ECE 418: Introduction to Image & Video Processing

Video compression

Shuai Huang

04/16/2019

University of Illinois at Urbana-Champaign
Overview

1. Video compression
   1.1 Basic compression techniques
   1.2 Motion estimation
   1.3 Color video compression
   1.4 Bidirectional prediction
   1.5 MPEG
Video compression
Video compression differs from still image compression:

- The data rates are higher. We need very high compression ratio, and relatively simple compression algorithms.
- The temporal characteristics of video are different from its spatial characteristics. Video also needs to be accessed at arbitrary time instants for operations such as play out, fast forward, rewind and editing.

Only lossy compression is of practical interested.
Basic compression techniques

Notation for video compression.

- The spatial coordinates are denoted by \( n_1 \) and \( n_2 \), the temporal coordinate is denoted by \( n_3 \).
- The original video signals is denoted by \( x(n_1, n_2, n_3) \), the decompressed video signal is denoted by \( \tilde{x}(n_1, n_2, n_3) \).
Basic compression techniques

Intraframe coding

Each frame is encoded separately using still image compression techniques.

- The main advantages are simplicity, fast encoding and decoding.
- Exploit spatial redundancies.
Temporal differential pulse code modulation (DPCM)

The temporal dependencies of adjacent video frames are exploited via temporal DPCM. The *temporal prediction error*

\[
e(n_1, n_2, n_3) := x(n_1, n_2, n_3) - \tilde{x}(n_1, n_2, n_3 - 1)
\]

is encoded using a feedback loop introduced earlier.

- Most pixels in the difference image \(e(n_1, n_2, n_3)\) have small values.
- Some spatial correlation are still apparent, the \(e(n_1, n_2, n_3)\) can be further encoded via transform coding.
Basic compression techniques

Figure 1: The difference between two video frames.
Basic compression techniques

Figure 2: Predictive encoding scheme with feedback.
## Basic compression techniques

### Motion compensation

Motion-compensated predictive coding is based on the elementary assumption about the scene: it contains moving objects.

- If motion of objects was known, we could track them to improve *temporal prediction*.
- Motion information needs to be transmitted to the decoder. The bit savings far overweight the additional motion information.
- Motion-compensated predictive coding is the state-of-the-art in video compression. Current international standards are based on it.
Motion compensation

Figure 3: One original frame in a video.
Figure 4: Differences between the original frame and the next frame.
Figure 5: Motion compensated difference.
Motion estimation

- Motion extraction is difficult CV task.
- Practical video compression uses an elementary approach based on simple signal-processing operations.

Translation block-motion model

- The current frame is partitioned into small blocks.
- Each block is assumed to undergo a translation.

Find a single motion vector $\vec{v}$ that describes the displacement of each block.

The decoder need not perform motion extraction, and is much simpler than the encoder.
Motion estimation

Figure 6: Translation block-motion model.
Motion estimation

Figure 7: Translation block-motion model.
Motion estimation

To simplify notation

- Let $C(n_1, n_2) = x(n_1, n_2, n_3)$ be the current frame, and $P(n_1, n_2) = x(n_1, n_2, n_3 - 1)$ be the previous frame;
- $\hat{C}(n_1, n_2)$ be the prediction of $C(n_1, n_2)$, and $\tilde{P}(n_1, n_2)$ be the previous decompressed frame.

Pixels in each block $B$ is predicted as

$$\hat{C}(n_1, n_2) = \tilde{P}(n_1 - v_1, n_2 - v_2), \quad (n_1, n_2) \in B$$  \hspace{1cm} (1)

where $(v_1, v_2)$ is the motion vector for block $B$. 
The Displaced Frame Difference (DFD), previously known as prediction error, is

\[ C(n_1, n_2) - \hat{C}(n_1, n_2) \]

Motion compensation significantly reduces error, it also reduces the variance and entropy of the error image.

- DFD contains mostly edges, and very small low-frequency components.
- Transform coding can be applied on DFD.
Motion estimation

As before, motion compensation is performed on the decompressed frame $\tilde{P}(n_1, n_2)$ rather than the original frame $P(n_1, n_2)$.

Figure 8: Motion-compensated predictive coder and decoder.
Matching criterion

We choose the mean-squared prediction error as the matching-criterion

\[
MSE = \sum_{(n_1, n_2) \in B} \left| C(n_1, n_2) - \tilde{P}(n_1 - v_1, n_2 - v_2) \right|^2
\]

(2)

for every block \( B \) in the current frame.

Block matching

- Find a block \( B' \) from the previous frame that best matches the current block \( B \).
- Block \( B \) and \( B' \) are related by the motion vector \( \vec{v} \)

The matching criterion is minimized over a set of candidate motion vectors.
An alternative matching criterion is the sum of absolute errors

$$MAE = \sum_{(n_1,n_2) \in B} \left| C(n_1, n_2) - \tilde{P}(n_1 - v_1, n_2 - v_2) \right|$$  \hspace{1cm} (3)

- $MSE$ is slightly better than $MAE$.
- $MAE$ is preferred in hardware implementation due to its lesser complexity: no multiplication is involved.
Search procedure

We assume the displacement is never larger than some maximum displacement.

- The search is restricted to a finite search window.
- Evaluating the matching criterion is repeated for each candidate motion vector.
Search procedure

Full search: all motion vectors inside the search window are evaluated.

Fast search: A limited number of motion vectors are evaluated. A popular method is the $n$-step search algorithm: the search window is the square

$$[-2^n + 1, 2^n - 1] \times [-2^n + 1, 2^n - 1]$$

- Fast search is used in practice due to computational limitations.
- It yields sub-optimal matches due to the limited search.
Search procedure

Figure 9: Three-step motion vector search.
Search procedure

Computing the prediction in (1) requires the motion vector \( \vec{v} \) has integer components. The prediction is often improved if \( \vec{v} \) has half-pixel accuracy.

○ We can interpolate the current and previous frames by a factor of 2, usually via bilinear interpolation.

○ In practice, the pixel-accuracy motion vector is estimation in the first step, then refined to half-pixel-accuracy motion vector via a local search.
Choosing the appropriate block size

- If the block size is too small, the match is unreliable. More motion vectors need to be transmitted to the decoder.
- If the block size is too large, the translation model is less likely to be valid.

Most international compression standards use a block size of $16 \times 16$. 

Motion model failure

The motion model is not perfect:

- The motion model breaks down when two objects with different velocities appear in the same block. The motion vector may capture the motion of one object but not both.
- The translation model could not handle rotation, zoom or illumination changes between two consecutive frames.

Motion vectors minimize some criterion, they do not necessarily correspond to true motion.

- Due to motion model failures and quantization noise effects.
- Accurate representation of true motion is not the objective in compression applications.
For color video compression, in the $YIQ$ or $YC_BC_R$ color space

- Only the luminance image is used for motion estimation.
- The same motion vectors is used for prediction of luminance and chrominance images.
- The prediction error image are encoded separately using transform coding.
Bidirectional prediction

- **Forward Motion Compensation (FMC):** the prediction is based on the previous frame.
- **Backward Motion Compensation (BMC):** the prediction is based on the next frame.

![Diagram showing past, current, and future frames for bidirectional prediction.]

This area can now be predicted using “future frame.”
Bidirectional prediction

Figure 10: FMC and BMC.

BMC deals well with occlusion or uncovered background.
Bidirectional prediction

Bidirectional prediction defines a choice among three predictors

- The FMC prediction $\hat{B}_f(n_1, n_2)$.
- The BMC prediction $\hat{B}_b(n_1, n_2)$.
- Their average $\frac{1}{2} (\hat{B}_f(n_1, n_2) + \hat{B}_b(n_1, n_2))$.

The predictor that yields the best match is selected.

- Advantage: better prediction, no error propagation, higher SNR.
- Disadvantage: increased computational complexity, increased buffer size, increased delay.
Moving Picture Experts Group (MPEG) is an organization formed in 1988 to set standards for audio and video compression and transmission.

- MPEG-1: a standard for storage and retrieval of audio and video on storage media.
- MPEG-4: a standard for multimedia applications.
- MPEG-7: a content representation standard for information search.
- MPEG-21: offer metadata information for audio and video files.
MPEG-1

**Input format:** the maximum frame size is $4095 \times 4095$.

**Groups of Pictures:** Intraframe coded (I), predicted using FMC (P), bidirectionally predicted (B).

- I-picture is coded without referencing any other pictures, used for fast forward/reverse searches.
- P-picture is coded referencing the last I- or P-picture.
- B-picture is coded referencing the last and the next non-B-pictures.
- The first GOP sequence must be an I-picture. Typical GOP sequence: IBBPBBPBBPBB.
Macroblocks: $16 \times 16$ luminance block.

Motion compensation: a motion vector is assigned to each macroblock.

Macroblock types: depends on the frame type.

DCT: on $8 \times 8$ blocks.

Quantization: uniform quantizer with frequency-dependent step sizes.

Entropy coding: zigzag scanned followed by variable-length-coding.

Entropy coding: limited bit rate to 1.862 Mbit/s.
MPEG-2 is extension to MPEG-1.

- Motion vectors are represented by half-pixel accuracy.
- Substantially reduces bandwidth required for high-quality transmission.
- Higher frame size, higher temporal resolution, improved quantization/coding, higher bit rate.

Widely used in multimedia communications, broadcasting, telecommunications, etc.
MPEG-4 is built on MPEG-2 and MPEG-1.

- Improved coding efficiency.
- Ability to encode mixed media data (video, audio, speech)
- Allows higher quality video at lower data rates and smaller file size.
- More robust and error resilience, interactive ability.

Used in digital TV, mobile multimedia, streaming videos, etc.
Acknowledgment: The slides are based on the lecture notes by Prof. Pierre Moulin, UIUC.