

UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE PATENT TRIAL AND APPEAL BOARD

Apple, Inc.
“Apple”
Petitioners

v.

Advanced Coding Technologies, LLC
“ACT”
Patent Owner

INTER PARTES REVIEW OF U.S. PATENT NO. 9,986,303

DECLARATION OF Dr. Sheila S. Hemami

I declare that all statements made in this declaration on my own knowledge are true and that all statements made on information and belief are believed to be true, and further, that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Date: May 9, 2025

By: 

Dr. Sheila S. Hemami

I, Dr. Sheila Hemami, Ph.D., do hereby declare:

1. I am making this declaration at the request of Apple in the matter of INTER PARTES REVIEW OF U.S. PATENT NO. 9,986,303, “the ’303 Patent.”

2. I am being compensated for my work in this matter at my standard hourly rate of \$850 for consulting services. My compensation is in no way contingent on the result of this proceeding or the content of my testimony.

3. I have considered information from numerous sources in forming the opinions expressed in this Declaration. I have employed methods and analyses of a type reasonably relied upon by experts in my field in forming opinions or inferences on the subject. To make these determinations, I have also drawn on my decades of experience and expertise in this field. In preparing this Declaration, I considered the following materials:

- (a) U.S. Patent No. 9,986,303 (“the ’303 Patent”) (Exhibit APPLE-1001)
- (b) Excerpts from the Prosecution History of the ’303 Patent (Exhibit APPLE-1002)
- (c) U.S. Patent Publication No. 2008/0181302 (“Demircin”) (Exhibit APPLE-1005)
- (d) U.S. Patent Publication No. 2007/0153916 (“Demircin916”) (Exhibit APPLE-1006)

- (e) U.S. Patent Publication No. 2011/0211631 (“Kimoto”) (Exhibit APPLE-1007)
- (f) Excerpts from the Prosecution History of Kimoto (“Kimoto File History”) (Exhibit APPLE-1008)
- (g) ITU-T Recommendation H.264 (11/07) (Exhibit APPLE-1009)
- (h) U.S. Patent Publication No. 2004/0196972 (“Zhu”) (Exhibit APPLE-1010)
- (i) U.S. Patent No. 8,335,252 (“Chen”) (Exhibit APPLE-1011)
- (j) U.S. Patent No. 8,175,154 (“Luthra”) (Exhibit APPLE-1012)
- (k) U.S. Patent No. 9,094,641 (“Ansley”) (Exhibit APPLE-1013)
- (l) ISO/IEC 11172-2: 1993 (E) (MPEG1 Video) (Exhibit APPLE-1015)
- (m) ISO/IEC 13818-2: 2000 (E) (MPEG2/H.262 Video) (Exhibit APPLE -1016)
- (n) Infringement Contentions, Advanced Coding Technologies LLC v. Samsung Electronics Co., Ltd. et al., 2:22-cv-00499-JRG, E.D. Tex., filed December 30, 2022 (Exhibit APPLE-1017)
- (o) Schwarz et al., Overview of the Scalable Video Coding Extension of the H.264/AVC Standard, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, v.17, n.9 (September 2007) (Exhibit APPLE-1018)
- (p) ITU-T Recommendation H.262 (02/98) (Exhibit APPLE -1019)
- (q) ITU-T Recommendation H.263 (02/98) (Exhibit APPLE -1020)

- (r) Radha et al., The MPEG-4 Fine-Grained Scalable Video Coding Method for Multimedia Streaming Over IP, IEEE TRANSACTIONS ON MULTIMEDIA, v.3, 53 (March 2001) (Exhibit APPLE -1024)
- (s) ISO/IEC 14496-2: 2004 (MPEG4) (Exhibit APPLE -1027)
- (t) ISO/IEC 13818-2: 1996 (MPEG2) (Exhibit APPLE -1028)
- (u) U.S. Patent No. 6,292,512 to Radha et al. (Exhibit APPLE -1031)
- (v) U.S. Patent No. 6,697,426 to van der Schaar et al. (Exhibit APPLE -1032)
- (w) U.S. Patent No. 5,886,736 to Chen (Exhibit APPLE-1033)
- (x) Apostolopoulos et al., Video Streaming: Concepts, Algorithms, and Systems, HP Laboratories (Sept. 18, 2002) (Exhibit APPLE-1034)
- (y) U.S. Serial No. 11/479,077 to Demircin, Application as Filed for Demircin916 (“Demircin077”) (Exhibit APPLE-1035)
- (z) Mokhtarian, et al., Authentication of Scalable Video Streams With Low Communication Overhead, IEEE Transactions on Multimedia, Vol. 12, No. 7, November 2010 (Exhibit APPLE-1043)
- (aa) U.S. Patent Publication No. US2009/0060035A1 (“He2009”) (Exhibit APPLE-1044)
- (bb) P. Amon et al., “File format for scalable video coding,” IEEE Transactions on Circuits and Systems for Video Technology, Vol 17, No. 9, September 2007 (Exhibit APPLE-1045)

- (cc) Tian et al., "Sub-sequence video coding for improved temporal scalability"
2005 IEEE International Symposium on Circuits and Systems (ISCAS), 23-26 May
2005 (Exhibit APPLE-1046)
- (dd) Iain Richardson, "The H.264 Advanced Video Compression Standard,
Second Edition," Wiley, 2010 (Exhibit APPLE-1047)
- (ee) Wien et al., "Real-Time System for Adaptive Video Streaming Based on
SVC," IEEE Transactions on Circuits and Systems for Video Technology, Vol 17,
No. 9, September 2007 (Exhibit APPLE-1048)
- (ff) Wiegand et al., "Introduction to the Special Issue on Scalable Video
Coding—Standardization and Beyond," IEEE Transactions on Circuits and
Systems for Video Technology, Vol 17, No. 9, September 2007 (Exhibit APPLE-
1049)
- (gg) Tektronix, A Guide to MPEG Fundamentals and Protocol Analysis –
Updated with QoS/QoE Diagnostics and Troubleshooting, Primer, *retrieved from*
<https://download.tek.com/document/25W-11418-10.pdf> (Exhibit APPLE-1050)
- (hh) Yao Wang, Scalable Video Coding, Slides for Video Processing &
Communications, Polytechnic Institute of NYU (2011) (Exhibit APPLE-1051)

I. Qualification and Professional Experience

4. I am qualified by education and experience to testify as an expert in
the field of image and video processing and compression and related technologies.

Additional details regarding my qualifications are provided in my curriculum vitae, which is attached as Appendix A to this Declaration (and incorporated by reference herein). My curriculum vitae documents the details of my professional experience in the area of image and video processing and compression and lists all of my publications, as well as certain of my Professional Service Activities, awards and honors, and speaking engagements. Additionally, the following overview of my background pertains to my qualifications for providing expert testimony in regarding the subject matter of this Declaration.

5. I received my B.S. in Electrical Engineering, summa cum laude, from the University of Michigan in 1990. I received an M.S. and Ph.D. in Electrical Engineering from Stanford University in 1992 and 1994, respectively. My graduate studies concentrated on aspects of image and video processing and transmission.

6. I have performed research and written on numerous aspects of image and video processing and compression, including as a graduate student at Stanford University, as a Member of the Technical Staff at Hewlett-Packard Laboratories, as a Professor of Electrical and Computer Engineering at Cornell University, as a Professor of Electrical and Computer Engineering at Northeastern University, and as Director of Strategic Technical Initiatives at the Charles Stark Draper Laboratory. I have published over 130 peer-reviewed journal and conference

papers in the area of image and video processing and compression. I have also supervised eleven (11) Ph.D. theses, two M.S. theses, and twenty-eight Masters of Engineering projects on topics in image and video processing and compression. I have taught graduate classes and seminars on digital image processing and on video coding standards, including MPEG-1, MPEG-2, MPEG-4, and H.264. I have received over \$1.8M in research funding from the National Science Foundation, the Department of the Navy, the Department of Energy, and companies such as Cisco, Intel, and Verizon over an 18-year period in support of this work.

7. I have served on various review panels for the National Science Foundation and on a number of program committees for numerous conferences related to video and image processing and compression.

8. I am a Fellow of the Institute of Electrical and Electronics Engineers (“IEEE”) for “contributions to robust and perceptual image and video communications” and a member of the Signal Processing Society therein. I was a Distinguished Lecturer for the IEEE Signal Processing Society in 2010–11. I have served as the Editor-in-Chief for the IEEE Transactions on Multimedia journal and as Associate Editor for signal representation, coding and compression for the IEEE Transactions on Signal Processing journal. I have also served as the Chair of the Image & Multidimensional Signal Processing Technical Committee (now known

as the Image, Video, & Multidimensional Signal Processing Technical Committee). In 2012, I was the Technical Program co-chair for IEEE's flagship image, video, and signal processing conference—the International Conference on Image Processing—which involved overseeing the peer review process for over 2000 submitted papers and organizing the technical program. In 2020, I was the Awards co-chair for the IEEE International Conference on Image Processing, and my responsibilities included selection of the best papers in the conference.

9. Additional information concerning my professional publications and presentations in the field of image and video processing and transmission are set forth in my current Curriculum Vitae, a copy of which is attached as Exhibit A. This Curriculum Vitae lists the publications authored or co-authored by me; and lists the cases in which I have testified via depositions and trials.

II. Digital Video Compression and Transmission

10. Below I describe the state of digital video compression and transmission as of the '303 Patent's earliest claimed priority date (March 31, 2014), and summarize the development of technologies underlying the features and methods recited in claims 1-2.

A. Basics of Video Compression

11. Video is a sequence of images intended for display at a regular time interval representing a scene or scenes in space and time and providing the illusion

of motion when viewed sequentially. Digital video is the term used for a video in which the frames are all digital images comprising pixels. The temporal (*i.e.*, frame rate) and spatial (*i.e.*, frame size in pixels) resolution of digital cameras and digital video recorders produces file sizes (if stored) and bandwidths required (if transmitted) that exceed the capabilities of practicality and/or feasibility without compression. This was true at the time of the Japanese priority application at issue in this matter (JP 2014-071975, filed March 31, 2014), it was true at the time the '303 Patent was filed (May 29, 2018), and it remains true today.

12. Data compression, or simply compression, reduces the number of bits (*i.e.*, the size of the file) required to represent the digital video. Compression can be performed in one of two regimes: lossless compression and lossy compression. Lossless compression permits an exact recovery of the original signal on a pixel-for-pixel basis. Lossy compression provides greater compression (smaller file sizes), but the original signal cannot be recovered in its original bit-for-bit form from the compressed representation. In the case of video information, lossy compression can be quite effective because a perfect pixel-for-pixel representation of the original is not necessary. Rather, the information can be represented in a manner such that it looks to the human eye equivalent to the original. A practical video coder employs lossy compression to significantly reduce the amount of information required to represent the video, and lossless compression to reduce the

size of the resulting information along with other data required to decode the video.

13. Coding (or encoding) is used as a noun to describe the operation of compression, as well as a verb in lieu of “compressing.” The corresponding operation to decompress is called decoding. More generally, a coder is an algorithm, system, or method that can be implemented in software or hardware to perform coding, and a decoder is an algorithm, system, or method that can be implemented in software or hardware to perform decoding. A coder is not necessarily guaranteed to compress; *i.e.*, it is not necessarily guaranteed to result in a smaller file size. In the case of video, the content is not known *a priori*, so coders are designed with the goal of providing good performance on expected content. In other words, coders are designed using assumptions about the statistics of the input data.

14. In order to provide the desired functionality of compressing video and subsequently decompressing it for use (for viewing, analysis, or for some other purpose), a video coder and a video decoder must correspond to one another. Specifically, an algorithm or process for encoding video implies an algorithm or process for decoding, and vice versa, because the decoder must be able to reverse the operations performed by the encoder in order to generate decoded video. Conversely, an encoder must perform operations that can be reversed by the

decoder (an exception is the operation of quantization; quantization is non-invertible and provides the “loss” in lossy compression; and the decoder performs its operations with the quantized data in lieu of the original data about which the encoder had knowledge).

15. Video coders of the type in the '303 Patent operate by removing redundancies in the video frames. Spatial redundancy reduction reduces the number of bits required to represent the pixels in a single frame (or portion of a frame). It operates on frames as if they were still images, and is often referred to as intraframe compression, intraframe coding, or simply intra coding. Temporal redundancy reduction exploits the similarities between adjacent frames (and those occurring close together in time) to achieve compression. It is often referred to as interframe compression, interframe coding, or simply inter coding, and a common example is using past frames to predict future frames. Spatial and temporal redundancy reduction are combined with a data discarding process (called quantization) and bit-level redundancy reduction into a complete video coder.

16. Bit-level redundancy reduction is achieved via entropy coding. Entropy coding is a technique to losslessly compress data that reduces the number of bits required to represent it. Entropy coding is based upon a statistical model of the input data. More frequently occurring data is represented by fewer bits, and less frequently occurring data is represented by more bits. These representations are not

all the same length (as implied by “fewer” and “more” in the previous sentence), and entropy coding is a form of variable-length coding (VLC). Entropy coding is not relevant to the ’303 Patent and it will not be further discussed.

17. A *hybrid video coder* uses an orthogonal transform for spatial redundancy reduction and motion-compensated prediction for temporal redundancy reduction. The term “hybrid” refers to the combination of these two techniques. The H.261 video compression standard adopted in 1988 was a hybrid video coder, and subsequent ITU and ISO standards have been hybrid video coders. The general structure has been well known for over 3 decades. Below, I will provide more details on both spatial and temporal redundancy reduction as is relevant to the discussion at hand.

18. Video compression was well understood at the time of the ’303 Patent, with multiple video compression standards having been defined by standards bodies, put into practice by hardware and software manufacturers of a variety of products, and used by consumers in a broad range of applications. These standards include MPEG-1, MPEG-2, H.261, H.263, H.264/MPEG-4, and H.265 (HEVC). Further information about these standards and the standards bodies that developed and adopted them is provided in another section below.

B. Spatial Redundancy Reduction

19. Video compression algorithms employ orthogonal transformations to

facilitate efficient representation of the video. An orthogonal transformation is simply a mathematical operation that maps input data (e.g., pixels, or a prediction residual of pixels) from the spatial domain into the frequency domain, in which the number of samples is conserved. Just as we can think of audio of consisting of different frequencies such as low tones, high tones, and those in between, two-dimensional signals such as images and video frames also consist of different frequencies which are now two-dimensional and also can be thought of as representing slowly changing signals or highly varying signals. The transformation itself is invertible. Transformation provides no compression by itself, but it does facilitate compression through quantization. A transform can be considered to be a spectral decomposition, providing a twofold insight into the operation of orthogonal transformation. First, image data is separated into spatial frequencies, which can be coded according to human perception. Secondly, the coefficient data is significantly less correlated than the corresponding pixel information, and hence is easier to represent efficiently, thereby facilitating a goal of compression.

C. Temporal Redundancy Reduction

20. Temporal redundancy reduction exploits the similarities between adjacent frames (and those occurring close together in time) to achieve compression. It is often referred to as *interframe* compression, and it is a form of *predictive* coding.

21. Predictive coding can use very sophisticated predictions, or very simple predictions. I will introduce the concept with a very simple example, and then describe motion compensation, which is the temporal prediction most commonly used in video coders and specifically used in coders of the type in the '303 Patent. Lastly, I will describe prediction structures which influence the order in which the compressed data is generated, transmitted, and decoded.

1. Generic predictive coding

22. Predictive coding attempts to predict the current frame, and forms the difference between the current frame and the prediction. The difference is also called a difference image, a prediction residual, or simply a residual. If the prediction is perfect, the difference is zero and very efficient to represent! In the absence of clairvoyance, if the prediction is accurate, the difference is small and can be represented using fewer bits than coding the current frame independently as a still image (i.e., than intra coding the frame).

2. A simple predictive coding example

23. To provide an intuitive understanding of temporal redundancy reduction via predictive coding, I first introduce the simplest approach. More sophisticated approaches are then easy to understand.

24. Consider the case of video at 30 frames/second, and two sequential frames. Typically, very few things have changed in the 1/30 second between

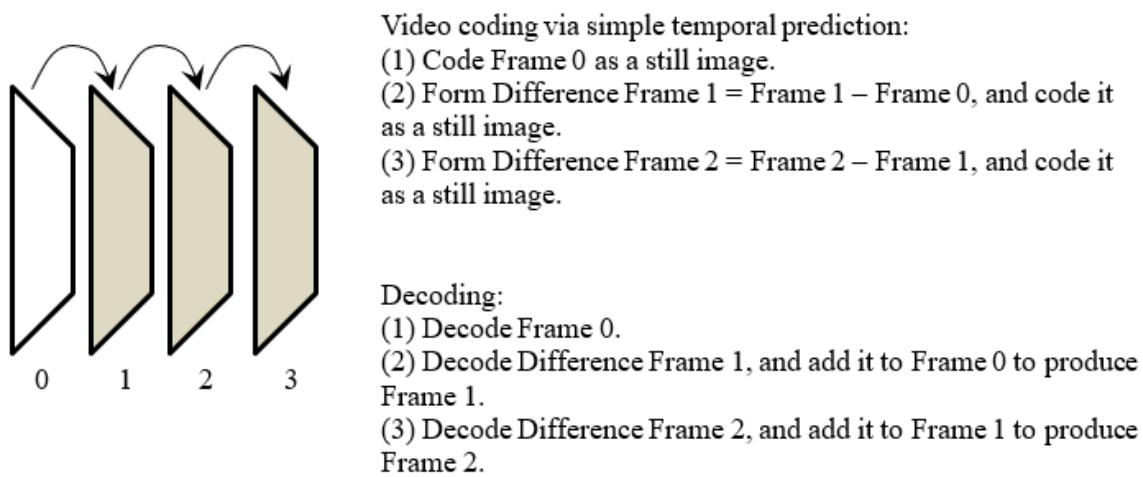
frames, so the two frames will be very similar.

25. The current frame is predicted by the previous frame.¹ A difference image is formed by subtracting the previous frame from the current frame. It is likely that most of the pixels in the difference image will be very small or zero. This difference image is a 2-D array of numbers which can take on both positive and negative values, and the difference image is coded as a still image (e.g., using some form of spatial redundancy reduction, generally combined with data discarding via quantization) and transmitted to the receiver. At the receiver, the decoder reconstructs² the current frame by adding the decoded difference image to the previous frame. The coding process is initiated by coding the first frame as a

¹ More accurately, the prediction is formed using previously encoded data (which is then decoded at the encoder for use in the prediction). Use of previously encoded data ensures that the decoder has the same data available to form the prediction that the encoder had, ensuring that the information can be decoded (in other words, the decoder is not clairvoyant). Additionally, in lossy encoding, use of previously encoded data to form the prediction prevents a catastrophic failure known as “prediction drift.” Appropriate implementation of predictive coding is crucial for it to work successfully.

² I note that my use of “reconstruct” is as in general compression. I understand that the term “reconstruct” was construed by the parties in a previous IPR to mean “rearrange.” Unless specified otherwise in this Declaration, I apply the common meaning to the term “reconstruct.”

still image. This process is illustrated below:



Hemami-FIG. 1

26. The process shown in the figure continues ad infinitum, and illustrates a challenge with practical use of predictive coding. In order to decompress (decode) the third frame, all previous frames must be decoded. This poses problems when a user wishes to skip ahead in viewing a video. Rather than simply beginning the decoding at the desired location, all intermediate frames must be decoded, which is an inefficient use of resources at best. In practical implementation of interframe video coding, periodic “reset” frames are included which are coded as still images. If a user wishes to begin viewing a video at an arbitrary location, the decoder must only begin decoding from the nearest reset frame occurring before the desired location. Reset frames also are important in that they support the decoder being able to restart in the event of data loss, when the

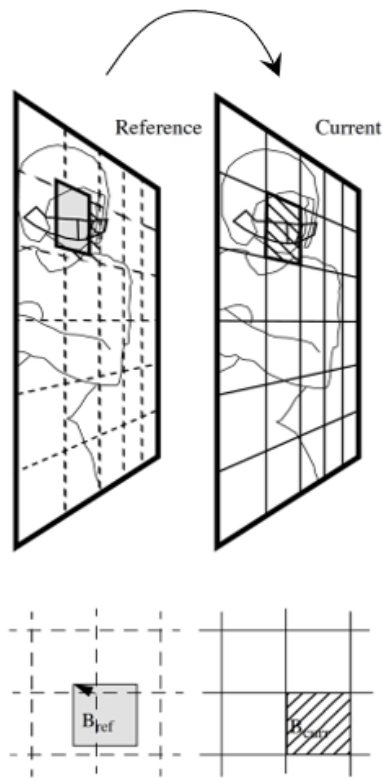
video is transmitted over a channel with loss (e.g., mobile or internet).

27. A collection of frames between reset frames (and including the first one) are often called a “group-of-pictures,” (GOP) and such a group-of-pictures can be independently decoded, without decoding other portions of the video. This term took on a specific meaning early in standards development, and generally was (and is) used to describe a collection of frames with a temporal prediction structure.

3. Better predictive coding: motion compensated prediction

28. More sophisticated interframe video compression techniques do not perform predictive coding on entire frames at once, but instead segment frames into regions or blocks of pixels and process each block separately. As the name implies, “motion compensation” attempts to compensate for motion that has occurred in the time period between two frames. It does so by attempting to align the content in the two frames, on a block-by-block basis, so that the resulting difference block will be as small as possible (or more specifically, so that spatial redundancy reduction can operate as efficiently as possible).

29. For a block in the current frame, the process of alignment begins by searching for a “best-match” block in a so-called reference frame, shown in Hemami-FIG. 2:



To code a block in the current frame:

- (1) Motion estimation is performed to find a best-match block in the reference frame. The offset in the location of the best-match block in the reference frame, relative to the location of the block being coded, is called a motion vector.
- (2) The best-match block is the prediction for the current block and is called a motion-compensated prediction.
- (3) The motion-compensated prediction (i.e., the best-match block) is subtracted from the current block to form a difference block, called a prediction residual (or simply a residual).
- (4) The residual is compressed (typically using a still image compression approach).
- (5) The current block is thus represented by a motion vector and a compressed residual.

$$\text{Residual to be coded} = \text{Current block} - \text{Best-match block}$$

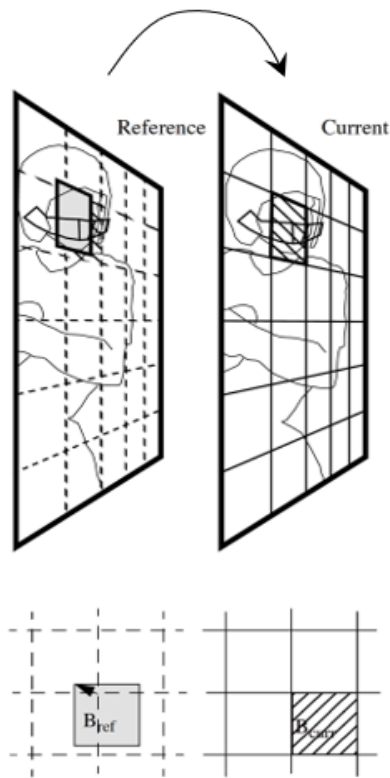
Hemami-FIG. 2

30. If there has been no movement at all between the two frames (no camera motion, and no object motion), the best-match for the current block will be in exactly the same location in the reference frame. However, if there has been a change, the best-match block will be at a different location. The process of finding the best-match block is called “motion estimation.” The location of the best-match block in the reference frame is specified as an offset from the location of the current block. This offset (difference in spatial position) is called a motion vector.

31. The best-match block in the reference frame that is indicated by the motion vector is then the prediction for the current block. The best-match block is

called a motion-compensated prediction (prediction referring to a block of pixels), and the process is also called motion-compensated prediction (prediction referring to the operation of predicting) or simply motion compensation. The predictive picture comprises all of the best-match blocks for the entire frame.

32. Coding proceeds by forming a difference image for each block (current block minus motion-compensated prediction). The difference image is coded like a still image (i.e., using an orthogonal transform followed by quantization). The motion vectors are also part of the coded data, because the decoder must be instructed how to form the prediction that was used to generate the difference image which it will decode. Other data may also be coded (e.g., an indication of which frame was used as a reference, the size of the block, and other information required in order to decode the block).



To code a block in the current frame:

- (1) Motion estimation is performed to find a best-match block in the reference frame. The offset in the location of the best-match block in the reference frame, relative to the location of the block being coded, is called a motion vector.
- (2) The best-match block is the prediction for the current block and is called a motion-compensated prediction.
- (3) The motion-compensated prediction (i.e., the best-match block) is subtracted from the current block to form a difference block, called a prediction residual (or simply a residual).
- (4) The residual is compressed (typically using a still image compression approach).
- (5) The current block is thus represented by a motion vector and a compressed residual.

$$\text{Residual to be coded} = \text{Current block} - \text{Best-match block}$$

Hemami-FIG. 2

33. At the decoder, the block is decoded by forming the motion-compensated prediction (i.e., copying the best-match block from the reference frame as indicated by the motion vector), and adding in the decoded difference block. Such video compression is called motion compensated predictive video coding, and was used in video coding algorithms starting around 1988.

4. *Prediction modes – forward prediction, backward prediction, and bi-directional prediction*

34. The figure and explanation above illustrates predicting a current frame from a reference frame in the past – this is called “forward prediction” because the arrow in the figure points from left to right (forward in time). Forward prediction

works well when the current frame is very similar to frames in the past (as shown in the figure – the football player has moved a little bit in the frame interval of 1/30 second). It does not work well, however, when the current frame coincides with a scene change. It also becomes less effective when the (temporal) distance between the reference frame and the current frame increases.

35. The situation of a scene change can be accommodated by allowing reference frames to be in the future relative to the current frame – a frame is predicted from a reference frame that occurs after it. Such “backward prediction” can be easily accomplished when coding video by including a sufficiently long delay in the encoder – frames are buffered, and the length of the buffer (in frame periods) can allow prediction of an earlier (buffered) frame using a later (buffered) frame.

36. The MPEG-1 standard (the first version of which was formally adopted in 1991) introduced temporal coding using motion compensation with a combination of forward and backward prediction, called bi-directional prediction, which was included in all later standards (discussed in this declaration). The MPEG1 standard is incorporated by reference in this declaration in full. In bi-directional prediction, multiple video frames are buffered prior to encoding and the coded data for multiple frames are buffered before decoding. Temporal coding via motion compensation using forward prediction is used first to predict and code the

blocks in a future frame that typically may be two or three frames ahead in the sequence of buffered frames. A frame coded using forward prediction is called a “P-frame.” Then the blocks in an intervening frame are coded via bi-directional prediction, by searching for matching blocks in both a previous frame and a future frame. A frame coded using bi-directional prediction is called a “B-frame.” The corresponding pixel values in the best match blocks are averaged and the averaged block is used to predict the block in the intervening frame. The resultant residual block and the two corresponding motion vectors are coded.

37. More sophisticated techniques for prediction use more reference frames, and from not only in the past but also from in the near future (which can be especially effective when, for example, scene changes occur). When multiple reference frames are used, the best-match from each reference frame can be averaged to form a prediction for a block. These techniques contribute to the concept of the group-of-pictures.

5. Temporal prediction structure: the group-of-pictures (GOP)

38. As previously explained, “reset frames” can facilitate random access into an encoded video bitstream. One goal of the MPEG-1 video compression standard development was to enable random access, or “trick play” including fast forward, fast rewind, and reverse play. The solution combined the idea of reset frames with temporal prediction and yielded a temporal prediction structure known

as a group-of-pictures, that employed intra- and inter-coded frames.

39. Specifically, the MPEG-1 and MPEG-2 standards defined three types of pictures. MPEG1, at *v*; MPEG2, at *vii*. Intra-coded pictures (known as I-pictures or I-frames) are coded independently, as still images and without reference to other frames. Predictively-coded pictures (known as P-pictures or P-frames) are coded using motion compensated prediction from a previously occurring I-picture or P-picture, and can be used as a reference for prediction. Bidirectionally predicted pictures (known as B-pictures or B-frames) are coded using motion compensated prediction from both a previous and a future reference frame. The organization of the three picture types in a sequence is flexible and left to the encoder to make the appropriate choice for a specific application. An example illustrating a series of coded pictures appears in both the MPEG-1 and MPEG-2 standards and is shown below:

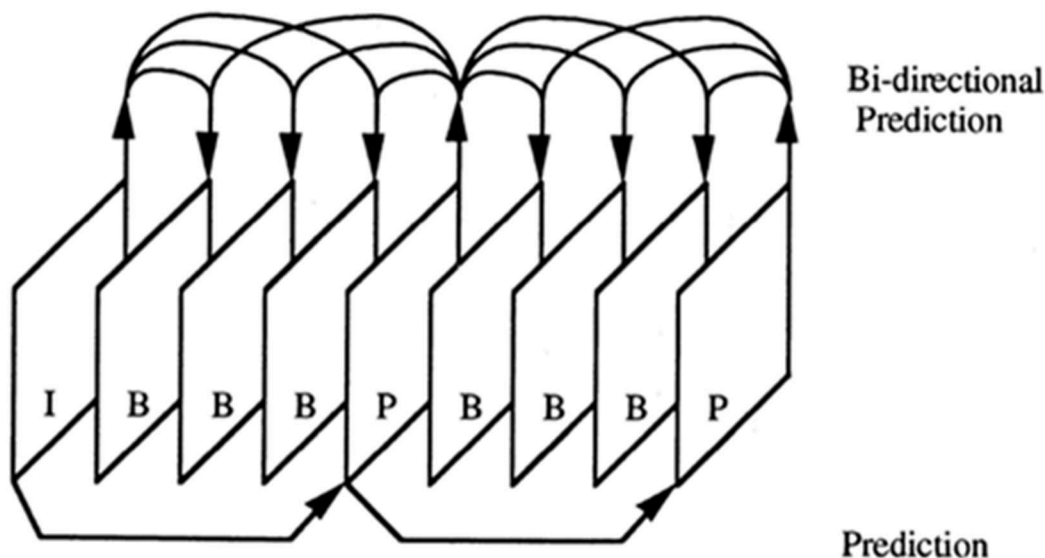


Figure 1 -- Example of temporal picture structure

E.g., MPEG1, FIG. 1.

40. In this example, the first frame is labeled as an I-frame and is intra-coded. The 5th frame is labeled as a P-frame, and the arrow from the first I-frame to the P-frame indicates that it is (forward) predicted from the I-frame. Similarly, the 9th frame is also labeled as a P-frame, and the arrow pointing to it indicates that its reference picture is the previous P-frame (the 5th frame). The B frames use the closest I and/or P frames (both past and future) for prediction. A second example illustrating a different organization of picture types is in Figure D.2 of the MPEG-1 standard. Arrows pointing from a first frame to a second frame indicate that the second frame is predicted from the first (reference) frame. Arrows going from left to right (past to future) indicate forward prediction; arrows going from right to left (future to past) indicate backward prediction. As such, intra-coded

frame I2 is not predicted from another frame; P-frames P5, P8 and P11 are forward-predicted from I-frames or other P-frames, and pairs of frames I2, P5, P8 and P11 are used to predict the corresponding interstitial B-frames.

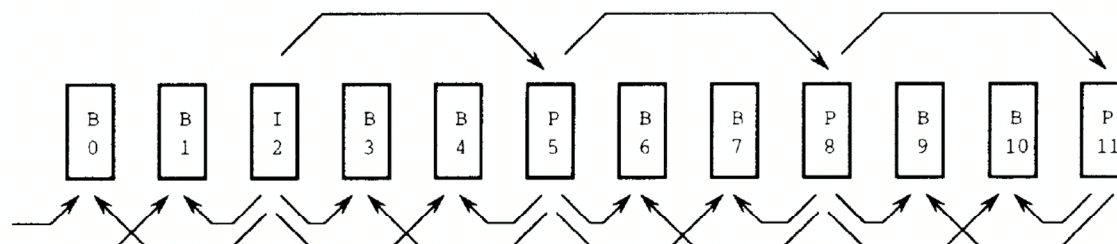


Figure D.2 -- Dependency relationship between I, B, and P-pictures

Figure D.2 illustrates a number of pictures in display order. The arrows show the dependency relationship of the predictive and bidirectionally predictive coded pictures.

MPEG1, 54

41. The above prediction structures support random access – each I-frame can be decoded without requiring any other frames, and hence they represent points-of-entry into decoding the video. MPEG1, at v. If a user wishes to start viewing at a frame that is coded as a P-frame (or a B-frame), it requires decoding of the previous I-frame and any P-frames on which it depends. In all cases – decoding of the entire sequence is not required.

42. The predictive structure imposes constraints on the order in which the frames are processed at the encoder (i.e., coded), and in which they are transmitted and subsequently decoded. Because B-frames depend on both past and future reference frames, these reference frames must be encoded first, and must be

present at the decoder, before encoding (decoding) of the B-pictures can occur.

The requirement for out-of-order decoding and transmission implies the need to store decoded frames at the decoder until they are no longer needed for prediction, and also that the frames are displayed in a different order than they are decoded.

This is illustrated and explained in the MPEG-1 standard's discussion of GOPs:

Group of pictures

A **group of pictures** is a series of one or more coded pictures intended to assist random access into the sequence. In the stored bitstream, the first coded picture in a group of pictures is an I-Picture. The order of the pictures in the coded stream is the order in which the decoder processes them in normal playback. In particular, adjacent B-Pictures in the coded stream are in display order. The last coded picture, in display order, of a group of pictures is either an I-Picture or a P-Picture.

The following is an example of groups of pictures taken from the beginning of a video sequence. In this example the first group of pictures contains seven pictures and subsequent groups of pictures contain nine pictures. There are two B-pictures between successive P-pictures and also two B-pictures between successive I- and P-pictures. Picture '1I' is used to form a prediction for picture '4P'. Pictures '4P' and '1I' are both used to form predictions for pictures '2B' and '3B'. Therefore the order of pictures in the coded sequence shall be '1I', '4P', '2B', '3B'. However, the decoder should display them in the order '1I', '2B', '3B', '4P'.

At the encoder input,

```

|| 1 2 3 4 5 6 7 || 8 9 10 11 12 13 14 15 16 || 17 18 19 20 21 22 23 24 25
| I B B P B B P | B B I B B P B B P | B B I B B P B B P

```

At the encoder output, in the stored bitstream, and at the decoder input,

```

|| 1 4 2 3 7 5 6 || 10 8 9 13 11 12 16 14 15 || 19 17 18 22 20 21 25 23 24
| I P B B P B B | I B B P B B P B B | I B B P B B P B B

```

where the double vertical bars mark the group of pictures boundaries. Note that in this example, the first group of pictures is two pictures shorter than subsequent groups of pictures, since at the beginning of video coding there are no B-pictures preceding the first I-Picture. However, in general, in display order, there may be B-Pictures preceding the first I-Picture in the group of pictures, even for the first group of pictures to be decoded.

MPEG1, 15

43. The MPEG-1 standard included the following general definition for its Group Of Pictures (GOP) data structure:

2.1.71 group of pictures [video]: A series of one or more coded pictures intended to assist random access. The group of pictures is one of the layers in the coding syntax defined in this part of ISO/IEC 11172.

APPLE-1015, 6. MPEG-1 formally defined a group-of-pictures (GOP) as a series of one or more coded pictures containing at least one I-frame and intended to assist

random access into the sequence (MPEG-1 std, pp. 15-16). The term was defined and used in the standard to represent a unit of the coded video. From largest to smallest, these units were sequence (all the frames), group-of-pictures (a series of pictures as described above), picture (a single frame), slice (a subset of a frame), macroblock (16x16 pixels, comprising 4 blocks), block (8x8 pictures).

44. Patent Owner's interpretation of GOP, which is a "grouping of a series of frames into sequences," is illustrated by Figure D.2 from the 1993 MPEG1 standard. This figure includes an I-frame (intra-coded picture) and several B- and P-frames (a plurality of inter-prediction coded pictures). The GOP structure claimed by the '303 Patent was described in the MPEG1 standard as early as 1993. APPLE-1015, 15. "Applications requiring random access, fast-forward playback, or fast and normal reverse playback may use relatively short groups of pictures. Groups of pictures may also be started at scene cuts or other cases where motion compensation is ineffective. The number of consecutive B-Pictures is variable." *Id.*

45. As a result of bi-directional prediction, the video frames are encoded and decoded out of order. This yields a requirement to store some decoded frames in memory (e.g., in buffers) at both the encoder and the decoder for some number of frame intervals, so that they can be used in encoding/decoding and so that all decoded frames can be displayed in the correct order. For example in the above illustration, frame P5 is encoded following encoding of frame I2, and then frames

B3 and B4 are encoded. This is further illustrated in Figures D.3 and D.4 of the MPEG1 Standard, shown below.

Note that because of the picture dependencies, the bitstream order, i.e. the order in which pictures are transmitted, stored, or retrieved, is not the display order, but rather the order which the decoder requires them to decode the bitstream. An example of a sequence of pictures, in display order, might be:

I	B	B	P	B	B	P	B	B	P	B	B	I	B	B	P	B	B	P
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Figure D.3 -- Typical sequence of pictures in display order

whereas the bitstream order would be as shown below:

I	P	B	B	P	B	B	P	B	B	I	B	B	P	B	B	P	B	B
0	3	1	2	6	4	5	9	7	8	12	10	11	15	13	14	18	16	17

Figure D.4 -- Typical sequence of pictures in bitstream order

Because the B-pictures depend on the following (in display order) I or P-picture, the I or P-picture must be transmitted and decoded before the dependent B-pictures.

MPEG1, 54

46. In Figure D.3, the pictures are numbered and sequenced in natural order, corresponding to the order in which the decoded pictures are displayed. In Figure D.4 it can be seen the sequence of picture numbers is not in natural order. The pictures are decoded in this order and then buffered for use in decoding and/or to support reordering for display.

47. Likewise, the GOP structure (IBBPBBPBBPBB) in Figure D.3 also satisfies the patent owner's definition of a GOP because D.3 shows a grouping of I-frames (intra-coded pictures) and a plurality of B- and P-frames (inter-prediction coded pictures).

48. It should be noted the MPEG specifications are fully flexible in the

number of B-frames between any two reference frames. The number could be 0, 1, 2, ... and may vary dynamically between any pair of predicting frames (i.e., it is not fixed within a GOP). In the MPEG standards, B-frames are not used to predict other frames. Examples are shown in the following figure:

```

      I
      I  P  P
      I  B  P  B  P
      B  B  I  B  P  B  P
      B  B  I  B  B  P  B  B  P  B  B  P
      B  I  B  B  B  B  P  B  I  B  B  I  I

```

Figure D.17 -- Examples of groups of pictures in display order

MPEG1, 68

49. The number of interstitial B-frames is determined by an encoder and is application-specific. It is common practice to have a constant number of B-frames, as shown in the above figures.

50. The MPEG1 standard was technically completed in 1991, and published by the standards setting organization in 1993. *E.g.*, APPLE-1015.

51. The later standards maintain the same concepts of reference frames, prediction structures, and I/P/B pictures though in a more sophisticated way. The term group-of-pictures was not formally defined in MPEG-2 or other standards, but the concept of a group-of-pictures was used (and is still used today) to describe prediction structures using the I/P/B frame concept in hybrid video coding that

support random access.

52. For example, MPEG-2 added flexibility by creating coding units called slices, which corresponded to rows of blocks in a picture (a picture comprised multiple slices). The MPEG-2 standard refers to I-slice, P-slice, and B-slice. A picture in which all slices were coded in the same way is referred to as an I-picture, a P-picture, or a B-picture (MPEG-2 std, page 12, section 6.1.1.4 Picture).

53. For example, the 2010 reference entitled "The H.264 Advanced Video Compression Standard, Second Edition" by Iain Richardson includes Section 6.4.7 entitled "Prediction Structures" opens with "H.264 offers many different options for choosing reference pictures for inter prediction." It then goes on to describe several GOP structures:

6.4.7.2 'Classic' Group of Pictures structure

Earlier standards such as MPEG-1 and MPEG-2 supported the Group of Pictures structure shown in Figure 6.42. The Group of Pictures (GOP) starts with an I slice. P slices are inserted at intervals, with B slices between the I and P slices. The I and P slices are used for reference; the B slices are not used for reference, i.e. all pictures are predicted from I or P slices. Each P slice is predicted from the preceding I or P slice and each B slice is predicted from the I and/or P slices on either side of it.

This structure provides higher compression efficiency than the structure shown in Figure 6.41 because prediction in a B slice tends to be more efficient than in a P slice due to the availability of two reference pictures. The disadvantage is increased delay and larger frame storage

requirements. The decoder must receive slices I0 and P3 before slices 1 and 2 can be decoded, i.e. the delay is a minimum of three frames in this example. Slices 0 and 3 must be stored by the decoder because they are reference frames for slices 1 and 2.

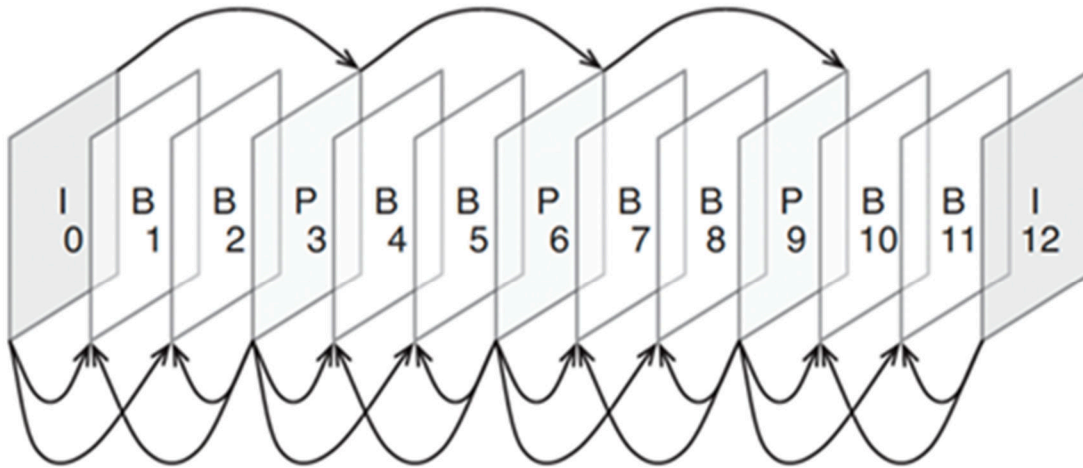


Figure 6.42 ‘Classic’ Group of Pictures prediction structure

6.4.7.4 Hierarchical prediction structures

Figure 6.44 shows an example of a hierarchical GOP structure. This type of structure is made possible by H.264’s flexible prediction options. The GOP starts and finishes with I slices 0 and 12. Next, slice B6 is predicted using I0 and I12 as references; note that B6 is halfway between I0 and I12 in display order. B3 is predicted from I0 and B6; B9 is predicted from B6 and I12. B1 and B2 are predicted from I0 and B3 and so on. The GOP is composed of a series of layers, a dyadic or pyramid decomposition, see Table 6.5. This type of structure can give improved compression performance [iii] if the quantization parameters of the layers are carefully controlled such that the QP increases as the layer number increases. Hierarchical or pyramid GOP structures were first proposed as a method of providing temporal scalability (Chapter 10) but they may be used in any H.264 codec that supports B slices.

Chapter 9 compares the compression performance of several prediction structures.

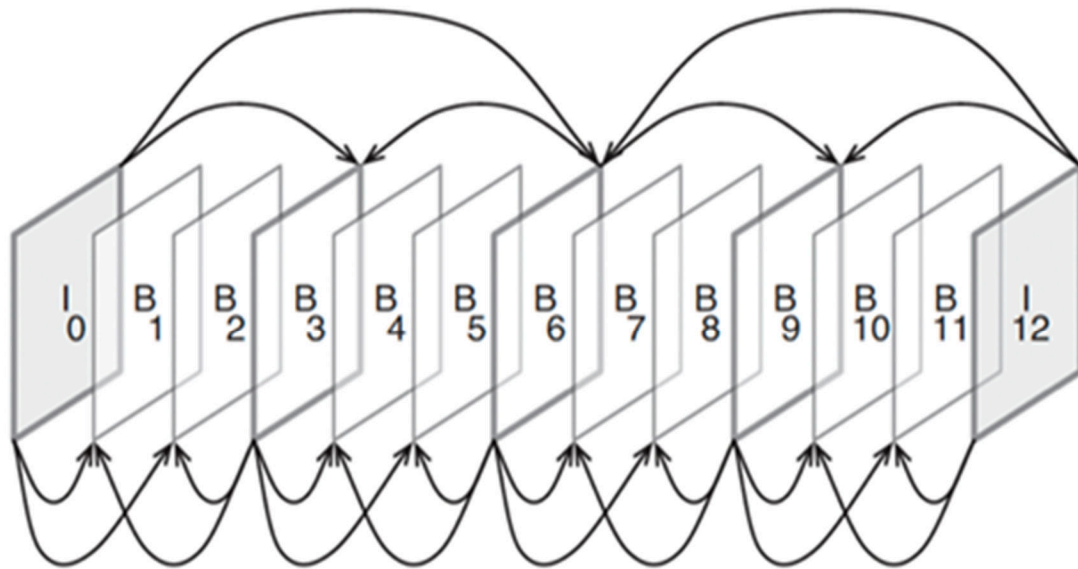


Figure 6.44 Hierarchical GOP structure

Richardson, pp. 169-171

54. Prediction structures given in Chapter 9 as referenced in the above excerpt include

- (i) IPPPPPPPPPPPP. . ., i.e. one I-slice followed by 11 P-slices, with one reference frame
- (ii) As (i) but with five reference frames
- (iii) IBBPBBPBBPBBIBBP. . ., i.e. 12-frame GOP with one I-slice, three P-slices and eight B-slices in each GOP.
- (iv) IBBBBBBBBBBBIBB., 12-frame GOP, hierarchical prediction.

Richardson, pp. 272-273.

55. As the standards evolved, the concept of a GOP was generalized to refer to reference pictures and predicted pictures. This generalization supported the case in which any picture could be used as a reference for prediction. The term GOP then more broadly referred to a collection of reference frames and predicted frames and their prediction structure.

56. As previously mentioned, predictive coding benefits from “reset frames” that support both random access and resilience to data loss during transmission, and the I-frames in MPEG-1 provided that function. As prediction structures became more sophisticated and as digital transmission became more prevalent over mobile and other time varying channels, the concept of a reset frame also became more sophisticated.

D. Scalability and Scalable Video Coding

57. As digital video use became more widespread (owing to the increase in available bandwidth, the reduction in cost of computing and memory and hence consumer electronics that could employ digital video, and the widespread adoption of smartphones, computers, and the internet), providing content to the large diversity of end users became challenging. A user on a 3G network with a smartphone had a lower available bandwidth, a smaller screen size, and lower computing power than a user on a home computer with a fiber optic connection, a large monitor, and a fast CPU with lots of memory. Video compressed for the smartphone user would be of a lower resolution and likely lower perceived quality than the home user would expect, while video compressed for the home user would likely be impossible to transmit, store, and decode on the smartphone. “One size” of compressed video did not fit all.

58. One solution to this issue was called “simulcast coding” or simply

simulcast, in which a single video would be independently compressed for each target user (or classes of target users). (Today's widely used HTTP adaptive bitrate streaming does this.)

59. An alternative solution, which was included in the MPEG-2 video compression standard and has been part of many subsequent standards, is to provide so-called “scalable video,” which can also be stated as providing “scalability.” The term “hierarchical” is also used to describe scalable coding. Scalable video, hierarchical video, scalable coding, layered coding, and hierarchical coding are all terms that were (and are) used for the same idea.

60. In scalable video, a video stream is conceptually thought of as being composed of layers. The *base layer* provides the lowest resolution (or quality), and can be decoded independently by any receiver. Additional layers each provide more data that is used to increase the resolution (or quality) of the video, i.e., to enhance the video beyond the base layer. These additional layers are generally known as *enhancement layers*. Scalable video will have at least one enhancement layer. If it has more than one, each layer adds increasing information, starting with the first layer and ending with the n^{th} layer. The complete set of information includes the base layer and all enhancement layers, which represents the video at the highest quality. This complete bitstream comprises a hierarchical set of substreams representing the video data.

61. A subset of the substreams may represent the video content at a lower level of detail, or at a lower spatial resolution, or at a lower temporal resolution, i.e., a lower frame rate.

62. Importantly, the enhancement layers cannot be decoded independently – by themselves, they do not provide enough information to decode the video. They must be decoded in conjunction with the base layer. More specifically, scalable video requires that the layers maintain specific dependency relationships. Frames in the base layer can only be predicted using other frames in the base layer – this ensures that the base layer can be independently decoded. Frames in the enhancement layer can be predicted from frames in the base layer, and/or frames in the enhancement layer. If more than one enhancement layer is used, frames in the n^{th} enhancement layer cannot be predicted using frames in a higher enhancement layer.

63. For example, if a video sequence was coded using scalable coding into 3 layers, those layers would be (1) the base layer, (2) the first enhancement layer, and (3) the second enhancement layer. These 3 layers would be generated once, by an encoder. A smartphone on a 3G network might receive and decode the base layer, a smartphone over a wifi connection might receive and decode the base layer + the first enhancement layer, and a home user with a HDTV might receive and decode the base layer + the first enhancement layer + the second enhancement

layer.

64. Temporal scalability supports transmission and decoding of video at reduced frame rates relative to the original sequence. If the original sequence is at 30 frames/second, reduced frame rates could be provided at half, quarter, and/or eight rate, corresponding to 15, 7.5, and 3.75 frames/second. These reduced rates are achieved by keeping only every other frame, every 4th frame, and every 8th frame, respectively. Reduced frame rate could alternatively be provided at third rate (corresponding to 10 frames/second, and achieved by keeping only every third frame). In a hybrid video coder, temporal scalability is enabled by the GOP prediction structure, with the structure enabling the reduced rates. The base layer comprises reference frames corresponding to the lowest-rate video, and one or more enhancement layers include more frames, with the decoding of base + all of the enhancement layers resulting in full-frame-rate video. Temporal scalability will be further described and illustrated below. Examples of practical applications for such temporal scalability include a heterogeneous network in which a node on a high-bandwidth network interfaces to a low bandwidth network, at which a full resolution, full bandwidth video bitstream may be forwarded on the low bandwidth network by stripping out the coded B-frames. Another example is encoding a HD television signal such that the base layer provides 30 frames/second video while the enhancement layer allows decoding at 60 frames/second. MPEG-2, at *ix*.

65. SNR (or quality) scalability supports transmission and decoding of video of varying qualities, where quality in this case refers to the perceived level of detail and sharpness in the frames. The base layer comprises the lower quality video (with less detail), and one or more enhancement layers include progressively more detail. The enhancement layer therefore contains information about every frame in the sequence, and the frame rate and frame size of the information in each layer are the same. The MPEG-2 standard from 2000 provides example uses of SNR scalability including two-tier service for HDTV, and also an example in which additional color information is included in the enhancement layer for video production or distribution applications. MPEG-2, at *ix*.

66. Spatial scalability supports transmission and decoding of video at frame sizes that are smaller than the original sequence. The base layer decodes at a smaller frame size, and the enhancement layer information enables decoding of the video at a larger frame size. The enhancement layer therefore contains information about every frame in the sequence, and the frame rate of the information in each layer are the same.

67. As is evident, the coded video data contained in the “base” and “enhancement” layers depends on the type of scalability. In spatial and SNR scalability, information about every frame is contained in both the base layer and the enhancement layer(s). In temporal scalability, each frame is wholly contained

in a single layer. More accurately, in temporal scalability, the enhancement information contributes to decoding a new, previously unavailable frame. Enhancement information contains no information about previously decoded frames in the base layer.

68. Two-layer temporal scalability can be intuitively understood to split the video into two sets of frames. The lower layer provides lower-rate video, and the upper layer contains the additional frames required to achieve full frame-rate.

69. Terminology used in scalability includes layers (base, enhancement, or simply numbered) and levels (especially as it pertains to temporal scalability). A POSITA would understand that these terms need to be interpreted in the context in which they appear.

70. Scalability was well known before 2000 and was explicitly included in H.263, MPEG-2, and later standards while being inherently provided in MPEG-1. Scalable coding was well understood at the time of the '303 priority date. A tremendous amount of complementary efforts were simultaneously directed to developing scalable video solutions, both within and outside of the standards. I provide some detail on the scalability supported in the standards in Section F, “Scalability Support in Video Coding Standards,” *infra*, because it provides a picture of the state-of-the-art. It by no means comprises all prior work on scalable video.

E. Transmission of Scalable Video

1. General video transmission

71. Video inherently has a notion of time associated with it – it is by definition frames (still images) occurring at a regular interval. Once a video decoder begins displaying decoded frames, the temporal nature of video requires that a new frame be displayed every T seconds, where T is the frame interval. Otherwise, the video will freeze until the next frame is available (i.e., a pause in the video will be observed). So for video with a frame rate of 30 frames/second (or 60 frames/second), a new frame must be presented to the display every $1/30$ second (or $1/60$ second).

72. The display requirement implies – obviously – that the decoder must have decoded the frame in order to present it to the display. A frame cannot be decoded until all of the bits representing it have been received.

73. This strict display timing places requirements on the availability of data at the decoder – all of the bits to decode a frame must be available (i.e., have arrived at the decoder) such that it can decode it in time for either its display, or its use in decoding frames that will be displayed before it. The time by which all of the bits representing a frame must be available such that the decoder can fetch them and decode the frame is the “decoding deadline” of the frame.

74. Video also has a bit rate, computed by dividing the number of bits

used to represent the compressed video (colloquially the “file size”) by the length of the video in seconds. The resulting units are bits/second (we often use kilobits/second or megabits/second). The nature of hybrid video coding inherently produces a variable number of bits per frame (i.e., a variable instantaneous bit rate), and as such, the bitrate computed as described above is an average over the entire length of the video. Because of the inherent variability of the instantaneous video bit rate, a more useful value for video bit rate is an average that is computed over a shorter time interval (multiple frames, commonly but not necessarily in units of GOPs).

75. For successful transmission and decoding of real-time video, the video bit rate cannot exceed the channel bandwidth (if it does, decoding deadlines will be missed). In order to ensure reliable transmission, the encoder and transmitter must ensure that the compressed video fits in the channel bandwidth. Adjusting the number of bits to represent the video such that it fits in the channel is called “rate control” and the target number of bits is called a “bit budget.” In a real-time transmission application, this adjustment is performed regularly, typically over given time intervals (again corresponding to multiple frames, commonly but not necessarily in units of GOPs).

76. This average can be compared with the currently available bandwidth and adjusted accordingly.

77. Another feature of general video transmission is the use of buffers at the receiver which can absorb the instantaneous variability of the received bits, to facilitate decoder operation. The receiver buffer “catches and buffers” all the bits as they arrive over the communication channel, so that the decoder can fetch coded data corresponding to complete units that it can process.

78. Many communication channels provide a variable bandwidth that changes based on, for example, use of a shared channel by many users, changes in the physical environment (this is especially the case with wireless channels, such as to a mobile device), and throttling by the carrier. A fixed-rate channel is simply a special case of a channel in which the variability is zero. In video transmission, knowledge of the channel bandwidth is important because the bandwidth influences the rate at which bits arrive at the decoder and are made available for decoding (i.e., enables coded frames to meet or fail to meet their decoding deadline). As such, matching the video data rate to the available channel bandwidth is imperative, and success relies on an accurate value for available channel bandwidth.

79. Channel bandwidth estimation is the operation of obtaining knowledge of the channel bandwidth. At the time of the '303 priority date, many techniques were proposed, known, and in use to do so.

2. Scalable video transmission – additional features and details

80. In the case of scalable video transmission, matching the video data rate to the available channel bandwidth is facilitated by the scalable nature of the bitstream. This matching is called “bit rate adaptation” or “bit rate scaling.” As the channel bandwidth changes, the information transmitted can be adapted by selecting and transmitting the appropriate layers of a scalably coded video; the encoder generating the scalable video does not fundamentally need to change its operation (though nothing precludes it from doing so). Selecting the appropriate number of layers requires an accurate measure of the available channel bandwidth.

81. Operating scalable video coders for time-varying channels and both real-time and streaming applications was well-understood by the ’303 priority date.

82. In 2005, Tian describes “...streaming servers can easily adapt the bitrate of the transmitted streams by deciding which sub-sequence layers and sub-sequences are transmitted.” Tian, at 3.

83. For example, Amon describes transmission of scalably coded video when the scalability is of a single type (e.g., temporal, spatial, or SNR) explaining that such single-type scalability “offers simple adaptation operations at defined qualities by discarding unneeded layers.” Amon then explains that “Adaptation decisions (e.g., during adaptation operations on a network node) can then be based on the scalar layer ID.” Amon, at 1173.

84. He2009 states “The Advanced Video Coding (AVC) standard, Part 10

of MPEG4 (Motion Picture Experts Group), otherwise known as H.264, includes advanced compression techniques that were developed to enable transmission of video signals at a lower bit rate or storage of video signals using less storage space. The newer standard outperforms video compression techniques of prior standards in order to support higher quality streaming video at lower bit-rates and to enable internet-based video and wireless applications and the like...” He2009, [0004]. He further explains that SVC is useful “particularly in scenarios where the downstream client capabilities, system resources, and network conditions are not known in advance, or dynamically changing from time to time.” He2009, [0005].

F. Scalability Support in Video Coding Standards

1. Background

85. Video compression was well understood at the time of the '303 Patent, with multiple video compression standards having been defined by standards bodies, put into practice by hardware and software manufacturers in a variety of products, and used by consumers in a broad range of applications. These standards include (but are not limited to) MPEG-1, MPEG-2, H.264/MPEG-4, H.261, H.263. The MPEG standards were defined by the ISO/IEC, while the H.261/H.263 standards were defined by the International Telecommunications Union (ITU-T). MPEG-4/H.264 was defined by a joint committee. Each of these standards bodies included representatives from numerous industry companies and

organizations. The representatives of these companies worked together and shared ideas and research for improving the standard under development.

86. The ITU was formed in 1865 to develop international standards for global telecommunications technologies. ITU-T refers to the telecommunications sector of the ITU. The ITU-T was also responsible for several video coding standards, including H.261 and H.263. The H.261 Standard was published in 1988. The H.263 Standard, which enabled transmission of video content over telephone lines, was first adopted in 1996, with a later iteration adopted in February 1998.

87. The ISO is an international standard development organization composed of representatives from the national standards organizations of member countries. The IEC is an international standards organization that prepares and publishes international standards for all electrical, electronic and related technologies. Computer and telecommunications standards are handled by joint ISO/IEC committees. As noted above, the ISO/IEC is responsible for the various MPEG standards.

88. Each standard was targeted to a different application and had different properties, as summarized in the chart below. Despite the origination from different standards bodies, the standards were based upon common operations (including those described above). This is not a surprise since they had a common

goal of removing redundancy from a video sequence while being implementable with the software and hardware capabilities of the time.

Standard	Earliest Adoption	Application targets
MPEG-1	1991	Storing movies on CD-ROM
MPEG-2	1996	Broadcast, and storing digital video on DVD
MPEG-4 / H.264	1999	Internet, video, interactive video, content manipulation, professional video, 2-D and 3-D graphics, and mobile video communications
H.261	1988	Videoconferencing using ISDN channels
H.263	1996	Video communication over analog telephone lines
H.265	2013	High resolution, supported via extreme efficiency improvements relative to H.264

2. Scalability in hybrid video coders and MPEG-1

89. The sophistication of the support for scalability increased between the MPEG-1 standard and the later standards. MPEG-1 did not explicitly include scalability, but its GOP structure inherently supported it by design, and the resulting video sequences that could be played in fast-forward or fast-rewind were inherently temporally scalable lower frame-rate versions of the original sequence. (The pre-1990 standard H.261 did not support scalability due to its prediction structure.)

90. The MPEG committee established requirements for the MPEG1 standard and called for proposals in 1989. Many of the resulting proposals were

based on the preceding work for the H.261 videoconferencing standard, and though it was a different standard from a different standards body, MPEG1 may be viewed as an extension of H.261. But for MPEG1, the requirements were focused on a digital replacement for the analog VCR, and included trick play modes such fast-forward, fast rewind and reverse play. Multiple responses to the call proposed groups of pictures as a mechanism to meet these requirements. GOPs as introduced by MPEG-1 have been earlier described.

91. In MPEG-1, a group of pictures must contain an I-picture, may contain P-pictures and B-pictures and can be any length greater than or equal to one I-picture. APPLE-1015, 15-16. In MPEG1, the coding order and display order of the pictures of the GOP can be different from each other. *Id.* In the MPEG-1, B-frames inherently support temporal scalability because they are not used to predict other frames. Their omission can therefore yield a decoded video of a lower frame rate. This fundamental concept of providing temporal scalability based on GOP structure continued and is supported by all subsequent standards.

92. That MPEG-1's GOP structure provided scalability was well recognized, because of the nature of its hybrid coding. In 2005, Tian stated that B-pictures "provide a way to achieve temporal scalability" in referring to H.263 and MPEG-2. Additionally, the same functionality that provides trick play also provides temporal scalability – "decoding only I pictures of each GOP and

decoding only the I and P pictures.” (Tian 2005, page 2) The H.264 Schwarz overview also recognized this fact: “For hybrid video codecs, temporal scalability can generally be enabled by restricting motion-compensated prediction to reference pictures with a temporal layer identifier that is less than or equal to the temporal layer identifier of the picture to be predicted. The prior video coding standards MPEG-1 [2], H.262 MPEG-2 Video [3], H.263 [4], and MPEG-4 Visual [5], all support temporal scalability to some degree. H.264/AVC [6] provides a significantly increased flexibility for temporal scalability because of its reference picture memory control. It allows the coding of picture sequences with arbitrary temporal dependencies, which are only restricted by the maximum usable DPB size. Hence, for supporting temporal scalability with a reasonable number of temporal layers, no changes to the design of H.264/AVC were required.” (Schwartz, page 1107)

93. Details of scalability support in these standards are now provided, with an emphasis on temporal scalability because of its relevance to the matter at hand.

3. The MPEG2 Standard

94. The MPEG2 standard, officially ISO/IEC 13818-2, which is incorporated by reference in this declaration in full, was technically completed in 1994, and was deployed rapidly in the global marketplace as the standard for

digital television broadcasting and digital video recording – DVD. *E.g.*, APPLE-1028, APPLE-1016. The MPEG2 standard also includes optional coding tools for scalable video coding. *E.g.*, APPLE-1028; APPLE-1016.

95. The MPEG2 standard supports three types of scalability – SNR scalability (i.e. quality scalability), spatial scalability and temporal scalability. APPLE-1016, 146. Section D.7.1.2 describes these video coding tools.

96. Temporal scalability in MPEG2 is illustrated in the figure below and summarized in section D.7.4 as follows:

A two layer temporally scalable coding structure consisting of a base and an enhancement layer is shown in Figure D.1. Consider video input at full temporal rate to temporal demultiplexer; in our example it is temporally demultiplexed to form two video sequences, one input to the base layer encoder and the other input to the enhancement layer encoder. The base layer encoder is a non-hierarchical encoder operating at half temporal rate, the enhancement layer encoder is like a MAIN profile encoder and also operates at half temporal rate except that it uses base layer decoded pictures for motion compensated prediction. The encoded bitstreams of base and enhancement layers are multiplexed as a single stream in the systems multiplexer. The systems demultiplexer extracts two bitstreams and inputs corresponding bitstreams to base and enhancement layer decoders. The output of the base layer decoder can be shown standalone at half temporal rate or after multiplexing with enhancement layer decoded frames and shown at full temporal rate.

97.

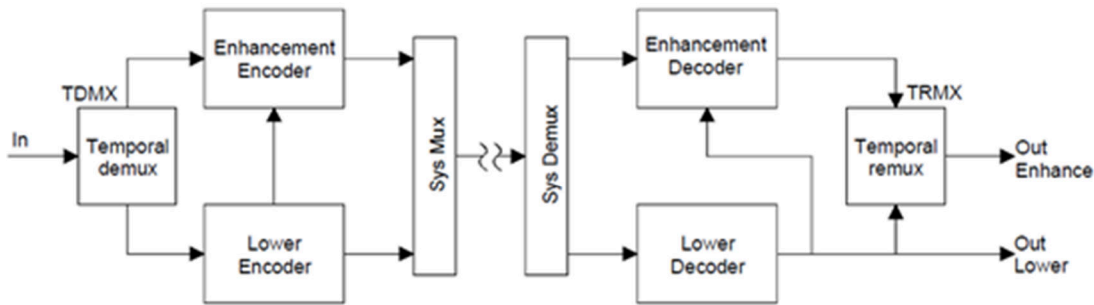


Figure D.1 – A two-layer codec structure for temporal scalability

APPLE-1016, 147

4. *The H.263 Standard*

98. H.263 is a standard developed by ITU-T for videoconferencing applications. The work commenced in 1993, leading to three versions of the standard in 1996, 1998 and 2001. The first version is an evolution of the earlier H.261 standard. The second version, which is incorporated by reference in this declaration in full, included a number of optional annexes that are relevant to the current matter. See, *e.g.*, APPLE-1020.

99. For example, APPLE-1020, Annex O specifies three scalability modes, similar to MPEG2: SNR, spatial and temporal. In brief, section O.1.1 entitled “Temporal Scalability” discloses temporal scalability as follows:

Temporal scalability is achieved using bidirectionally predicted pictures, or B-pictures. B-pictures allow prediction from either or both a previous and subsequent reconstructed picture in the reference layer. This property generally results in improved compression efficiency as compared to that of P-pictures. B-pictures (and the B-part of PB- or Improved PB-frames) are not used as reference pictures for the prediction of any other pictures. This property allows for B-pictures to

be discarded if necessary without adversely affecting any subsequent pictures, thus providing temporal scalability. Figure O.1 illustrates the predictive structure of P- and B-pictures.

100.

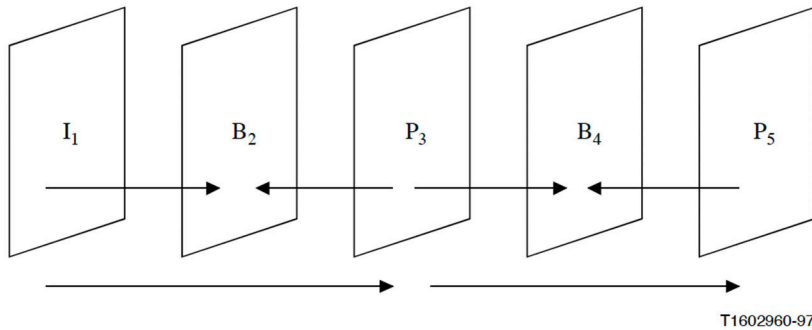


Figure O.1/H.263 – Illustration of B-picture prediction dependencies

APPLE-1020, 102-103; *see also*, APPLE-1020, § 3.4.13.

5. *The H.264 Standard*

101. H.264 is a standard jointly developed by ITU-T for videoconferencing applications and ISO for consumer video applications. (The ISO designation is 14496-10, i.e., MPEG4 Part 10 also known as AVC – Advanced Video Codec, APPLE-1027.) The work commenced in 1998 in the ITU-T workgroup VCEG (Video Coding Experts Group), with MPEG joining in December 2001 to form the Joint Video Team – JVT. The first version was published in May 2003. Scalable Video Coding (SVC) was included from version 3 in November 2007 (ITU-T Rec. H.264 (11/2007), APPLE-1009, which is incorporated by reference in this declaration in full.

102. H.264/AVC provides a significant improvement in video coding

performance over MPEG2 and has achieved substantial worldwide use. Notably it is used for High Definition Television broadcasting (HDTV). The IEEE Transactions on Circuits and Systems for Video Technology published a special issue on scalable coding in H.264 in September 2007.

103. As described in the introduction to the special issue, “The fundamental concept of SVC is to enable the creation of a compressed bit stream that can be rapidly and easily adapted to fit with the bit rate of various transmission channels and with the display capabilities and computational resource constraints of various receivers. This is accomplished by structuring the data of a compressed video bit stream into layers. The base layer is decodable by an ordinary H.264/AVC decoder, while one or more enhancement layers provides improved quality for those decoders that are capable of using it.

104. Three fundamental types of scalability were enabled. The first is temporal scalability, in which the enhancement layer provides an increase of the frame rate of the base layer. To a large extent this was already supported in the original H.264/AVC standard—but new supplemental information has been designed to make such uses more powerful.” Intro to special issue – (page 1099)

105. The Schwarz IEEE paper “Overview of the scalable video coding extension of the H. 264 AVC standard” by three of the principle contributors to the H.264 scalability specification, APPLE-1018, which is incorporated by reference

in this declaration in full, provides a thorough overview of the scalable video coding extension of the H. 264 AVC standard. One new feature in H.264/AVC was that any picture could be used as a reference picture for motion compensated prediction. In other words, B-pictures could now be used to predict other B-pictures. This enabled new prediction structures that support temporal scalability, even beyond those in earlier standards that supported temporal scalability.

106. Like the earlier standards, H.264/AVC supports the three types of scalability – temporal, spatial and quality (SNR).

107. H.264 provides temporal scalability and uses a “temporal layer identifier” to facilitate use. “A bit stream provides temporal scalability when the set of corresponding access units can be partitioned into a temporal enhancement layers with the following property. Let the temporal layers be identified by a temporal layer identifier T , which starts from 0 for the base layer and is increased by 1 from one temporal layer to the next. Then for each natural number k , the bit stream that is obtained by removing all access units of all temporal layers with a temporal layer identifier T greater than k forms another valid bit stream for the given decoder.” H.264, at 1107.

108. Temporal scalability may be implemented using hierarchical B-Pictures. In H.264/AVC B-pictures can be used as reference pictures. Figure 1 of the paper illustrates different prediction structures that enable temporal scalability:

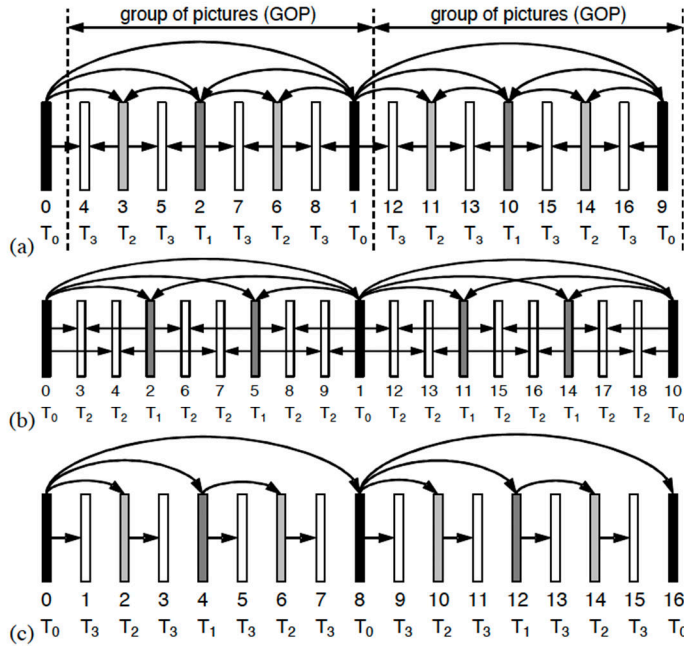


Fig. 1. Hierarchical prediction structures for enabling temporal scalability: (a) coding with hierarchical B pictures, (b) non-dyadic hierarchical prediction structure, (c) hierarchical prediction structure with a structural encoder/decoder delay of zero. The numbers directly below the pictures specify the coding order, the symbols T_k specify the temporal layers with k representing the corresponding temporal layer identifier.

APPLE-1018, FIG. 1

109. Note the pictures are arranged as GOPs in which base-layer pictures demarcate the GOP boundaries.

Temporal scalability with dyadic temporal enhancement layers can be very efficiently provided with the concept of hierarchical B pictures [28][29] as illustrated in Fig. 1a2. The enhancement layer pictures are typically coded as B pictures, where the reference picture lists 0 and 1 are restricted to the temporally preceding and succeeding picture, respectively, with a temporal layer identifier less than the temporal layer identifier of the predicted picture. Each set of temporal layers $\{T_0, \dots, T_k\}$ can be decoded independently of all layers with a temporal layer identifier $T > k$. In the following, the set of pictures between two successive pictures of the temporal base layer together with the

succeeding base layer picture is referred to as a group of pictures (GOP).

Although the described prediction structure with hierarchical B pictures provides temporal scalability and also shows excellent coding efficiency as will be demonstrated later, it represents a special case. In general, hierarchical prediction structures for enabling temporal scalability can always be combined with the multiple reference picture concept of H.264/AVC. This means that the reference picture lists can be constructed by using more than one reference picture, and they can also include pictures with the same temporal level as the picture to be predicted. Furthermore, hierarchical prediction structures are not restricted to the dyadic case. As an example, Fig. 1b illustrates a non-dyadic hierarchical prediction structure, which provides 2 independently decodable sub-sequences with 1/9-th and 1/3-rd of the full frame rate. It should further be noted that it is possible to arbitrarily modify the prediction structure of the temporal base layer, *e.g.*, in order to increase the coding efficiency. The chosen temporal prediction structure does not need to be constant over time.

Note that it is possible to arbitrarily adjust the structural delay between encoding and decoding a picture by restricting motion-compensated prediction from pictures that follow the picture to be predicted in display order. As an example, Fig. 1c shows a hierarchical prediction structure, which does not employ motion-compensated prediction from pictures in the future. Although this structure provides the same degree of temporal scalability as the prediction structure of Fig. 1a, its structural delay is equal to zero compared to 7 pictures for the prediction structure in Fig. 1a. However, such low-delay structures typically decrease coding efficiency.

APPLE-1018.

110. H.264 specifies a mechanism called the Network Abstraction Layer – NAL – for encapsulating the coded video data and associated parametric data into data units that may be stored or transmitted. In temporal scalability, the coded

information for a single frame comprises multiple NALs which collectively form an *access unit*.

3.89 NAL unit: A syntax structure containing an indication of the type of data to follow and *bytes* containing that data in the form of an RBSP interspersed as necessary with *emulation prevention bytes*.
APPLE-1009 clause 3.97

Scalable video coding is specified in Annex G allowing the construction of bitstreams that contain sub-bitstreams that conform to this Specification. For temporal bitstream scalability, i.e., the presence of a sub-bitstream with a smaller temporal sampling rate than the bitstream, complete access units are removed from the bitstream when deriving the sub-bitstream. In this case, high-level syntax and inter prediction reference pictures in the bitstream are constructed accordingly. For spatial and quality bitstream scalability, i.e., the presence of a sub-bitstream with lower spatial resolution or quality than the bitstream, NAL units are removed from the bitstream when deriving the sub-bitstream. In this case, inter-layer prediction, i.e., the prediction of the higher spatial resolution or quality signal by data of the lower spatial resolution or quality signal, is typically used for efficient coding.

APPLE-1009.

111. H.264 further formalized the concept of a reset frame, with the IDR frame (instantaneous decoder reset frame). An IDR frame is An IDR frame is a special type of I-frame in H.264, which is defined such that no frame after the IDR frame can reference any frame before it, and which signals to the decoder that it can completely reset its decoding operation. These were created to add error resilience, since one goal of H.264 was created to facilitate video transmission

over digital channels exhibiting variable conditions. For the purposes of scalability, an IDR frame can be understood simply as an I-frame, since the I-pictures of MPEG-1 and MPEG-2 and H.263 also restarted the prediction process in a GOP. Richardson, at 101, 114, 117.

G. Temporal Scalability in Hybrid Video Coding Was Well Understood by the Time of the '303 Patent

112. Providing temporal scalability via a GOP prediction structure as used in hybrid video coding was well understood by the time of the '303 Patent. Because of the nature of the prediction structure, layers are inherently set in units of GOPs.

113. For example, Tian explained in 2005 that “A known method in today’s streaming systems to cope with drastically dropped channel throughput is to transmit Intra pictures only. When the network throughput is restored, Inter pictures can be transmitted again from the beginning of the next Group of Pictures (GOP).” Tian further explains “Generally, any chain of Inter pictures can be safely disposed, if no other picture is predicted from them. This fact can be utilized to treat Inter pictures at the end of a prediction chain as less important than other Inter pictures. The known layered coding techniques put some pictures into enhancement layers for temporal scalability...” Tian, at 2.

114. Tian further provides Table III which shows temporal frame rates achievable with 3 different GOP structures that are taught on page 2, as well as the

bit rates of those sequences. Tian, at 4. A POSITA would understand that the GOP structures are the “classic” GOP structures described in Richardson, but written in an abbreviated manner.

115. Tian primarily described his proposed technique using numbered layers, but also mentioned enhancement layers: “The known layered coding techniques put some pictures into enhancement layers for temporal scalability.” Tian, at 2.

116. As another example, “Authentication of scalable video streams with low communication overhead” by K. Mokhtarian appeared in IEEE Transactions on Multimedia in 2010 explains “temporal scalability is achieved by employing a hierarchical prediction structure among video frames belonging to the same group-of-pictures (GOP), as shown in Fig. 2. In this structure, frames of higher temporal layers can only be predicted from lower temporal layers. A GOP consists of one frame in the temporal base layer, which is generally coded as P-frame, and several hierarchically coded B-frames that are located between the temporal base layer frames.” Mokhtarian, at 732. In this figure, the GOP structure comprises P- and B- frames, taking advantage of H.264’s more sophisticated prediction structures as mentioned previously.

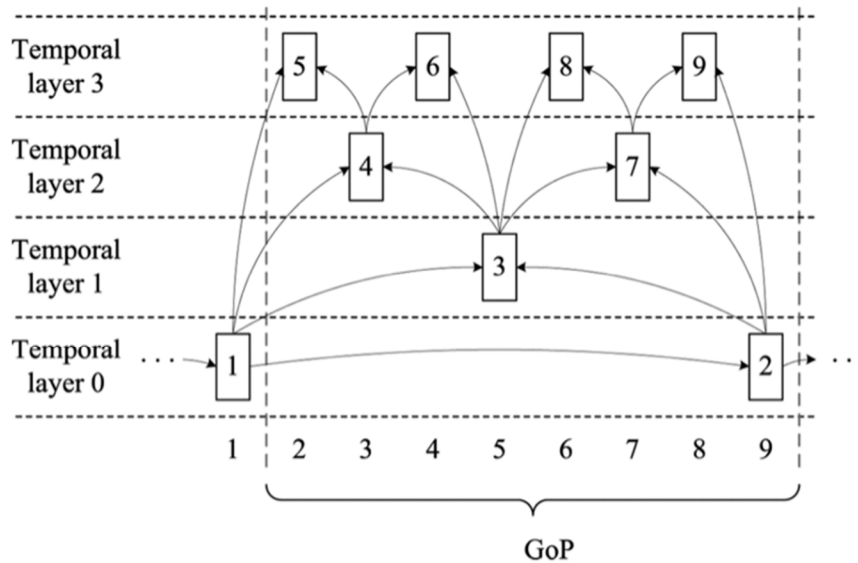


Fig. 2. Hierarchical prediction structure of H.264/SVC temporal scalability. Arrows represent prediction. Numbers listed in the bottom row show the displaying order, while numbers inside frames show the coding order.

Mokhtarian, FIG. 2

117. The paper explains that “Each GOP consists of a number of video frames, each of which belongs to a certain temporal level.” Mokhtarian, at 732, also *id.* Figure 2). Figure 2 uses the equivalent “temporal layer” for the same entity, and later the paper uses that term in “Each encoded stream consists of four temporal layers (GoP size 8).” Mokhtarian, at 739. The paper also uses the terms “base layer” and “enhancement layer” to indicate two layer scalability.

118. As another example, US Patent Application US 2009/0060035 to He, filed Aug 28, 2007 and entitled “Temporal Scalability for Low Delay Scalable Video Coding” (He2009) teaches a GOP of size 8 with a forward-only prediction structure:

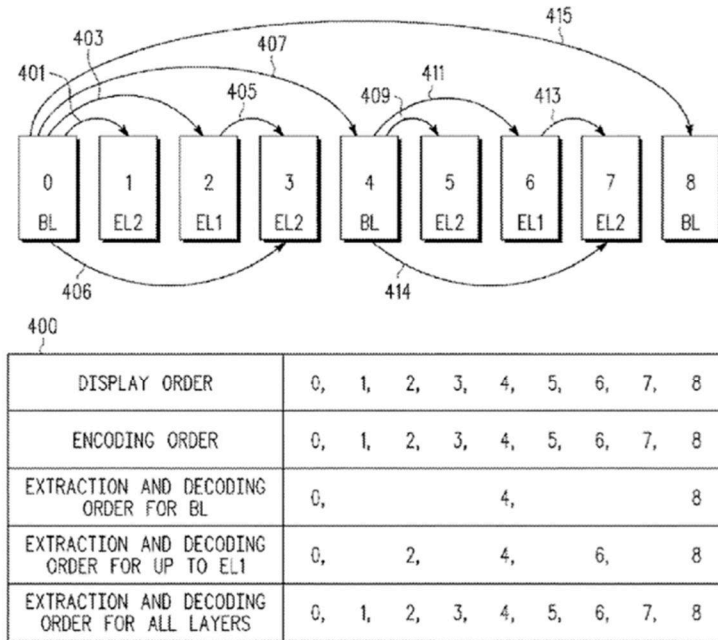


FIG. 4

He2009, FIG. 4

119. He2009's GOP shows 2 frames in the base layer (0 and 4), 2 frames in enhancement layer 1 (2 and 6), and 4 frames in enhancement layer 2 (1, 3, 5, and 7). Figure 4 above shows how quarter-rate, half-rate, and full-rate video are decoded by using the base layer only, the base layer and enhancement layer 1, and all three layers. He2009, [0039]. I note that He2009 uses the term "coding structure" to represent the prediction structure. He2009, [0009-0011]. He2009's prediction structure is selected to support scalable coding as well as the desired low-delay coding: base layer frames are only predicted from base layer frames, while frames at uniform spacings are predicted from frames in the base layer and/or lower enhancement layers.

120. He2009 explains that a GOP includes a base layer frame (BL) and one or more enhanced layer frames (EL). He2009, [0039] The first frame 0 is an IDR frame and is provided before the first GOP. He2009 explains that the prediction structure is not limited to the arrows illustrated in Figure 4; for example, frame 3 could be predicted using frame 2, or frame 0, both accomplishing the goal of low-delay coding. He2009, [0039].

121. As a further example, lecture notes from a course taught at NYU in 2012 illustrate the labeling of temporal layers in a hybrid video coding GOP, and the options for creating 2 temporal layers or 4 temporal layers to provide temporal scalability. These labels are also used in conjunction with terms “base” and “enhancement.” These GOPs illustrate a base layer with both intra- and inter-coded frames, as well as an alternate “layer 0” only having I-frames. Yao Wang slides, citing Amy Reibman.

122. Richardson’s book includes an entire section on temporal scalability in H.264. The term is defined as “The base layer is coded at a low temporal resolution or low frame rate; adding enhancement layers increases the frame rate of the decoded sequence.” Richardson, at 290. Section 10.2.5 is entitled Temporal Scalability and I quote it in its entirety:

In a sequence coded with temporal scalability, the base layer 0 is coded at the lowest temporal resolution, i.e. the lowest frame rate. Successive enhancement layer(s), when decoded with the base layer, provide

progressively higher decoded frame rates. Figure 10.4 shows a sequence encoded as three temporally scalable layers. Layer 0 is coded at a frame rate F_0 and consists of coded frames 0, 6, 12, etc. An H.264/AVC decoder can decode layer 0 in isolation.

Layer 1 consists of frames 3, 9, 15, . . . etc. A decoder may decode layers 0 and 1 to produce a higher-rate sequence (Figure 10.5) at $2F_0$ frames per second. Layer 2 consists of frames 1, 2, 4, 5, 7, 8, ... etc. and a decoder that decodes layers 0, 1 and 2 can deliver an output sequence (Figure 10.5) at $6F_0$ frames per second.

Temporal scalability can be achieved using the P- and/or B-slice coding tools available in H.264/AVC. The examples in Figure 10.4 and Figure 10.5 are developed using the ‘Hierarchical’ or ‘Pyramid’ Group of Pictures structure discussed in Chapter 6 (section 6.4.7). For completeness, Figure 10.6 shows the prediction directions for this hierarchical structure. The base layer consists of the I-slices every 12 slices and the first set of B-slices, B6, etc. Layer 1 consists of the second set of B-slices (3, 9, 15, etc) which are predicted from the base layer. Layer 2 consists of the remaining B-slices which are predicted from layers 0 and 1. Hence the following subsets may be independently decoded:

Layer 0, I0, B6, I12, etc.

Layer 0 + Layer 1, I0, B3, B6, B9, etc.

Layer 0 + Layer 1 + Layer 2, I0, B1, B2, B3, etc.

Because the necessary prediction tools, i.e. P or B pictures used for reference, are supported in the Main and High Profiles of H.264/AVC, temporal scalability can be achieved without the need for any extensions to the core H.264/AVC standard.

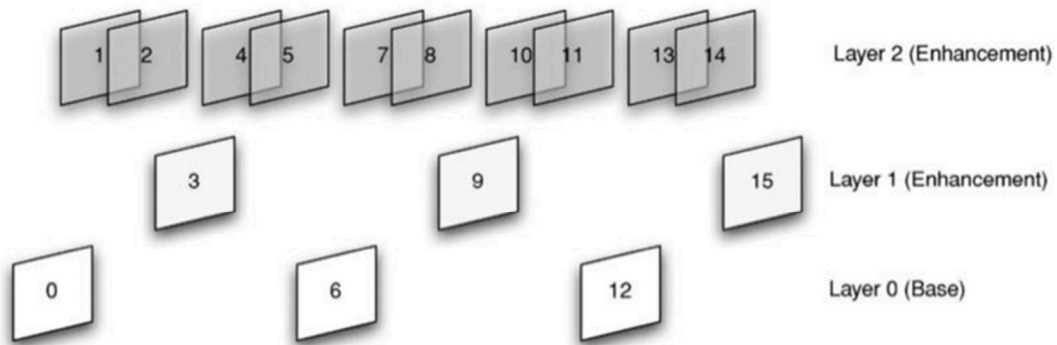


Figure 10.4 Temporally scalable sequence, 3 layers

Richardson, FIG. 10.4

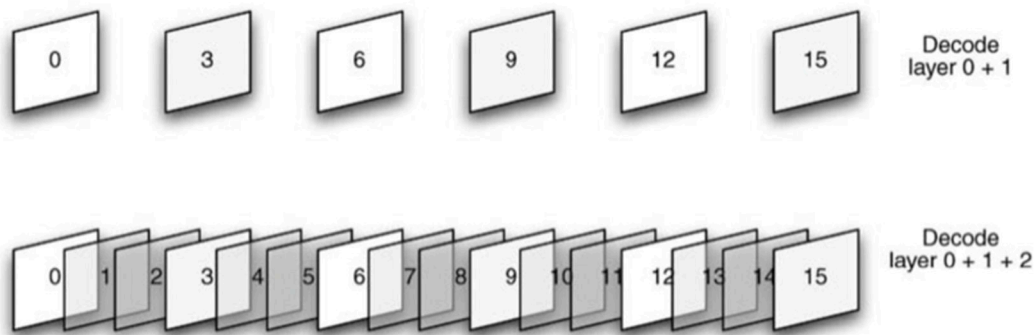


Figure 10.5 Decoding a temporally scalable sequence

Richardson, FIG. 10.5

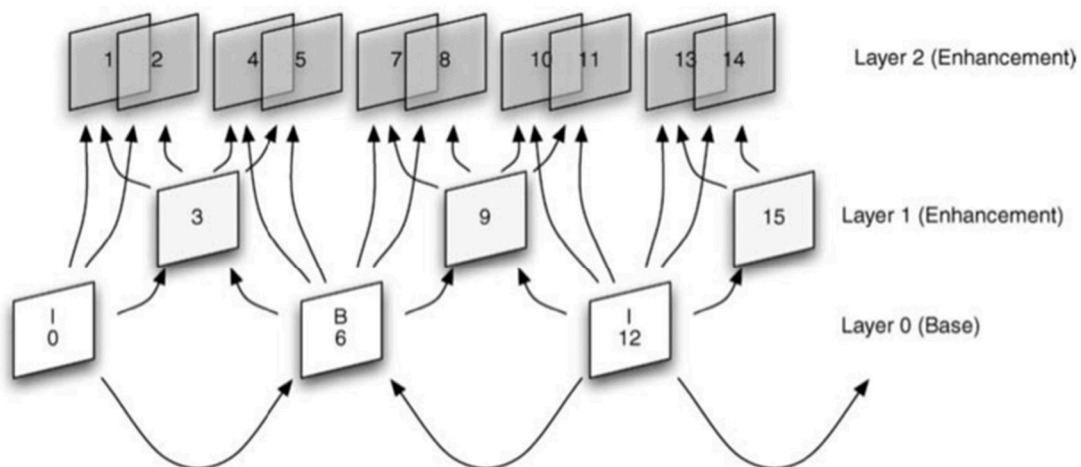


Figure 10.6 Hierarchical prediction structure

Richardson, FIG. 10.6

Richardson, 292-283. I note that here, the language “layer 0,” which includes an I-

frame and a B-frame, is called a “base” layer.

H. Transmission of Scalable Video Was Well Understood at the Time of the '303 Patent

123. Transmission of temporally scalable video obtained via a GOP prediction structure as used in hybrid video coding was well understood by the time of the '303 Patent.

124. An example appears in “Real-time System for adaptive video streaming based on SVC” by Wien et al., which appeared in the aforementioned special issue of the IEEE Transactions on Circuits & Systems for Video Technology in September 2007. It presents a platform for the transmission of scalably coded video, for a variety of usage scenarios including surveillance (which is real-time). An illustration of the suitability of scalable coding to transmission over time-varying channels, with a specific callout to a surveillance application, is also described.

125. Wien describes one example application for transmission of scalably coded video as video surveillance in which different “views” with different constraints, such as quality, are needed. “Again, a single scalable encoding in combination with adaptation serves all needs, avoiding the effort of multiple encoder runs for the same view.” Here, Wien is simply stating that a single run of the encoder can provide all layers needed for the scalable video. It describes different network connection bandwidths, including limited data-rate networks

such as 2.5/3G network. A POSITA would understand that a cellular network could also have a time-varying connection bandwidth. Wien, at 1227.

126. Wien further explains “The platform allows for adaptation of real-time encoded video streams” and further explains how scalably coded video supports such adaptation, as “Bandwidth consuming content like video may be adapted according to the actual bandwidth available.” It further explains that the network bandwidth currently available can be dynamic and hence vary with time. Wien, at 1228.

127. Specifically, for live video, the video being transmitted to a receiver is adapted according to parameters including the receiver’s current bandwidth. The adaptation specifically includes, “e.g., lower quality or frame rate.” Wien specifically states “Temporal scaling is achieved by using the concept of hierarchical B-frames”, which a POSITA would understand to refer to a GOP being coded using a prediction structure that facilitated temporal scalability by discarding bidirectionally predicted frames spaced at uniform intervals within a GOP. Wien, at 1229. A POSITA would also understand that bidirectional prediction was not necessary for hierarchical prediction, and that (unidirectional) prediction could also provide the same scalability.

128. The adaptation is taught to be determined based upon a variety of constraints, including “network capabilities like the maximum bandwidth and, as

an example for a dynamically changing constraint, network status information like the currently available bandwidth or the packet loss ratio;” “optionally also user related constraints like a personal preference indication for temporal resolution versus spatial detail;” and “extractable spatial and temporal resolutions” that are supported by the scalable encoding. Use of temporal scalability alone is therefore envisioned. Wien, at 1230-31.

129. Wien explains that “the constraints from the former group need to be matched to the constraints from the second group.” A POSITA would understand that such matching would take the form of applying a user preference for lower temporal resolution in the case of a change to a lower bandwidth implies transmitting a lower-frame-rate video from the scalable stream. Wien explicitly explains that such matching must be repeated if conditions change: “In case of possibly dynamically changing constraints, this matching process must be repeated during the session. For each update, the process will result in clear decisions for the adaptation process itself, i.e., which NAL units should be sent or dropped.” And a POSITA would understand that frames are described by multiple NAL units, so dropping of frames corresponds to dropping NAL units.

III. Relevant Legal Standards

130. I have been asked to provide my opinions as to whether claims 1-2 of the '303 Patent would have been obvious to a person of ordinary skill in the art as

of the earliest claimed priority date of the '303 Patent (March 31, 2014) ("Critical Date"). The opinions I express in this declaration involve the application of my technical knowledge and experience to the evaluation of certain prior art with respect to the '303 Patent.

131. I am not a lawyer or legal expert, and I am not offering any opinions regarding the law. In forming my opinions, I have relied upon certain legal principles that counsel for Petitioner have explained to me. My understanding of these concepts and principles are summarized below.

132. It is my understanding that, in determining whether claims of the '303 Patent are obvious in this proceeding, the claim terms are generally given their ordinary and customary meaning as understood by a person of ordinary skill in the relevant art. A person of ordinary skill in the art would read the claim terms in the context of the entire patent specification in which they appear, as well as the prosecution history of the patent.

133. It is my understanding that a claim is unpatentable under 35 U.S.C. § 103 if the claimed subject matter as a whole would have been obvious to a person of ordinary skill in the art at the time of the alleged invention. I also understand that an obviousness analysis takes into account the scope and content of the prior art, the differences between the claimed subject matter and the prior art, and the level of ordinary skill in the art at the time of the invention.

134. In determining the scope and content of the prior art, it is my understanding that a reference is considered relevant prior art if it falls within the field of the patent, or a field similar to the patent. In addition, a reference is prior art if it is reasonably pertinent to the particular problem with which the inventor was involved. A reference is reasonably pertinent if the inventor would have considered it when attempting to solve their problem. If a reference relates to the same problem as the claimed invention, that supports use of the reference as prior art in an obviousness analysis.

135. To assess the differences between prior art and the claimed subject matter, it is my understanding that 35 U.S.C. § 103 requires the claimed invention to be considered as a whole. This “as a whole” assessment involves showing that one of ordinary skill in the art at the time of invention, confronted by the same problems as the inventor and with no knowledge of the claimed invention, would have selected the elements from the prior art and combined them in the claimed manner.

136. It is my further understanding that several rationales may be applied for combining references or modifying a reference to show obviousness of claimed subject matter. These rationales include: combining prior art elements according to known methods to yield predictable results; simple substitution of one known element for another to obtain predictable results; a predictable use of prior art

elements according to their established functions; applying a known technique to a known device (method or product) ready for improvement to yield predictable results; choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success; and some teaching, suggestion, or motivation in the prior art that would have led one of ordinary skill to modify a prior art reference or to combine prior art teachings to arrive at the claimed invention.

IV. Person of Ordinary Skill in the Art

137. It is my understanding that when interpreting the claims of the '303 Patent I must do so based on the perspective of one of ordinary skill in the art at the time of the alleged invention ("POSITA"). For purposes of my analysis in this declaration, I have taken the Critical Date (March 31, 2014) as the date of the alleged invention. I understand that the factors considered in determining the ordinary level of skill in a field of art include the level of education and experience of persons working in the field; the types of problems encountered in the field; the teachings of the prior art, and the sophistication of the technology at the time of the alleged invention. I understand that a POSITA is not a specific real individual, but rather is a hypothetical individual having the qualities reflected by the factors above. I understand that a POSITA would also have knowledge from the teachings of the prior art, including the art cited below.

138. Taking these factors into consideration, it is my opinion that one of

ordinary skill in the art in the field of digital video coding as of the Critical Date would have had a bachelor's degree in electrical engineering, computer engineering, or computer science, with 2-3 years of experience in digital video processing and familiarity with standards. Additional graduate education could substitute for professional experience, or additional professional experience could substitute for formal education.

139. I have possessed the qualifications of a POSITA since the Critical Date of the '303 Patent and significantly earlier.

V. Overview of the Subject Patent US Pat. No. 9,986,303

140. The '303 Patent is titled Video Image Coding Data Transmitter, Video Image Coding Data Transmission Method, Video Image Coding Data Receiver, and Video Image Coding Data Transmission and Reception System. It was filed March 13, 2015, and derived from a Japanese application, filed March 31, 2014. '303 Patent, Cover.

A. Overview

1. Field of Art

141. The stated field of art is “video image coding data transmission technology and, more particularly, to a video image coding data transmission technology for transmitting video image coding data in real time.” '303 Patent, 1:17-21. The stated purpose of the '303 Patent is “to provide a video image transmission technology capable of achieving real-time video image transmission

and viewing of high-quality video images at the same time.” ’303 Patent, 1:64-67.

2. Background - Problem Statement, Admitted Prior Art

142. The patent discusses the extensive work on video coding standardization since the 1990s. Two sets of standards are cited:

- MPEG-2 Video (ISO/IEC 13818-2)
- MPEG-4 AVC/H.264 (14496-10 in ISO IEC, H.264 in ITU-T)

3. Summary of the Alleged ’303 Patent Invention

143. In brief, the purpose of the alleged invention is summarized as:

“to provide a video image transmission technology capable of achieving real-time video image transmission and viewing of high-quality video images at the same time.”

’303 Patent, 4:22-42.

144. The ’303 Patent is directed to techniques for transmitting encoded video information across time-varying communications channels and decoding and reconstructing the video information at a receiver. *See, e.g., id.*, 1:18-21, 1:23-30, 5:6-12, 5:19-25.

4. Brief Description

145. The ’303 Patent is directed to techniques for transmitting encoded video information across time-varying communications channels and decoding and reconstructing the video information at a receiver. *See, e.g., id.*, 1:18-21, 1:23-30, 5:6-12, 5:19-25.

146. The ’303 Patent describes the “transmission structure of the video

image.” ’303 Patent, 3:63-64. Video data in the ’303 Patent is assigned to one of two transmission structures based upon the transmission rate: a “basic” structure comprising “basic hierarchical pictures” and a “supplementary” structure comprising “supplementary hierarchical pictures.” APPLE-1001, 3:58-67, 9:27-36; *see also id.*, 4:34-42, 4:59-65, 5:42-44, 8:39-42, 9:57-67, 11:22-36. According to the ’303 Patent, a “basic hierarchy is a hierarchy for transmitting an overview video image, and a supplementary hierarchy is a hierarchy for transmitting an extension video image.” The ’303 Patent explains that the basic hierarchical pictures “refer to only basic hierarchical pictures,” which “allows for decoding in the presence of only basic video image coding data.” ’303 Patent, 9:61-64.

“Meanwhile, allowing supplementary hierarchical pictures to refer to both basic hierarchical pictures and supplementary hierarchical pictures increases the coding efficiency.” *Id.*, 9:64-67. This means that the basic hierarchical pictures of the basic video image coding data are video data that can be decoded on their own to reproduce a video. Claim 1 recites “decoding the basic video image coding data so as to reproduce a video image.” ’303 Patent, claim 1. The supplementary video image coding data, and the supplementary hierarchical pictures included in the supplementary video image coding data, are therefore used for enhancement purposes and rely on the basic video image coding data for decoding.

147. The ’303 Patent always transmits the basic hierarchical pictures, while

only sometimes transmitting the supplementary hierarchical pictures. When the supplementary hierarchical pictures are transmitted, they are delayed.

148. The '303 Patent teaches assigning frames as basic or supplementary using a prediction structure that supports a variety of transmission structures. This assignment is based on the acquired transmission rate. '303 Patent, 4:4-7, 9:49-56. "Prediction structure" refers to inter coding as described in Section II.C, Temporal Redundancy Reduction. "Transmission structure" refers to assigning frames to be basic hierarchical pictures or supplementary hierarchical pictures. The '303 Patent teaches that the ratio of the number of pictures in the basic hierarchical structure to number of pictures in the supplementary hierarchical structure is not limited to those shown but is general, and also that the ratio can be fixed and non-adaptive to transmission rate. '303 Patent, 4:4-7, 9:37-48.

149. The basic structure is coded as basic video image coding data and the supplementary structure is coded as supplementary video image coding data. '303 Patent, 3:58-67, 9:27-36. Importantly, the basic hierarchy can be decoded independent of the supplementary hierarchy to reproduce the video for playback at a basic quality level. *Id.*, 9:57-67. When the receiver later obtains the supplementary data, it can be combined with the basic data to reproduce the coded video stream for playback at a higher quality. *Id.*, 4:50-5:44, 8:39-42, FIGS. 1-4. For example, the '303 Patent describes how supplementary data can be applied to

increase the frame rate of the basic video stream, decrease coding errors in the basic stream through use of higher-frequency error data, or add color to an otherwise monochromatic stream coded in the basic hierarchy. *Id.*, 8:39-42, 14:28-15:27, 15:35-48.

150. At the transmitter, the video image coding unit “generates basic video image coding data that corresponds to the basic hierarchy and supplementary video image coding data that corresponds to the supplementary hierarchy.” ’303 Patent, 4:11-14. The transmitter sends the basic video image coding data to the receiver and “stores the supplementary video image coding data in the memory unit 103.” *Id.*, 4:15-18. The receiver “receives the basic video image coding data transmitted” from the transmitter and “stores the received basic video image coding data in a memory of the coding data reconstruction unit 202.” *Id.*, 4:50-51, 5:8-9. The receiver “decodes the basic video image coding data provided from the receiving unit 200 and outputs the decoded basic video image coding data to terminal 2.” *Id.*, 5:14-18; *see also id.*, 7:30-34 (describing generating a “reproduction video image” from the basic video image coding data and outputting the “reproduction video image from the terminal”). At the transmitter, a “memory unit stores supplementary video image coding data of the supplementary hierarchy.” *Id.*, Abstract. The transmitter later transmits the supplementary video image coding data. *E.g., Id.*, 4:22-33, 4:39-42. To summarize, the transmitter first

transmits the basic video image coding data, which is stored (and in some cases reproduced) at the receiver; and stores the supplementary video image coding data in a memory at the transmitter to be sent later. *Id.*; *see also id.*, 4:40-42.

151. The '303 Patent further explains how the use of basic and supplementary transmission structures can allow basic data to be transmitted in *real-time* during periods of lower channel bandwidth while transmission of supplementary data is *delayed* until channel bandwidth increases sufficiently. *Id.*, 11:38-12:25, FIG. 5. According to this framework, the receiver can decode and reproduce a video stream from the basic video image coding data in real-time at the basic quality level. *Id.* The receiver can then later reproduce the video at a higher quality by reconstructing pictures from both the basic and supplementary hierarchies once the supplementary data is later received. *Id.*

152. According to the '303 Patent specification, the transmitter “belatedly transmits supplementary video image coding data of the past when there is a margin in the transmission rate.” '303 Patent, 4:40-42. The '303 Patent explains that, while the basic video image coding data is “transmitted in real time” by the transmitter, “the supplementary video image coding data is belatedly-transmitted video image coding data of the past stored” at the transmitter. *Id.*, 4:34-42. Therefore, “the supplementary hierarchical picture included in the supplementary video image coding data . . . has a coding order and a display order that precede

those of the basic hierarchical picture included in the basic video image coding data transmitted in real time.” *Id.*, 5:25-30.

153. After the transmitter transmits the supplementary video image coding data, the receiver “receives the supplementary video image coding data” from the transmitter. ’303 Patent, 4:52-53. The receiver “provides the received supplementary video image coding data to the coding data reconstruction unit 202.” *Id.*, 5:10-12. After the receiver receives the supplementary video image coding data, the receiver acquires from memory the basic video image coding data of the past. The receiver “rearranges the order of the pictures between the basic video image coding of the past stored in the memory and the supplementary video image coding data received” by the receiver “so as to reconstruct video image coding data.” *Id.*, 5:19-25. The coding data reconstruction unit “rearranges the order of pictures between the basic video image coding data of the past stored in the memory and the supplementary video image coding data received by the receiving unit 200 so as to reconstruct video image coding data.” *Id.*, 19-23.

154. To summarize, the ’303 Patent describes first sending basic video image coding data with basic hierarchical pictures to a receiver, storing the basic video image coding data in a memory, and in some cases reproducing it. The ’303 Patent then describes sending supplementary video image coding data with supplementary hierarchical pictures to the receiver. The receiver reconstructs

video image coding data using the basic video image coding data of the past stored in memory with newly received supplementary video image coding data.

155. The '303 Patent explains that “a supplementary hierarchical picture included in the supplementary video image coding data transmitted by the second transmission unit 104 has a coding order and a display order that precede those of a basic hierarchical picture included in the basic video image coding data transmitted in real time by the first transmission unit 102.” '303 Patent, 4:34-39.

156. Because “the supplementary hierarchical picture included in the supplementary video image coding data transmitted from the second transmission unit 104 has a coding order and a display order that precede those of the basic hierarchical picture included in the basic video image coding data transmitted in real time from the first transmission unit 102” as described previously, the video image coding data cannot be reconstructed in combination with the supplementary video image coding data “if the basic video image coding data of the past is not stored in the memory.” '303 Patent, 5:25-36. This means that, in the '303 Patent, the basic video image coding data has been received and is already stored in memory by the time the supplementary video image coding data is received.

157. According to the specification, the transmission of supplementary hierarchical pictures in a video sequence can be delayed relative to the transmission of basic hierarchical pictures in the video sequence. *See* '303 Patent,

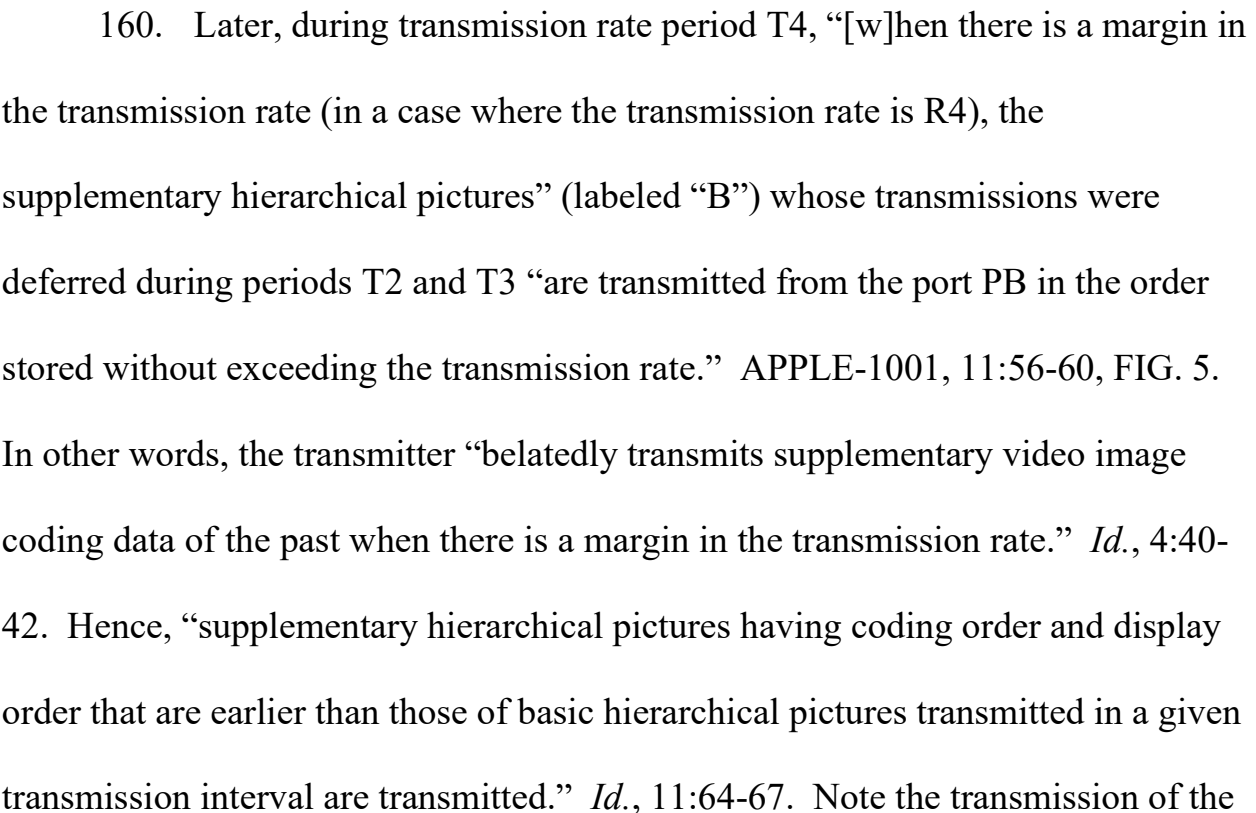
10:25-37. The earlier transmission of basic hierarchical pictures allows the video sequence to be decoded and displayed by the receiver earlier (*e.g.*, in real time) at a lower quality when the channel bandwidth is sufficient to support transmission of only the basic hierarchical pictures, while the delayed transmission of the supplementary hierarchical pictures allows those frames or pictures to be decoded and displayed by the receiver at a later time when the channel bandwidth increases. The combination of (i) the basic hierarchical pictures that was earlier stored at the receiver and (ii) the supplementary hierarchical pictures that are later received then yield higher frame-rate video. *Id.*, 10:25-37, 11:38-12:25.

158. For example, as shown in FIG. 5, “[t]he basic hierarchical pictures” (labeled “A”) “are transmitted in real time from the port PA.” ’303 Patent, 11:55-56; *see also id.*, 9:27-36. “When there is a margin in the transmission rate (in a case where the transmission rate is R4), the supplementary hierarchical pictures are transmitted from the port PB in the order stored without exceeding the transmission rate.” *Id.*, 11:56-60. In this manner, “supplementary hierarchical pictures having coding order and display order that are earlier than those of basic hierarchical pictures transmitted in a given transmission interval are transmitted,” as claimed.

159. To illustrate, FIG. 5 below depicts an example use of basic and supplementary transmission structures in the ’303 Patent for transmitting coded

video data to a receiver over a channel with a time-varying transmission rate or bandwidth. APPLE-1001, 11:38-12:25; *see also id.*, 8:60-10:36, FIG. 4. During transmission rate period T1, the channel bandwidth is sufficient for all pictures in the video stream to be transmitted in real-time and all coded pictures in the stream are therefore allocated to the basic hierarchy (labeled “A” and all shown in the second row labeled “Port PA Transmission”) for real-time transmission. *Id.*, 11:55-56, 9:27-36. During transmission rate periods T2 and T3, the channel bandwidth decreases from R1 to R2 and from R2 to R3, respectively such that the channel is no longer able to support transmission of every picture coded from the input video stream in real time. *Id.*, 11:38-12:25. Accordingly, the system changes the transmission structure of the video stream by allocating a subset of coded pictures in the stream during periods T2 and T3 to a supplementary hierarchy, thereby leaving fewer pictures allocated to the basic hierarchy. *Id.* Only the basic hierarchical pictures (labeled A), and not the supplementary hierarchical pictures, are transmitted during periods T2 and T3 due to limited channel bandwidth. This limited transmission of basic hierarchical pictures, however, still allows the receiver to decode and reproduce the video at a basic quality level from only the basic video coding data. *Id.*

'303 Patent, FIG. 5 (annotated)



odd-numbered frames (“B” frames) 11, 13, 15, 17, and 19 forming the supplementary hierarchical pictures during T2 that correspond to the basic hierarchical pictures comprising the even-numbered frames 10, 12, 14, 16, 18 in T2. This is followed by transmission of frame 21 which is a supplementary hierarchical picture in T3. If the receiver stored the earlier received basic video coding data, the receiver can then use the stored basic and received data together to reconstruct and replay the video stream from periods T2 and T3 at a higher quality (e.g., at a higher frame rate or with additional color depth) at a later time.³ *Id.*

161. The ’303 Patent further discloses that “basic video image coding data (basic hierarchical GOP [group of pictures]) and supplementary video image coding data (supplementary hierarchical GOP) can be set in units of GOPs” rather than being “set in units of pictures.” ’303 Patent, 12:27-36. As described above in Section II.C.5, the use of GOPs for structuring hierarchical video data was a well known technique as early as 1993, and were included in the MPEG standards, as

³ Patent Owner previously argued in IPR2024-00327 that “basic video image coding data corresponds to a higher transmission rate (due to its comprising pictures occupying more bandwidth) than separate supplementary video image coding data.” APPLE-1037, 17-18. These arguments reflected a flawed and incomplete understanding of the ’303 Patent, however, which the Board correctly rejected in its institution decision. APPLE-1040, 6-10, 16-17; *see also* APPLE-1038, 1-4.

well as in the patent literature.

162. The received supplementary video image coding data is synthesized with the stored basic video image coding data of the past by the coding data reconstruction unit 202 so as to reconstruct the video image coding data. APPLE-1001, 5:19-25. Note that after the video image coding data is reconstructed, the “coding data reconstruction unit 202 deletes the basic video image coding data of the past from memory.” *Id.*, 5:35-38.

163. The '303 Patent describes a crime prevention system as an example use case that demonstrates the concepts of basic and supplemental hierarchical pictures. APPLE-1001, 12:12-25. The '303 Patent describes “reducing a transmission quantity by transmitting only basic hierarchical pictures in a 3G/LTE area having a band limitation and transmitting supplementary hierarchical pictures in a WiFi area having no band limitation.” *Id.* Thus, an overview video image can be reproduced even in an environment with a band limitation by transmitting basic hierarchical pictures, and a video image with a high frame rate can be reproduced in an environment with no band limitation by transmitting the supplementary hierarchical pictures. *Id.* As shown in Figure 5 above, the supplementary hierarchical pictures that correspond to already-received basic hierarchical pictures are transmitted when there is a higher transmission rate, along with new basic hierarchical pictures, so that the video can be viewed in real-time at a lower quality

and on rewind at higher quality. *E.g., id.*, 12:6-11.

5. *Comments*

164. The '303 Patent is directed to real-time transmission, reception, and decoding of digital video, and more specifically to transmission of scalably coded video over a channel whose bandwidth varies with time, in which the scalability of the video enables adaptation to time-varying bandwidth real-time transmission of a low-quality video and delayed transmission of high-quality video. The '303 specification describes an application in the context of a video surveillance system, in which a monitoring station receives and decodes video in low quality in real time, and receives and decodes high-quality video corresponding to previously viewed low-quality video (i.e., high quality video from the past) at a later time. “Quality” corresponds to frame rate in the first embodiment which teaches temporal scalability; to frame size in the second embodiment which teaches spatial scalability; or to an increased level of detail in the third embodiment which teaches SNR scalability (with a variation in which color information is offered as an enhancement layer to a gray-scale base layer).

165. The claims both refer to “basic hierarchical pictures” and “supplemental hierarchical pictures” which are the terms for the base layer and enhancement layer given in embodiment 1, which provides temporal scalability. Further discussion will focus on Embodiment 1 because it is reflected in the

claims.

166. The '303 specification describes setting a transmission structure, which corresponds to assigning frames to be in the base layer (a frame in the base layer is called a “basic hierarchical picture”) or in a single enhancement layer (a “supplementary hierarchical picture”). In other words, the '303 Patent employs temporally scalable coding with one enhancement layer, and the basic hierarchical pictures determine the frame rate of the lower quality video. The '303 specification describes different prediction structures for the basic and the supplementary data '303 Patent, 9:57-67, which repeat the well-known understanding that predictive coding for frames in the base layer must be restricted to be from reference frames that are also in the base layer, while predictive coding for frames in the enhancement layer can use frames from the base layer, from the enhancement layer, or both. As previously described, such restrictions were well recognized to be requirements of scalable video coding.

167. The '303 Patent explains that the proportion of frames in the base layer is changed depending upon the available bandwidth, toward ensuring that the base layer can be transmitted in real time even when the bandwidth (“transmission rate”) is small. '303 Patent, 9:49-56, 10:16-24. A POSITA would recognize this as bit rate adaptation, which was enabled easily by scalable coding. The '303 Patent also offers an alternative in which the proportion of frames in the base layer is

fixed independent of bandwidth. '303 Patent, 4:5-7. A POSITA would recognize this as a simpler implementation if bandwidth variations are limited, or for higher reliability channels.

168. The '303 specification provides examples in Figure 4 of one general prediction structure for a GOP, comprising an I-frame followed by P-frames, while teaching that B frames may also be used instead of P frames. '303 Patent, 9:25-26. Within those examples, it also provides examples of two GOP lengths (20 and 10). While the term GOP does not appear in the specification, it does appear in the claims, where it is defined generally as “including an intra coded picture and a plurality of inter prediction coded pictures.” '303 Patent, 18:48-50, 19:9-10. A POSITA would understand that the '303 purported invention was not limited to the GOPs shown in Figure 4 but rather encompassed a broad range of prediction structures. I remind the reader that the video compression standards do not require any specific prediction structure.

169. A purported invention of the '303 Patent is delayed transmission of the enhancement layer. The '303 specification explains that full-frame-rate video is transmitted when the transmission bandwidth can accommodate it, while lower-frame-rate video is transmitted when the bandwidth cannot accommodate full-frame-rate video. Lower-rate video is achieved by discarding frames, as illustrated by Figure 5. In transmission periods T1 and T4, in which the bandwidth

is described as sufficiently high, all of the frames are labeled “A” and transmitted. In transmission period T2, every other frame is transmitted, resulting in half-rate video. In T3, every third frame is transmitted, resulting in 1/3-rate video.

170. The frames that are not transmitted in real-time correspond to the enhancement layer, and are stored for later transmission. This later transmission is illustrated in Figure 5 in transmission period T4. T4 is described as having excess bandwidth available, such that it can transmit not only the current base layer but also the stored enhancement layers corresponding to earlier periods. Figure 5 shows that the frames labeled B from the transmission periods T2 and T3 (i.e., the enhancement layers corresponding to the base layers that were transmitted in periods T2 and T3) are transmitted after the full-rate video has been transmitted in T4. A POSITA would understand that the GOP prediction structure as far as as MPEG-1 could accommodate this type of temporal scalability simply by labeling frames as corresponding to the base layer and the enhancement layer, based upon the desired frame rate of the low-rate video.

171. Nothing in the teachings of the '303 Patent summarized above is new. Scalable coding was well known at the time and the '303 Patent simply describes an application case. The prior art teaches first receiving basic video image coding data and storing it, and then, later, receiving supplementary video image coding data, and reconstructing a video using stored basic video image coding data with

received supplementary video coding data. The prior art teaches that it was well known to structure hierarchical video image coding data – scalable video data – into GOPs, with each GOP having an intra-coded picture and a plurality of inter-prediction coded pictures.

172. Furthermore, the use of groups of pictures to structure the basic hierarchical pictures and supplementary hierarchical pictures was not new, as evidenced by the prior art discussed in this declaration, including prior art used to invalidate the challenged claim and the prior art discussed as background art. For example, Demircin teaches structuring base layer data and enhancement later data in groups of pictures that include one intra-coded picture and a plurality of inter-prediction coded pictures. *E.g.*, Demircin, FIG. 1. I discuss Demircin at length in the analysis sections of the claims.

173. In another example, Chen, filed in the United States on March 5, 2009, makes it clear that the use of GOPs in video coding was commonplace at the time. First, Chen teaches that “known MPEG/H.26x video coding standards (*e.g.* MPEG-1, MPEG-2, MPEG-4, MPEG-4 AVC/H.264, H.263, VC-1), there are basically three types of pictures: I (intraframe coded) pictures, P (interframe coded) pictures and B (bi-directionally predicted) pictures.” Chen, 1:20-24. According to Chen, “[v]ideo sequences are usually coded in a group of picture (GOP) structure.” *Id.*, 1:33-35 (emphasis added).

174. Not only were GOPs commonplace, the “typical MPEG GOP structure” included an “intra encoded picture and more than two predicted encoded pictures.” *E.g.*, Chen, 1:12-16 (emphasis added). FIG. 1 of Chen illustrates the “typical MPEG GOP structure” that includes an I picture and a plurality of P and B pictures:

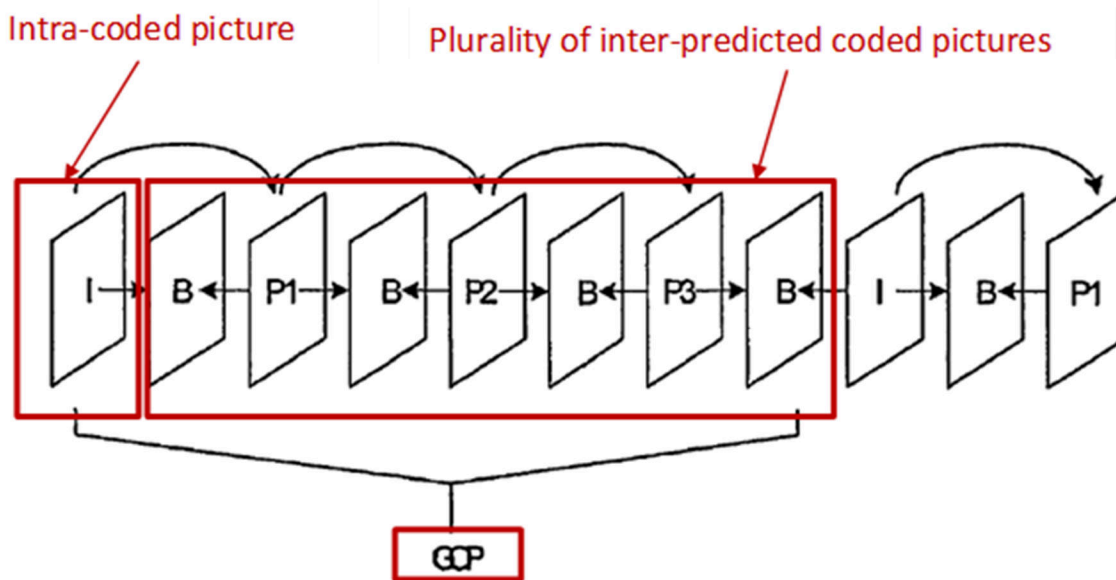


Fig.1

Chen, FIG. 1 (annotated), 4:40

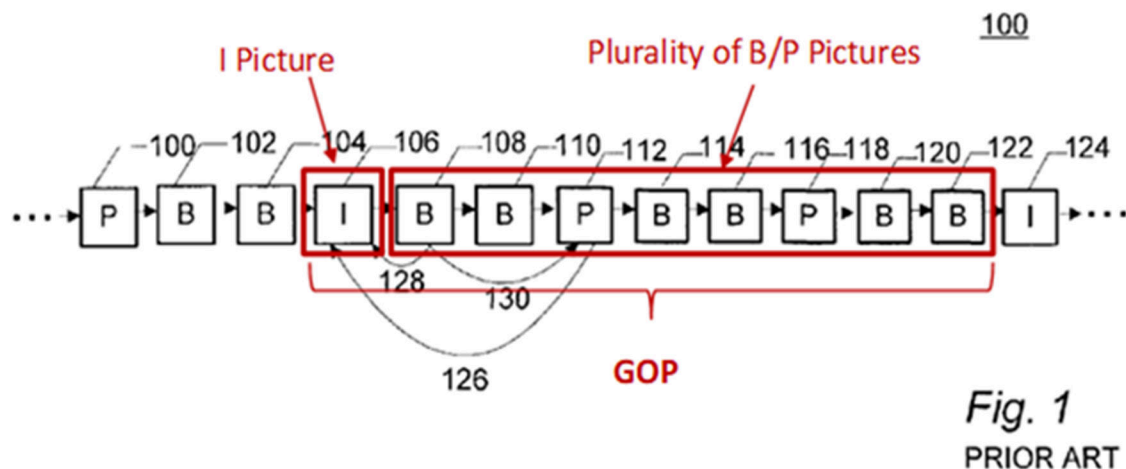
175. As shown in Chen, it was “typical” in MPEG video coding to structure an I picture and a plurality of B/P pictures into a GOP, as shown above.

Id.

176. Luthra, filed on June 3, 2003, is another example of prior art that performs video coding using a GOP with an I picture and a plurality of B/P pictures. Luthra teaches, as early as 2003:

Video content typically comprises one or more groups of pictures, each of which are a collection of pictures comprised of one or more macroblocks. Each of the pictures in the group of pictures are either I (intra-coded) pictures, B (bi-directional) pictures, or P (predicted) pictures.

177. Luthra, 1:27-31 (emphasis added). As confirmed by Luthra, an I picture is “intra-coded” and B and P pictures are reference pictures that rely on other pictures for coding and decoding. *Id.*, 1:35-59. Luthra illustrates a prior art group of pictures as including an I picture and a plurality of B pictures and P pictures. *Id.*, FIG. 1:



Luthra, FIG. 1 (annotated)

178. “FIG. 1 is a simplified block diagram illustrating a group of pictures under the MPEG 2 standard.” Luthra, 1:35-38. As described in Luthra, “each individual group of pictures is defined by the collection of pictures beginning at a first I frame and ending at the frame previous to the next I frame.” Luthra, 1:31-

34. Thus, in Luthra, the prior art GOP included an intra-coded picture and a plurality of inter-prediction coded pictures.

179. In addition to structuring video coding data into GOPs having I pictures and a plurality of B/P pictures, it was known that the size of the GOP was adjustable. For example, Ansley, filed on June 8, 2011, teaches the ability to adjust the group of picture size. *E.g.*, Ansley, Abstract. FIG. 2 of Ansley clearly shows that it was known at the time that the size of the group of pictures was not only a well know construct, but that it was also know that the size of the group of pictures could be selected to suit different circumstances or implementation choices.

180. As explained earlier, GOPs are multifunctional and well adapted to the requirements of specific applications. When first conceived, GOPs facilitated trick modes, such as fast-forward and reverse play. Subsequently, for broadcast TV, GOPs facilitate channel surfing. GOPs also provide error resilience for use in error-prone environments.

181. Prior to the priority date of the '303 Patent, there existed numerous publications and patents that described the functionality claimed in the '303 Patent. Thus, the methodology of the '303 Patent was well-known in the prior art. Further, to the extent there was any problem to be solved in the '303 Patent, it had already been solved in the prior art systems before the priority date of the '303 Patent.

B. Claims

182. The '303 Patent includes two claims: the first directed to a video image coding data receiver and the second to a video image coding transmission and reception system. Claims 1 and 2 of the '303 Patent are as follows:

1. A video image coding data receiver comprising:

a processor; and

a memory unit having instructions stored which, when executed by the processor, cause the processor to perform operations comprising:

receiving basic video image coding data;

decoding the received basic video image coding data so as to reproduce a video image;

receiving supplementary video image coding data including a supplementary hierarchical picture whose coding order and display order are earlier by a factor of a group of pictures including an intra coded picture and a plurality of inter prediction coded pictures than those of a basic hierarchical picture included in the basic video image coding data, a basic hierarchy and a supplementary hierarchy being set in units of the group of pictures;

acquiring basic video image coding data received before supplementary video image coding data that has been received at the moment; and

reconstructing video image coding data from the basic video image coding data and the supplementary video image coding data.

2. A video image coding data transmission and reception system comprising:

a video image coding data transmitter that codes and transmits a video image; and

a video image coding data receiver that receives and reconstructs video image coding data,

wherein the video image coding data transmitter includes:

a first processor; and

a first memory unit having instructions stored which,

when executed by the first processor, cause the first processor to perform operations comprising:

acquiring the transmission rate of a network;

setting a transmission structure including a basic hierarchy and a supplementary hierarchy, the basic hierarchy and the supplementary hierarchy being set in units of a group of pictures including an intra coded picture and a plurality of inter prediction coded pictures;

transmitting basic video image coding data of the basic hierarchy, wherein the first memory unit is further configured to store supplementary video image coding data of the supplementary hierarchy;

transmitting the supplementary video image coding data stored in the memory unit; and

controlling the transmission of the supplementary video image coding data according to the transmission rate, and

wherein the video image coding data receiver includes:

a second processor; and

a second memory unit having instructions stored which, when executed by the second processor, cause the second processor to perform operations comprising:

receiving the basic video image coding data;
decoding the received basic video image coding data so as to reproduce a video image;
receiving the supplementary video image coding data including a supplementary hierarchical picture having coding order and display order that are earlier by a factor of the group of pictures than those of a basic hierarchical picture included in the basic video image coding data;
acquiring basic video image coding data received before supplementary video image coding data that is currently received; and
reconstructing video image coding data from the basic video image coding data and the supplementary video image coding data.

C. Prosecution History of the '303 Patent

183. I have reviewed excerpts from the Prosecution History of the '303 Patent, which issued from U.S. Patent Serial No. 14/657,359, which was filed on March 13, 2015.

184. The application resulting in the '303 Patent was filed on March 13, 2015, by JVC Kenwood Corporation. '303 Patent, Face. The '303 Patent claims the benefit of foreign priority to JP 2014-0071975, dated March 31, 2014. The '303 Patent issued on May 29, 2018. For the purposes of my analysis, I assume that the priority date of March 31, 2014 is the so-called "Critical Date" for the '303 Patent. The prior art I discuss for invalidity and background all predate the Critical Date of the '303 Patent.

185. The Applicant of the 14/657,359 patent application originally

submitted 11 claims, 5 of which were independent claims. The claims underwent 4 rejections, each set of rejections compelling the Applicant to make amendments to the claims (or to cancel claims). The '303 Patent includes two independent claims: claim 1 was originally claim 8 and is generally directed to a video image coding data receiver; claim 2 was originally claim 11 and is generally directed to a video image coding data transmission and reception system.

186. During examination of the application, the Examiner first found that Radha (U.S. Pat. 6,292,512) taught both base layer (basic hierarchy) and enhancement layer (supplementary hierarchy). The examiner then rejected claim 8 as obvious in view of Radha and Van Der Schaar (U.S. Pat. 6,697,426). APPLE-1002, 139-140, 146-147, 188-189. Regarding a limitation in claim 8 that recited similar language to that now recited in elements [1.5.1]-[1.5.2], the Examiner noted that Van Der Schaar discloses an enhancement (supplemental) picture E2 having a coding order and display order earlier than a basic picture P3. *Id.*, 102. In response, the Applicant argued that claim 8 required “that the supplementary hierarchical picture whose coding order and display order are earlier than those of a basic hierarchical picture included in the basic video image coding data.” APPLE-1002, 133-134. The Applicant sought to distinguish Van Der Schaar on the basis that it only described “rearranging the transmission order of the enhancement frames ‘to be the same as the display order of the enhancement

frames on the decoder side.’” *Id.*

187. The Examiner subsequently rejected claim 8 as obvious based on Radha (U.S. Pat. 6,292,512) in view of Chen (U.S. Pat. 5,886,736). *Id.*, 93-95. In response, the Applicant amended claim 8 to require that the coding order and display order are earlier “by a factor of a group of pictures including an intra coded picture and a plurality of inter prediction coded pictures” than those of a basic hierarchical picture included in the basic video image coding data, “the basic hierarchy and a supplemental hierarchy being set in units of the groups of pictures.” *Id.*, 66. Following this amendment, the Examiner allowed claim 8. *Id.*, 51.

1. Comments

188. Throughout the prosecution of the ’303 Patent application, the examiner found that the prior art taught basic hierarchy as a “base layer” and supplementary hierarchy as an “enhancement layer.” *E.g.*, APPLE-1002, 69-70, 103-04, 138-39, 174-75. As discussed below, Demircin also uses the terms “base layer” and “enhancement layer” to refer to the hierarchical pictures. *E.g.*, Demircin [0035]. As evidenced by the examiner’s rejections of the ’303 Patent claims and application of the prior art, the examiner concluded, and I agree, that the terms “base layer” and “basic hierarchy” have the same meaning; and the terms “enhancement layer” and “supplemental hierarchy” have the same meaning.

189. Kimoto uses the terms “lower layer” and “upper layer” to describe the hierarchical structure of the content data. *E.g.*, Kimoto [0019]. During prosecution of Kimoto, the examiner interpreted the term “lower layer” to mean “base quality layer” and upper layer to include one or more “quality layers.” *E.g.*, Kimoto File History at NFOA, 5-6. As noted by the cited prior art from Kimoto, a scalable multi-layer encrypted stream includes “a low quality base layer video (encrypted or non-encrypted)” and “enhancement layer data via different quality layers, each corresponding to a service level.” Zhu, [0026]-[0027]. In addition, the synthesis of lower layer data with upper layer data in Kimoto results in high quality data. *E.g.*, Kimoto [0065]. Thus, the examiner concluded, and I agree, that the lower layer data is base layer data, and the upper layer data is enhancement layer data.

190. As evidenced by the Kimoto File History, the lower layer is the same as a base quality layer, and the upper layer is the same as an enhancement layer. In short, the base layer of Demircin and the lower layer of Kimoto read onto the basic hierarchical coding data of the ’303 Patent; and the enhancement layer of Demircin and the upper layer of Kimoto read onto the supplementary video image coding data.

191. According to the file history of the ’303 Patent, the prior art discussed and analyzed in this declaration were not considered by the examiner when

rejecting the claims. If Demircin and Kimoto had been considered, as I have explained here, the '303 Patent should not have issued.

D. IPR History

192. I understand that the '303 Patent was previously the subject of *inter partes* review (IPR2024-00327) that terminated soon after institution due to settlement. IPR2024-00327-18 (Sept. 20, 2024). *See* APPLE-1036, 1040. The petitioner in the -00327 proceeding was Samsung Electronics Co., Ltd. *Id.* Apple was not a party to the -00327 proceeding. I understand that Samsung's petition in the -00327 proceeding challenged claims 1 and 2 of the '303 Patent as obvious based on Demircin and challenged claim 1 as obvious based on Kimoto in view of Demircin. *Id.* (The present Grounds also include obviousness under Demircin in view of Kimoto.) The Board's institution decision found a reasonable likelihood that Demircin renders claim 1 obvious sufficient to support institution. APPLE-1040, 17. The institution decision did not address the remaining challenges in Samsung's petition. *Id.*

193. IPR2024-00327 involved Demircin as a first obviousness ground and Kimoto in view of Demircin as a second obviousness ground. I agree that the teachings of Demircin and the teachings of Kimoto in view of Demircin render the claimed subject matter of the '303 Patent obvious. Grounds 1A and 2 of this declaration, which I address further below, build on substantially the same

obviousness theories presented in Samsung's instituted IPR2024-00327.

VI. Summary of Prior Art

194. The following pieces of prior art render obvious the alleged invention of the '303 Patent in the combinations addressed further below (Grounds 1-2).

A. Overview of Demircin

195. Demircin (US2008/0181302) is a US Patent Application publication entitled "Methods and Systems for Rate-adaptive Transmission of Video," filed January 25, 2007 by M. Demircin et al. ("Demircin"). It incorporates by reference US patent application 11/479077 entitled "Wireless Video Transmission System," filed June 30, 2006 by M. Demircin et al ("Demircin916").

196. Demircin and Demircin916 together disclose a sophisticated, highly detailed system and method for transmitting scalably-coded video over channels with time-varying bandwidth. The Demircin and Demircin916 references provide approaches for estimating transmission channel bandwidth, selecting scalably-coded data to transmit in a given time period over the available bandwidth and transmitting that data, and opportunistically making use of newly available channel bandwidth to transmit data that was previously not transmitted. These approaches are within the context of real-time video transmission, in which data for coded video must be available for decoding and reproduction by strict deadlines.

197. Demircin, like the '303 Patent, describes techniques for transmitting

coded “audiovisual (AV) data from a source device (e.g., server) to one or more destination devices (e.g., clients) over a packet network.” Demircin, [0023].

Demircin explains that “high-quality AV data streams require a very high bandwidth continuously,” but “time-varying channel conditions” can make the available bandwidth of a communication channel “unpredictable” and can lead to degradation in video quality when the bandwidth decreases. Demircin, [0024].

198. Demircin describes techniques for transmitting coded “audiovisual (AV) data from a source device (e.g., server) to one or more destination devices (e.g., clients) over a packet network.” Demircin, [0023]. Demircin explains that “[a] network linking a video transmitter to a receiver can often be considered a transmission channel with time-varying channel conditions” in which the bandwidth “may vary over time and can be unpredictable” due to various factors. Demircin, [0024]. “Transmission of compressed AV data over such networks is challenging because high-quality AV data streams require a very high bandwidth continuously, and impose stringent delay constraints on delivery of AV data,” and “[f]urthermore, a small number of lost or late packets may have a large effect on video quality seen at the receiver, due to the dependencies between parts of the compressed AV stream.” *Id.*

199. Demircin proposes to address these challenges in the context of a Scalable Video Coding (SVC) system. Demircin, [0027]. SVC was well-known

by the Critical Date and had been defined in standards since at least MPEG-2. MPEG2 (APPLE-1016), 10, 17, 109 (describing temporal scalability). Demircin provides a basic overview of SVC and explains how “SVC enables fine-grained adaptation of a bit stream in spatial, temporal and SNR dimensions.” Demircin, [0026]. “SVC enables coding video in a scalable manner, resulting in a scalable bit stream, for example by coding the video data in several layers.” *Id.* “[T]he base layer provides a minimum level of video quality,” while the “enhancement layers provide additional quality in an incremental manner.” Demircin, [0035]. “Decoding of the first enhancement layer requires availability of the base layer.” *Id.* “Decoding of each subsequent enhancement layer requires availability of all previous enhancement layers, as well as the base layer.” *Id.* Figure 1 (reproduced below) “illustrates an exemplary prediction structure, GOP structure and layering structure of a sample SVC bit stream” that includes a base layer 10 and enhancement layers 11, 12. *Id.*, [0036].

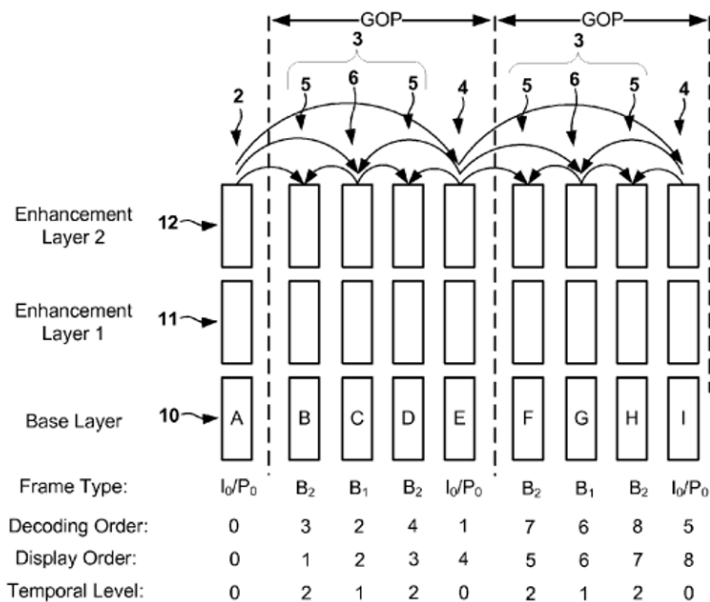


FIG. 1

Demircin, FIG. 1

200. SVC defines different forms of scalability that can be used individually or in combination. These forms of scalability include temporal scalability for scaling the frame rate, spatial scalability for scaling the resolution of video frames, and signal-to-noise ratio (SNR) scalability for scaling the visual quality of video frames. Demircin, [0026]. “[T]he base layer provides a minimum level of video quality,” while the “enhancement layers provide additional quality in an incremental manner, also referred to as progressive refinement.” *Id.*, [0035]. “Decoding of the first enhancement layer requires availability of the base layer.” *Id.* “Decoding of each subsequent enhancement layer requires availability of all previous enhancement layers, as well as the base layer.” *Id.*

201. On the other hand, temporal scalability “utilizes the concept of a

temporal level, which indicates the level of a video frame in a hierarchical temporal prediction structure.” *Id.*, [0035]-[0036] (“Temporal scalability is achieved using an IBP prediction structure with hierarchical B-frames.”).

Demircin also explains temporal scalability in the context of FIG. 1.

FIG. 1 illustrates an exemplary prediction structure, GOP structure and layering structure of a sample SVC bit stream. Temporal scalability is achieved using an IBP prediction structure with hierarchical B-frames. This structure uses I-frames 2, which are intra-predicted from within the frame; B-frames 3, which may be predicted from one or more other frames; and P-frames 4 which are predicted from another I-frame. A hierarchical B-frame 5 may be predicted from other B-frames as well as I-frames and P-frames while a standard B-frame 6 may only be predicted from I-frames and P-frames. In this example, the GOP size is equal to 4 frames. The first frame of the GOP (in encoding/decoding order) is a key picture 2, 4. The key pictures are either intra-coded (I-frame) 2 or inter-coded using a previous (key) picture (P-frame) 4 as a reference for motion-compensated prediction. Key pictures may be considered a base layer for temporal scalability and may be assigned temporal level 0. The remaining pictures of a GOP are hierarchically predicted as illustrated by arrows in FIG. 1. These non-key pictures form successive temporal levels for temporal scalability, each layer incrementing the temporal level by 1. In some embodiments, the temporal level may be a function of the number of frames through which a current frame is predicted. For example, frame A is intra-predicted and is not predicted from any other frame, accordingly, its temporal level is 0. As another example, frame C is predicted only from key frames, which have a temporal level of 0. Accordingly, frame C has a temporal level of 1 indicating that it has been predicted through one key frame. As yet another example, frame B is predicted from frame C, which was predicted from frame A. Since frame B has been predicted through two other frames successively, it is assigned a temporal level of 2.

Demircin, [0036].

202. To illustrate the terminology and explanation in [0036], I have modified FIG. 1 to focus on how the prediction structure supports temporal scalability:

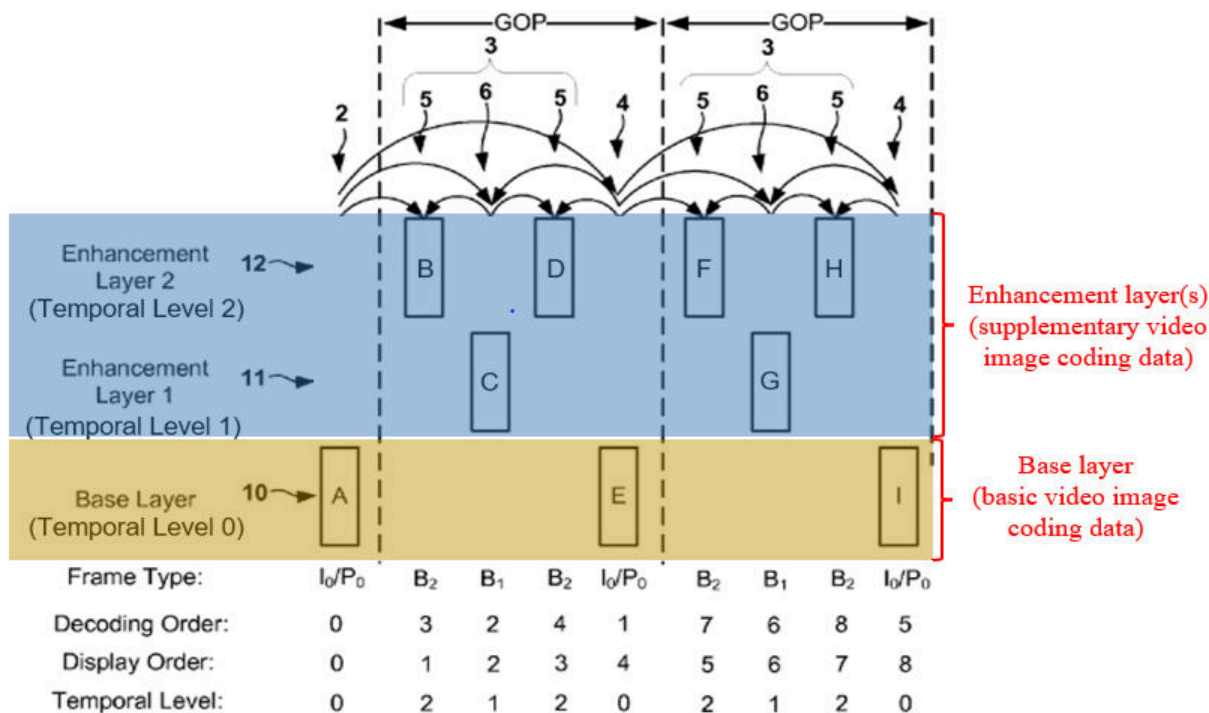


FIG. 1

Demircin, FIG. 1 (modified annotated⁴)

⁴ The prediction structure shown in FIG. 1 is merely exemplary, and a POSITA would have known that many different prediction structures were known and commonly employed in the art before the Critical Date. For example, FIG. 1 depicts a prediction structure with hierarchical B-frames, but a POSITA would have known (and Demircin expressly acknowledges) that P-frames could be used in place of B-frames at higher temporal levels. Demircin, [0036], [0039], [0043],

203. The temporal levels are indicated in the last line of labels below the frame types, and relate to numbers of prediction dependencies. In the example in FIG. 1, frames in display order 0, 4, and 8 are explained to be key pictures, and either I or P frames, which are either intra-coded (I-frames) or predicted from other key pictures (P-frames), and are explained to be a “base layer for temporal scalability.” These frames are assigned a temporal level of 0. Other frames are assigned temporal levels in increasing order based upon the prediction structure, and as such the frames in each non-zero temporal level n comprise enhancement layer n . Demircin, [0036] [0095].

204. The temporally scalable structure shown in FIG. 1 consists of (i) I-frames at the lowest temporal level (e.g., level 0) that are intra-predicted without reference to other pictures and (ii) B-frames at higher temporal levels (e.g., levels 1

[0095] (“a P-frame 164[] ... may be assigned a temporal level value of 1”); Demircin916, FIGS. 4-5 (depicting additional prediction structures and GOP sizes). Demircin’s FIG. 1 also depicts “open” GOPs in which pictures are predicted with reference to decoded picture in a prior GOP. But a POSITA would have appreciated that Demircin’s techniques are equally suitable for “closed” GOPs that do not reference pictures from other GOPs. APPLE-1050, 21, 46, 104. Both open and closed GOPs were permissible in MPEG-based video coders and it would have been obvious to a POSITA to use closed GOPs to create splice points in a video sequence, e.g., for scene changes and to provide random access points at which a user could choose to begin playing the video. Cf. APPLE-1001, 12:37-50.

or 2) that are inter-predicted with reference to other frames at a lower temporal level. Demircin, [0036], FIG. 1. In the example shown in Figure 1, “the GOP size is equal to 4 frames.” *Id.* The frames at lower temporal level(s) (e.g., the I-frames at level 0) constitute a “base” layer while the frames at higher temporal level(s) constitute “enhancement” layers in the context of temporal scalability (e.g., B-frames at levels 1 and 2). Demircin, [0036] (“The key pictures are [] intra-coded (I-frame 2 Key pictures may be considered a base layer for temporal scalability and may be assigned temporal level 0.”). The collection of temporal levels could be $\{(0), (1\ 2)\}$ or $\{(0\ 1), (2)\}$, when using Demircin’s length-4 GOP depicted in FIG. 1.

205. Temporal scalability is achieved by a subset of layers. In the example prediction structure shown in FIG. 1, transmission of the base layer for temporal scalability (i.e., frames with a temporal level of 0) results in the transmission of every 4th frame, resulting in decoded video at quarter-frame-rate (e.g., if the original video was at 60 frames/sec, the decoded base layer would result in video at 15 frames/sec. Practically speaking, each decoded frame would be displayed 4 times, yielding a visual effect of 15 frames/second while the display would be functioning at its normal rate of 60 frames/second). Transmission of the base layer and the first enhancement layer – corresponding to temporal level 1 – would result in video presenting every other frame in the original sequence, which is decoded

video at half-frame rate (again, if the original video was at 60 frames/second, the decoded base layer + first enhancement layer would result in video at 30 frames/sec. Again, practically speaking, each decoded frame would be displayed twice, yielding a visual effect of 30 frames/second while the display would be functioning at its normal rate of 60 frames/second). Transmission of the base layer and both enhancement layers would result in decoding full-frame-rate video.

206. Demircin explains that coded video data has decoding deadlines. *E.g.*, Demircin, [0063], [0105], [0111]; *see also* Demircin916, [0296]. The number of bits that can be transmitted in a given time interval is given by the currently available channel bandwidth (or an estimate of the currently available channel bandwidth). This number of bits is a “bit budget.” Dynamically adapting what is transmitted to the currently available bandwidth is facilitated by scalable coding. Demircin teaches ways to achieve a bit budget using scalably coded video.

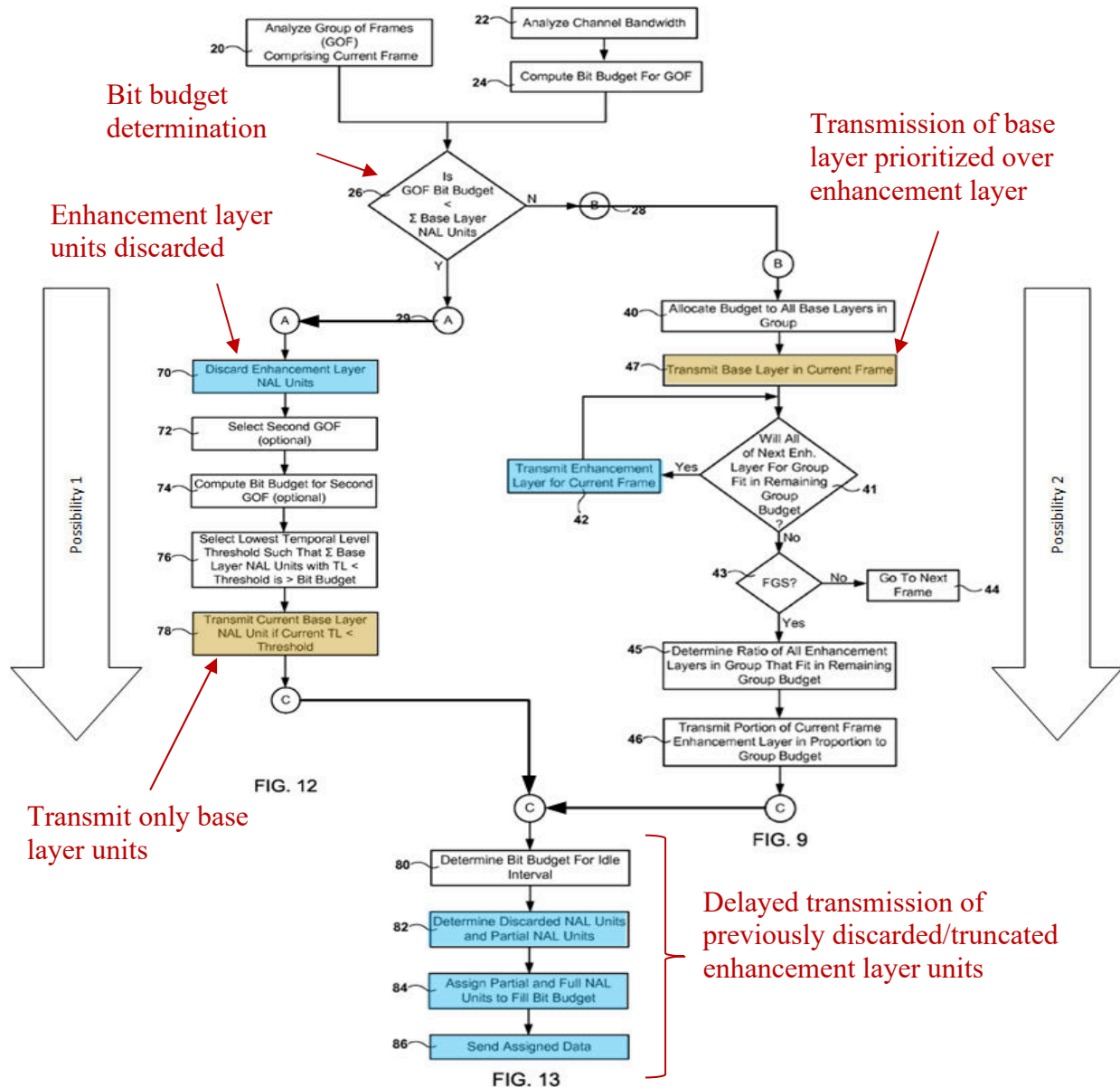
207. Based on the bit budget, Demircin decides whether all or only some of the layers for the video will be transmitted, Demircin [0027]. After transmission of the frames in the current unit starts, Demircin monitors the channel, and if the expectation for channel conditions is incorrect, takes action. If for example there is congestion and the channel bandwidth is reduced, Demircin may discard enhancement layers, thereby reducing the volume of bits transmitted. The discarded layers may be held in memory. If on the other hand more than expected

bandwidth opens up on the channel, Demircin discloses that the scheduled data will be delivered ahead of the deadline, creating idle time, in which (any) discarded data may now be transmitted.

208. Demircin's transmission scheme involves "dynamically adapt[ing] a given SVC bit stream for high-quality video streaming, by intelligently selecting which parts of the bit stream should be transmitted, which parts should be discarded entirely, and which parts should be transmitted partially (truncated), in order to meet time-varying rate and/or delay constraints." *Id.*, [0027]. In this way, Demircin adapts the bit rate to varying channel conditions. *Id.* Demircin teaches performing this adaptation over a time interval called a group-of-frames (GOF). Demircin teaches that the GOF length can be greater than or equal to the GOP length. Examples of GOF lengths given are 2x or 1x (Figures 3, 4, 5), or 1.5x or 3x (Fig 6). A POSITA would understand the benefits of a longer adaptation time for temporal scalability.

209. For example, Demircin's system can detect current channel conditions, and based on the current conditions, can establish a "bit budget" for a group of frames (GOF) that is to be transmitted to a remote decoding device (receiver). *Id.*, [0061]-[0062]. The system then selectively transmits a portion of the coded data for a GOF that can be accommodated within the allocating bit budget, prioritizing base layer data over enhancement layer data. *Id.*, [0129]-

[0131]. If the bit budget cannot accommodate transmission of a frame in the enhancement layer, for example, the enhancement layer frame can be discarded for the moment and later transmitted on a delayed basis during an “idle interval.” *Id.*, [0104]-[0122]. The flowcharts in Figures 7, 9, 12, and 13 illustrate these concepts:



210. Demircin describes a highly general approach in which SNR scalability is combined with temporal scalability. A POSITA would recognize that Demircin's general approach could be used to provide bit-rate adaptation for temporally scalable video with maximum flexibility, without having to combine it with SNR scalability. A POSITA would recognize the advantages of such a technique – it would be substantially simpler to implement, and could be applied to video from any hybrid video coder using an appropriate prediction structure (i.e., application to H.264/AVC video or video from MPEG-2 or MPEG-1), as has been previously described.

211. Demircin explains temporal scalability, both alone and in combination with SNR scalability. Demircin, [0095]-[0103]. Temporal scalability alone is first explained by use of a temporal level value based on the frame's status in the prediction scheme. Demircin assigns a temporal level value that corresponds to the number of dependent frames in the prediction scheme, explaining that this results in I-frames being assigned a level of 0, P-frames predicted from an I-frame are assigned a level of 1, B-frames predicted from a P-frame that is in turn predicted from another frame are assigned a level of 2, and B-frames predicted from other B-frames are assigned a level of 3. This is illustrated in Figure 6. Demircin, [0095] I note that Demircin's FIG. 6 also labels all frames as "layer 0" because this explanation is provided as part of an explanation of combining SNR and temporal

scalability. A POSITA would recognize that temporal scalability could be used on its own, using the temporal level values taught by Demircin.

212. Demircin teaches that the rate of the video can be reduced through temporal scalability, by entirely discarding some frames. Demircin, [0096], Demircin explains that the “SNR base layer data of certain video frames” is discarded. A POSITA would understand that when temporal scalability is used by itself, entire frames are discarded, which is exactly what Demircin teaches, specifically explaining that frames can be discarded based upon their temporal levels, implemented either alone or in conjunction with SNR scalability. Demircin, [0097]-[0098].

213. Demircin teaches that when reducing the frame rate through temporal scalability, the rate reduction (i.e., the bit rate adaptation) can be performed over a time interval that is longer than the time interval that might be selected for SNR scalability. Demircin teaches that this time interval corresponds to a group of frames G’. A POSITA would understand that temporal scalability is effected over multiple frames. A POSITA would also understand that visual effects caused by changing the frame rate are mitigated when the time period that the frame rate remains constant is longer rather than shorter. Indeed, Demircin calls such a longer interval advantageous. Demircin, [0099]

214. Demircin explains that frames are dropped within a group

corresponding to their temporal level. In other words, frames are dropped at a group-level, by examining whether the bit budget can accommodate all of the frames at and below a temporal level. Demircin, [0098]-[0103]

215. While Demircin explains dropping NAL units to effect a rate change at [0101]-[0102], a POSITA would understand that when only temporal scalability is used, a frame is represented in an access unit comprising multiple NAL units, and as such the singular “NAL unit” mentioned in [0101] and [0102] simply becomes plural, referring to all of the NAL units in each dropped frame.

216. A POSITA would understand that Demircin is teaching a subtractive process at [0095]-[0103] – for temporal scalability, if the full frame rate cannot be transmitted, then frames must be dropped.

217. Demircin further explains temporal scalability in the context of combining it with SNR scalability at Demircin, [0144]-[0147], which describe the general subtractive process shown in Figure 12. Figure 12 is invoked when there is insufficient bit rate to transmit the target video and the size of the bitstream to be transmitted must be reduced. Figure 12 represents the process that was previously described in greater detail at Demircin, [0095]-[0103].

218. In more detail, the bit budget constraint that starts the bit rate adaptation process is described in [0128] in conjunction with FIG. 7. Specifically, “a group of frames (GOF) may be analyzed 20 to determine the size of the GOF in

bits, bytes, or some other measure of memory needed to store or transmit the GOF. In some embodiments, the size of each frame within the GOF and each layer within each frame may also be determined. In some embodiments, the size of each frame's base layer and any enhancement layers may be determined.” A POSITA would understand “size of each frame within the GOP and each layer within each frame” is applicable to temporal scalability, in which each frame is coded as a single entity (i.e., without a combination with SNR scalability).

219. Demircin also teaches additive processes for forming scalable bitstreams. FIG. 8 illustrates a frame-level process which is described at Demircin, [0131]-[0134], while FIG. 9 illustrates a process suitable for operating on multiple frames at once (i.e., for a group of frames, and hence for implementing temporal scalability) and is described at Demircin, [0135]-[0138]. Both provide an additive process by which layers are added to the bitstream to increase the amount of coded video data transmitted.

220. By describing both additive and subtractive processes for adapting a scalably coded bitstream, Demircin thus presents a very general approach to provide maximum adaptivity.

221. Demircin's general approach is described for video that is coded using both SNR scalability and temporal scalability. Demircin's process starts with a GOF, a bit budget computed over that GOF, and an initial “target” video to encode

and transmit. Demircin’s initial “target” video is the lowest-quality SNR layer, at full-frame rate (i.e., with all temporal levels). The bit budget computed over the GOF is compared to the bit rate required for the target video, in 26 of FIG. 7. If the bit budget can accommodate the target video (“N” at 26), the additive process of FIG. 9 is invoked at a GOF level. If the bit budget cannot accommodate the target video (“Y” at 26), the subtractive process of FIG. 12 is invoked.

222. As such, to provide maximum flexibility in adaptation, Demircin’s starting target video is not the lowest-rate video that is possible from the SVC bitstream (which would be the lowest-quality SNR layer at minimum frame rate, i.e., for only frames with temporal level 0), because otherwise, a subtractive process would not be necessary. Neither is it the highest-rate video possible (which would be the highest-quality frames at full frame rate, i.e., for all temporal levels), because otherwise an additive process would not be necessary. Demircin’s target video – the starting point of the bit-rate adaptation algorithm – is an intermediate-rate video produced by the SVC coder, as described in the previous paragraph. A POSITA would recognize the benefits of the taught process – by starting “in the middle,” the algorithm is limiting the maximum number of computations it must perform in order to fit the video into the bit budget.

223. A POSITA would recognize that Demircin’s approach is completely and trivially applicable to bit rate adaptation for temporal scalability, instead of to

the combination of SNR and temporal scalability at the same time. A POSITA would recognize that Demircin's teachings of an additive and a subtractive process can be applied to produce temporal scalability in several ways. I remind the reader that in temporal scalability used alone, frames are coded as single entities.

224. First of all, Demircin's subtractive process can be used to reduce the frame rate of video to a lower frame rate that can fit in target bit budget, using FIG. 12. In other words, the frames are each coded in a single layer (as shown in Figure 6), labeled with their temporal level, and the temporal levels are selected for transmission on a GOF-basis that can fit within the bit budget. This process will result in lower-frame-rate video to be transmitted, received, and decoded. By applying this on Demircin's teachings of using a hybrid video coder with a GOP structure including I, P, and/or B frames, this lower-frame rate video will comprise basic hierarchical pictures, while the frames that were not selected for transmission will comprise supplementary hierarchical pictures.

225. Secondly, Demircin's additive process can be used in a dual manner, to increase the frame rate of the lowest-rate video to a higher frame rate that can fit in a target bit budget. In this case, Demircin's starting point is video at temporal level 0, or the base layer of any scalably coded video resulting from a hybrid video coder operating using a GOP prediction structure. Demircin then teaches an additive process in which additional layers are added on a GOF basis. Specifically,

in temporal scalability a POSITA would understand that Demircin adds frames corresponding to temporal layers over units of groups-of-frames as illustrated in FIG. 9. This process will result in higher-frame-rate video to be transmitted, received, and decoded, where the actual frame rate corresponds to the highest temporal level that was included in the bitstream. By applying this on Demircin's teachings of using a hybrid video coder with a GOP structure including I, P, and/or B frames, the transmitted frames form the basic hierarchical pictures and the nontransmitted frames corresponding to temporal levels beyond the highest level selected comprise supplementary hierarchical pictures.

226. Lastly, A POSITA would understand that Demircin's approach of supporting both additive and subtractive processes for bit rate adaptation can be used advantageously if the video frame rate taken as the starting point for adaptation corresponds to neither the lowest possible frame rate nor the highest possible frame rate but rather to an intermediate frame rate.

227. For example, Demircin provides a GOP example with 4 temporal levels in FIG. 1 that can provide frame rates that are full-rate, half-rate, quarter-rate, or eighth-rate. Starting the process with target video that is half-rate supports an additive process of transmitting more frames to achieve full frame rate, or a subtractive process of removing frames to achieve quarter- or eighth-rate. Similarly, starting the process with target video that is quarter-rate supports an

additive process of transmitting more frames to achieve either half- or full frame rate, or a subtractive process of removing frames to achieve eighth-rate.

228. In each of these cases, the resulting temporal frame rate selected by Demircin's process to fit the temporally scalable video into a bit budget results in frames whose coded data is transmitted for a GOF – this will by definition include an I-frame and may include P-frames. It may result in frames whose coded data is not transmitted; these nontransmitted frames are at temporal levels greater than the highest level transmitted, and are the supplementary hierarchical pictures corresponding to the basic hierarchical pictures.

229. Importantly, each of these cases will result in the same set of basic and supplementary hierarchical pictures. There will only be “one answer” for how many temporal levels fit in a bit budget – Demircin teaches how to determine these pictures.

230. The bit budget constraint that starts the bit rate adaptation process is described in [0128] in conjunction with FIG. 7. Specifically, “a group of frames (GOF) may be analyzed 20 to determine the size of the GOF in bits, bytes, or some other measure of memory needed to store or transmit the GOF. In some embodiments, the size of each frame within the GOF and each layer within each frame may also be determined. In some embodiments, the size of each frame's base layer and any enhancement layers may be determined.” A POSITA would

understand “size of each frame within the GOP and each layer within each frame” is applicable to temporal scalability, in which each frame is coded as a single entity (i.e., without a combination with SNR scalability).

B. Overview of Kimoto

231. Kimoto discloses a “moving image data distribution system for distributing reproduction content data that is hierarchically structured so that lower layer data is synthesized with upper layer data to form high quality data.” *E.g.*, Kimoto, Abstract, [0019]-[0024], [0063], [0069]-[0070]. Kimoto perceived that with “moving image coding methods defined in MPEG (Moving Picture Experts Group) and the like, the amount of code to be generated varies greatly from one frame to another.” *Id.*, [0003]. In a viewing situation where bandwidth is limited, “there can occur fluctuations in the reception time of the frame data.” *Id.* Kimoto observed that caching techniques had been proposed “to absorb fluctuations in arrival time of the frames of the first data” received and “to absorb changes in the reproduction time of some of the frame.” *Id.*, [0014]. However, as noted by Kimoto, the “data increases in size as the reproduction contents have a higher bit rate.” *Id.*, [0015]. “This gives rise to the problem that the first data transmission time increases if the first data is transmitted over a limited bandwidth.” *Id.*

232. Thus, Kimoto describes “a moving image data distribution system for distributing moving images” that allows for “quick data initialization even when

the reproduction contents have a high bit rate.” Kimoto, [0018]. Kimoto achieves this by “transmitting only the upper layer data when low quality data has already been viewed on a receiving terminal that contains the lower layer data.” *Id.*, [0025]. “When reproducing the contents,” Kimoto describes synthesizing the upper layer data and “the corresponding part 1005 of the lower layer data to generate high quality data.” Kimoto, [0069].

233. Kimoto explains that its techniques are broadly “applicable” to “international standards of the hierarchical coding methods” including “MPEG-2 scalable, MPEG-4 scalable, MPEG-4 FGS, H.264 / MPEG-4 AVC scalable video coding, and Motion JPEG 2000,” as well as “other hierarchical coding methods.” Kimoto, [0038]. Kimoto teaches scalability via hierarchical coding:

The transmission data storing unit 100 contains data on reproduction contents. The content data stored is *hierarchically coded*. The data is *scalable* in one or a plurality of directions including time (frame rate), space (resolution), and image quality (bit rate). Some of the data can be extracted for reduced frame rate, resolution, and/or bit rate.

Kimoto, [0037] (emphasis added).

234. Kimoto refers to two layers resulting from this coding as a “lower layer” and an “upper layer,” and describes such layers as being “hierarchically structured.”

235. To provide fast start-up on playback, Kimoto teaches determining a length of the upper layer data based on transmission bit rate. For example, the “the

upper layer first data [has] a data length determined based on a transmission bit rate of the upper layer first data and a reproduction time of the first data of the upper layer.” Kimoto claim 42, [0051]. “The upper layer first data length determining unit 110 calculates a reproduction delay time for the case of transmitting only the upper layer data to the receiving unit 20, and determines an upper layer first data length 1100 (S10). The upper layer first data transmitting unit 101 reads upper layer data as long as the upper layer first data length 1100 from the transmission data storing unit 100, and transmits the upper layer data to the receiving unit 20 (S11).” *Id.*, [0047].

236. To reproduce moving image contents, “the hierarchical data synthesizing unit 203 synthesizes the upper layer first data 1004 stored in above S03 and the corresponding part 1005 of the lower layer data to generate high quality data 1006 (S04).” Kimoto, [0069]. “The reproducing unit 201 then decodes the high-quality data 1006” to reproduce the content at the receiver. *Id.*, [0070]. Accordingly, Kimoto’s receiving unit 20 receives video image coding data in the form of lower layer data and upper layer data from the transmitting unit 10, and reconstructs (i.e., synthesizes) video image coding data.

237. Thus, Kimoto teaches first receiving, at a receiver, lower layer data, and storing the lower layer data. Kimoto, [0041]-[0042] (describing that the low quality contents are viewed at a mobile terminal or the like prior to the receiver

receiving the upper layer data). Later in time, Kimoto teaches receiving and storing the upper layer data at the receiver. *Id.*, [0067]. The, when reproducing the contents, “the hierarchical data synthesizing unit 203 synthesizes the upper layer first data 1004 stored in above S03 and the corresponding part 1005 of the lower layer data to generate high quality data 1006 (S04).” *Id.*, [0069].

VII. Claim Construction

238. I have not been asked to opine on the special meaning of claim terms within the ’303 Patent. I have been informed that the claim terms should be afforded their ordinary meaning in light of the language of the claims as a whole, the specification, and file history.

239. I also understand that the parties in a prior litigation involving the ’303 Patent—a litigation in which Apple was *not* a party—stipulated to constructions of the following claim terms:

Claim Term	Claims	Stipulated Construction
“a factor of a group of pictures”	1-2	“one or more groups of pictures”
“reconstructing video image coding data from the basic video image coding data and the supplementary video image coding data”	1-2	“rearranging basic video image coding data with supplementary video image coding data”

240. While I do not take a position on the stipulated constructions from the

prior litigation at this time, this declaration nonetheless shows that the claims would have been unpatentable under Grounds 1 and 2 even if those constructions were to be applied.

VIII. Ground 1A – Claims 1-2 Are Obvious Based on Demircin

241. In my opinion, a person of ordinary skill in the art would have found claims 1-2 of the '303 Patent to have been obvious based on Demircin. I have reviewed the Petition. In my opinion, Demircin renders each of the limitations of claims 1 and 2 obvious for at least the reasons set forth in the Petition, which I fully endorse. To streamline the analysis in this declaration, I do not repeat all of the reasons stated in the Petition as to why a POSITA would have found claims 1-2 obvious based on Demircin. Nonetheless, by way of example and to highlight certain points, I make the following additional observations.

242. Demircin renders obvious [1.P]. In Demircin, a destination device (a video image coding data receiver) receives video image coding data transmitted from a source / server (a video image coding data transmitter) over a packet network. Demircin, [0023] (“transmission of audiovisual (AV) data from a source device (e.g., server) to one or more destination devices (e.g., clients) over a packet network”), [0024] (“a receiver”), [0119] (“client station”). For similar reasons, Demircin also renders obvious [2.P]-[2.2], [2.9].

243. Demircin renders [1.1]-[1.2] obvious. A POSITA would have

understood, and it would have been obvious, that Demircin's "computational resources" would have included a processor and memory for decoding and performing other processing on coded video data at the receiver. Demircin, [0019]; Demircin 917, [0061]. For similar reasons, Demircin also renders obvious [2.3.1]-[2.3.2], [2.10]-[2.11].

244. Demircin renders [1.1] obvious for the additional reason that it was well known at the time that a receiver for decoding video data would include a processor. For example, US 6,292,512 to Radha, cited during prosecution, includes a processor 38, which includes a microprocessor or the link for executing applications including scalable video decoder 22 from memory 20. Radha, 5:20-32. Likewise, van der Shaar, US 6,697,426 and Chen, US 5,886,736, both cited during prosecution, describe a processor for executing instruction. *E.g.*, van der Schaar, 5:54-64; Chen, 20:8-20 (processor 640).

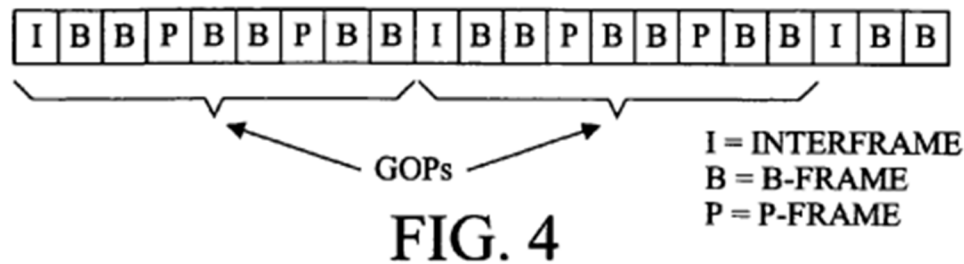
245. Likewise, Demircin renders [1.2] obvious for the additional reason that it was well known at the time that a receiver for decoding video data would include a memory to store instructions, including instruction for decoding video data. For example, US 6,292,512 to Radha, cited during prosecution, includes a memory 30 that stores applications 25, that includes scalable video coder 21 and scalable video decoder 22." Radha, 5:5-19. Likewise, van der Shaar, US 6,697,426, cited during prosecution, describes memory for storing instructions.

van der Schaar, 6:1-7 (memory 68). As evidenced by the prior art, the use of a memory for storing instructions at the receiver for decoding video data was well known and obvious at the time.

246. Demircin renders [1.3] obvious. In Demircin, the transmitter transmits and the receiver receives coded video data including base layer coding data. Demircin, [0095]-[0103], [0129]-[0132], FIGS. 6-9 & 12-13. For similar reasons, Demircin also renders obvious [2.6.1], [2.12].

247. Demircin renders [1.3] obvious for the additional reason that, for temporal scalability, a group of frames can be decoded and played at a higher frame rate if both base and enhancement layer frames have been received by the deadline for frames in the group to be decoded; otherwise, the frames may be decoded and played at a lower frame rate using only base layer frames (e.g., I-frames at temporal level 0). Demircin, [0036]-[0036], [0099], [0145], FIGS. 1 & 12.

248. While the GOP size depicted in FIG. 1 accommodates 4 frames, Demircin broadly contemplates other GOP sizes and structures as well. For example, Demircin further discloses through Demircin916, which is incorporated by reference (*see* APPLE-1006, [0063]), an arbitrarily sized GOP of length N_{GOP} (e.g., 9 frames): “The length of a GOP in number of pictures is denoted by N_{GOP} ... in FIG. 4 where $N_{\text{GOP}}=9$ for illustration purposes.” Demircin916, [0084], FIG. 4.



Demircin916, FIG. 4

A POSITA would understand that GOP size is a variable that can be set by the system designer.

249. Demircin renders [1.4] obvious. In Demircin, the received basic video image coding data would be decoded so as to reproduce video images even when the receiver has received coded data for frames at a higher temporal level since playing video at a higher frame rate involves decoding frames both at the lower and higher temporal levels. Demircin, [0035]-[0036], [0061], [0095]-[0103], FIG. 12; *see also id.*, [0131], FIGS. 6-9, 13. For similar reasons, Demircin also renders obvious [2.13].

250. Demircin renders [1.5.1]-[1.5.3] obvious. Demircin's transmitter transmits, and the receiver thus receives, coded data for video frames at higher temporal levels (e.g., P- or B-frames at temporal levels 1, 2), which corresponds to "supplementary video image coding data" including a "supplementary hierarchical picture" (e.g., a picture for a P- or B-frame) as recited in claim 1. Demircin, [0022], [0035]-[0036]. Demircin's GOP structure "uses I-frames 2, which are

intra-predicted from within the frame; B-frames 3, which may be predicted from one or more other frames,” and optionally “P-frames 4 which are predicted from another I-frame.” Demircin, [0036]; *see also id.*, [0039], [0043], [0095], FIG.1. the source/server “determine[s] the bandwidth of the communication medium over which a signal may be sent” and “[b]ased on the available bandwidth, a GOF bit budget may be computed 24.” Demircin, [0129]. For similar reasons, Demircin also renders obvious [2.5], [2.7]-[2.8], [2.14].

251. The plain language of claim [1.5.1]-[1.5.2] recites the basic hierarchical picture (BHP) to be received before the supplementary hierarchical picture (SHP) (e.g., due to a delay in transmitting the supplementary hierarchical picture). Demircin renders obvious the case in which transmission of the SHP is delayed such that it is received after the BHP, as well as the case in which transmission of the SHP is not delayed and is received before the BHP.

Non-Delayed Transmission of SHP (e.g., P- or B-Frame)

252. To the extent claim 1 does not require receipt of the BHP before the SHP (e.g., no requirement for delayed transmission of the SHP), this scenario is rendered obvious by Demircin. In scenarios where temporal enhancement layer frames are transmitted, and hence received, without delay—which Demircin permits when the channel bandwidth and corresponding bit budgets are sufficiently large—it would have been understood and obvious to a POSITA that, per

Demircin, the blue-shaded B-frame for picture C (an exemplary SHP) identified in the figure above would be transmitted and received before the respective transmission and reception of the gold-shaded I-frame for picture I (an exemplary BHP).

253. For example, Demircin teaches that frames are grouped into groups of frames (GOFs) for purposes of computing “bit budgets” and setting transmission parameters. Demircin, [0029] (“[A] group of frames comprising a current frame and several subsequent frames is considered, a bit budget for that group of frames is calculated, and then bits are allocated ... according to several methods. ... The use of a moving window comprising a group of multiple frames reduces video quality variations over time due to the variations in the bit sizes of individual frames.”); *see also id.*, [0099], [0128]. Demircin’s system then determines which frames to transmit within a GOF based on the frames’ respective temporal levels, where “a lower temporal level indicates greater importance.” *Id.*, [0146]. Frames are selected for transmission “in order of most important temporal level until the bit budget [for the GOF] is exhausted,” which “may be done by selecting a temporal level threshold or by other methods.” *Id.*, [0147]; *see also id.*, [0061], [0144]-[0147], [0097] (“[T]he system may decide whether the current frame is transmitted or dropped based on the importance of the frame in terms of its temporal level $\tau(i)$.”), [0095]-[0103], FIGS. 6, 12. It follows that the B-frame for

picture C (supplementary hierarchical picture) in the above example, which is assigned a temporal level of 1, would be transmitted and received with its current GOF without delay when the bit budget for the GOF is sufficiently large to accommodate transmission of frames with temporal level 1. The I-frame for picture I (basic hierarchical picture), which is assigned a temporal level of 0, would then be transmitted and received after picture C when picture I is included in a GOF that follows the GOF of picture C. *Id.*; see also Demircin, FIGS. 7-9, 12. The same result would be similarly attained for any other temporal enhancement layer frame (SHP) transmitted and received without delay in a GOF preceding a subsequent GOF for which a temporal base layer frame (BHP) is then transmitted and received.

254. For example, Demircin teaches that frames are grouped into groups of frames (GOFs) for purposes of computing “bit budgets” and setting transmission parameters. Demircin, [0029] (“[A] group of frames comprising a current frame and several subsequent frames is considered, a bit budget for that group of frames is calculated, and then bits are allocated ... according to several methods. ... The use of a moving window comprising a group of multiple frames reduces video quality variations over time due to the variations in the bit sizes of individual frames.”); see also *id.*, [0099], [0128]. Demircin’s system then determines which frames to transmit within a GOF based on the frames’ respective temporal levels,

where “a lower temporal level indicates greater importance.” *Id.*, [0146]. Frames are selected for transmission “in order of most important temporal level until the bit budget [for the GOF] is exhausted,” which “may be done by selecting a temporal level threshold or by other methods.” *Id.*, [0147]; see also *id.*, [0061], [0144]-[0147], [0097] (“[T]he system may decide whether the current frame is transmitted or dropped based on the importance of the frame in terms of its temporal level $\tau(i)$.”), [0095]-[0103], FIGS. 6, 12. It follows that the B-frame for picture C (supplementary hierarchical picture) in the above example, which is assigned a temporal level of 1, would be transmitted and received with its current GOF without delay so long as the bit budget for the GOF is sufficiently large to accommodate transmission of frames with temporal level 1. The I-frame for picture I (basic hierarchical picture), which is assigned a temporal level of 0, would then be transmitted and received after picture C when picture I is included in a GOF that follows the GOF of picture C. *Id.*; see also Demircin, FIGS. 7-9, 12.

The same result would be similarly attained for any other temporal enhancement layer frame (SHP) transmitted and received without delay in a GOF preceding a subsequent GOF for which a temporal base layer frame (BHP) is then transmitted and received.

255. Demircin highlights an example embodiment in paragraphs [0061], [0095]-[0103], and [0144]-[0147], and in FIGS. 1 and 12, that involves both SNR

and temporal scalability. For this reason, Demircin first discards the SNR-based enhancement layer NAL units for all frames in a GOF before using temporal scalability to reduce the frame rate selecting frames according to temporal level. As explained above, however, Demircin also discloses and a POSITA would have understood that SVC further allows for temporal scalability even in the absence of SNR scalability. Demircin, [0035]-0036]; *supra*, §VI.A. These embodiments would not require first discarding SNR-based enhancement layer NAL units before prioritizing frames for transmission based on temporal level. While I focus for simplicity on the embodiment involving temporal scalability alone, Demircin's teachings under both embodiments are equally relevant and both render the features of claim 1 obvious.

256. Demirin renders [1.6]-[1.7] obvious. In view of Demircin prioritizing the immediate transmission of key frames (e.g., I-frames at temporal level 0) (basic video image coding data) and delay transmission of non-key frames at higher temporal levels (supplementary video image coding data) when the channel bandwidth decreases, it would have been obvious to delay the transmission of frames in a temporal enhancement layer (e.g., the B-frame for picture C in FIG. 1) by a factor of one or more GOPs after transmission of the key frame for that GOP (e.g., the I-frame for picture E in FIG. 1). Demircin, [0061], [0095]-[0103], [0126], [0144]-[0147], FIGS. 6, 12; *see also id.*, [0104]-[0122], [0127], [0148]-

[0150], FIGS. 13-14. For similar reasons, Demircin also renders obvious [2.15]-[2.16].

257. Demircin also renders [1.6]-[1.7] obvious based on its descriptions of techniques for “Delayed NAL Unit Transmission.” Demircin, [0104]; [0104]-[0122], [0123]-[0128], FIGS. 13-14. According to these techniques, “when the system detects there may be additional channel bandwidth available, transmission of NAL units that were previously discarded or transmitted only partially may be considered and performed” and “[e]ligible NAL units may be reconsidered for transmission in delayed manner.” *Id.*, [0034] (“This method results in high video quality because it increases channel bandwidth utilization.”). In particular, the system “may detect idle channel intervals (instances when the channel is not fully utilized), and subsequently transmit previously dropped or truncated NAL units whose decoding deadlines have not yet expired.” *Id.*, [0105]; *see also id.*, [0106]-[0108], [0148]-[0149].

258. In the context of a temporal scalability implementation, Demircin teaches that transmission of discarded frames can occur during these idle intervals, and that the oldest frame (whose decoding deadline has not yet expired) should be transmitted. Demircin, [0109], [0114]. Specifically, Demircin states “In some embodiments, SNR base layers are always more important than the enhancement layers.” *Id.*, [0109] A POSITA would understand this to indicate that when

frames have been dropped entirely, they should be reinserted (if possible) via the delayed transmission approach. While Demircin's discussion is in terms of NAL units to align with its general application of combined SNR and temporal scalability, a POSITA would understand that a single frame comprises multiple NAL units and that the unit for transmission is the access unit for the eligible frame (i.e., all of the NAL units comprising that frame).

259. Rearrangement of data can occur in a variety of ways. Demircin does not provide details of memory management in a receiver. A POSITA would understand that there has to be a mechanism for the decoder to decode the data in the correct order according to the prediction structure, and that a variety of memory management approaches could be used. For example, the receiver could partition memory for base layer reception and enhancement layer reception, appropriately reading from each memory partition based upon a tabulation of the data that has been received by the decoding deadline. Alternatively, the receiver could linearly fill its buffer, while indexing the data locations for each frame and reading coded data appropriately as needed for decoding frames in the appropriate order. If enhancement frames exist, they will be decoded and displayed. If they do not exist, the base layer frame(s) will be decoded and then displayed for the correspondingly appropriate number of frame intervals.

260. Demircin renders [2.4] obvious. In Demircin, the source/server

“determine[s] the bandwidth of the communication medium over which a signal may be sent” and “[b]ased on the available bandwidth, a GOF bit budget may be computed 24.” Demircin, [0129].

261. Demircin renders [2.6.2] obvious. It would have been obvious to a POSITA for the first memory unit of Demircin’s source/server to store coded enhancement data (supplementary video image coding data) such as B-frames encoded at a non-zero temporal level or other enhancement layer NAL units. Storing coded enhancement data until an appropriate transmission time would have been obvious to ensure that any enhancement layer data that was initially discarded due to insufficient bit budget would be available for delayed transmission during a subsequent idle interval. Demircin, [0034] (“Eligible NAL units may be reconsidered for transmission in delayed manner.”), [0104]-[0106], FIGS. 7, 12-14. It would have been obvious that the discarded pictures would be stored so as to be available for delayed transmission, and a POSITA would have known and found obvious that the storage of the discarded pictures, like much other digital content, would be in a memory unit. Demircin, [0061]-[0062].

262. Demircin renders [2.8] obvious. Demircin’s system determines whether to transmit enhancement layer data for frames in a GOF, and thus controls the transmission of the enhancement layer data for frames in a GOF, based on a bit budget for the GOF calculated based on the acquired bandwidth (i.e., transmission

rate) of the network. Demircin, [0129], FIGS. 7-9.

IX. Ground 1B – Claims 1-2 Are Obvious Based on Demircin in Combination with Kimoto

263. Beyond the relevance of Demircin’s teachings alone, as described in Ground 1A, the features recited in [1.6]-[1.7] and [2.15]-[2.16] also would have been obvious to a POSITA based on the combined teachings of Demircin and Kimoto.⁵ For example, Kimoto, addressed in further detail below in Ground 2, describes a receiving unit 20 that receives “lower layer” coding data for a video long before receiving any “upper layer” data for the video. Kimoto, [0041]-[0043]; *infra*, §III.C.1, [1.6]-[1.7]. Like Demircin and the ’303 Patent, Kimoto explains that its video is “hierarchically coded” and “is scalable in one or a plurality of directions including time (frame rate).” Kimoto, [0037]-[0038]. In the context of temporal scalability, a POSITA thus would have understood that Kimoto’s lower layer data would comprise key frames (e.g., I-frames at temporal level 0) while the upper layer data would comprise non-key frames (e.g., B- or P-frames at higher temporal levels) in some examples. Demircin, [0035]-[0036]; *see also infra* §X.C.1.[1.5.1] for a discussion on different options for organizing temporal layers. Kimoto further explains that, in preparation for decoding, the

⁵ Demircin-Kimoto render obvious all other elements of claims 1-2 for the reasons described above with respect to those elements in Ground 1A.

lower layer data received earlier in time is acquired from a lower layer data storing unit, and is provided to a hierarchical data synthesizing unit to be synthesized with upper layer data received at the moment, thereby forming high quality data.

Kimoto, [0069]-[0070], [0080], [0083], [0109], Abstract, FIGS. 1-2, 6-7.

264. A POSITA would have found it obvious to apply Kimoto's teachings to Demircin such that Demircin's receiver would likewise acquire temporal base layer frame(s) received earlier in time from memory, and would provide the temporal base layer frames(s) to a hierarchical data synthesizing unit to be synthesized with temporal enhancement layer frame(s) received at the moment (e.g., frame(s) whose transmission was delayed per Demircin's teachings), thereby forming high quality data suitable for decoding at a high frame rate. It would have been apparent to a POSITA, or at least obvious, that synthesizing temporal base/lower layer data (*basic video image coding data*) with temporal enhancement/upper layer data (*supplementary video image coding data*) as taught in Kimoto would involve reconstructing the video image coding data by re-arranging the base/lower layer frame data and the enhancement/upper layer frame data in the correct decoding order as opposed to the order in which the frames were received, for reasons discussed in detail above—and especially since Kimoto explains that the result of the synthesizing operation is to produce high-quality data that is suitable for decoding. Kimoto, [0070] (“reproducing unit 201 then decodes

the high quality data 1006”); *supra*, Ground 1A, [1.6]-[1.7]. A POSITA would have been motivated to implement the combination to ensure Demircin’s temporal base and enhancement layer frames are properly and reliably arranged at the receiver in a manner that would ensure they are decoded and reproduced in the correct order, despite the delayed transmission of enhancement layer frames. A POSITA would have reasonably expected success, especially since Demircin and Kimoto describe similar hierarchical coding techniques and rearranging coded frames was a known technique in the art.

265. A POSITA would have sought to integrate Kimoto’s teachings in Demircin in this regard to better ensure Demircin’s temporal base and enhancement layer frames are properly ordered for decoding, despite the delayed transmission of enhancement layer frames. This would serve as a reliable way of ensuring that frames are not inadvertently sent to the decoder before their respective reference picture(s) have been decoded, and would further ensure that the receiver does not inadvertently bypass decoding delayed enhancement layer frames that were not received contemporaneously with the base layer frame in a GOP. A POSITA would have reasonably expected success implementing the combination, especially since Demircin and Kimoto describe similar hierarchical coding techniques and rearranging coded frames was a known technique in the art.

266. Additionally, while Demircin describes a real-time streaming system

in which the delayed transmission of enhancement layer data is limited by real-time decoding deadlines at the receiver, Kimoto describes another option in which all lower layer data for a video is transmitted to and received by the receiver before the upper layer data. Kimoto, [0041]-[0043]. A POSITA would have applied Kimoto's teachings in this regard to Demircin such that Demircin's receiver would receive all temporal base layer frames (e.g., I-frames) for a video before temporal enhancement layer frames (e.g., B- or P-frames). In the combination, and to the extent required by the claims, the resulting system would thus delay transmission and receipt, by at least a factor of a group of pictures, of supplementary video image coding data including a supplementary hierarchical picture (e.g., B- or P-frames) relative to basic video image coding data including a basic hierarchical picture (e.g., I-frames). *Supra*, [1.5.1]-[1.5.3], [1.6]-[1.7]. A POSITA would have sought to temporally separate transmission of basic and enhancement layer data in the combination as taught in Kimoto to achieve predictable benefits, including allowing users to view video at a lower quality even before any enhancement layer data is received. This would have been desirable in various circumstances, such as where the transmission channel is initially bandwidth limited, where the user wishes to view the video first on a smaller display, or where the user may first wish to preview the video at a lower quality without consuming additional bandwidth. Kimoto, [0042], [0091]-[0095]. Kimoto's approach would also reduce complexity

of Demircin's transmitter by avoiding the need to actively manage transmission of enhancement layer frames while transmitting base layer frames in the first instance. A POSITA would have reasonably expected success implementing the combination since it merely involves techniques that were well known before the Critical Date for transmitting base and enhancement layer data at different times.

X. Ground 2 – Claim 1 Is Obvious Based on Kimoto in Combination with Demircin

267. Kimoto in combination with Demircin renders claim 1 obvious.

A. Combination of Kimoto and Demircin

268. Kimoto describes techniques for coding, transmitting, and reproducing scalable video, referred to as “reproduction content data that is hierarchically structured so that lower layer data is synthesized with upper layer data to form high quality data.” Kimoto Abstract, *see also id.*, [0037]-[0045]. More specifically, Kimoto teaches receiving, storing, and displaying the lower layer data earlier in time; then, later in time, receiving upper layer data, and synthesizing the upper layer data with corresponding stored lower layer data to generate high quality data. Kimoto, [0069]-[0070]. As such, Kimoto teaches substantially the same features as claim 1 of the '303 Patent. Kimoto does not, however, expressly disclose setting the lower layer data and upper layer data as GOPs that have an intra coded picture and a plurality of (i.e., at least two) inter prediction coded pictures.

269. Doing so, however, would have been an obvious modification to Kimoto's system in view of Demircin, as described in detail below. In the combination, the hierarchically structured lower and upper data of Kimoto are each set in GOPs, each GOP having an intra coded picture and a plurality of inter prediction coded pictures. For example, a GOP like that depicted in Demircin's Figure 1 (which contains an intra-coded picture and multiple inter-prediction coded pictures) would have been obvious to apply to Kimoto's system in the combination, in an example. A POSITA would have found it obvious to modify the lower and upper layer data of Kimoto to be set in units of GOP, in a manner similar to Demircin, to achieve the predictable results of accurate and efficient synthesis of newly received upper layer data with stored lower layer data, and to support delayed playback described in Kimoto.

B. Reasons to Combine Kimoto and Demircin

270. A POSITA would have been motivated to modify Kimoto so that Kimoto's lower and upper layer data would be organized in GOPs consisting of an intra coded picture (*e.g.*, an I-frame) and a plurality of inter-prediction coded pictures (*e.g.*, hierarchical B-frames), consistent with the structure depicted in FIG. 1 of Demircin, for example, to achieve accurate and efficient synthesis of newly received upper layer data with stored lower layer data, which supports Kimoto's delayed playback of high quality contents. Additionally, a POSITA would have

recognized the advantages of using a GOP structure, as taught by Demircin, to support random playback and to support bandwidth considerations.

Accurate and Efficient Synthesis to Support Delayed Playback

271. At the outset, a POSITA would have turned to Demircin as analogous art because both Kimoto and Demircin were concerned with prioritizing the transmission of base layer (or lower layer) data over the enhancement layer (or upper layer) data, and both involved providing a reduced-rate representation of the video. Demircin, [0035]-[0036], [0135]-[0138]; Kimoto, [0041]-[0042]. And both Kimoto and Demircin describe synthesizing high-quality data at the receiver by reconstructing lower layer data with upper layer data. *E.g.*, Kimoto, [0069]; Demircin, [0035].

272. Kimoto describes the use case in which a viewer “views low quality contents for some period of time before viewing the contents in high quality again.” Kimoto, [0095] – either on a different device, *id.*, [0042], or on the same device. *Id.*, [0095]. The delay between the initial (low quality) viewing and the start of the (high quality) re-viewing is either fixed, *id.*, [0072], or at the request of the user. *Id.*, [0073]. A POSITA also would have been motivated to turn to Demircin for structure of the lower and upper layer data in the form of GOPs because Demircin describes a mechanism to assign data coded using that GOP structure to lower and upper layers. *E.g.*, Demircin, [0035]-[0036]. Such an

assignment facilitates the matching of receiving upper layer data with the “corresponding part” of the stored lower layer data during synthesis, thereby increasing efficiency and accuracy of the synthesis. Kimoto, [0070]. Additionally, such assignments also facilitate random playback. Thus, a POSITA would have recognized the tangible benefits to Kimoto of applying Demircin’s GOP structure. Kimoto, [0070].

Variable Bandwidth

273. Kimoto’s first embodiment involves sending the lower layer (low-quality video) to a mobile terminal, and a POSITA would have recognized such transmissions may have required sending lower layer data over a link that has varying bandwidth. Kimoto, [0015], [0042]. A POSITA would have understood that, in some cases, the mobile unit may have supported a lower bandwidth than the receiving unit; and in that scenario, Kimoto would have benefited from the rate adaptation techniques, as taught by Demircin. Demircin, [0026]-[0027], [0035]-[0036], [0135]-[0138].

274. Because Kimoto teaches that the low quality contents still need to be reproduced on the mobile terminal, a POSITA would be motivated to look for ways to ensure that the perceived quality of video (e.g., frame rate) at the mobile terminal was as high as possible to improve the user’s experience. A POSITA would have understood that use of Demircin’s approach for bit rate adaptation

using an appropriate GOP prediction structure would have resulted in the highest-frame-rate-possible video being transmitted to the mobile terminal and would improve the user experience. Demircin, [0035]-[0036], [0135]-[0138].

275. Moreover, Demircin teaches flexibility in the GOP structure to adapt to varying bandwidths that are often experienced by mobile terminals. Demircin, [0036] (length 4 GOPs), FIG. 6 (length 8 GOP). A POSITA would have recognized the benefits of Demircin's adaptability for providing low quality contents for viewing on Kimoto's mobile terminal. Likewise, a POSITA would have looked to Demircin916's teachings to estimate the available channel bandwidth. A POSITA would then use Demircin's teachings to map the scalable video provided by a hybrid video coder with an appropriate GOP prediction structure into Kimoto's lower layer and upper layer. Accordingly, a POSITA would have recognized the obvious benefits of using Demircin's approach for bit rate adaptation as resulting in the highest-frame-rate-possible video being transmitted to the mobile terminal.

276. Demircin teaches transmission of scalable video over channels with limited and possibly fluctuating bandwidth. Specifically, Demircin addresses how to select (existing, coded) "units" of the scalably coded video that are to be transmitted [in a specific time frame (i.e., given current channel conditions)] over the existing, limited-bandwidth channel. In order to do this, Demircin needs to

know the channel bandwidth, which is provided by extensively described techniques to estimate it in Demircin916.

277. Furthermore, Demircin would improve Kimoto. First, Demircin teaches 2 GOP structures (which a POSITA would have known) that provide more flexibility for temporal scalability than Kimoto's illustration in Figure 4. More details are provided about this below. Use of Demircin's length-8 GOP that allows for 4 frame rates provides more flexibility in the frame rate received by the mobile terminal than Kimoto's GOP – the lower rate could be 1/8-rate, 1/4-rate, or 1/2-rate. Use of Demircin's length-4 GOP that allows for 3 frame rates also provides more flexibility in the frame rate received by the mobile terminal than Kimoto's GOP – the lower rate could be 1/4-rate or 1/2-rate. In both cases, half-rate video is higher frame rate than Kimoto's GOP allows, and thus of higher perceived quality. And the lower-frame-rates support broader functionality of Kimoto's system than with Kimoto's lowest-rate of 1/3-rate video.

278. Video itself has variable rate over time, and hence must be appropriately fit into the channel. Kimoto's system requires that the scalably coded data be partitioned into the lower layer (to be transmitted to the mobile device and to the receiving unit) and into the upper layer (to be transmitted only to the receiving unit). Kimoto specifically uses the terms “lower” and “upper” – not “lowest,” and also not “base” and “enhancement.” Kimoto is not limiting the

mobile terminal to get the lowest-possible frame rate that a scalable stream can provide; Kimoto simply envisions that the mobile terminal gets less data. But Kimoto does not provide details on how to specifically map the scalable coded video to the lower and upper layers that are transmitted.. This mapping is crucial, because the lower layer defines the user experience at the mobile terminal. Thus, A POSITA would have looked to Demircin to provide a method by which scalable video provided by a hybrid video coder with an appropriate GOP prediction structure could be mapped to Kimoto's upper and lower layer. Kimoto, [0095-0103], [0128-0130], [0135-0136], [0144-0148].

Reasonable Expectation of Success

279. Setting the lower and upper layer data of Kimoto in GOPs, like that of Demircin, would have been met with a reasonable likelihood of success considering that the use of GOPs with intra coded pictures and a plurality of inter prediction coded pictures was a well-known video coding technique at the time. *E.g.*, MPEG1, FIGS. 1, 15. GOP prediction structures were well known and used as early as 1993, and were understood to provide temporal scalability. *Id.* A POSITA would have been able to structure Kimoto's lower and upper layer data in units of GOPs, like that described by Demircin, because doing so was a common and standard practice at that time. *E.g.*, MPEG1, FIGS. 1, 15.

280. Kimoto already explains that its "content data" is "hierarchically

coded” and is “scalable in one or a plurality of directions including time (frame rate).” Kimoto, [0037]. Kimoto is broadly applicable to hybrid video coding. *Id.*, [0038]. Likewise, Demircin involves a hybrid video coder providing hierarchical scalable video coding, such as that prescribed by MPEG and other standards bodies. *E.g.*, Demircin [0035] (citing H.264/MPEG AVC/SVC).

281. Also, Kimoto already contemplates structuring pictures as GOPs, as shown in Kimoto, FIG. 4. Kimoto describes that its hierarchical structure includes a non-reference frame and a plurality of corresponding reference frames. *E.g.*, Kimoto [0056]. Thus, the underlying frame structure for intra- and inter-prediction in Kimoto already includes the same number of frames necessary to satisfy the claim limitation of the ’303 Patent. *See* ’303 Patent, 9:33-36 (describing reference frames). The GOP of Demircin already included one intra-coded picture and a plurality of inter-prediction coded pictures, which was a commonplace GOP structure at the time as discussed at length above. *E.g.*, Demircin, FIG. 1, [0036].

282. Additionally, Demircin’s description of structuring data in units of GOPs, as well as its description of adaptability, is thorough and comprehensive. Thus, a POSITA reading Demircin would not have had to undertake undue experimentation to structure the lower and upper layer data of Kimoto in a GOP structure that allowed both 1) sufficiently high quality viewing of lower layer data on the mobile terminal, 2) adaptive lower (and upper) layer data structuring, and 3)

efficient and accurate synthesis of received upper layer data with the corresponding part of the stored lower layer data to produce high quality contents. The GOP prediction structure taught in Demircin naturally supported scalability and hence the mapping of frames to lower and upper layers for each GOP. *See, e.g., supra* §§ II.D, V.A.5, & VI.A.

283. The use of a GOP in Kimoto would have been met with a reasonable expectation of success by a POSITA because the use of GOPs in video image coding was a well-known, standard practice at the time. The application of known techniques for organizing a video stream into GOPs, as taught in Demircin, would have improved the similar type of scalable video coding system of Kimoto in a similar way, which would have yielded no more than predictable results. Implementing GOPs and frame structures like those disclosed in Demircin in other scalable video contexts such as that described in Kimoto would have been well within the norms of standard practice at the time. Kimoto, [0038]; Demircin, [0020], [0026].

C. Analysis

284. In my opinion, a person of ordinary skill in the art would have found claim 1 of the '303 Patent to have been obvious based on Kimoto in view of Demircin. I have reviewed the Petition. In my opinion, the Kimoto/Demircin combination renders each of the limitations of claim 1 obvious for at least the

reasons set forth in the Petition, which I fully endorse. To streamline the analysis in this declaration, I do not repeat all of the reasons stated in the Petition as to why a POSITA would have found claim 1 obvious based on Kimoto/Demircin. Nonetheless, by way of example and to highlight certain points, I make the following additional observations.

285. [1.P] Kimoto/Demircin renders obvious a receiving unit 20 (video image coding data receiver) that receives lower layer and upper layer video image coding data from a transmitting unit 10 (video image coding data transmitter). Kimoto, [0043], [0047].

286. In more detail, Kimoto explains that “FIG. 1 is a block diagram showing a first embodiment of the moving image distribution system according to the present invention.” Kimoto, [0035]. “The transmitting unit 10 includes a transmission data storing unit 100, an upper layer first data transmitting unit 101, an upper layer subsequent data transmitting unit 102, and an upper layer first data length determining unit 110.” Kimoto [0036]. The transmitting unit codes video image coding data according to “international standards of the hierarchical coding methods include MPEG-2 scalable, MPEG-4 scalable, MPEG-4 FGS, H.264/MPEG-4 AVC scalable video coding, and Motion JPEG 2000. The present invention is applicable to such hierarchical coding methods illustrated and other hierarchical coding methods.” Kimoto [0038]. “The transmission data storing unit

100 contains data on reproduction contents. The content data stored is hierarchically coded.” Kimoto [0037]. The transmitting unit coded a video image that would be decoded at the receiver. Kimoto [0037]-[0038].

287. In addition, the transmitting unit 10 transmits first upper layer and subsequent upper layer data to the receiving unit 20. “The transmitting unit 10 includes a transmission data storing unit 100, an upper layer first data transmitting unit 101, an upper layer subsequent data transmitting unit 102, and an upper layer first data length determining unit 110.” Kimoto [0036]. “The upper layer first data transmitting unit 101 reads upper layer data as long as the upper layer first data length 1100 from the transmission data storing unit 100, and transmits the upper layer data to the receiving unit 20 (S.11).” Kimoto [0047].

288. The receiving unit 20 includes “an upper layer first data storing unit 200, a reproducing unit 201, a lower layer data storing unit 202, and a hierarchical data synthesizing unit 203.” *Id.*, [0040].

289. To reproduce moving image contents, “the hierarchical data synthesizing unit 203 synthesizes the upper layer first data 1004 stored in above S03 and the corresponding part 1005 of the lower layer data to generate high quality data 1006 (S04).” *Id.*, [0069]. “The reproducing unit 201 then decodes the high-quality data 1006” to reproduce the content at the receiver. *Id.*, [0070]. Therefore, Kimoto’s receiving unit 20 receives video image coding data in the

form of lower layer data and upper layer data from the transmitting unit 10, and reconstructs (i.e., synthesizes) video image coding data. (Note according to the '303 Patent, “the basic video image coding data and the supplementary video image coding data of the coding data reconstruction unit 202” are synthesized for reproduction. See '303 Patent, 12:51-56.)

290. [1.1] Kimoto/Demircin renders obvious that the receiver, which can be a computer, has a processor. Kimoto, [0024], [0044]. A POSITA would have found it obvious that a receiver, such as a computer, includes a processor for carrying out various functions of the receiving unit as described in Kimoto and provided in the Kimoto-Demircin combination, including receiving, storing, retrieving, decoding, synthesizing, and reproducing moving images, as was commonly practiced at the time. FIG. 2 of Kimoto shows the “procedure of processing of the transmitting unit 10.” Kimoto [0044]. Structurally, the transmitting unit includes an upper layer first data length determining unit 110, and the upper layer first data length determining unit 110 would have been implemented using a processor. Using a processor was widely known at the time to implement image processing, storing, and reproduction, so using a processor in Kimoto would have been met by a reasonable expectation of success, especially in view of the common use of processors for carrying out processing functions in receiving units and other types of computers as of the Critical Date.

291. [1.2] Kimoto/Demircin renders obvious that the receiving unit (computer) has a memory with instructions that cause the processor to perform operations. Kimoto, [0024], [0047], [0111]. Likewise, the transmitting unit 10 would have included a memory unit for storing instructions, including for calculating a “reproduction delay time for the case of transmitting only the upper layer data to the receiving unit 20, and determin[ing] an upper layer first data length 1100 (S10),” among other things. Kimoto [0047].

292. [1.3] Kimoto/Demircin renders obvious receiving basic video image coding data. Kimoto’s receiver receives and stores “lower layer data” such that “it is assumed that the low quality content data, or lower layer data, has already been transferred from a server to the receiving unit 20.” Kimoto [0043]-[0044], [0025].

293. In addition, Kimoto describes hierarchical coding via hybrid video coding, specifically mentioning several international standards, including H.264/MPEG-4 advanced video coding (AVC) scalable video coding (SVC). *E.g.*, Kimoto, [0003], [0038]; *e.g.*, Demircin, [0020]-[0021], [0026], [0035]-[0037]. Demircin provides methods for setting lower and upper layer data to provide scalability in hybrid video coding, specifically mentioning H.264/MPEG4 AVC and the SVC extension. [0020], [0022], [0026], [0035]-[0037]. According to both Kimoto and Demircin, hierarchical coding of the video coding data (including lower layer data) can take different forms, including temporal scalability and

signal-to-noise ratio (SNR) scalability (also referred to as quality scalability). *E.g.*, Demircin, [0026]; Kimoto, [0037]. Kimoto teaches that reproduction content data is hierarchically structured so that the lower layer data and the upper layer data are synthesized to form high quality data. *E.g.*, Kimoto, [0019]-[0024]. As discussed above, lower layer data reads onto basic video image coding data.

294. The lower layer includes the pictures at temporal level 0, which include an intra-prediction coded I-frame. Demircin, [0036]; *see also id.*, FIG. 1. The I-frames of the base layer are each basic hierarchical pictures. Demircin's disclosure of a base layer can correspond to the lower layer of Kimoto in the combination, and is organized into groups of pictures (GOPs), where the "GOP size is equal to 4 frames." Demircin, [0036]. When performing decoding with temporal scalability, decoding the "first enhancement layer requires availability of the base layer," and decoding "each subsequent enhancement layer requires availability of all previous enhancement layers, as well as the base layer." Demircin, [0035].

295. [1.4] Kimoto/Demircin renders obvious decoding basic video data so as to reproduce a video image. Kimoto describes that a situation where a user *who has viewed low quality contents* on a mobile terminal such as a cellular phone and a PDA (Personal Data Assistant) views the same contents in high quality on the receiving unit 20 which is a part of the embodiment of the present invention."

Kimoto, [0041]-[0043]. Viewing low quality contents involves decoding lower layer data (basic video image coding data), *e.g.*, by decoding temporal base layer frames at a lower frame rate.

296. In the combination, Kimoto's lower layer data (for temporal scalability), corresponding to the temporal level 0 and selected higher temporal levels sequentially, not including the highest temporal level as taught in Demircin) "has already been transferred from the server to the receiving unit 20," renders obvious the received basic video image coding data. Kimoto, [0043].

297. Kimoto teaches that the lower layer data is stored in the lower layer data storing unit 202, *e.g.*, Kimoto, [0043], and when reproducing the higher quality contents, "the hierarchical data synthesizing unit 203 synthesizes the upper layer first data 1004 stored in above S03 and the corresponding part 1005 of the lower layer data to generate high quality data 1006 (S04)" stored in the lower layer storing unit. *Id.*, [0069]. Thus, in Kimoto, the lower layer data is decoded and viewed as low quality contents at one point in time and stored in its entirety at the receiver; and later, after the upper layer data is received, the lower layer data is acquired from the lower layer data storing unit for reproduction as a higher quality content with the upper layer data. *Id.*

298. In greater detail, the data at the receiver comprises (1) the entire lower layer data for the entire video, which has already been received and is in memory,

and (2) the higher layer data which is being transmitted in real time. The receiver begins reproducing contents following reception of a portion of the upper layer data (the “upper layer first data”). At this point, the first and second GOPs’ lower layer data will be decoded so that all reference frames are available for use in decoding the first GOP’s upper layer data. Thus, the lower layer is decoded and reproduced. (The process as described assumes the GOPs are coded as “open,” in which a reference frame for the next GOP is needed to decode the current GOP. If the GOPs are instead coded as “closed,” in which a GOP can be decoded as a completely independent entity, then only the first GOP’s lower layer data need be decoded.)

299. [1.5.1] Kimoto/Demircin renders obvious receiving supplementary video image coding data. In the combination, Kimoto describes receiving “upper layer data,” which the receiver synthesizes with the lower layer data to reproduce high quality contents. Kimoto, [0045]-[0047], [0069]-[0070]. In addition, Kimoto teaches that the upper layer data results from hierarchical coding, meaning that the upper layer data is decoded using the lower layer data using hierarchical decoding techniques based upon the temporal dependencies that exist in a GOP structure supporting temporal scalability. *E.g.*, Kimoto, [0019]-[0024], [0038].

300. Before the receiving unit 20 can synthesize the stored lower layer data with the upper layer data to reproduce high-quality contents, “the transmitting unit

10 transmits upper layer first data” to the receiving unit. Kimoto, [0045]. As a first step, at the transmitter, the “upper layer first data length determining unit 110 calculates a reproduction delay time” for “transmitting only the upper layer data to the receiving unit 20.” Kimoto, [0047]. The upper layer first data length determining unit then “determines an upper layer first data length” in order to determine the size of the upper layer first data *Id.*

301. In addition, Kimoto teaches non-reference pictures (basic hierarchical picture) at the beginning of the contents of the upper layer data, *e.g.*, Kimoto, [0063], and reference pictures (supplementary hierarchical picture). *E.g.*, Kimoto, [0055]-[0057].

302. In the combination, the upper layer data corresponds to coded frames at consecutive temporal levels, starting at a minimum value of 1, where the temporal levels are assigned to frames per Demircin, and where the specific upper layer data corresponds to a complementary selection of temporal levels for the lower level data. For example, the {(lower), (upper)} collection of temporal levels could be {(0), (1 2 3)}, or {(0 1), (2 3)}, or {(0 1 2), (3)}, when using Demircin’s length-8 GOP depicted in FIG. 6. A length-4 GOP as shown in FIG. 1 would have lower and upper layers corresponding to temporal levels {(0), (1 2)} or {(0 1), (2)}. The exact selection of the lower and upper levels would be performed according to Demircin’s teachings.

303. Kimoto teaches hierarchical coding via hybrid video coding, specifically mentioning several international standards, including H.264/MPEG-4 advanced video coding (AVC) scalable video coding (SVC). *E.g.*, Kimoto, [0003], [0038]; *e.g.*, Demircin, [0020]-[0021], [0026], [0035]-[0037]. Demircin provides methods for setting lower and upper layer data to provide scalability in hybrid video coding, specifically mentioning H.264/MPEG4 AVC and the SVC extension. [0020], [0022], [0026], [0035]-[0037]. According to both Kimoto and Demircin, hierarchical coding of the video coding data (including lower layer data) can take different forms, including temporal scalability and signal-to-noise ratio (SNR) scalability (also referred to as quality scalability). *E.g.*, Demircin, [0026]; Kimoto, [0037]. Kimoto teaches that reproduction content data is hierarchically structured so that the lower layer data and the upper layer data are synthesized to form high quality data. *E.g.*, Kimoto, [0019]-[0024].

304. [1.5.2]-[1.5.3] Kimoto/Demircin renders obvious that the supplementary hierarchical picture has a coding order and display order earlier by a factor of a GOP. In Kimoto, all of the lower layer data is received and stored before the upper layer data. Kimoto, [0041]-[0043]. Thus, a POSITA would have recognized that the upper layer data has a coding order and display order earlier than at least some of the pictures that make up the lower layer data. In the combination, the upper and lower layer data are structured in GOPs, where each

GOP has one intra coded picture and a plurality of interprediction coded pictures, such as that shown in Demircin, FIG. 1.

305. Kimoto/Demircin also renders [1.5.3] obvious for at least the reasons that, according to both Kimoto and Demircin, hierarchical coding of the video coding data (including lower layer data) can take different forms, including temporal scalability and signal-to-noise ratio (SNR) scalability (also referred to as quality scalability). *E.g.*, Demircin, [0026]; Kimoto, [0037]. Kimoto teaches that reproduction content data is hierarchically structured so that the lower layer data and the upper layer data are synthesized to form high quality data. *E.g.*, Kimoto, [0019]-[0024].

306. The upper layer data and the lower layer data in the combination are complementary sets of the temporal levels assigned to frames in a GOP per Demircin, where the specific upper layer data starts with a minimum temporal level of 1 and corresponds to a complementary selection of temporal levels for the lower level data. The exact selection of the lower and upper levels would be performed according to Demircin's teachings. Kimoto teaches hierarchical coding via hybrid video coding, specifically mentioning several international standards, including H.264/MPEG-4 advanced video coding (AVC) scalable video coding (SVC). *E.g.*, Kimoto, [0003], [0038]; *e.g.*, Demircin, [0020]-[0021], [0026], [0035]-[0037]; *supra* [1.3]. Demircin's FIG. 1 also discloses hierarchical B-frame

coding. Demircin, FIG. 1. As taught in both Kimoto and Demircin, hierarchical coding of the video coding data, yielding both lower (basic) and upper/enhancement (supplementary) data, can take different forms, including temporal scalability and signal-to-noise ratio (SNR) scalability (also referred to as quality scalability). *E.g.*, Demircin, [0026]; Kimoto, [0037]. Kimoto teaches that reproduction content data is hierarchically structured so that the lower layer data and the upper layer data are synthesized to form high quality data. *E.g.*, Kimoto, [0019]-[0024].

307. In view of the above, in the combination, the lower layer data and the upper layer data in the combination are set in units of the group of pictures, such as the units of GOP described by Demircin, each unit of which includes an intra-coded picture and a plurality of inter-prediction coded pictures. Demircin, [0036] (“FIG. 1 illustrates an exemplary ... GOP structure and layering structure of a sample SVC bit stream. ... The first frame of the GOP (in encoding/decoding order) is a key picture 2, 4. The key pictures are either intra-coded (I-frame) 2 or inter-coded using a previous (key) picture (P-frame) 4 as a reference for motion-compensated prediction. ... The remaining pictures of a GOP are hierarchically predicted ...”), [0037], FIG. 1. Demircin’s adaptation relies on the GOP structure to set the lower and upper layers.

308. [1.6]-[1.7] Kimoto/Demircin renders obvious that the basic video

image coding data received before the supplementary video image coding data is acquired from memory. “[w]hen reproducing the contents, the hierarchical data synthesizing unit 203 synthesizes the upper layer first data 1004 stored in above S03 and the corresponding part 1005 of the lower layer data to generate high quality data 1006 (S04).” Kimoto, [0067]-[0070].

XI. Additional Remarks

309. I currently hold the opinions set expressed in this declaration. But my analysis may continue, and I may acquire additional information and/or attain supplemental insights that may result in added observations.

APPENDIX A

Sheila S. Hemami, Ph.D.

Boston MA sheila.hemami@gmail.com www.linkedin.com/in/sheilahemami +1.607.280.9739

Education

STANFORD UNIVERSITY **PALO ALTO, CA**
Ph.D. in Electrical Engineering 1994
M.S. in Electrical Engineering 1992

UNIVERSITY OF MICHIGAN **ANN ARBOR, MI**
B.S. in Electrical Engineering, Summa Cum Laude 1990

Professional Experience

TRIPLE RING TECHNOLOGIES **BOSTON, MA & NEWARK, CA**
Senior Director, Growth Initiatives 2021-present

CHARLES STARK DRAPER LABORATORY **CAMBRIDGE, MA**
Director, Strategic Technical Opportunities and Global Challenges Initiative 2016-2020
Director, Internal Research & Development 2017-2018
Director, Draper Fellowship Program 2016-2020

NORTHEASTERN UNIVERSITY **BOSTON, MA**
Professor and Chair, Department of Electrical and Computer Engineering 2013-2016

CORNELL UNIVERSITY (LEADERSHIP POSITIONS) **ITHACA, NY**
Associate Director, School of Electrical & Computer Engineering 2010-2013
Programming Director & Principal Investigator, CU-ADVANCE Center 2006-2012
Women in Science & Engineering Committee Chair 2002-2005

CORNELL UNIVERSITY (ACADEMIC POSITIONS) **ITHACA, NY**
Professor, School of Electrical & Computer Engineering 2006-2013
Associate Professor, School of Electrical & Computer Engineering 2000-2006
Assistant Professor, School of Electrical Engineering 1995-2000

HEWLETT PACKARD LABORATORIES **PALO ALTO, CA**
Member of Technical Staff, Interactive Video Initiative: Multimedia Systems & Community Networking Group 1994

Visiting Positions

INRIA Rennes – Bretagne Atlantique Visiting Researcher (France) Summer 2016
Chaire Régionale d'Excellence, Pays de la Loire, Université de Nantes (France) 2009 – 2011
École Polytechnique Fédérale de Lausanne WISH Distinguished Visiting Professor (Switzerland) Fall 2008 – Winter 2009
Rice University, Texas Instruments Visiting Professor Spring 2002
Princeton University Department of Electrical Engineering, Visiting Fellow Fall 2001
University of Rabat, Fulbright Distinguished Lecturer (Morocco) May 2001

Externally Sponsored Research at Triple Ring Technologies

US Environmental Protection Agency, “A fieldable, portable, reagent-free microplastic sensor enabling rapid readout and modular operation,” \$100,000, December 2021-May 2022 (PI, with a subaward to Woods Hole Oceanographic Institution)

Programs Developed at Draper in the Global Challenges Initiative

Externally Facing Partnered Projects

- Measuring microplastics in the world’s waterways (with US Environmental Protection Agency, Wallace Research Foundation, 11th Hour Racing, Mystic Aquarium (CT), Patagonia)
- Accelerating coral restoration (with National Atmospheric and Oceanic Administration, Florida Keys coral restoration organizations, Sea Ventures, and Henkel)
- Comprehensive marine protected area monitoring/“Counting whales from space” (with New England Aquarium, Dalhousie University, SUNY Stony Brook, WildMe, Maxar, Planet, Revive & Restore, and Oceankind)
- Monitoring of human impact on the Luangwa River Valley/“Counting hippos from space” (with Community Markets for Conservation (COMACO), Zambia National Parks & Wildlife, Maxar, and Cornell University Atkinson Center for Sustainability)
- Coral rescue toolkit and technologies for ocean conservation (multiple projects, with Oceankind, Revive & Restore, SECORE, University of California/Berkeley, San Diego State University, The Smithsonian, Pennsylvania State University, and Vulcan)
- Accelerating road safety interventions for pedestrians & bicyclists (with the Cities of Cambridge & Somerville, MA and Miovision)
- Pediatric heart valves with adaptation to growth (with Boston Children’s Hospital and Seattle Children’s Hospital, childhood disease advocacy groups, medical device companies and venture firms, medical device entrepreneurs, and multiple FDA offices)
- Maternal ultrasound in ultra-low resource environments (with the Bill & Melinda Gates Foundation and Philips)

Research Programs and Sponsored Research at Cornell

Sponsored Research

- National Science Foundation, “Nanoengineering the future of bits and bytes,” \$200,000, November 2012-October 2014 (1 of 5 Co-PIs)
- Cisco/Intel/Verizon, “Meeting the Video Traffic Demand of 2014: Opportunities, Strategies, and Prototypes,” \$1,000,000, January 2011-December 2013 (1 of 4 Co-PIs)
- National Science Foundation, “Visual Information Measures for Task-Based Imaging Applications,” \$268,261, August 2009-August 2013 (sole PI)
- National Science Foundation, “ACCEL: Advancing Cornell’s Commitment to Excellence and Leadership (Cornell University ADVANCE Proposal),” \$3,300,000, November 2006-October 2011 (1 of 5 Co-PIs)
- National Science Foundation, “A Signal Processing. Approach to Modeling Visual Masking,” \$255,873, July 2005-June 2008 (sole PI)
- National Science Foundation, “A Framework for Encoding American Sign Language and Other Structured Video,” joint with University of Washington, \$269,584 for Cornell, July 2005-June 2008 (1 of 3 Co-PIs)

National Science Foundation, “Workshop Series on Starting Successful Faculty Careers,” \$20,000, September 2004-December 2005 (1 of 2 Co-PIs)

National Science Foundation, “Nomination of Professor Robert M. Gray for a Presidential Award for Excellence in Science, Mathematics, and Engineering Mentoring,” \$10,000 (to Stanford, matched by Stanford to fund a mentoring workshop) (1 of 4 Co-PIs)

Department of the Navy, “Channel and QOS Adaptive Wireless Multimedia Ad-hoc Networks,” \$4,500,000, 1999-2004 (1 of 7 Co-PIs)

Department of Energy, “Visually Optimized Scalable Image Compression,” May 1999-April 2002, \$309,000 (sole PI)

National Science Foundation, “Visually Optimized Supra-threshold Image Compression,” May 1999-April 2002, \$205,000 (sole PI)

National Science Foundation Career Award, “Robust Visual Communications for Packet Networks,” May 1997 - April 2001, \$209,000 (sole PI)

Department of Energy, “Visual Communications for Heterogeneous Networks,” September 1995 - August 1998, \$375,000 (sole PI)

National Science Foundation, “A Next Generation Computing and Communication Substrate,” July 1997 - June 2002, \$1,000,000 (1 of 16 Co-PIs)

Corporate Sponsors

Hewlett-Packard, Lockheed-Martin, Tektronix Corporation, Eastman Kodak Company, AT&T, Cisco, Verizon, GTE, Intel, Center for Electronic Imaging & Science (New York State)

Selected Professional Service Activities

Service to the profession

Special Sessions Chair, IEEE International Conference on Image Processing, 2026

Co-Chair, IEEE Signal Processing Society Fellow Evaluation Committee, 2022, 2023

Chair, IEEE Publishing Products & Services Committee, 2017-18

IEEE Vice President, Publication Services and Products, 2015-2016

Organizing committee, Electrical & Computer Engineering Department Heads Associations (ECEDHA) Workshop on Vision, Branding, and Advocacy, January 2016

ECE Vision & Branding Task Force, Electrical & Computer Engineering Department Heads Association (ECEDHA), 2015-6

Chair, IEEE Publications Services and Products Board Strategic Planning Committee, 2013, 2014 (Member 2011-2016)

Member-at-large, IEEE Publication Services and Products Board, 2012-14

IEEE Signal Processing Society Board of Governors Member-at-Large, January 2009-December 2011

Technical Service

Awards co-chair, IEEE International Conference on Image Processing, 2020

Symposium Co-Chair, SPIE Electronic Imaging 2014

Technical Program Co-Chair, IEEE International Conference on Image Processing, 2012

General Co-Chair, IEEE Image & Multidimensional Signal Processing Workshop, June 2011

Special Sessions Chair, IEEE International Conference on Image Processing, 2009

Editor-in-Chief, IEEE Transactions on Multimedia, January 2008 - December 2010

Chair, Image & Multidimensional Signal Processing Technical Committee (IMDSP TC) of the IEEE, 2006-7; general member 2001-2007 (includes program committee duties for ICASSP & ICIP); Awards Subcommittee Chair, 2005

Associate Editor, IEEE Transactions on Signal Processing, August 2000-April 2006

Service in the interest of mentoring and diversity

Senior Mentor, The Valedictorian Project, Boston, MA, 2022-present

Co-organizer, Mentoring for Engineering Academia II, Banff, Canada, 2007

Organizing/Program Committee, PAESMEM/Stanford School of Engineering Workshop on Mentoring in Engineering, June 2004

Other external service

Brown University College of Engineering External Advisory Committee, 2017-2020

Worcester Polytechnic University Electrical & Computer Engineering External Advisory Committee, 2019-present

French National Laboratories (INRIA) Visiting Review Committee, October 2014

University of California, Santa Barbara ECE Advisory Committee on Faculty Diversity, 2008-2011

Banff International Research Station Scientific Advisory Committee, 2010-2014

Memberships

IEEE (elected Fellow in 2009), Tau Beta Pi, Eta Kappa Nu

Awards and Honors

Scholarship and International Awards

IEEE Signal Processing Society Distinguished Lecturer, 2010-11

IEEE Fellow, 2009, *for contributions to robust and perceptual image and video communications*

Eta Kappa Nu Outstanding Young Electrical Engineer Finalist, 2002

Fulbright Distinguished Lecturer, Morocco, 2001

National Science Foundation CAREER Award, 1997

Kodak Term Professor of Electrical Engineering, 1997, 1998, 1999

Teaching Awards

Cornell University Menschel Teaching Fellow, 2013-4

Cornell University Stephen H. Weiss Presidential Fellow, 2012, *in recognition of inspiring teaching of undergraduate students*

Joel & Ruth Spira Excellence in Teaching Award (Cornell), 2004

HKN C. Holmes MacDonald Outstanding Teaching Award, 2000 (a national award)

Honorable Mention, HKN C. Holmes MacDonald Outstanding Teaching Award, 1999

Lilly Teaching Fellowship, 1996-7

Cornell College of Engineering Michael Tien '72 Teaching Award, 1996-7, 1999-2000

Service Awards

Constance E. Cook and Alice H. Cook Recognition Award, 2005 (from the Cornell University Advisory

Council on the Status of Women)
Cornell University College of Engineering Faculty Diversity Award, 2005

Graduate Student Supervision (at Cornell)

Frank Ciaramello, Ph.D., May 2011, “A computation model of the intelligibility of compressed American Sign Language, and Applications to Video Coding”
David Rouse, Ph.D., December 2010, “Subjective and objective utility assessment for images”
Matt Gaubatz, Ph.D., May 2006, “Practical rate-control tools for wavelet-based image compression”
Chao Tian, Ph.D., August 2005, “Multiple Description Quantization: Improvement and New Approach”
Yegnaswamy Sermadevi, Ph.D., August 2005, “Optimal bit allocation for efficient encoding and transmission of video”
Damon M. Chandler, Ph.D., May 2005, “Visual Detection and Perception of Distortion in Wavelet-Compressed Images”
Mark A. Masry, Ph.D., January 2004, “Perceptual Metrics for Video Quality Evaluation and Image Watermarking”
Aaron Deever, Ph.D. (Applied Math), November 2000, “Projection-based Techniques for Lossy and Lossless Wavelet Image Compression”
Marcia Ramos (Ms.), Ph.D., May 2000, “Perceptually-Based Image Coding”
Yan Yang (Ms.), Ph.D., January 2000, “Rate Control for Video Coding and Transmission”
W. Knox Carey, Ph.D., August 1999, “Applications of Wavelet Coefficient Decay”
2 M.S. graduates, 30 M.Eng. graduates

Teaching **University**

Digital Signal Processing at sophomore, junior, and senior levels
Statistical Signal Processing — senior/1st-year-graduate course
Digital Image Processing — graduate course
Detection & Estimation — advanced graduate course
Various topical seminars including Image & Video Coding Standards; Visual Motion

Technical short courses and tutorials

Visual Perception Models and Metrics – 8-hour tutorial, co-taught (with Prof. Thrason Pappas) in 2015 at IEEE Signal Processing Summer School, Garda Lake, Italy, Sep. 2015
Perceptual Metrics for Image and Video Quality in a Broader Context: From Perceptual Transparency to Structural Equivalence – 3-4 hour tutorial, co-taught 14 times in 2014-24 in IEEE, SPIE, and IS&T conferences (with Prof. Thrason Pappas)
Perceptual Metrics for Quality Evaluation – 3 hour tutorial; co-taught 7 times in 2007-10 in both IEEE and SPIE conferences (with Prof. Thrason Pappas)

Publications

Refereed Journal Publications

1. Morgan G. Blevins, Harry L. Allen, Beckett C. Colson, Anna-Marie Cook, Alexandra Z. Greenbaum, Sheila S. Hemami, Joseph Hollmann, Ernest Kim, Ava A. LaRocca, Kenneth A. Markoski, Peter Miraglia, Vienna L. Mott, William M. Robberson, Jose A. Santos, Melissa M. Sprachman, Patricia Swierk, Steven Tate, Mark F. Witinski, Louis B. Krachman and Anna P. M. Michel, “Perspective:

- Field-Portable Microplastic Sensing in Aqueous Environments: A Perspective on Emerging Techniques,” *Sensors* 2021, Volume 21, No. 10, May 2021.
2. H. R. Wu, A. R. Reibman, W. Lin , F. Pereira, S. S. Hemami, “Perceptual visual signal compression and transmission,” *Proceedings of the IEEE*, October 2013.
 3. R. Vanam, E. Riskin, R. Ladner, S. S. Hemami, “Fast algorithms for designing nearly optimal lookup tables for complexity control of the H.264 encoder,” *Signal, Image and Video Processing: Vol. 7, No. 5* (2013), pp. 991-1003.
 4. F. M. Ciaramello, S. S. Hemami, “A computational intelligibility model for assessment and compression of American Sign Language video,” *IEEE Transactions on Image Processing*, November 2011.
 5. D. M. Rouse, S. S. Hemami, Romuald P  pion, and Patrick Le Callet, “Estimating the usefulness of distorted natural images using an image contour degradation measure,” *Journal of the Optical Society of America: A*, Vol. 28, Issue 2, pp. 157-188 (2011).
 6. S. S. Hemami, A. R. Reibman, “No-reference image and video quality estimation: applications and human motivated design,” *Signal Processing: Image Communications, Special Issue on Image and Video Quality Assessment*, Volume 25, Issue 7, July 2010, pp. 469-81.
 7. D. M. Chandler, M. D. Gaubatz, S. S. Hemami, “Visual Masking in Natural Image Patches: A Model and Compression Application,” *EURASIP Journal on Image and Video Processing/ Patches in Vision*, July 2009.
 8. D. M. Chandler, S. S. Hemami, “VSNR: A visual signal-to-noise ratio for natural images based on near-threshold and supra-threshold vision,” *IEEE Transactions on Image Processing*, Volume 16, Issue 9, Sept. 2007, pp. 2284 - 2298 .
 9. M. D. Gaubatz, S. S. Hemami, “Efficient entropy estimation with doubly stochastic models for quantized wavelet image data,” *IEEE Transactions on Image Processing*, Volume 16, Issue 4, April 2007, pp. 967 - 981 .
 10. M. D. Gaubatz, S. S. Hemami, “Ordering for Embedded Coding of Wavelet Image Data Based on Arbitrary Scalar Quantization Schemes,” *IEEE Transactions on Image Processing*, Volume 16, Issue 4, April 2007, pp. 982 - 996.
 11. M. D. Gaubatz, S. S. Hemami, “Robust rate-control for wavelet-based image coding via Bootstrapping with Probability Models,” *IEEE Transactions on Image Processing*, Volume 16, Issue 3, March 2007, pp. 649 - 663.
 12. J. Chen, C. Tian, T. Berger, S. S. Hemami, “Multiple description quantization via Gram- Schmidt Orthogonalization,” *IEEE Transactions on Information Theory*, Volume 52, Issue 12, Dec. 2006, pp. 5197 - 5217.
 13. Y. Sermadevi, S. S. Hemami, “Convex programming formulations for rate allocation in video coding,” *IEEE Transactions on Circuits and Systems for Video Technology*, Volume 16, Issue 8, Aug. 2006, pp. 947 - 959.
 14. M. A. Masry, S. S. Hemami, “A wavelet-based scalable video quality metric and applications,” *IEEE Transactions on Circuits and Systems for Video Technology*, Volume 16, Issue 2, Feb. 2006, pp. 260 - 273.
 15. C. Tian, S. S. Hemami, “A new class of multiple description scalar quantizer and its application to image coding,” *IEEE Signal Processing Letters*, Vol. 12, No. 4, pp. 329-332, April 2005.
 16. D. M. Chandler, S. S. Hemami, “Dynamic contrast-based quantization for lossy wavelet image compression,” *IEEE Transactions on Image Processing*, Vol. 14, No. 4, pp. 397-410, April 2005.
 17. C. Tian, S. S. Hemami, “On the analysis of multiple description lattice vector quantization,” *IEEE Transactions on Information Theory*, Vol. 50, No. 10, pp. 2458-70, October 2004.

18. C. Tian, S. S. Hemami, "Universal multiple description scalar quantization: analysis and design," *IEEE Transactions on Information Theory*, Vol. 50, No. 9, pp. 2089-2102, September 2004.
19. M. A. Masry, S. S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions," *Signal Processing: Image Communication*, Vol. 19, No. 2, pp. 133-46, February 2004.
20. W. Xu, S. S. Hemami, "Robust adaptive transmission of images and video over multiple channels," *Signal Processing: Image Communication*, Vol. 18, No. 10, pp. 981-1000, November 2003.
21. D. M. Chandler, S. S. Hemami, "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions," *Journal of the Optical Society of America:A*, July 2003.
22. A. T. Deever, S. S. Hemami, "Efficient sign coding and estimation of zero-quantized coefficients in embedded wavelet image codecs," *IEEE Transactions on Image Processing*, Vol. 12, No. 4, pp. 420-30, April 2003.
23. A. T. Deever, S. S. Hemami, "Projection-based modelling and context modelling for lossless image compression," *IEEE Transactions on Image Processing*, Vol. 12, No. 5, pp. 489-99, May 2003.
24. M. G. Ramos, S. S. Hemami, "Supra-threshold wavelet coefficient quantization in complex stimuli: psychophysical evaluation and analysis," *Journal of the Optical Society of America:A*, October 2001.
25. Y. Yang, S. S. Hemami, "Rate control for VBR video over ATM networks: simplification and implementation," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 11, No. 9, pp. 1045-58, September 2001.
26. Y. Yang, S. S. Hemami, "Generalized rate-distortion optimization for motion-compensated video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 10, No. 6, pp. 942-55, September 2000.
27. S. S. Hemami, "Robust image transmission using resynchronizing variable-length codes," *IEEE Journal on Selected Areas in Communications*, Special Issue on Error Resilient Image and Video Transmission, Vol. 18, No. 6, pp. 927-39, August 2000.
28. W. K. Carey, S. S. Hemami, P. N. Heller, "Smoothness-Constrained Wavelet Image Compression," *IEEE Transactions on Image Processing*, Vol. 8, No. 12, pp. 1807-1811, December 1999.
29. G. C. Conklin, S. S. Hemami, "A Comparison of Temporal Scalability Techniques," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 9, No. 6, pp. 909-19, September 1999.
30. W. K. Carey, D. B. Chuang, S. S. Hemami, "Regularity-Preserving Image Interpolation," *IEEE Transactions on Image Processing*, Vol. 8, No. 9, pp. 1293-7, September 1999.
31. M. G. Ramos, S. S. Hemami, "Perceptually-Based Scalable Image Compression for Packet Networks," *Journal of Electronic Imaging*, special issue on Image/Video Compression and Processing for Visual Communications, Vol. 7, No. 3, pp. 453-64, July 1998.
32. S. S. Hemami, R. M. Gray, "Subband Coded Image Reconstruction for Lossy Packet Networks," *IEEE Transactions on Image Processing*, Vol. 6, No. 4, pp. 523-39, April 1997.
33. S. S. Hemami, R. M. Gray, "Subband Filters Optimized for Lost Coefficient Reconstruction," *IEEE Transactions on Signal Processing*, Vol. 45, No. 3, pp. 763-7, March 1997.
34. S. S. Hemami, "Reconstruction-Optimized Lapped Orthogonal Transforms for Robust Image Transmission," *IEEE Trans. Circuits and Systems for Video Technology*, special issue on Wireless Visual Communications, Vol. 6, No. 2, pp. 168-81, April 1996.
35. S. S. Hemami, T. H.-Y. Meng, "Transform Coded Image Reconstruction Exploiting Interblock Correlation," *IEEE Transactions on Image Processing*, Vol. 4, No. 7, pp. 1023-6, July 1995.

Conference Publications/Presentations

1. Edward Scott, S. S. Hemami, "No-Reference Utility Estimation with a Convolutional Neural Network," Electronic Imaging (Intelligent Robotics and Industrial Applications using Computer Vision), San Bruno, CA, February 2018.
2. X. Jiang, M. Le Pendu, R. Farrugia, S. Hemami, C. Guillemot, "Homography-based low rank approximation of light fields for compression," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2017.
3. Edward Scott, S. S. Hemami, "Interactions Between Saliency and Utility," Human Vision & Electronic Imaging, San Bruno, CA, March 2017.
4. Edward Scott, S. S. Hemami, "Image Utility Estimation Using Difference-of-Gaussian Scale Space," IEEE International Conference on Image Processing, Phoenix, AZ, September 2016.
5. Linda McLaughlin, Sheila S. Hemami, "Reduced-reference video quality assessment with scalable overhead for video with packet loss," IEEE International Conference on Image Processing, September 2013.
6. Sheila S. Hemami, Francis M. Ciaramello, Sean S. Chen, Nathan G. Drenkow, Dae Yeol Lee, Seonwoo Lee, Evan G. Levine, Adam J. McCann, "Comparing User Experiences in 2D and 3D Videoconferencing," IEEE International Conference on Image Processing, Orlando, FL, October 2012.
7. Albert Wang, Sheila S. Hemami, Alyosha Molnar, "Angle-sensitive pixels: a new paradigm for low-power, low-cost 2D and 3D sensing," Stereoscopic Displays and Applications XXIII, San Francisco, CA, January 2012.
8. Guilherme O. Pinto, Sheila S. Hemami, "Image quality assessment in the low quality regime," Human Vision and Electronic Imaging, San Francisco, CA, January 2012.
9. Guilherme O. Pinto, David M. Rouse, Sheila S. Hemami, "Measuring contour degradation in natural image utility assessment: methods and analysis," Human Vision and Electronic Imaging, San Francisco, CA, January 2011.
10. Darshan Siddalinga Swamy, Damon M. Chandler, Kellen J. Butler, Sheila S. Hemami, "Parametric quality assessment of synthesized textures," Human Vision and Electronic Imaging, San Francisco, CA, January 2011.
11. F. M. Ciaramello, S. S. Hemami, "Quality versus intelligibility: studying human preferences for American Sign Language video," Human Vision and Electronic Imaging, San Francisco, CA, January 2011.
12. David Rouse, Yiran Wang, Fan Zhang, Sheila Hemami, "A Novel Technique to Acquire Perceived Utility Scores from Textual Descriptions of Distorted Natural Images," IEEE International Conference on Image Processing, Hong Kong, October 2010.
13. F. M. Ciaramello, S. S. Hemami, "The influence of space and time varying distortions on objective intelligibility estimators for region-of-interest video," IEEE International Conference on Image Processing, Hong Kong, October 2010.
14. F. M. Ciaramello, Jung Ko, S. Hemami, "Quality versus intelligibility: Evaluating the coding trade-offs for American Sign Language video," 2010 44th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, March 2010.
15. D. M. Rouse, R. Pepion, P. Le Callet, S. S. Hemami, "Tradeoffs in subjective testing methods for image and video quality assessment," Human Vision and Electronic Imaging, San Jose, CA, January 2010.
16. D. M. Rouse, S. S. Hemami, "Natural Image Utility Assessment Using Image Contours," IEEE International Conference on Image Processing, Cairo, Egypt, November 2009.
17. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned Salient Region Detection,"

Computer Vision and Pattern Recognition (CVPR), Miami, FL, June 2009.

18. D. M. Rouse, R. Pepion, S. S. Hemami, Patrick Le Callet, "Image utility assessment and the relationship with image quality assessment," Human Vision and Electronic Imaging, San Jose, CA, January 2009.
19. F. M. Ciaramello, S. S. Hemami, "Modeling temporal coherence in compressed American Sign Language Video," Human Vision and Electronic Imaging, San Jose, CA, January 2009.
20. D. M. Rouse, S. S. Hemami, "Understanding and Simplifying the Structural Similarity Metric," IEEE International Conference on Image Processing, San Diego, CA, October 2008.
21. M. D. Gaubatz, S. S. Hemami, "On the Nearly Scale-Independent Rank Behavior of Image Quality Metrics," IEEE International Conference on Image Processing, San Diego, CA, October 2008.
22. S. Ghouzali, S. Hemami, M. Aziza, D. Aboutajdine, E. M. Mouaddib, "A skin detection algorithm based on discrete cosine transform and generalized Gaussian density," IEEE International Conference on Image Processing, San Diego, CA, October 2008.
23. D. M. Rouse, S. S. Hemami, "Analyzing the Role of Visual Structure in the Recognition of Natural Image Content with Multi-Scale SSIM," Human Vision and Electronic Imaging, San Jose, CA, January 2008.
24. F. M. Ciaramello, S. S. Hemami, "Complexity constrained rate-distortion optimization of sign language video using an objective intelligibility metric," Visual Communication and Image Processing, San Jose, CA, January 2008.
25. Rahul Vanam, Eve A. Riskin, Sheila S. Hemami, Richard E. Ladner, "Distortion-Complexity Optimization of the H.264/MPEG-4 AVC Encoder using the GBFOS Algorithm," Data Compression Conference, 2007, March 2007.
26. D. M. Rouse, S. S. Hemami, "Quantifying the use of structure in cognitive tasks," Human Vision and Electronic Imaging, San Jose, CA, January 2007.
27. F. Ciaramello, S. S. Hemami, "Can you see me now? An objective metric for predicting intelligibility of compressed American Sign Language Video, Human Vision and Electronic Imaging, San Jose, CA, January 2007.
28. M. D. Gaubatz, S. Kwan, S. S. Hemami, "The role of spatially adapted versus non-spatially adapted structural distortion in supra-threshold compression," Human Vision and Electronic Imaging, San Jose, CA, January 2007.
29. M. D. Gaubatz, S. Kwan, B. Chern, D. M. Chandler, S. S. Hemami, "Spatially adaptive wavelet image compression via structural masking," ICIP 2006, Atlanta, GA.
30. C. Tian, S. S. Hemami, "Visually optimized multiple description image coding," ICASSP 2006, Toulouse, France.
31. M. Gaubatz, A. Vosoughi, A. Scaglione, S. S. Hemami, "Efficient, low complexity encoding of multiple burred noisy downsampled images via distributed source coding principles," ICASSP 2006, Toulouse, France.
32. D. M. Chandler, K. L. Lim, S. S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," SPIE Human Vision and Electronic Imaging, San Jose, CA, January 2006.
33. S. S. Hemami, D. M. Chandler, B. G. Chern, J. A. Moses, "Suprathreshold visual psychophysics and structure-based visual masking," SPIE Visual Communication and Image Processing, San Jose, CA, January 2006.
34. M. D. Gaubatz, S. S. Hemami, "Fast Accurate Rate Control for Low-Rate Wavelet-Based Image Coding via Bootstrapping," IEEE International Conference on Image Processing, Genoa, Italy, September

2005.

35. J. Chen, C. Tian, S. Hemami, T. Berger, "A new class of universal multiple description lattice quantizers," IEEE International Symposium on Information Theory, Adelaide, Australia, September 2005.
36. M. D. Gaubatz, D. M. Chandler, S. S. Hemami, "Spatially selective quantization and coding for wavelet-based image compression," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, March 2005.
37. S. S. Hemami, "Robust Video Coding — An Overview," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, March 2005.
38. Y. Sermadevi, J. Chen, S. S. Hemami, T. Berger, "When is bit allocation for predictive video coding easy?" Data Compression Conference, Snowbird, Utah, March 2005.
39. C. Tian, S. S. Hemami, "Staggered Lattices in Multiple Description Quantization," Data Compression Conference, Snowbird, Utah, March 2005.
40. C. Tian, J. Chen, S. S. Hemami, T. Berger, "Multiple descriptions with central refinement," 39th Annual Conference on Information, Sciences, and Systems (CISS), Baltimore, MD, March 2005.
41. D. M. Chandler, N. L. Dykes, S. S. Hemami, "Visually lossless compression of digital radiographs based on contrast sensitivity and visual masking," SPIE Medical Imaging: Image Perception, Observer, and Technology Assessment, San Diego, CA, February 2005.
42. M. D. Gaubatz, D. M. Chandler, S. S. Hemami, "Spatial quantization via local texture masking," SPIE Human Vision and Electronic Imaging 2005, Santa Clara, CA, January 2005.
43. Y. Sermadevi, S. S. Hemami, "Convexity results for a predictive video coder," Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, November 2004.
44. C. Tian, S. S. Hemami, "An embedded image coding system based on tarp filter with classification," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, CA, May 2004.
45. M. Gaubatz, S. S. Hemami, "Scalable image embeddings from arbitrary wavelet-based perceptual models," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, CA, May 2004.
46. W. Xu, S. S. Hemami, "Distortion optimized multiple channel transmission under delay constraints," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, CA, May 2004.
47. C. Tian, S. S. Hemami, "Sequential design of multiple description scalar quantizers," Data Compression Conference, Snowbird, Utah, March 2004.
48. Y. Sermadevi, S. S. Hemami, "Efficient bit allocation for dependent video coding," Data Compression Conference, Snowbird, Utah, March 2004.
49. M. D. Gaubatz, S. S. Hemami, "Local entropy estimation for low rate wavelet image coding," 38th Annual Conference on Information, Sciences, and Systems (CISS), Princeton, NJ, March 2004.
50. Y. Sermadevi, M. A. Masry, S. S. Hemami, "Rate control with a perceived quality metric," Visual Communications and Image Processing, Santa Clara, CA, January 2004.
51. D. M. Chandler, M. A. Masry, S. S. Hemami, "Quantifying the visual quality of wavelet compressed images based on local contrast, visual masking, and global precedence," Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, November 2003.
52. M. A. Masry, D. M. Chandler, S. S. Hemami, "Digital watermarking using local contrast based texture masking," Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, November 2003.
53. G. Yadavalli, M. A. Masry, S. S. Hemami, "Frame rate preferences in low bit rate video," IEEE

- International Conference on Image Processing, Barcelona, Spain, September 2003.
54. Y. Sermadevi, S. S. Hemami, "Linear approximations and linear programming for video rate control," IEEE International Conference on Image Processing, Barcelona, Spain, September 2003.
 55. W. Xu, S. S. Hemami, "Delay-optimized robust transmission of images over multiple channels," International Conference on Multimedia and Exposition, Baltimore, MD, March 2003.
 56. C. Tian, S. S. Hemami, "Universal multiple description scalar quantization: analysis and design," Data Compression Conference 2003, Snowbird, Utah, March 2003.
 57. Y. Sermadevi, S. S. Hemami, "Linear programming optimizations for video coding under multiple constraints," Data Compression Conference 2003, Snowbird, Utah, March 2003.
 58. C. Tian, S. S. Hemami, "On the asymptotic analysis of multiple description scalar quantization," 37th Annual Conference on Information, Sciences, and Systems, Baltimore, MD, March 2003.
 59. D. M. Chandler, S. S. Hemami, "Suprathreshold image compression based on contrast allocation and global precedence," SPIE Human Vision and Electronic Imaging 2003, San Jose, CA, January 2003.
 60. M. A. Masry, S. S. Hemami, "CVQE: a continuous video quality evaluation model for low bit rates," SPIE Human Vision and Electronic Imaging 2003, San Jose, CA, January 2003.
 61. D. M. Chandler, S. S. Hemami, "Contrast-based quantization and rate control for wavelet coded images," Proceedings IEEE Intl. Conf. Image Processing, Rochester, NY, September 2002. Winner (2nd prize), Best Student Paper Award.
 62. M. A. Masry, S. S. Hemami, "Perceived quality metrics for low bit rate compressed video," Proceedings IEEE Intl. Conf. Image Processing, Rochester, NY, September 2002.
 63. W. Xu, S. S. Hemami, "Efficient partitioning of unequal error protected MPEG video streams for multiple channel transmission," Proceedings IEEE Intl. Conf. Image Processing, Rochester, NY, September 2002.
 64. D. M. Chandler, S. S. Hemami, "Additivity models for suprathreshold distortion in quantized, wavelet-coded images," Proceedings SPIE Human Vision and Electronic Imaging, San Jose, CA, January 2002.
 65. M. Masry, S. S. Hemami, "Subjective quality evaluation of low-rate video," Proceedings IEEE Intl. Conf. Image Processing, Thessaloniki, Greece, October 2001.
 66. S. S. Hemami, M. G. Ramos, "Quantifying Visual Distortion in Low-Rate Wavelet-Coded Images," Proceedings IEEE Intl. Conf. Image Processing, Thessaloniki, Greece, October 2001.
 67. S. S. Hemami, "Supra-threshold wavelet coefficient quantization in natural images: analysis and application to compression," Int. Conf. on Image and Signal Processing, Agadir, Morocco, May 2001.
 68. M. Masry, S. S. Hemami, W. Osberger, A. M. Rohaly, "Subjective quality evaluation of low rate video," Proceedings SPIE Human Vision and Electronic Imaging, San Jose, CA, January 2001.
 69. Y. Yang, S. S. Hemami, "Rate-distortion optimization for region- and object-based wavelet video coding," Proceedings Thirty-fourth Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, October 2000.
 70. A. T. Deever, S. S. Hemami, "Scalable Image Coding with Projection-Based Context Modelling," Proceedings IEEE Intl. Conf. Image Processing, Vancouver, BC, September 2000.
 71. M. Masry, S. S. Hemami, "Data Hiding in Images with Psychovisual Thresholding," Proceedings IEEE Intl. Conf. Image Processing, Vancouver, BC, September 2000.
 72. M. Eoin Buckley, M. G. Ramos, S. S. Hemami, S. B. Wicker, "Perceptually-based Robust Image Transmission over Wireless Channels," Proceedings IEEE Intl. Conf. Image Processing, Vancouver, BC, September 2000.
 73. M. G. Ramos, S. S. Hemami, "Perceptual Quantization for Wavelet-Based Image Coding," Proceed-

- ings IEEE Intl. Conf. Image Processing, Vancouver, BC, September 2000.
74. A. T. Deever, S. S. Hemami, "What's Your Sign? Efficient Sign Coding in an Embedded Wavelet Image Coder," Data Compression Conference, Snowbird, UT, March 2000.
 75. Y. Yang, S. S. Hemami, "Separate Source and Channel Rate Selection for Video over ATM," Data Compression Conference, Snowbird, UT, March 2000.
 76. S. S. Hemami, M. G. Ramos, "Wavelet coefficient quantization to produce equivalent visual distortion in complex stimuli," Proceedings of SPIE Human Vision & Electronic Imaging, San Jose, CA, Jan. 2000.
 77. Y. Yang, S. S. Hemami, "Minmax Frame Rate Control using a Rate-Distortion-Optimized Wavelet Coder," Proceedings of IEEE Int. Conf. on Image Processing, Kobe, Japan, Oct. 1999.
 78. S. S. Hemami, "Distortion Analyses for Temporal Scalability Coding Techniques," Proceedings of IEEE Int. Conf. on Image Processing, Kobe, Japan, Oct. 1999.
 79. Y. Yang, S. S. Hemami, "VBR Rate-distortion optimization for a motion compensated video coder," Proceedings of IEEE Int. Conf. on Image Processing, Kobe, Japan, Oct. 1999.
 80. Y. Yang, S. S. Hemami, "A rate-distortion optimized wavelet video coder," Proceedings of Picture Coding Symposium, Portland, OR, April 1999, pp. 61-4.
 81. S. S. Hemami, T. Chang, R. Lau, "Resynchronizing Variable-Length Codes for Robust Image Transmission," Data Compression Conference, Snowbird Utah, April 1999, pp. 529.
 82. M. G. Ramos, S. S. Hemami, "Activity-Selective SPIHT Coding," Proceedings SPIE Visual Communications & Image Processing, Vol. 3653, Pt. 1, San Jose, CA, Jan. 1999, pp. 315-26.
 83. A. T. Deever, S. S. Hemami, "Dense Motion Field Reduction for Motion Estimation," Thirty second Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, Nov. 1998, Vol. 2, pp. 944-8.
 84. Y. Yang, S. S. Hemami, "Rate-distortion-based Combined Motion Estimation and Segmentation," IEEE Int. Conference on Image Processing, Chicago, IL, Oct. 1998, Vol. 3, pp. 920-4.
 85. W. K. Carey, L. A. Von Pischke, S. S. Hemami, "Rate-distortion Based Scalable Progressive Image Coding," SPIE Conf. on Mathematics of Data and Image Coding, Compression, and Encryption, San Diego, CA, July 1998, Vol. 3456, pp. 197-208.
 86. S. S. Hemami, "Visual Sensitivity Considerations for Subband Coding," Proceedings of Thirty-first Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, Nov. 1997, Vol. 1, pp. 652-6.
 87. Y. Yang, S. S. Hemami, "Rate-Constrained Motion Estimation and Perceptual Coding," Proceedings of IEEE Int. Conference on Image Processing, Santa Barbara, CA, Oct. 1997, Vol. 1, pp. 81-4.
 88. M. G. Ramos, S. S. Hemami, "Psychovisually-Based Multiresolution Image Segmentation," Proceedings of IEEE Int. Conference on Image Processing, Santa Barbara, CA, Oct. 1997, Vol. 3, pp. 66-9.
 89. G. J. Conklin, S. S. Hemami, "Evaluation of Temporally Scalable Video Coding Techniques," Proceedings of IEEE Int. Conference on Image Processing, Santa Barbara, CA, Oct. 1997, Vol. 2, pp. 61-4.
 90. W. K. Carey, D. B. Chuang, S. S. Hemami, "Regularity-Preserving Image Interpolation," Proceedings of IEEE Int. Conference on Image Processing, Santa Barbara, CA, Oct. 1997, Vol. 1, pp. 901-4.
 91. G. Conklin, S. S. Hemami, "Multiresolution Motion Estimation," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Munich, Germany, April 1997, pp. 2873-6.
 92. M. G. Ramos, S. S. Hemami, "Robust Image Coding with Perceptual-Based Scalability," Proceedings Data Compression Conference '97, Snowbird, Utah, March 1997, p. 466.

93. W. K. Carey, S. S. Hemami, P. Heller, "Smoothness-Constrained Wavelet Image Compression," Proceedings IEEE International Conference on Image Processing, Lausanne, Switzerland, September 1996, pp. 569-72.
94. S. S. Hemami, "Reconstruction-Optimized Lapped Orthogonal Transforms," Proceedings of IEEE ICASSP '96, Atlanta, Georgia, May 1996, pp. 1542-5.
95. M. Ramos, S. S. Hemami, "Edge-Adaptive JPEG Image Compression," Proceedings of SPIE Conference on Visual Communications and Image Processing, Orlando, Florida, March 1996, pp. 1082-93.
96. M. Ramos, S. S. Hemami, "Eigenfeatures Coding of Videoconferencing Sequences," Proceedings of SPIE Conference on Visual Communications and Image Processing, Orlando, Florida, March 1996, pp. 100-10.
97. S. S. Hemami, "Digital Image Coding for Robust Multimedia Transmission," Proceedings of Symposium on Multimedia Communications and Video Coding, New York, October 1995, pp. 491-98.
98. S. S. Hemami, R. M. Gray, "Subband Filters Designed for Lost Coefficient Reconstruction," Proceedings of IEEE ICASSP '95, Detroit, Michigan, April 1995, pp. 1520-3.
99. S. S. Hemami, R. M. Gray, "Subband Coded Image Reconstruction for Lossy Packet Networks," Proceedings of The Twenty Eighth Asilomar Conference on Signals, Systems, and Computers, October 1994, pp. 497-91.
100. S. S. Hemami, R. M. Gray, "Image Reconstruction Using Vector Quantized Linear Interpolation," Proceedings of IEEE ICASSP '94, Adelaide, Australia, April 1994, pp. 629-32.
101. T.-H. Meng, E. K. Tsern, A. C. Hung, S. S. Hemami, B. M. Gordon, "Video Compression for Wireless Communications," Proceedings of Virginia Tech's Third Symposium on Wireless Personal Communications, Blacksburg, VA, June 1993, pp. 1-17.
102. S. S. Hemami, T. H.-Y. Meng, "Reconstruction of Lost Transform Coefficients Using a Smoothing Criterion," Proceedings of 1993 Picture Coding Symposium, Lausanne, Switzerland, March 1993.

Book Chapters & Other

1. S. S. Hemami, M. C. H. van der Meulen, "But you don't look like an engineer..." Chronicle of Higher Education, January 24, 2010.
2. E. Riskin, M. Ostendorf, P. Cosman, M. Effros, J. Li, S. Hemami, R. M. Gray, ed., "Mentoring for Academic Careers in Engineering: Proceedings of the PAESMEM/Stanford School of Engineering Workshop," Graphics Publishing, 2005.
3. S. S. Hemami, "Robust image communication over wireless channels," IEEE Communications Magazine, November 2001.
4. S. S. Hemami, "Image Compression," Image Databases: Search and Retrieval of Digital Imagery, L. Bergman and V. Castelli, ed., John Wiley & Sons, 2000.
5. S. S. Hemami, R. M. Gray, "Subband Coding for Lossy Packet Networks," in Recovery Techniques for Image and Video Compression and Transmission, A. Katsaggelos and N. Galatsanos, ed., Kluwer Academic Publishers, October 1998.

Workshop Presentations

1. International Union for the Conservation of Nature (IUCN) World Commission on Protected Areas, Marine Workshop, "What high resolution remote sensing can bring to monitoring of marine protected areas," Washington, DC, February 4, 2020.
2. John M. Irvine, Joshua Nolan, Nathaniel Hofmann, Dale Lewis, Twakundine Simpamba, Paul Zyambo, Alexander J. Travis, Sheila Hemami, "Estimating the Population of Large Animals in the Wild Using

- Satellite Imagery: A Case Study of Hippos in Zambia's Luangwa River," IEEE Artificial Intelligence and Pattern Recognition, Washington DC, October 2019.
3. Guilherme O. Pinto, Sheila S. Hemami, "Image Quality Assessment in the Low Quality Regime," Western New York Image Processing Workshop, Rochester, NY, November 2011.
 4. F. Ciaramello, S. Hemami, "Quality versus Intelligibility: Studying Human Preferences for American Sign Language Video," Western New York Image Processing Workshop, Rochester, NY, September 2010.
 5. R. Vanam, Jaehon Chon, F. Ciaramello, S. S. Hemami, E. Riskin, R. Ladner, "Rate-Distortion- Complexity Optimization of an H.264/AVC Encoder for Real-Time Videoconferencing on a Mobile Device," Video Processing and Quality Monitoring Workshop, Scottsdale, AZ, January 2010.
 6. F. Ciaramello, S. S. Hemami, "Joint rate-intelligibility-complexity optimization of an H.264 video encoder for American Sign Language," Western New York Image Processing Workshop, Rochester, NY, September 2009, Winner, best student presentation.
 7. D. M. Rouse, S. S. Hemami, "The role of edge information to estimate the perceived utility of natural images," Western New York Image Processing Workshop, Rochester, NY, September 2009.
 8. F. Ciaramello, S. S. Hemami, "Real-time face and hand detection for videoconferencing on a mobile device," Video Processing and Quality Monitoring Workshop, Scottsdale, AZ, January 2009.
 9. D. M. Rouse, S. S. Hemami, "How to use and misuse image assessment algorithms," Western New York Image Processing Workshop, Rochester, NY, September 2008.
 10. F. Ciaramello, S. S. Hemami, "Complexity constrained rate-distortion optimization of sign language video using an objective intelligibility metric," Western New York Image Processing Workshop, Rochester, NY, October 2007, Winner, best student presentation.
 11. D. M. Rouse, S. S. Hemami, "Analyzing the Role of Visual Structure in the Recognition of Natural Image Content with Multi-Scale SSIM," Western New York Image Processing Workshop, Rochester, NY, October 2007.
 12. F. Ciaramello, S. S. Hemami, "Can you see me now? An objective metric for predicting intelligibility of compressed American Sign Language video," Video Processing and Quality Monitoring Workshop, Scottsdale, AZ, January 2007.
 13. D. M. Rouse, S. S. Hemami, "Quantifying the Use of Structure in Cognition," Western New York Image Processing Workshop, Rochester, NY, September 2006, Winner, best student presentation.
 14. F. Ciaramello, S. S. Hemami, "Can you see me now? An Objective Metric for Predicting the Intelligibility of Compressed American Sign Language Video," Western New York Image Processing Workshop, Rochester, NY, September 2006.
 15. F. Ciaramello, S. S. Hemami, A. Cavender, R. Ladner, E. Riskin, "Predicting Intelligibility of Compressed American Sign Language Video with Objective Quality Metrics," Video Processing and Quality Monitoring Workshop, Scottsdale, AZ, January 2006.
 16. S. S. Hemami, "A Signal Processor's Approach to Modeling the Human Visual System, and Applications," Mathematics and Multimedia Workshop, Banff, Alberta, Canada, July 2005.
 17. S. S. Hemami, M. A. Masry, "A scalable video quality metric and applications," Video Processing and Quality Monitoring Workshop, Scottsdale, AZ, January 2005.
 18. C. Tian, S. S. Hemami, "A special class of multiple description scalar quantizers," IEEE Information Theory Workshop, Austin, TX, October 2004.
 19. M. D. Gaubatz, S. S. Hemami, "Embedding from arbitrary perceptual models," Western New York Image Processing Workshop, Rochester, NY, September 2004.
 20. Y. Sermadevi, S. S. Hemami, "Convexity results for predictive video coding," Western New York

- Image Processing Workshop, Rochester, NY, September 2004, Winner, best student presentation.
21. D. M. Chandler, S. S. Hemami, "Quantifying the visual quality of wavelet-compressed images based on contrast signal-to-noise ratios and global precedence," Western New York Image Processing Workshop, Rochester, NY, October 2003.
 22. M. D. Gaubatz, S. S. Hemami, "Perceptual embedded coding," Western New York Image Processing Workshop, Rochester, NY, October 2003, Best Student Poster Presentation.
 23. W. Xu, S. S. Hemami, "Rate control for video transmission over multiple channels in ATM Networks," Western New York Image Processing Workshop, Rochester, NY, October 2003.
 24. C. Tian, S. S. Hemami, "An embedded image coding system based on Tarp filtering with classification," Western New York Image Processing Workshop, Rochester, NY, October 2003.
 25. Y. Sermadevi, S. S. Hemami, "Optimal bit allocation for video coding under multiple rate constraints," Western New York Image Processing Workshop, Rochester, NY, October 2003, Best Student Oral Presentation.
 26. S. S. Hemami, "Robust visual communications: Possibilities for sensor networks," IEEE Upstate New York Workshop on Sensor Networks, Syracuse, NY, October 2002.
 27. W. Xu, S. S. Hemami, "Spectrally efficient partitioning of MPEG video streams for efficient transmission over multiple channels," Proceedings of the International Packet Video Workshop, Pittsburgh, PA, April 2002.
 28. M. Masry, S. S. Hemami, "Modeling the subjective quality of low bitrate coded video," Western New York Image Processing Workshop, Rochester, NY, Sept. 2001.
 29. D. Chandler, S. S. Hemami, "Additivity models for suprathreshold distortion in quantized, wavelet-coded images," Western New York Image Processing Workshop, Rochester, NY, Sept. 2001.
 30. A. Deever, S. S. Hemami, "What's Your Sign? Efficient Sign Coding in an Embedded Wavelet Image Coder," Western New York Image Processing Workshop, Rochester, NY, Sept. 1999.
 31. Y. Yang, S. S. Hemami, "Rate control for region-based coding using a wavelet video coder," Western New York Image Processing Workshop, Rochester, NY, Sept. 1999.
 32. M. Ramos, S. S. Hemami, "Towards a perceptual model for supra-threshold subband image coding," Western New York Image Processing Workshop, Rochester, NY, Sept. 1999.
 33. S. S. Hemami, "Rate-Distortion Optimizations for Content-based Coding," IBM Interactive TV Workshop, Yorktown Heights, NY, August 1999.
 34. Y. Yang, S. S. Hemami, "Rate-distortion-based Combined Motion Estimation and Segmentation," Western New York Image Processing Workshop, Rochester, NY, Sept. 1998.
 35. M. G. Ramos, S. S. Hemami, "Activity-Selective SPIHT Coding," Western New York Image Processing Workshop, Rochester, NY, Sept. 1998.
 36. S. S. Hemami, "Human Sensitivity to Quantization Noise in Image Subbands," Western New York Image Processing Workshop, Rochester, NY, Sept. 1997.
 37. M. G. Ramos, S. S. Hemami, "Activity-Selective JPEG Coding," Western New York Image Processing Workshop, Rochester, NY, Sept. 1997.
 38. M. G. Ramos, S. S. Hemami, "Robust Image Coding with Perceptual-Based Scalability," Data Compression Conference Industry Workshop, Snowbird, Utah, March 1997.
 39. S. S. Hemami, T. H.-Y. Meng, "Spatial and Temporal Video Reconstruction for Non-layered Transmission," Proceedings of the Fifth International Workshop on Packet Video, Berlin, Germany, March 1993.

Invited Talks

1. Santa Clara University, Intel-sponsored women's mentoring program, "Personal Ethics - Navigating the Gray Areas at Work," Santa Clara, CA (virtual), May 6, 2021.
2. Senator Sheldon Whitehouse's Energy, Environment, and Ocean Leaders Day, "Microplastic pollution and mitigation," Providence, RI, October 15, 2019.
3. United States Senate Oceans Caucus, "Microplastic pollution," Washington DC, April 4, 2019.
4. World Bank Workshop on Water Resource Management in Tanzania, "Smart sensors and distributed measurement," Washington DC, December 13, 2018.
5. Cornell University Atkinson Center for Sustainability, "Remote sensing for socio-economic and environmental applications," Ithaca, NY, May 2018.
6. World Bank Workshop for Technical Women in Development, "How we "see" women in engineering: reflections on a career path," Washington DC, November 8, 2017.
7. S. S. Hemami, "OneHealth: a common theme for both biosecurity & global health," Gates Grand Challenges Meeting, Washington DC, October 2017.
8. World Bank Conference of Sustainable Development Leaders, "Remote sensing for socio-economic analysis," Washington DC, September 12, 2017.
9. Cornell University, Ithaca, NY, October 5, 2015
10. Auburn University, Alabama, April 6, 2015
11. University of Washington, Seattle, December 8, 2014
12. IEEE GlobalSIP Keynote, December 5, 2014
13. Rensselaer Polytechnic Institute Mercer Distinguished Lecture, Rensselaer, NY, May 7, 2014.
14. UCLA, February 12, 2014.
15. Keynote presentation, Southwest Symposium on Image Analysis and Interpretation, April 24, 2012.
16. University of Arizona, Tucson, AZ, November 21, 2011 (SPS DL)
17. Ryerson University, Toronto, Canada, November 17, 2011 (SPS DL)
18. University of British Columbia, Vancouver, Canada, May 7, 2011 (SPS DL)
19. University of Texas, Dallas, May 5, 2011 (IEEE SPS DL)
20. International Conference on Multimedia Computing and Systems Keynote, Ouarzazate, Morocco, April 7, 2011
21. Faculte de Sciences, Rabat, Morocco, April 5, 2011 (IEEE SPS DL)
22. University of Tetuan, Tetuan, Morocco, April 4, 2011 (IEEE SPS DL)
23. Indian Institute of Sciences, Bangalore, India, March 28, 2011 (IEEE SPS DL)
24. IEEE Hyderabad Section, Hyderabad, India, March 24, 2011 (IEEE SPS DL)
25. Indian Institute of Technology, Hyderabad, March 23, 2011
26. University of Pune, India, March 22, 2011 (IEEE SPS DL)
27. University of California, Davis, February 25, 2011
28. Rochester Institute of Technology, February 9, 2011
29. University of Illinois, Urbana-Champaign, January 28, 2011, "A Signal-Processing Approach to Modeling Vision, and Applications."
30. SIGRAD 2010 – Content Aggregation and Visualization Conference, Vasteras, Sweden, November 24, 2010 (keynote) (IEEE SPS DL)
31. Iowa State University, Ames, IA, November 10, 2010 (IEEE SPS DL)
32. Mayo Clinic, Rochester, MN, November 9, 2010 (IEEE SPS DL)
33. University of Wisconsin, November 8, 2010

34. Madison, Wisconsin IEEE Chapter, November 8, 2010 (IEEE SPS DL)
35. University of Maryland, October 29, 2010 (IEEE SPS DL)
36. Faculte de Sciences, Sfax, Tunisia, October 15 & 16, 2010 (IEEE SPS DL)
37. Boston University, September 15, 2010
38. Ecole Nationale Supérieure de Telecommunications, Paris, May 20, 2010 (IEEE SPS DL)
39. North Carolina State University, April 23, 2010 (IEEE SPS DL)
40. Purdue University, April 16, 2010 (IEEE SPS DL)
41. University of Illinois, Chicago, April 9, 2010 (IEEE SPS DL)
42. IEEE Rochester Joint Chapters Meeting, March 31, 2010 (IEEE SPS DL)
43. Mississippi State University, February 5, 2010 (IEEE SPS DL)
44. Drexel University, February 26, 2010 (IEEE SPS DL)
45. AT&T Shannon Labs, February 24, 2010
46. University of Minnesota, January 21, 2010 (IEEE SPS DL)
47. University of Bristol, Bristol, England, May 13, 2009
48. National University of Singapore, April 28, 2009
49. Nanyang Technological University, Singapore, April 27, 2009
50. Hewlett Packard Laboratories, April 10, 2009
51. Stanford University, April 9 2009, Information Systems Laboratory seminar
52. Stanford University, April 9, 2009, Clayman Institute for Gender Studies
53. Stanford University, April 8, 2009, Women in Electrical Engineering
54. Qualcomm, San Diego, California, April 7, 2009
55. Facute de Sciences, Rabat, Morocco, March 17, 2009
56. University of Fez, Fez, Morocco, March 16, 2009
57. Agency for Science, Technology, and Research (ASTAR), Singapore, February 10 & 11, 2009
58. Pacific Rim Conference on Multimedia, December 10, 2008, "From Single Media to Multimedia - Perception, Coding, and Quality" (Keynote)
59. National Tsing-Hua University, Hsinchu, Taiwan, December 9, 2008, "Incorporating Perception and Cognition into Image and Video Coding: A Task-based Approach"
60. University of Grenoble, Grenoble, France, October 23, 2008, "Incorporating Perception and Cognition into Image and Video Coding: A Task-based Approach"
61. Yale University, New Haven, CT, April 14, 2008, "Incorporating Perception and Cognition into Image and Video Coding: A Task-based Approach"
62. Boeing Corporation, Renton, WA, November 30, 2007, "Incorporating Perception and Cognition into Image and Video Coding: A Task-based Approach"
63. University of Washington, Seattle, WA, November 29, 2007, "Incorporating Perception and Cognition into Image and Video Coding: A Task-based Approach"
64. North Carolina State University, Raleigh, NC, April 26, 2007, "A Signal Processing Approach to Modeling Vision, and Applications"
65. University of Washington, Seattle, WA, April 19, 2007, "A Signal Processing Approach to Modeling Vision, and Applications"
66. IEEE Joint Chapters Meeting, Rochester, NY, March 14, 2007, "A Signal Processing Approach to Modeling Vision, and Applications"
67. North Carolina State University, Raleigh, NC, April 26, 2007, "A Signal Processing Approach to

Modeling Vision, and Applications”

68. University of Washington, Seattle, WA, April 19, 2007, “A Signal Processing Approach to Modeling Vision, and Applications”
69. IEEE Joint Chapters Meeting, Rochester, NY, March 14, 2007, “A Signal Processing Approach to Modeling Vision, and Applications”
70. Massachusetts Institute of Technology, Cambridge, MA, December 5, 2005, “Natural Image Perception: Models and Signal Processing Algorithms”
71. Harvard University, Cambridge, MA, December 2, 2005, “Edges, Structures, and Errors: A Cognition-based Approach to Visual Masking”
72. Wells College, Aurora, NY, October 24, 2005, “Understanding How We See: A Woman Electrical Engineer’s Perspective”
73. University of California, San Diego, June 22, 2005, “A Signal Processor’s Approach to Modeling the Human Visual System, and Applications”
74. Ecole Nationale Supérieure de Telecommunications, Paris, France, April 13, 2005, “A Signal Processor’s Approach to Modeling the Human Visual System, and Applications”
75. Ecole Nationale Supérieure d’Electroniques et Applications, Paris, France, April 12, 2005, “Applying Human Visual System Properties to Imaging Systems”
76. INRIA Institut de Recherche des Images et Signaux Analysis, Rennes, France, April 11, 2005, “A Signal Processor’s Approach to Modeling the Human Visual System, and Applications”
77. University of Texas, Center for Perceptual Systems, Austin, TX, March 25, 2005, “A Signal Processor’s Approach to Modeling the Human Visual System, and Applications”
78. University of Wisconsin, Madison, WI, February 7, 2005, “A Signal Processor’s Approach to Modeling the Human Visual System, and Applications”
79. University of California, Davis, CA, March 22, 2004, “Achieving “Maximum Visual Quality” — An EE’s Approach to Understanding the Human Visual System”
80. Cornell University, T&AM Seminar, November 2003, “Achieving “Maximum Visual Quality” — An EE’s Approach to Understanding the Human Visual System”
81. University of Washington, Seattle, WA, May 29, 2003, “Suprathreshold visual psychophysics and applications to image compression”
82. Microsoft Research Laboratories, Bellevue, WA, May 29, 2003, “How Low Can You Go? Images & Video for Very Low Bit Rates”
83. Stanford University, January 31, 2003, “Suprathreshold visual psychophysics and applications to image compression”
84. IBM T. J. Watson Research Labs, Yorktown Heights, NY, November 2002, “Suprathreshold visual psychophysics and applications to image compression”
85. University of Michigan, Ann Arbor, MI, October 25, 2002, “Suprathreshold visual psychophysics and applications to image compression”
86. University of Rochester, Rochester, NY, October 23, 2002, “Suprathreshold visual psychophysics and applications to image compression”
87. Rice University, Houston, TX, April, 2002, “Psychophysics and image compression”
88. Princeton University, Princeton, NJ, December 2001, “Anytime, Anywhere Visual Communications — Issues in Providing Universal Access”
89. Faculte des Sciences, Rabat, Morocco, May 2001, “Anytime, anywhere visual communications — issues in providing universal access”

90. Al Akhawayn University, Ifrane, Morocco, May 2001, "Robust image communications using resynchronizing variable-length codes and error concealment"
91. Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN, January 31, 2001, "Perception of extremely low-rate images and video: psychophysical evaluation and analysis"
92. Princeton University, Princeton, NJ, December 2000, "Better Imaging Through Psychophysics: Visually Optimized Supra-threshold Compression"
93. University of Michigan, April 2000, "Anytime, Anywhere Visual Communications — Issues in Providing Universal Access"
94. Lucent Technologies — Bell Labs, Murray Hill, NJ, March 3, 2000, "Better Imaging through Psychophysics: Visually Optimized Suprathreshold Compression"
95. Rice University, Houston, TX, February 25, 2000, "Better Imaging through Psychophysics: Visually Optimized Suprathreshold Compression"
96. IBM Interactive TV Workshop, Yorktown Heights, NY, August 17, 1999, "Rate-Distortion Optimizations for Content-based Coding"
97. Eastman Kodak Research Labs, Rochester, NY, June 1, 1999, "Incorporating Human Visual System Properties into Wavelet Image and Video Coders"
98. Tektronix Convergence Distinguished Speaker, Portland, OR, May 28, 1999, "Incorporating Human Visual System Characteristics into Digital Imaging and Video — Mathematical and Experimental Approaches"
99. Real Networks, Seattle, WA, May 26, 1999, "Rate-Distortion Optimizations for Video Coding"
100. Motorola Satcom Group, Chandler, AZ, March 26, 1999, "Image and Video Compression and Applications to Broadband Transmission Problems"
101. Rensselaer Polytechnic Institute, Troy, NY, March 18, 1999, "Rate-Distortion Optimizations for Video Coding"
102. UCLA, Los Angeles, CA, January 28, 1999, "Rate-Distortion Optimizations for Video Coding"
103. California Institute of Technology, Pasadena, CA, January 27, 1999, "Rate-Distortion Optimizations for Video Coding"
104. Texas Instruments Signal Processing Research Labs, Dallas, TX, May 28, 1998, "Perceptually-Based Imaging and Motion Estimation"
105. Southern Methodist University, Dallas, TX, May 27, 1998, "Perceptually-Based Imaging and Motion Estimation"
106. Eastman Kodak Research Labs, Rochester, NY, May 8, 1998, "Wavelet Coefficient Decay and Applications"
107. Eastman Kodak Research Labs, Rochester, NY, February 10, 1998, "Activity-Selective Image and Video Compression"
108. Progressive Networks [now Real Networks], Seattle, WA, October 13, 1997, "Internet-driven Scalable Video"
109. Los Alamos National Lab, Los Alamos, NM, January 6, 1997, "Visual Communications for Heterogeneous Networks"
110. Lockheed Martin, Owego, NY, November 21, 1996, "Scalable Multimedia for Next- Generation Networks"
111. Eastman Kodak Research Labs, Rochester, NY, September 19, 1996, "Imaging and Video Algorithms for Transmission Applications"

112. University of California, San Diego, CA, May 8, 1996, "Reconstruction-Optimized Lapped Orthogonal Transforms for Robust Wireless Image Transmission"
113. Intel Corporation, Portland, OR, June 12, 1995, "Visual Communications for Heterogeneous Networks"

Other relevant experience

Excellent communication skills including explaining highly sophisticated technical concepts to non-experts.

- Depositions, trial testimony (Tyler, TX), Markman tutorials (Northern District of California)
- Live media: Sky News (UK television), WGBH (Boston Public Radio)
- Selected recorded and print media: WBZ television (Boston CBS), Public Radio International, TedX Beacon Street, Canadian Broadcasting Corporation, German Public Radio, the New York Times