CTT310: Digital Image Processing

Video Processing Fundamentals

Dr. Nguyen Ngoc Thao
Department of Computer Science, FIT, HCMUS

Outline

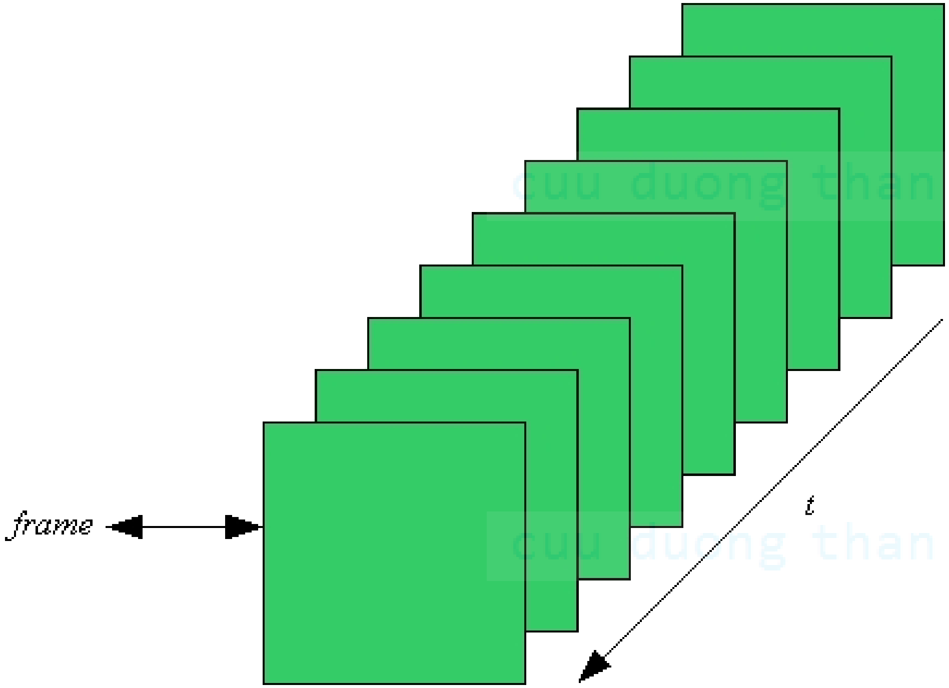
- Introduction to Video processing
- Video coding

Section 8.1

VIDEO PROCESSING

Video signals

- **Video signal** is a sequence of 2-D images captured from the projection of a 3-D scene onto an image plane.



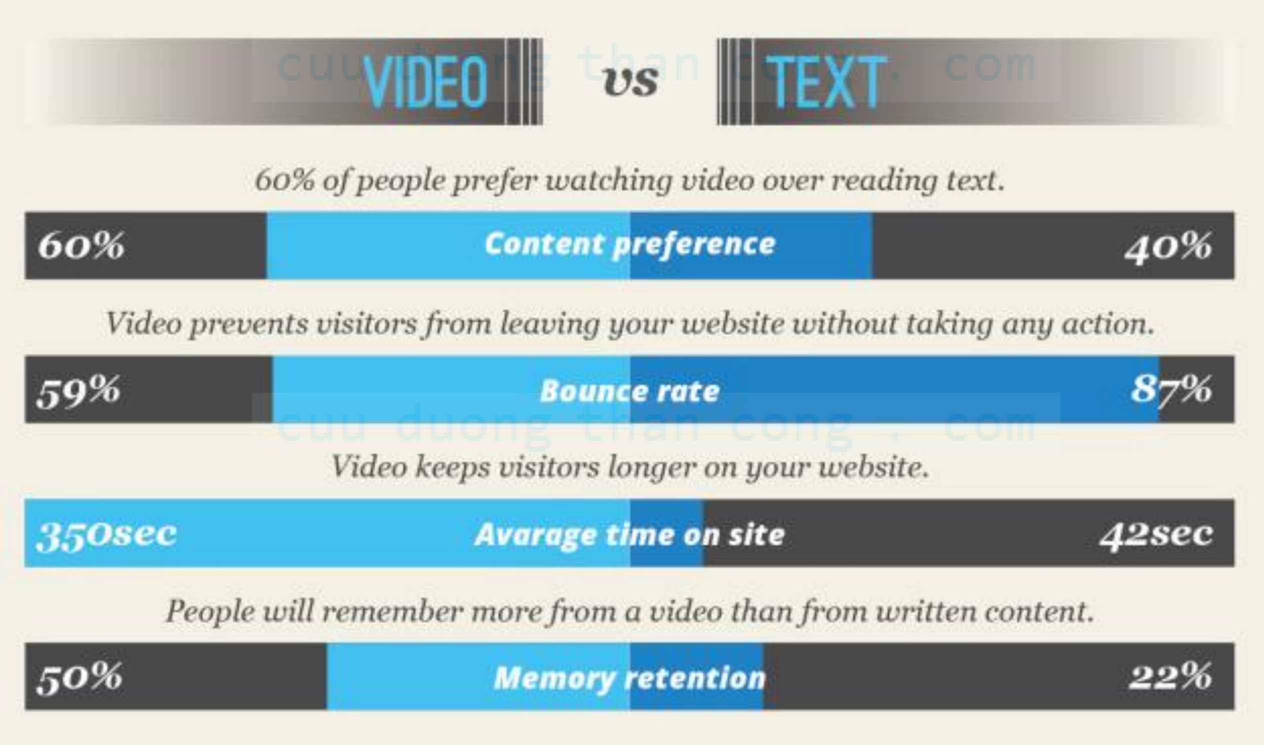
Temporal sampling of video signals

- A video consists of a sequence of images, displayed in rapid succession, to give an illusion of continuous motion.
  - If the time gap between successive frames is too large, the viewer will observe jerky motion.

https://www.youtube.com/watch?v=2r6YpMzNyMk

Why do we care video?

- Visual representations are often the most efficient way to represent information.
  - Most information (85%) are acquired by human vision system
  - Telling stories (movie), sharing ideas (demo), scouting scene (surveillance), communicating effectively (video conferencing).



60% of people prefer watching video over reading text.

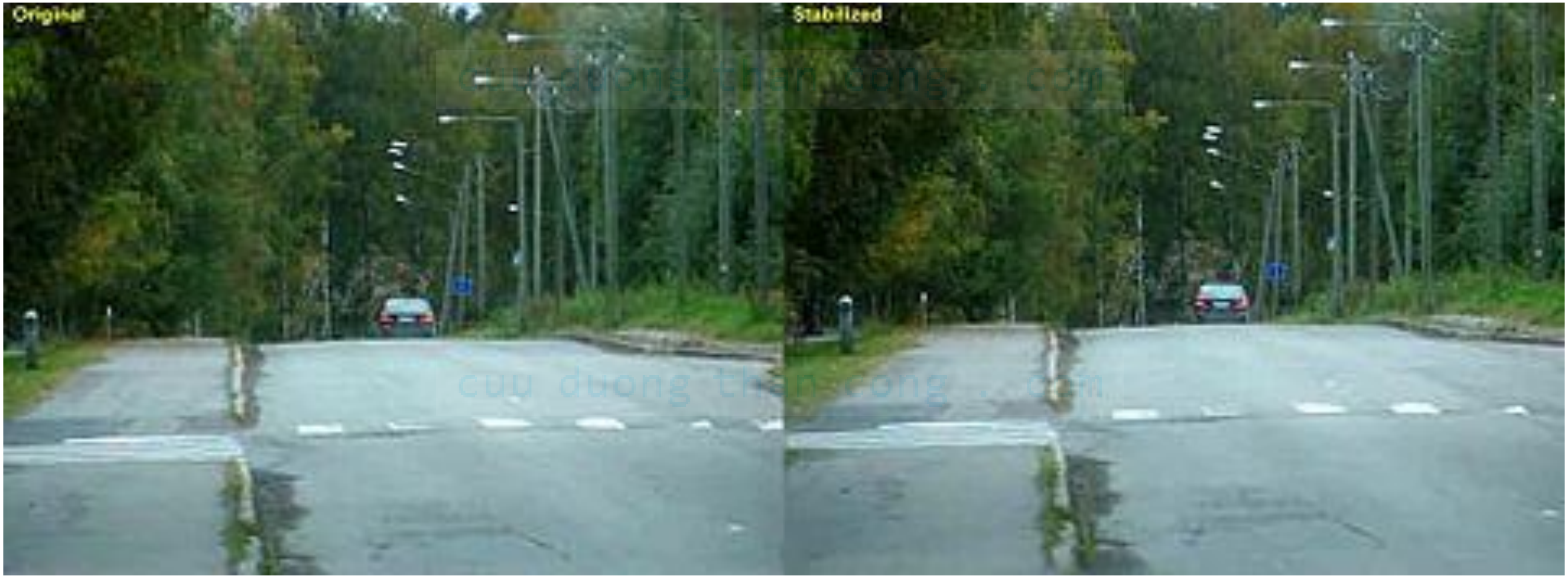
Video prevents visitors from leaving your website without taking any action.

Video keeps visitors longer on your website.

People will remember more from a video than from written content.

Video processing

- **Video processing** is a particular case of signal processing, which often employs video filters and where the input and output signals are video files or video streams.
- Purpose: improves the apparent definition of video signals



https://en.wikipedia.org/wiki/Image_stabilization

Why do we process video?

- Tasks in video processing involve the manipulation of videos' characteristics, such as
  - Deinterlacing
  - Aspect ratio control, digital zoom and pan, frame rate conversion
  - Brightness/contrast/hue/saturation/sharpness/gamma adjustments
  - Color point conversion (601 to 709 or 709 to 601), color space conversion (YPBPR/YCBCR to RGB or RGB to YPBPR/YCBCR)
  - Primary and secondary color calibration (including hue/saturation/luminance controls independently for each)
  - Mosquito noise reduction, block noise reduction
  - Detail enhancement, edge enhancement
  - Motion compensation

Why do we process video?

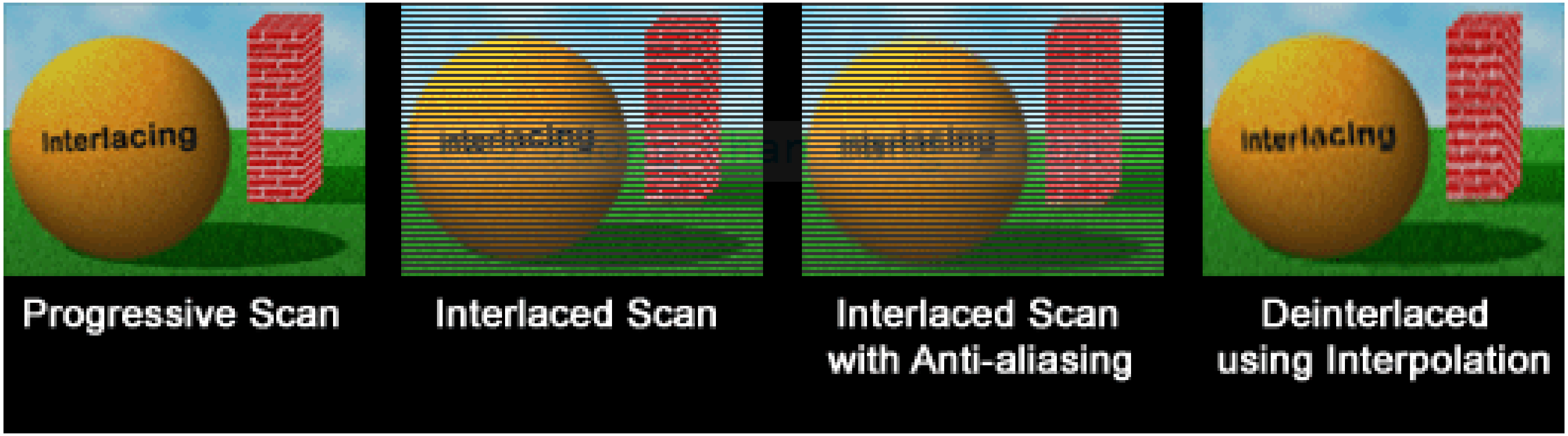
- They are also generalization of (static) image techniques into the (dynamic) video environment
  - Detail enhancement, denoising, deblurring, restoration,…
  - Visual mosaicking, inpainting
  - Motion tracking, face recognition, segmentation…

Why do we process video?

- Video formation systems are not ideal
  - Videos can be corrupted, resolutions are limited.
- Vast video data are challenging for storage/transmission.
  - To see more will less storage/bandwidth
- Autonomous systems are desirable by making computer to understand videos

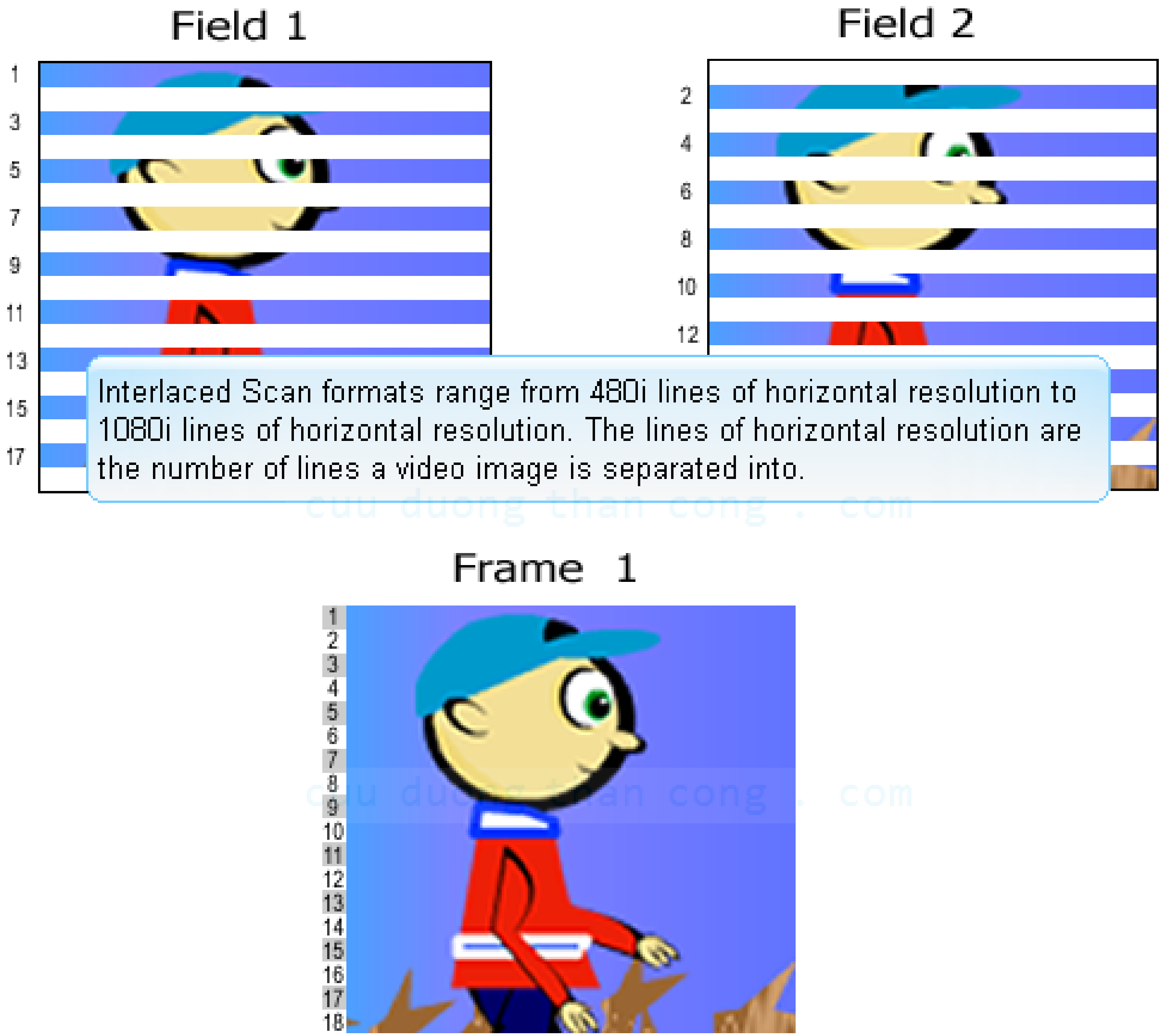
Video deinterlacing

- **Deinterlacing** is the process of converting interlaced video, such as common analog television signals or 1080i format HDTV signals, into a non-interlaced form.


Interlacing
Progressive Scan
Interlaced Scan
Interlaced Scan with Anti-aliasing
Deinterlaced using Interpolation

Interlaced vs. Progressive scanning

- **Interlaced scanning** uses two fields captured at two different times to create a video frame.
  - One field contains all odd-numbered lines in the image while the other contains all even-numbered lines.
  - E.g., a television scans 60 fields/sec (30 odd and 30 even) → these two sets of 30 fields are combined to create a 30fps full frame.
- **Progressive scanning** is where the lines are drawn in one at a time in sequential order.
  - E.g., the entire single 30fps frame image is painted every 1/60th of a second, allowing for twice the detail to be sent.



Interlaced Scan formats range from 480i lines of horizontal resolution to 1080i lines of horizontal resolution. The lines of horizontal resolution are the number of lines a video image is separated into.

Interlacing scanning

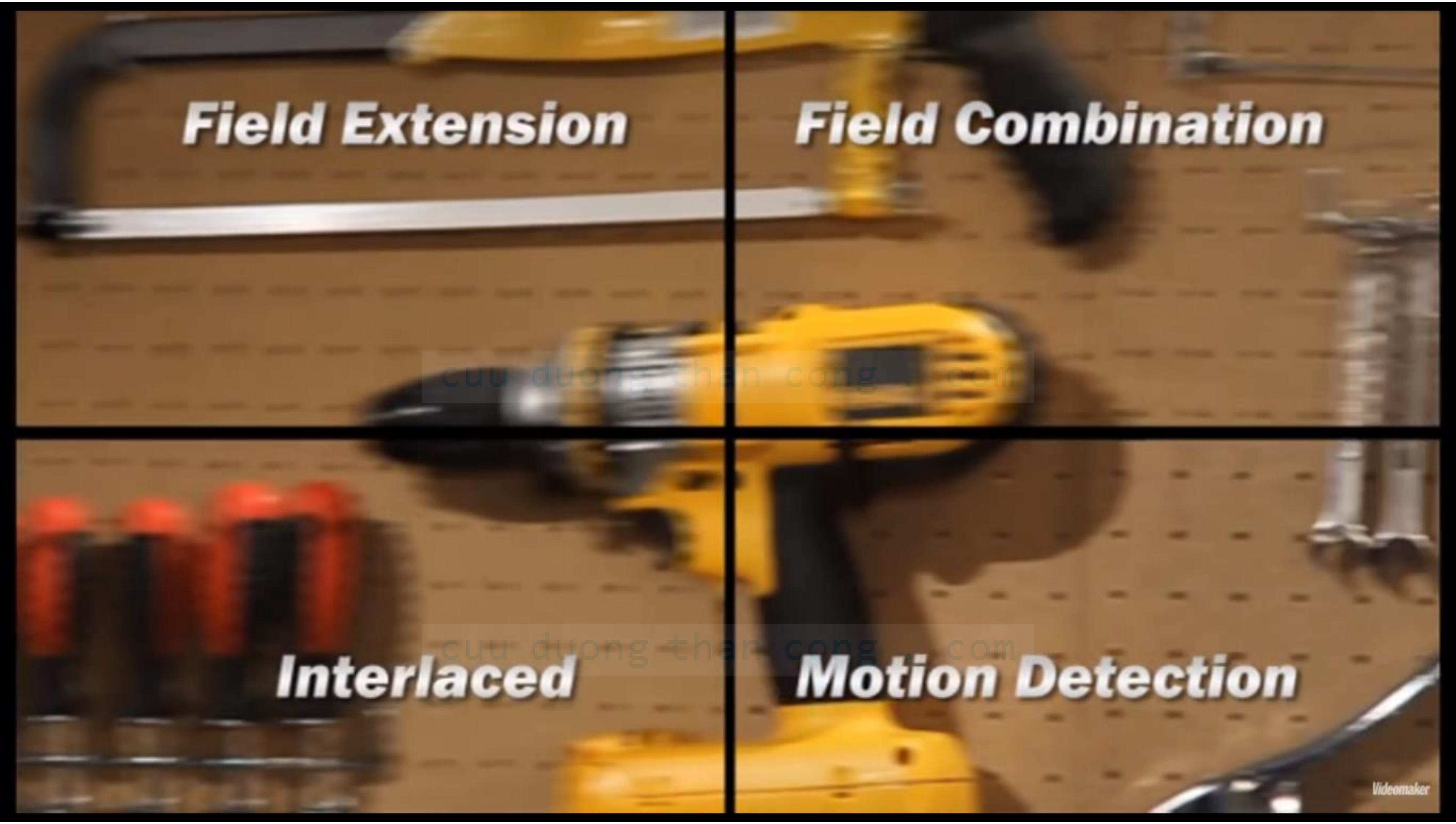
- Interlaced scanning allows for doubling the perceived frame rate of a video display without consuming extra bandwidth, but suffers flicker, lower resolution and quality issues
- Analog TV and old CRT-based displays were able to display interlaced video correctly due to their complete analogue nature.
- Newer digital displays require the two fields to be combined into a single frame, which leads to various visual defects.

Video deinterlacing

- There are various methods to deinterlace video, each producing different problems or artifacts of its own.
- **Field combination:** the even and odd fields are combined into one frame which is then displayed
- **Field extension:** each field (with only half the lines) is extended to the entire screen to make a frame
- **Motion compensation** and others: a combination of both

- The problem has been researched for decades and employs complex processing algorithms, yet consistent results have been very hard to achieve

Example of video deinterlacing

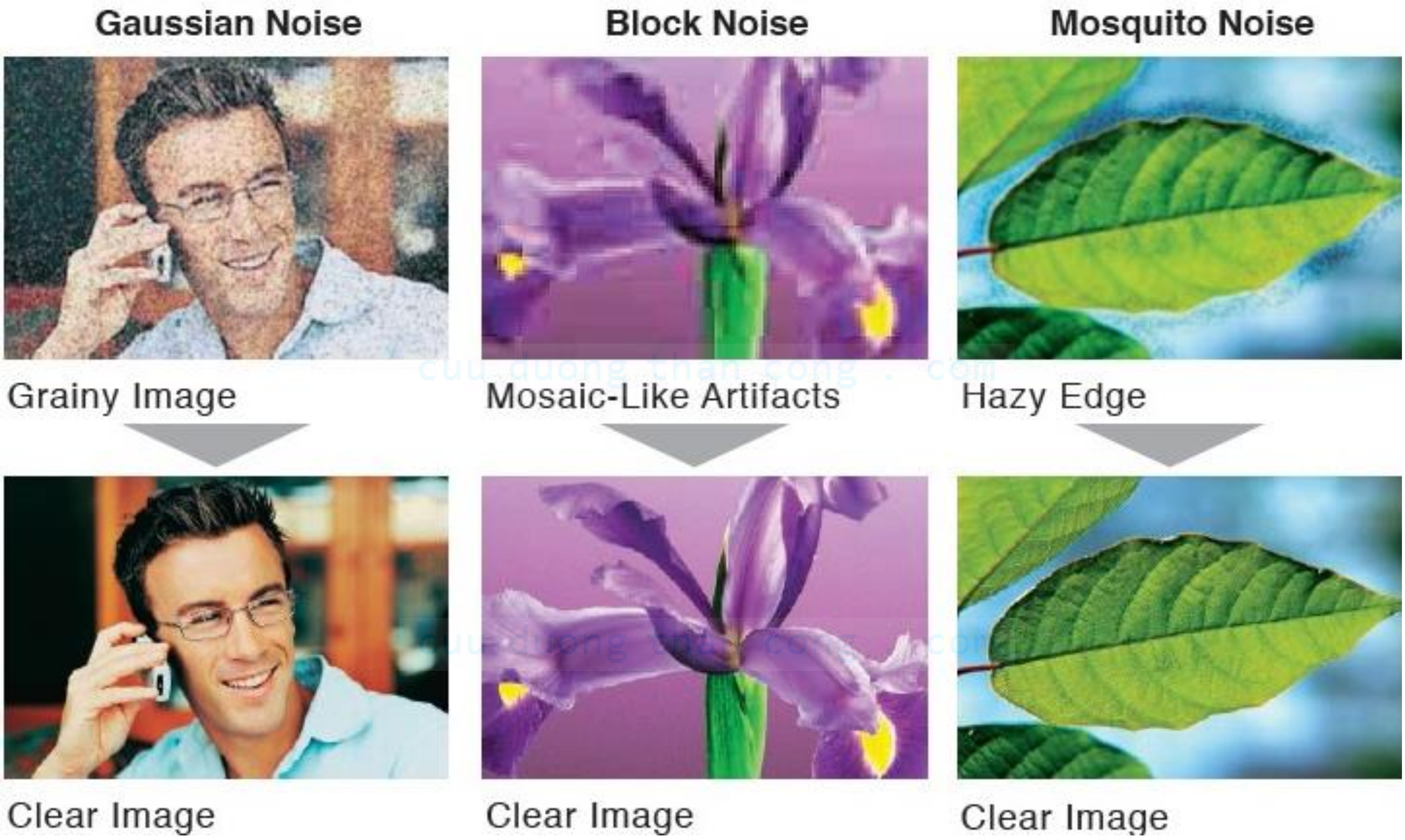


https://www.youtube.com/watch?v=YczLRshnxQ8

Noise reduction

- Video compression artifacts include cumulative results of compression of the comprising still images.
- **Block boundary discontinuities:** occur at edges during the motion compensated video compression.
  - The current picture is predicted by shifting blocks of pixels from previously decoded frames. If two neighboring blocks use different motion vectors, there is a discontinuity at the edge between blocks.
- **Mosquito noise:** ringing or other edge busyness in successive still images appear in sequence as a shimmering blur of dots around edges

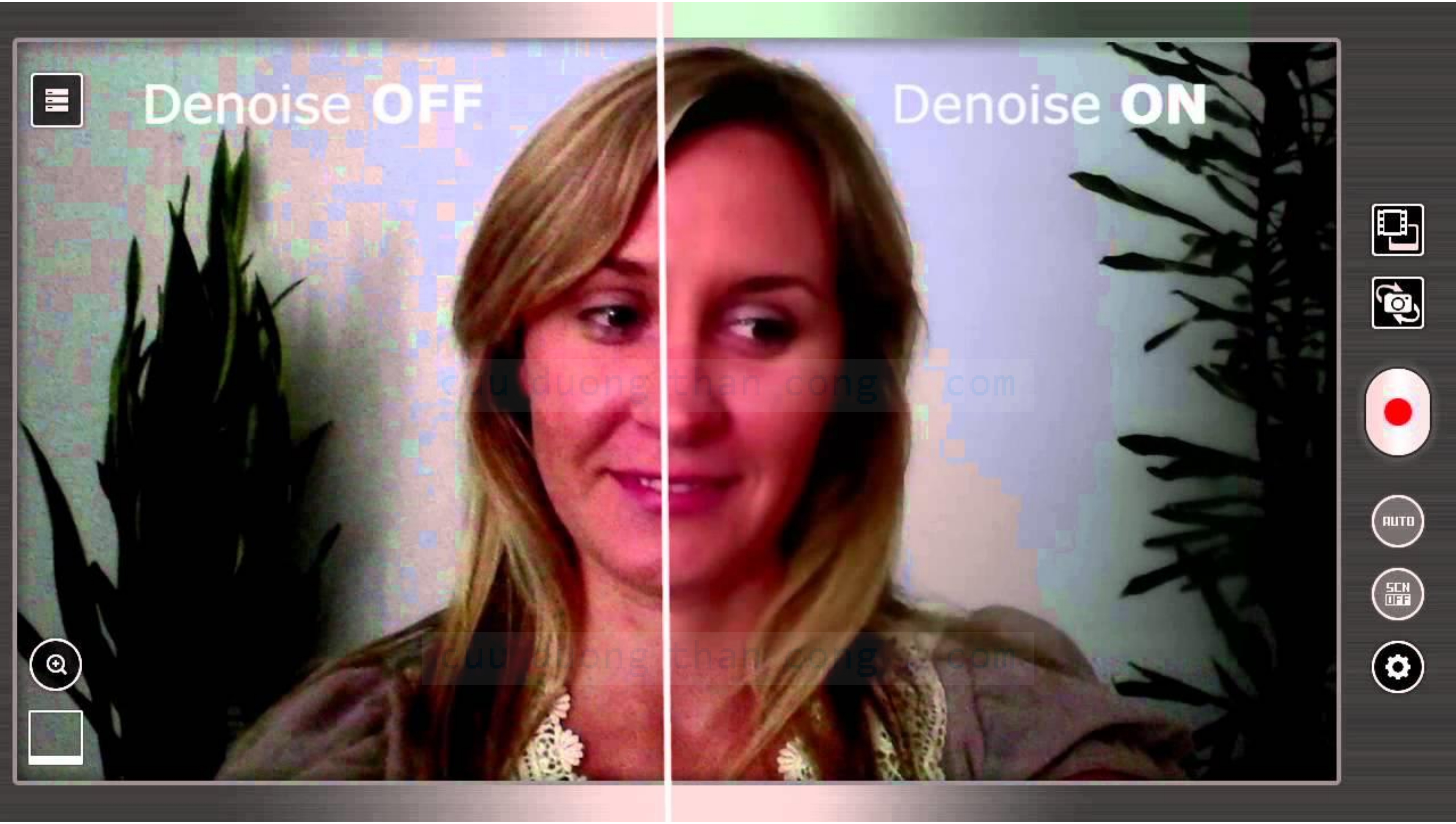
Example of noise reduction



Video denoising

- **Video denoising** is the process of removing noise from a video signal.
  - **Chroma noise** is where one see color fluctuations while **luminance noise** is where one see light/dark fluctuations.
  - Generally, the luminance noise looks more like film grain while chroma noise looks more unnatural or digital like.
- **Spatial denoising:** apply image noise reduction to individual frames.
- **Temporal denoising:** reduce noise between frames
  - Motion compensation may be used to avoid ghosting artifacts when blending together pixels from several frames.
- **Spatial-temporal (or 3D) denoising:** a combination of both

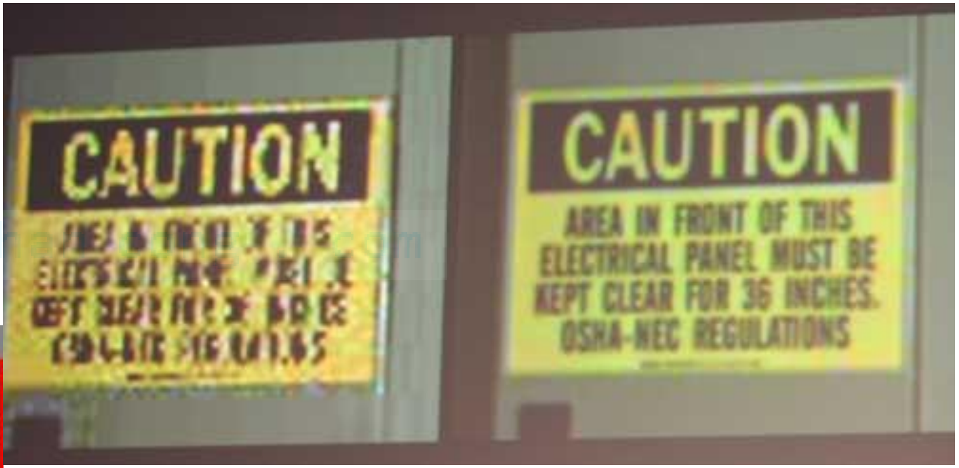
Example of video denoising

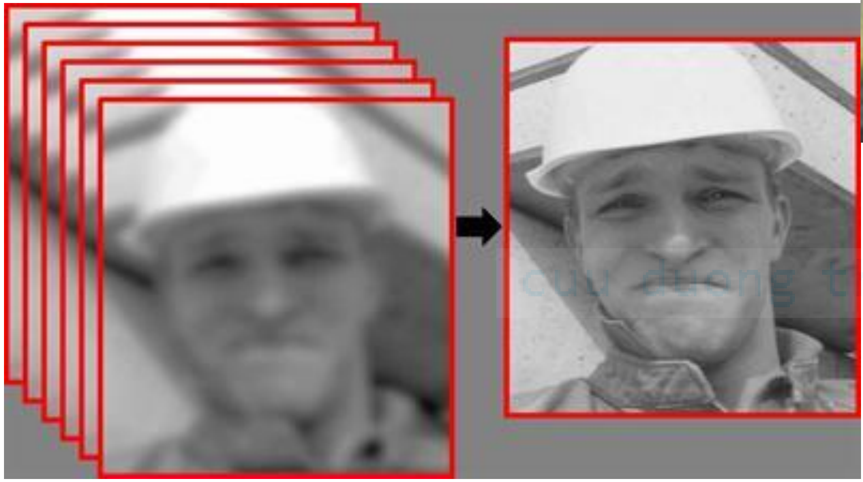


https://www.youtube.com/watch?v=sGc9qDjU9AQ

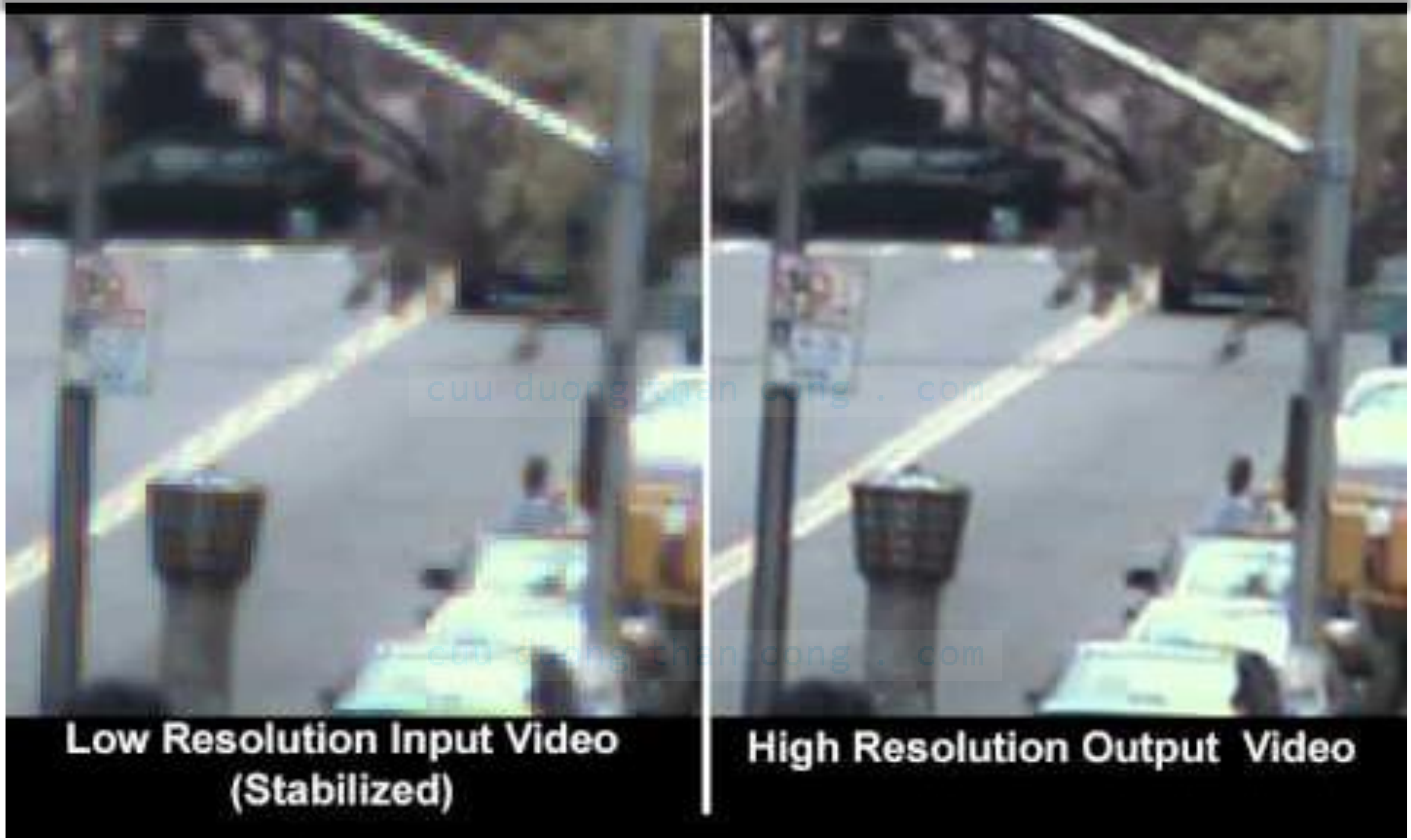
Video super-resolution

- **Video super-resolution** aims for exploiting additionally the information from multiple low resolution images to provide higher resolution images.





Example of video super-resolution

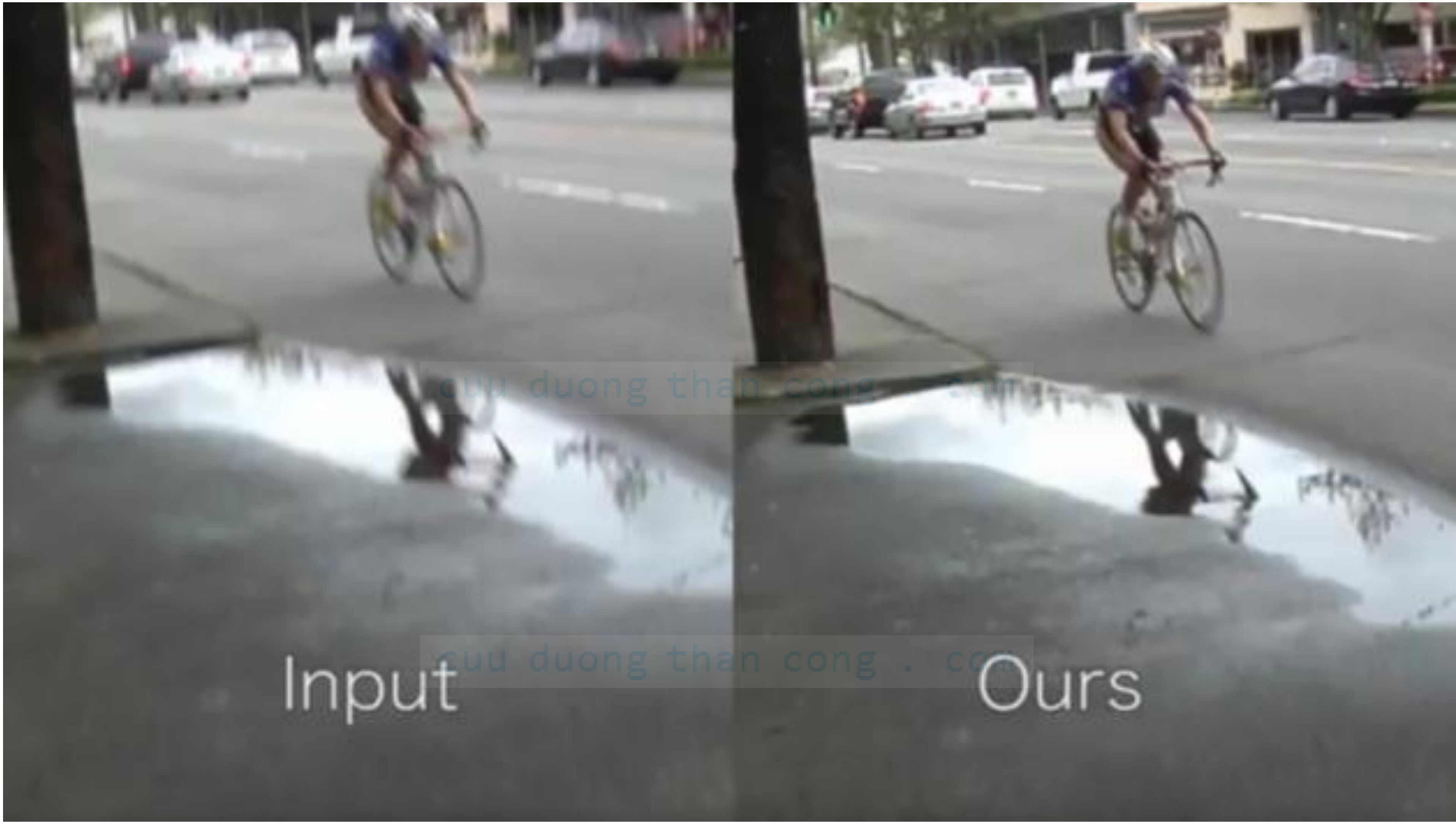


https://www.youtube.com/watch?v=QdK5-gNf4Wg

Video deblurring

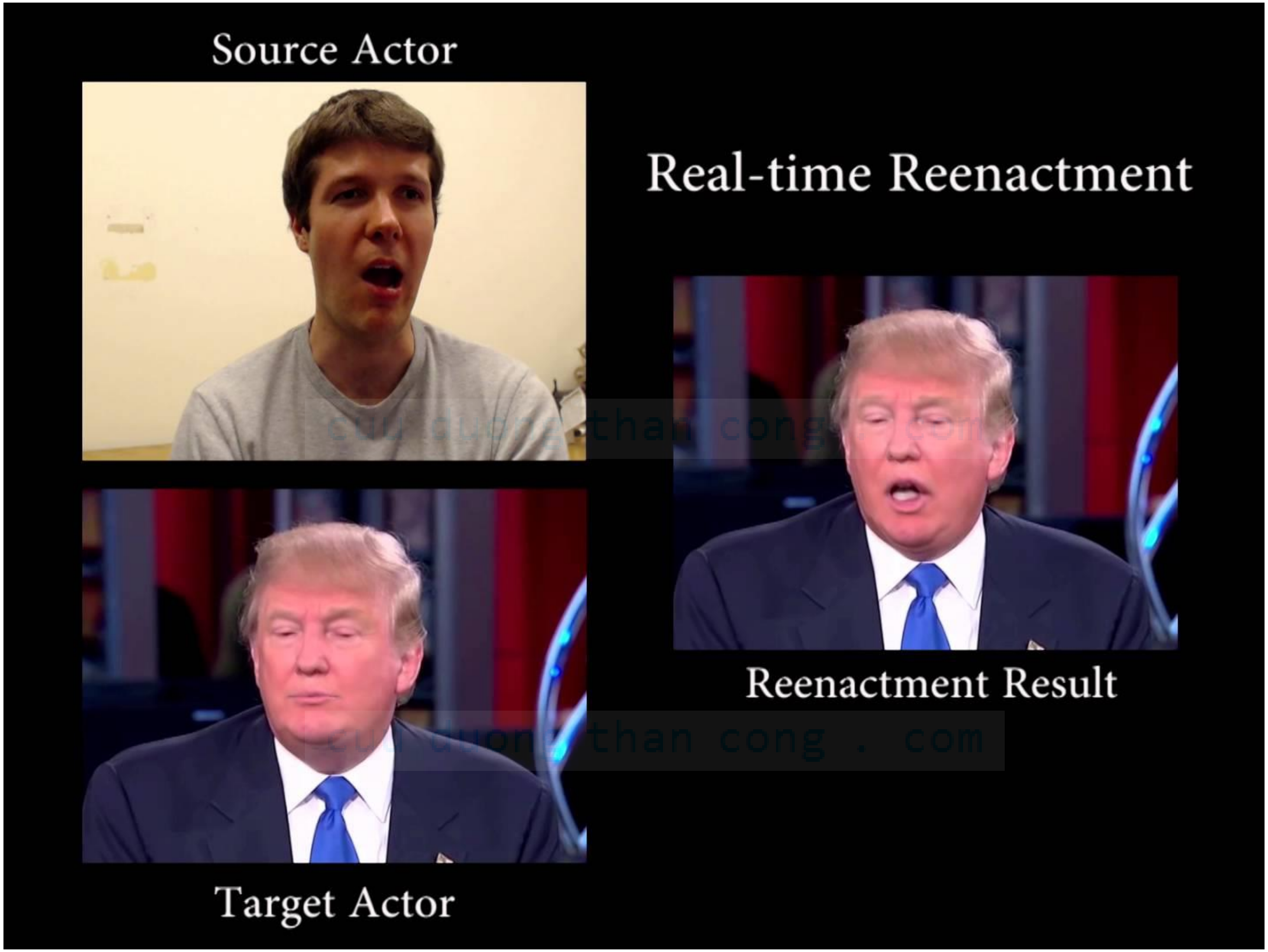
- Blurry frames often cause a flickering effect when viewed in real time, degrading the quality of visual perception.
- Camera motion within the capture of each individual frame leads to motion blur.
  - The blurring is often more pronounced after stabilization, due to inconsistencies with the modified stabilization-induced motion path
- Motion between frames yields inter-frame misalignment that can be exploited for blur removal.

Example of video deblurring



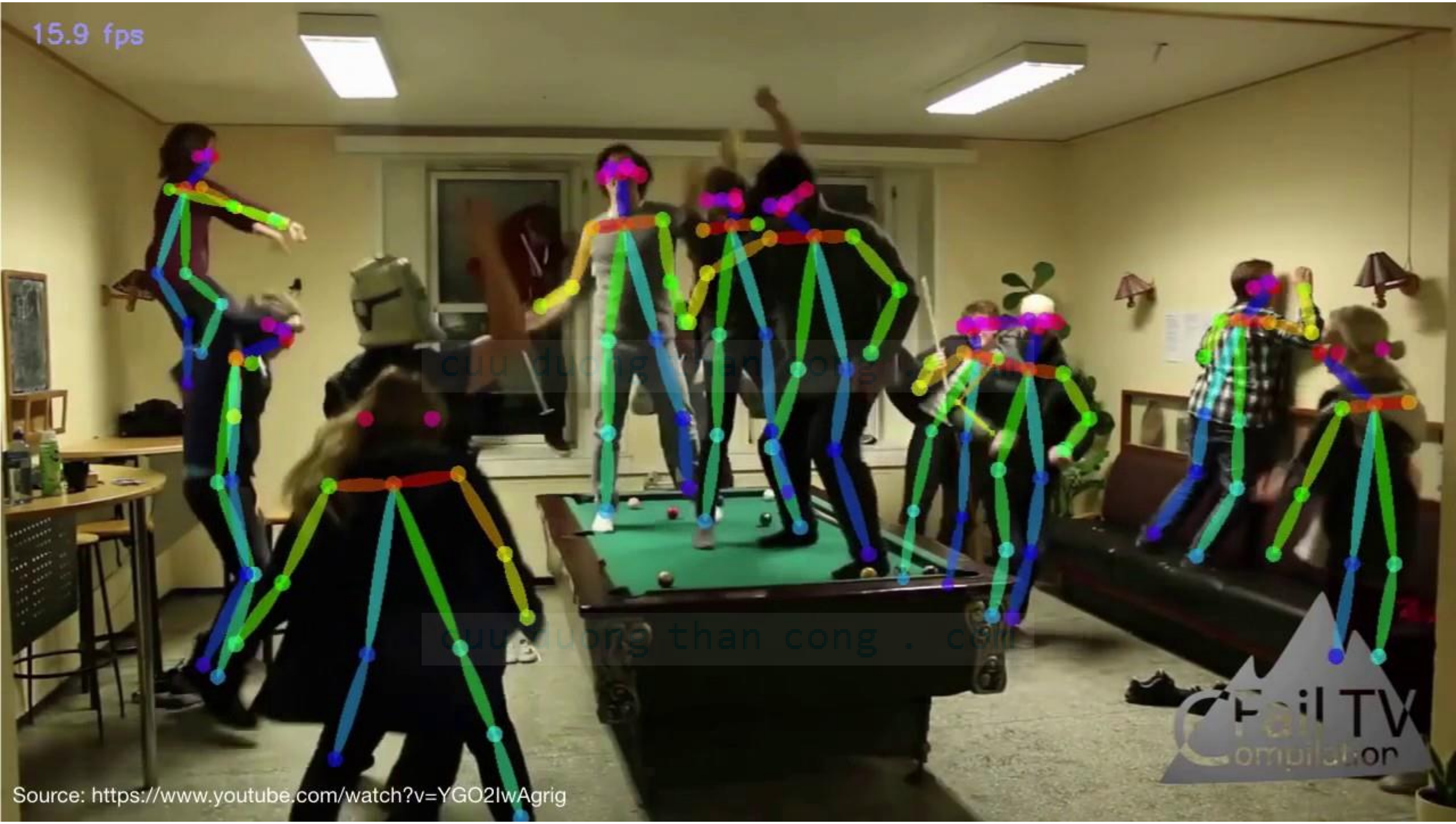
https://www.youtube.com/watch?v=NoqRMlbqgaQ&t=60s

Face capture and reenactment



Face2Face: Real-time Face Capture and Reenactment of RGB Videos (CVPR 2016)

Human pose estimation



Source: https://www.youtube.com/watch?v=YGO2IwAgrig

Realtime Multi-Person 2D Human Pose Estimation using Part Affinity Fields (CVPR 2017)

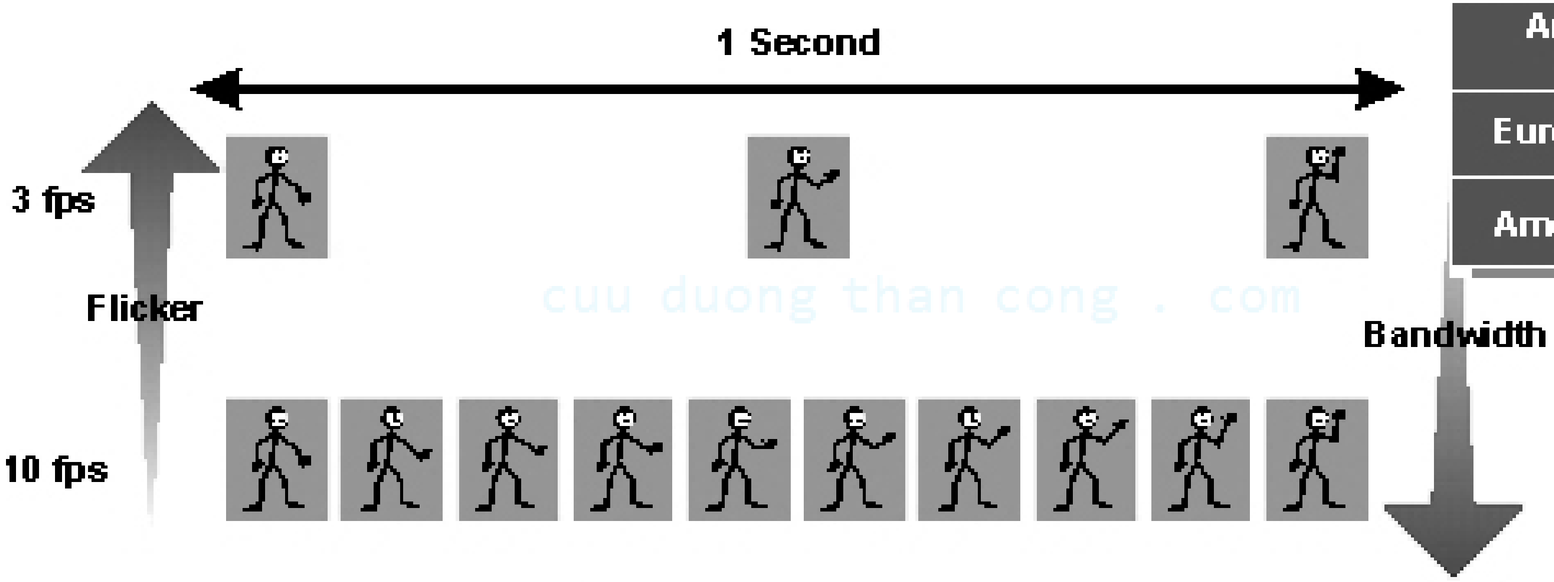
Section 9.2

VIDEO CODING

Video frame rate

- **Frame rate** (fps, in *hertz*) is the frequency (rate) at which consecutive frames are displayed in an animated display.
  - In practice, most video formats use temporal sampling rates of 24 frames per second and above.
  - E.g., NTSC video has a frame rate of 30 frames/sec.

Medium	fps
Film	24
European Video	25
American Video	30
European TV	50
American TV	60

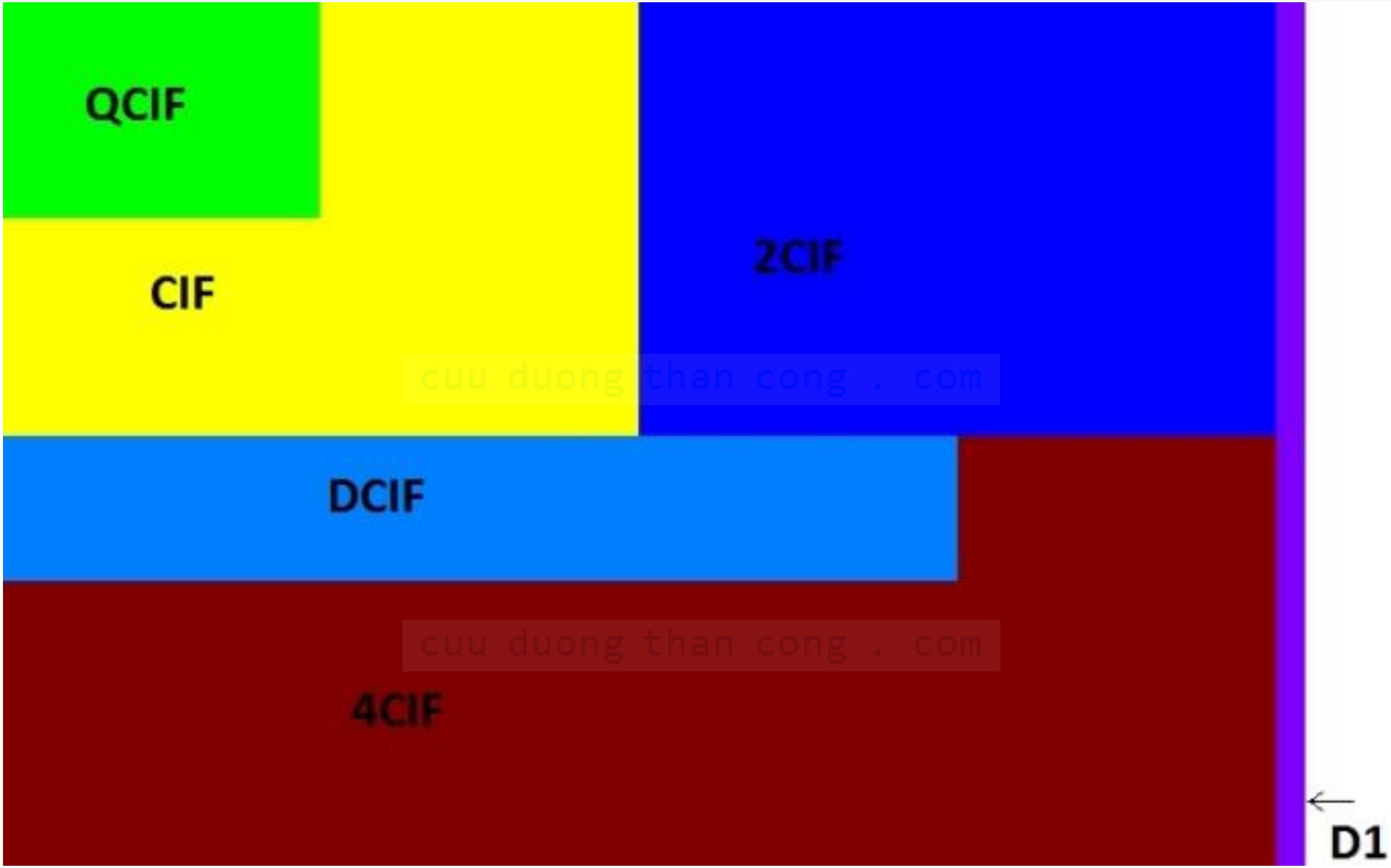


Common Intermediate Format

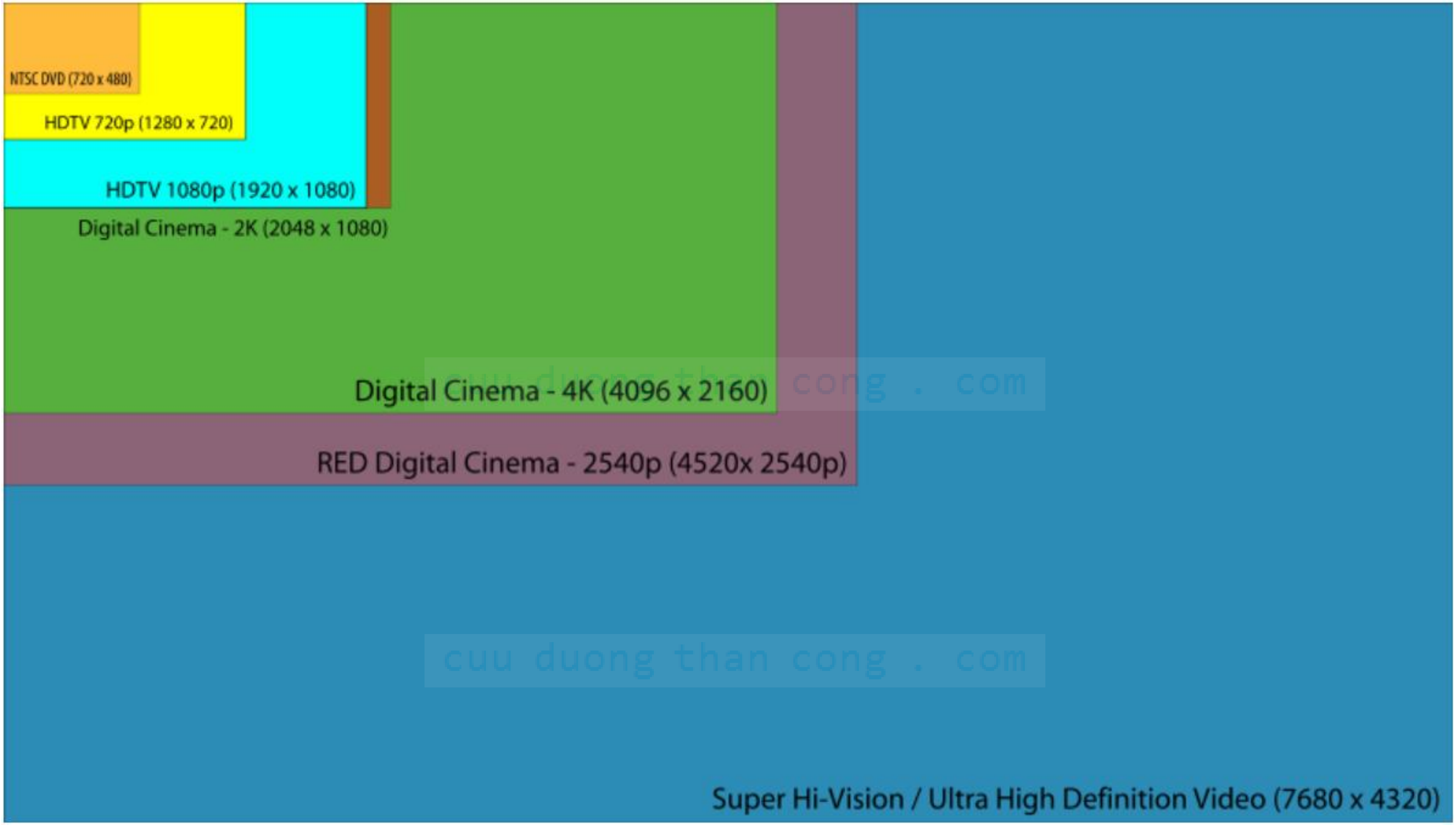
- The **Common Intermediate Format** (CIF) is used to standardize the horizontal and vertical resolutions in pixels of YC_bC_r sequences in video signals
- It is commonly used in video teleconferencing systems.

Format	**Luminance pixel resolution**	**Typical Applications**
Sub-QCIF	128 × 96	Mobile Multimedia
QCIF	176 × 144	Video conferencing and Mobile Multimedia
CIF	352 × 288	Video conferencing
4CIF	704 × 576	SDTV and DVD-Video
16CIF	1408 × 1152	HDTV and DVD-Video

Common Intermediate Format



HD video formats in pixel resolution


NTSC DVD (720 x 480)
HDTV 720p (1280 x 720)
HDTV 1080p (1920 x 1080)
Digital Cinema - 2K (2048 x 1080)
Digital Cinema - 4K (4096 x 2160)
RED Digital Cinema - 2540p (4520x 2540p)
Super Hi-Vision / Ultra High Definition Video (7680 x 4320)

* HD = high definition

Color space

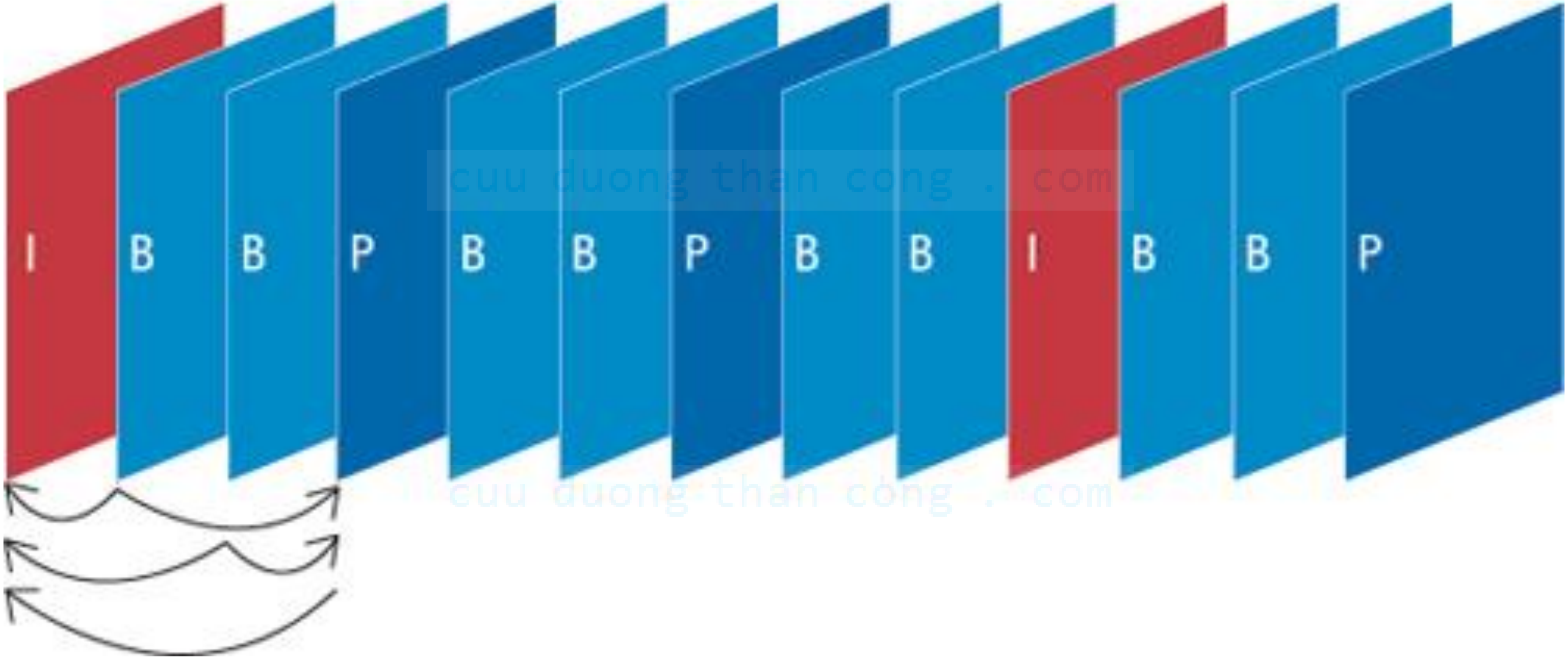
- Each pixel is represented by three components: the **luminance component $\boldsymbol{Y}$**, and the two **chrominance components $\boldsymbol{C_b}$ and $\boldsymbol{C_r}$**.
- **RGB to YC_bC_r conversion**

$$[Y \quad C_b \quad C_r] = [R \quad G \quad B]\begin{bmatrix} 0.299 & -0.168935 & 0.499813 \\ 0.587 & -0.331665 & -0.418531 \\ 0.114 & 0.50059 & -0.081282 \end{bmatrix}$$

- **Video quality** is commonly evaluated by using PSNR in the Y channel, which is referred to as the Y-PSNR (dB).
  - Peak signal-to-noise ratio: the maximum possible power of a signal over the power of corrupting noise that affects the fidelity of its representation

Video frame types

- There are three types of video frames are **I-frame**, **P-frame** and **B-frame**.

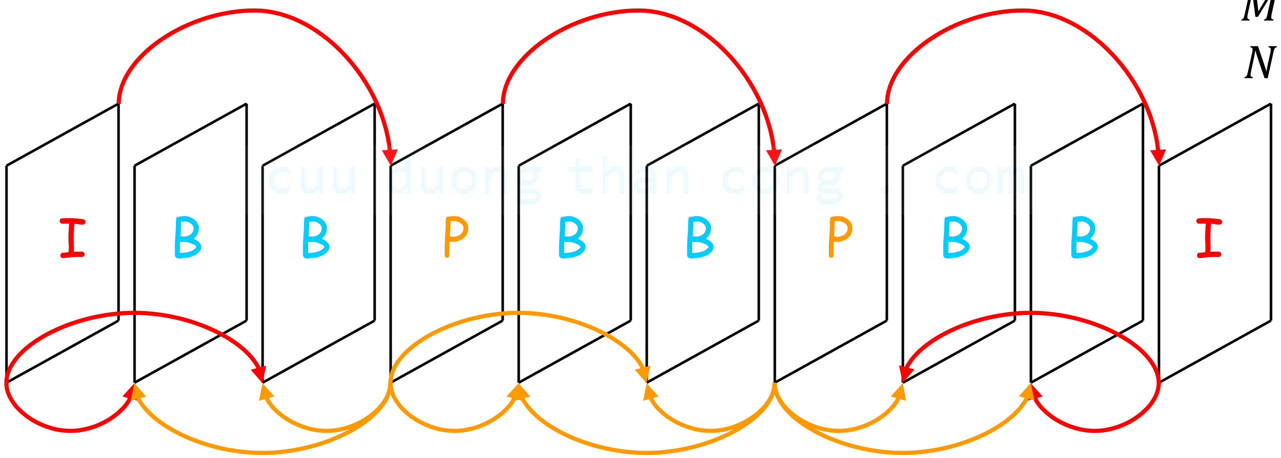


Video frame types

- **I-frames** (intra-coded frame) are encoded **without any motion compensation** and are used as a reference for future predicted **P** and **B** type frames.
  - **I** frames require a relatively large number of bits for encoding
- **P-frames** (predictive frame) are encoded **using motion compensated prediction** from a reference frame which can be either **I** or **P** frame.
  - **P** frames are more efficient in terms of number of bits required compared to **I** frames, but still require more bits than **B** frames.
- **B-frames** (bidirectional predictive frame) require the lowest number of bits compared to both **I** and **P** frames but incur computational complexity.

Group of pictures (GOP)

- The **Group of pictures** (GOP) includes successive pictures within a coded video stream, starting with an **I** frame.
  - Encountering a new GOP means that the decoder does not need any previous frames to decode the next ones → fast seeking.
- The GOP is often referred by two numbers, M and N.
  - M: the distance between two nearest **P** frames or **P** and **I** frame
  - N: the distance between two nearest **I** frames, called **GOP size**

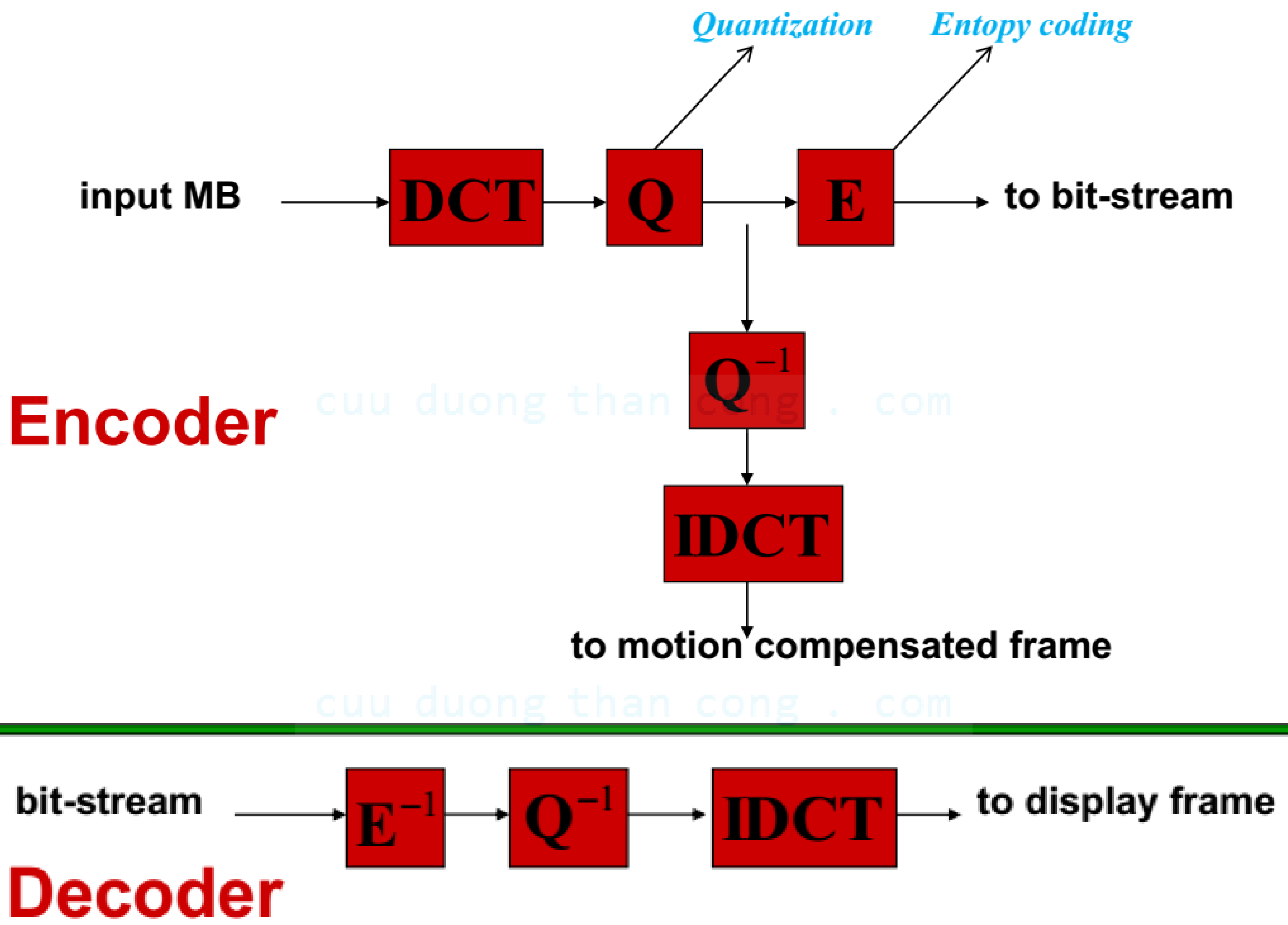


$$M = 3$$
$$N = 9$$

Intra-frame coding

- **Intra-frame coding** removes the spatial redundancy within a frame by using transform (commonly used is DCT).
- **I**-coding
  - **MB (Macro Block)** is encoded as is, without motion compensation.
  - DCT followed by Q (Quantization), zig-zag, run-length, Huffman Coding.

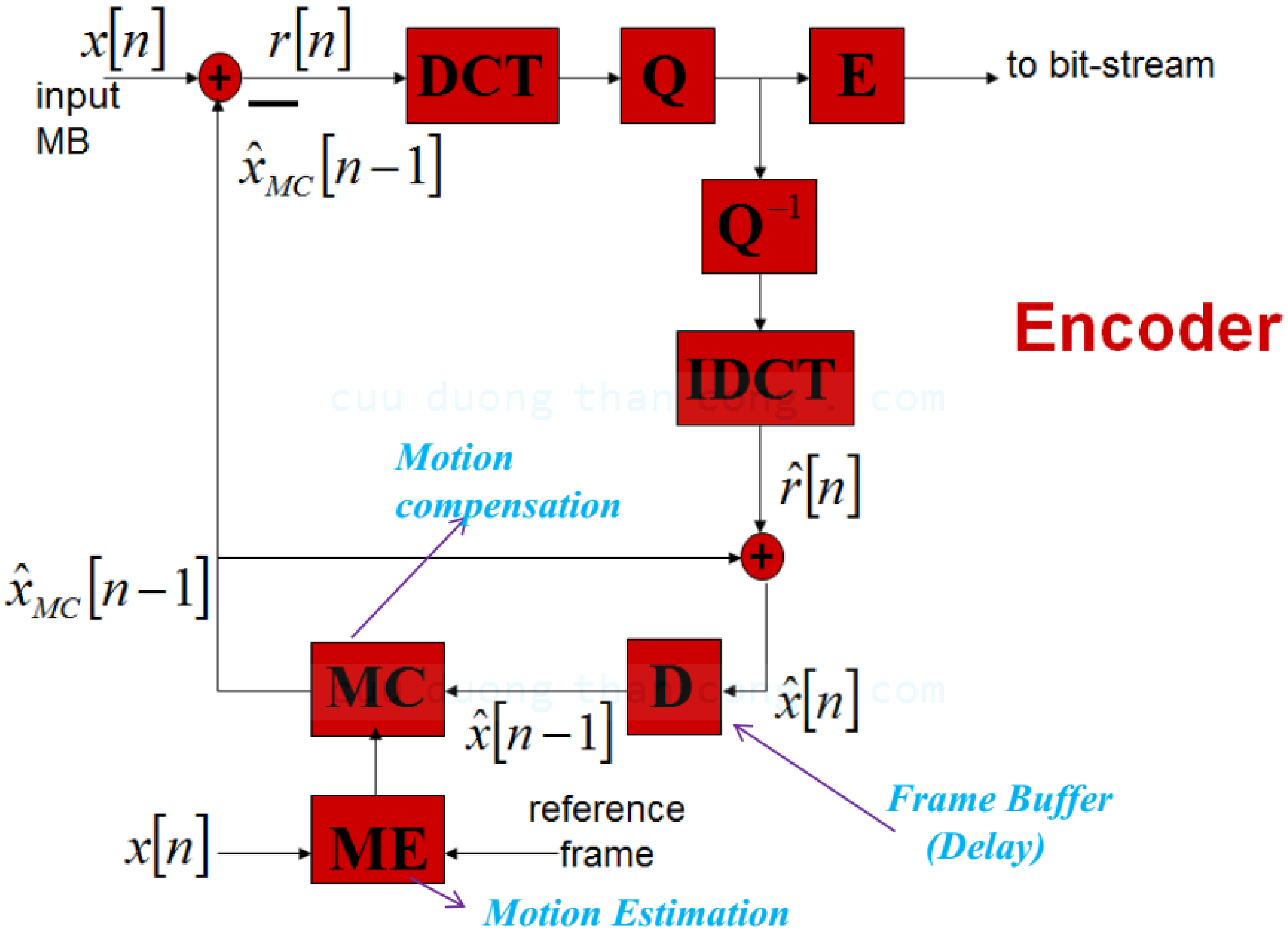
Intra-frame coding



Inter-frame coding

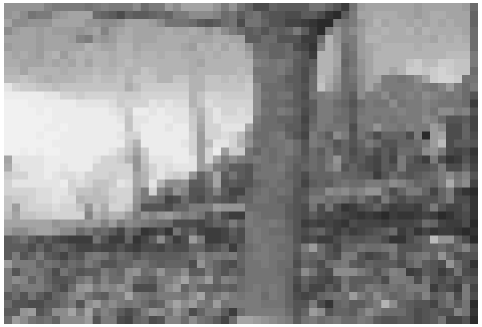
- **Inter-frame coding** removes the temporal redundancy between successive frames by exploiting the inter-dependencies of video frames.
  - It relies on the fact that adjacent pictures in a video sequence have high temporal correlation
- **Inter (P**- and **B**-coding**)**
  - **Block-matching – motion estimation**
  - Predictive motion residue from best-match block is DCT encoded (similarly to intra mode)
  - Motion vector is differentially encoded

Inter-frame coding



Video sequence and picture

- **Intra Picture (I**-Picture)
  - Encoded without referencing others
  - All MBs are intra coded
- **Inter Picture (P**-Picture, **B**-Picture)
  - Encoded by referencing other pictures
  - Some MBs are intra coded, and some are inter coded



Intra 0

Inter 1

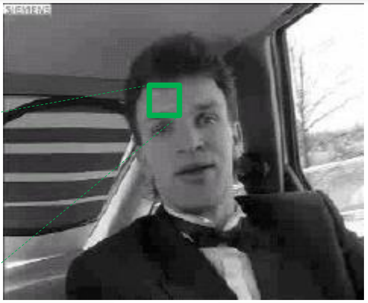
Inter 2

Inter 3

Inter 4

Inter 5

Coding of I slice



139	144	149	153	155	155	155	155
144	151	153	156	159	156	156	156
150	155	160	163	158	156	156	156
159	161	162	160	160	159	159	159
159	160	161	162	162	155	155	155
161	161	161	161	160	157	157	157
162	162	161	163	162	157	157	157
162	162	161	161	163	158	158	158

Original block

DCT

235.6	-1.0	-12.1	-5.2	2.1	-1.7	-2.7	1.3
-22.6	-17.5	-6.2	-3.2	-2.9	-0.1	0.4	-1.2
-10.9	-9.3	-1.6	1.5	0.2	-0.9	-0.6	-0.1
-7.1	-1.9	0.2	1.5	0.9	-0.1	0.0	0.3
-0.6	-0.8	1.5	1.6	-0.1	-0.7	0.6	1.3
1.8	-0.2	1.6	-0.3	-0.8	1.5	1.0	-1.0
-1.3	-0.4	-0.3	-1.5	-0.5	1.7	1.1	-0.8
-2.6	1.6	-3.8	-1.8	1.9	1.2	-0.6	-0.4

Transformed block

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

Quantization matrix

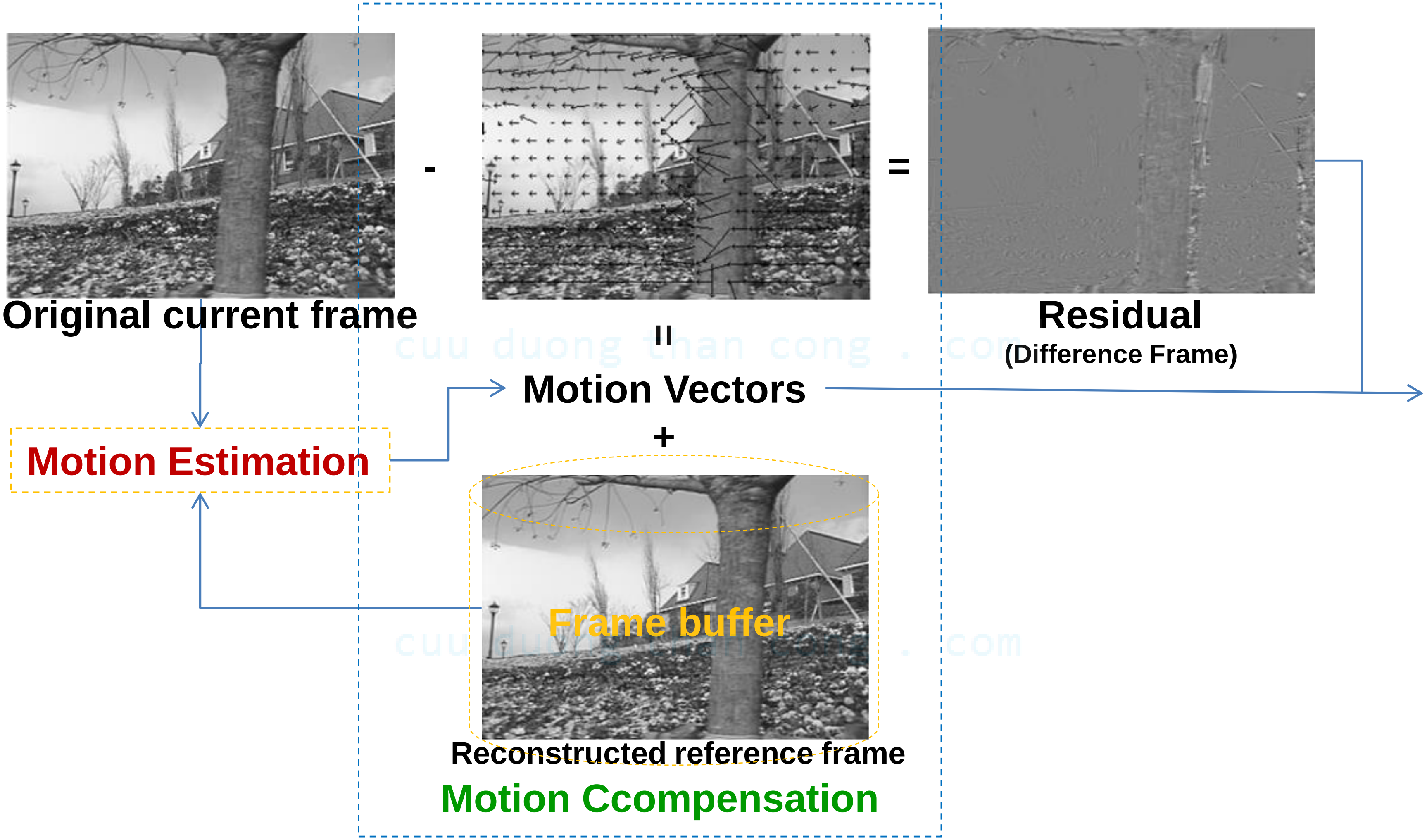
15	0	-1	0	0	0	0	0
-2	-1	0	0	0	0	0	0
-1	-1	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Bit-stream ← **15 0 -2 -1 -1 -1 0 …** ←

Entropy coding

Zig-zag scan

Coding of P-slice



Original current frame

-

=

Residual

(Difference Frame)

=

Motion Vectors

+

Motion Estimation

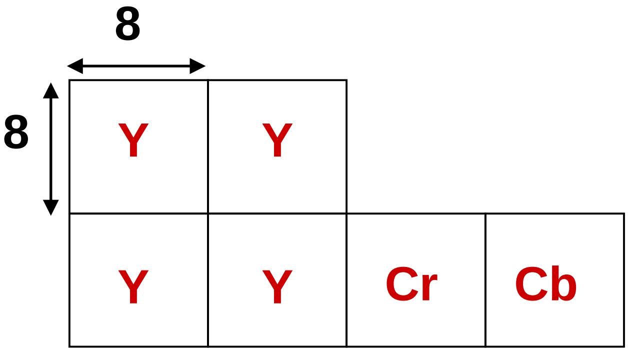
Frame buffer

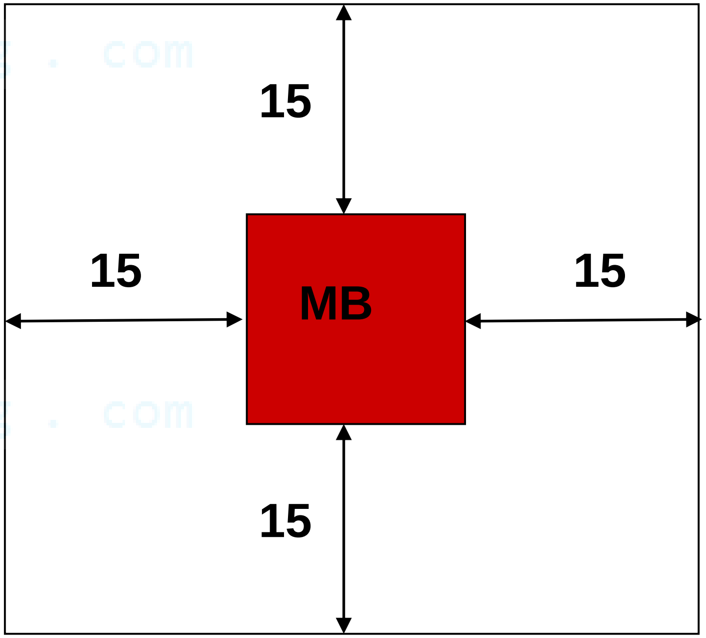
Reconstructed reference frame

Motion Ccompensation

Motion estimation in H.261 standard

- Macro-block
  - Luminance: 16×16, four 8×8 blocks
  - Chrominance: two 8x8 blocks
- Motion estimation only performed for luminance component
- Motion vector range $\in [-15, 15]$





Search Area in Reference Frame

Coding of motion vectors

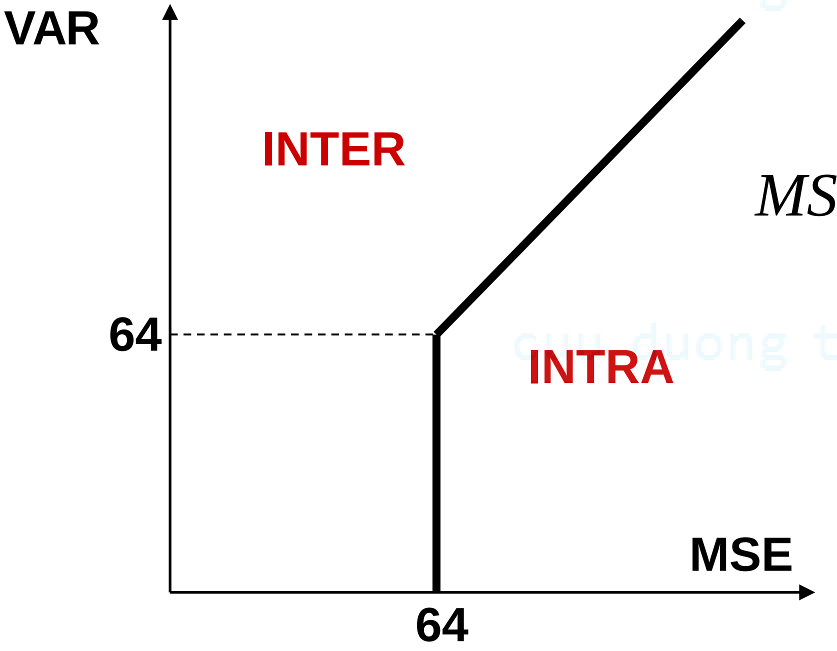
- Integer pixel motion estimation search only
- Motion vectors are differentially and separably encoded

$$MVD_x = MV_x[n] - MV_x[n-1]$$
$$MVD_y = MV_y[n] - MV_y[n-1]$$

- 11-bit VLC (Variable Length Coding) for MVD
- For example
  - MV = 2 2 3 5 3 1 -1…
  - MVD = 0 1 2 -2 -2 -2…
  - Binary: 1 010 0010 0011 0011 0011…

Inter/Intra switching

- It is based on energy of prediction error
- Intra mode is used for high energy while inter mode is used for low energy.
  - High energy: scene change, occlusions, uncovered areas…
  - Low energy: stationary background, translational motion …



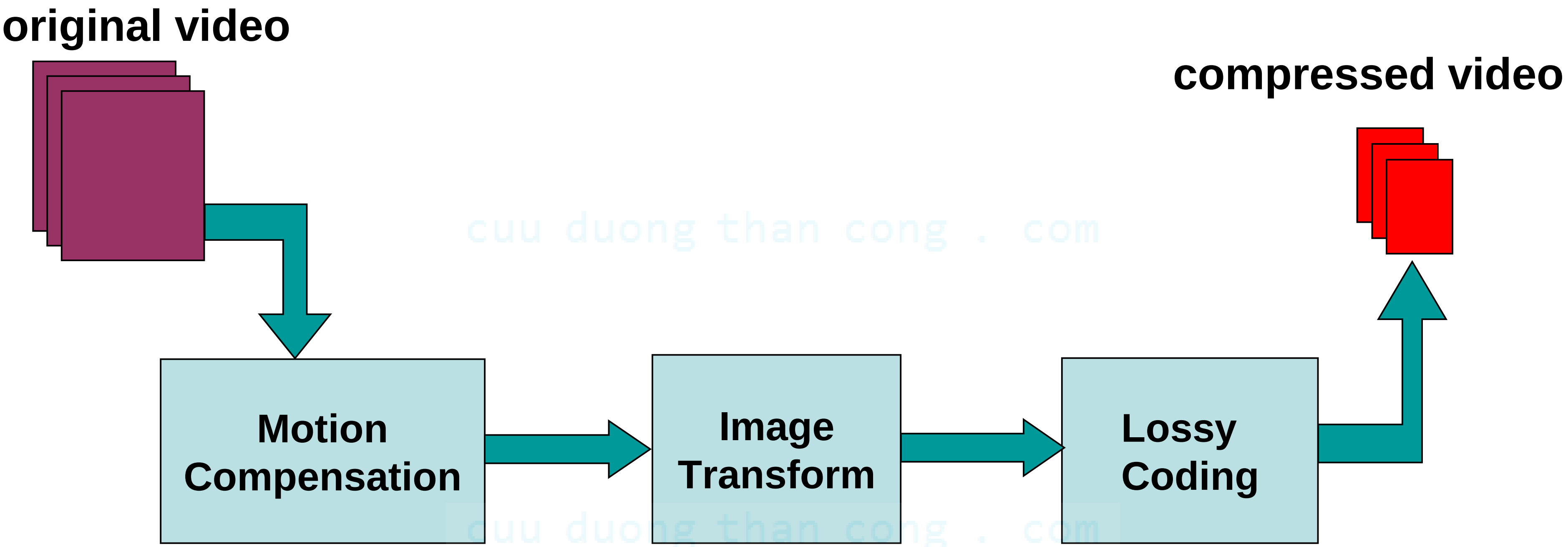
$$VAR = \frac{1}{256}\sum_{MB}\left(c[x,y]-\bar{c}\right)^2$$

$$MSE = \frac{1}{256}\sum_{MB}\left(c[x,y]-r[x+dx,y+dy]\right)^2$$

H.263 standard

- Standardization effort started Nov 1993
- Aim:
  - Low bit-rate video communications, less than 64 kbps
  - Target PSTN and mobile network: 10-32 kbps
- Developed as an evolutionary improvement based on experience from H.261, MPEG-1 and MPEG-2 standards.
- Inherited by H.264 (also known as MPEG-4 part 10).
- Main properties
  - H.261 with many MPEG features optimized for low bit rates
  - Performance: 3-4 dB improvements over H.261 at less than 64 kbps; 30% bit rate saving over MPEG-1.

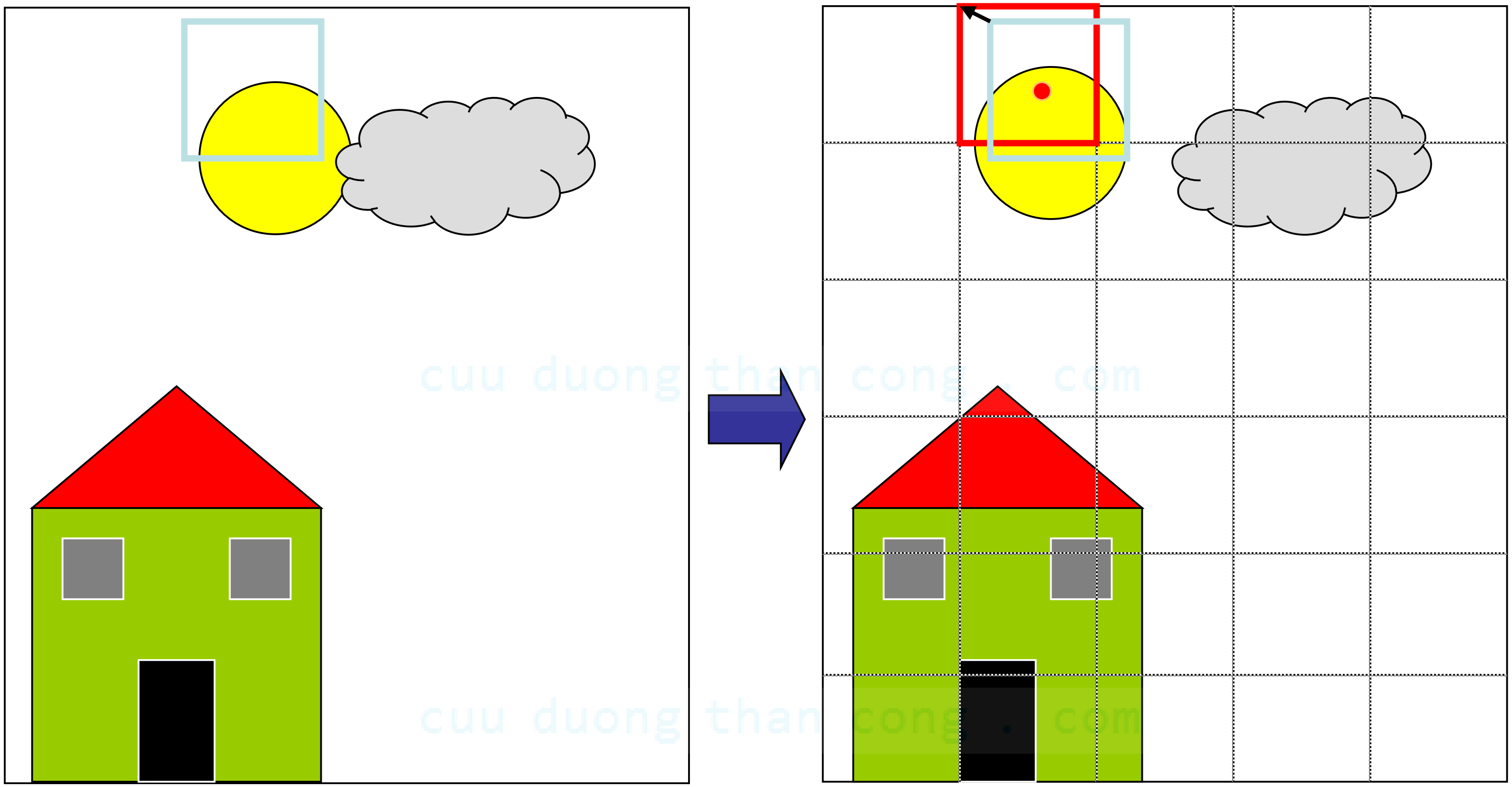
H.263 standard coder



H.263 motion compensation

- Image is divided into 16×16 macroblocks,
- Each macro block is matched against nearby blocks in previous frame (called reference frame),
  - “Nearby” = within 15-pixel horizontal/vertical range
  - Half-pixel accuracy (with bilinear pixel interpolation)
- Best match is used to predict the macro block,
  - The relative displacement, or motion vector, is encoded and transmitted to decoder
- Prediction error for all blocks constitute the **residual**.

Example of motion compensation



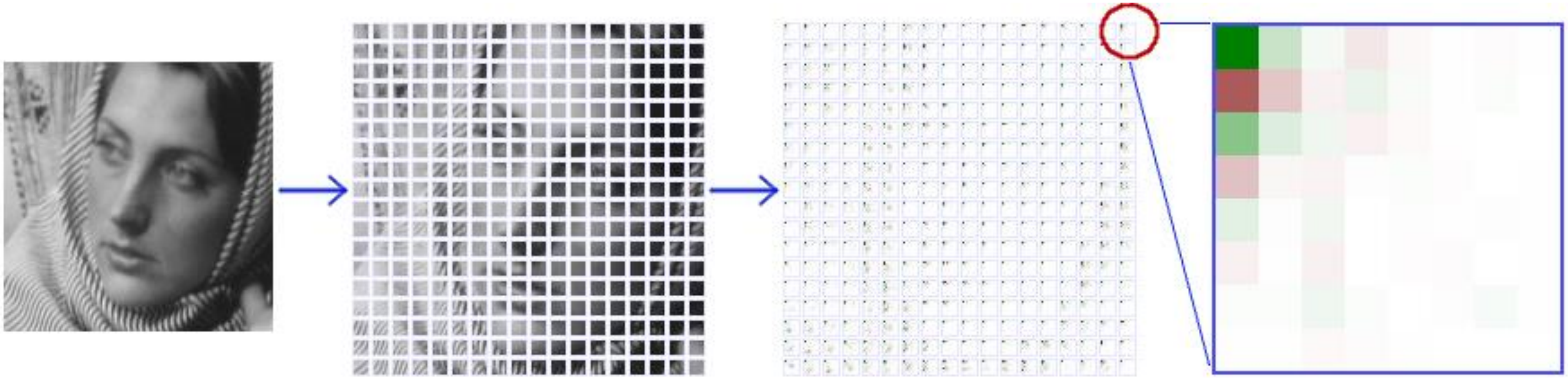
T=1 (reference)

T=2 (current)

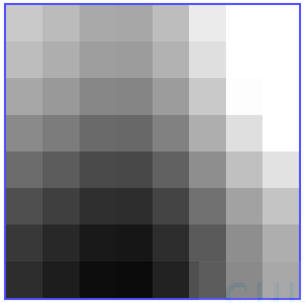
H.263 Image transform

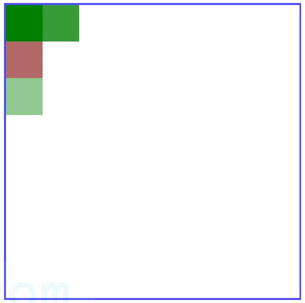
- Residual is divided into 8x8 blocks,
- 8 × 8 2-D Discrete Cosine Transform (DCT) is applied to each block independently
- DCT coefficients describe spatial frequencies in the block
  - High frequencies correspond to small features and texture
  - Low frequencies correspond to larger features
  - Lowest frequency coefficient, called DC, corresponds to the average intensity of the block

Example of DCT



Original image *Pixel blocks* *DCT coefficient blocks* *Single coefficient block*





DCT

Original pixel data

114	108	100	99	109	129	152	166
109	102	95	94	104	124	146	161
99	93	85	84	94	114	137	151
86	80	72	71	82	102	124	138
73	66	58	57	68	88	110	125
60	53	46	45	55	75	97	112
50	43	36	35	45	65	88	102
45	38	31	30	40	60	82	97

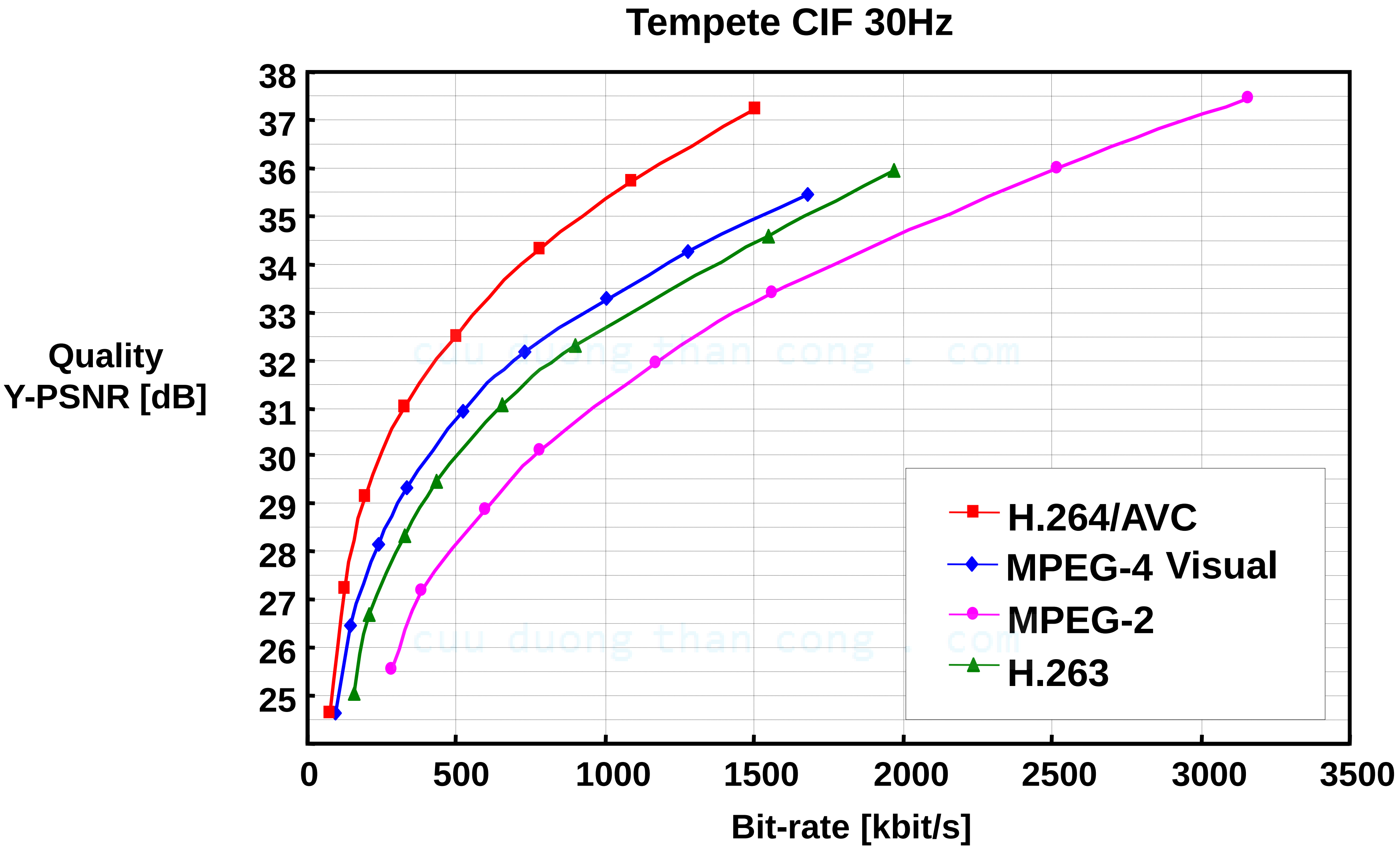
DCT coefficient data

700	200	0	0	0	0	0	0
-150	0	0	0	0	0	0	0
110	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

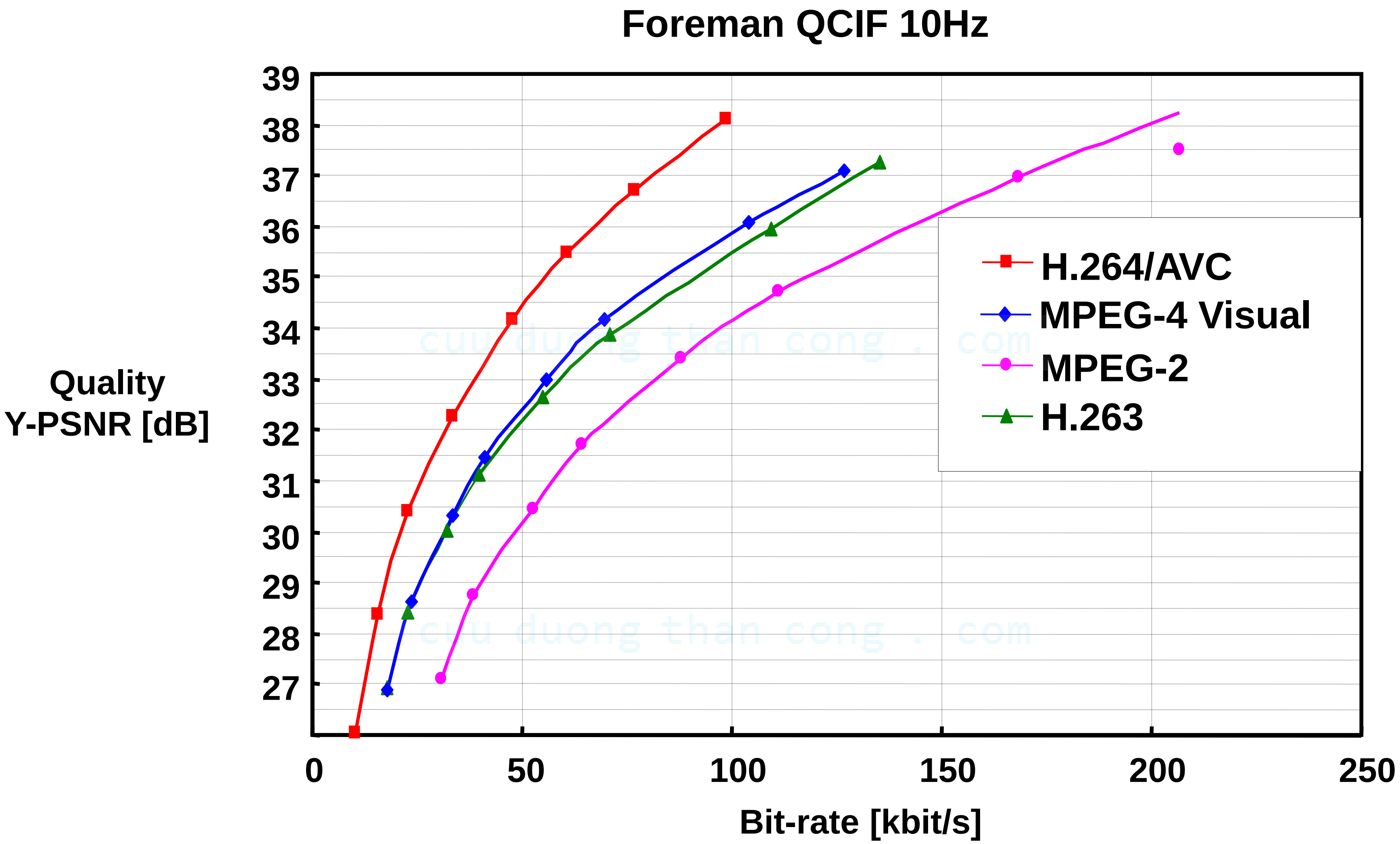
H.263 Lossy coding

- Transform coefficients are quantized
  - Some less-significant bits are dropped, remaining bits are encoded
- For inter-frames, all coefficients get the same number of bits, except for the DC which gets more.
- For intra-frames, lower-frequency coefficients get more bits
  - To preserve larger features better
- The actual number of bits used depends on the quantization parameter (QP), whose value depends on the bit-allocation policy
- Finally, bits are encoded using entropy (lossless) code, which is traditionally Huffman-style code

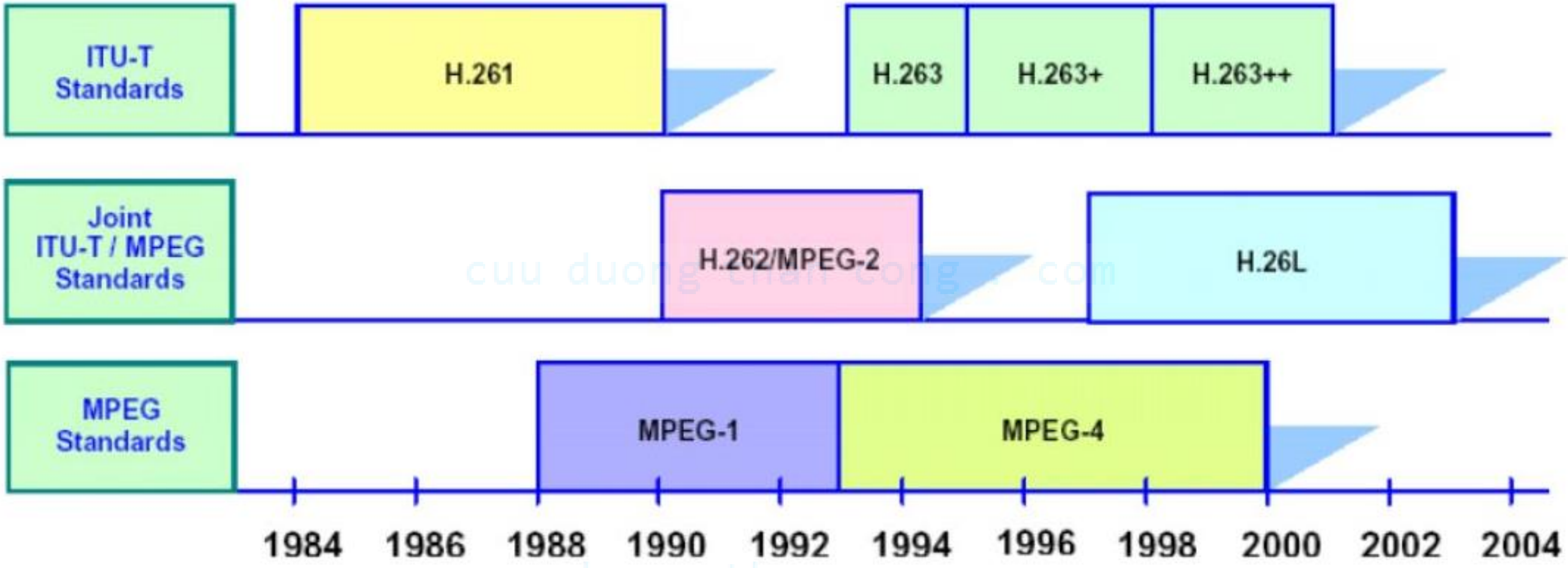
Comparison of video coding standards



Comparison of video coding standards



History of video coding standards



Video coding stands

ISO **(Int. Organization for Standardization)**

MPEG-1 (1992)
1.5Mbps, VCD

MPEG-2 (1996)
2-10Mbps, DVD

MPEG-4 (2000)
8-1024Kbps, videophone

Digital cinema (ongoing)

ITU **(Int. Telecommunication Union)**

H.261 (1990)
p ×64Kbps

H.263
8-64Kbps, videophone

H.263+/++
8-64Kbps, videophone

H.264/AVC

windows media player(Microsoft)

real player(Real-Networks)

Skype Video

References

- Rafael C. Gonzalez, Richard E. Woods, “Digital Image Processing”, 3rd edition, 2008. Chapter 6

- EE583 – Digital Image Processing, Dr.

http://faraday.ee.emu.edu.tr/ee583/Lectures/EE%20583-Lecture11.pdf