IMPROVED INTER-LAYER PREDICTION FOR EXTENDED SPATIAL SCALABILITY IN VIDEO CODING

FIELD OF THE INVENTION

[0001] This invention relates generally to the field of video coding. More particularly, the present invention relates to scalable video coding that supports extended spatial scalability.

BACKGROUND OF THE INVENTION

[0002] This section is intended to provide a background or context to the invention that is recited in the claims. The description herein may include concepts that could be pursued, but are not necessarily ones that have been previously conceived or pursued. Therefore, unless otherwise indicated herein, what is described in this section is not prior art to the description and claims in this application and is not admitted to be prior art by inclusion in this section.

[0003] Different standards have been specified for different technologies. Video coding standards include the ITU-T H.261, ISO/IEC MPEG-1 Visual, ITU-T H.262 or ISO/IEC MPEG-2 Visual, ITU-T H.263, ISO/IEC MPEG-4 Visual and ITU-T H.264 (also known as ISO/IEC MPEG-4 Advanced Video Coding (AVC) or, in short, H.264/AVC). In addition, there are currently efforts underway to develop new video coding standards. One such standard under development is the scalable video coding (SVC) standard, which will become the scalable extension to the H.264/AVC standard. The latest draft of the SVC is Annex F (now Annex G) of the H.264/Advanced Video Coding (AVC) standard. In particular, Annex F includes a feature known as extended spatial scalability (ESS), which provides for the encoding and decoding of signals in situations where the edge alignment of a base layer macroblock (MB) and an enhancement layer macroblock is not maintained. When spatial scaling is performed with a ratio of 1 or 2 and a macroblock edge is aligned across different layers, it is considered to be a special case of spatial scalability.

[0004] For example, when utilizing dyadic resolution scaling (i.e., scaling resolution by a power of 2), the edge alignment of macroblocks can be maintained. This phenomenon is illustrated in Figure 1, where a half-resolution frame on the left (the base layer frame 1000) is upsampled to give a full resolution version of the frame on the right (an enhancement layer frame 1100). Considering the macroblock MB_0 in the base layer frame 1000, the boundary of this macroblock after upsampling is shown as the outer boundary in the enhancement layer frame 1100. In this situation, it is noted that the upsampled macroblock encompasses exactly four full-resolution macroblocks— MB_1, MB_2, MB_3 and MB_4—at the enhancement layer. The edges of the four enhancement layer macroblocks MB_1, MB_2, MB_3 and MB_4 exactly correspond to the upsampled boundary of the macroblock MB_0. Importantly, the identified base layer macroblock macroblock is the only base layer macroblock covering each of the enhancement layer macroblocks MB_1, MB_2, MB_3 and MB_4. In other words, no other base layer macroblock is needed to form a prediction for MB_1, MB_2, MB_3 and MB_4.

[0005] In the case of non-dyadic scalability, on the other hand, the situation is quite different. This is illustrated in Figure 2 for a scaling factor of 1.5. In this case, the base layer macroblocks MB_{10} and MB_{20} in the base layer frame 100 are upsampled from 16x16 to 24x24 in the higher resolution enhancement layer frame 110. However, considering the enhancement layer macroblock MB_{30}, it is clearly observable that this macroblock is covered by two different up-sampled macroblocks— MB_{10} and MB_{20}. Thus, two base-layer macroblocks, MB_{10} and MB_{20}, are required in order to form a prediction for the enhancement layer macroblock MB_{30}. In fact, depending upon the scaling factor that is used, a single enhancement layer macroblock may be covered by up to four base layer macroblocks.

[0006] In the current draft of Annex F of the H.264/AVC standard, it is possible for an enhancement layer macroblock to be coded relative to an associated base layer frame, even though several base layer macroblocks may be needed to form the prediction. Because coding efficiency is closely related to prediction accuracy, it is desirable to form an accurate prediction of the enhancement layer macroblock to improve coding efficiency.

[0007] According to the current draft of Annex F of the H.264/AVC standard, a number of aspects of a current enhancement layer macroblock can be predicted from its corresponding base layer macroblocks. For example, intra-coded macroblocks (also referred to as intra-macroblocks or intra-MBs) from the base layer are fully decoded and reconstructed so that they may be upsampled and used to directly predict the luminance and chrominance pixel values at a corresponding enhancement layer. Additionally, inter-coded macroblocks (also referred to as inter-macroblocks or inter-MBs) from the base layer are not fully reconstructed. Instead, only a prediction residual of each base layer inter-MB is decoded and may be used to predict an enhancement layer prediction residual, but no motion compensation is performed on the base layer inter-MB. This is referred as "residual prediction". Furthermore, for inter-MBs, base layer motion vectors are also upsampled and used to predict enhancement layer motion vectors.

[0008] In addition to the above, in Annex F of the H.264/AVC standard, a flag named *base_mode_flag* is defined for each enhancement layer macroblock. When this flag is equal to 1, then the type, mode and motion vectors of the enhancement layer macroblock should be fully-predicted (or inferred) from its base layer MB(s). Because the same method for deriving macroblock type, mode and motion vectors of an enhancement layer macroblock from base layer MB(s) is known to both the encoder and the decoder, it is unnecessary to further code the macroblock type and mode, as well as its motion vector information into bitstream in this case. If the *base_mode_flag* is equal to 0, then the macroblock type and mode information of an enhancement layer macroblock is not inferred.

[0009] As discussed above, the macroblock type and mode information of an enhancement layer macroblock can be fully predicted from its base layer MB(s) in certain situations. According to the current draft of Annex F of the H.264/AVC standard, when enhancement layer macroblocks are not edge-aligned with base layer macroblocks, for each enhancement layer macroblock, a virtual base layer macroblock is derived based on the base layer macroblocks that cover the enhancement layer macroblock. The type, mode and motion vectors of the virtual base layer macroblock are all determined based on the base layer MB(s). The virtual

base layer macroblock will then be considered as the only macroblock from base layer that exactly covers this enhancement layer macroblock. If the base_mode_flag is equal to 1 for the current enhancement layer macroblock, then its type, mode and motion vectors are set as the same as those of the virtual base layer macroblock.

[0010] The method defined in the current draft of Annex F of the H.264/AVC standard for determining the type, mode and motion vectors of the virtual base layer macroblock is a bottom-up process. First, for each 4x4 block of the virtual base layer macroblock, one pixel located in the second row and second column in the block is used as a representative point for the block, which is shown in Figure 3. In Figure 3, the macroblock is represented at 300. The 4x4 blocks inside the macroblock are represented at 310, and the representative pixel within each 4x4 block is represented at 320. The use of one pixel in each 4x4 block of the virtual base layer macroblock has the advantage of simplicity when the current 4x4 block in virtual base layer macroblock is covered by only one 4x4 block from base layer. But when it is covered by multiple 4x4 blocks from base layer, such a method may not be accurate.

[0011] Figures 4(a) and 4(b) show the relationship between the virtual base layer macroblock 300 and corresponding base layer macroblock(s). The area in the base layer that, after upsampling, would exactly cover the current enhancement layer macroblock is represented at 410 in Figure 4(b). This is also the area that corresponds to the virtual base layer macroblock 300. A representative pixel in a 4x4 block in the virtual base layer macroblock 300 is labeled as p_e. Its corresponding pixel at the base layer is p_b. According to the current draft of Annex F of the H.264/AVC standard, the macroblock partition information of the 4x4 block at the base layer, denoted as 420 in Figure 4(b), in which p_b is located is used as the partition information for the 4x4 block at enhancement layer in which p_e is located. In other words, the partition information of the 4x4 block at the base layer that covers the pixel p_e is used as the partition information for the 4x4 block in which p_e is located. In this way, each 4x4 block in the virtual base layer macroblock 300 can have partition information. Motion vectors associated with the partition information are also used as predictors for enhancement layer motion vectors.

[0012] Within each of the four 8x8 blocks in the virtual base layer macroblock, a block merging process is activated at the 4x4 block level. As shown in Figure 5, if block 1, 2, 3 and 4 all derive their partition from the same single partition from the base layer, then the mode of the 8x8 block is set as 8x8. Otherwise, if block 1 and block 2 derive their partition from a same one partition from the base layer, and block 3 and block 4 also derive their partition from another same one partition from the base layer, then the mode of the 8x8 block is determined as 8x4. Similarly, if block 1 and block 3 have the same partition, and block 2 and block 4 also have the same partition from the base layer, then the mode of the 8x8 block is determined as 4x8. Otherwise, the mode of the 8x8 block is determined as 4x4. This process is repeated separately inside all of the other three 8x8 blocks.

[0013] If all four 8x8 blocks are in 8x8 mode, a block merging process is also performed at the 8x8 block level as shown in Figure 6. In Figure 6, blocks 1, 2, 3 and 4 all represent an 8x8 block. If block 1, 2, 3 and 4 all derive their partition from the same single partition from the base layer, then the mode of the virtual base layer macroblock is determined to be 16x16. If block 1 and block 2 have the same partition, and block 3 and block 4 also have the same partition from the base layer, then the mode of the virtual base layer macroblock is determined as 16x8. If block 1 and block 3 have the same partition, and block 2 and block 4 also have the same partition, then the mode of the virtual base layer macroblock is set as 8x16. Otherwise, the mode of virtual base layer macroblock is set as 8x8.

[0014] According to the current draft of Annex F of the H.264/AVC standard, the predicting of macroblock mode is solely based on the partition information from the base layer. In this arrangement, blocks can only be merged when the blocks share the same partition information from the base layer. However, in the case of extended spatial scalability, it is quite common for different partitions from the base layer to have the same reference frame index and motion vectors. For example, two neighboring macroblocks from the base layer can have the same reference frame index and motion vectors. Additionally, in the case of extended spatial scalability, it is very common for an enhancement layer macroblock to be covered by multiple macroblocks from the base layer. Therefore the use of only partition information in

determining if two blocks should be merged or not often unnecessarily creates small partitions inside of a macroblock. Such small partitions increase computation complexity during sample interpolation processes in motion compensation.

[0015] In light of the above, it would be desirable to provide a system for improved inter-layer prediction for macroblock mode, as well as motion vectors for the case of extended spatial scalability.

SUMMARY OF THE INVENTION

[0016] Various embodiments of the present invention provide a system and method for improving inter-layer prediction for the macroblock mode for the above-described case of extended spatial scalability. The various embodiments also provide a system and method for improving inter-layer prediction for motion vectors in the case of extended spatial scalability. Conventional systems for predicting macroblock mode are based solely on partition information from the base layer; only when blocks share the same partition information from the base layer can they can be merged. In embodiments of the present invention, on the other hand, the merging process comprises relying on reference frame index and motion vector information. When blocks share the same reference frame index and motion vector information from the base layer, these blocks may be merged when applicable. Therefore, even when two blocks have different partitions from the base layer, it may be possible for them to be merged together. This system and method can avoid unnecessarily creating small block partitions, which will in turn reduce computation complexity in the interpolation process of motion compensation.

[0017] Various embodiments provide a method, computer program product and apparatus for encoding an enhancement layer block representing at least a portion of a video frame within a scalable bit stream, whereby a coding mode indicator specifying a partition of the enhancement layer block based at least in part on a plurality of reference frame index values corresponding to a plurality of reference layer blocks is derived, wherein the reference layer blocks that have the same reference frame index are merged to derive the partition size. The enhancement layer block is encoded utilizing the coding mode indicator. Embodiments also provide a method, computer

program product and apparatus for decoding an enhancement layer block representing at least a portion of a video frame within a scalable bit stream, whereby a coding mode indicator specifying a partition of the enhancement layer block based at least in part on a plurality of reference frame index values corresponding to a plurality of reference layer blocks is derived, wherein the reference layer blocks that have the same reference frame index are merged to derive the partition size. The enhancement layer block is decoded utilizing the coding mode indicator.

[0018] Other embodiments provide a method, computer program product and apparatus for encoding an enhancement layer block representing at least a portion of a video frame within a scalable bit stream, whereby partition and motion vector information is derived for the enhancement layer block based upon partition and motion vector information from at least one of a plurality of reference blocks. The enhancement layer block is encoded utilizing the partition and motion vector information. Still other embodiments provide a method, computer program product and apparatus for decoding an enhancement layer block representing at least a portion of a video frame within a scalable bit stream, whereby partition and motion vector information is derived for the enhancement layer block based upon partition and motion vector information from at least one of a plurality of reference blocks. The enhancement layer block is decoded utilizing the partition and motion vector information.

[0019] In another embodiment of the present invention, the condition for block merging can be loosened. In this embodiment, so long as the blocks share the same reference frame index and similar motion vectors, neighboring blocks can be merged together. Motion vectors for the merged larger block are derived from the motion vectors of the blocks to be merged (pre-merging).

[0020] Additionally, each 4x4 block in a virtual base layer macroblock has conventionally been essentially represented by a single pixel located at the second row and second column in the block, and partition information for the current block has been obtained by mapping the representative pixel to base layer and then locating a 4x4 block at base layer. In contrast, in the various embodiments of the present invention, multiple representative pixels in a 4x4 block can be used to represent the

block. When multiple 4x4 blocks are available at the base layer for a current block in the virtual base layer macroblock, the partition and motion vector information for the current block in the virtual base layer macroblock can be derived from all of the partition information and motion vectors of those 4x4 blocks. When using more than one representative pixel in a 4x4 block in deriving partition as well as motion vector information from the base layer, it is possible to obtain a more accurate prediction of motion vectors at the enhancement layer. Such an improved prediction of motion vectors can thereby improve coding efficiency.

[0021] The various embodiments of the present invention can be implemented directly in software using any common programming language, e.g. C/C++ or assembly language. The embodiments of the present invention can also be implemented in hardware and used in consumer devices.

[0022] These and other advantages and features of the invention, together with the organization and manner of operation thereof, will become apparent from the following detailed description when taken in conjunction with the accompanying drawings, wherein like elements have like numerals throughout the several drawings described below.

BRIEF DESCRIPTION OF THE DRAWINGS

[0023] Figure 1 is shows the positioning of macroblock boundaries in dyadic resolution scaling;

[0024] Figure 2 is a shows the positioning of macroblock boundaries in non-dyadic resolution scaling;

[0025] Figure 3 is a representation of a virtual base layer macroblock, with a representative pixel in each of sixteen 4x4 blocks;

[0026] Figure 4(a) is a representation of a virtual base layer macroblock, and Figure 4(b) shows the relationship between the virtual base layer macroblock of Figure 4(a) and a plurality of base layer macroblocks;

[0027] Figure 5 shows how blocks may be merged at the 4x4 block level in a virtual base layer macroblock in accordance with Annex F of the H.264/AVC standard

[0028] Figure 6 shows how blocks may be merged at the 8x8 block level in a virtual base layer macroblock in accordance with Annex F of the H.264/AVC standard;

[0029] Figure 7 is a representation showing a 4x4 block being covered by multiple 4x4 blocks from a base layer, when a single representative pixel is used in accordance with Annex F of the H.264/AVC standard;

[0030] Figure 8 is a representation showing the use of multiple representative pixels in a 4x4 block in accordance with an embodiment of the present invention;

[0031] Figure 9 shows a generic multimedia communications system for use with the various embodiments of the present invention;

[0032] Figure 10 is a perspective view of a communication device that can be used in the implementation of the present invention; and

[0033] Figure 11 is a schematic representation of the telephone circuitry of the communication device of Figure 10.

DETAILED DESCRIPTION OF VARIOUS EMBODIMENTS

[0034] Various embodiments of the present invention provide a system and method for improving inter-layer prediction for the macroblock mode for the above-described case of extended spatial scalability. The various embodiments also provide a system and method for improving inter-layer prediction for motion vectors in the case of extended spatial scalability. Conventional systems for predicting macroblock mode are based solely on partition information from the base layer; only when blocks share the same partition information from the base layer can they can be merged. In embodiments of the present invention, on the other hand, the merging process comprises relying on reference frame index and motion vector information. When blocks share the same reference frame index and motion vector information from the base layer, these blocks may be merged when applicable. Therefore, even when two blocks have different partitions from the base layer, it may be possible for them to be merged together.

[0035] In another embodiment of the present invention, the condition for block merging can be loosened. In this embodiment, so long as the blocks share the same reference frame index and similar motion vectors, neighboring blocks can be merged

together. Motion vectors for the merged larger block are derived from the motion vectors of the blocks to be merged (pre-merging).

[0036] Additionally, each 4x4 block in a virtual base layer macroblock has conventionally been essentially represented by a single pixel located at the second row and second column in the block, and partition information for the current block has been obtained by mapping the representative pixel to base layer and then locating a 4x4 block at base layer. In contrast, in the various embodiments of the present invention, multiple representative pixels in a 4x4 block can be used to represent the block. When multiple 4x4 blocks are available at the base layer for a current block in the virtual base layer macroblock, the partition and motion vector information for the current block in the virtual base layer macroblock can be derived from all of the partition information and motion vectors of those 4x4 blocks. When using more than one representative pixel in a 4x4 block in deriving partition as well as motion vector information from the base layer, it is possible to obtain a more accurate prediction of motion vectors at the enhancement layer.

[0037] According to the various embodiments of the present invention, the method for determining the type, mode and motion vectors of the virtual base layer macroblock is still a bottom-up process. However, a number of changes are implemented to achieve the improvements discussed above.

[0038] In terms of block merging, two blocks are merged if they have the same reference frame index and similar or identical motion vectors from the base layer. As shown in Figure 4(b), it can be assumed for exemplary purposes that base layer MB_1 and base layer MB_2 both have an inter 16x16 mode and in addition, have the same reference frame index and motion vectors. According to the relationship shown in Figures 4(a) and 4(b), the 4x4 blocks M and P in virtual base layer macroblock gets partition information from base layer MB1, while 4x4 block N and Q get partition information from base layer MB2. According to one embodiment of the present invention, it is possible for blocks M, N, P and Q to be merged because they all have the same reference frame index and motion vector information from the base layer. After merging, the mode for the 8x8 block is 8x8. However, according to the method currently defined in Annex F of the H.264/AVC standard, such a merging would not

be permitted because they are from different base layer macroblocks and thus belong to different partitions. After merging using the conventional rules, the mode for the 8x8 block would be 4x8, even though such a sub-partition is unnecessary. Furthermore, due to such unnecessary sub-partitioning, further block merging could also be impossible. For example, if it is assumed in Figure 4 that base layer macroblocks MB3 and MB4 also have an inter 16x16 mode and have the same reference frame index and motion vectors as MB1 and MB2, then according to embodiments of the present invention, the blocks in the virtual base layer macroblock 300 will eventually be all merged together, and a mode of inter 16x16 will be assigned to the virtual base layer. However, with the method currently defined in Annex F of the H.264/AVC standard, the mode for the virtual base layer macroblock 300 would be 8x8, with some 8x8 blocks having further sub-partitions.

[0039] Although the use of unnecessarily small partitions may not affect coding efficiency, it may affect computation complexity. During the motion compensation process, sample interpolation is generally performed on a partition or sub-partition basis. Interpolation on a larger partition generally has less complexity than on a number of smaller partitions that have the same total size as the larger partition. This is primarily because, when performing interpolation on a larger partition, more intermediate data can be shared and utilized to reduce computation complexity.

[0040] In another embodiment of the present invention, the condition for block merging can be loosened to a certain extent. As long as they share the same reference frame index and similar motion vectors, neighboring blocks can be merged together. The similarity of motion vectors can be measured through a predetermined threshold T_{mv}. Assuming two motion vectors are $(\Delta x_1, \Delta y_1)$, $(\Delta x_2, \Delta y_2)$, respectively, the difference between the two motion vectors can be expressed as: $\mathbf{D}((\Delta x_1, \Delta y_1), (\Delta x_2, \Delta y_2))$. In this instance, **D** is a certain distortion measure. For example, the distortion measure can be defined as the sum of the squared differences between the two vectors. The distortion measure can also be defined as the sum of absolute differences between the two vectors. As long as $\mathbf{D}((\Delta x_1, \Delta y_1), (\Delta x_2, \Delta y_2))$ is not larger than the threshold T_{mv}, then the two motion vectors are considered to be similar.

[0041] The threshold T_{mv} can be defined as a number, e.g. T_{mv} =0, 1 or 2, etc. T_{mv} can also be defined as a percentage number, such as within 1% of (Δx_1, Δy_1) or (Δx_2, Δy_2) etc. Some other forms of definition of T_{mv} are also allowed. When T_{mv} is equal to 0, it requires that (Δx_1, Δy_1) and (Δx_2, Δy_2) be exactly the same, which is the condition described in the embodiment of the invention discussed previously.

[0042] Motion vectors for the merged larger block are derived from the motion vectors of those blocks before merging. The derivation method can be based on different criteria. For example, the derived motion vector can be an average or a weighted average of the motion vectors of those blocks before merging. It can also be the median of those motion vectors.

[0043] The following describes the derivation of partition information and motion vectors for a 4x4 block. As discussed previously, in the case of extended spatial scalability, it is possible for a 4x4 block in an enhancement layer macroblock (as well as a virtual base layer MB) to be covered by up to four base layer 4x4 blocks. If a single pixel is used in the 4x4 block as a representative for the block, then one of the base layer 4x4 blocks is selected, and the selected block's partition and motion vector information is used as prediction of the current 4x4 block. Such a method may not be accurate, as is demonstrated in Figure 7. In Figure 7, the borders of upsampled 4x4 blocks from the base layer are represented at 700. It can be seen that the first 4x4 block 710 in the virtual base layer macroblock is actually covered by four 4x4 blocks 720, 722, 724 and 726 from the base layer. When using a representative pixel (indicated at 730 in Figure 7), one of the four 4x4 blocks from the base layer (720 in Figure 7) is chosen, and its partition and motion vector information is used as prediction for the first 4x4 block 710, even if the other three 4x4 blocks may have different partition and motion vector information.

[0044] In an embodiment of the present invention, instead of using one pixel as a representative for the 4x4 block, a number of pixels can be used as representatives in deriving the partition and motion vector information from the base layer for the current 4x4 block. For example, as indicated in Figure 8, four samples 800, 810, 820 and 830 can be used as representative pixels for a 4x4 block 840. For each representative samples, a corresponding base layer 4x4 block can be located, with the

base layer 4x4 blocks being 840, 850, 860 and 870 in Figure 8. In total, there can be up to four 4x4 blocks from the base layer. The partition and motion vector information for the current 4x4 block can then be derived from that of the four 4x4 blocks from the base layer.

[0045] To derive partition and motion vector information from multiple 4x4 blocks from the base layer, the following rules are applied in one embodiment. First, when the 4x4 blocks from the base layer have different reference frame indexes, then the block with smaller reference frame index value should be selected. Second, when more than one 4x4 block from the base layer have the smallest reference frame index value, an average or a weighted average of the motion vectors of the 4x4 blocks should be calculated and used as the prediction for the motion vectors in the current 4x4 block. Third, when the weighted average operation described above is used, the weighting coefficient for each motion vector from the base layer can be determined based on a number of factors. These factors include, but not limited to, the size of the area the motion vector represents inside the current 4x4 block, its delta motion vector (i.e. the differential motion vector between a motion vector and its neighboring motion vectors), the motion vector's partition size, the motion vector's macroblock type and mode, etc.

[0046] Figure 9 shows a generic multimedia communications system for use with the present invention. As shown in Figure 4, a data source 100 provides a source signal in an analog, uncompressed digital, or compressed digital format, or any combination of these formats. An encoder 110 encodes the source signal into a coded media bitstream. The encoder 110 may be capable of encoding more than one media type, such as audio and video, or more than one encoder 110 may be required to code different media types of the source signal. The encoder 110 may also get synthetically produced input, such as graphics and text, or it may be capable of producing coded bitstreams of synthetic media. In the following, only processing of one coded media bitstream of one media type is considered to simplify the description. It should be noted, however, that typically real-time broadcast services comprise several streams (typically at least one audio, video and text sub-titling stream). It should also be noted that the system may include many encoders, but in

the following only one encoder 110 is considered to simplify the description without a lack of generality.

[0047] The coded media bitstream is transferred to a storage 120. The storage 120 may comprise any type of mass memory to store the coded media bitstream. The format of the coded media bitstream in the storage 120 may be an elementary self-contained bitstream format, or one or more coded media bitstreams may be encapsulated into a container file. Some systems operate "live", i.e. omit storage and transfer coded media bitstream from the encoder 110 directly to the sender 130. The coded media bitstream is then transferred to the sender 130, also referred to as the server, on a need basis. The format used in the transmission may be an elementary self-contained bitstream format, a packet stream format, or one or more coded media bitstreams may be encapsulated into a container file. The encoder 110, the storage 120, and the sender 130 may reside in the same physical device or they may be included in separate devices. The encoder 110 and sender 130 may operate with live real-time content, in which case the coded media bitstream is typically not stored permanently, but rather buffered for small periods of time in the content encoder 110 and/or in the sender 130 to smooth out variations in processing delay, transfer delay, and coded media bitrate.

[0048] The sender 130 sends the coded media bitstream using a communication protocol stack. The stack may include but is not limited to Real-Time Transport Protocol (RTP), User Datagram Protocol (UDP), and Internet Protocol (IP). When the communication protocol stack is packet-oriented, the sender 130 encapsulates the coded media bitstream into packets. For example, when RTP is used, the sender 130 encapsulates the coded media bitstream into RTP packets according to an RTP payload format. Typically, each media type has a dedicated RTP payload format. It should be again noted that a system may contain more than one sender 130, but for the sake of simplicity, the following description only considers one sender 130.

[0049] The sender 130 may or may not be connected to a gateway 140 through a communication network. The gateway 140 may perform different types of functions, such as translation of a packet stream according to one communication protocol stack to another communication protocol stack, merging and forking of data streams, and

manipulation of data stream according to the downlink and/or receiver capabilities, such as controlling the bit rate of the forwarded stream according to prevailing downlink network conditions. Examples of gateways 140 include multipoint conference control units (MCUs), gateways between circuit-switched and packet-switched video telephony, Push-to-talk over Cellular (PoC) servers, IP encapsulators in digital video broadcasting-handheld (DVB-H) systems, or set-top boxes that forward broadcast transmissions locally to home wireless networks. When RTP is used, the gateway 140 is called an RTP mixer and acts as an endpoint of an RTP connection.

[0050] The system includes one or more receivers 150, typically capable of receiving, de-modulating, and de-capsulating the transmitted signal into a coded media bitstream. The coded media bitstream is typically processed further by a decoder 160, whose output is one or more uncompressed media streams. It should be noted that the bitstream to be decoded can be received from a remote device located within virtually any type of network. Additionally, the bitstream can be received from local hardware or software. Finally, a renderer 170 may reproduce the uncompressed media streams with a loudspeaker or a display, for example. The receiver 150, decoder 160, and renderer 170 may reside in the same physical device or they may be included in separate devices.

[0051] Figures 10 and 11 show one representative communication device 12 within which the present invention may be implemented. It should be understood, however, that the present invention is not intended to be limited to one particular type of communication device 12 or other electronic device. The communication device 12 of Figures 10 and 11 includes a housing 30, a display 32 in the form of a liquid crystal display, a keypad 34, a microphone 36, an ear-piece 38, a battery 40, an infrared port 42, an antenna 44, a smart card 46 in the form of a UICC according to one embodiment of the invention, a card reader 48, radio interface circuitry 52, codec circuitry 54, a controller 56, a memory 58 and a battery 80. Individual circuits and elements are all of a type well known in the art, for example in the Nokia range of mobile telephones.

[0052] Communication devices may communicate using various transmission technologies including, but not limited to, Code Division Multiple Access (CDMA), Global System for Mobile Communications (GSM), Universal Mobile Telecommunications System (UMTS), Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), Transmission Control Protocol/Internet Protocol (TCP/IP), Short Messaging Service (SMS), Multimedia Messaging Service (MMS), e-mail, Instant Messaging Service (IMS), Bluetooth, IEEE 802.11, etc. A communication device may communicate using various media including, but not limited to, radio, infrared, laser, cable connection, and the like.

[0053] The present invention is described in the general context of method steps, which may be implemented in one embodiment by a program product including computer-executable instructions, such as program code, executed by computers in networked environments. A computer-readable medium may include removable and non-removable storage devices including, but not limited to, Read Only Memory (ROM), Random Access Memory (RAM), compact discs (CDs), digital versatile discs (DVD), etc. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. Computer-executable instructions, associated data structures, and program modules represent examples of program code for executing steps of the methods disclosed herein. The particular sequence of such executable instructions or associated data structures represents examples of corresponding acts for implementing the functions described in such steps.

[0054] Software and web implementations of the present invention could be accomplished with standard programming techniques with rule based logic and other logic to accomplish the various database searching steps, correlation steps, comparison steps and decision steps. It should also be noted that the words "component" and "module," as used herein and in the claims, is intended to encompass implementations using one or more lines of software code, and/or hardware implementations, and/or equipment for receiving manual inputs.

[0055] The foregoing description of embodiments of the present invention have been presented for purposes of illustration and description. It is not intended to be

exhaustive or to limit the present invention to the precise form disclosed, and modifications and variations are possible in light of the above teachings or may be acquired from practice of the present invention. The embodiments were chosen and described in order to explain the principles of the present invention and its practical application to enable one skilled in the art to utilize the present invention in various embodiments and with various modifications as are suited to the particular use contemplated. The features of the embodiments described herein may be combined in all possible combinations of methods, apparatus, computer program products and systems.

WHAT IS CLAIMED IS:

1. A method for encoding an enhancement layer block representing at least a portion of a video frame within a scalable bit stream, comprising:

deriving a coding mode indicator specifying a partition of the enhancement layer block based at least in part on a plurality of reference frame index values corresponding to a plurality of reference layer blocks, wherein the reference layer blocks that have the same reference frame index are merged to derive the partition size; and

encoding the enhancement layer block utilizing the coding mode indicator.

2. The method of claim 1, wherein the reference layer blocks that have the same motion vector information are merged to derive the partition size.

3. The method of claim 1, wherein the reference layer blocks that have motion vector information having a difference metric less than a predetermined threshold are merged to derive the partition size.

4. The method of claim 3, wherein the difference metric is derived by a sum of absolute difference of motion vector components, and wherein the threshold value is equal to one.

5. The method of claim 1, wherein a plurality of representative pixels are used to select partition and motion vector information for each respective reference layer block.

6. The method of claim 1, wherein, if the plurality of reference blocks have different reference frame indices, then the motion vector information from the reference block with the lowest reference frame indices is selected.

7. A computer program product, embodied in a computer-readable medium, comprising computer code for performing the processes of claim 1.

8. An apparatus, comprising:

a processor; and

a memory unit communicatively connected to the processor and including:

computer code for deriving a coding mode indicator specifying a partition of an enhancement layer block based at least in part on a plurality of reference frame index values corresponding to a plurality of reference layer blocks, wherein the reference layer blocks that have the same reference frame index are merged to derive the partition size; and

computer code for encoding the enhancement layer block utilizing the coding mode indicator.

9. The apparatus of claim 8, wherein the reference layer blocks that have the same motion vector information are merged to derive the partition size.

10. The apparatus of claim 8, wherein the reference layer blocks that have motion vector information having a difference metric less than a predetermined threshold are merged to derive the partition size.

11. The apparatus of claim 10, wherein the difference metric is derived by a sum of absolute difference of motion vector components, and wherein the threshold value is equal to one.

12. The apparatus of claim 8, wherein a plurality of representative pixels are used to select partition and motion vector information for each respective reference layer block.

13. The apparatus of claim 8, wherein, if the plurality of reference blocks have different reference frame indices, then the motion vector information from the reference block with the lowest reference frame indices is selected.

14. An apparatus, comprising:

means for deriving a coding mode indicator specifying a partition of an enhancement layer block based at least in part on a plurality of reference frame index values corresponding to a plurality of reference layer blocks, wherein the reference layer blocks that have the same reference frame index are merged to derive the partition size; and

encoding the enhancement layer block utilizing the coding mode indicator.

15. A method for decoding an enhancement layer block representing at least a portion of a video frame within a scalable bit stream, comprising:

deriving a coding mode indicator specifying a partition of the enhancement layer block based at least in part on a plurality of reference frame index values corresponding to a plurality of reference layer blocks, wherein the reference layer blocks that have the same reference frame index are merged to derive the partition size; and

decoding the enhancement layer block utilizing the coding mode indicator.

16. The method of claim 15, wherein the reference layer blocks that have the same motion vector information are merged to derive the partition size.

17. The method of claim 15, wherein the reference layer blocks that have motion vector information having a difference metric less than a predetermined threshold are merged to derive the partition size.

18. The method of claim 17, wherein the difference metric is derived by a sum of absolute difference of motion vector components, and wherein the threshold value is equal to one.

19. The method of claim 15, wherein a plurality of representative pixels are used to select partition and motion vector information for the respective reference layer block.

20. The method of claim 15, wherein, if the plurality of reference blocks have different reference frame indices, then the motion vector information from the block with the lowest reference frame indices is selected.

21. A computer program product, embodied in a computer-readable medium, comprising computer code for performing the processes of claim 15.

22. An apparatus, comprising:

a processor; and

a memory unit communicatively connected to the processor and including:

computer code for deriving a coding mode indicator specifying a partition of an enhancement layer block based at least in part on a plurality of reference frame index values corresponding to a plurality of reference layer blocks, wherein the reference layer blocks that have the same reference frame index are merged to derive the partition size; and

decoding the enhancement layer block utilizing the coding mode indicator.

23. The apparatus of claim 22, wherein the reference layer blocks that have the same motion vector information are merged to derive the partition size.

24. The apparatus of claim 22, wherein the reference layer blocks that have motion vector information having a difference metric less than a predetermined threshold are merged to derive the partition size.

25. The apparatus of claim 24, wherein the difference metric is derived by a sum of absolute difference of motion vector components, and wherein the threshold value is equal to one.

26. The apparatus of claim 22, wherein a plurality of representative pixels are used to select partition and motion vector information for the respective reference layer block.

27. The apparatus of claim 22, wherein, if the plurality of reference blocks have different reference frame indices, then the motion vector information from the block with the lowest reference frame indices is selected.

28. An apparatus, comprising:

means for deriving a coding mode indicator specifying a partition of an enhancement layer block based at least in part on a plurality of reference frame index values corresponding to a plurality of reference layer blocks, wherein the reference layer blocks that have the same reference frame index are merged to derive the partition size; and

means for decoding the enhancement layer block utilizing the coding mode indicator.

29. A method for encoding an enhancement layer block representing at least a portion of a video frame within a scalable bit stream, comprising:

deriving partition and motion vector information for the enhancement layer block based upon partition and motion vector information from a plurality of reference blocks; and

encoding the enhancement layer block utilizing the partition and motion vector information.

30. The method of claim 29, wherein, if each of the mapped reference blocks have different reference frame indices, then the motion vector information from the reference block with the lowest reference frame index is selected to derive the partition and motion vector information for the enhancement layer block.

31. The method of claim 29, wherein, if more than one of the mapped blocks have the smallest reference frame indices, then an average of the motion vector information for the respective reference blocks with the smallest reference frame indices is selected to derive the partition and motion vector information for the enhancement layer block.

32. The method of claim 31, wherein the average comprises a weighted average using weighting coefficients, and wherein the weighting coefficient for the motion vector in each mapped reference block is determined using at least one of a size of the area the motion vector represents within the reference block; a delta motion vector for the motion vector; a partition size for the reference block; a block type of the reference block; and a block mode of the reference layer block.

33. A computer program product, embodied in a computer-readable medium, comprising computer code for performing the processes of claim 29.

34. An apparatus, comprising:

a processor; and

a memory unit communicatively connected to the processor and including:

computer code for deriving partition and motion vector information for an enhancement layer block based upon partition and motion vector information from a plurality of reference blocks; and

computer code for encoding the enhancement layer block utilizing the partition and motion vector information.

35. The apparatus of claim 34, wherein, if each of the mapped reference blocks have different reference frame indices, then the motion vector information from the reference block with the lowest reference frame index is selected to derive the partition and motion vector information for the enhancement layer block.

36. The apparatus of claim 34, wherein, if more than one of the mapped blocks have the smallest reference frame indices, then an average of the motion vector information for the respective reference blocks with the smallest reference frame indices is selected to derive the partition and motion vector information for the enhancement layer block.

37. The apparatus of claim 36, wherein the average comprises a weighted average using weighting coefficients, and wherein the weighting coefficient for the

motion vector in each mapped reference block is determined using at least one of a size of the area the motion vector represents within the reference block; a delta motion vector for the motion vector; a partition size for the reference block; a block type of the reference block; and a block mode of the reference layer block.

38. An apparatus, comprising:

means for deriving partition and motion vector information for an enhancement layer block based upon partition and motion vector information from a plurality of reference blocks; and

means for encoding the enhancement layer block utilizing the partition and motion vector information.

39. A method of decoding an enhancement layer block representing at least a portion of a video frame within a scalable bit stream, comprising:

deriving partition and motion vector information for the enhancement layer block based upon partition and motion vector information from a plurality of reference blocks; and

decoding the enhancement layer block utilizing the partition and motion vector information.

40. The method of claim 39, wherein, if each of the mapped reference blocks have different reference frame indices, then the motion vector information from the reference block with the lowest reference frame index is selected to derive the partition and motion vector information for the enhancement layer block.

41. The method of claim 39, wherein, if more than one of the mapped blocks have the smallest reference frame indices, then an average of the motion vector information for the respective reference blocks with the smallest reference frame indices is selected to derive the partition and motion vector information for the enhancement layer block.

42. The method of claim 41, wherein the average comprises a weighted average using weighting coefficients, and wherein the weighting coefficient for the motion vector in each mapped reference block is determined using at least one of a

size of the area the motion vector represents within the reference block; a delta motion vector for the motion vector; a partition size for the reference block; a block type of the reference block; and a block mode of the reference layer block.

43. A computer program product, embodied in a computer-readable medium, comprising computer code for performing the processes of claim 39.

44. An apparatus, comprising:

a processor; and

a memory unit communicatively connected to the processor and including:

computer code for deriving partition and motion vector information for an enhancement layer block based upon partition and motion vector information from a plurality of reference blocks; and

computer code for decoding the enhancement layer block utilizing the partition and motion vector information.

45. The apparatus of claim 44, wherein, if each of the mapped reference blocks have different reference frame indices, then the motion vector information from the reference block with the lowest reference frame index is selected to derive the partition and motion vector information for the enhancement layer block.

46. The apparatus of claim 44, wherein, if more than one of the mapped blocks have the smallest reference frame indices, then an average of the motion vector information for the respective reference blocks with the smallest reference frame indices is selected to derive the partition and motion vector information for the enhancement layer block.

47. The apparatus of claim 46, wherein the average comprises a weighted average using weighting coefficients, and wherein the weighting coefficient for the motion vector in each mapped reference block is determined using at least one of a size of the area the motion vector represents within the reference block; a delta motion vector for the motion vector; a partition size for the reference block; a block type of the reference block; and a block mode of the reference layer block.

48. An apparatus, comprising:

means for deriving partition and motion vector information for an enhancement layer block based upon partition and motion vector information from a plurality of reference blocks; and

means for decoding the enhancement layer block utilizing the partition and motion vector information.