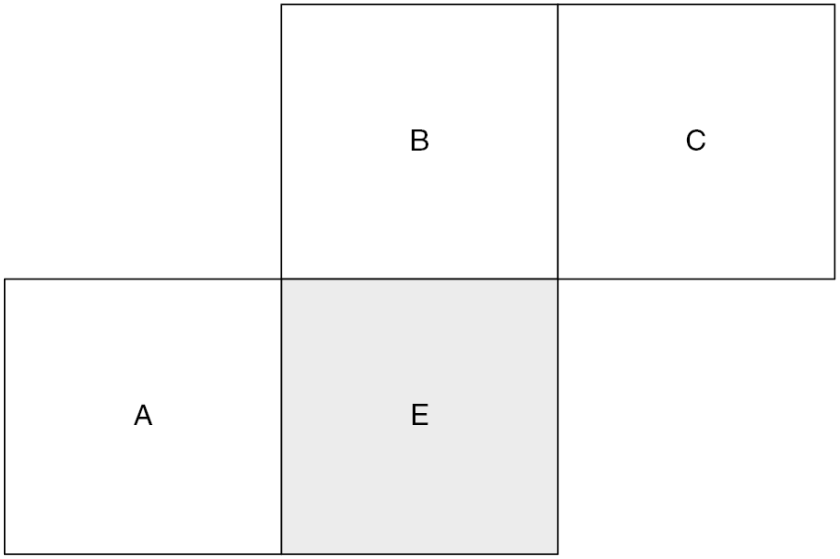


Figure 6.18 Current and neighbouring partitions (same partition sizes)

Neighboring Blocks

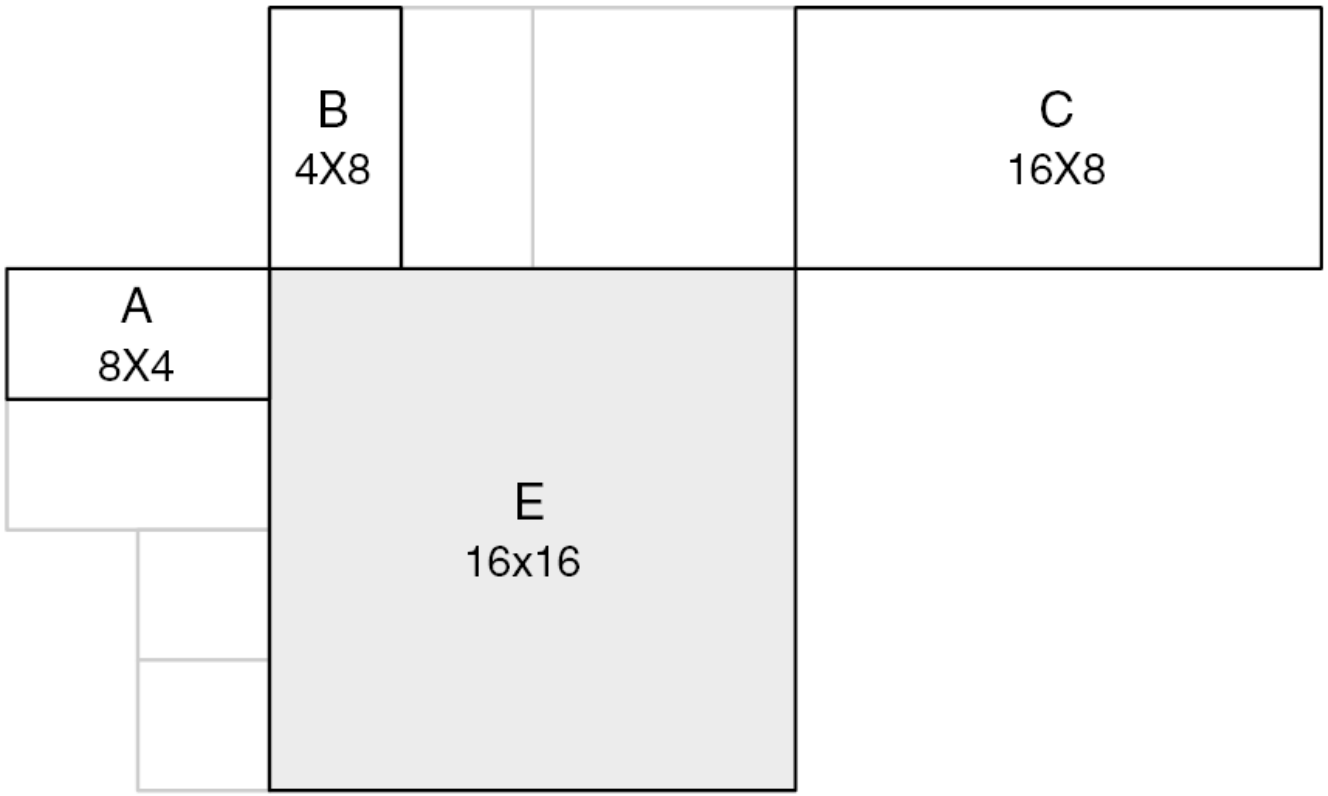


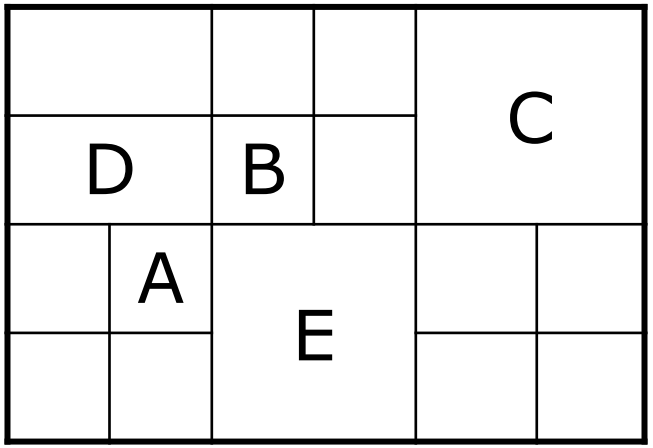
Figure 6.19 Current and neighbouring partitions (different partition sizes)

Method for MV Prediction

- **Median prediction**
  - Use for all block sizes excluding 16×8 and 8×16
  - If variable blocks exist, use topmost and left most instead

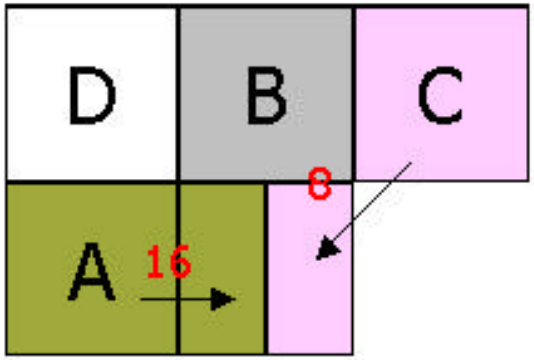
Median prediction

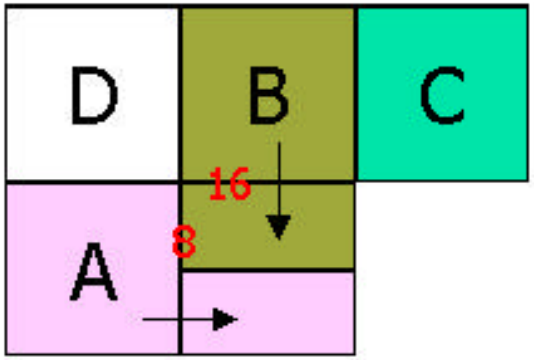
Prediction = median(V_A,V_B,V_C)
If C not exist then C=D
If B, C not exist then prediction = V_A
If A, C not exist then prediction = V_B
If A, B not exist then prediction = V_C



Directional Segmentation Prediction

- Use only for 16×8 and 8×16

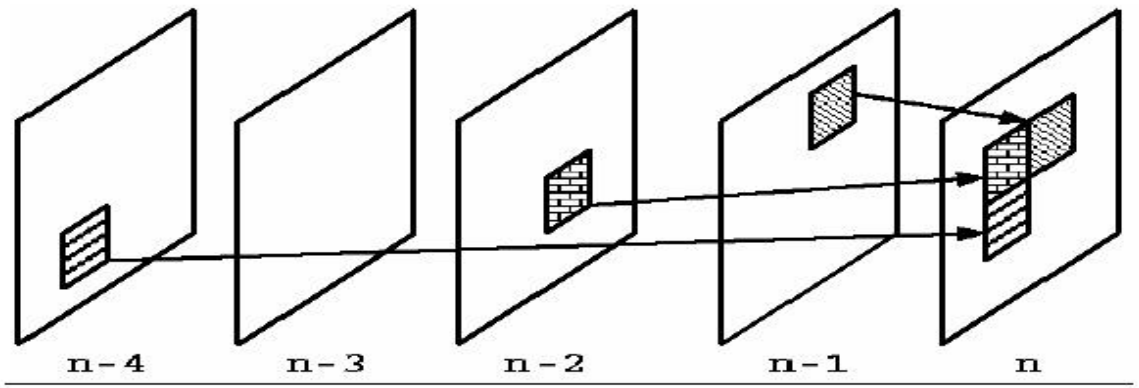




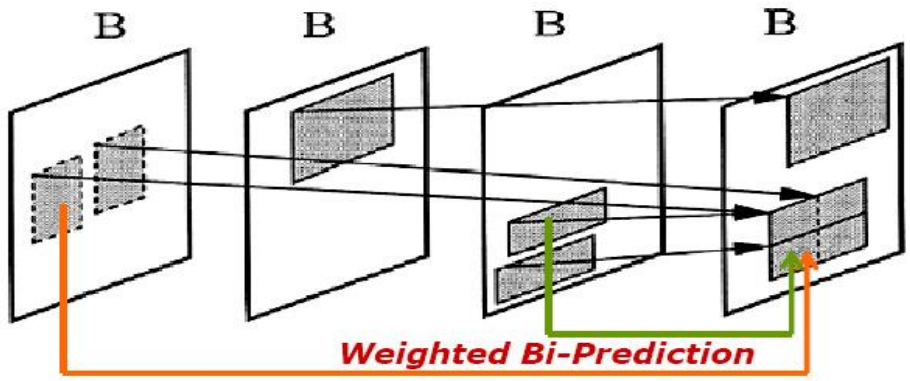
- Vector block size 8×16
  Left: prediction = V_A
  Right: prediction = V_C
- Vector block size 16×8
  Up: prediction = V_B
  Down: prediction = V_A

Multiple Reference Frames

- Reference blocks

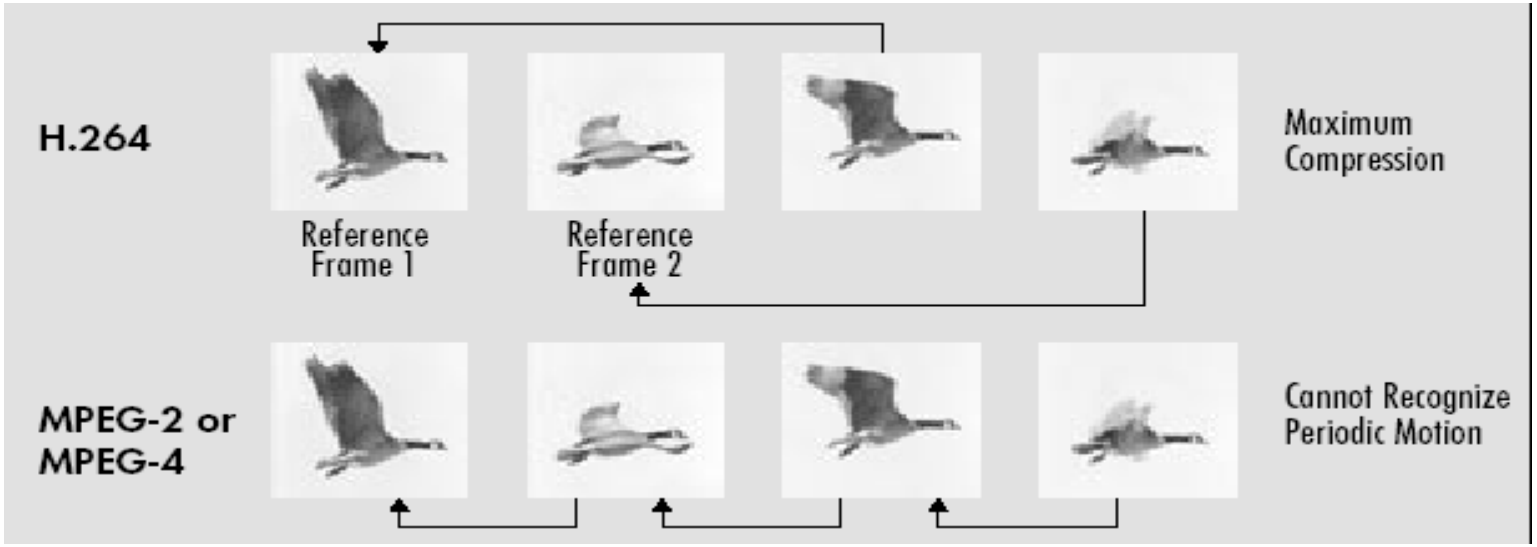


- Weighted bi-prediction

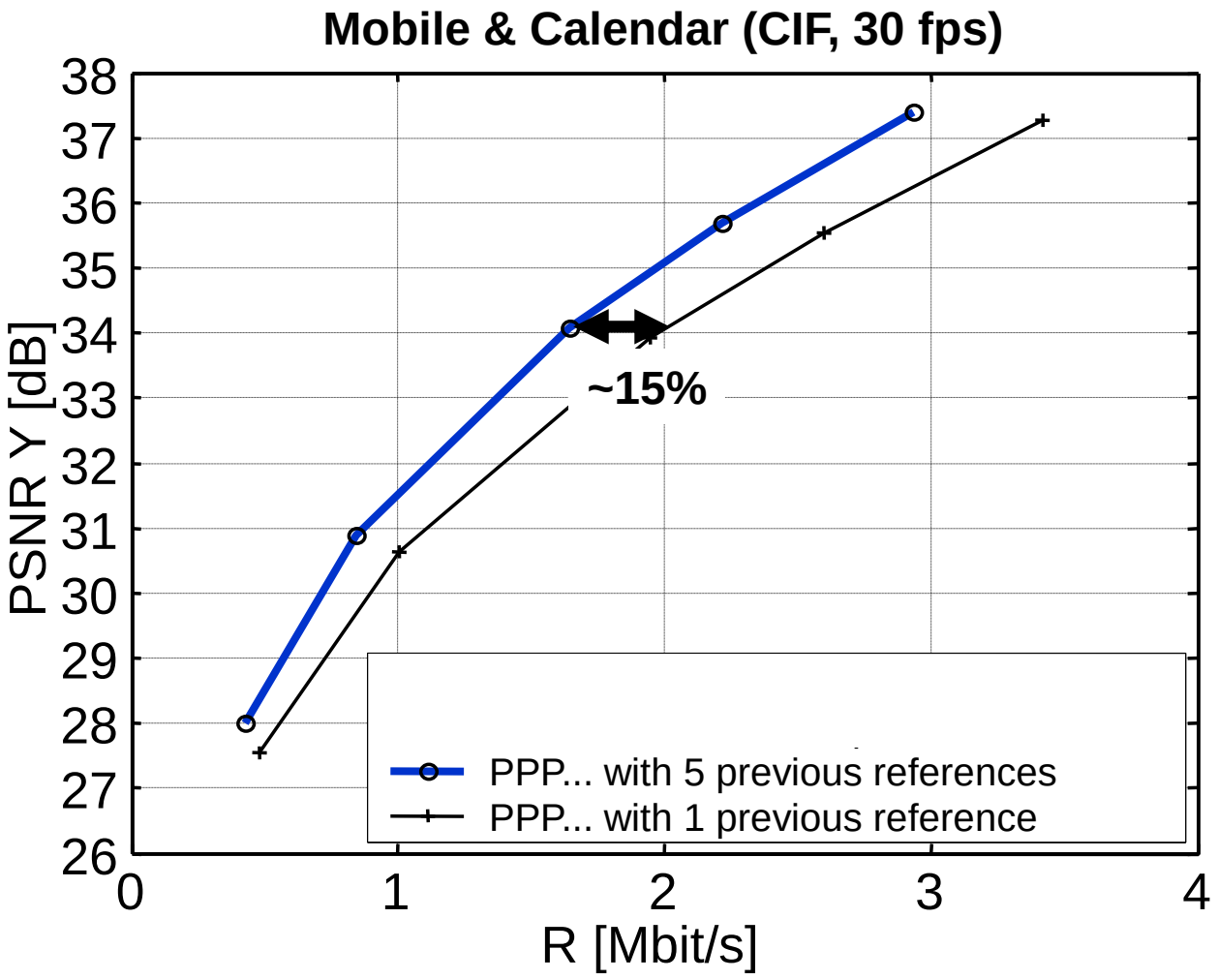


Multiple Reference Frames

- The motion estimation techniques based on multiple reference frame technique provides opportunities for more precise inter-prediction, and also improved robustness to lost picture data.
- The drawback of multiple reference frames is that both the encoder and decoder have to store the reference frames used for Inter-frame prediction and high computational complexity.

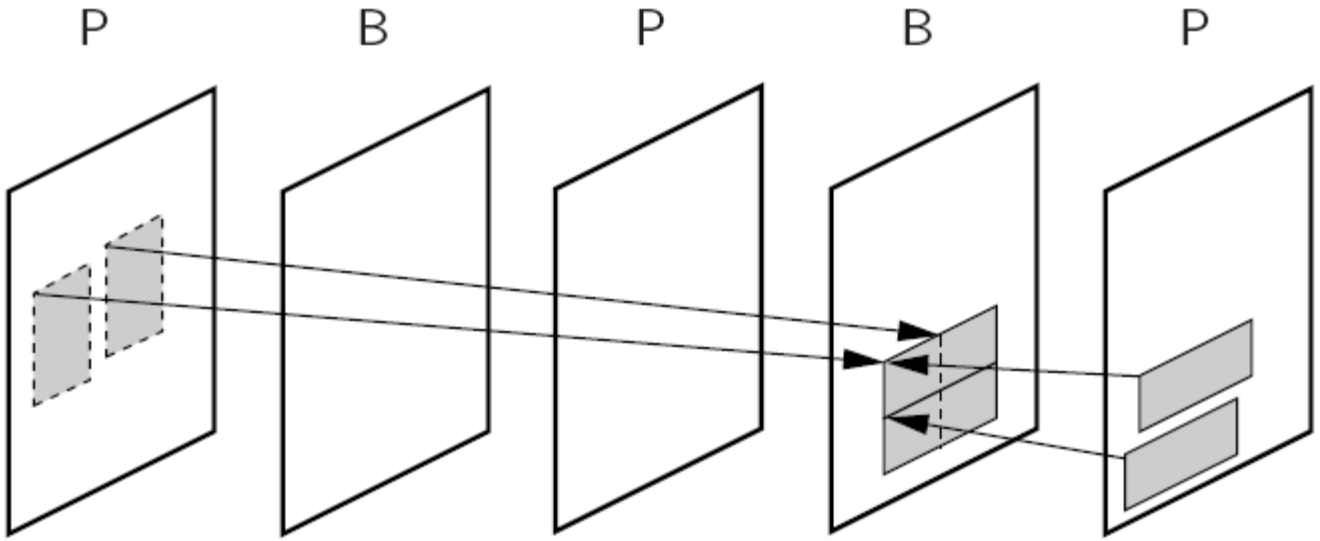


Improvement



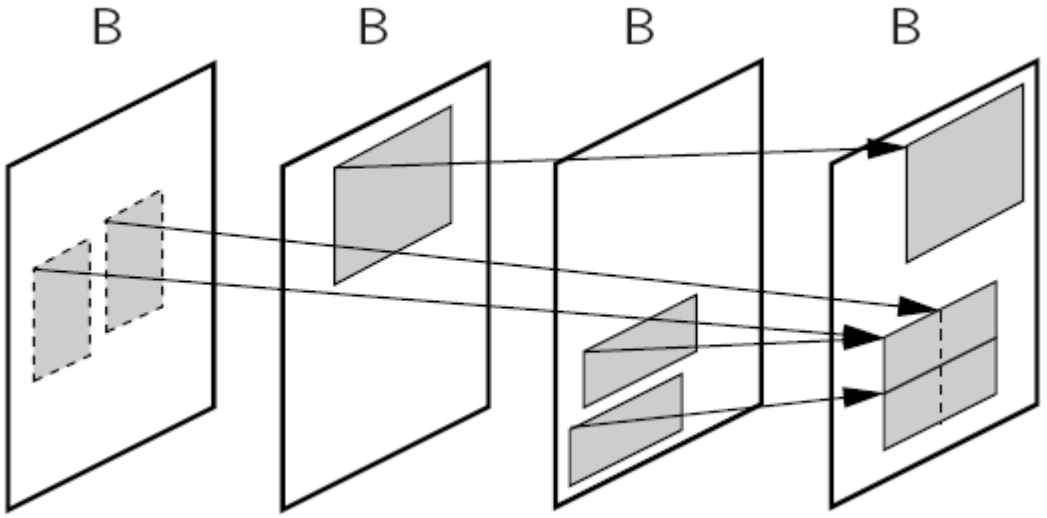
Generalized B Frames

- Basic B-frames:
  - The basic B-frames cannot be used as reference frames.

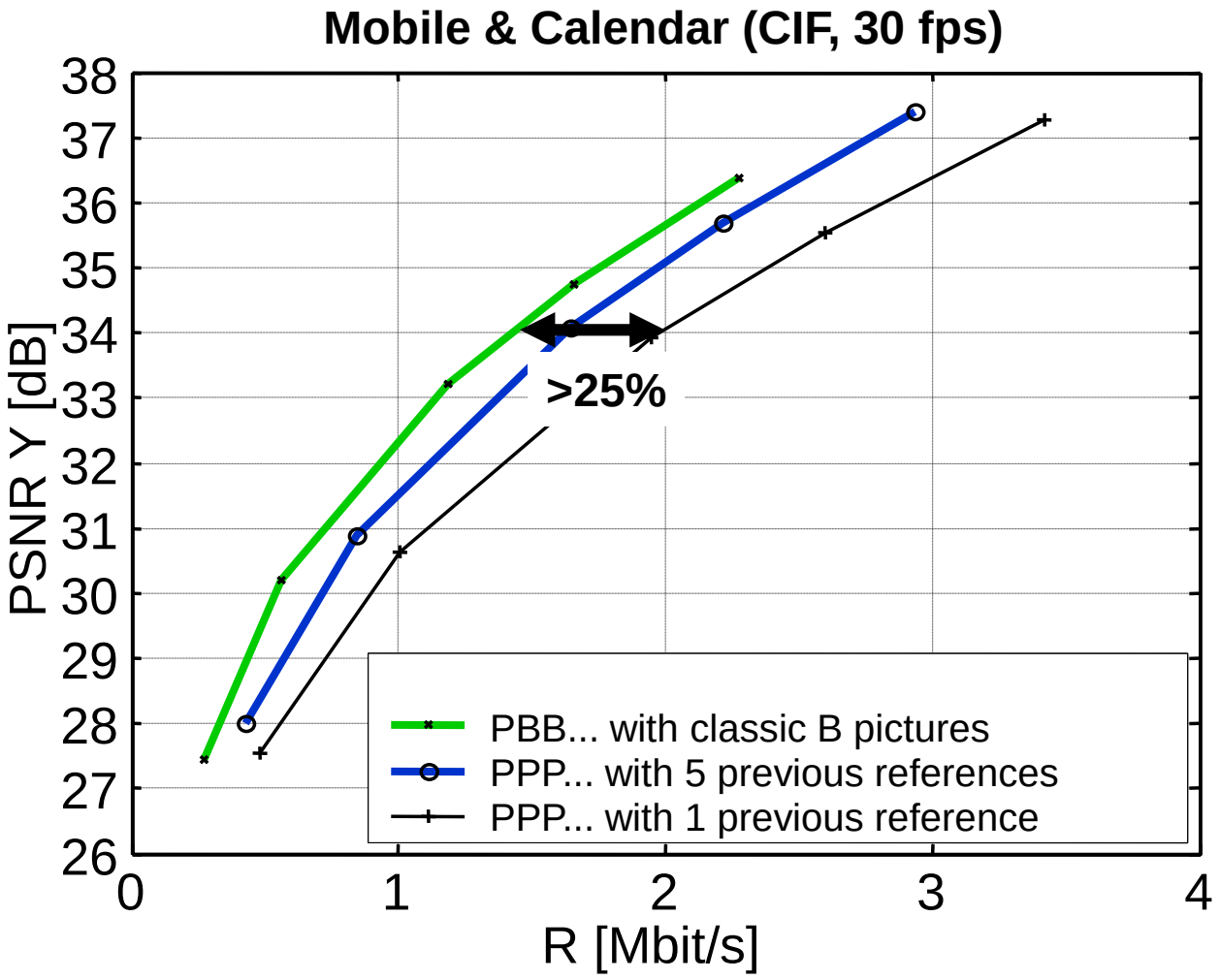


Generalized B-frames

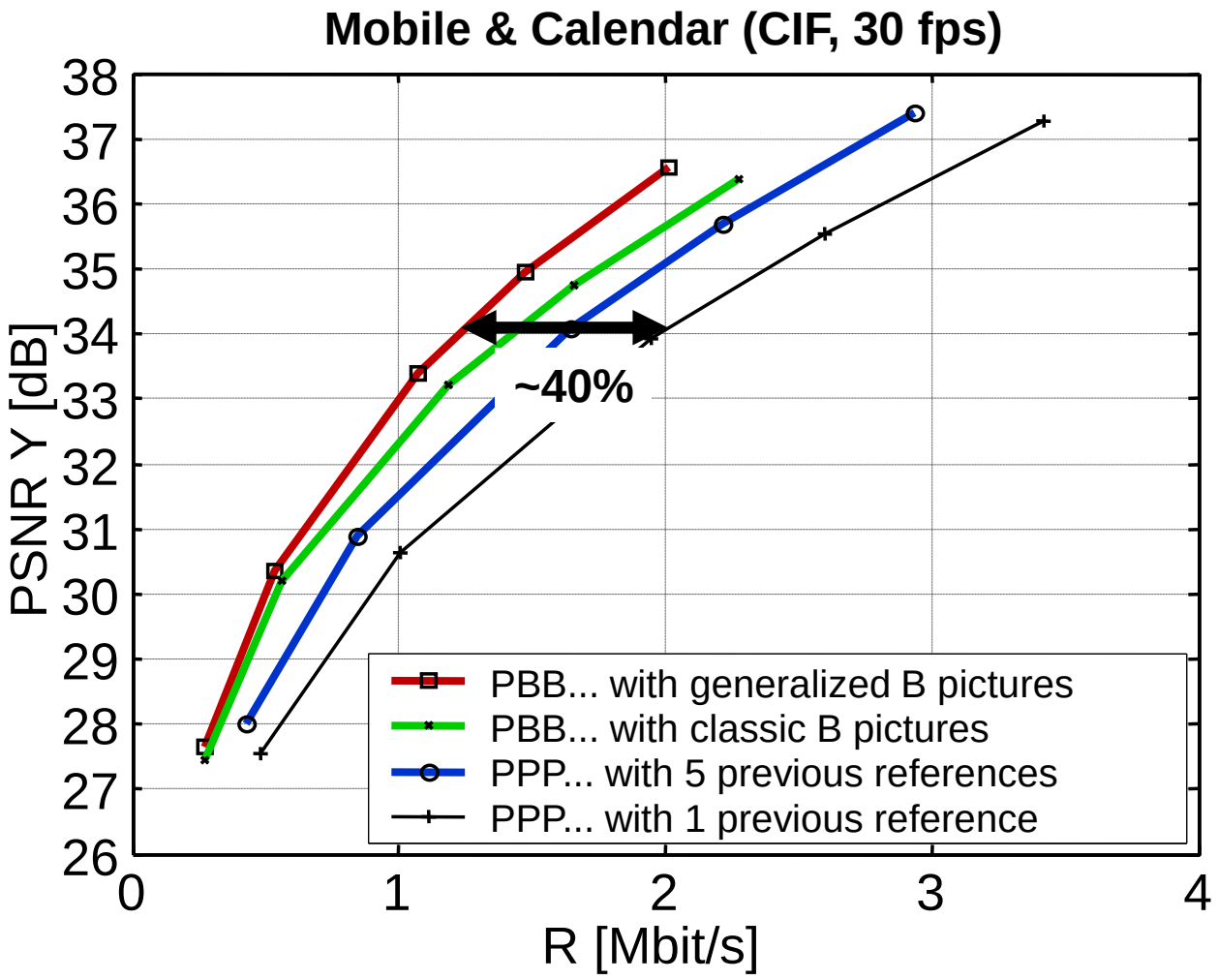
- The generalized B-frames can be used as reference frames



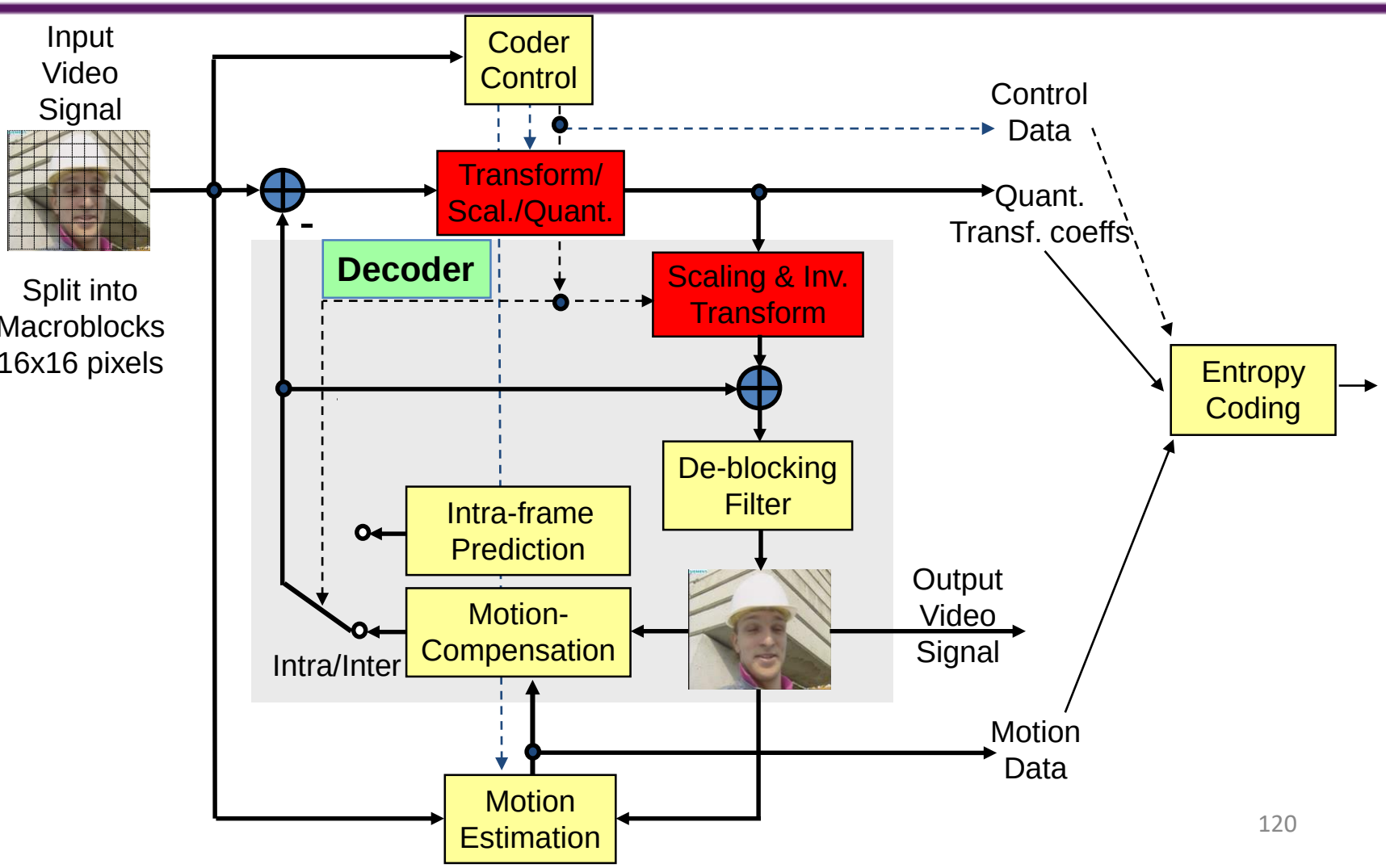
Improvement



Improvement

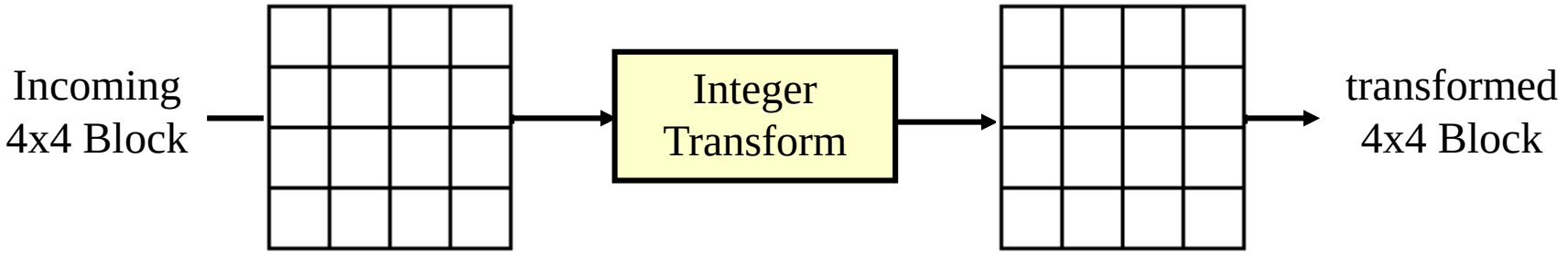


Transformation/Quantization

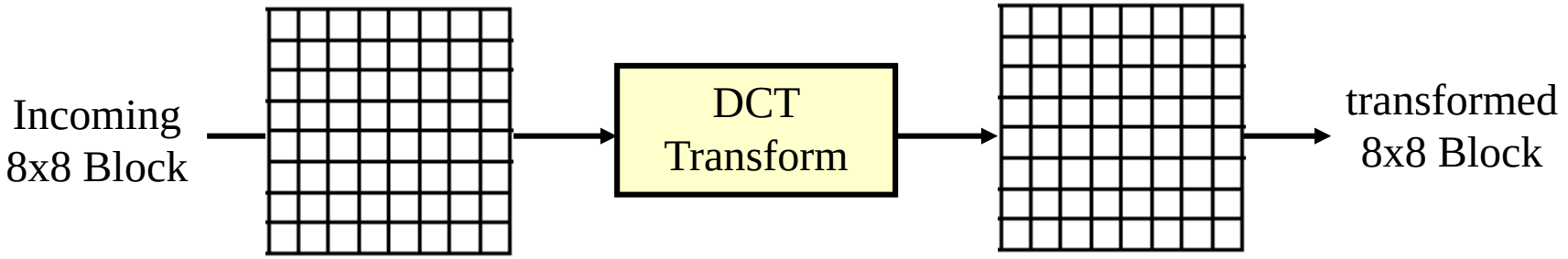


Transform

- H.264 / AVC



- MPEG-2 / MPEG-4



H.264 Transform

- Small block-size transform
  - The new H.264/AVC design is based primarily on a 4×4 transform.
  - Allowing the encoder to represent signals in a more locally-adaptive fashion, which reduces artifacts known as "ringing".

Transformation

- The Discrete Cosine transform (DCT) operates on **X**, a block of $N \times N$ samples and creates **Y**, an $N \times N$ block of coefficients.

The forward DCT:

$$Y = AXA^T$$

The reverse DCT:

$$X = AYA^T$$

DCT Transform

The elements of **A** are: $A_{ij} = C_i \cos\frac{(2j+1)i\pi}{2N}$

where

$$C_i = \sqrt{\frac{1}{N}}, \quad i = 0 \qquad C_i = \sqrt{\frac{2}{N}}, \quad i > 0$$

That is,

$$Y_{xy} = C_x C_y \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} X_{ij} \cos\frac{(2j+1)y\pi}{2\times N}\cos\frac{(2i+1)x\pi}{2\times N}$$

$$X_{ij} = \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} C_x C_y Y_{xy} \cos\frac{(2j+1)y\pi}{2\times N}\cos\frac{(2i+1)x\pi}{2\times N}$$

Example

- The transform matrix **A** for a 4×4 DCT is:

$$\mathbf{A} = \begin{bmatrix} \frac{1}{2}\cos(0) & \frac{1}{2}\cos(0) & \frac{1}{2}\cos(0) & \frac{1}{2}\cos(0) \\ \sqrt{\frac{1}{2}}\cos\left(\frac{\pi}{8}\right) & \sqrt{\frac{1}{2}}\cos\left(\frac{3\pi}{8}\right) & \sqrt{\frac{1}{2}}\cos\left(\frac{5\pi}{8}\right) & \sqrt{\frac{1}{2}}\cos\left(\frac{7\pi}{8}\right) \\ \sqrt{\frac{1}{2}}\cos\left(\frac{2\pi}{8}\right) & \sqrt{\frac{1}{2}}\cos\left(\frac{6\pi}{8}\right) & \sqrt{\frac{1}{2}}\cos\left(\frac{10\pi}{8}\right) & \sqrt{\frac{1}{2}}\cos\left(\frac{14\pi}{8}\right) \\ \sqrt{\frac{1}{2}}\cos\left(\frac{3\pi}{8}\right) & \sqrt{\frac{1}{2}}\cos\left(\frac{9\pi}{8}\right) & \sqrt{\frac{1}{2}}\cos\left(\frac{15\pi}{8}\right) & \sqrt{\frac{1}{2}}\cos\left(\frac{21\pi}{8}\right) \end{bmatrix}$$

Example

That is,

$$\mathbf{A}=\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \sqrt{\frac{1}{2}}\cos\left(\frac{\pi}{8}\right) & \sqrt{\frac{1}{2}}\cos\left(\frac{3\pi}{8}\right) & -\sqrt{\frac{1}{2}}\cos\left(\frac{3\pi}{8}\right) & -\sqrt{\frac{1}{2}}\cos\left(\frac{\pi}{8}\right) \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \sqrt{\frac{1}{2}}\cos\left(\frac{3\pi}{8}\right) & -\sqrt{\frac{1}{2}}\cos\left(\frac{\pi}{8}\right) & \sqrt{\frac{1}{2}}\cos\left(\frac{\pi}{8}\right) & -\sqrt{\frac{1}{2}}\cos\left(\frac{3\pi}{8}\right) \end{bmatrix}$$

or

$$\mathbf{A}=\begin{bmatrix} a & a & a & a \\ b & c & -c & -b \\ a & -a & -a & a \\ c & -b & b & c \end{bmatrix}$$

where

$$a=\frac{1}{2} \qquad c=\sqrt{\frac{1}{2}}\cos\left(\frac{3\pi}{8}\right)$$

$$b=\sqrt{\frac{1}{2}}\cos\left(\frac{\pi}{8}\right)$$

H.264 Transform

- The H.264 transform is based on the 4×4 DCT but with some fundamental differences:
  - It is an integer transfer.
  - The core part of the transform can be implemented using only additions and shifts.
  - A scaling multiplication is integrated into the quantizer, reducing the total number of multiplications.

H.264 Transform

Recall that

$$Y = AXA^T = \begin{bmatrix} a & a & a & a \\ b & c & -c & -b \\ a & -a & -a & a \\ c & -b & b & -c \end{bmatrix} \begin{bmatrix} \\ X \\ \\ \end{bmatrix} \begin{bmatrix} a & b & a & c \\ a & c & -a & -b \\ a & -c & -a & b \\ a & -b & a & -c \end{bmatrix}$$

where

$$a = \frac{1}{2} \qquad b = \sqrt{\frac{1}{2}}\cos\left(\frac{\pi}{8}\right) \qquad c = \sqrt{\frac{1}{2}}\cos\left(\frac{3\pi}{8}\right)$$

H.264 Transform

$$\mathbf{X} = (\mathbf{C}\,\mathbf{x}\,\mathbf{C}^T) \otimes \mathbf{E}$$ Post-scaling

$$= \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & d & -d & -1 \\ 1 & -1 & -1 & 1 \\ d & -1 & 1 & -d \end{bmatrix} \begin{bmatrix} \mathbf{x} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & d \\ 1 & d & -1 & -1 \\ 1 & -d & -1 & 1 \\ 1 & -1 & 1 & -d \end{bmatrix} \right) \otimes$$

$$\begin{bmatrix} a^2 & ab & a^2 & ab \\ ab & b^2 & ab & b^2 \\ a^2 & ab & a^2 & ab \\ ab & b^2 & ab & b^2 \end{bmatrix}$$

(where $d = c/b$)

1. We call (**CxC**^{T}) the core 2D transform.
2. **E** is a matrix of scaling factors.
3. ⊗ indicates that each element of (**CxC**^{T}) is multiplies by the scaling factor in the same position in matrix **E**

H.264 Transform

- To simplify the implementation of the transform
  - *d* is approximated by 0.5.
- In order to ensure that the transform remains orthogonal, *b* also needs to be modified so that:

$$a = \frac{1}{2}, b = \sqrt{\frac{2}{5}}, d = \frac{1}{2}$$

H.264 Transform

- The final forward transform

$$\mathbf{X} = (\mathbf{C}_f \ \mathbf{x}\, \mathbf{C}_f{}^T) \otimes \mathbf{E}_f$$

$$= \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \begin{bmatrix} \ \ \\ \mathbf{x} \\ \ \ \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 & 1 \\ 1 & 1 & -1 & -2 \\ 1 & -1 & -1 & 2 \\ 1 & -2 & 1 & -1 \end{bmatrix} \right) \otimes$$

$$\begin{bmatrix} a^2 & ab/2 & a^2 & ab/2 \\ ab/2 & b^2/4 & ab/2 & b^2/4 \\ a^2 & ab/2 & a^2 & ab/2 \\ ab/2 & b^2/4 & ab/2 & b^2/4 \end{bmatrix}$$

De-block Filter

Input Video Signal

Split into Macroblocks 16x16 pixels

Coder Control

Control Data

Transform/ Scal./Quant.

Quant. Transf. coeffs

Decoder

Scaling & Inv. Transform

Entropy Coding

De-blocking Filter

Intra-frame Prediction

Motion-Compensation

Intra/Inter

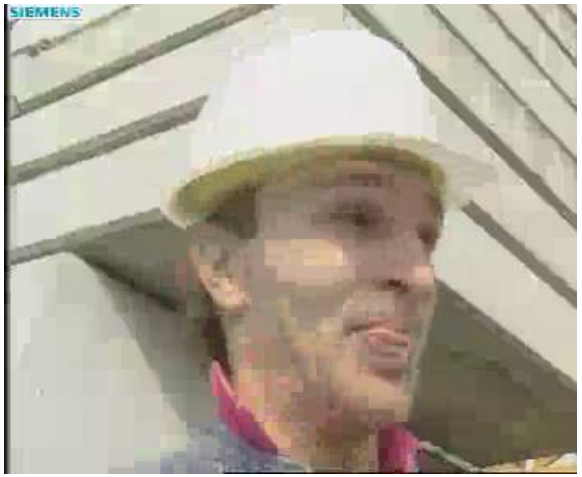
Output Video Signal

Motion Data

Motion Estimation

Adaptive Deblocking Filter

- There are severe blocking artifacts
  - 4*4 transforms and block-based motion compensation
- Result in bitrate savings of around 6~9%
- Improve subjective quality and PSNR of the decoded picture

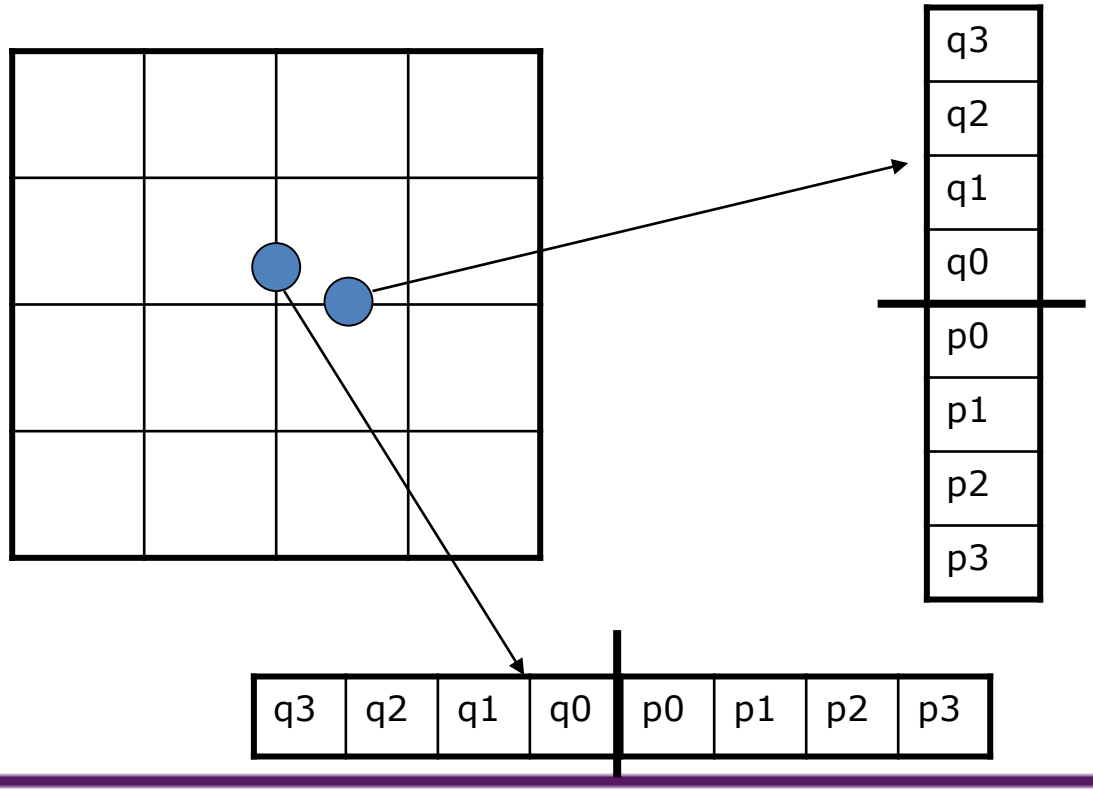


Without Filter

With AVC Deblocking Filter

Deblocking Filter

- The deblocking filter improves subjective visual quality.
- The filter is highly context adaptive.
- It operates on the boundary of 4×4 subblock as shown below.

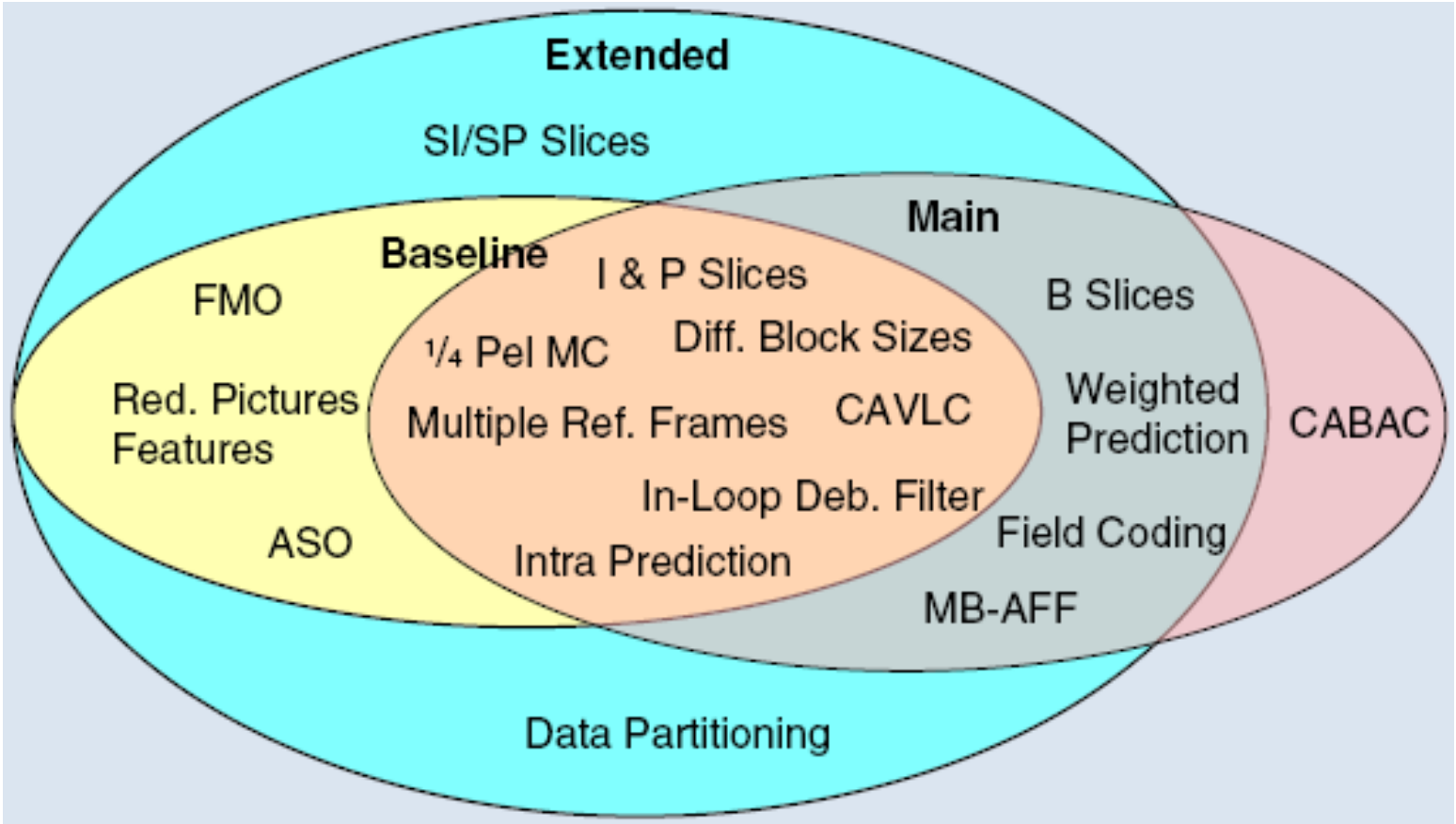


Deblocking Filter

- The choice of filtering outcome depends on the boundary strength and on the gradient of image samples across the boundary.
- The boundary strength parameter Bs is selected according to :

p or q is intra coded **and** boundary is a macroblock boundary	Bs=4 (strongest filtering)
p or q is intra coded and boundary is **not** a macroblock boundary	Bs=3
neither p or q is intra coded; p or q contain coded coefficients	Bs=2
neither p or q is intra coded; neither p or q contain coded coefficients; p and q have different reference frames **or** a different number of reference frames **or** different motion vector values	Bs=1
neither p or q is intra coded; neither p or q contain coded coefficients; p and q have same reference frame and identical motion vectors	Bs=0 (no filtering)

H.264/AVC Profiles



Conclusions

- Some important differences relative to prior standards.
  - Enhanced motion-prediction capability
  - Use of a small block-size exact –match transform
  - Adaptive in-loop deblocking filter
  - Enhanced entropy coding methods
- When used well together, a approximately 50% bit rate savings for equivalent perceptual quality relative to the performance of prior standards.

Computation Consumption

- Use large amount of computation to achieve high quality, reliability and compression ratio
  - Intra/inter frame prediction: ~60%
  - Integer transform: ~10%
  - Error resilience: ~20%
  - De-blocking filter: ~10%
- Encoding: ~10x of MPEG-2's CPU usage