MITEL SEMICONDUCTOR

AN206

An Overview of the H.261

Application Note

AN206 - 1.1 October 1995

THE REQUIREMENT

H.261 is an ITU recommendation concerned with providing an international standard for video codecs which will allow inter regional compatibility for video telephony. The recommendation covers the bitstream contents and the accuracy of the decoder, but leaves room for differentiation in the encoder and pre/post processing. It also allows users with either PAL, NTSC, or SECAM local cameras and monitors to communicate freely without knowing the TV standards of the remote participant. It was also concerned in specifying a compression method which could economically be mechanized at real time camera rates. This implies that the processing requirements for coding and decoding should not be vastly different, as is the case with the MPEG standard.

The Integrated Services Digital Network (ISDN) provides the means of communication, and thus compressed video bandwidths are based on increments of 64 Kbits/second. The basic rate ISDN interface provides two 64 Kbits/second information (B) channels and the primary rate interface (PRI) in Europe provides scope for 30 B channels. H.261 thus describes video coding and decoding methods at rates of p x 64 Kbits/second, where p is in the range of 1 to 30. For this reason the specification is often known as the px64 compression method.

Once a call has been established the transmission must continue at a fixed bitrate regardless of the level of activity within the picture. Since movement within a picture generally results in more bits being generated for that picture, then quality usually has to suffer in some way in order to keep the bitrate constant. This is in spite of system level buffering to smooth out the peaks and troughs in the bitrate. The use of broadband ISDN technology (B-ISDN), with its use of asynchronous transfer mode (ATM), would remove this restriction. The basic compression standard would not be affected, but the variable bitrate available with ATM allows the picture quality to remain constant when movement occurs. There may still be an upper bound to the bitrate, but only above this limit will quality start to suffer.

H.261 only covers the video side of a video phone system. Audio can use any of the adopted standards but typically would be either G.722 wideband coding at 48/56/64 Kbits/second or G.728 narrowband coding at 16 Kbits/second. The combined video and audio bitrate must comply with the number of information (B) channels provided at a given time. Thus in a basic rate system providing 2 B channels, and which uses G.728 audio coding, only 112 Kbits/second would be available for the video path.

Several other standards exist to cover a complete narrowband video telephony system. The complete top level specification is covered by H 320; H.221 covers the framing structure for the multiplexing of H.261 video, audio, data and signalling; H.242 covers the communication procedure (call set up, capability exchange, etc.); and H.230 covers control and identification signals.

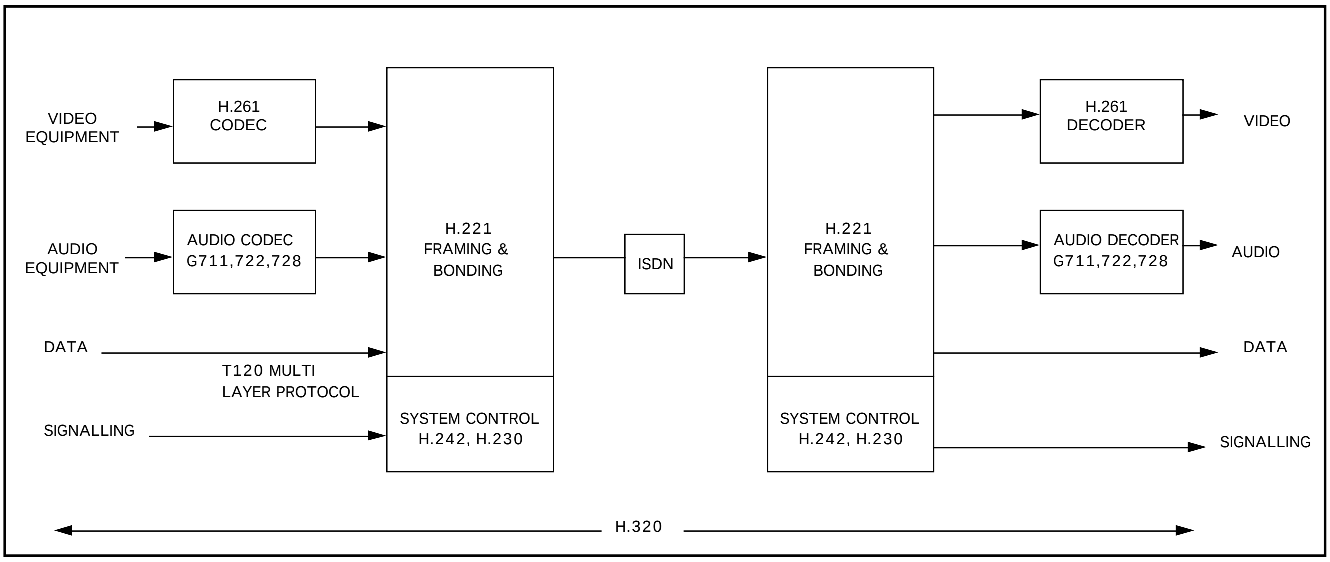

Fig.1 Video Phone Architecture Specified by H.320

THE PROBLEM

Studio quality PAL requires 720 luminance pixels per line and 576 active lines at 25 frames per second. Chrominance pixels can have half the horizontal resolution of luminance pixels without any perceived loss in quality; this results in 4:2:2 video with each of the 720 x 576 pixel positions requiring a 16 bit word to represent an 8 bit luminance component plus one of the 8 bit chrominance components. All this equates to the need for bitrates of around 166 M bits/second for studio quality video.

Videophones do not, of course, have to transmit studio quality video, and in fact a subjective quality of that achieved by domestic video recorders would be perfectly adequate. For this reason a Common Intermediate Format (CIF) was defined which in many ways was a political compromise between Europe and USA/Japan. It had the line rate of PAL/SECAM and the frame rate of NTSC, and is defined to have a luminance resolution of 352 x 288 lines with an update rate of 30 frames per second. Chrominance resolution is half that of the luminance in both the horizontal and vertical directions i.e. it has a resolution of 176 x 144 lines. These resolutions result in a bandwidth requirement of 31 Mbits/second to transmit uncompressed CIF frames at 30 Hz rates.

The problem then is to reduce this 31 Mb/s rate down to one within the range defined by H.261. Thus at the top end of the range a compression ratio of around 16:1 is needed, but for a 64 Kb/s line a compression ratio of 480:1 is required.

When the received picture is only to be displayed in a small window on a PC monitor, it is possible to get acceptable quality with lower spatial resolution. A quarter CIF (QCIF) standard was thus defined which is half the luminance and chrominance resolution in both the vertical and horizontal directions. This requires one quarter of the compression ratios given above, but is normally only used in low bitrate applications.

Since video phone systems still use conventional PAL or NTSC cameras, and off the shelf digital composite video decoders, it is necessary to use digital filters when decimating down to CIF or QCIF resolutions. Simply discarding alternate pixels produces unacceptable aliasing effects which soon annoy the user. There is a problem in producing a CIF frame from an NTSC source which is usually glossed over by people offering single chip or software solutions to H.261 compression. PAL is relatively easy since one of the fields out of an interlaced pair provides the correct number of luminance lines. Simple low pass filtering is then needed to reduce the bandwidth to match the required horizontal resolution. With NTSC, however, a single field only produces 240 active lines and polyphase interpolating vertical filters are needed to get acceptable quality. These require six sets of coefficients to produce six CIF luminance lines for every 5 NTSC lines.

THE COMPRESSION SOLUTION AND ITS EMBODIMENT IN THE MITEL CHIPSET

One fairly obvious way of reducing the amount of data needed to transmit a sequence of related frames is to code only the differences between frames. This is known as inter frame coding as opposed to intra frame coding which codes each frame in isolation. It is particularly useful for a video phone comprising a head and shoulders view of a talking person. Only the lips move plus a slight movement of the head between frames. Even so these movements plus camera noise and lighting effects can still cause significant frame to frame differences.

The first step is thus to minimize the secondary effects (caused by camera noise etc.) and then to recognize that movement has occurred. It is then possible to code vectors defining the movement of blocks of pixels rather than re-coding the complete frame every time. Both these steps are considered enhancements to the H.261 specification, although the decoder must be capable of using the coded motion vectors when they are transmitted.

The secondary frame to frame differences can be minimized by a combination of spatial low pass filtering and temporal filtering. The latter produces an average of the present and previously captured camera frame which is dependent on the difference between individual pixels in each of those frames. Thus if the difference is small the previous pixel is used and if it is large the new pixel is used. In between large and small some fraction of the actual difference is added to the previous pixel to produce a new pixel which is closer to the previous pixel than it would have been. The fraction applied varies in a non linear way, and above some difference threshold the new pixel is always used i.e. the whole difference is added to the previous pixel.

In the present Mitel chipset the spatial noise filtering is done with the VP520. Decimation down to CIF or QCIF is done at the same time. It uses 8 tap horizontal filters and 5 tap vertical filters for producing CIF, and 16 tap horizontal and 7 tap vertical filters for producing QCIF. It also provides the six phase filters previously identified as being needed to produce 288 CIF lines from a 240 line NTSC field. Temporal filtering is presently not implemented but will be featured in the next generation solution.

Movement can only be estimated by considering manageable blocks of pixels. The assumption then being that all pixels in that block move in the same direction. The pixels must thus be strongly correlated. For simplicity only luminance pixels are used, and the difference between pixels in the present and the previous block is calculated by summing the absolute differences between each pair of pixels. The block of pixels is then moved around in a search area the dimensions of which represent the maximum value of the motion vectors to be produced. The difference summation is done at each pixel position in the search window, and the position with the minimum sum of differences defines the best fit.

The size of the luminance block used to detect movement is defined to be 16 x 16 pixels in the H.261 specification. The maximum search displacement is ±15 from the centre position, giving a total search window size of 46 x 46 pixels. Even though the position of the best fit has been found there will still be errors between individual pixels in the 16 x 16 block. These errors are coded but, as will be shown later, still represent a bitrate saving over simply re-coding every new frame as it occurs. If the error in the best fit position is outside a threshold then motion compensation is aborted, and the new block is coded without reference to the previous frame.

The H.261 specification does not demand that the encoder uses motion compensation, only that the decoder can use any transmitted vectors plus the errors. An encoder without such facilities would, of course, produce worse quality video for a given bitrate. The VP2611 Encoder incorporates a motion estimator but displacement vectors are limited to ±7 pixels i.e. the search window is limited to 23 x 23 pixels. This was considered adequate in its intended traditional video conferencing environment, where the camera is a long way from the speakers and little movement occurs at 30 Hz frame rates. In a PC based video phone, however, the camera is only a couple of feet from the user and lower frame rates are used. Much more movement is then possible between frames and the full ±15 search area would be advantageous in reducing bitrates. This is being addressed in the next generation design.

The calculations required to find the best fit, using an exhaustive search of all the positions that a 16 x 16 block can occupy over a ±15 range, amount to approximately 256 thousand 8 bit differences and 256 thousand 16 bit accumulations. In a CIF frame there are 396 such blocks, and at a 30 Hz frame rate the calculation must be done in about 75 microseconds. This is equivalent to a computation rate of approximately 6.5 giga operations per second (GOPS). Several algorithms have been put forward in an attempt to reduce the computation rate, but with all it is possible to never find the optimum fit. It should be noted that since the errors between previous and present blocks are always coded then not finding the best fit does not actually cause any corruption of the picture. It simply means that a higher bitrate is produced when coding that part of the picture , which does however mean that lower quality is obtained for a fixed average bitrate.

A two step algorithm has been developed for the new VP600 Encoder which has been shown to give good performance with a standard set of test sequences. In the first pass a fit is calculated at every third pixel position in both the horizontal and vertical directions. In the second pass a search is done using the twenty four positions round the first pass best fit. This reduces the computation rate to less than one GOP, and is implemented with 20 eight bit subtractors and 20 sixteen bit accumulators. With a 54 MHz clock rate these allow 30 Hz frame rates to be achieved.

TRANSFORM CODING AND QUANTIZATION

To achieve the high compression rates needed in a video phone system it is necessary to use transform coding plus lossy compression techniques i.e. the reconstructed image will contain errors when compared to the original image. This in turn requires that the encoder keeps a copy of the image which has been encoded and then decoded when inter frame coding is to be implemented. This re-constructed image is then used to calculate frame to frame differences.

Transform coding is a way of compressing the energy present in a two dimensional array of pixels such that most of the information contained in the original number of spatial pixels is reduced to a smaller number of transform coefficients. An array of closely correlated pixels is converted to one of de-correlated coefficients, and redundancy is removed. Once these coefficients have been calculated the bits used to represent each value can be reduced by a division process (quantization) and most of the smaller values should then go to zeros. At this point the compression method becomes lossy since it is impossible to re-construct the original image.

In the H.261 specification the Direct Cosine Transform (DCT) was chosen to do the energy compression. Although other transforms are more optimal the DCT was chosen since it is relatively easy to implement at video scan rates and does not use complex numbers. To the non mathematical user it has the advantage that the coefficients can be likened to frequency components, since the DCT can be compared to the real part of an FFT calculation. Thus conceptually the transform and quantization process is simply converting from the spatial to the frequency domain and then filtering out the higher frequencies. The eye is known to be less sensitive to the higher frequencies in a video frame, and thus they can be removed without the effect being too visible. A visualisation of the mathematical transform process is thus possible.

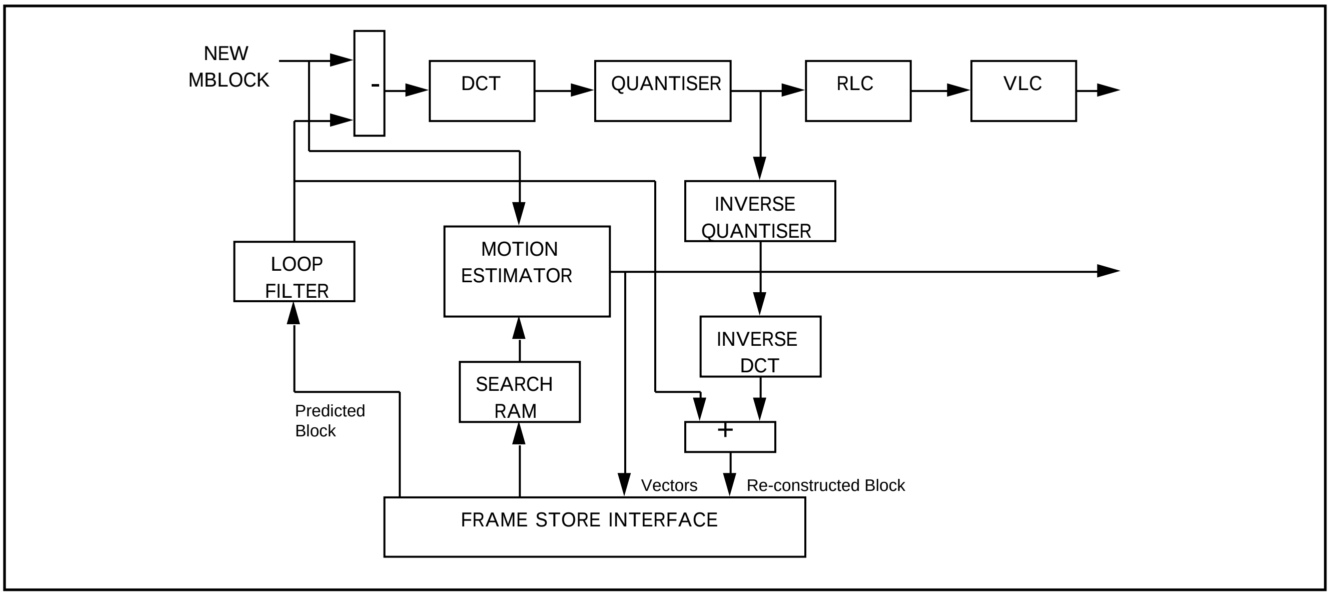

Fig.2 H.261 Coding Kernel

The pixel array size was chosen to be 8 x 8, and was a compromise between the computation power needed to calculate the DCT and selecting the largest possible group of correlated pixels. Earlier proprietary standards had chosen a 16 x 16 block as being optimal, but beyond this pixels are not as well correlated in a typical picture. The actual accuracy of the DCT calculation is not specified, but the error tolerance in the output of the Inverse DCT is well defined. The DCT is performed either on absolute pixels or signed pixel differences if inter frame coding is being performed. The latter are represented by 9 bit two's complement numbers in the VP2611 Encoder, and a sign bit is added to the 8 bit unsigned absolute pixel value. A 12 bit two's complement number is obtained form the DCT circuit and this is then quantized.

The quantization value is defined by the H.261 specification to be an even number in the range 2 to 62, it can thus be expressed as a 5 bit value in the coded bitstream. The 12 bit signed value is thus divided by a variable unsigned number, and then clipped to an 8 bit signed number. For quantization levels below 8 it is possible that clipping will occur with high coefficient values from the DCT.

The actual quantization value is determined by the system operating conditions. Since the number of bits generated for a frame depends on the amount of movement since the last frame, it is necessary to have a buffer at the output of the system. This attempts to smooth out the peaks and troughs and will be loaded at a variable rate but read at a fixed rate determined by the chosen constant line rate. When this buffer starts to fill the system reacts by increasing the quantization level in an attempt to produce more zero coefficients out of the quantizer. As the buffer gets even fuller the system is allowed to completely skip coded blocks and ultimately complete frames.

Since the clipped output from the quantizer is expected to contain a large amount of zero's, run length coding can be used to reduce the number of bits needed to represent the coefficients. This is made more effective if the coefficients are first re-arranged in ascending frequency order since we expect that the higher frequencies values will have disappeared. This is known as zig zag scan order.

For every original array of 8 x 8 pixels we have now obtained 64 quantized coefficients expressed as runs of zero's and actual values. We normally only expect three or four actual values to remain, and several combinations of runs and coefficients have a higher probability of occurring. There is thus something to be gained by using variable length coding, and the codes to be used are defined in the H.261 specification. Motion vectors are also variable length coded.

H.261 STRUCTURE

We have established that H.261 video coding is based on quantizing the output of a DCT which is done on an array of 8 x 8 pixels. Inter frame coding is supported in which the DCT is done on pixel differences, and the performance can be further improved by optionally applying motion compensation. The output is a series of run length coded coefficients which can be variable length coded in commonly occurring sequences. This data must then be combined with any motion vectors, and also with information amount the coded blocks such as quantization values and whether they were inter or intra coded. A serial bitstream must then be produced.

It is beyond the scope of this paper to fully define the H.261 bitstream but essentially the 8 x 8 basic DCT blocks (sub blocks) are firstly combined into macroblocks. These consist of four luminance sub blocks plus two chrominance sub blocks (one for each component). Remember that the definition of CIF called for chrominance to be half the resolution of luminance in both directions. A chrominance sub block thus corresponds to the same spatial screen area as four luminance sub blocks. These four luminance sub blocks are considered as one 16 x 16 entity when motion estimation is calculated.

Each macroblock is given a relative address since macroblocks can be skipped if they contain no coded inter mode data (fixed backgrounds) or they can be forcibly skipped to keep the generation of bits under control (the decoder then uses the same macroblock in the last received frame and hopes it is relevant). A macroblock header defines its quantization value if this has changed from the last macroblock, plus whether it was inter or intra coded and whether it was motion compensated etc.

Macroblocks are then combined into 11 x 3 groups (GOBS) and finally a picture header is added. For a CIF picture there are 12 GOBS and for a QCIF picture there are 3 GOBS. All headers are unique sequences of bits and cannot be produced by the data coding scheme.

The serial bitstream contains all this information plus the actual coded data. This is divided into frames of 512 bits which should not be confused with original video frames. The 512 bits consist of 492 bits of the above data; plus 18 parity bits for error correction; plus one bit of framing code which allows synchronization of the decoder; plus a fill bit which indicates whether the data is valid or whether the transmission buffer was empty and the system is just keeping the line busy. The parity bits allow one or two bit errors to be corrected within that particular frame, and the framing code is 8 bits long in total. The decoder initially searches for this framing code by examining 8 bits which are 512 apart in the received bitstream. Once the code has been detected it must be repeated three times to ensure that it was not found by chance in the data. Having obtained frame lock the receiver then knows the position of the fill bit and the error correction bits. It can then do any error correction and finally search for unique picture start codes. Once a picture start code has been found the decoder can progress through the various layers and finally decode the run length coded coefficients using the inverse DCT transform.

IMPLEMENTATION WITH THE MITEL CHIPSET

The encode path, from CCIR601 video data out of a composite video decoder through to an error corrected H.261 bitstream, takes three devices in the present Mitel one micron solution. The VP520 contains all the vertical and horizontal filters necessary to produce CIF or QCIF macroblock data from both PAL and NTSC line video.

The VP2611 contains the complete coding kernel comprising of the ±7 motion estimator; the loop filter defined in the specification; the inter/intra decision processor; the DCT and quantizer; and finally zig zag re-ordering and run length coding. It also contains a complete reverse path so that quantized coefficients can be inverse quantized and passed through an inverse DCT in the same manner as one at the far end. This re-constructed data is then written to an external DRAM frame store and used in inter frame coding.

The VP2612 takes the run length coded coefficients and motion vectors from the VP2611 and does variable length coding. These are then written to an external transmit buffer, and read out at the required line rate. The H.261 header information, framing structure, and error correction bits are then added to produce an H.261 compliant bitstream.

The quantization is not done by dedicated hardware, but all the information for an external software algorithm is provided on a host controller bus. This allows each user to add differentiation to their product, since experienced providers of video conferencing system have their own proprietary algorithms.

Three devices provide the reverse operations in the decoder. The VP2614 De-Multiplexer needs the support of an external 32K x 8 static RAM. It searches for H.261 frame lock and then does error correction. Picture start codes are then identified and variable length decoding takes place. A state machine interprets the H.261 coding structure and produces run length coded coefficients for the VP2615 Decoder. The VP2615 then completes the decoding operation and performs the inverse DCT operation. It fetches the previous block from an external DRAM if inter frame coding was specified. The output of the VP2615 consists of macroblocks of pixel data which are supplied as inputs to the VP520. This device has both encode and decode modes of operation; in the decode mode it interpolates the CIF or QCIF data up to normal CCIR601 resolutions. Two sets of vertical filter coefficients allow interlaced fields to be produced.

This chipset is now in production and the next generation system is now being designed. This comprises two devices; the VP600 contains all the present functionality of the VP520, VP2611, and VP2612 and only requires a single external DRAM rather than two DRAM's and an SRAM. Likewise the VP610 contains all the functionality of the VP2614, VP2615, and VP520 in decode mode. It also only requires a single external DRAM rather than three RAM's as at present.

The new devices are specifically aimed at single board PC videophone designs, rather than it raditional video conferencing systems. They thus contain system level enhancements aimed at meeting the needs of multimedia PC's and offer increased performance with a ±15 search window and temporal filtering.

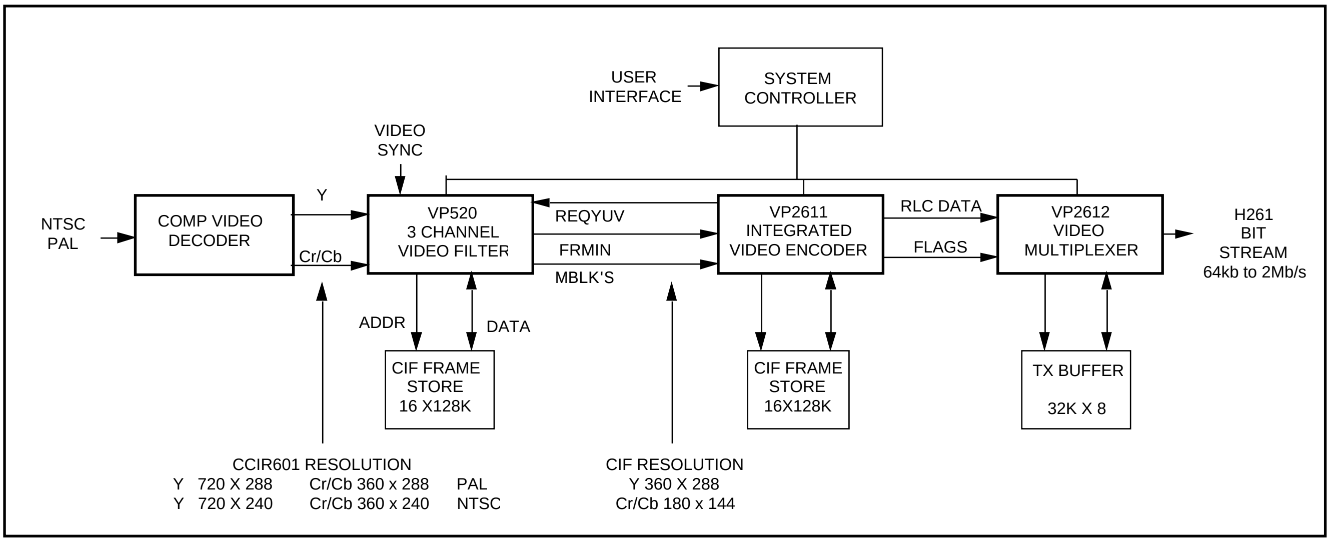

Fig.3 Complete H.261 Encoder Using the Mitel Chipset

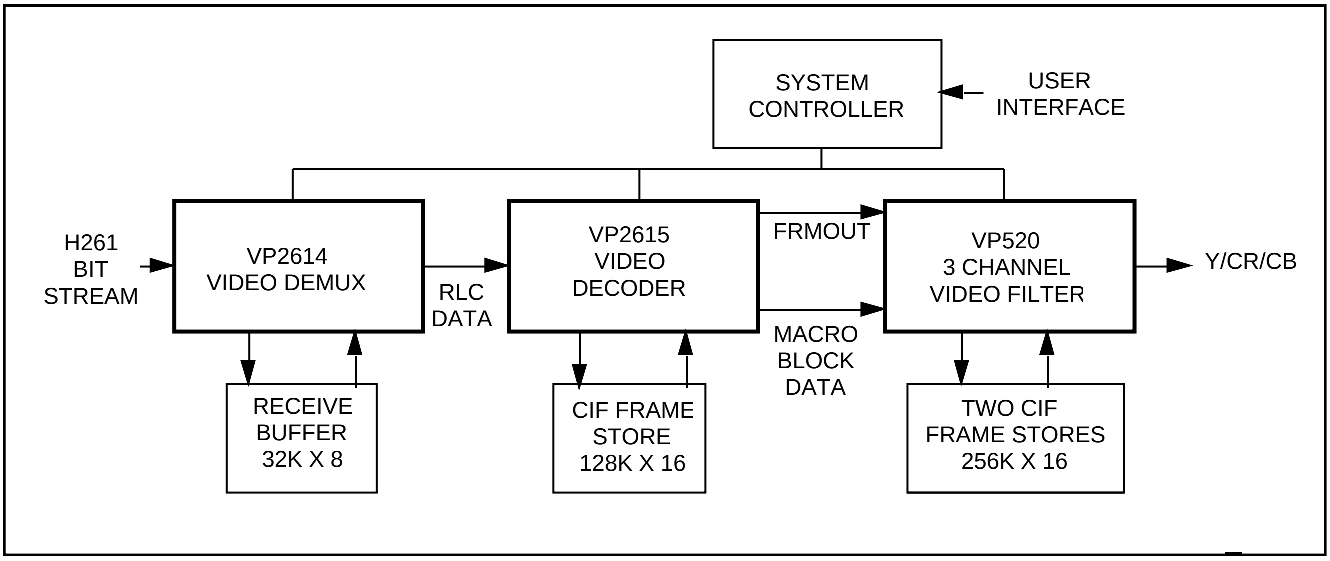

Fig.4 Complete H.261 Decoder Using the Mitel Chipset

MITEL
SEMICONDUCTOR

HEADQUARTERS OPERATIONS
MITEL SEMICONDUCTOR
Cheney Manor, Swindon,
Wiltshire SN2 2QW, United Kingdom.
Tel: (01793) 518000
Fax: (01793) 518411

MITEL SEMICONDUCTOR
1500 Green Hills Road,
Scotts Valley, California 95066-4922
United States of America.
Tel (408) 438 2900
Fax: (408) 438 5576/6231

Internet: http://www.gpsemi.com
CUSTOMER SERVICE CENTRES
- **FRANCE & BENELUX** Les Ulis Cedex Tel: (1) 69 18 90 00 Fax : (1) 64 46 06 07
- **GERMANY** Munich Tel: (089) 419508-20 Fax : (089) 419508-55
- **ITALY** Milan Tel: (02) 6607151 Fax: (02) 66040993
- **JAPAN** Tokyo Tel: (03) 5276-5501 Fax: (03) 5276-5510
- **KOREA** Seoul Tel: (2) 5668141 Fax: (2) 5697933
- **NORTH AMERICA** Scotts Valley, USA Tel: (408) 438 2900 Fax: (408) 438 5576/6231
- **SOUTH EAST ASIA** Singapore Tel:(65) 3827708 Fax: (65) 3828872
- **SWEDEN** Stockholm Tel: 46 8 702 97 70 Fax: 46 8 640 47 36
- **TAIWAN, ROC** Taipei Tel: 886 2 25461260 Fax: 886 2 27190260
- **UK, EIRE, DENMARK, FINLAND & NORWAY**
  Swindon Tel: (01793) 726666 Fax : (01793) 518582

These are supported by Agents and Distributors in major countries world-wide.

This publication is issued to provide information only which (unless agreed by the Company in writing) may not be used, applied or reproduced for any purpose nor form part of any order or contract nor to be regarded as a representation relating to the products or services concerned. No warranty or guarantee express or implied is made regarding the capability, performance or suitability of any product or service. The Company reserves the right to alter without prior notice the specification, design or price of any product or service. Information concerning possible methods of use is provided as a guide only and does not constitute any guarantee that such methods of use will be satisfactory in a specific piece of equipment. It is the user's responsibility to fully determine the performance and suitability of any equipment using such information and to ensure that any publication or data used is up to date and has not been superseded. These products are not suitable for use in any medical products whose failure to perform may result in significant injury or death to the user. All products and materials are sold and services provided subject to the Company's conditions of sale, which are available on request.