Principles of Digital Video Coding

Harilaos Koumaras
University of the Aegean, Greece

Evangellos Pallis
Technological Educational Institute of Crete, Greece

Anastasios Kourtis
National Centre for Scientific Research "Demokritos", Greece

Drakoulis Martakos
National and Kapodistrian University of Athens, Greece

INTRODUCTION

Multimedia applications and services have already possessed a major portion of today's traffic over communication networks. The revolution and evolution of the World Wide Web has enabled the wide provision of multimedia content over the Internet and any other autonomous network.

Among the various types of multimedia, video services (transmission of moving images and sound) are proven dominant for present and future communication networks. Although the available network bandwidth and the corresponding supporting bit rates continue to increase, the raw video data necessitate high bandwidth requirements for its transmission. For example, current commercial communication networks throughput rates are insufficient to handle raw video in real time, even if low spatial and temporal resolution (i.e., frame size and frame rate) has been selected. Towards alleviating the network bandwidth requirements for efficient transmission of audiovisual content, coding techniques have been applied on raw video data, which perform compression by exploiting both temporal and spatial redundancy in video sequences.

Video coding is defined as the process of compressing and decompressing a raw digital video sequence, which results in lower data volumes, besides enabling the transmission of video signals over bandwidth-limited means, where uncompressed video signals would not be possible to be transmitted. The use of coding and compression techniques leads to better exploitation and more efficient management of the available bandwidth.

Video compression algorithms exploit the fact that a video signal consists of sequence series with high similarity in the spatial, temporal, and frequency domain. Thus, by removing this redundancy in these three different domain types, it is possible to achieve high compression of the deduced data, sacrificing a certain amount of visual information, which however it is not highly noticeable by the mechanisms of the *human visual system*, which is not sensitive at this type of visual degradation (Richardson, 2003).

Thus, the research area of video compression has been a very active field during the last few years by proposing various algorithms and techniques for video coding (International Telecommunications Union [ITU], 1993; ITU 2005a, 2005b; Moving Picture Experts Group [MPEG], 1998; MPEG, 2005a, 2005b). In general video compression techniques can be classified into two classes: (1) the *lossy* ones and (2) information preserving (*lossless*). The first methods, although maintaining the video quality of the original/uncompressed signal, do not succeed high compression ratios, while the lossless ones compress more efficiently the data volume of initial raw video signal with the cost of degrading the perceived quality of the video service.

The lossy video coding techniques are widely used, in contrast to lossless ones, due to their better performance. More specifically, by enhancing the encoding algorithms and techniques, the latest proposed coding methods try to perform in a more efficient way both the data compression and the maintenance of the deduced perceived quality of the encoded signal at high levels. In this framework, many of these coding techniques and algorithms have been standardized, encouraging by this way the interoperability between various products designed and developed by different manufacturers.

This article deals with the fundamentals of the video coding process of the lossy encoding techniques that are common on the great majority of today's video coding standards and techniques.

BACKGROUND

The majority of the compression standards have been proposed by the ITU and the International Organization for Standardization (ISO) bodies, by introducing the following standards H.261, H.263, H.263+, H.263++, H.264, MPEG-1, MPEG-2, MPEG-4 and MPEG-4 Advanced Video Coding (AVC).

Some of the aforementioned standards were developed in partnership of ITU with MPEG, exploiting similar coding techniques developed by each body separately.

Each standard was designed and proposed targeting a specific service and application, featuring therefore specific parameters and characteristics. For example H.261 was proposed in 1990 for transmission of video signals over Integrated Services Digital Network (ISDN) lines on which data rates are multiples of 64 kbit/sec. The H.263 standard was designed as a low bit rate encoding solution for video-conferencing applications.

Similarly MPEG-1 was proposed by MPEG in order to be used by the video compact disc (VCD) medium, which stores digital video on a compact disc (CD) with a quality almost similar to that of an analog VHS video. In 1994 MPEG-2 was proposed for encoding audio and video for broadcast signals, exploiting interlace format. MPEG-2 is also the coding format used by the widely successful commercial digital versatile disc (DVD) medium.

Regarding the latest H.264, or MPEG-4 Part 10 AVC, it was proposed in common by the ITU Telecommunication Standardization Sector (ITU-T) Video Coding Experts Group (VCEG) and the ISO MPEG as the outcome of a joint venture effort known as the Joint Video Team (JVT). The scope of H.264/AVC project is to create a standard that would be capable of providing broadcast video quality at very low bit rates on a wide variety of applications, networks and systems.

Finally, in order to create a framework, which will reassure the interoperability of the codec implementation among the various developers, the standards include the concept of profiles and levels, defining a specific set of capabilities to be defined and implemented for a specific subset of applications and services.

VIDEO CODING

All the aforementioned video coding standards are based on the same basic coding scheme, which briefly consists of the following stages: (1) the temporal, (2) the spatial, (3) the transform, (4) the quantization and (5) the entropy coding stage.

The temporal stage exploits the similarities between successive frames with scope to reduce the temporal redundancy in a video sequence. The spatial stage exploits spatial similarities located on the same frame, reducing by this way the spatial redundancy. Then the output parameters of the temporal and spatial stages are further quantized and compressed by an entropy encoder, which removes the statistical redundancy in the data, producing an even more compressed video stream.

More analysis of each stage follows.

TEMPORAL STAGE

As input to the temporal stage of the encoding process, the uncompressed video sequence is used, which contains a lot of redundancy between its successive frames. The scope of this stage is to remove this redundancy by constructing a prediction of each frame based on previous or future frames, enhanced by compensating for fine differences between the selected reference frames. Depending on the prediction level, by which each frame is constructed, each frame is classified to three discrete types, namely:

(1) Intra-frame (I), (2) Predictive (P) and (3) Bidirectional predictive (B), widely referred as I, P, and B. The I-frames are also called Intra frames, while B and P are known as Inter frames.

- I-frames do not contain any prediction from any other coded frame.
- P-frames are coded based on prediction from previously encoded I- or P-frames.
- B- frames are coded based on prediction from previously or future encoded I- or P-frames.

The pattern of successive types of frames like IBBPBBPBBP… forms a Group of Pictures (GOP), whose length is mainly described by the distance of two successive I-frames.

Therefore, in order to perform this temporal compression, two discrete processes are performed at this stage—the motion estimation and motion compensation. Both these processes are usually applied on specific rectangular regions of a frame, called blocks if their size is 8 x 8 pixels or macroblocks if they are 16 x 16 rectangular pixel regions. At the latest standards (i.e., H.264) as Figure 1 depicts, variable block sizes are used for motion compensation depending on the content, achieving better coding efficiency.

During motion estimation, the encoding algorithm searches for an area in the reference frame (past or future frame) in order to find a corresponding matching region. The process of specifying the best match between a current frame and a reference one, which will be used as a predictor of the current frame, is called motion estimation. This is performed by comparing specific rectangular areas (i.e., blocks/macroblocks) in the reference and current frame, until the best match is detected. Due to this, their spatial differences are calculated, using the Sum of Absolute Differences (SAD):

$$SAD(d_x,d_y)=\sum_{i=0}^{15}\sum_{j=0}^{15} |f(i,j)-g(i-d_x, j-d_y)|$$