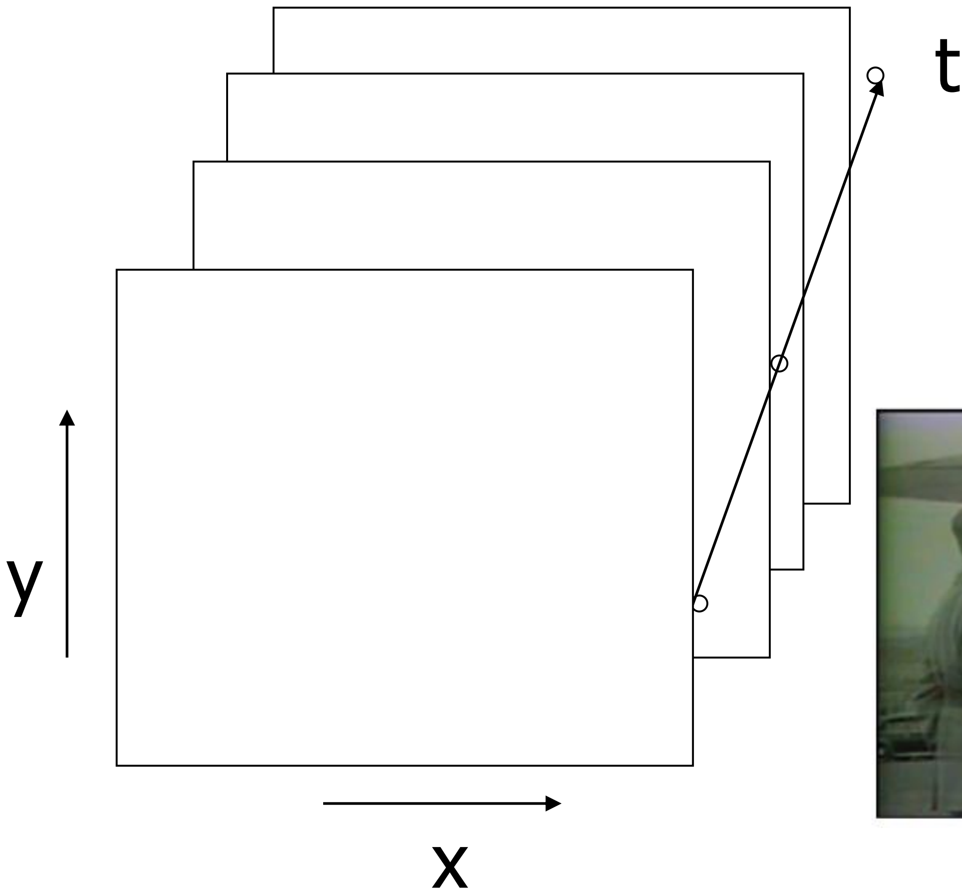
Fundamentals of Video Processing

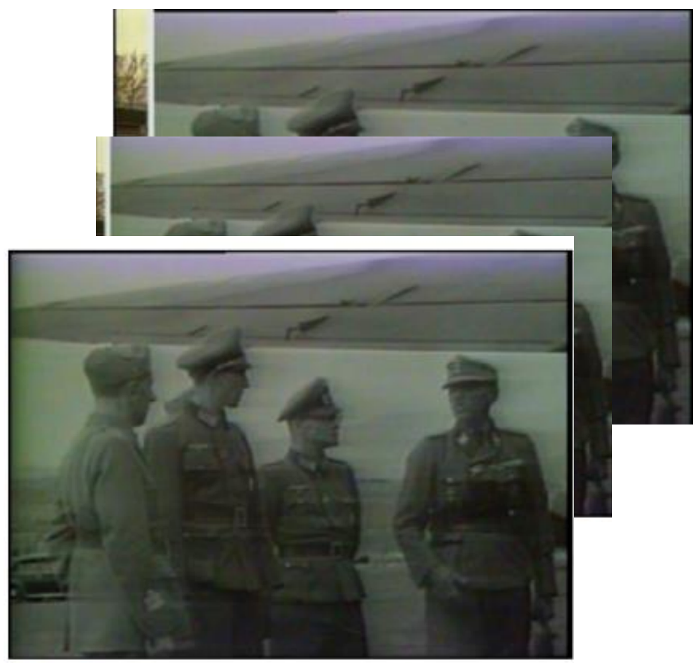
Jayanta Mukhopadhyay
Dept. of Computer Science and Engg.
Indian Institute of Technology, Kharagpur

Video

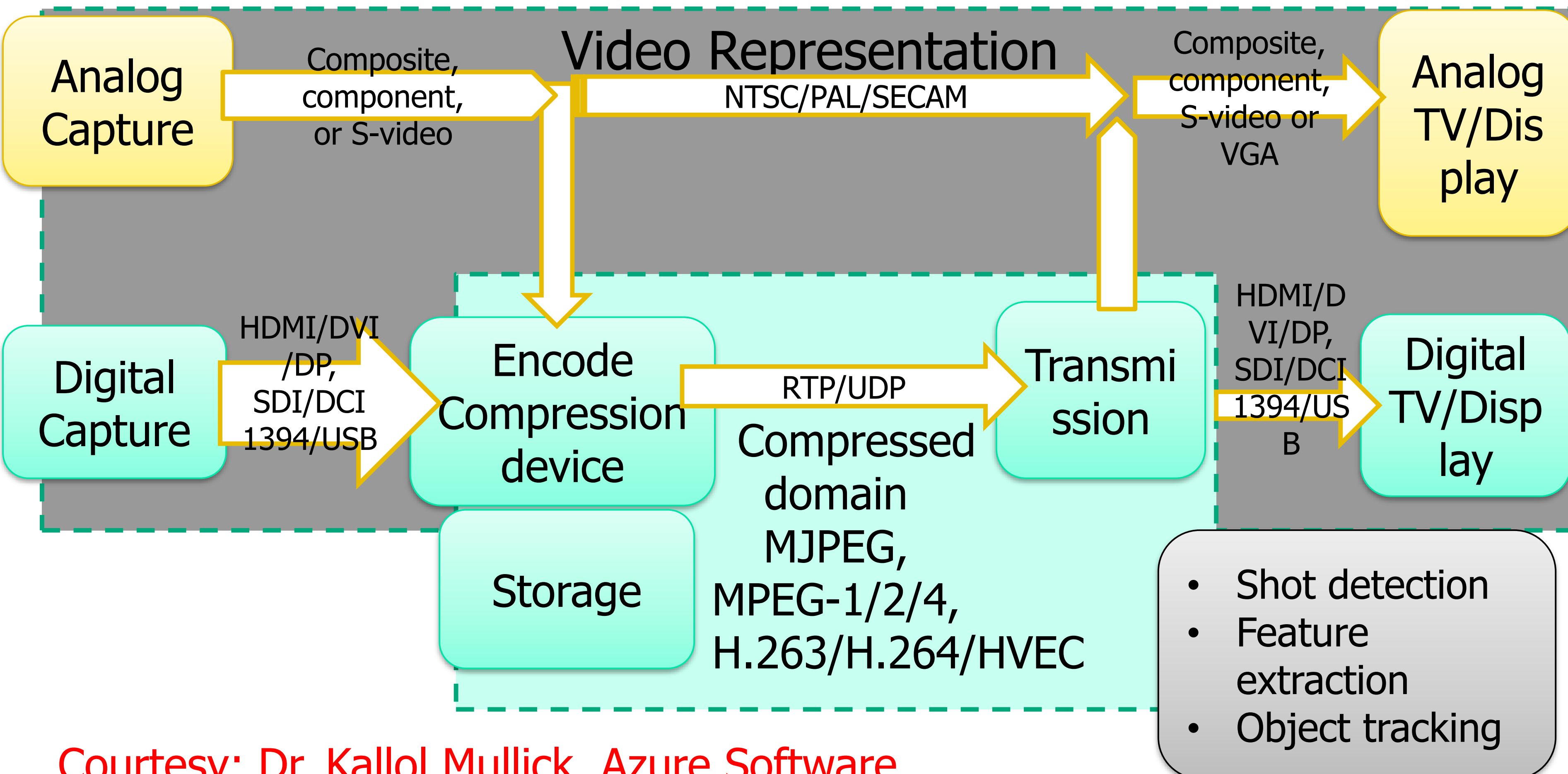
A sequence of frames

$f(x,y,t)$





Video acquisition



Courtesy: Dr. Kallol Mullick, Azure Software

Analog video scanning

A process of conversion of 3D spatio-temporal signal into a 1D temporal signal by periodic vertical-temporal sampling.

- The analog-video signal: a one-dimensional signal s(t) of time
- Obtained by sampling V(x, y, t) in the vertical(y) and temporal(t) coordinates.
- $s(t) = \text{Sampling}_{y,t}(V(x,y,t))$

- Video (3D)
- Spatial 2D+temporal 1D

Temporal sampling

- Frame (2D)
- Spatial 2D

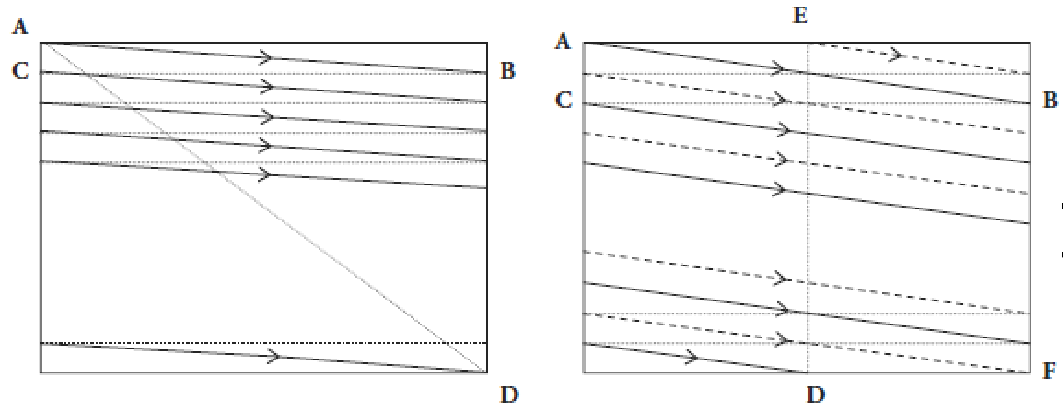
- Scan signal (1D)

Analog video

Courtesy: Dr. Kallol Mullick, Azure Software

Progressive and interlaced scan

Progressive



Interlaced

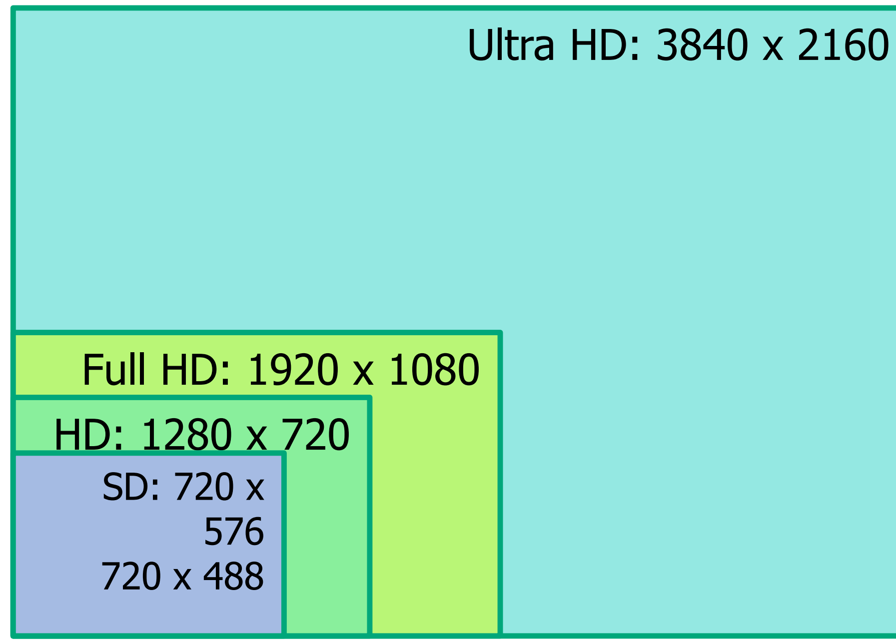
- Scans a frame in every Δt sec
- B->C : horizontal retrace
- D-> A: Vertical retrace
- Computer monitor, with $\Delta t = 1/72$ sec.

- A frame: Odd field + even field
- Solid line: odd field
  - F->A : vertical trace
- Dotted line: even field
  - D->E: vertical trace
- To reduce flickering in TV without extra transmission bandwidth

Courtesy: Dr. Kallol Mullick, Azure Software

Digital video

- Spatial digitization:
  - Pixels in each frame/field
    - Horizontal & vertical resolution
  - HDTV
    - wider aspect ratio 16:9
- Temporal digitization:
  - Frames per second
    - Typically 50/60 Hz
      - 50i : 50 *fields* per second of *interlaced* fields
      - 60p: 60 progressive *frames* per second

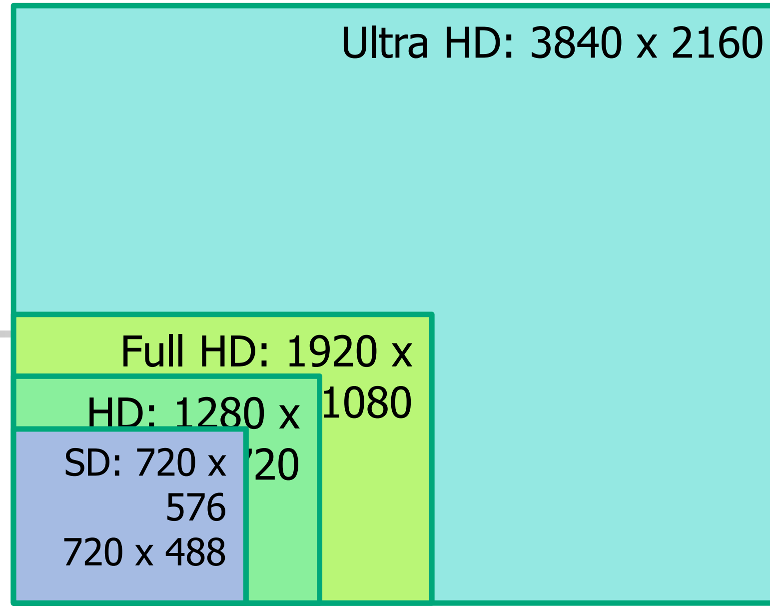


Advantages

Storage, compression
Processing (access, edit, filtering, noise reduction, etc.)
Error-free transmission – noise tolerance, encryption

Courtesy: Dr. Kallol Mullick, Azure Software

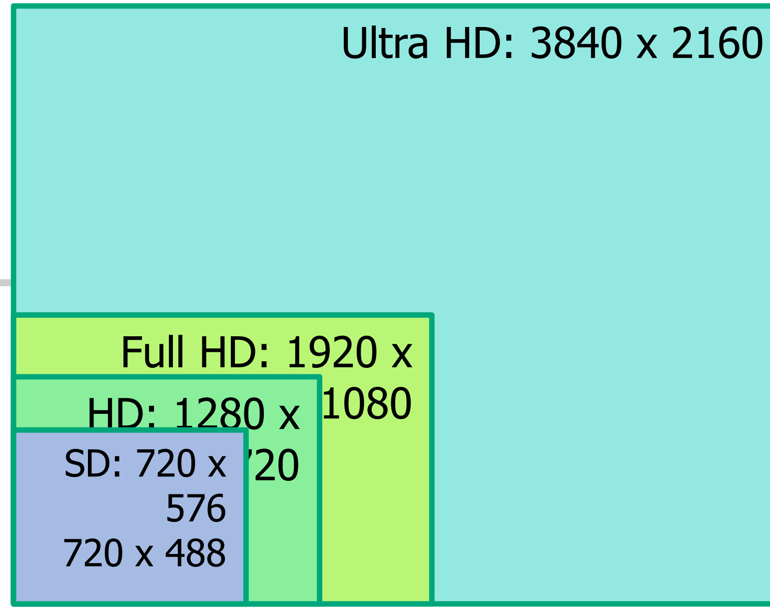
Digital Video Format-I



ITU-R TV standards	Pixels	
VGA	640 x 480	4:3
WXGA	1366 x 768	16:9
SDTV (BT.601-7 480i)	720 x 486	2:1 interlaced, 30 Hz
SDTV (BT.601-7 576i)	720 x 576	2:1 interlaced, 25 Hz
HDTV (BT.709-5 720p)	1280 x 720	Progressive, 50/60 Hz
Full HD (BT.709-5 1080i)	1920 x 1080	Interlaced, 25/30Hz
Full HD (BT.709-5 1080p)	1920 x 1080	Progressive

Courtesy: Dr. Kallol Mullick, Azure Software

Digital Video Format-II



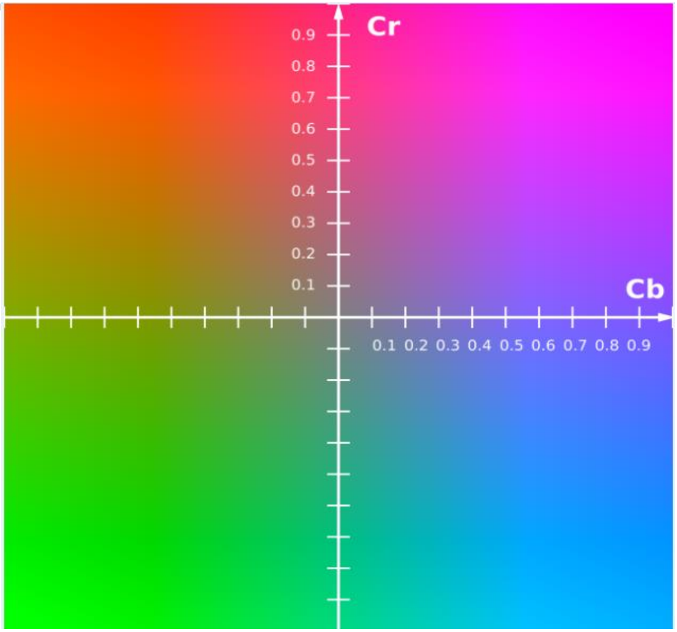
- Used for digital teleconferencing
- Common compromise and easily convertible from NTSC and PAL

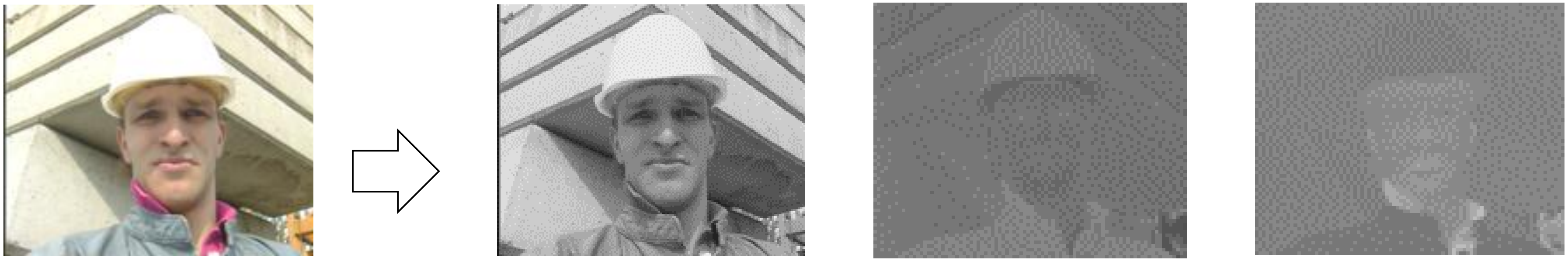
Common standards	Pixels	
QCIF (Quarter CIF)	176 x 144	30 fps, 4:2:0
CIF (Common Intermediate Format)	352 x 288	30 fps, 4:2:0
SIF (Source Input Format) (625/525)	352 × 288 or 352 x 240	25/30 fps, 4:2:0
4CIF	704 x 576	Corresponds to SDTV (BT.601-7 576i)
16CIF	1408 x 1152	

Courtesy: Dr. Kallol Mullick, Azure Software

Digital Video: Color

- Luminance and Chroma
  - U and V for analog = Cr and Cb for digital video
- ITU-R BT.709 defines
  - Y = 0.299 R + 0.587 G + 0.114 B
  - Cr = 0.499 R - 0.418 G - 0.0813 B + 128
  - Cb = - 0.169 R - 0.331 G + 0.499 B + 128





Courtesy: Dr. Kallol Mullick, Azure Software

Digital Video: Chroma sampling

- Sub-sample Chroma to reduce bandwidth
- Sub-sampling schemes
  - A:B:C - Given 4 pixels width (A = 4),
    - B = number of samples of color pixels in row 1
    - C = number of samples of color pixels in row 2

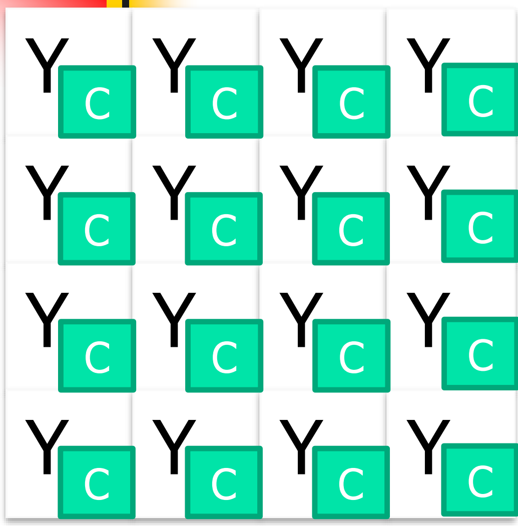
  - 4:4:4 – each pixel Y, Cr and Cb
  - 4:2:2 – each pixel Y; in each row alternate pixel Cr, Cb
  - 4:2:0 - each pixel Y; in alternate row, every alternate pixel Cr, Cb
  - 4:1:1 - each pixel Y; in each row every 4^{th} pixel Cr, Cb
  - 4:1:0 - each pixel Y; in alternate row every 4^{th} pixel Cr, Cb

- most commonly used
- Reduces image size by approx. 17%

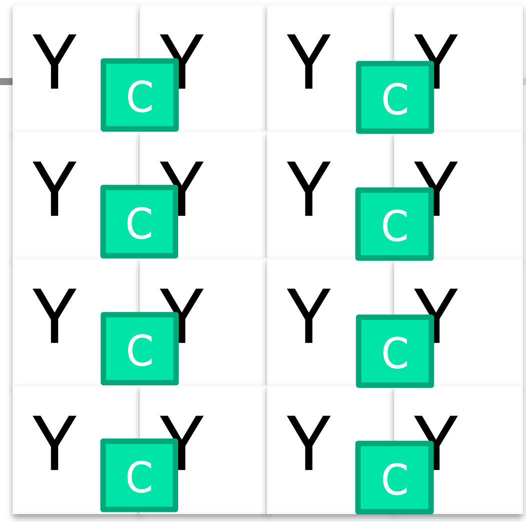
Courtesy: Dr. Kallol Mullick, Azure Software

Digital Video – Chroma sampling

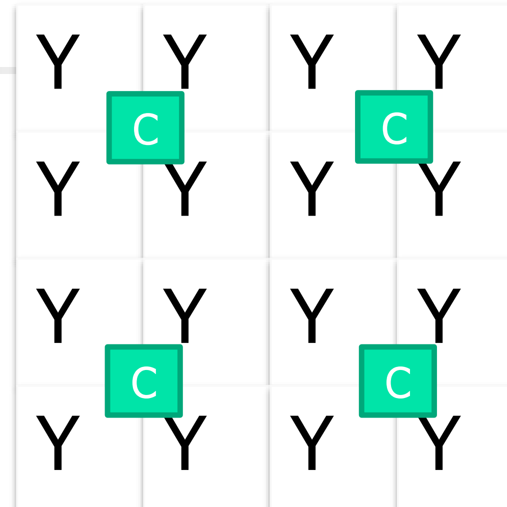
Courtesy: Dr. Kallol Mullick, Azure Software



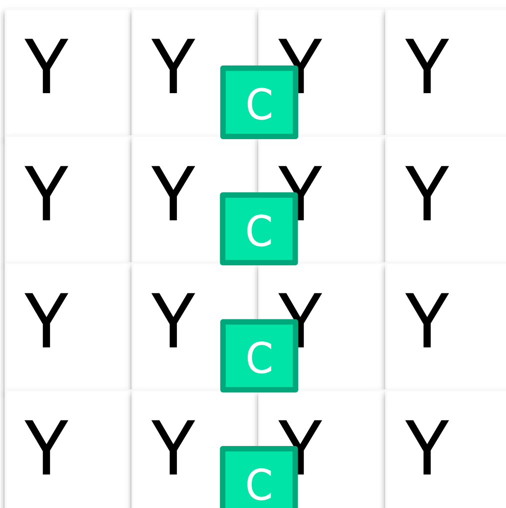
4:4:4



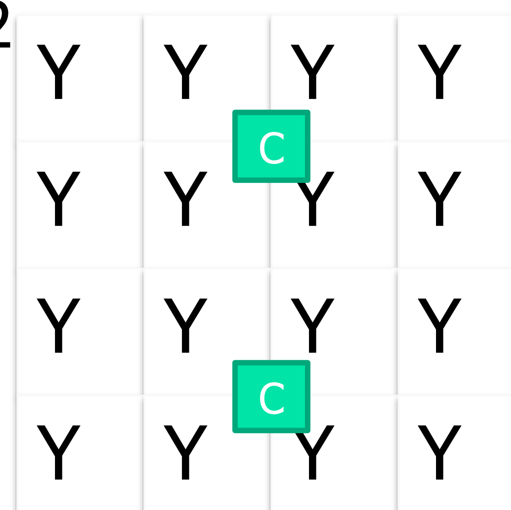
4:2:2



4:2:0



4:1:1



4:1:0

Digital Video: Bit rate and Bandwidth

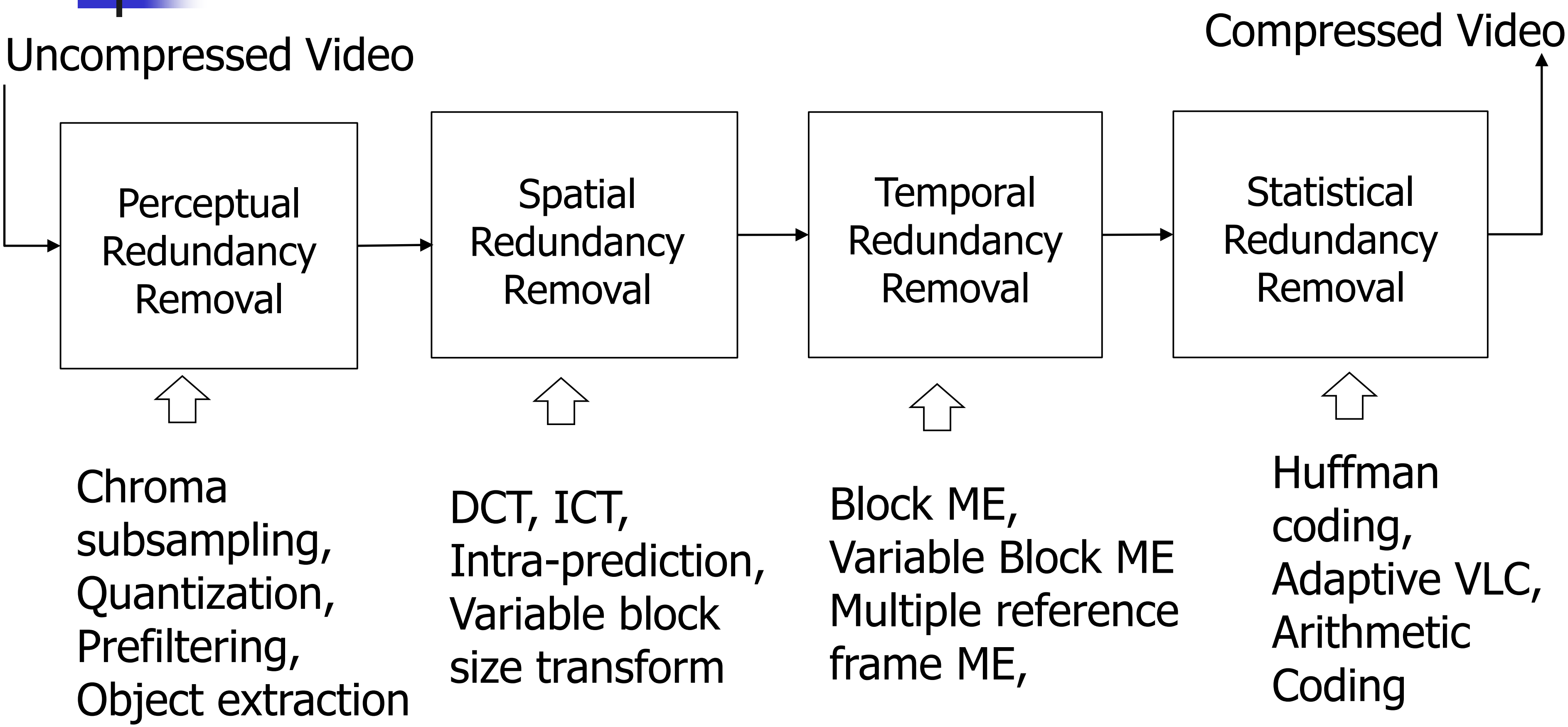
Pixels/frame	640 x 480	
Frames/sec	25	fps
Bits/pixel	24	bits
Duration	1	Hr

- Huge bandwidth requirement for transmission and space requirement for storage.
  - Need Compression ….

Resolution : Pixels/frame	640 * 480	307,200	pixels
Bits/frame :	307,200 * 24	7,372,800	bits
Bit rate:	7,372,800 * 25	184.25	Mbps
Video size:	184.25 * 3600	662,400	Mbits
		82.8	Gbytes

Courtesy: Dr. Kallol Mullick, Azure Software

Video compression



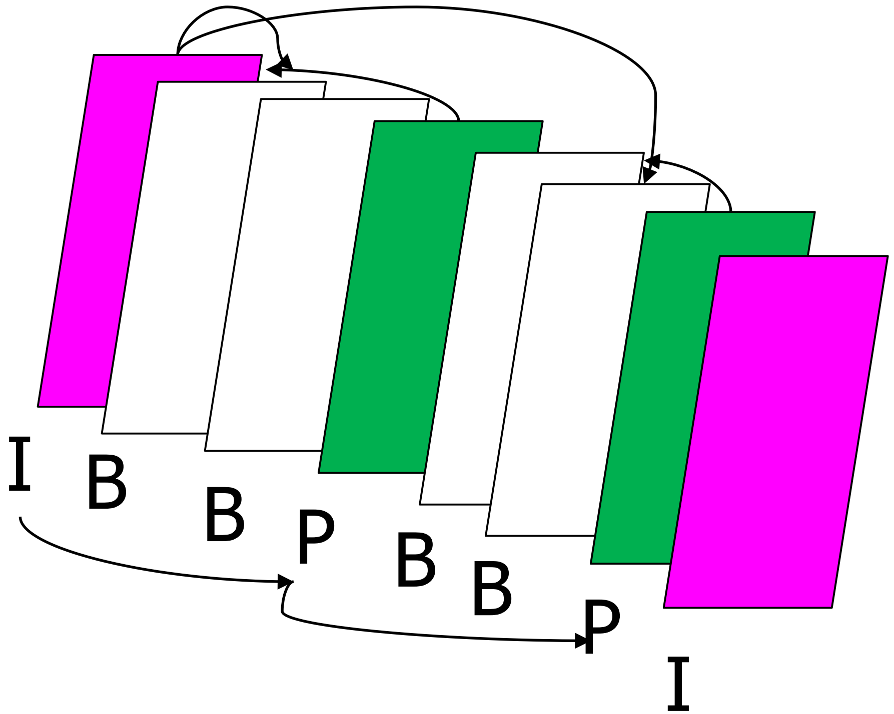
The H.264 Video Coding Standard by Hari Kalva, in IEEE multimedia 13.4 (2006): 86-90.

Video compression

- Video frame organization
  - GOP, Frame, Slice, MB
- Exploit Redundancy
  - Spatial –
    - Intra-frame
  - Temporal –
    - Inter-frame
- Motion estimation
  - Motion Vector (MV) – Displacement between the current block and the reference block.
  - Transmit (MV, ref) and diff. between current block and ref block
  - Block Matching Algorithms - find closest match block in spatial and temporal neighborhood.

Courtesy: Dr. Kallol Mullick, Azure Software

Video Frames: I, P and B



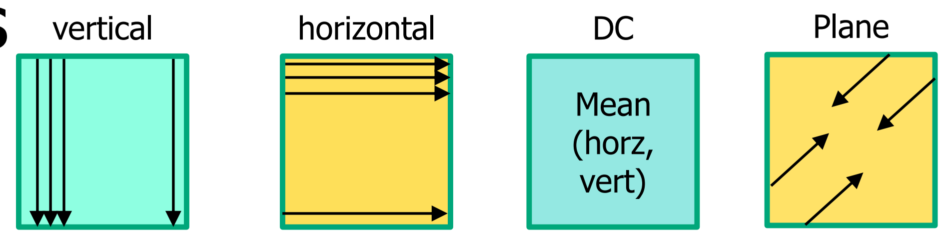
I: Intra Coded Frame
P: Predicted Coded Frame
B: Bi-directionally
Predicted Coded Frame

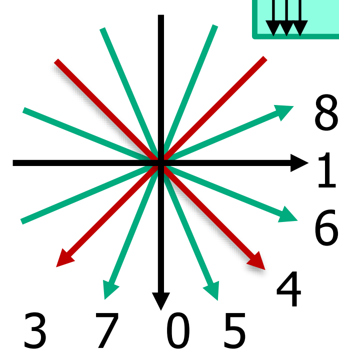
I Frame used for fast forwarding

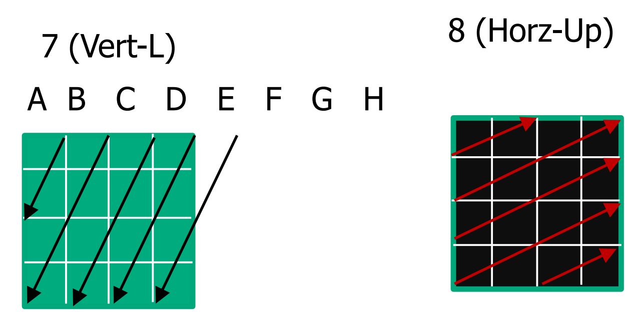
A typical GOP: IBBPBBP..

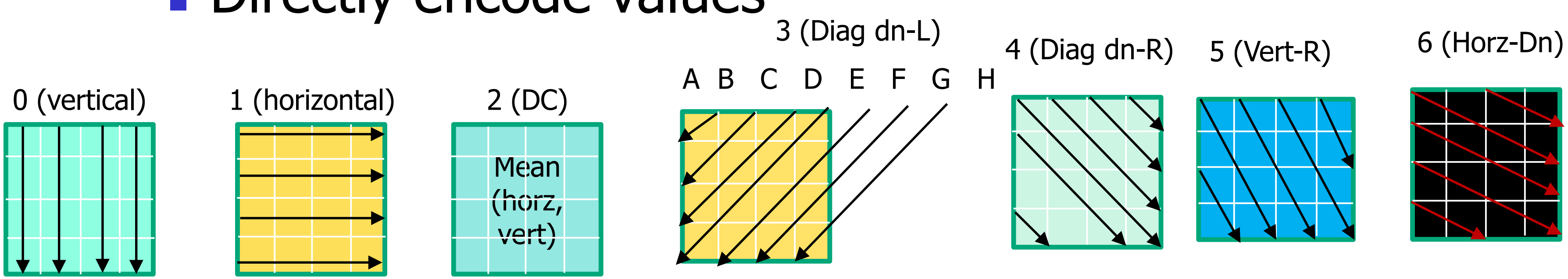
Intra-frame prediction

- Intra prediction modes
  - Intra 16x16
    - 4 modes
  - Intra 4x4
    - 9 modes
  - I_PCM
    - Directly encode values



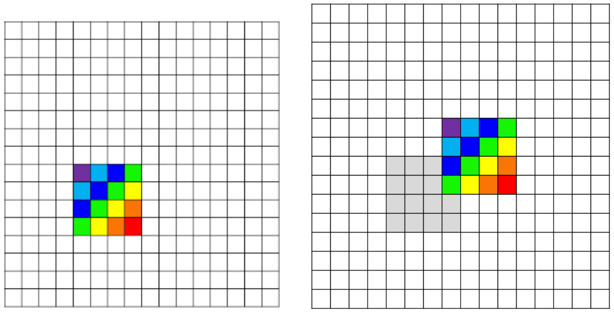






Courtesy: Dr. Kallol Mullick, Azure Software

Motion Estimation



$\Psi_1 = \Psi(x, y, t_1)$ $\Psi_2 = \Psi(x, y, t_2)$

- Reference frame: $\Psi_1 = \Psi(x, y, t_1)$
- Current frame $\Psi_2 = \Psi(x, y, t_2)$
  - Point (x, y) in Ψ_1 moved to $\left(x + d_x, y + d_y\right)$ in Ψ_2
  - $MV_{t_1}^{t_2}(x, y) = \left(d_x, d_y\right)$
- Motion estimation types
  - $t_1 < t_2$: forward motion estimation
  - $t_1 > t_2$: backward motion estimation

Courtesy: Dr. Kallol Mullick, Azure Software

"Video Processing and Communication", by Y. Wang, Y.-Q. Zhang, and Ostermann

Courtesy: Dr. Kallol Mullick, Azure Software

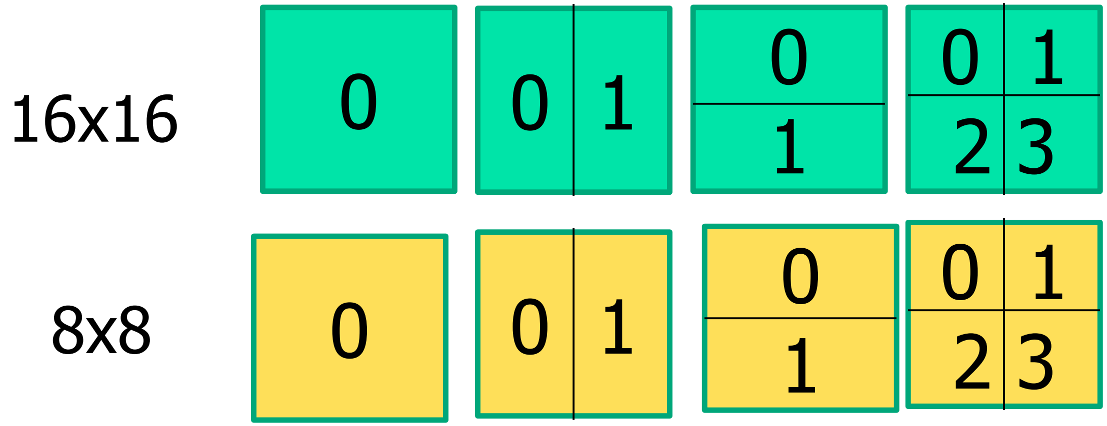
Motion search

- With motion vector $MV_{t_1}^{t_2}(x,y)$, motion compensated estimate,
  - $\Psi_2' = MC\left(\Psi_1, MV_{t_1}^{t_2}(x,y)\right)$
  - Residual for a motion vector $MV_{t_1}^{t_2}(x,y)$
    - $R(x,y) = \Psi_2(x,y) - \Psi_2'(x,y)$
- Motion search: $D(x,y) = \underset{\Psi_1,t_1}{\operatorname{argmin}} D\left(\Psi_2, MV_{t_1}^{t_2}(x,y)\right)$
  - $D\left(\Psi_2, MV_{t_1}^{t_2}(x,y)\right) = \sum_{x,y\in A} |\Psi_2(x,y) - \Psi_2'(x,y)|^p$, $p>0$
    - $p = 1$: MAD; $p = 2$: MSE
  - Selection over reference frame, reference block

Courtesy: Dr. Kallol Mullick, Azure Software

Motion estimation-inter-frame

- Different MV for each partition of MB
- Flexible partitioning (modes)



- SKIP mode
  - Neither MV, nor residual is transmitted
  - MV is predicted from neighboring MBs

Motion Search: Block Matching

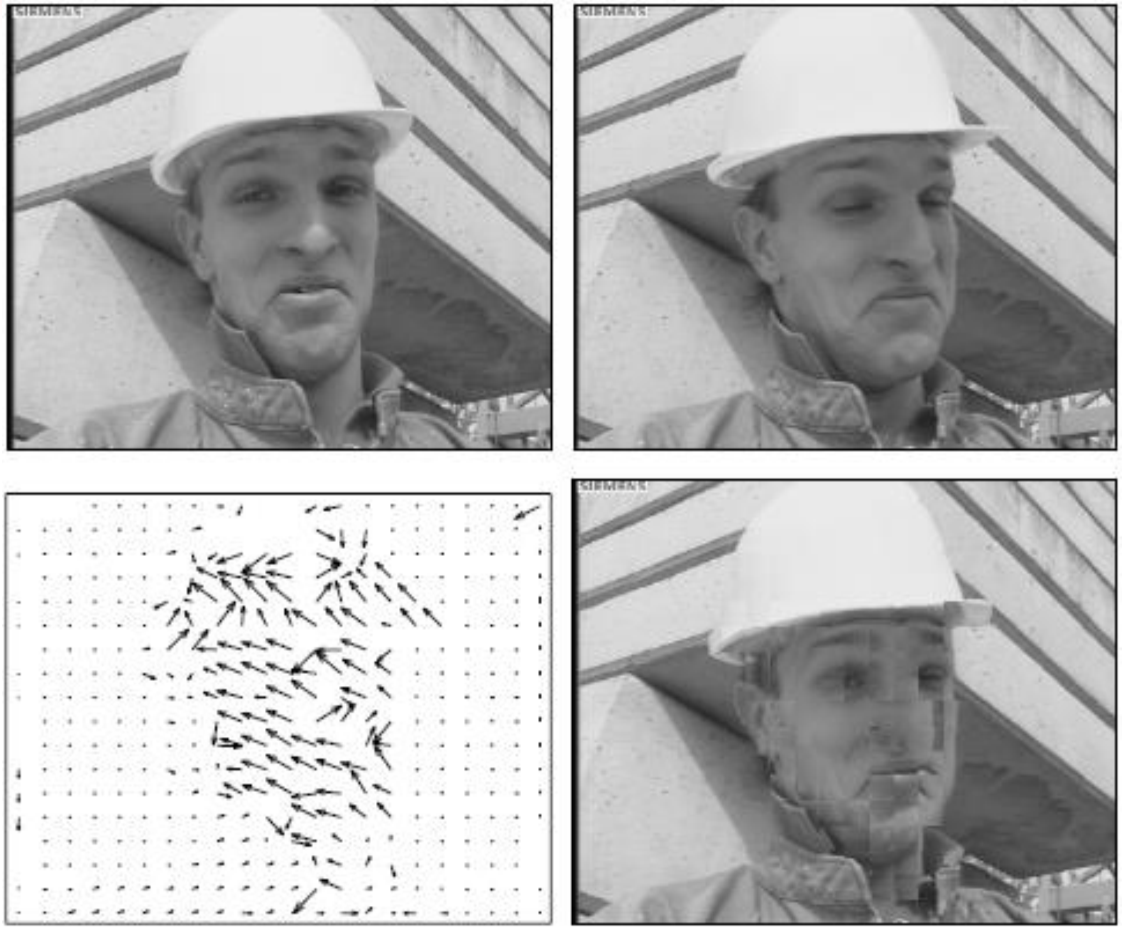
- Exhaustive search Block Matching Algorithm (EBMA)
- Reduced neighborhood search
- Fractional pixel Accuracy

- Bi-directional MV
- Multiple reference frames
  - Weighted average

Courtesy: Dr. Kallol Mullick, Azure Software

Courtesy: Dr. Kallol Mullick, Azure Software

Motion compensation



- [LT] Current frame
- [RT] Reference frame
- [LB] Motion vectors
- [RB] Motion compensated image

"Video Processing and Communication", by Y. Wang, Y.-Q. Zhang, and Ostermann

Video Shot

- Shot – a set of consecutive frames
    - Maximal continuous partition of video sequence in temporal domain
    - Without interruption
    - Single capture device (camera)
- GOP – fixed length syntactical segment;
- Shot – variable length (max-length) semantical segment

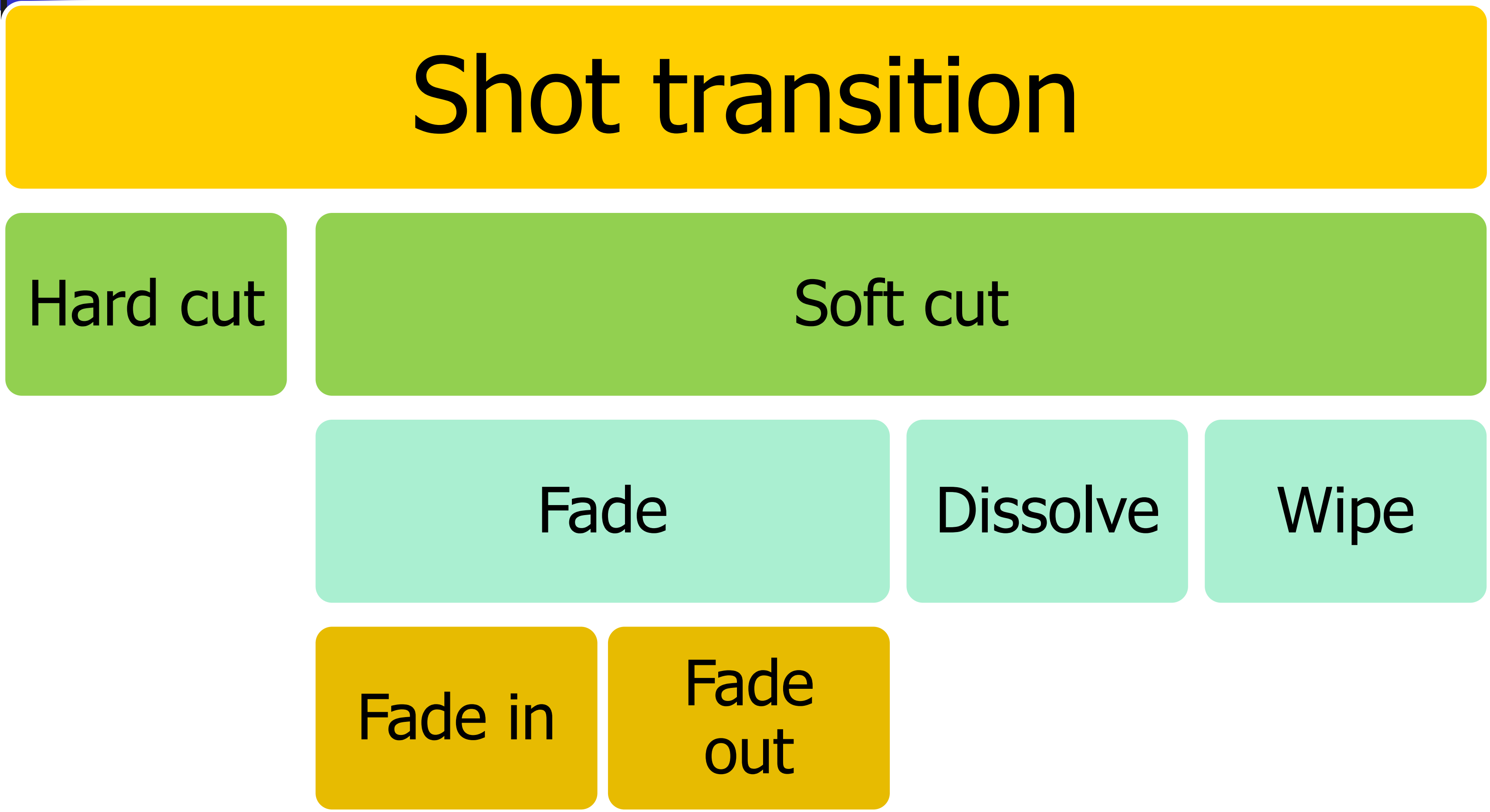
Courtesy: Dr. Kallol Mullick, Azure Software

Shot detection: Objective and Applications

- Objective
  - Information about specific objects or events
- Video analysis
    - Video abstraction or summarization,
    - Annotation, indexing
    - Video semantic analysis,
    - Content based retrieval,
    - Classification.

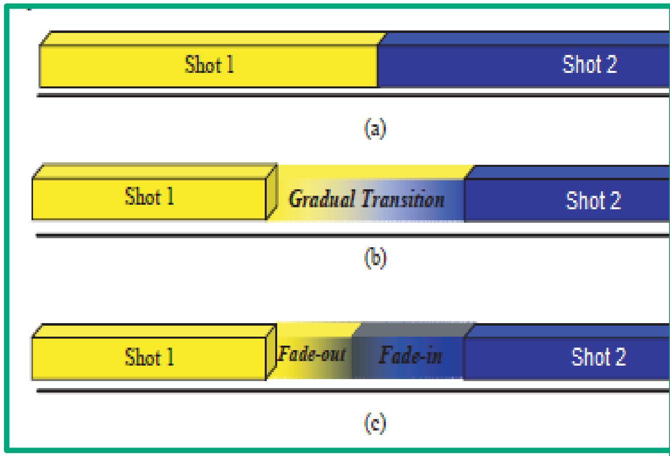
Courtesy: Dr. Kallol Mullick, Azure Software

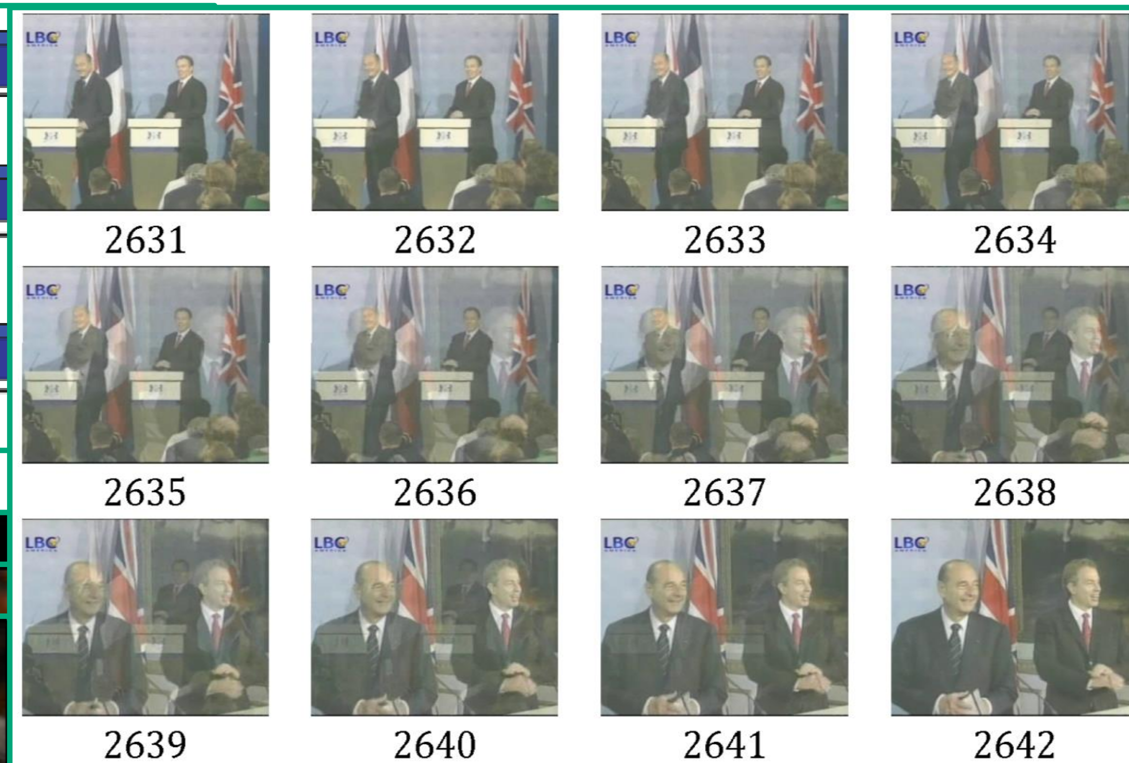
Shot Boundary



Courtesy: Dr. Kallol Mullick, Azure Software

Shot transition types





Fade

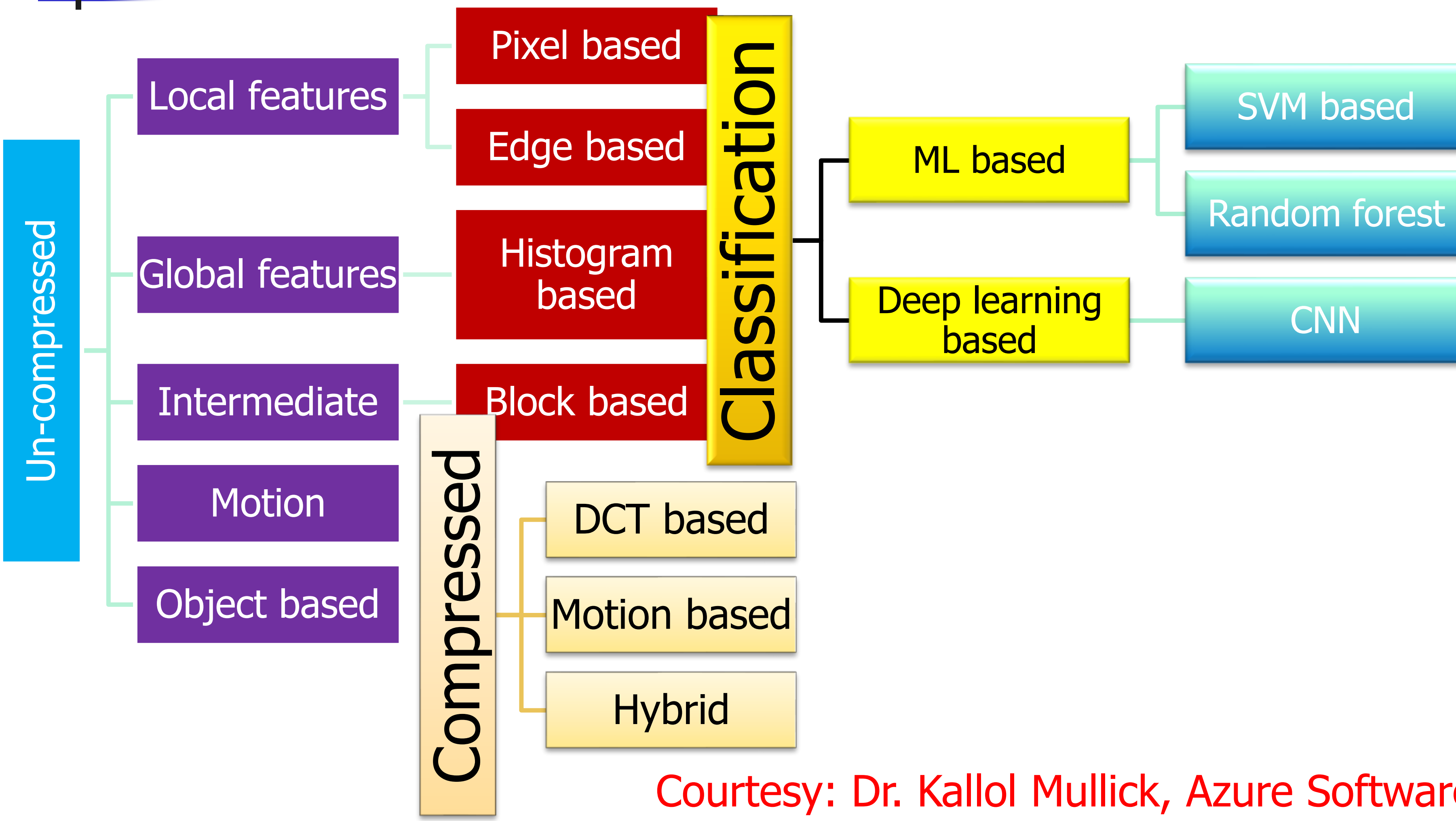
Courtesy: Dr. Kallol Mullick, Azure Software

Shot Boundary Detection: Approaches

- Extract features – intensity, color, edge, object
- Threshold Based Approach
  - $d(n, m)$ = measure of dissimilarity between frame m and n
  - $f(n)$ = discontinuity feature value for n-th frame:
    - $f(n) = d(n-1, n)$
  - Detect a cut if $f(n) > Th$, where Th is some threshold
    - Predefined or adaptive (e.g. sliding window)
- Classify frame as normal frame, hard cut of soft cut
  - Different approaches for classification

Courtesy: Dr. Kallol Mullick, Azure Software

Shot Boundary Detection Approaches



Courtesy: Dr. Kallol Mullick, Azure Software

Courtesy: Dr. Kallol Mullick, Azure Software

Dissimilarity functions

- Minkowski distance:
  - $d(f_i, f_{i-1}) = \left(\sum_k |Z_i(k) - Z_{i-1}(k)|^p\right)^{\frac{1}{p}}$, where $Z(.)$ is a feature function
  - $p = 1$: City-block distance; $\mathrm{p} = 2$: Euclidean distance
- [Local] Pixel based: $\left(\sum_x \sum_y \left|I_{f_i}(x,y) - I_{f_{i-1}}(x,y)\right|\right) > T$
- [Local] Edge based: compare the number of edge pixels between consecutive frames
  - Missed spatial information on edges though

"Methods and Challenges in Shot Boundary Detection: A Review" by Abdulhussain, et.al. March 2018, Entropy 20(4)

Dissimilarity functions

- [Global] Histogram based: $\left(\sum_v \left| H_{f_i}(v) - H_{f_{i-1}}(v) \right| \right) > T$
  - [v is the v-th bin in histogram]
  - Less sensitive to object and camera motion
- [Intermediate] Block based: use of statistical measures (mean/var), histogram, likelihood ratio

"Methods and Challenges in Shot Boundary Detection: A Review" by Abdulhussain, et.al. March 2018, Entropy 20(4)

Courtesy: Dr. Kallol Mullick, Azure Software

SVM for shot boundary detection

- Support Vector Machine (SVM) classification
  - two-pass hierarchical supervised approach –
    - 1st pass for cut detection, and 2nd pass for gradual transition
  - Formation of feature vectors
    - combining color histograms and few moment measures
    - Dissimilarity measures – overall or with temporal neighbors (1, 2, 6)
  - SVM classifier – 3 class: (normal sequences, abrupt cuts and gradual transitions)
    - K-means cluster or one-against-one
  - Active learning
    - first minutes of a video used for training

"Simultaneous detection of abrupt cuts and dissolves in videos using support vector machines" by Chasanis, Likas and Galatsanos, PRL (2009) 55-65

Courtesy: Dr. Kallol Mullick, Azure Software

Spatio-Temporal Features

- 2D frame + time → 3D volume
  - 3D volume assumes 3 orthogonal filters.
    - the orthogonality not required for motion analysis
  - Gaussian filtering "smoothes out" motions regions
- STIP – Spatio temporal interest point [Laptev 2003]

"Survey of Spatio-Temporal Interest Point Detection Algorithms in Video" by Y. Li, et. al.

Courtesy: Dr. Kallol Mullick, Azure Software

Spatio-Temporal Corners (STC) – Harris 3D

- Harris3D detector
  - linear scale-space representation $L = f * g$
    - Anisotropic Gaussian kernel,
    - $g(x, y, t, \sigma^2, \tau^2) = \frac{1}{\sqrt{(2\pi)^3 . \sigma^4 . \tau^3}} . e^{-\frac{x^2+y^2}{2\sigma^2} - \frac{t^2}{2\tau^2}}$
    - independent spatial variance $\sigma^2$and temporal variance τ^2
  - $\mu = g(.) * \begin{bmatrix} L_{xx} & L_{xy} & L_{xt} \\ L_{yx} & L_{yy} & L_{yt} \\ L_{tx} & L_{ty} & L_{tt} \end{bmatrix}$, $\sigma_\mu = s_1\sigma$, $\tau_\mu = s_2\tau$
    - where L_{xx} = 2nd order derivative of L

• non-maximum suppression.

• STC points by imposing some constraints

- Response function : $H = det(\mu) - k.trace^3(\mu)$, $k<1/27$

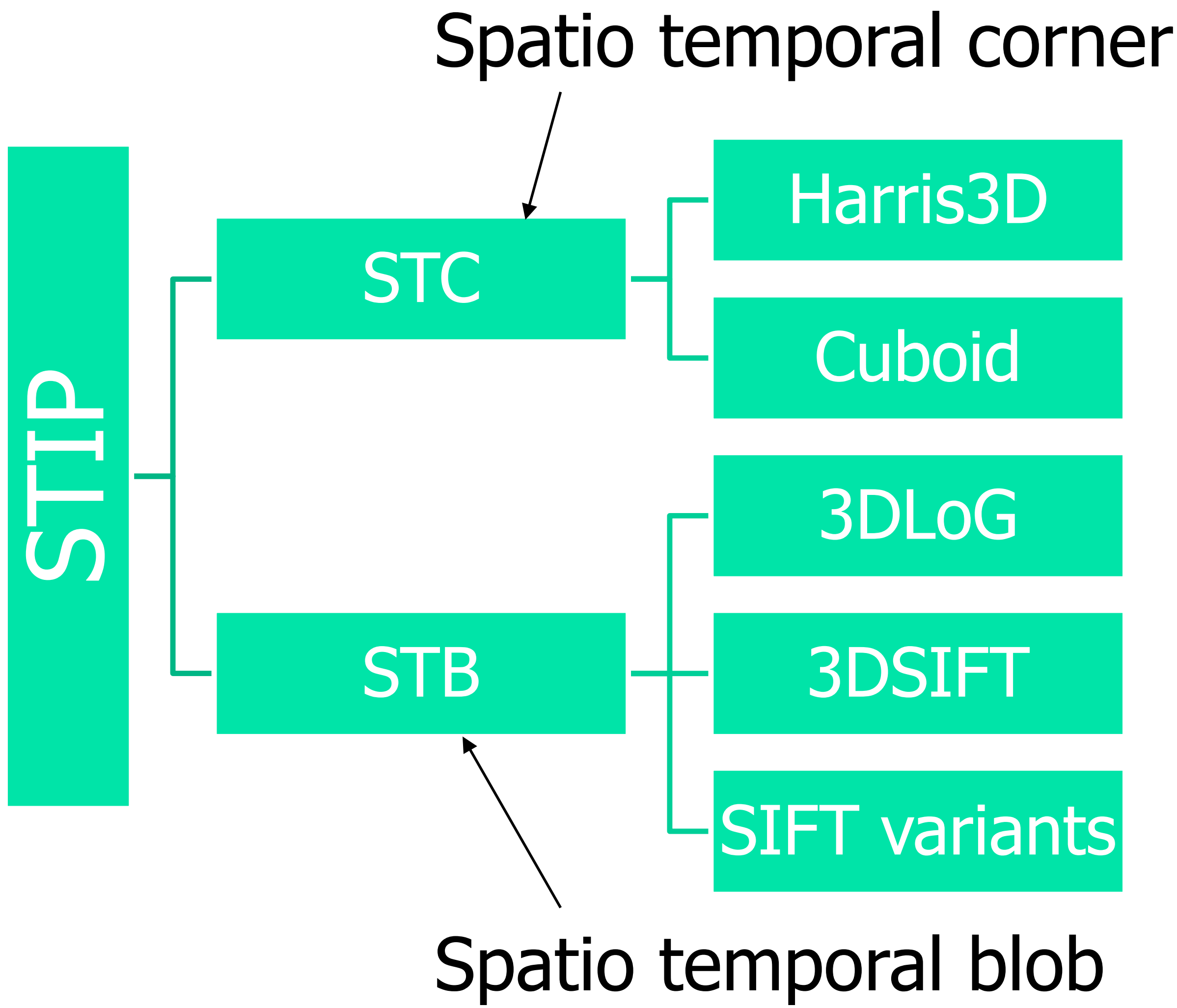
Courtesy: Dr. Kallol Mullick, Azure Software

STB – Spatio temporal blob

- 3DLoG
  - extended from Laplace of Gaussian (LoG)
  - $3DLoG(x, y, t, \sigma, \tau) = \sigma^3 \nabla^2 g = \frac{x^2+y^2+t^2 - 3\sigma\tau}{2\pi\tau^3} e^{-\frac{x^2+y^2+t^2}{2\sigma\tau}}$
  - g – 3D Gaussian kernel with the spatial and temporal scale parameters
  - σ - Spatial scale parameter, τ - temporal scale parameter
  - $\nabla^2 g$ - is the Laplace operator of g
- May be approximated by 3D DoG.

The extreme values of the second-order derivative.

Spatio-Temporal Features: Summary



Summary

- Video: A sequence of frames: $I(x,y,t)$
- Compressed representation required
  - Motion compensation, Residual error encoding + Motion Vectors + Intra frame prediction
- Shot boundary detection
  - Dissimilarity among consecutive frames
  - Classification of a frame to transition types
- Spatio Temporal Features
  - Spatio Temporal Corner (STC): Harris-3D detector
  - Spatio Temporal Blob (STB): 3D LoG, 3D DoG

Thank you!