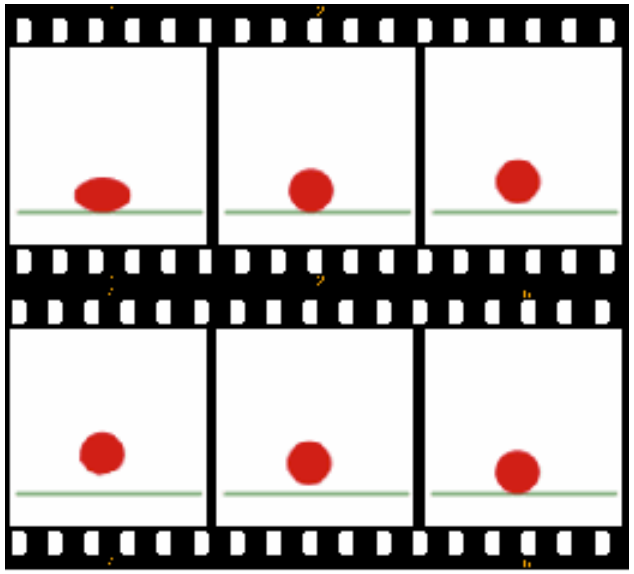
Using “animation” to motivate motion

- In computer generated animation, we take an object and mathematically render where it will be in the different frames



Courtesy: Wikipedia

- Given the rendered frames (or real life objects), we are trying to identify objects and their trajectories

10.3 Search for Motion Vectors

- Macroblock based (rather than pixel based or object based (MPEG-4). The goal is to find vector that maps block between reference and target frame
- The difference between two macroblocks measured by their *Mean Absolute Difference (MAD)*:

$$MAD(i, j) = \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \left| C(x+k, y+l) - R(x+i+k, y+j+l) \right|$$

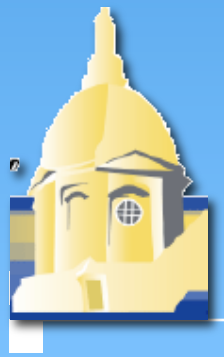
N — size of the macroblock,

k and l — indices for pixels in the macroblock,

i and j — horizontal and vertical displacements,

$C(x+k, y+l)$ — pixels in macroblock in Target frame,

$R(x+i+k, y+j+l)$ — pixels in macroblock in Reference frame.

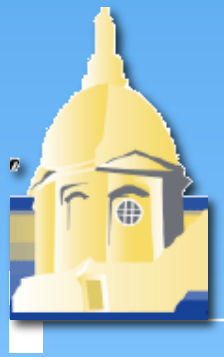


- The goal of the search is to find a vector (i, j) as the motion vector $\mathbf{MV} = (\mathbf{u}, \mathbf{v})$, such that $MAD(i, j)$ is minimum:

$$(u, v) = \left[(i, j) \mid MAD(i, j) \text{ is minimum}, \ i \in [-p, p], j \in [-p, p] \right]$$

Sequential Search

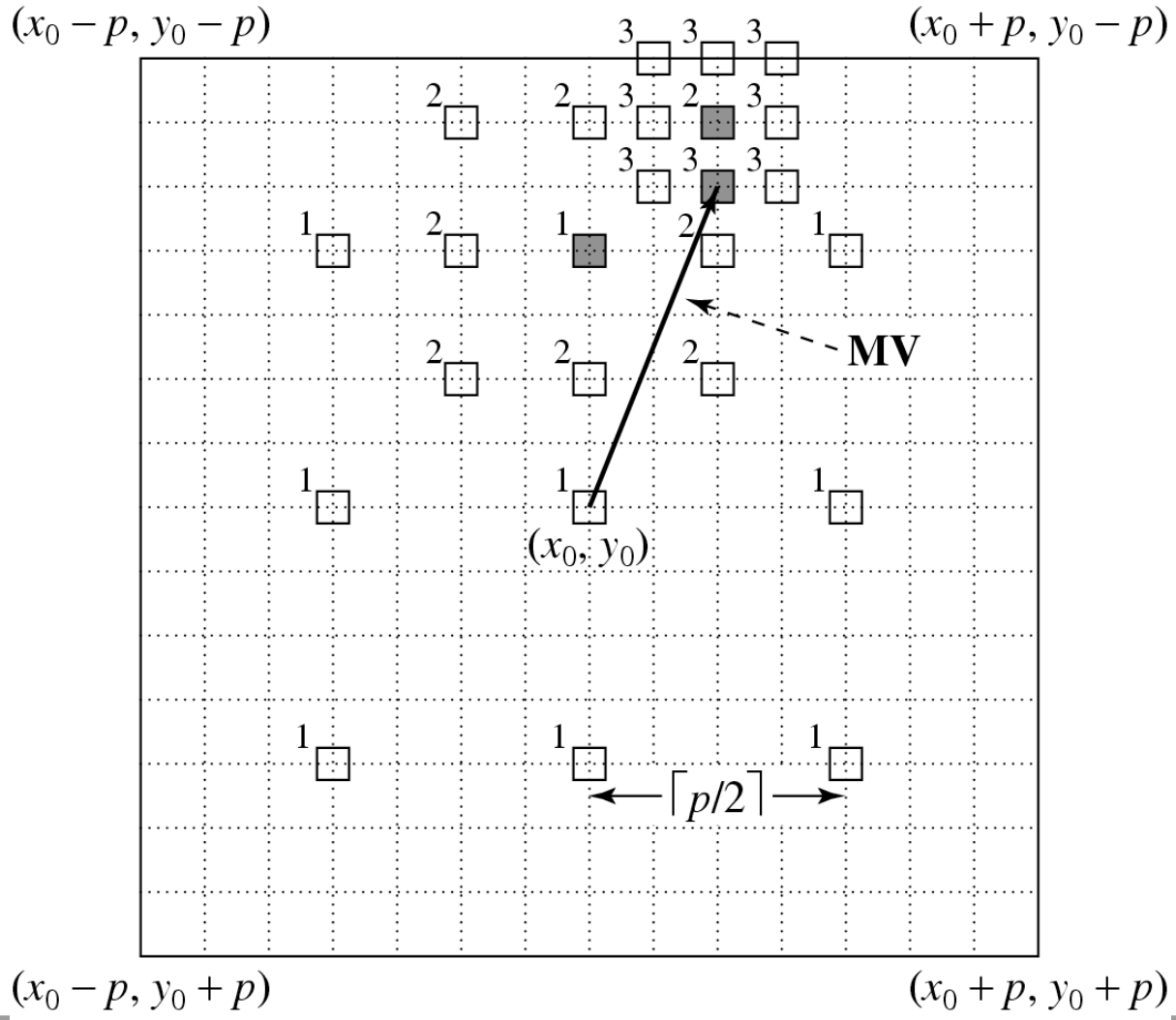
- **Sequential search**: sequentially search the whole $(2p + 1)$ x $(2 + 1)$ window in the Reference frame (referred to as Full search)
  - a macroblock centered at each of the positions within the window is compared to the macroblock in the Target frame pixel by pixel and their respective *MAD*
  - The vector (i, j) that offers the least MAD is designated as the **MV** (u, v) for the macroblock in the Target frame
  - sequential search method is very costly — assuming each pixel comparison requires three operations (subtraction, absolute value, addition), the cost for obtaining a motion vector for a single macroblock is $O(p^2 N^2)$



2D Logarithmic Search

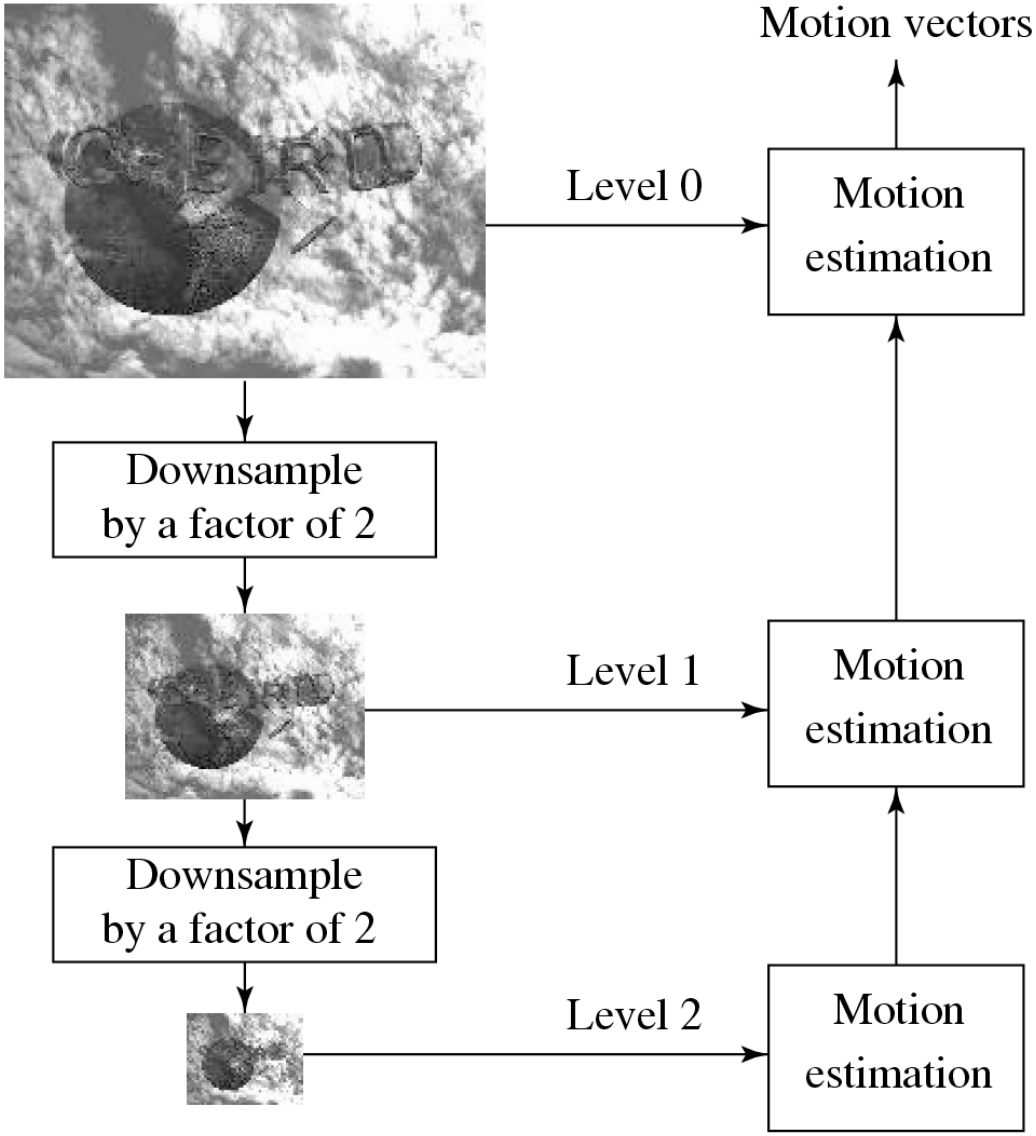
- Logarithmic search: a cheaper version, that is suboptimal but still usually effective
- The procedure for 2D Logarithmic Search of motion vectors takes several iterations and is akin to a binary search:
  - initially only nine locations in the search window are used as seeds for a MAD-based search; they are marked as ‘1’
  - - After the one that yields the minimum MAD is located, the center of the new search region is moved to it and the step-size (“offset”) is reduced to half
  - - In the next iteration, the nine new locations are marked as ‘2’ and so on

- 2D Logarithmic Search for Motion Vectors.



Hierarchical Search

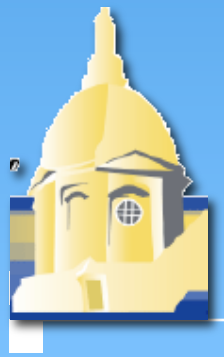
- The search can benefit from a hierarchical (multiresolution) approach in which initial estimation of the motion vector can be obtained from images with a significantly reduced resolution.
- a three-level hierarchical search in which the original image is at Level 0, images at Levels 1 and 2 are obtained by down-sampling from the previous levels by a factor of 2, and the initial search is conducted at Level 2
- Since the size of the macroblock is smaller and p can also be proportionally reduced, the number of operations required is greatly reduced


Motion vectors
Level 0
Motion estimation
Downsample by a factor of 2
Level 1
Motion estimation
Downsample by a factor of 2
Level 2
Motion estimation

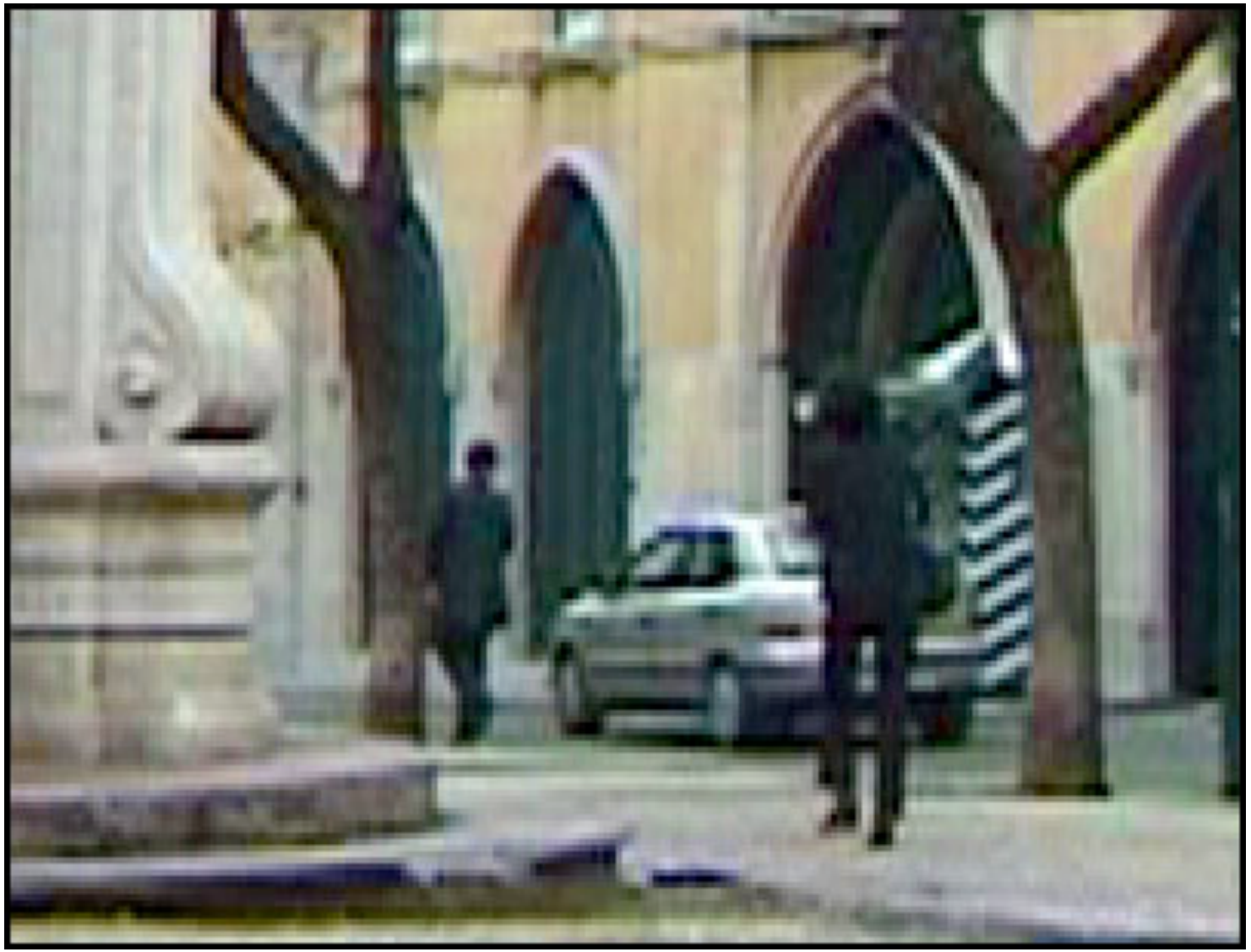
Cost of Motion Vector Search

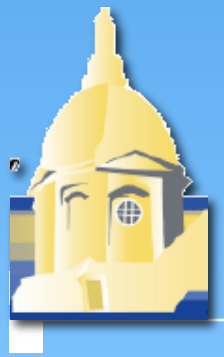
Search Method	OPS_per_second for 720×480 at 30 fps	
	$p = 15$	$p = 7$
Sequential search	29.89×10^9	7.00×10^9
2D Logarithmic search	1.25×10^9	0.78×10^9
3-level Hierarchical search	0.51×10^9	0.40×10^9

http://dvd-hq.info/

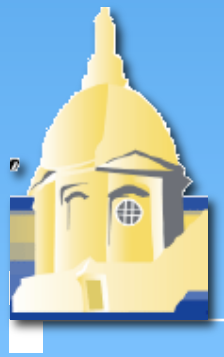


Frame 2



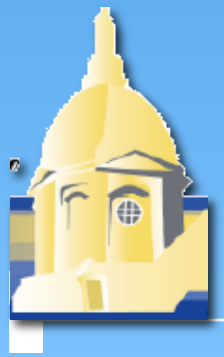


Macro blocks

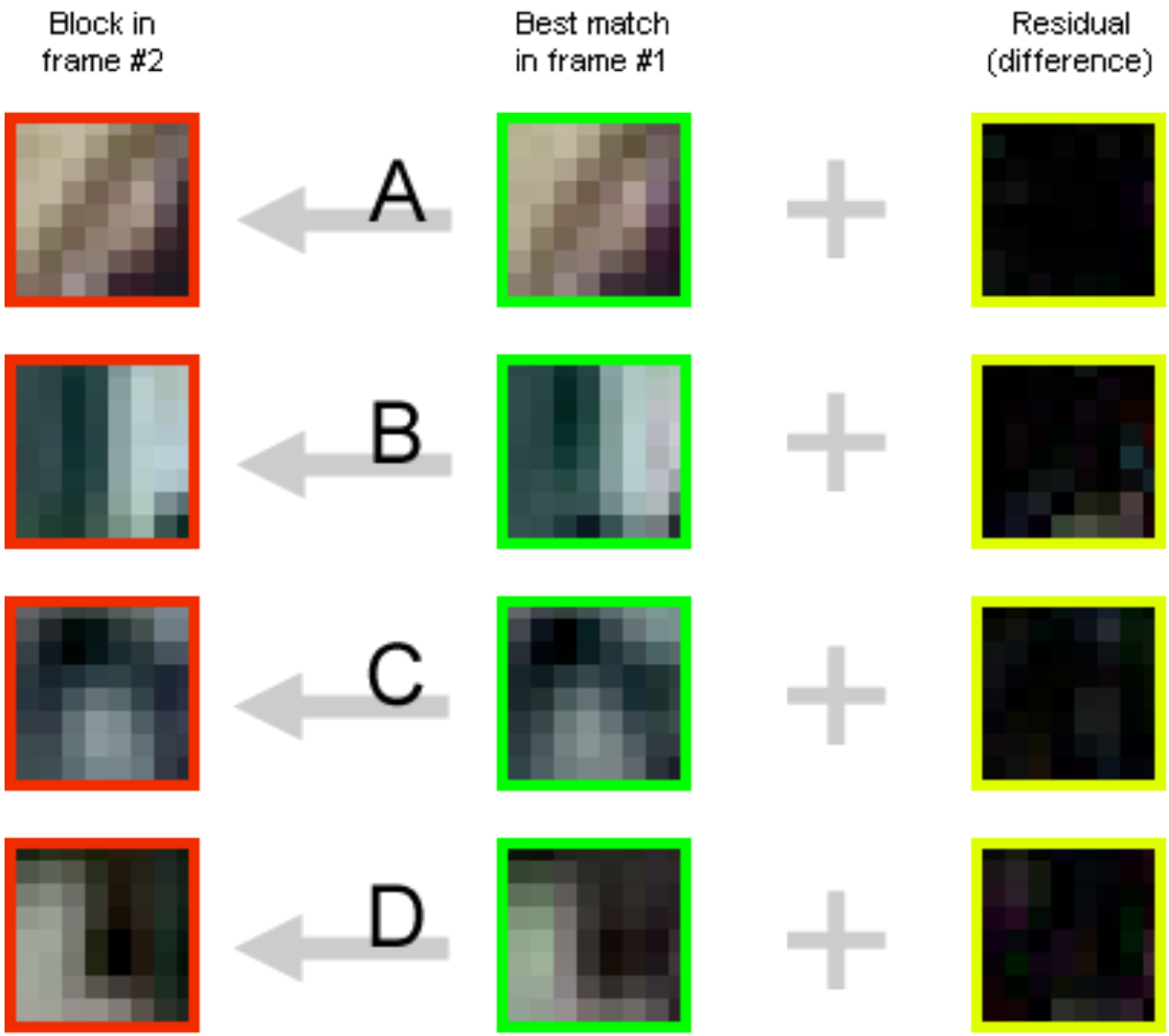


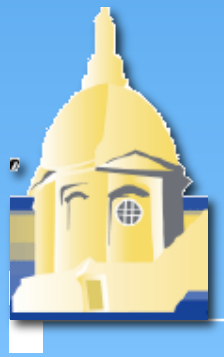
Focusing on blocks A B C & D

Best match in reference frame

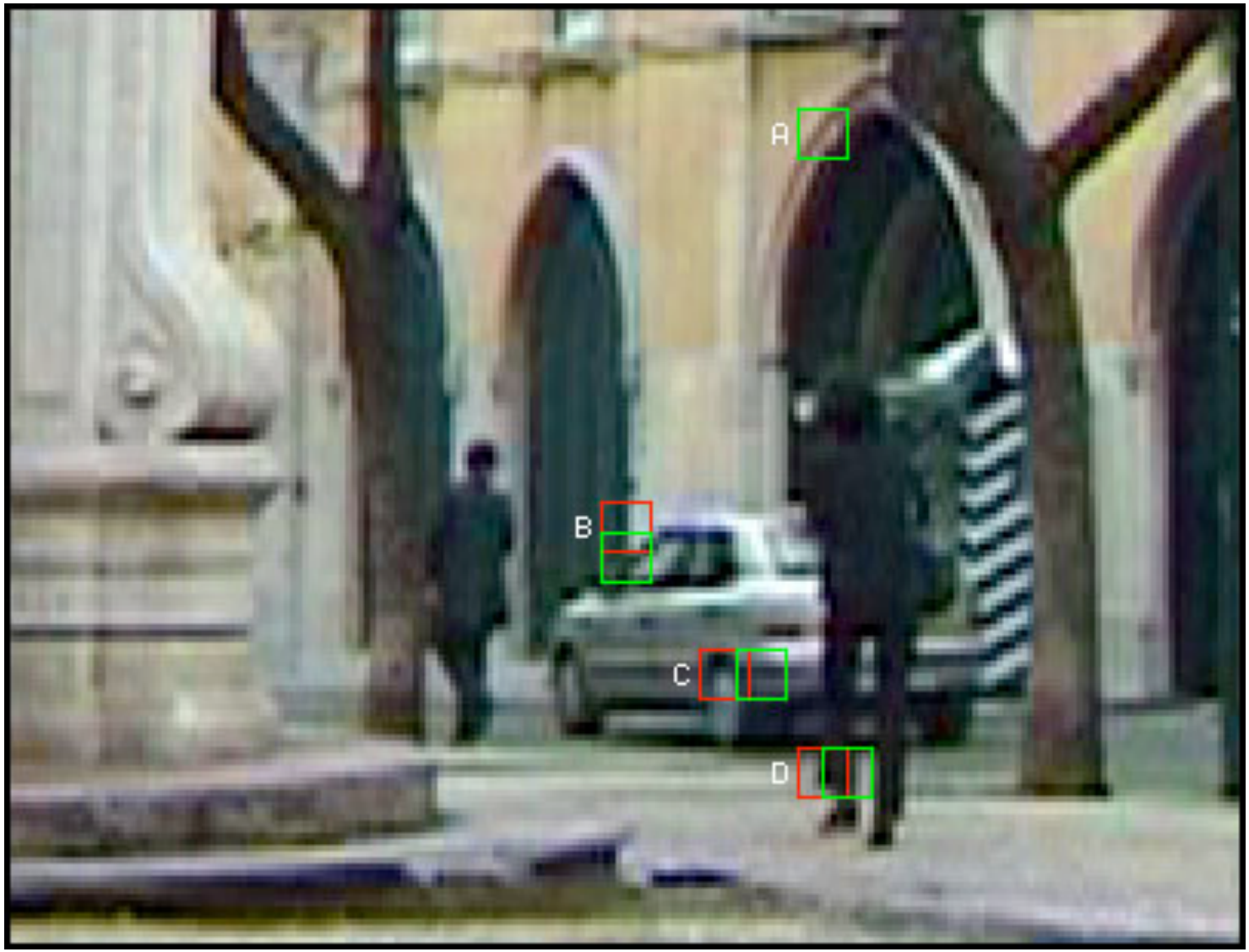


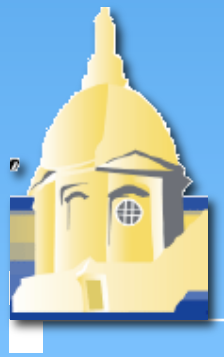
Detail


Block in frame #2
Best match in frame #1
Residual (difference)
A
B
C
D



Motion vector





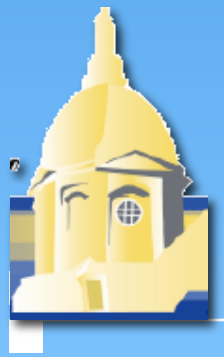
10.4 H.261 digital video compression

- designed for videophone, video conferencing and other audiovisual services over ISDN (pre-DSL telephony/broadband service)
  - The video codec supports bit-rates of p x 64 kbps, where p ranges from 1 to 30 (Hence also known as p * 64)
  - Require that the delay of the video encoder be less than 150 msec so that the video can be used for real-time bidirectional video conferencing

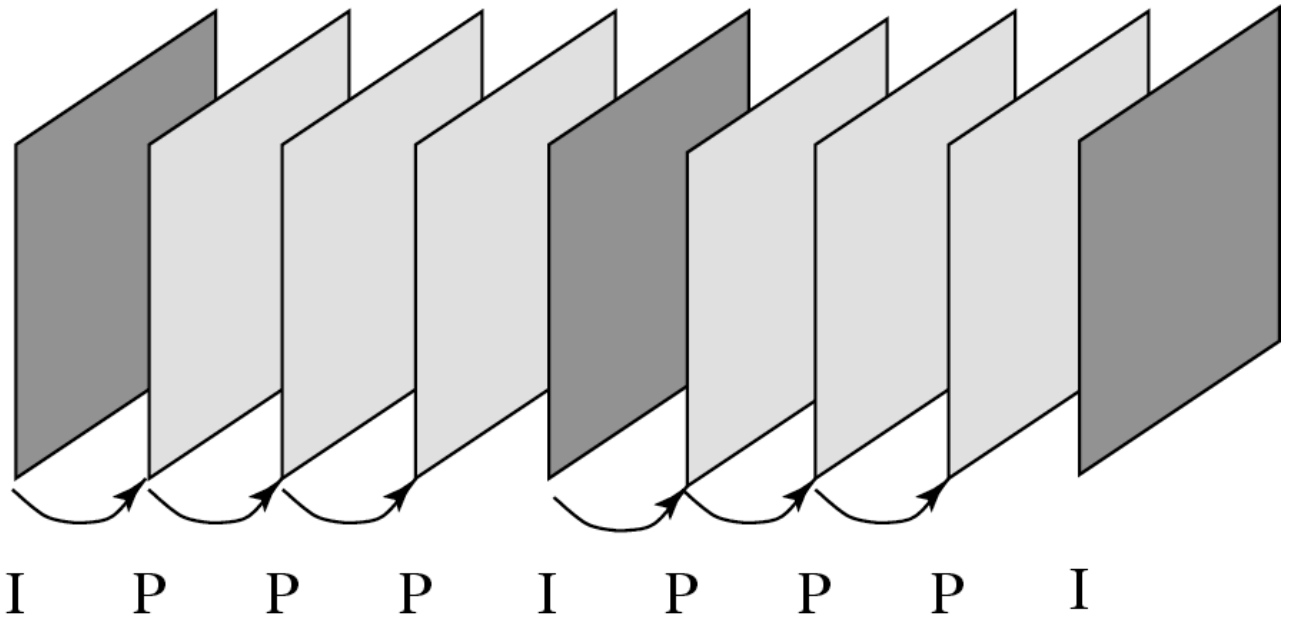
Video format	Luminance image resolution	Chrominance image resolution	Bit-rate (Mbps) (if 30 fps and uncompressed)	H.261 support
QCIF	176×144	88×72	9.1	required
CIF	352×288	176×144	36.5	optional

H.261 Frame Sequence

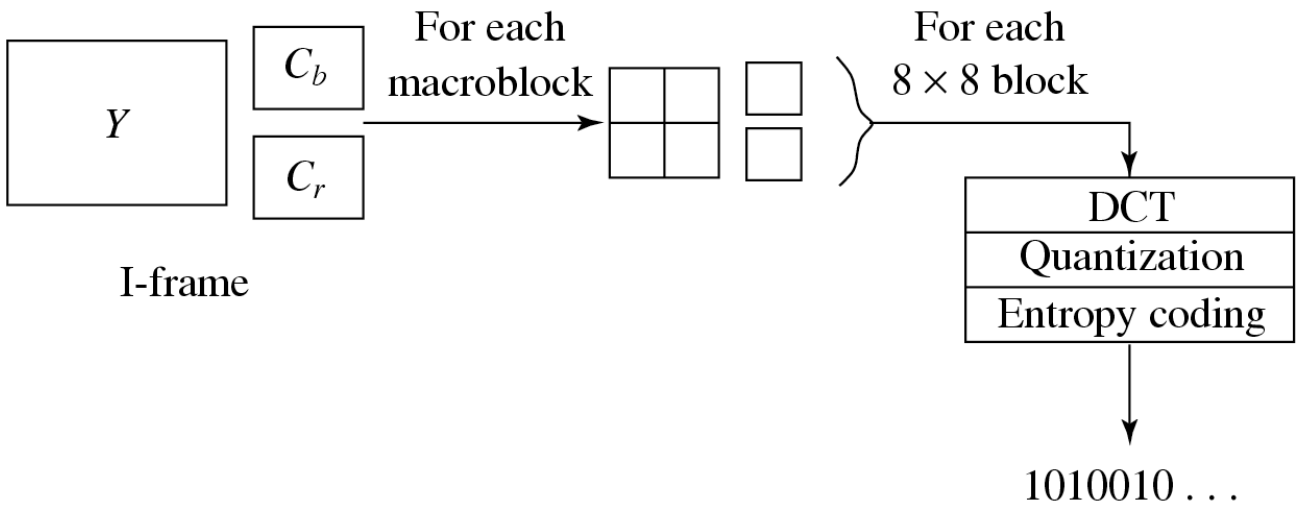
- Two types of image frames are defined: Intra-frames (I-frames) and Inter-frames (P-frames):
  - I-frames are treated as independent images. Transform coding method similar to JPEG is applied within each I-frame, hence “Intra”
  - P-frames are not independent: coded by a forward predictive coding method (prediction from a previous P-frame is allowed — not just from a previous I-frame)
  - Temporal redundancy removal is included in P-frame coding, whereas I-frame coding performs only spatial redundancy removal
  - To avoid propagation of coding errors, an I-frame is usually sent a couple of times in each second of the video



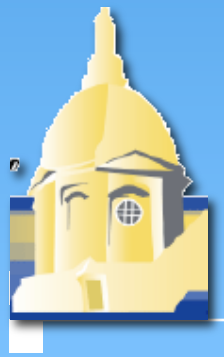
H.261 Frame Sequence.



Intra-frame (I-frame) Coding



- Macroblocks are of size 16 x 16 pixels for the Y frame, and 8 x 8 for Cb and Cr frames, since 4:2:0 chroma subsampling is employed. A macroblock consists of four Y, one Cb, and one Cr 8 x 8 blocks.
- For each 8 x 8 block a DCT transform is applied, the DCT coefficients then go through quantization zigzag scan and entropy coding.



Quantization in H.261

- The quantization in H.261 uses a constant step_size, for all DCT coefficients within a macroblock
- If we use DCT and QDCT to denote the DCT coefficients before and after the quantization, then for DC coefficients in Intra mode:

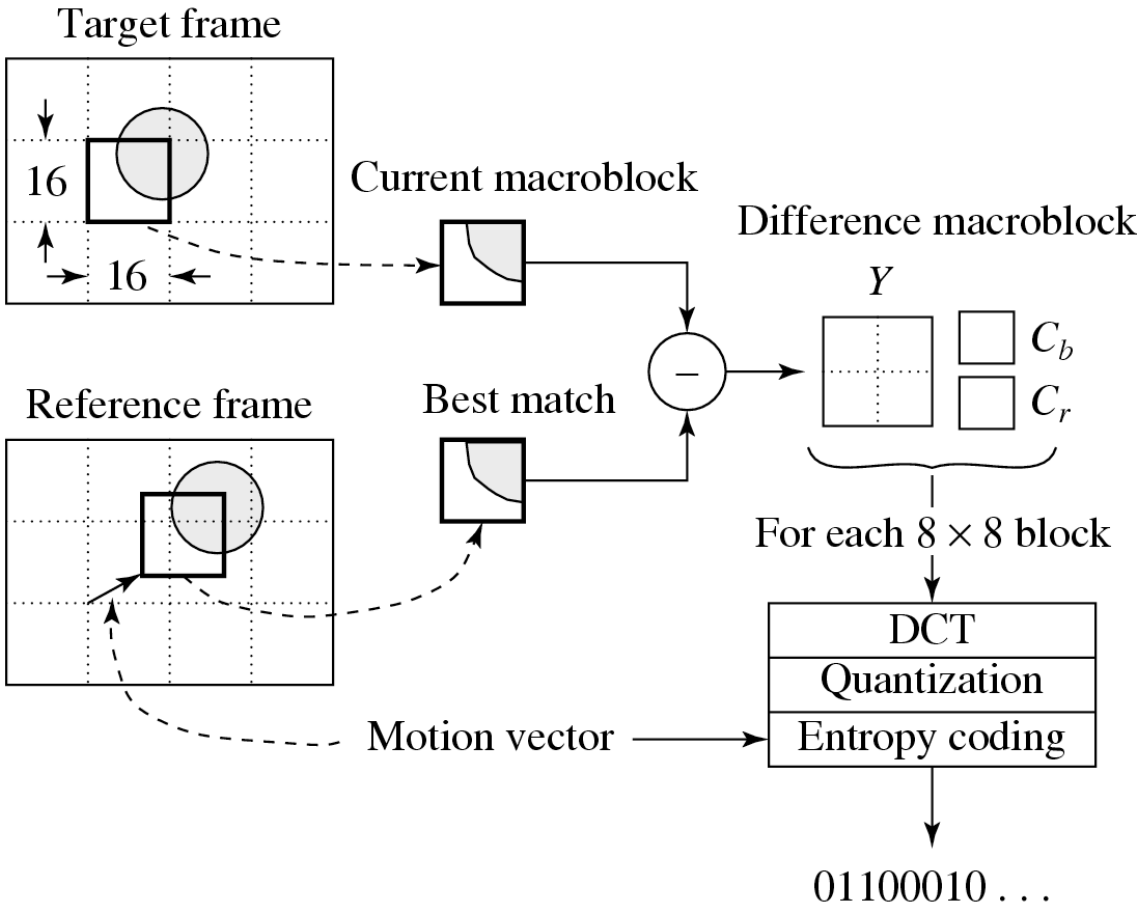
$$QDCT = round\left(\frac{DCT}{step\_size}\right) = round\left(\frac{DCT}{8}\right)$$

- for all other coefficients:

$$QDCT = \left\lfloor \frac{DCT}{step\_size} \right\rfloor = \left\lfloor \frac{DCT}{2 * scale} \right\rfloor$$

scale — an integer in the range of [1, 31]

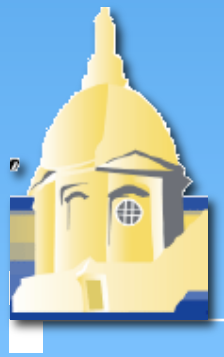
Inter-frame (P-frame) Predictive Coding



Motion vectors in H.261 are measured in units of full pixel and they have a limited range of ± 15 pixels, i.e., p = 15.

- For each macro block in the Target frame, a motion vector is allocated using one of the search methods
  - the difference MVD is sent for entropy coding:

    MVD = MVPreceding − MVCurrent
- After the prediction, a difference macro block is derived to measure the prediction error
- Sometimes, a good match cannot be found, i.e., prediction error exceeds a certain acceptable level
  - MB itself is encoded (treated as an Intra MB) referred as non-motion compensated MB



Syntax of H.261 Video Bitstream

- a hierarchy of four layers: Picture, Group of Blocks (GOB), Macroblock, and Block.
  - The Picture layer: PSC (Picture Start Code) delineates boundaries between pictures. TR (Temporal Reference) provides a time-stamp for the picture.
  - The GOB layer: H.261 pictures are divided into regions of 11 x 3 macroblocks, each of which is called a Group of Blocks (GOB).

GOB 0	GOB 1
GOB 2	GOB 3
GOB 4	GOB 5
GOB 6	GOB 7
GOB 8	GOB 9
GOB 10	GOB 11

CIF

GOB 0
GOB 1
GOB 2

QCIF