ITCT Lecture 11.1: An Overview of Video Coding Algorithms

Prof. Ja-Ling Wu

Department of Computer Science
and Information Engineering
National Taiwan University

1

- Video coding can be viewed as image compression with a temporal component since video consists of a finite sequence of images.

Video coding =
Image coding
+
Strategy to take advantage of temporal correlation

- Of all the different modalities of data, video is the one that produces the largest amount of data.
- A video is a sequence of correlated images.

2

- Video compression can be viewed as the compression of a sequence of images, images with a temporal component. However, there are limitations to this approach:
  - We do not perceive motion video in the same manner as we perceive still images.
  - Motion video may mask coding artifacts that would be visible in still images. On the other hand, artifacts that may not be visible in reconstructed images can be very annoying in reconstructed motion video sequences.

EX:

- (1) A compression scheme that introduces a modest random amount of change in the average intensity of the pixels in the image.
  Unless a reconstructed still image was being compared side by side with the original image, this artifact may go totally unnoticed.
  However, in a motion video sequence, especially one with low activity, random DC variations can be quite annoying.
- (2) Poor reproduction of edges can be a serious problem in the compression of still images. However, if there is some temporal activity in the video sequence, error in the reconstruction of edges may go unnoticed.

- Most of the video compression algorithms make use of the temporal correlation to remove redundancy:
  - The previous reconstructed frame is used to generate a prediction of the current frame.
  - The different between the prediction and the current frame, the prediction error or residue, is encoded and transmitted to the receiver.
  - ------- Predictive Coding based Approach!

5

- The previous reconstructed frame is also available at the receiver.
- If the receiver knows the manner in which the prediction was performed, it can use this information to generate the prediction values and add them to the prediction error to generate the reconstruction.

—— Please recall the DPCM coding scheme!

The prediction operation in video coding has to take into account motion of the objects in the frame and is known as "Motion Compensation"

6

- **Symmetric Video Coding Algorithms:** H.261, H.263, H.263+
- When the compression algorithm is being designed for two-way communication
  - It is necessary for the coding delay to be minimal, and the compression and decompression should have about the same level of complexity.
- **Asymmetric Video Coding Algorithms:** MPEG-1/2,4 H.264/AVC
- When the compression algorithm is being designed one-way (broadcasting) applications – the complexity can be unbalanced.
- There is one transmitter and many receivers, and the communication is essentially one way.

- For an asymmetric video coding application, the encoder can be much more complex (≈10-fold) than the receiver, and there is more tolerance for encoding delays.
- In applications where the video is to be decoded on mobile devices, the decoding complexity has to be extremely low in order for the decoder to decode a sufficient number of images to give the illusion of motion (≥25fps). --- Distributed Video Coding:

  Lightweight-client devices

  User Generated Video

  Cloud Computing Environment

- In general, the encoding can not be done in real time due to its complexity.
- When the video is to be transmitted over "Error-prone" channels (such as wireless networks), the effects of channel noises (e.g., interference or packet loss) have to be taken into account when designing the compression algorithm (such as issues of Error-Correction and Error-Recovery).

Each application will present its own unique requirements and demand a solution that fits those requirements.

9

I. Motion Compensation:

- In most video sequences there is little change in the contents of the image from one frame to the next. Even in sequences that depict a great deal of activity, there are significant portions of the image that do not change from frame to frame.
- Most video compression schemes take advantage of this redundancy by using the previous frame to generate a prediction for the current frame.

10

- If we try to apply the differential coding techniques blindly to video compression by predicting the value of each pixel by the value of the pixel at the same location in the previous frame, we will run into trouble because we would not be taking into account the fact that objects tend to move between frames.
- The object in one frame that was providing the pixel at a certain location (i_o, j_o) with its intensity value might be providing the same intensity value in the next frame to a pixel at location (i_1, j_1).
- If we don't take this into account, we can actually increase the amount of information that needs to be transmitted.

- In order to use a previous frame to predict the pixel values in the frame being encoded, we have to take the motion of objects in images into account.
- Although a number of approaches have been investigated, the method that has worked best is the approach called: block-based motion compensation.
- In this approach, the frame being encoded is divided into blocks of size MxM. For each block, we search the previous reconstructed frame for the block of size MxM that most closely matched the block being encoded.

- Usually, we measure the closeness of a match, or distance, between two blocks by the sum of absolute differences between corresponding pixels in the two blocks. We would obtain similar results if we used the sum of squared differences between the corresponding pixels as a measure of distance.
- Generally, if the distance of the block being encoded to the closest block in the previous reconstructed frame is greater than some prespecified threshold, the block is declared uncompensable and is encoded without the benefit of prediction. This decision is also transmitted to the receiver.

13

- If the distance is below the threshold, then a motion vector is transmitted to the receiver. The motion vector is the relative location of the block to be used for prediction obtained by subtracting the coordinates of the upper left corner pixel of the block being encoded from the coordinates of the upper left corner pixel of the block being used for prediction.

14

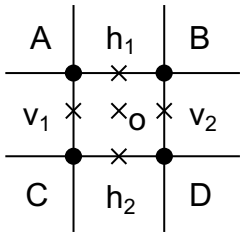
- Suppose the block being encoded is an 8x8 block between pixel locations (24,40) and (31,47); that is, the upper left corner pixel of the 8x8 block is at (24,40). If the block that best matches it in the previous frame is located between pixels at location (21,43) and (28,50), then the motion vector would be {(21,43)-(24,40)} = (-3,3).
- Note that the blocks are numbered starting from the top left corner. Therefore, a positive x component means that the best matching block in the previous frame is to the right of the location of the block being encoded.

15

- If the displacement between the block being encoded and the best matching block is not an integer
  - ➔ $\frac{1}{2}(\frac{1}{4})$-pixel motion compensation algorithms.
- In order to do this, pixels of the coded frame being searched are interpolated to obtain twice (four times) as many pixels as in the original frame. This interpolated image is then searched for the best matching block.

16

- Example: Half-pixel motion compensation

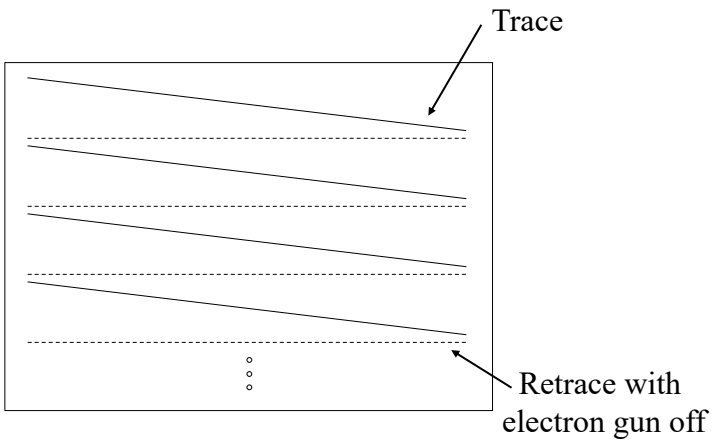


$$h_1 = \left\lfloor \frac{A+B}{2} + 0.5 \right\rfloor, \quad h_2 = \left\lfloor \frac{C+D}{2} + 0.5 \right\rfloor$$

$$v_1 = \left\lfloor \frac{A+C}{2} + 0.5 \right\rfloor, \quad v_2 = \left\lfloor \frac{B+D}{2} + 0.5 \right\rfloor$$

$$o = \left\lfloor \frac{1}{4}(A+B+C+D) + 0.5 \right\rfloor$$

II. Video Signal Representation



The path traversed by the electron beam in a TV

- A black-and-white TV picture is generated by exciting the phosphor (磷光劑) on the TV screen using an electron beam whose intensity is modulated to generate the image we see.
- The path that the modulated electron beam traces is shown in the above. The line created by the horizontal traversal of the electron beam is called a line of the image. In order to trace the second line, the electron beam has to be deflected back to the left of the screen. During this period, the gun is turned off in order to prevent the retrace from being visible.

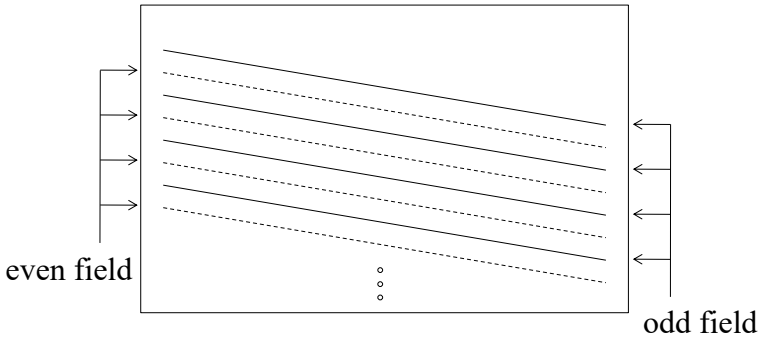
19

- The image generated by the traversal of the electron gun has to be updated rapidly enough for persistence of vision to make the image appear stable.
- However, higher rates of information transfer require higher bandwidths, which translate to higher costs.
- To keep the cost of bandwidth low, it was decided to send 525 lines, 30 times a second (30 fps).
- These 525 lines are said to constitute a "frame".

20

- Unfortunately, a thirtieth of a second between frames is long enough for the image to appear to flicker. To avoid the flicker, it was decided to divide the image into "Interlaced Fields". Even more serious in HDTV!
- A field is sent once every sixtieth of a second.
- First, one field consisting of 262.5 lines is traced by the electron beam. Then, the second field consisting of the remaining 262.5 lines is traced between the lines of the first field. The situation is shown schematically in the following.

21



A frame and its constitute fields

22

- The first field is shown with solid lines; the second, with dashed lines. The first field begins on a full line and ends on a half line, while the second field begins on a half line and ends on a full line.
- Not all 525 lines are displayed on the screen. Some are lost because of the time required for the electron gun to position the beam from the bottom to the top of the screen (time for fly-back).
- We actually see about 486 lines per frame.

23

- In a color TV, instead of a single electron gun, we have three electron guns that act in unison. These guns excited red, green, and blue phosphor dots imbedded in the screen.
- The beam from each gun strikes only one kind of phosphor and the gun is named according to the color of the phosphor it excites.
- Each gun is prevented from hitting a different type of phosphor by an aperture mask.

24

- In order to control the 3 guns we need 3 signals, a red signal, a green signal, and a blue signal. If we transmit each separately, we would need 3 times the bandwidth. With the advent of color TV, there was also the problem of backward compatibility:
  - Some people had black-and-white TV sets, and TV stations did not want to broadcast using a format that some of the viewing audience could not see on their existing sets.
  - Both issues were resolved with the creation of a composite color signal.

25

- In USA, the specifications for the composite signal were created by the National Television Systems Committee, and the composite signal is offer called an NTSC signal. The corresponding signals in Europe are PAL (Phase Alternating Lines), developed in Germany, and SECAM (Séquential Couleur Avec Mémoire), developed in France.

nicknames:

NTSC: Never Twice the same Color.

SECAM: System Essentially Against the Americans.

26

- The composite color signal consists of a 'Luminance' signal, corresponding to the black-and white TV signal, and two 'chrominance' signals.
- The luminance is denoted by Y:

$$Y = 0.299R + 0.587\,G + 0.114B$$

- The weighting of the 3 components was obtained through extensive testing with human observers. The two chrominance signals are:

$$Cb = B\text{-}Y \text{ and } Cr = R\text{-}Y.$$

- Y, Cb and Cr can be used by the color TV to generate R, G, B signals to control the electron guns. The Y signal can be used directly by the B/W TV. Color conversion!

27

- Because the eye is much less sensitive to changes of chrominance in an image, the chrominance signals do not need to have higher-frequency components. Thus, lower bandwidth of the chrominance signal along with a clever use of modulation techniques permits all three signals to be encoded without need of any bandwidth expansion.
- Chrominance signals are always subsampled!!

28

Digital TV:

- International Consultative Committee on Radio (CCIR), also known as ITU-R: CCIR 601-2 or ITU-R recommendation BT.601-2. ➔ CCIR-601.
- The standard proposes a family of sampling rates based on the sampling frequency 3.725MHz. Each component can be sampled at an integer multiple of 3.725MHz up to a maximum of 4 times this frequency.
- The sampling rate is represented as a triple of integers, with the first integer corresponding to the sampling of the luminance signal and the remaining two corresponding to the chrominance signals.

- 4:4:4 sampling means that all signals were sampled at 13.5MHz. The most popular sampling format is the 4:2:2 format, in which the luminance is sampled at 13.5MHz, while the lower bandwidth chrominance signals are sampled at 6.75MHz.
- If we ignore the samples of the portion of the signal that do not correspond to active video, sampling rate translates to 720 samples per line for the luminance and 360 samples per line for the chrominance.

- The luminance component of the digital video signal is also denoted by Y, while the chrominance components are denoted by U and V.
- The sampled analog values are converted to digital values as follows.
- The sampled values of YCbCr are normalized so that the sampled Y values, Ys, taken on values between 0 and 1 and the sampled chrominance values, Crs and Cbs, taken on values between $-\frac{1}{2}$ and $\frac{1}{2}$.

- These normalized values are converted to 8-bit numbers according to the transformations:

  Y = 219Ys + 16

  U = 224Cbs + 128

  V = 224Crs + 128

- Thus, the Y components takes on values between 16 and 235, and the U and V components take on values between 16 and 240.

- The YUV data can also be arranged in other formats.
- In the Common Interchange Format (CIF), used for video conferencing, the luminance of the image is represented by an array of 288×352 pixels, and the two chrominance signals are represented by two arrays consisting of 144×176 pixels. In the QCIF (quarter CIF) format, we have half the number pixels in both the rows and columns.

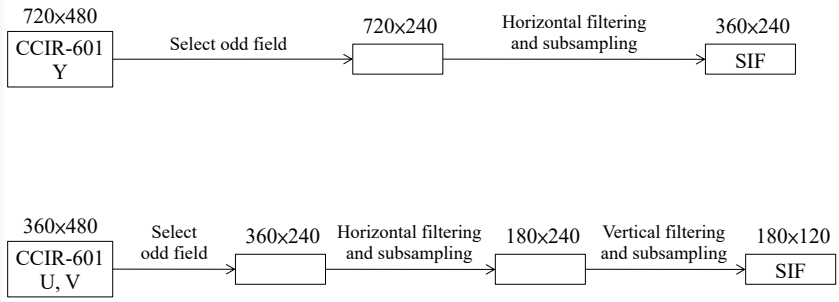
33

- The MPEG-1 algorithm, which was developed for encoding video at rates up to 1.5Mbits per second, used a different subsampling of CCIR-601 format to obtain the MPEG-SIF format. Starting from a 4:2:2, 480-line CCIR-601 format, the vertical resolution is first reduced by taking only the odd field for both the luminance and the chrominance components. The horizontal resolution is then reduced by filtering (to prevent aliasing) and then subsampling by a factor of two in the horizontal direction. ➔

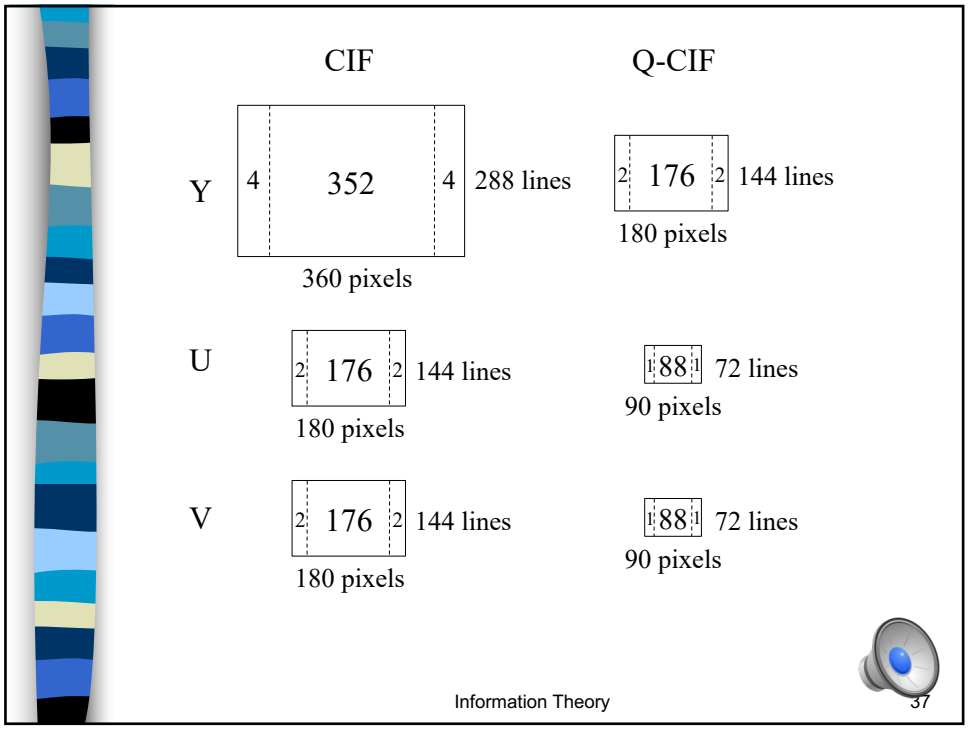
34

- Y: 360×240 samples; U, V: 180×240 samples.
- The vertical resolution of the chrominance sample is further reduced by filtering and subsampling in the vertical direction by a factor of two to obtain 180×120 samples for each of the chrominance signals. The conversion process is shown in the following:

35



36



37

30 frames/sec, 8bit/pixel:

- CIF: 36.5Mb/s ÷ 64kb/s = (ISDN) 570
- QCIF: 9.12Mb/s 142.5

- QCIF, 10 frames/sec : 142.5/3 = 47.5 : compression ratio.

- This analysis gives us an example for estimating the required **compression ratio** of a specific application! Compression of **VR/AR** contents?

38

Video Coding Algorithm:

- A hybrid transform/DPCM with motion estimation
  - Intra-frame coding: transform coding + VLC
  - Inter-frame coding: predictive coding (Motion estimation / compensation)

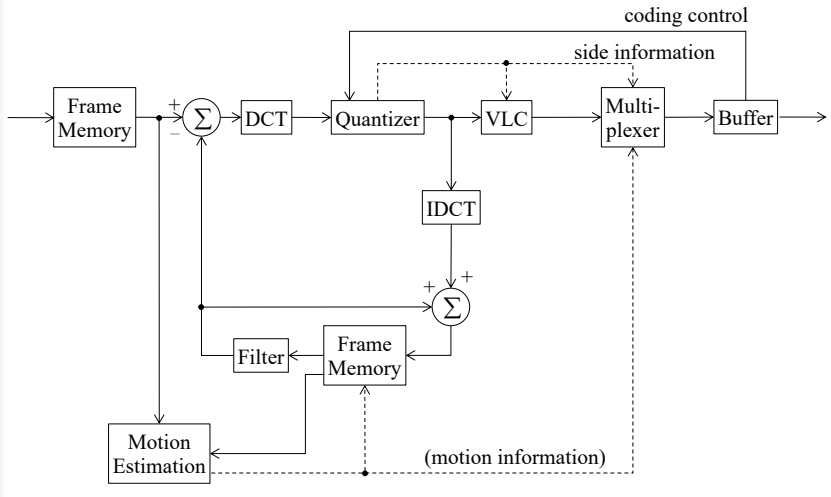
- Motion Estimation: Block Matching

$$\text{MotionVector}(V,H) = \min_{v,h} \sum_{i=1}^{16} \sum_{j=1}^{16} |a(i,j) - b(i+v, j+h)|$$

$a(i,j)$: the luminance pixel value in a 16×16 macroblock (MB) in the current frame

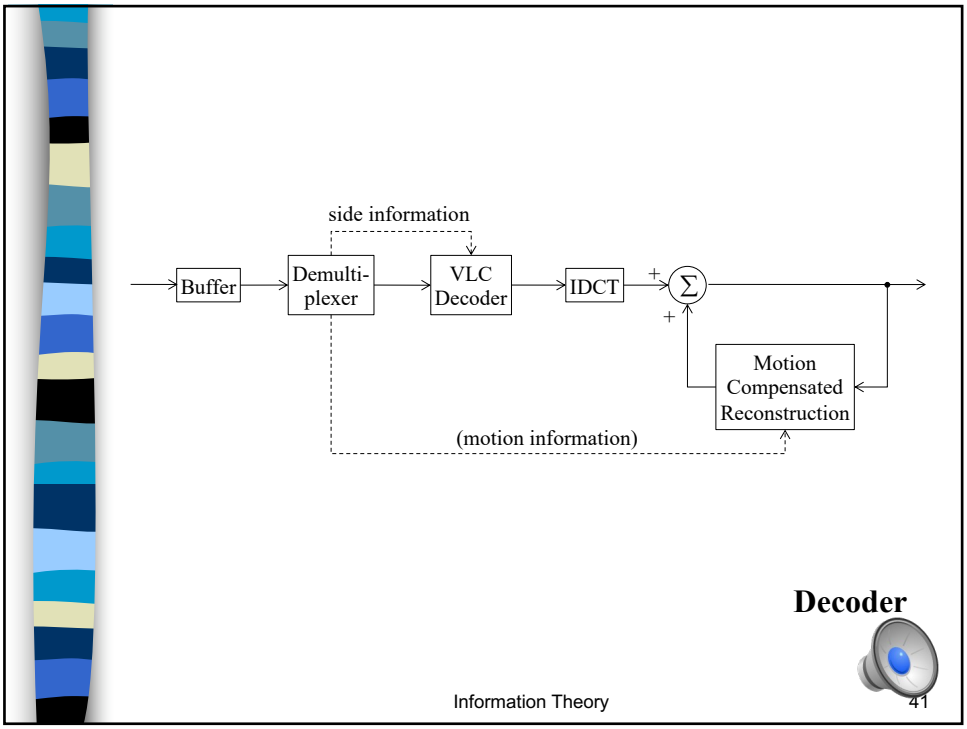
$b(i+v, j+h)$: the corresponding luminance pixel value in a 16×16 MB that is shifted (v,h) in the previous fra

39



Encoder

40



Decoder

41

The (Loop) filter:

- Sometimes sharp edges in the block used for prediction can result in the generation of sharp changes in the prediction error. This in turn can cause high values for the high-frequency coefficients in the transforms, which can increase the transmission rate.
- To avoid this, prior to taking the difference, the prediction block can be smoothed by using a 2-D spatial filter. The filter is separable: it can be implemented as 1-D filters along the rows and columns alternatively. The filter coefficients are (1/4, 1/2, 1/4), except at block boundaries where one of the filter taps would fall outside the block. To prevent this from happening, the block boundaries remain unchanged by the filtering operation.

42