

US00RE44285E

(19) **United States**
(12) **Reissued Patent**
Kerofsky

(10) **Patent Number:** **US RE44,285 E**
(45) **Date of Reissued Patent:** ***Jun. 11, 2013**

(54) **SYSTEMS AND METHODS FOR REDUCED BIT-DEPTH PROCESSING IN VIDEO-RELATED DATA WITH FREQUENCY WEIGHTING MATRICES**

(75) Inventor: **Louis Joseph Kerofsky**, Camas, WA (US)

(73) Assignee: **Sharp Kabushiki Kaisha**, Osaka (JP)

(*) Notice: This patent is subject to a terminal disclaimer.

(21) Appl. No.: **13/301,430**

(22) Filed: **Nov. 21, 2011**

Related U.S. Patent Documents

Reissue of:

(64) Patent No.: **7,400,682**
Issued: **Jul. 15, 2008**
Appl. No.: **10/931,558**
Filed: **Aug. 31, 2004**

U.S. Applications:

(60) Division of application No. 12/837,154, filed on Jul. 15, 2010, now Pat. No. Re. 43,091, which is a continuation of application No. 12/689,897, filed on Jan. 19, 2010, now Pat. No. Re. 42,745, which is a continuation of application No. 10/326,459, filed on Dec. 20, 2002, now Pat. No. 7,170,942, which is a continuation of application No. 10/139,036, filed on May 2, 2002, now Pat. No. 7,123,655.

(60) Provisional application No. 60/319,018, filed on Nov. 30, 2001, provisional application No. 60/311,436, filed on Aug. 9, 2001.

(51) **Int. Cl.**
H04N 7/12 (2006.01)

(52) **U.S. Cl.**
USPC **375/240.18**; 375/240.01; 375/E7.028; 375/E7.093; 375/E7.139; 375/E7.14; 375/E7.211; 375/E7.226; 375/E7.232

(58) **Field of Classification Search**
USPC 375/240, 240.01, 240.12, 240.18, 375/E7.028, E7.093, E7.139, E7.14, E7.211, 375/E7.226, E7.232
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,230,038 A 7/1993 Fielder et al.
5,345,408 A 9/1994 Hoogenboom
5,471,412 A 11/1995 Shyu
5,479,364 A 12/1995 Jones et al.

(Continued)

FOREIGN PATENT DOCUMENTS

CA 222118 A1 12/1993
GB 2264609 A 9/1993

(Continued)

OTHER PUBLICATIONS

Bjontegaard, "H.26L Test Model Long Term No. 7 (TML-7) Draft0", ITU—Telecommunications Standardization Sector, Study Group 16, Video Coding Experts Group (VCEG), Document VCEG-M81, Thirteenth Meeting, Austin, Texas, USA, Apr. 2-4, 2001, pp. 1-36.

(Continued)

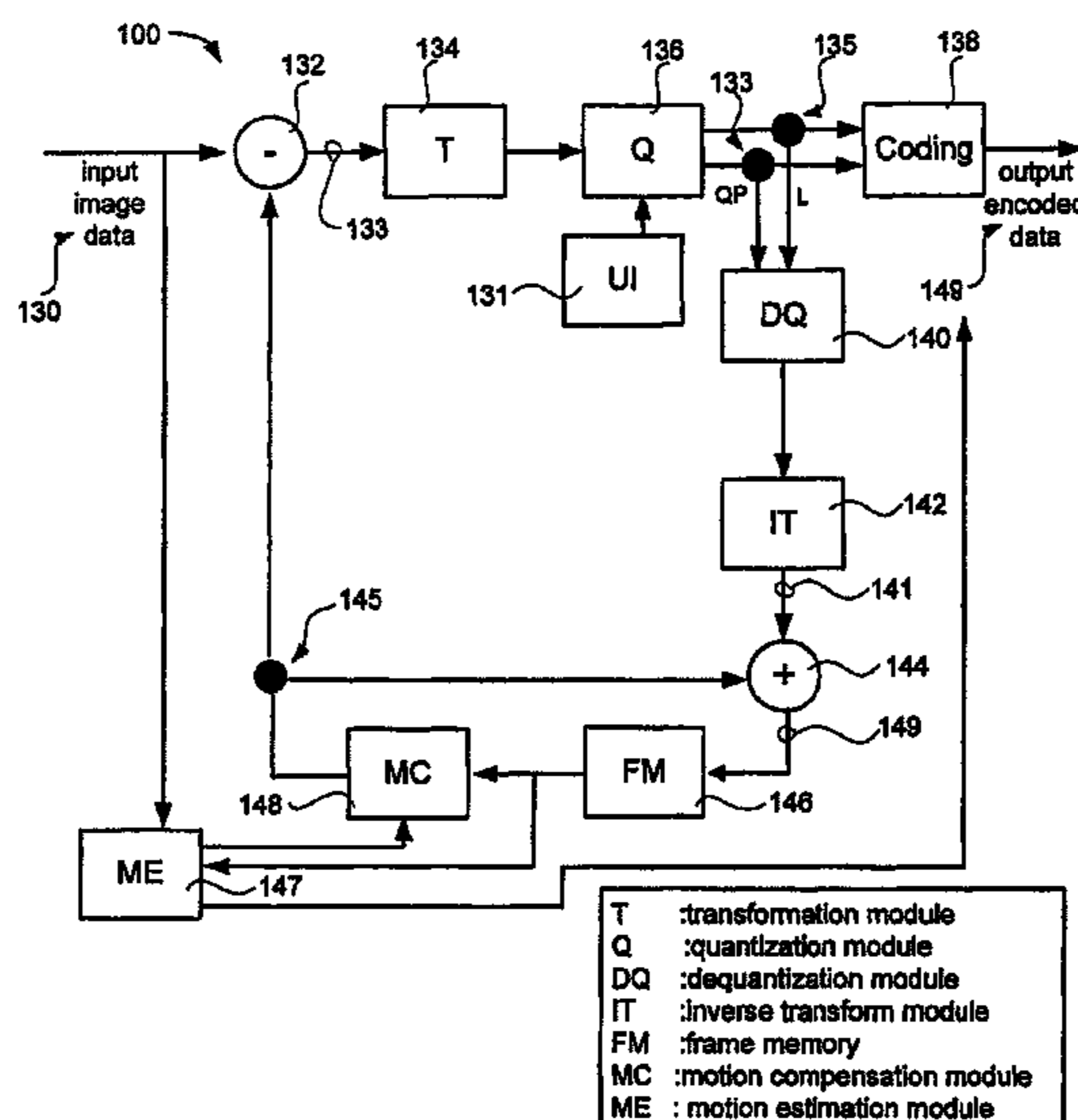
Primary Examiner — Allen Wong

(74) *Attorney, Agent, or Firm* — Birch, Stewart, Kolasch & Birch, LLP

(57) **ABSTRACT**

Embodiments of the present invention comprise systems and methods for processing of data related to video wherein reduced bit depth intermediate calculations are enabled.

2 Claims, 10 Drawing Sheets



U.S. PATENT DOCUMENTS

5,590,067	A	12/1996	Jones et al.
5,594,678	A	1/1997	Jones et al.
5,596,517	A	1/1997	Jones et al.
5,640,159	A	6/1997	Furlan et al.
5,748,793	A	5/1998	Sanpei
5,754,457	A	5/1998	Eitan et al.
5,764,553	A	6/1998	Hong
5,822,003	A	10/1998	Girod et al.
5,845,112	A	12/1998	Nguyen et al.
6,081,552	A	6/2000	Stevenson et al.
6,160,920	A	12/2000	Shyu
6,212,236	B1	4/2001	Nishida et al.
6,856,262	B2	2/2005	Mayer et al.
6,876,703	B2	4/2005	Ismaeil et al.
2003/0099291	A1	5/2003	Kerofsky
2003/0112876	A1	6/2003	Kerofsky
2003/0123553	A1	7/2003	Kerofsky
2004/0013202	A1	1/2004	Lainema
2004/0046754	A1	3/2004	Mayer et al.

FOREIGN PATENT DOCUMENTS

JP	3-270573	A	12/1991
JP	4-503136	A	6/1992
JP	4-504192	A	7/1992
JP	04-222121		8/1992
JP	4-222121	A	8/1992
JP	06-046269		2/1993
JP	5-95483	A	4/1993
JP	07-099578		9/1993
JP	05-307467		11/1993
JP	5-307467	A	11/1993
JP	6-46269	A	2/1994
JP	06-053839		2/1994
JP	6-53839	A	2/1994
JP	06-077842		3/1994
JP	6-77842	A	3/1994
JP	7-99578	A	4/1995
JP	2004-506990	A	3/2004
KR	0172902		3/1999
KR	0172902	B1	3/1999
WO	WO 90/09022	A1	8/1990
WO	WO 90/09064	A1	8/1990

OTHER PUBLICATIONS

Bjontegaard, "H.26L Test Model Long Term No. 8 (TML-8) Draft0", ITU—Telecommunications Standardization Sector, Study Group 16, Video Coding Experts Group (VCEG), Document VCEG-Nxx, Apr. 2, 2001 (Generated: Jun. 28, 2001), pp. 1-46.

Extended European Search Report, dated Nov. 4, 2010, for European Application No. 10179630.8.

Extended European Search Report, dated Nov. 4, 2010, for European Application No. 10179636.5.

Hallapuro et al., "Low Complexity (I)DCT", ITU-Telecommunications Standardization Sector, Study Group 16 Question 6, Video Coding Experts Group (VCEG), 15th Meeting: Pattaya, Thailand, Dec. 4-6, 2001, pp. 1-11.

Hallapuro et al., "Low Complexity Transform and Quantization—Part II: Extensions," Document: JVT-B039, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6), 2nd Meeting: Geneva, CH, Jan. 29—Feb. 1, 2002, pp. 1-14.

Hallapuro, et al., "Low Complexity Transform and Quantization-Part I: Basic Implementation," Joint Video Team (JVT) of ISO/IEC

MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6), Document: JVT-B038, XP030005037, Jan. 14, 2002, pp. 1-18.

Joint Video Team, T. Wiegand (Contact), "Draft ITU-T Recommendation H.264 (a.k.a. "H.26L")", ITU-Telecommunications Standardization Sector, Study Group 16 Question 6, Video Coding Experts Group (VCEG), Document: VCEG-P07, 16th Meeting: Fairfax, Virginia, USA, May 6-10, 2002, 141 pages.

Kerofsky et al., "Reduced Bit-Depth Quantization," ITU—Telecommunications Standardization Sector, Study Group 16 Question 6, Video Coding Experts Group (VCEG), Document VCEG-N20, Fourteenth Meeting: Santa Barbara, CA, USA, Sep. 24-27, 2001, pp. 1-14.

Kerofsky, "Modifications to the JVT IDCT", Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6), Document: JVT-C25, 3rd Meeting: Fairfax, Virginia, USA, May 6-10, 2002, pp. 1-14.

Kerofsky, "Notes on JVT IDCT", Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6), Document: JVT-C24, 3rd Meeting: Fairfax, Virginia, USA, May 6-10, 2002, pp. 1-8.

Klomp et al., "TE1: Cross-Checking Results of DMVD Proposal JCTVC-B076 (MediaTek Inc.)", Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Document: JCTVC-B119, 2nd Meeting: Geneva, CH, Jul. 21-28, 2010, 1 page.

Korean Office Action, dated Jan. 13, 2006, for Korean Application No. 10-2004-7001865 (English translation only provided).

Lepley et al., "Report on Core Experiment CodEff9: Integer Quantization", Coding of Still Pictures, ISO/IEC JTC1/SC29/WG1 (ITU-T SG8), XP017205064, Oct. 21, 1998, 5 pages.

Liang et al., "A 16-bit Architecture for H.26L, Treating DCT Transforms and Quantization", ITU—Telecommunications Standardization Sector, Study Group 16 Question 6, Video Coding Experts Group (VCEG), Document VCEG-M16, XP002332050, Austin, TX, USA, Apr. 2-4, 2001, pp. 1-17.

Sun et al., "Global Motion Vector Coding (GMVC)", ITU—Telecommunications Standardization Sector, Study Group 16 Question 6, Video Coding Experts Group (VCEG), Document: VCEG-O20, Fifteenth Meeting: Pattaya, Thailand, Dec. 4-7, 2001, pp. 1-10.

"H.26L Test Model Long Term No. 8(TML-8) Draft0" ITU-T Telecommunication Standardization Sector of ITU, Geneva CH, Apr. 2, 2001, pp. 1-54 XP001089814.

Liang J. , Tran T. , Topiwala P.: "a 16-bit architecture for H.26L, treating DCT transforms and quantization" Document VCEG-M16, ITU—Telecommunications Standardization Sector, Study Group 16 Question 6, Video Coding Experts Group (VCEG), 'Online! Apr. 2, 2001, pp. 1-17, XP002332050 Austin, Texas, USA Retrieved from the Internet ; URL: http://www.ensc.sfu.ca/people/faculty/jiel/papers/H26L_Proposal.doc !retrieved on Jun. 8, 2005 abstract Section "Introduction", p. 1 Section 1.2 "Quantization in TML 5.2", pp. 2-3 Section 2.4 "Scaling factors and Quantization", pp. 6-7.

Kerofsky L., Lei S.; "Reduced bit-depth quantization" Document VCEG-N20, ITU-Telecommunications Standardization Sector, Study Group 16 Question 6: Video Coding Experts Group (VCEG), 'Online! Sep. 24, 2001 , pp. 1-14 , XP002332051 Santa Barbara, CA USA Retrieved from the internet: URL: ftp3.itu.int/av-arch/video-site/0109_SAN/vceg-n20-DOC> Retrieved on Jun. 8, 2008! Section "Introduction", pp. 1-2 Section "Proposed Quantization values", pp. 3-4 Section "Appendix", Table 10 and Equations 4,5.

Gisele Bjontegard, "H.26L Test Model Long Term No. 8 (TML-8) draft0", document VCEG-M81, ITU-T Video Coding Experts Group (VCEG), Austin, Texas, Apr. 2001.

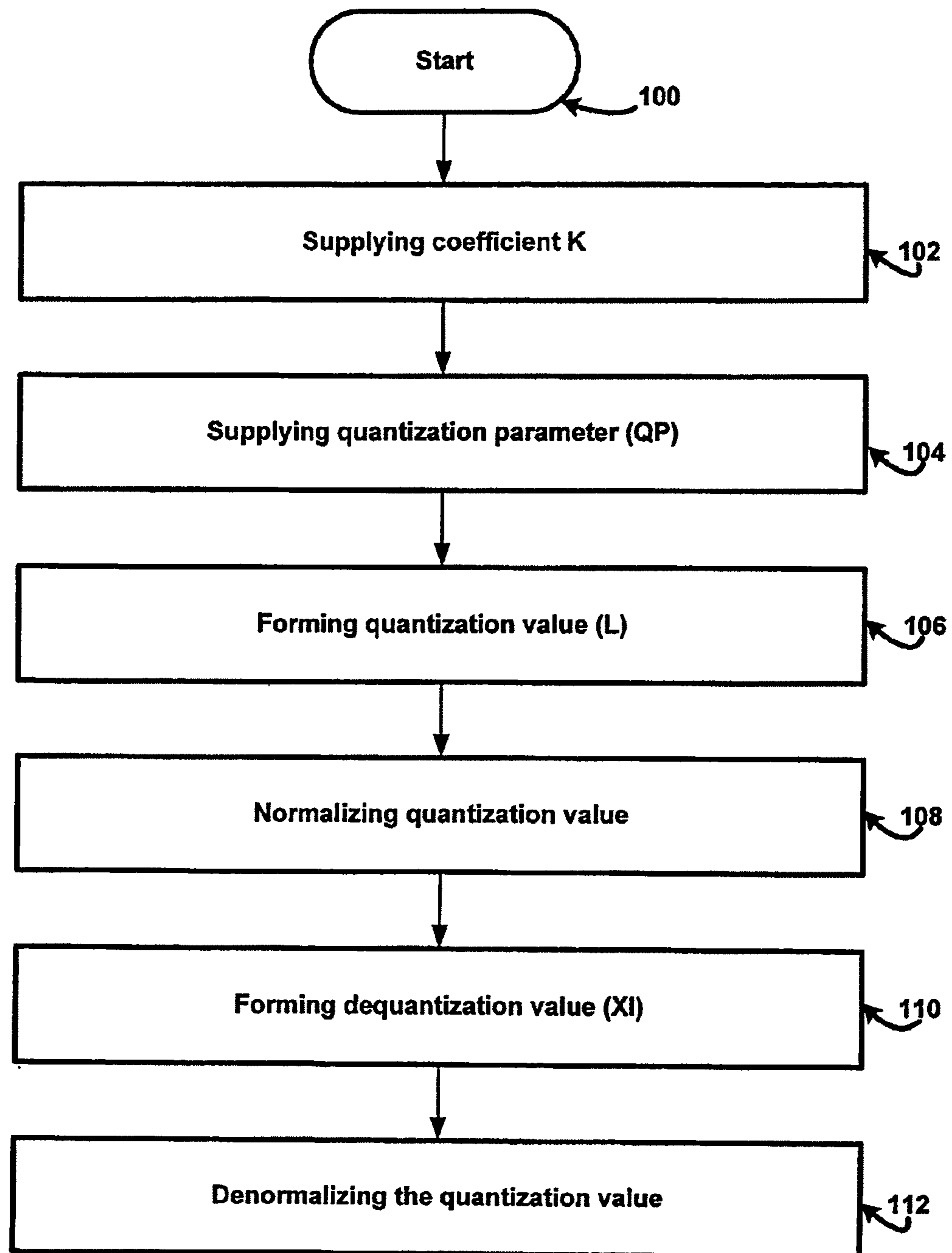


Figure 1

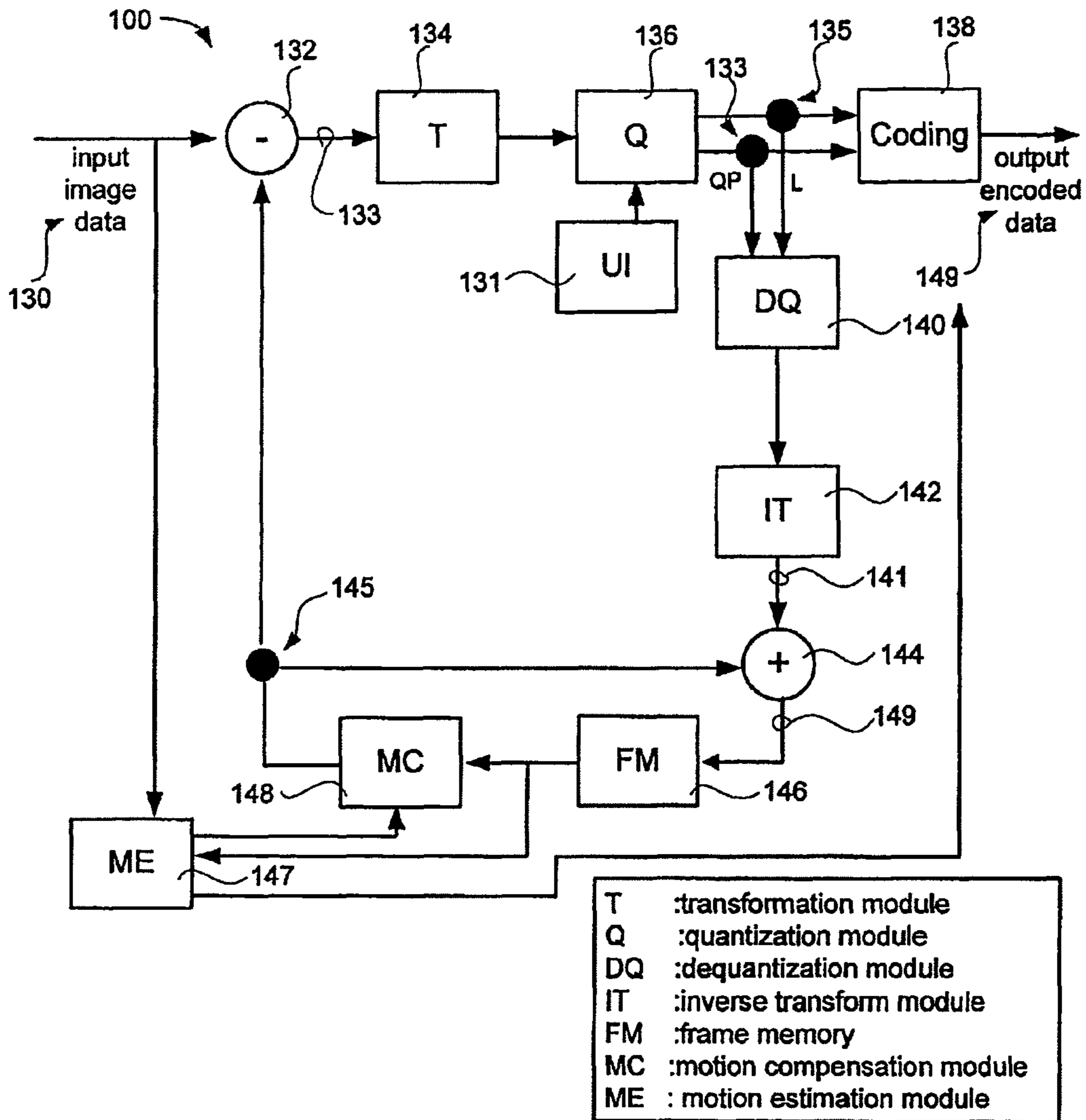


Figure 2

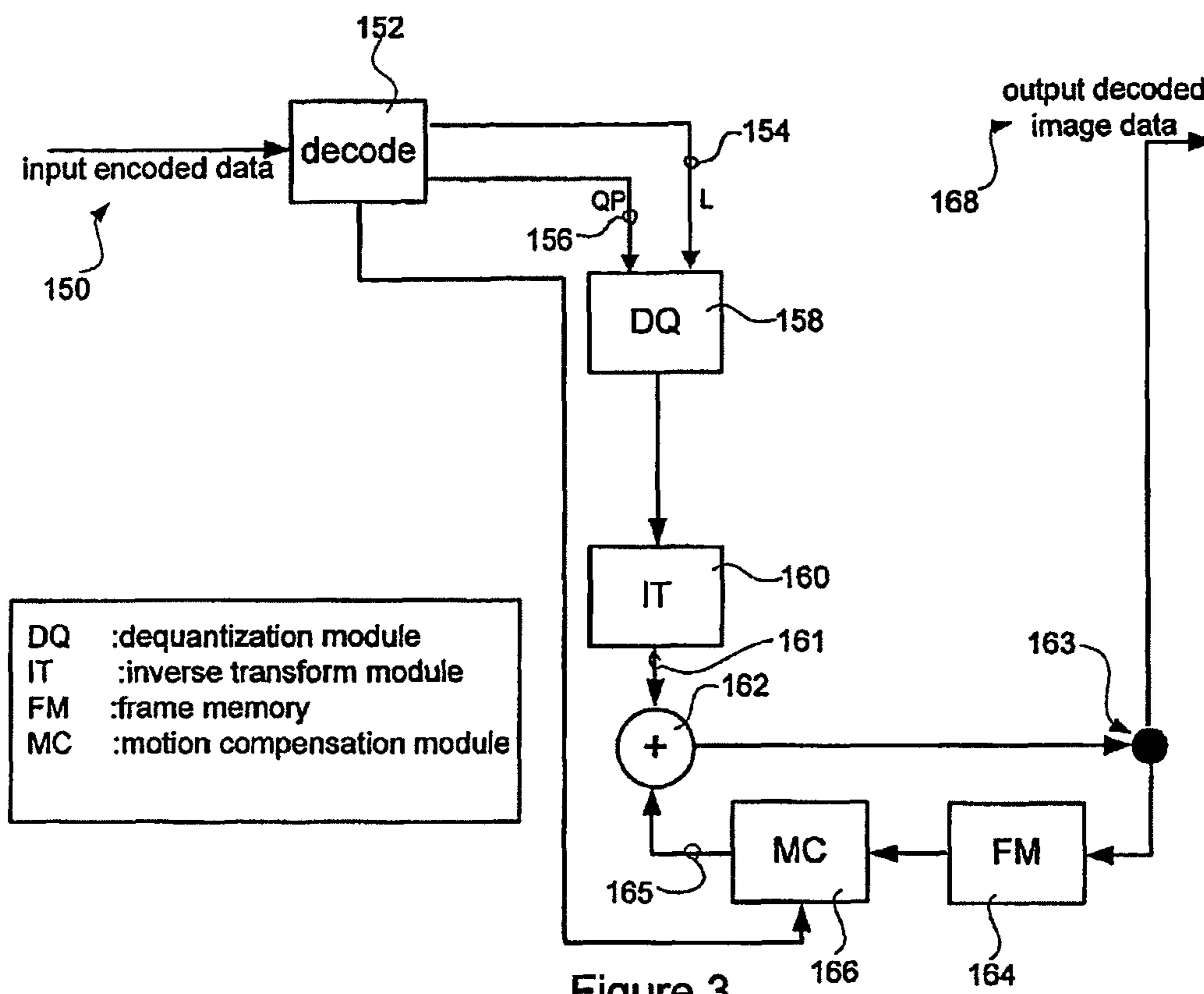


Figure 3

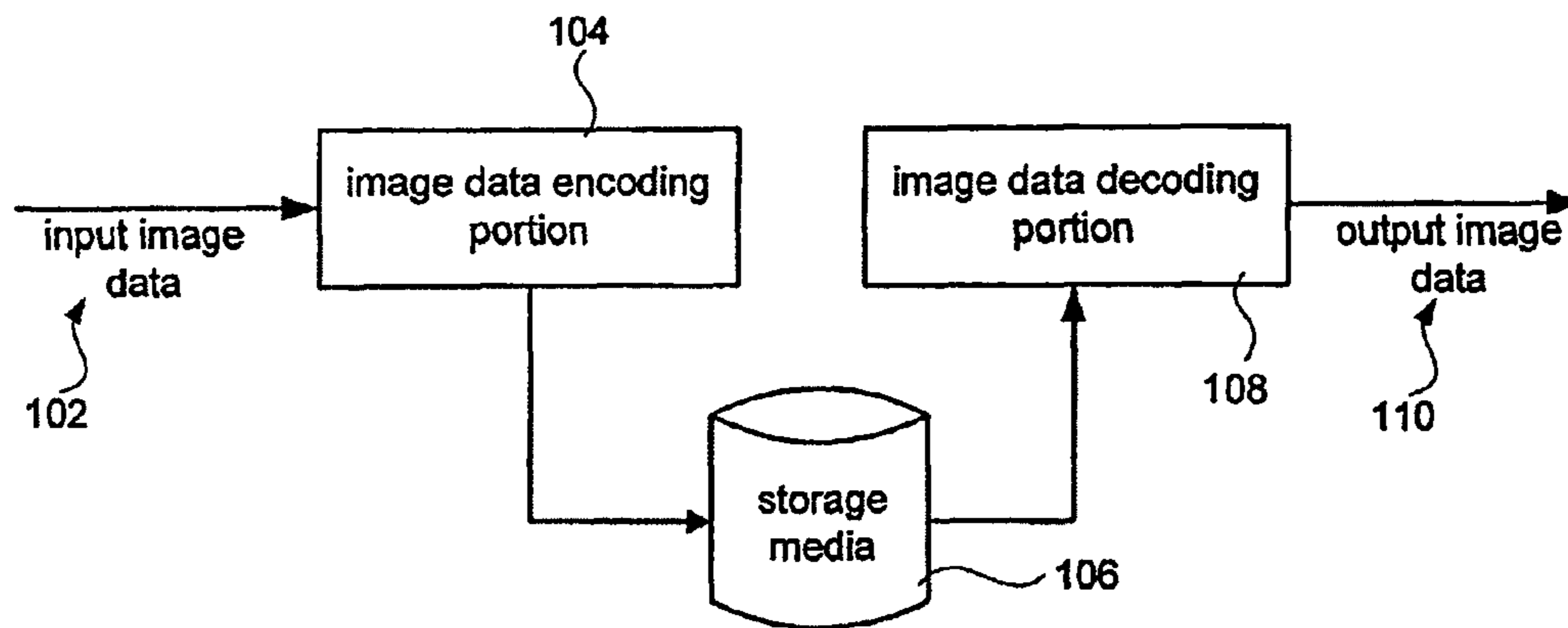


Figure 4

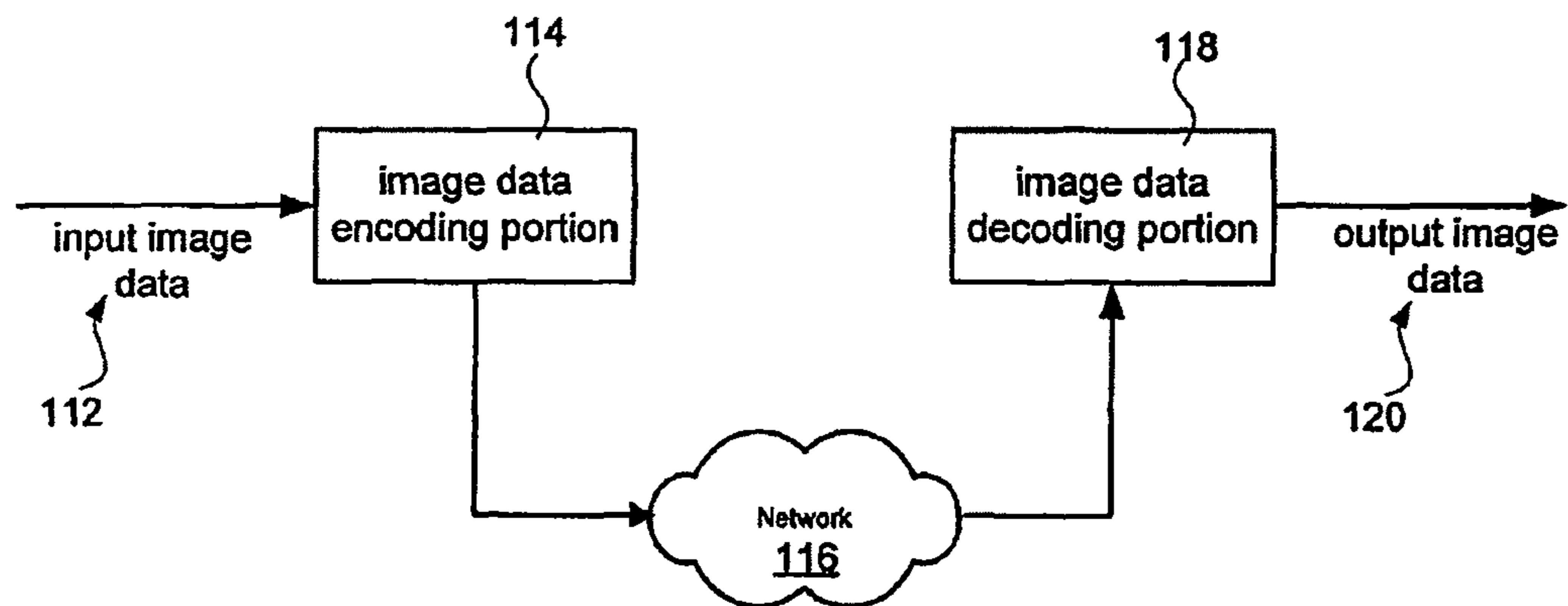


Figure 5

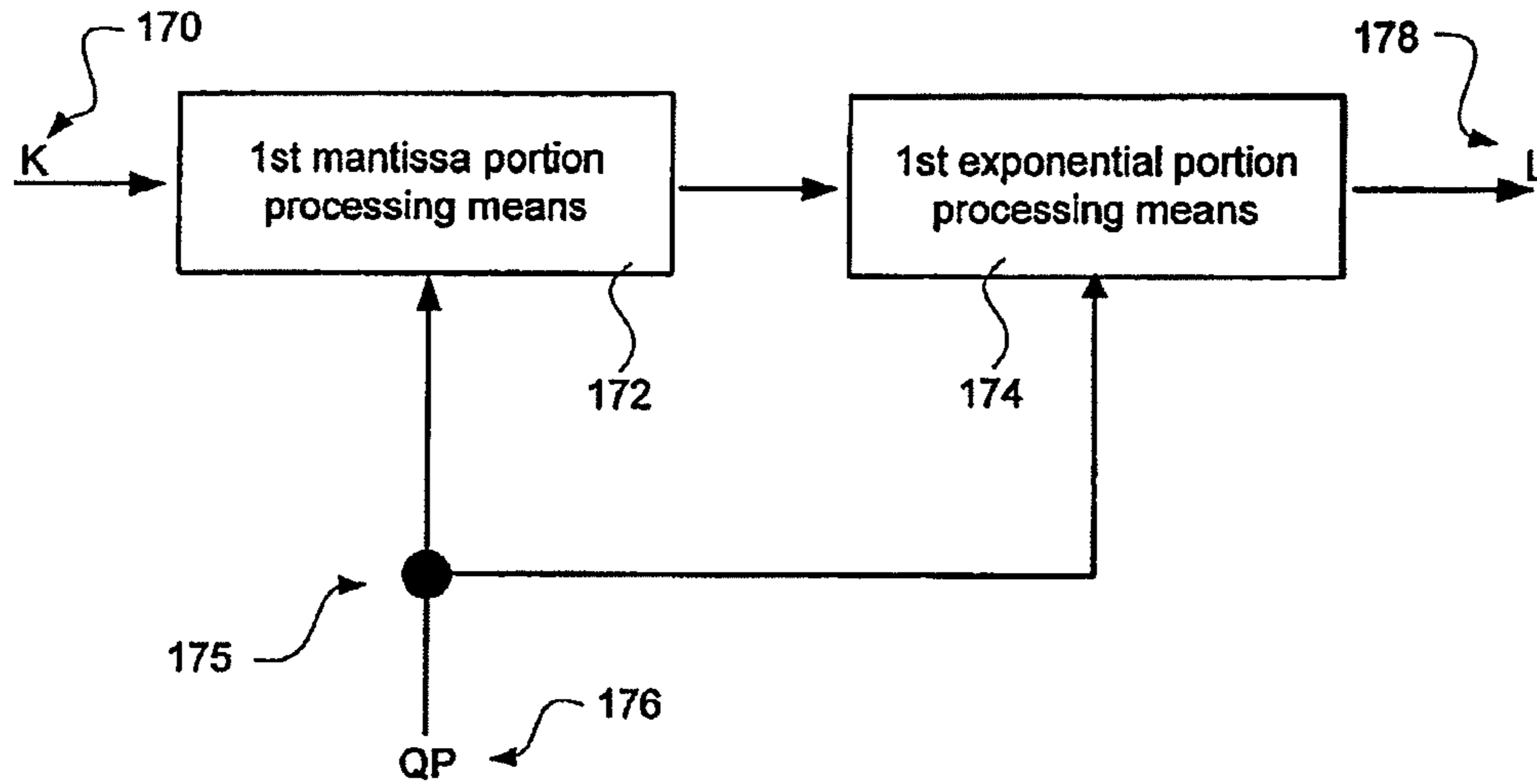


Figure 6

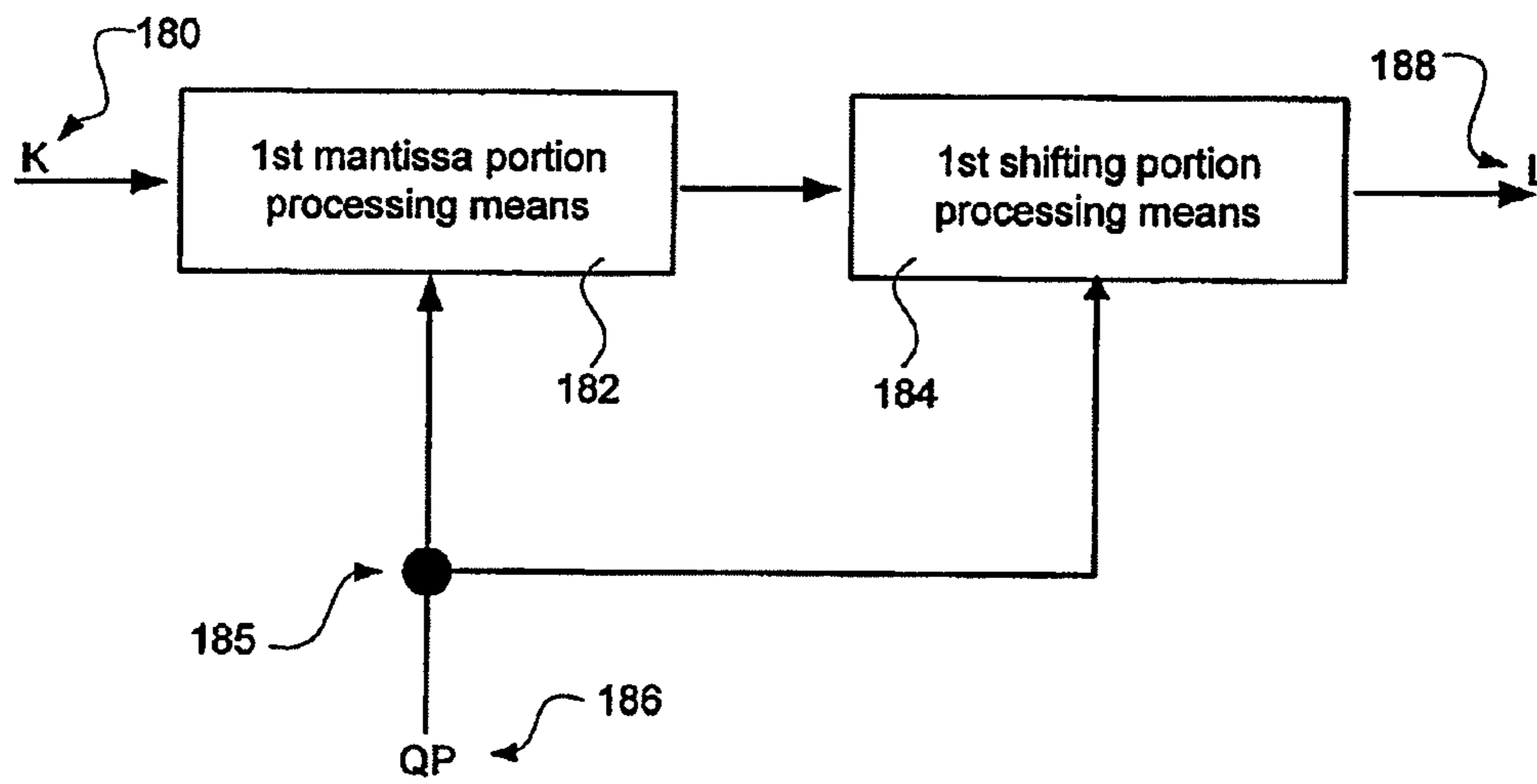


Figure 7

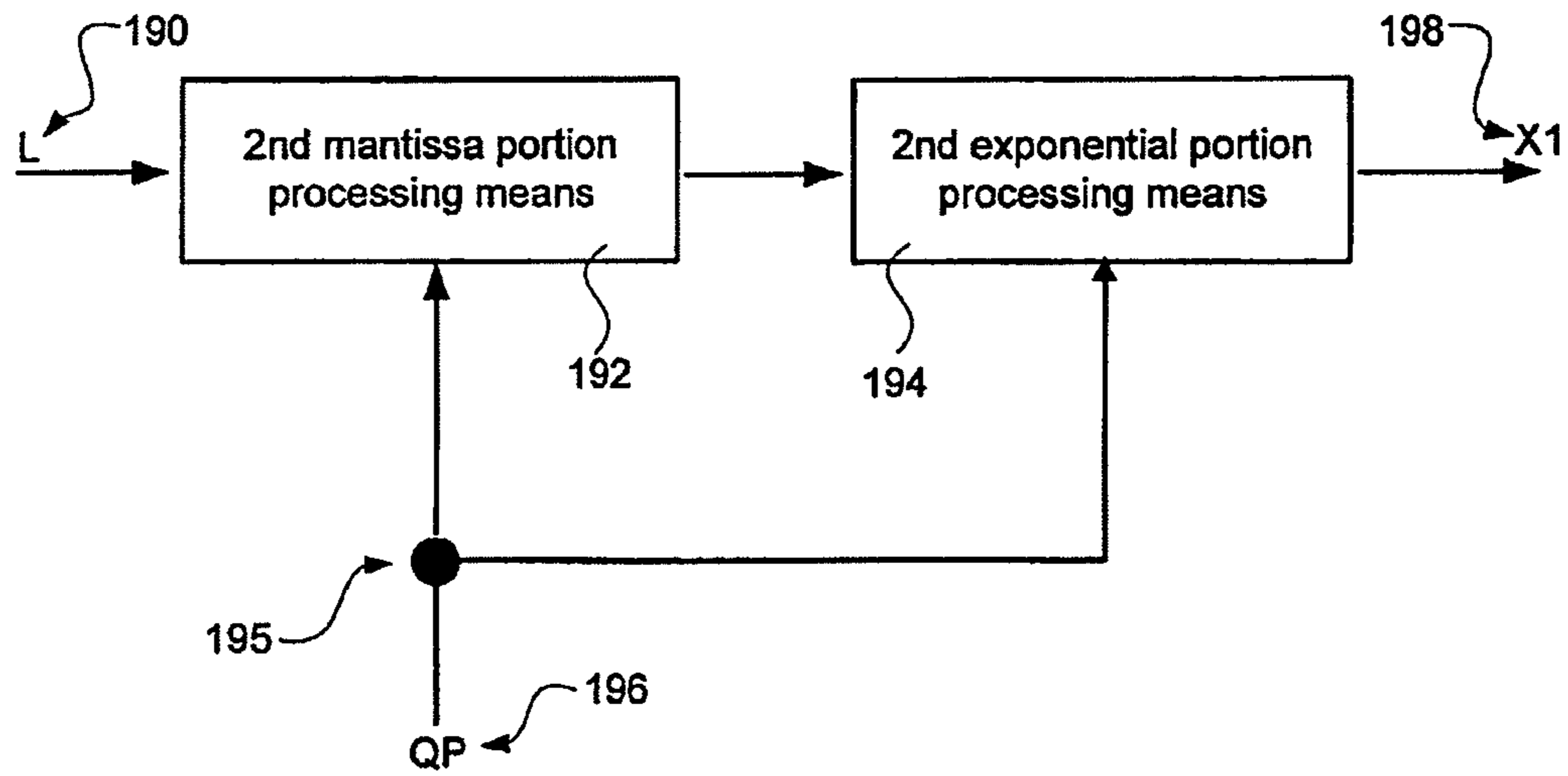


Figure 8

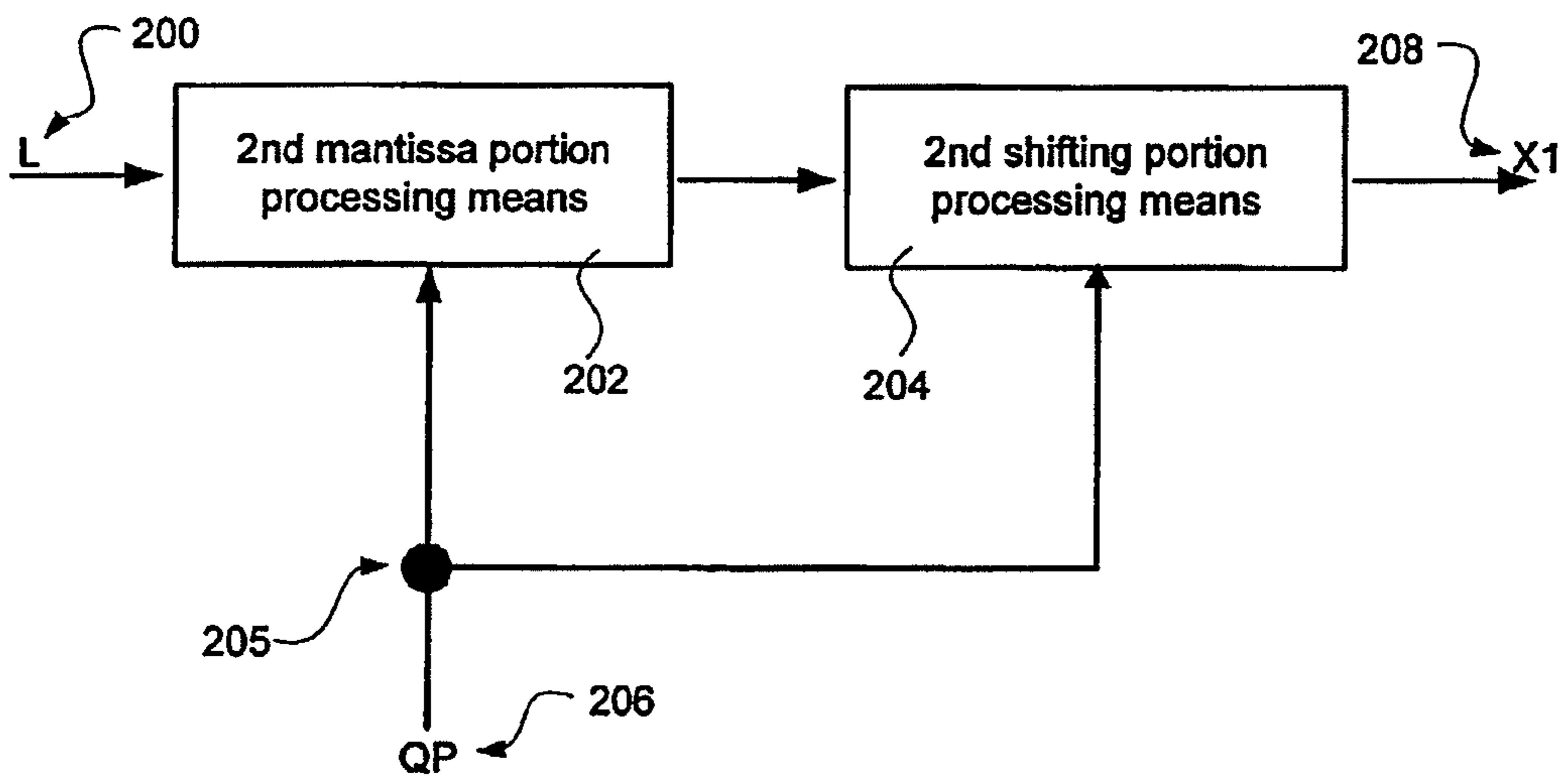
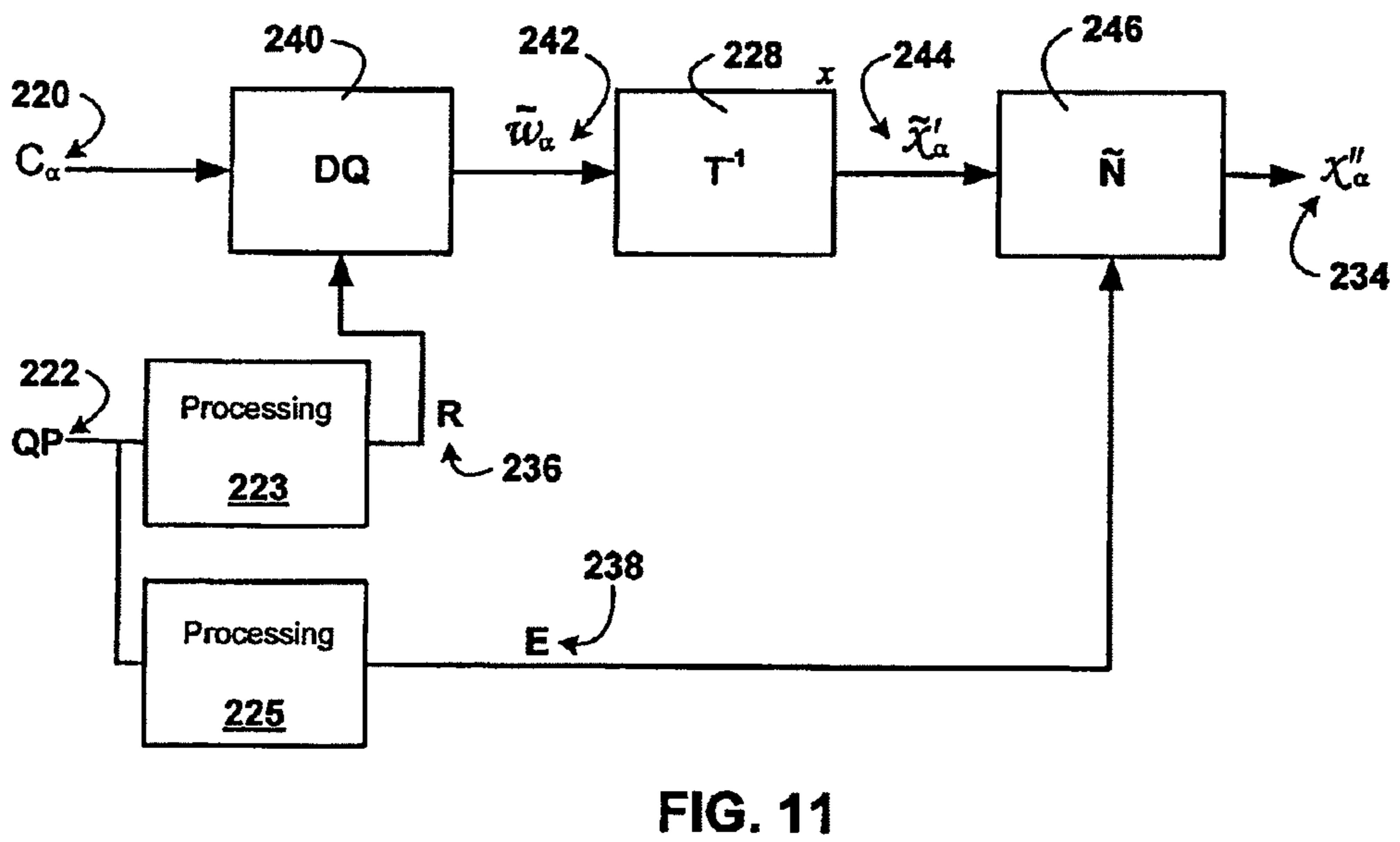
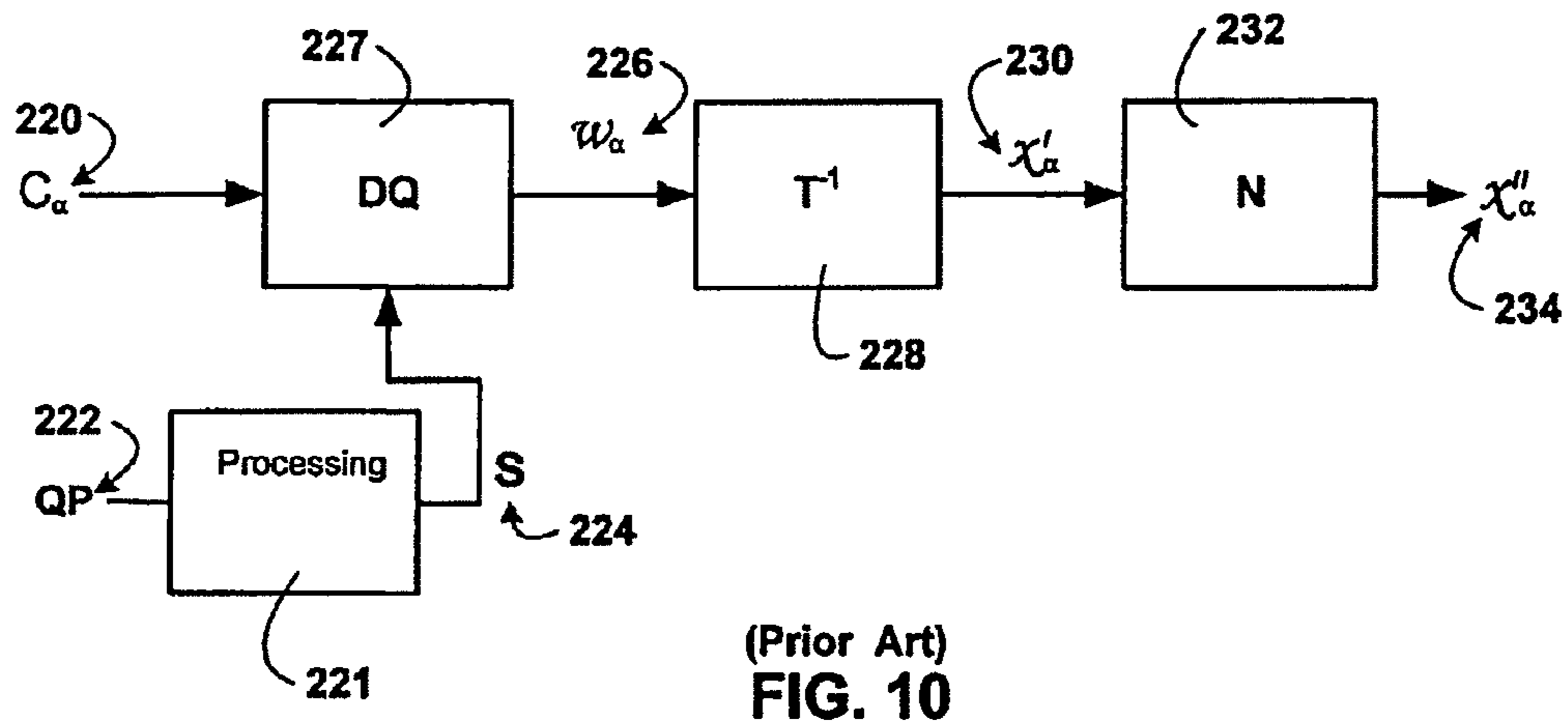


Figure 9



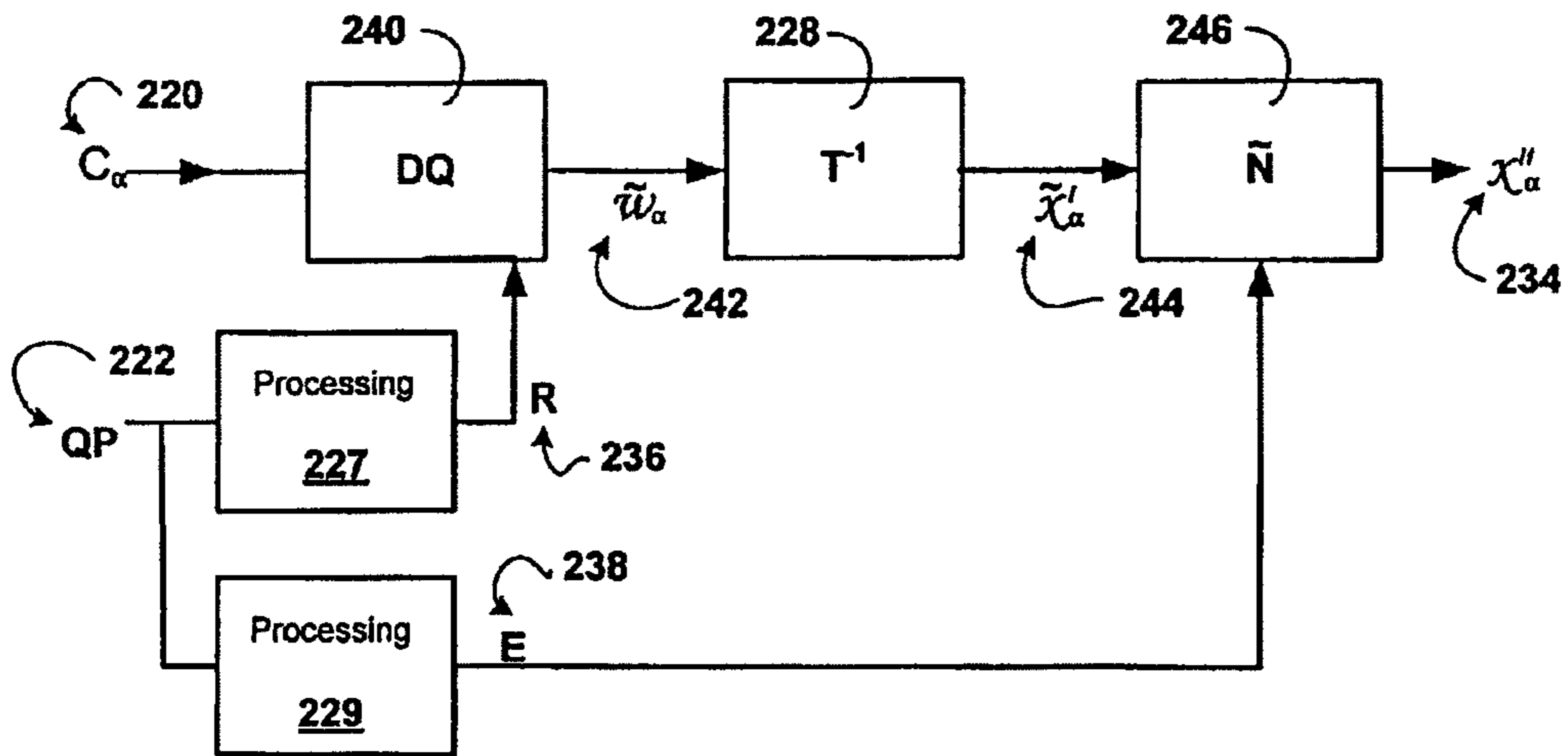


FIG. 12

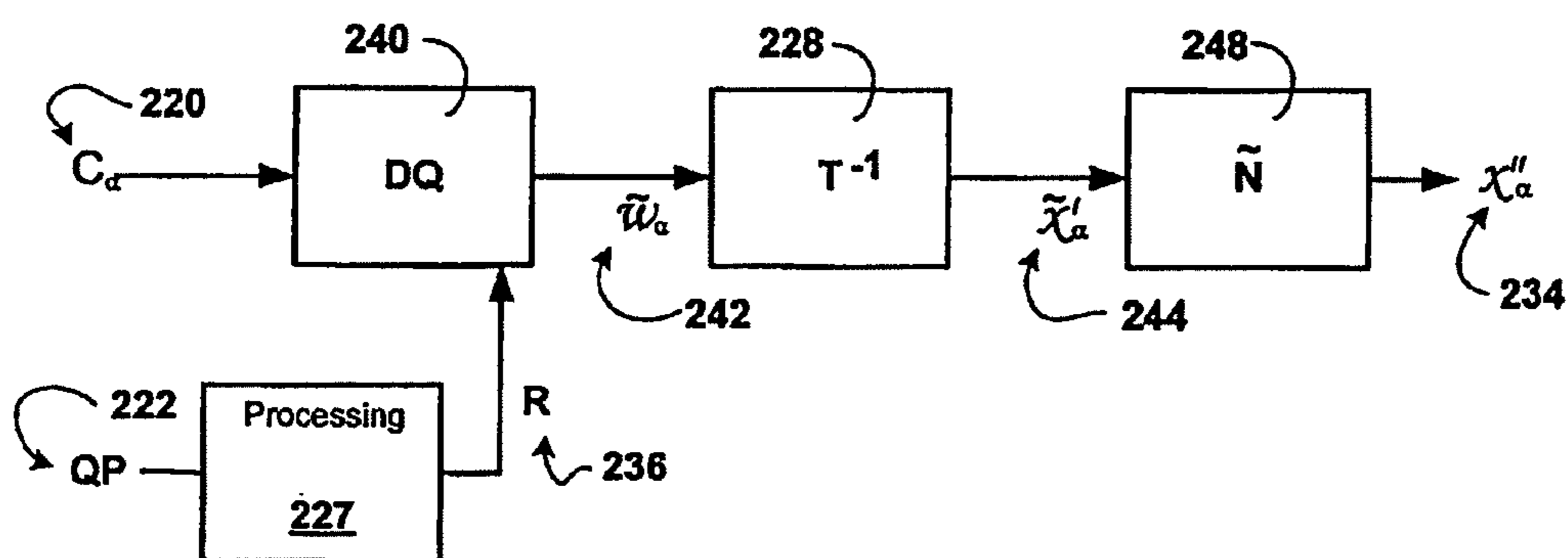


FIG. 13

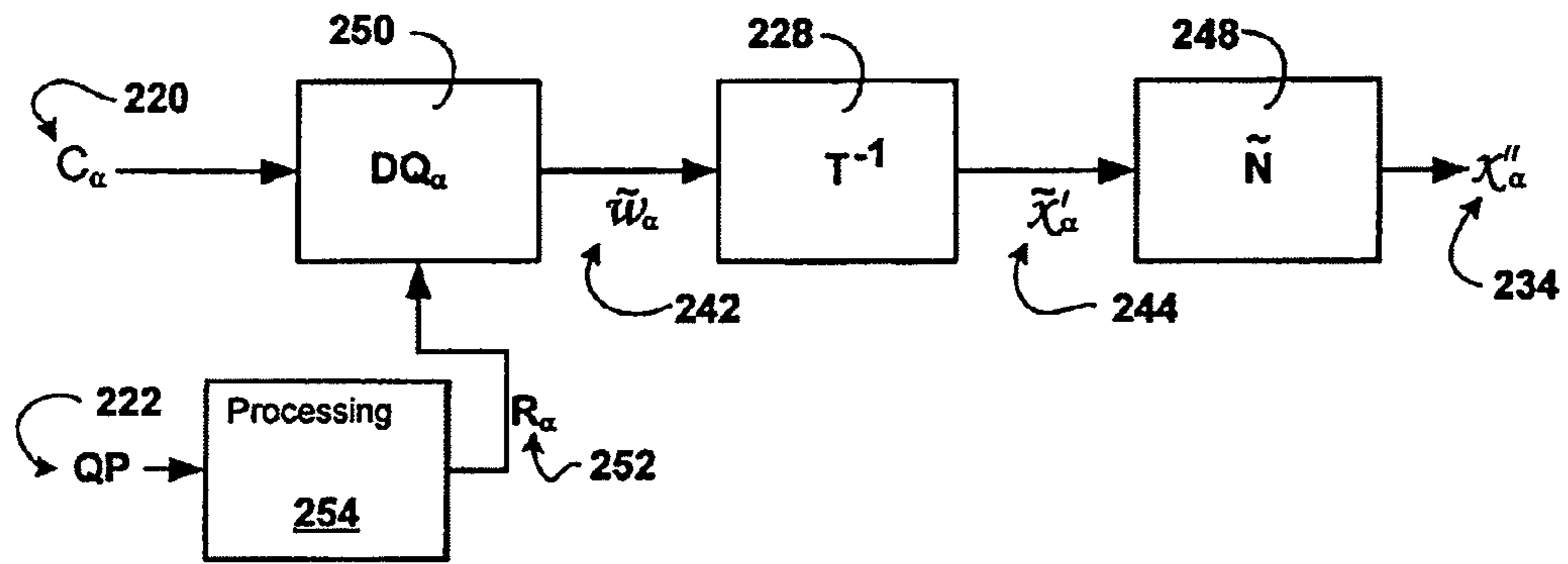
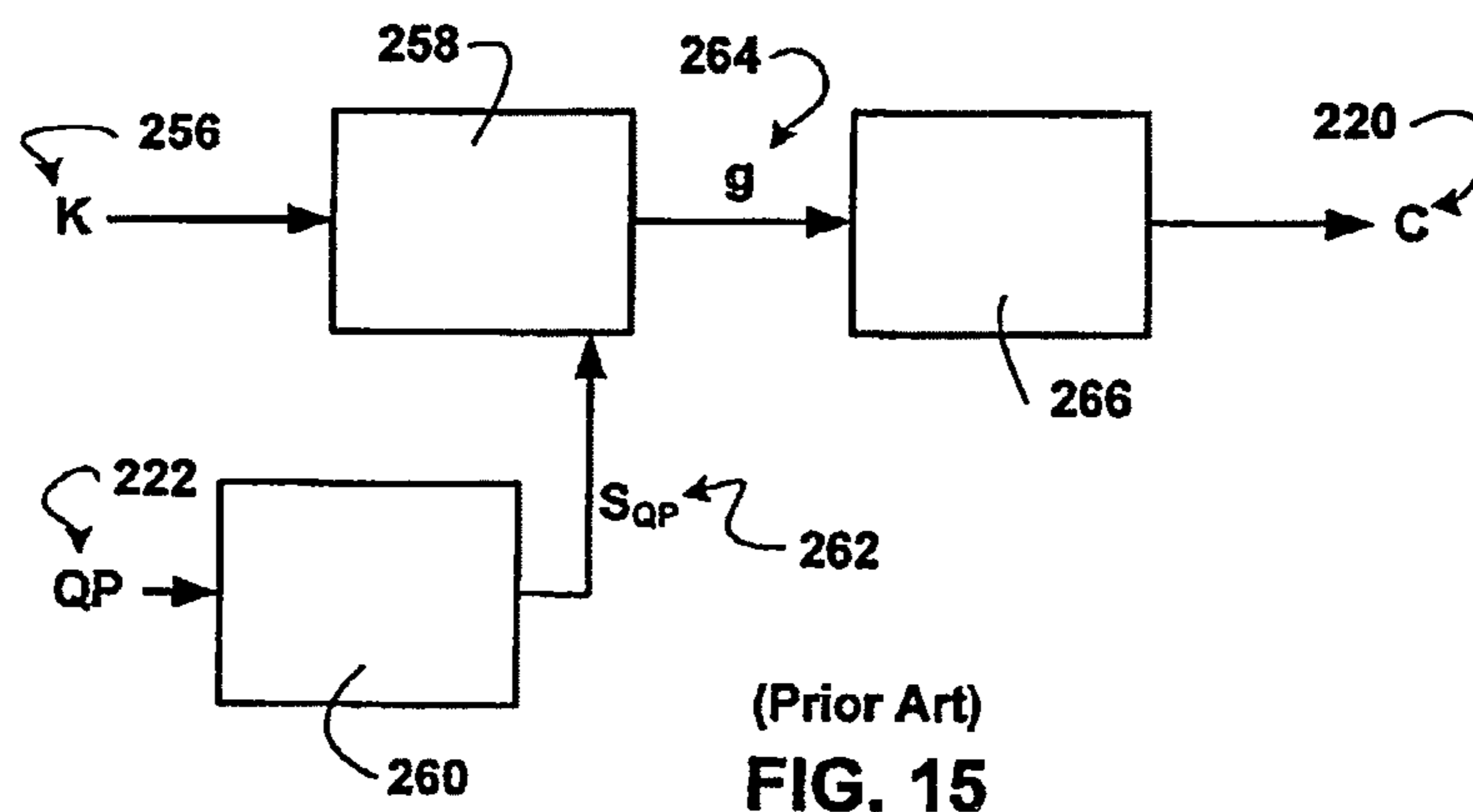


FIG. 14



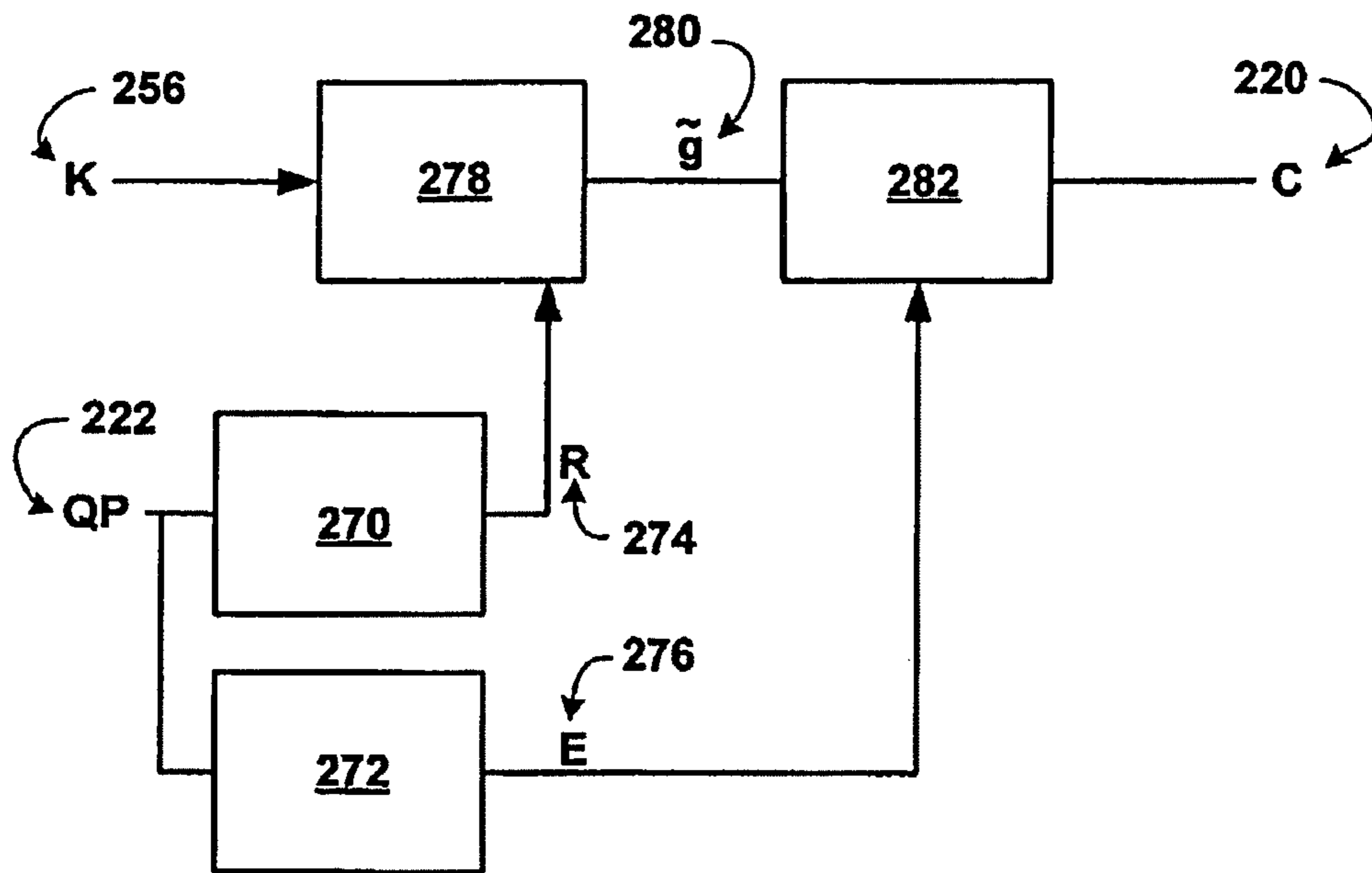


FIG. 16

1

**SYSTEMS AND METHODS FOR REDUCED
BIT-DEPTH PROCESSING IN
VIDEO-RELATED DATA WITH FREQUENCY
WEIGHTING MATRICES**

Matter enclosed in heavy brackets [] appears in the original patent but forms no part of this reissue specification; matter printed in italics indicates the additions made by reissue.

RELATED APPLICATIONS

[This application is a continuation of U.S. patent application Ser. No. 10/326,459 entitled, METHODS AND SYSTEMS FOR EFFICIENT VIDEO-RELATED DATA PROCESSING, invented by Louis Kerofsky, filed Dec. 20, 2002 now U.S. Pat. No. 7,170,942, which is a continuation of U.S. patent application Ser. No. 10/139,036 entitled, METHOD FOR REDUCED BIT-DEPTH QUANTIZATION, invented by Louis Kerofsky, filed May 2, 2002 now U.S. Pat. No. 7,123,655, which claims the benefit of U.S. Provisional Patent Application Ser. No. 60/319,018 entitled, METHODS AND SYSTEMS FOR VIDEO CODING WITH JOINT QUANTIZATION AND NORMALIZATION PROCEDURES, invented by Louis Kerofsky, filed Nov. 30, 2001, and also claims the benefit of U.S. Provisional Patent Application Ser. No. 60/311,436 entitled, REDUCED BIT-DEPTH QUANTIZATION, invented by Louis Kerofsky, filed Aug. 9, 2001.] *This Application is a Divisional Reissue Application of Continuation Reissue application Ser. No. 12/837,154, filed Jul. 15, 2010 (now U.S. Pat. No. Re. 43,091), of Reissue application Ser. No. 12/689,897, filed Jan. 19, 2010 (now U.S. Pat. No. Re. 42,745), which itself is a reissue of U.S. Pat. No. 7,400,682 entitled SYSTEMS AND METHODS FOR REDUCED BIT-DEPTH PROCESSING IN VIDEO-RELATED DATA WITH FREQUENCY WEIGHTING MATRICES, which is a continuation of U.S. patent application Ser. No. 10/326,459 entitled METHODS AND SYSTEMS FOR EFFICIENT VIDEO-RELATED DATA PROCESSING, invented by Louis Kerofsky, filed Dec. 20, 2002, now U.S. Pat. No. 7,170,942, which is a continuation of U.S. patent application Ser. No. 10/139,036 entitled METHOD FOR REDUCED BIT-DEPTH QUANTIZATION, invented by Louis Kerofsky, filed May 2, 2002, now U.S. Pat. No. 7,123,655, which claims the benefit of U.S. Provisional Patent Application Ser. No. 60/319,018 entitled METHODS AND SYSTEMS FOR VIDEO CODING WITH JOINT QUANTIZATION AND NORMALIZATION PROCEDURES, invented by Louis Kerofsky, filed Nov. 30, 2001, and also claims the benefit of U.S. Provisional Patent Application Ser. No. 60/311,436 entitled REDUCED BIT-DEPTH QUANTIZATION, invented by Louis Kerofsky, filed Aug. 9, 2001. More than one Divisional Reissue Application has been filed for the reissue of application Ser. No. 12/837,154. The reissue applications are application Ser. Nos. 13/301,430 (the present application), 13/301,472, 13/301,502, and 13/301,526, all of which are divisional reissue applications of application Ser. No. 12/837,154, and all of which were filed on Nov. 21, 2011.*

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention generally relates to video compression techniques and, more particularly, to a method for reducing the bit size required in the computation of video coding transformations.

2

2. Description of the Related Art

A video information format provides visual information suitable to activate a television screen, or store on a video tape. Generally, video data is organized in a hierarchical order. A video sequence is divided into group of frames, and each group can be composed of a series of single frames. Each frame is roughly equivalent to a still picture, with the still pictures being updated often enough to simulate a presentation of continuous motion. A frame is further divided into slices, or horizontal sections which helps system design of error resilience. Each slice is coded independently so that errors do not propagate across slices. A slice consists of macroblocks. In H.26P and Motion Picture Experts Group (MPEG)-X standards, a macroblock is made up of 16×16 luma pixels and a corresponding set of chroma pixels, depending on the video format. A macroblock always has an integer number of blocks, with the 8×8 pixel matrix being the smallest coding unit.

Video compression is a critical component for any application which requires transmission or storage of video data. Compression techniques compensate for motion by reusing stored information in different areas of the frame (temporal redundancy). Compression also occurs by transforming data in the spatial domain to the frequency domain. Hybrid digital video compression, exploiting temporal redundancy by motion compensation and spatial redundancy by transformation, such as Discrete Cosine Transform (DCT), has been adapted in H.26P and MPEG-X international standards as the basis.

As stated in U.S. Pat. No. 6,317,767 (Wang), DCT and inverse discrete cosine transform (IDCT) are widely used operations in the signal processing of image data. Both are used, for example, in the international standards for moving picture video compression put forth by the MPEG. DCT has certain properties that produce simplified and efficient coding models. When applied to a matrix of pixel data, the DCT is a method of decomposing a block of data into a weighted sum of spatial frequencies, or DCT coefficients. Conversely, the IDCT is used to transform a matrix of DCT coefficients back to pixel data.

Digital video (DV) codecs are one example of a device using a DCT-based data compression method. In the blocking stage, the image frame is divided into N by N blocks of pixel information including, for example, brightness and color data for each pixel. A common block size is eight pixels horizontally by eight pixels vertically. The pixel blocks are then “shuffled” so that several blocks from different portions of the image are grouped together. Shuffling enhances the uniformity of image quality.

Different fields are recorded at different time incidents. For each block of pixel data, a motion detector looks for the difference between two fields of a frame. The motion information is sent to the next processing stage. In the next stage, pixel information is transformed using a DCT. An 8-8 DCT, for example, takes eight inputs and returns eight outputs in both vertical and horizontal directions. The resulting DCT coefficients are then weighted by multiplying each block of DCT coefficients by weighting constants.

The weighted DCT coefficients are quantized in the next stage. Quantization rounds off each DCT coefficient within a certain range of values to be the same number. Quantizing tends to set the higher frequency components of the frequency matrix to zero, resulting in much less data to be stored. Since the human eye is most sensitive to lower frequencies, however, very little perceptible image quality is lost by this stage.

The quantization stage includes converting the two-dimensional matrix of quantized coefficients to a one-dimensional

linear stream of data by reading the matrix values in a zigzag pattern and dividing the one-dimensional linear stream of quantized coefficients into segments, where each segment consists of a string of zero coefficients followed by a non-zero quantized coefficient. Variable length coding (VLC) then is performed by transforming each segment, consisting of the number of zero coefficients and the amplitude of the non-zero coefficient in the segment, into a variable length codeword. Finally, a framing process packs every 30 blocks of variable length coded quantized coefficients into five fixed-length synchronization blocks.

Decoding is essentially the reverse of the encoding process described above. The digital stream is first deframed. Variable length decoding (VLD) then unpacks the data so that it may be restored to the individual coefficients. After inverse quantizing the coefficients, inverse weighting and an inverse discrete cosine transform (IDCT) are applied to the result. The inverse weights are the multiplicative inverses of the weights that were applied in the encoding process. The output of the inverse weighting function is then processed by the IDCT.

Much work has been done studying means of reducing the complexity in the calculation of DCT and IDCT. Algorithms that compute two-dimensional IDCTs are called "type I" algorithms. Type I algorithms are easy to implement on a parallel machine, that is, a computer formed of a plurality of processors operating simultaneously in parallel. For example, when using N parallel processors to perform a matrix multiplication on N×N matrices, N column multiplies can be simultaneously performed. Additionally, a parallel machine can be designed so as to contain special hardware or software instructions for performing fast matrix transposition.

One disadvantage of type I algorithms is that more multiplications are needed. The computation sequence of type I algorithms involves two matrix multiplies separated by a matrix transposition which, if N=4, for example, requires 64 additions and 48 multiplications for a total number of 112 instructions. It is well known by those skilled in the art that multiplications are very time-consuming for processors to perform and that system performance is often optimized by reducing the number of multiplications performed.

A two-dimensional IDCT can also be obtained by converting the transpose of the input matrix into a one-dimensional vector using an L function. Next, the tensor product of constant a matrix is obtained. The tensor product is then multiplied by the one-dimensional vector L. The result is converted back into an N×N matrix using the M function. Assuming again that N=4, the total number of instructions used by this computational sequence is 92 instructions (68 additions and 24 multiplications). Algorithms that perform two-dimensional IDCTs using this computational sequence are called "type II" algorithms. In type II algorithms, the two constant matrices are grouped together and performed as one operation. The advantage of type II algorithms is that they typically require fewer instructions (92 versus 112) and, in particular, fewer costly multiplications (24 versus 48). Type II algorithms, however, are very difficult to implement efficiently on a parallel machine. Type II algorithms tend to reorder the data very frequently and reordering data on a parallel machine is very time-intensive.

There exist numerous type I and type II algorithms for implementing IDCTs, however, dequantization has been treated as an independent step depending upon DCT and IDCT calculations. Efforts to provide bit exact DCT and IDCT definitions have led to the development of efficient integer transforms. These integer transforms typically increase the dynamic range of the calculations. As a result, the

implementation of these algorithms requires processing and storing data that consists of more than 16 bits.

It would be advantageous if intermediate stage quantized coefficients could be limited to a maximum size in transform processes.

It would be advantageous if a quantization process could be developed that was useful for 16-bit processors.

It would be advantageous if a decoder implementation, dequantization, and inverse transformation could be implemented efficiently with a 16-bit processor. Likewise, it would be advantageous if the multiplication could be performed with no more than 16 bits, and if memory access required no more than 16 bits.

SUMMARY OF THE INVENTION

The present invention is an improved process for video compression. Typical video coding algorithms predict one frame from previously coded frames. The error is subjected to a transform and the resulting values are quantized. The quantizer controls the degree of compression. The quantizer controls the amount of information used to represent the video and the quality of the reconstruction.

The problem is the interaction of the transform and quantization in video coding. In the past the transform and quantizer have been designed independently. The transform, typically the discrete cosine transform, is normalized. The result of the transform is quantized in standard ways using scalar or vector quantization. In prior work, MPEG-1, MPEG-2, MPEG-4, H.261, H.263, the definition of the inverse transform has not been bit exact. This allows the implementer some freedom to select a transform algorithm suitable for their platform. A drawback of this approach is the potential for encoder/decoder mismatch damaging the prediction loop.

To solve this mismatch problem portions of the image are periodically coded without prediction. Current work, for example H.26L, has focused on using integer transforms that allow bit exact definition. Integer transforms may not be normalized. The transform is designed so that a final shift can be used to normalize the results of the calculation rather than intermediate divisions. Quantization also requires division. H.26L provides an example of how these integer transforms are used along with quantization.

In H.26L Test Model Long-term 8, normalization is combined with quantization and implemented via integer multiplications and shifts following forward transform and quantization and following dequantization and inverse transform. H.26L TML uses two arrays of integers A(QP) and B(QP) indexed by quantization parameter (QP), see Table 1. These values are constrained by the relation shown below in Equation 1.

TABLE 1

TML quantization parameters		
QP	A _{TML} (QP)	B _{TML} (QP)
0	620	3881
1	553	4351
2	492	4890
3	439	5481
4	391	6154
5	348	6914
6	310	7761
7	276	8718
8	246	9781
9	219	10987
10	195	12339

5

TABLE 1-continued

TML quantization parameters		
QP	$A_{TML}(QP)$	$B_{TML}(QP)$
11	174	13828
12	155	15523
13	138	17435
14	123	19561
15	110	21873
16	98	24552
17	87	27656
18	78	30847
19	69	34870
20	62	38807
21	55	43747
22	49	49103
23	44	54683
24	39	61694
25	35	68745
26	31	77615
27	27	89113
28	24	100253
29	22	109366
30	19	126635
31	17	141533

Equation 1 Joint Normalization/Quantization Relation

$$A(QP) \cdot B(QP) \cdot 676^2 = 2^{40}.$$

Normalization and quantization are performed simultaneously using these integers and divisions by powers of 2. Transform coding in H.26L uses a 4x4 block size and an integer transform matrix T, Equation 2. For a 4x4 block X, the transform coefficients K are calculated as in Equation 3. From the transform coefficients, the quantization levels, L, are calculated by integer multiplication. At the decoder the levels are used to calculate a new set of coefficients, K'. Additional integer matrix transforms followed by a shift are used to calculate the reconstructed values X'. The encoder is allowed freedom in calculation and rounding of the forward transform. Both encoder and decoder must compute exactly the same answer for the inverse calculations.

Equation 2 H.26L Test Model 8 Transform Matrix

$$T = \begin{pmatrix} 13 & 13 & 13 & 13 \\ 17 & 7 & -7 & -17 \\ 13 & -13 & -13 & 13 \\ 7 & -17 & 17 & -7 \end{pmatrix}$$

Equation 3 TML DCT_LUMA and iDCT_LUMA

$$Y = T \cdot X$$

$$K = Y \cdot T^T$$

$$L = (A_{TML}(QP) \cdot K) / 2^{20}$$

$$K' = B_{TML}(QP) \cdot L$$

$$Y' = T^T \cdot K'$$

$$X' = (Y' \cdot T) / 2^{20}$$

6

Where the intermediate result Y is the result of a one dimensional transform and the intermediate result Y' is the result of a one dimensional inverse transform.

The dynamic range required during these calculations can be determined. The primary application involves 9-bit input, 8 bits plus sign, the dynamic range required by intermediate registers and memory accesses is presented in Table 2.

TABLE 2

Dynamic range of TML transform and inverse transform (bits)		
9-bit input	LUMA Transform	Inverse Transform
Register	30	27
Memory	21	26

To maintain bit-exact definitions and incorporate quantization, the dynamic range of intermediate results can be large since division operations are postponed. The present invention combines quantization and normalization, to eliminate the growth of dynamic range of intermediate results. With the present invention the advantages of bit exact inverse transform and quantization definitions are kept, while controlling the bit depth required for these calculations. Reducing the required bit depth reduces the complexity required of a hardware implementation and enables efficient use of single instruction multiple data (SIMD) operations, such as the Intel MMX instruction set.

Accordingly, a method is provided for the quantization of a coefficient. The method comprises: receiving a coefficient K; receiving a quantization parameter (QP); forming a quantization value (L) from the coefficient K using a mantissa portion ($A_m(QP)$) and an exponential portion ($x^{A_e(QP)}$). Typically, the value of x is 2.

In some aspects of the method, forming a quantization value (L) from the coefficient K includes:

$$L = K * A(QP) \\ = K * A_m(QP) * (2^{A_e(QP)}).$$

In other aspects, the method further comprises: normalizing the quantization value by 2^N as follows:

$$L_n = L / 2^N \\ = K * A_m(QP) / 2^{(N - A_e(QP))}.$$

In some aspects, forming a quantization value includes forming a set of recursive quantization factors with a period P, where $A(QP+P) = A(QP)/x$. Therefore, forming a set of recursive quantization factors includes forming recursive mantissa factors, where $A_m(QP) = A_m(QP \text{ mod } P)$. Likewise, forming a set of recursive quantization factors includes forming recursive exponential factors, where $A_e(QP) = A_e(QP \text{ mod } P) - QP/P$.

More specifically, receiving a coefficient K includes receiving a coefficient matrix $K[i][j]$. Then, forming a quantization value (L) from the coefficient matrix $K[i][j]$ includes forming a quantization value matrix ($L[i][j]$) using a mantissa portion matrix ($A_m(QP)[i][j]$) and an exponential portion matrix ($x^{A_e(QP)[i][j]}$).

Likewise, forming a quantization value matrix ($L[i][j]$) using a mantissa portion matrix ($A_m(QP)[i][j]$) and an expo-

ponential portion matrix ($x^{Ae(QP)[i][j]}$) includes, for each particular value of QP, every element in the exponential portion matrix being the same value. Every element in the exponential portion matrix is the same value for a period (P) of QP values, where $Ae(QP)=Ae(P*(QP/P))$.

Additional details of the above-described method, including a method for forming a dequantization value (X1), from the quantization value, using a mantissa portion (Bm(QP)) and an exponential portion ($x^{Be(QP)}$), are provided below.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a flowchart illustrating the present invention method for the quantization of a coefficient.

FIG. 2 is a diagram showing embodiments of the present invention comprising systems and methods for video encoding wherein a quantization parameter may be established based on user inputs;

FIG. 3 is a diagram showing embodiments of the present invention comprising systems and methods for video decoding;

FIG. 4 is diagram showing embodiments of the present invention comprising storing encoder output on a computer-readable storage media;

FIG. 5 is a diagram showing embodiments of the present invention comprising sending encoder output over a network;

FIG. 6 is a diagram showing embodiments of quantization methods and apparatuses of the present invention comprising a first mantissa portion processing means and a first exponential portion processing means;

FIG. 7 is a diagram showing embodiments of quantization methods and apparatuses of the present invention comprising a first mantissa portion processing means and a first shifting portion processing means;

FIG. 8 is a diagram showing embodiments of dequantization methods and apparatuses of the present invention comprising a second mantissa portion processing means and a second exponential portion processing means;

FIG. 9 is a diagram showing embodiments of dequantization methods and apparatuses of the present invention comprising a second mantissa portion processing means and a second shifting portion processing means;

FIG. 10 is diagram showing prior art methods comprising dequantization, inverse transformation, and normalization (Prior Art);

FIG. 11 is a diagram showing embodiments of the present invention comprising factorization of an equivalent of a dequantization scaling factor;

FIG. 12 is a diagram showing embodiments of the present invention comprising factorization thereby achieving a reduction in bit depth for inverse transformation calculations and reduce memory requirements for dequantization parameter storage;

FIG. 13 is a diagram showing embodiments of the present invention comprising a normalization process independent of quantization parameter (QP);

FIG. 14 is diagram showing embodiments of the present invention comprising frequency dependent quantization;

FIG. 15 is a diagram showing prior art methods comprising quantization (Prior Art); and

FIG. 16 is a diagram showing embodiments of the present invention comprising factorization of an equivalent of a quantization scaling factor.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The dynamic range requirements of the combined transform and quantization is reduced by factoring the quantiza-

tion parameters A(QP) and B(QP) into a mantissa and exponent terms as shown in Equation 4. With this structure, only the precision due to the mantissa term needs to be preserved during calculation. The exponent term can be included in the final normalization shift. This is illustrated in the sample calculation Equation 5.

Equation 4 Structure of Quantization Parameters

$$A_{proposed}(QP)=A_{mantissa}(QP)\cdot 2^{A_{exponent}(QP)}$$

$$B_{proposed}(QP)=B_{mantissa}(QP)\cdot 2^{B_{exponent}(QP)}$$

Equation 5 Reduced Bit_Depth LUMA Transform

$$Y=T\cdot X$$

$$K=Y\cdot T^T$$

$$L=(A_{mantissa}(QP)\cdot K)/2^{20-A_{exponent}(QP)}$$

$$K'=T^T\cdot L$$

$$Y'=K'\cdot T$$

$$X'=(B_{mantissa}(QP)\cdot Y')/2^{20-B_{exponent}(QP)}$$

To illustrate the present invention, a set of quantization parameters is presented that reduce the dynamic range requirement of an H.26L decoder to 16-bit memory access. The memory access of the inverse transform is reduced to 16 bits. Values for $A_{mantissa}$, $A_{exponent}$, $B_{mantissa}$, $B_{exponent}$, $A_{proposed}$, $B_{proposed}$ are defined for QP=0-5 as shown in Table 3. Additional values are determined by recursion, as shown in Equation 6. The structure of these values makes it possible to generate new quantization values in addition to those specified.

TABLE 3

Quantization values 0-5 for TML						
QP	$A_{mantissa}$	$A_{exponent}$	$B_{mantissa}$	$B_{exponent}$	$A_{proposed}$	$B_{proposed}$
0	5	7	235	4	640	3760
1	9	6	261	4	576	4176
2	127	2	37	7	508	4736
3	114	2	165	5	456	5280
4	25	4	47	7	400	6016
5	87	2	27	8	348	6912

Equation 6 Recursion Relations

$$A_{mantissa}(QP+6)=A_{mantissa}(QP)$$

$$B_{mantissa}(QP+6)=B_{mantissa}(QP)$$

$$A_{exponent}(QP+6)=A_{exponent}(QP)-1$$

$$B_{exponent}(QP+6)=B_{exponent}(QP)+1$$

Using the defined parameters, the transform calculations can be modified to reduce the dynamic range as shown in Equation 5. Note how only the mantissa values contribute to the growth of dynamic range. The exponent factors are incorporated into the final normalization and do not impact the dynamic range of intermediate results.

With these values and computational method, the dynamic range at the decoder is reduced so only 16-bit memory access is needed as seen in Table 4.

TABLE 4

Dynamic range with low-bit depth quantization (OP > 6)		
8-bit	LUMA Transform	LUMA Inverse Transform
Register	28	24
Memory	21	16

Several refinements can be applied to the joint quantization/normalization procedure described above. The general technique of factoring the parameters into a mantissa and exponent forms the basis of these refinements.

The discussion above assumes all basis functions of the transform have an equal norm and are quantized identically. Some integer transforms have the property that different basis functions have different norms. The present invention technique has been generalized to support transforms having different norms by replacing the scalars $A(QP)$ and $B(QP)$ above by matrices $A(QP)[i][j]$ and $B(QP)[i][j]$. These parameters are linked by a normalization relation of the form shown below, Equation 7, which is more general than the single relation shown in Equation 1.

Equation 7 Joint Quantization/Normalization of Matrices

$$A(QP)[i][j] \cdot B(QP)[i][j] = N[i][j]$$

Following the method previously described, each element of each matrix is factored into a mantissa and an exponent term as illustrated in the equations below, Equation 8.

Equation 8 Factorization of Matrix Parameters

$$A(QP)[i][j] = A_{mantissa}(QP)[i][j] \cdot 2^{A_{exponent}(QP)[i][j]}$$

$$B(QP)[i][j] = B_{mantissa}(QP)[i][j] \cdot 2^{B_{exponent}(QP)[i][j]}$$

A large number of parameters are required to describe these quantization and dequantization parameters. Several structural relations can be used to reduce the number of free parameters. The quantizer growth is designed so that the values of A are halved after each period P at the same time the values of B are doubled maintaining the normalization relation. Additionally, the values of $A_{exponent}(QP)[i][j]$ and $B_{exponent}(QP)[i][j]$ are independent of i, j and (QP) in the range $[0, P-1]$. This structure is summarized by structural equations, Equation 9. With this structure there are only two parameters $A_{exponent}[0]$ and $B_{exponent}[0]$.

Equation 9 Structure of Exponent Terms

$$A_{exponent}(QP)[i][j] = A_{exponent}[0] - QP/P$$

$$B_{exponent}(QP)[i][j] = B_{exponent}[0] - QP/P$$

A structure is also defined for the mantissa values. For each index pair (i, j) , the mantissa values are periodic with period P . This is summarized by the structural equation, Equation 10. With this structure, there are P independent matrices for $A_{mantissa}$ and P independent matrices for $B_{mantissa}$ reducing memory requirements and adding structure to the calculations.

Equation 10 Structure of Mantissa Terms

$$A_{mantissa}(QP)[i][j] = A_{mantissa}(QP \% P)[i][j]$$

$$B_{mantissa}(QP)[i][j] = B_{mantissa}(QP \% P)[i][j]$$

The inverse transform may include integer division that requires rounding. In cases of interest, the division is by a power of 2. The rounding error is reduced by designing the dequantization factors to be multiples of the same power of 2, giving no remainder following division.

Dequantization using the mantissa values $B_{mantissa}(QP)$ gives dequantized values that are normalized differently depending upon QP . This must be compensated for following the inverse transform. A form of this calculation is shown in Equation 11.

Equation 11 Normalization of Inverse Transform I

$$K[i][j] = B_{mantissa}(QP \% P)[i][j] \cdot Level[i][j]$$

$$X = (T^{-1} \cdot K \cdot T) / 2^{(N - QP/P)}$$

To eliminate the need for the inverse transform to compensate for this normalization difference, the dequantization operation is defined so that all dequantized values have the same normalization. The form of this calculation is shown in Equation 12.

Equation 12 Normalization of Inverse Transform II

$$K[i][j] = B_{mantissa}(QP \% P)[i][j] \cdot Level[i][j]$$

$$X = (T^{-1} \cdot K \cdot T) / 2^N$$

An example follows that illustrates the present invention use of quantization matrices. The forward and inverse transforms defined in Equation 13 need a quantization matrix rather than a single scalar quantization value. Sample quantization and dequantization parameters are given. Equation 14 and 16, together with related calculations, illustrate the use of this invention. This example uses a period $P=6$.

Equation 13 transforms

$$T_{forward} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{pmatrix}$$

$$T_{reverse} = \begin{pmatrix} 2 & 2 & 2 & 1 \\ 2 & 1 & -2 & -2 \\ 2 & -2 & -2 & 2 \\ 2 & -1 & 2 & -1 \end{pmatrix}$$

Equation 14 quantization parameters

$$Q(m)[i][j] = M_{m,0} \text{ for } (i, j) = \{(0, 0), (0, 2), (2, 0), (2, 2)\}$$

$$Q(m)[i][j] = M_{m,1} \text{ for } (i, j) = \{(1, 1), (1, 3), (3, 1), (3, 3)\}$$

$$Q(m)[i][j] = M_{m,2} \text{ otherwise}$$

11

$$M = \begin{bmatrix} 21844 & 8388 & 13108 \\ 18724 & 7625 & 11650 \\ 16384 & 6989 & 10486 \\ 14564 & 5992 & 9532 \\ 13107 & 5243 & 8066 \\ 11916 & 4660 & 7490 \end{bmatrix}$$

Equation 16 Dequantization parameters

$$R(m)[i][j]=S_{m,0} \text{ for } (i,j)=\{(0,0), (0,2), (2,0), (2,2)\}$$

$$R(m)[i][j]=S_{m,1} \text{ for } (i,j)=\{(1,1), (1,3), (3,1), (3,3)\}$$

$$R(m)[i][j]=S_{m,2} \text{ otherwise}$$

$$S = \begin{bmatrix} 6 & 10 & 8 \\ 7 & 11 & 9 \\ 8 & 12 & 10 \\ 9 & 14 & 11 \\ 10 & 16 & 13 \\ 11 & 18 & 14 \end{bmatrix}$$

The description of the forward transformation and forward quantization, Equation 18, are given below assuming input is in X, quantization parameter QP.

Equation 17 Forward Transform

$$K=T_{forward} X \cdot T_{forward}^T$$

Equation 18 Forward Quantization

$$\text{period}=QP/6$$

$$\text{phase}=QP-6\text{-period}$$

$$\text{Level}[i][j]=(Q(\text{phase})[i][j] \cdot K[i][j])/2^{17+\text{period}}$$

The description of dequantization, inverse transform, and normalization for this example is given below, Equation 19 and 20.

Equation 19 Dequantization

$$\text{period}=QP/6$$

$$\text{phase}=QP-6\text{-period}$$

$$K[i][j]=R(\text{phase})[i][j] \cdot \text{Level}[i][j] \cdot 2^{\text{period}}$$

Equation 20 IDCT and Normalization

$$X'=T_{reverse} K \cdot T_{reverse}^T$$

$$X''[i][j]=X'[i][j]/2^7$$

FIG. 1 is a flowchart illustrating the present invention method for the quantization of a coefficient. Although this method is depicted as a sequence of numbered steps for clarity, no order should be inferred from the numbering unless explicitly stated. It should be understood that some of these steps may be skipped, performed in parallel, or performed without the requirement of maintaining a strict order of

12

sequence. The methods start at Step 100. Step 102 supplies a coefficient K. Step 104 supplies a quantization parameter (QP). Step 106 forms a quantization value (L) from the coefficient K using a mantissa portion (Am(QP)) and an exponential portion ($x^{Ae(QP)}$). Typically, the exponential portion ($x^{Ae(QP)}$) includes x being the value 2.

In some aspects of the method, forming a quantization value (L) from the coefficient K using a mantissa portion (Am(QP)) and an exponential portion ($x^{Ae(QP)}$) in Step 106 includes:

$$L = K * A(QP)$$

$$= K * \text{Am}(QP) * (2^{Ae(QP)}).$$

Some aspects of the method include a further step. Step 108 normalizes the quantization value by 2^N as follows:

$$L_n = L / 2^N$$

$$= K * \text{Am}(QP) / 2^{(N-Ae(QP))}.$$

In other aspects, forming a quantization value in Step 106 includes forming a set of recursive quantization factors with a period P, where $A(QP+P)=A(QP)/x$. Likewise, forming a set of recursive quantization factors includes forming recursive mantissa factors, where $\text{Am}(QP)=\text{Am}(QP \bmod P)$. Then, forming a set of recursive quantization factors includes forming recursive exponential factors, where $\text{Ae}(QP)=\text{Ae}(QP \bmod P)-QP/P$.

In some aspects, forming a quantization value includes forming a set of recursive quantization factors with a period P, where $A(QP+P)=A(QP)/2$. In other aspects, forming a set of recursive quantization factors includes forming recursive mantissa factors, where $P=6$. Likewise, forming a set of recursive quantization factors includes forming recursive exponential factors, where $P=6$.

In some aspects of the method, receiving a coefficient K in Step 102 includes receiving a coefficient matrix $K[i][j]$. Then, forming a quantization value (L) from the coefficient matrix $K[i][j]$ using a mantissa portion (Am(QP)) and an exponential portion ($x^{Ae(QP)}$) in Step 106 includes forming a quantization value matrix ($L[i][j]$) using a mantissa portion matrix ($\text{Am}(QP)[i][j]$) and an exponential portion matrix ($x^{Ae(QP)[i][j]}$). Likewise, forming a quantization value matrix ($L[i][j]$) using a mantissa portion matrix ($\text{Am}(QP)[i][j]$) and an exponential portion matrix ($x^{Ae(QP)[i][j]}$) includes, for each particular value of QP, every element in the exponential portion matrix being the same value. Typically, every element in the exponential portion matrix is the same value for a period (P) of QP values, where $\text{Ae}(QP)=\text{Ae}(P*(QP/P))$.

Some aspects of the method include a further step. Step 110 forms a dequantization value (X1) from the quantization value, using a mantissa portion (Bm(QP)) and an exponential portion ($x^{Be(QP)}$). Again, the exponential portion ($x^{Be(QP)}$) typically includes x being the value 2.

In some aspects of the method, forming a dequantization value (X1) from the quantization value, using a mantissa portion (Bm(QP)) and an exponential portion ($2^{Be(QP)}$) includes:

13

$$\begin{aligned} X1 &= L * B(QP) \\ &= L * Bm(QP) * (2^{Be(QP)}). \end{aligned}$$

Other aspects of the method include a further step, Step 112, of denormalizing the quantization value by 2^N as follows:

$$\begin{aligned} X1d &= X1 / 2^N \\ &= X1 * Bm(QP) / 2^N. \end{aligned}$$

In some aspects, forming a dequantization value in Step 110 includes forming a set of recursive dequantization factors with a period P, where $B(QP+P)=x*B(QP)$. Then, forming a set of recursive dequantization factors includes forming recursive mantissa factors, where $Bm(QP)=Bm(QP \bmod P)$. Further, forming a set of recursive dequantization factors includes forming recursive exponential factors, where $Be(QP)=Be(QP \bmod P)+QP/P$.

In some aspects, forming a set of recursive quantization factors with a period P includes the value of x being equal to 2, and forming recursive mantissa factors includes the value of P being equal to 6. Then, forming a set of recursive dequantization factors includes forming recursive exponential factors, where $Be(QP)=Be(QP \bmod P)+QP/P$.

In some aspects of the method, forming a dequantization value (X1), from the quantization value, using a mantissa portion ($Bm(QP)$) and an exponential portion ($x^{Be(QP)}$) in Step 110 includes forming a dequantization value matrix ($X1[i][j]$) using a mantissa portion matrix ($Bm(QP)[i][j]$) and an exponential portion matrix ($x^{Be(QP)[i][j]}$). Likewise, forming a dequantization value matrix ($X1[i][j]$) using a mantissa portion matrix ($Bm(QP)[i][j]$) and an exponential portion matrix ($x^{Be(QP)[i][j]}$) includes, for each particular value of QP, every element in the exponential portion matrix being the same value. In some aspects, every element in the exponential portion matrix is the same value for a period (P) of QP values, where $Be(QP)=Be(P*(QP/P))$.

Another aspect of the invention includes a method for the dequantization of a coefficient. However, the process is essentially the same as Steps 110 and 112 above, and is not repeated in the interest of brevity.

A method for the quantization of a coefficient has been presented. An example is given illustrating a combined dequantization and normalization procedure applied to the H.26L video coding standard with a goal of reducing the bit-depth required at the decoder to 16 bits. The present invention concepts can also be used to meet other design goals within H.26L. In general, this invention has application to the combination of normalization and quantization calculations.

Embodiments of the present invention may be implemented as hardware, firmware, software and other implementations. Some embodiments may be implemented on general purpose computing devices or on computing devices specifically designed for implementation of these embodiments. Some embodiments may be stored in memory as a means of storing the embodiment or for the purpose of executing the embodiment on a computing device.

Some embodiments of the present invention comprise systems and methods for video encoding, as shown in FIG. 2. In these embodiments, image data 130 is subtracted from 132 with data representing prior video frames 145 resulting in a differential image 133, which is sent to a transform module

14

134. Transform module 134 may use DCT or other transform methods to transform the image. Generally, the result of the transform process will be coefficients (K), which are then sent to a quantization module 136 for quantization.

5 Quantization module 136 may have other inputs, such as user inputs 131 for establishing quantization parameters (QPs) and for other input. Quantization module 136 may use the transformation coefficients and the quantization parameters to determine quantization levels (L) in the video image. 10 Quantization module 136 may use methods employing a mantissa portion and an exponential portion, however, other quantization methods may also be employed in the quantization modules 136 of embodiments of the present invention. These quantization levels 135 and quantization parameters 15 133 are output to a coding module 138 as well as a dequantization module (DQ) 140.

Output to the coding module 138 is encoded and transmitted outside the encoder for immediate decoding or storage. Coding module 138 may use variable length coding (VLC) in its coding processes. Coding module 138 may use arithmetic coding in its coding process.

Output from quantization module 136 is also received at dequantization module 140 to begin reconstruction of the image. This is done to keep an accurate accounting of prior frames. Dequantization module 140 performs a process with essentially the reverse effect as quantization module 136. Quantization levels or values (L) are dequantized yielding transform coefficients. Dequantization modules 140 may use methods employing a mantissa portion and an exponential portion as described herein.

The transform coefficients output from dequantization module 140 are sent to an inverse transformation (IT) module 142 where they are inverse transformed to a differential image 141. This differential image 141 is then combined with data from prior image frames 145 to form a video frame 149 that may be input to a frame memory 146 for reference to succeeding frames.

Video frame 149 may also serve as input to a motion estimation module 147, which also receives input image data 130. These inputs may be used to predict image similarities and help compress image data. Output from motion estimation module 147 is sent to motion compensation module 148 and combined with output data from coding module 138, which is sent out for later decoding and eventual image viewing.

Motion compensation module 148 uses the predicted image data to reduce frame data requirements; its output is subtracted from input image data 130.

Some embodiments of the present invention comprise systems and methods for video decoding, as shown in FIG. 3. A decoder of embodiments of the present invention may receive encoded image data 150 to a decoder module 152. Encoded image data 150 may comprise data that has been encoded by an encoder 100 such as that described with reference to FIG. 2.

Decoder module 152 may employ variable length decoding methods if they were used in the encoding process. Other decoding methods may also be used as dictated by the type of encoded data 150. Decoding module 152 performs essentially the reverse process as coding module 138. Output from decoding module 152 may comprise quantization parameters 156 and quantization values 154. Other output may comprise motion estimation data and image prediction data that may be sent directly to a motion compensation module 166.

Typically, quantization parameters 156 and quantization values 154 are output to a dequantization module 158, where quantization values are converted back to transform coeffi-

icients. These coefficients are then sent to an inverse transformation module **160** for conversion back to spatial domain image data **161**.

The motion compensation unit **166** uses motion vector data and the frame memory **164** to construct a reference image **165**.

Image data **161** represents a differential image that must be combined **162** with prior image data **165** to form a video frame **163**. This video frame **163** is output **168** for further processing, display or other purposes and may be stored in frame memory **164** and used for reference with subsequent frames.

In some embodiments of the present invention, as illustrated in FIG. **4**, image data **102** may be sent to an encoder or encoding portion **104** for the various transformation, quantization, encoding and other procedures typical of video encoding as described above for some embodiments of the present invention. Output from the encoder may then be stored on any computer-readable storage media **106**. Storage media **106** may act as a short-term buffer or as a long-term storage device.

When desired, encoded video data may be read from storage media **106** and decoded by a decoder or decoding portion **108** for output **110** to a display or other device.

In some embodiments of the present invention, as illustrated in FIG. **5**, image data **112** may be sent to an encoder or encoding portion **114** for the various transformation, quantization, encoding and other procedures typical of video encoding as described above for some embodiments of the present invention. Output from the encoder may then be sent over a network, such as a LAN, WAN or the Internet **116**. A storage device such as storage media **106** may be part of a network. Encoded video data may be received and decoded by a decoder or decoding portion **118**, which also communicates with network **116**. Decoder **118** may then decode the data for local consumption **120**.

In some embodiments of the present invention, as illustrated in FIG. **6**, a quantization method or apparatus comprises a mantissa portion **172** and an exponential portion **174**. Quantization parameters **176** are input to both portions **172** & **174**. A coefficient **K 170** is input to the mantissa portion **172** where it is modified using the quantization parameter and other values as explained above. The result of this operation is combined with the result produced in the exponential portion using the quantization parameter thereby producing a quantization level or value **L 178**.

In some embodiments of the present invention, as illustrated in FIG. **7**, a quantization method or apparatus comprises a mantissa portion **182** and a shifting portion **184**. Quantization parameters **186** are input to both portions **182** & **184**. A coefficient, **K 180** is input to the mantissa portion **182** where it is modified using the quantization parameter and other values as explained above. The result of this operation is further processed in the shifting portion using the quantization parameter thereby producing a quantization level or value, **L 188**.

Some embodiments of the present invention, as illustrated in FIG. **8**, comprise a dequantization method or apparatus with a mantissa portion **192** and an exponential portion **194**. Quantization parameters **196** are input to both portions **192** & **194**. A quantization value, **L 190** is input to the mantissa portion **192** where it is modified using the quantization parameter and other values as explained above. The result of this operation is further processed in the exponential portion using the quantization parameter thereby producing a coefficient, **X1 198**.

Some embodiments of the present invention, as illustrated in FIG. **9**, comprise a dequantization method or apparatus with a mantissa portion **202** and a shifting portion **204**. Quantization parameters **206** are input to both portions **202** & **204**. A quantization value, **L 200** is input to the mantissa portion **202** where it is modified using the quantization parameter and other values as explained above. The result of this operation is further processed in the exponential portion using the quantization parameter thereby producing a coefficient, **X1 208**.

Some embodiments of the present invention may be stored on computer-readable media such as magnetic media, optical media, and other media as well as combinations of media. Some embodiments may also be transmitted as signals across networks and communication media. These transmissions and storage actions may take place as part of operation of embodiments of the present invention or as a way of transmitting the embodiment to a destination.

Typical methods of dequantization, inverse transformation, and normalization may be expressed mathematically in equation form. These methods, as illustrated in FIG. **10**, may begin with input in the form of an array of quantized coefficient levels c_α **220**, and a quantization parameter **QP 222**. A dequantization scaling value S^{QP} **224** is then calculated **221** using the quantization parameter **QP 222**. Quantized coefficient levels **220** are scaled **227** by S^{QP} **224** to give transform coefficients w_α **226** according to Equation 21. These transform coefficients **226** are then inverse transformed **228** to compute scaled samples x'_α **230** as shown in Equation 22. The scaled samples **230** may then be normalized **232** to give reconstructed samples, x''_α **234** according to Equation 23.

$$w_\alpha = c_\alpha \cdot S^{QP} \quad \text{Equation 21}$$

$$x'_\alpha = \sum_{\beta} T_{\alpha\beta}^{-1} \cdot w_\beta \quad \text{Equation 22}$$

$$x''_\alpha = (x'_\alpha + f) \gg M \quad \text{Equation 23}$$

In embodiments of the present invention, a reduction in bit depth for inverse transformation calculations is achieved. The processes of these embodiments, illustrated in FIG. **11**, begin with input in the form of an array of quantized coefficient levels c_α **220**, and a quantization parameter **QP 222** similar to typical prior art methods. However, in these embodiments, the equivalent of a dequantization scaling factor S^{QP} is factored **223** & **225** into a mantissa portion R^{QP} **236** and an exponential portion E^{QP} **238**. The mantissa portion **236** is used during dequantization **240** to calculate the reconstructed transform coefficients (\tilde{w}_α) **242**, which are used in the inverse transformation process **228** to calculate reconstructed samples (\tilde{x}'_α) **244**. These reconstructed samples may then be normalized using the exponential portion **238** according to Equation 26, thereby yielding reconstructed samples (x''_α) **234**. Using these methods, the values of \tilde{w}_α and \tilde{x}'_α require E^{QP} fewer bits for representation than the corresponding values w_α and x'_α . This factorization enables mathematically equivalent calculation of the reconstructed samples using lower intermediate precision as shown in Equations 24-26.

$$\tilde{w}_\alpha = C_\alpha \cdot R^{QP} \quad \text{Equation 24}$$

$$\tilde{x}'_\alpha = \sum_{\beta} T_{\alpha\beta}^{-1} \cdot \tilde{w}_\beta \quad \text{Equation 25}$$

by quantizing a two-dimensional transform coefficient matrix K , utilizing a processor and memory, the method comprising steps of:

inputting an element $K[i][j]$ of the transform coefficient matrix; 5

inputting a quantization parameter QP ; and
determining the quantization value matrix L ,
wherein the element $L[i][j]$ of the quantization value matrix is determined by using a mantissa portion matrix element $A(QP)[i][j]$ being a function of QP according to 10
the following equation:

$$L[i][j] = K[i][j] \times A(QP \bmod P)[i][j] \times 2^{A0 - QP/P},$$

where $A0$ and P are constants.

15. A computer-implemented dequantization method utilized in video decoding for determining a two-dimensional transform coefficient matrix K by dequantizing a quantization value matrix L , utilizing a processor and memory, the method comprising steps of: 15

inputting an element $L[i][j]$ of the quantization value matrix; 20

inputting a quantization parameter QP ; and
determining the transform coefficient matrix K ,
wherein the element $K[i][j]$ of the transform coefficient matrix is determined by using a mantissa portion matrix element $B(QP)[i][j]$ being a function of QP according to 25
the following equation:

$$K[i][j] = L[i][j] \times B(QP \bmod P)[i][j] \times 2^{B0 + QP/P},$$

where $B0$ and P are constants. 30

* * * * *