(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 February 2008 (14.02.2008)

PCT

(10) International Publication Number
WO 2008/019156 A2

(51) International Patent Classification:
G06T 13/00 (2006.01)

(21) International Application Number:
PCT/US2007/017718

(22) International Filing Date: 8 August 2007 (08.08.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/836,467 8 August 2006 (08.08.2006) US
60/843,266 7 September 2006 (07.09.2006) US

(71) Applicant *(for all designated States except US)*: **DIGITAL MEDIA CARTRIDGE, LTD.** [—/US]; Water Garden, 1601 Cloverfield Blvd., 2nd Floor, South Tower, Santa Monica, CA 90404 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **HSIUNG, Ping-Kang**; Taipiei (TW). **KUO, Chung Chieh** [US/US]; Arcadia, California (US). **YANG, Sheng** [CN/CN]; Shen-Zhen (CN).

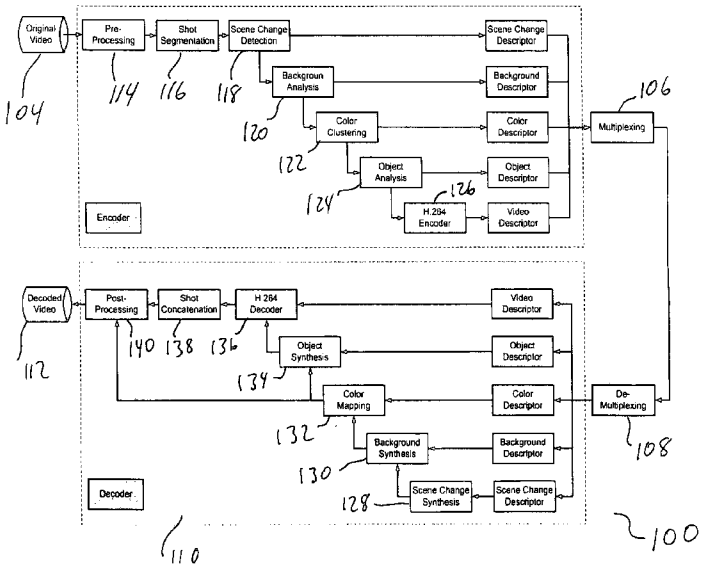
(74) Agent: **MONROE, Wesley W.**; Christie, Parker & Hald, LLP, P.O. Box 7066, Pasadena, CA 91109-7068 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

(54) **Title:** SYSTEM AND METHOD FOR CARTOON COMPRESSION



(57) **Abstract:** A system, specialized for encoding video of animated or cartoon content, encodes a video sequence. The system includes a background analyzer that removes moving objects from a series of video frames and generates a background definition for a static background used in a plurality of sequential video frames, a color clusterer that analyzes the colors contained in a video stream and creates a major color list of colors occurring in the video stream, an object identifier that identifies one or more objects that are constant within a series of video frames except for their position and rotational orientation within the series of video frames, and a hybrid encoder that encodes backgrounds and objects derived from a video sequence according to one of a plurality of encoding techniques depending on the compression achieved by each of the plurality of encoding techniques.

WO 2008/019156 A2

SYSTEM AND METHOD FOR CARTOON COMPRESSION

CROSS REFERENCE TO RELATED APPLICATION(S)

This application is based upon and claims priority to U.S. Provisional Application No. 60/836,467, filed August 8, 2006 and U.S. Provisional Application No. 60/843,266, filed September 7, 2006 the entire contents of which are hereby expressly incorporated by reference.

BACKGROUND OF THE INVENTION

Various video compression techniques are known in the art, such as MPEG-3, MPEG-4, H.264. Generally, these video compression techniques are good at compressing "live action" content, such as content shot with a conventional film or video camera. There is a need for a compression technique that takes into account unique features of animated, and particularly cartoon based video.

SUMMARY OF THE INVENTION

Animation, and particularly cartoon animation, has many characteristics that set it apart from "natural" or "live action" film or video. The present invention takes advantage of some of the characteristics and provides more flexible compression techniques to improve the coding gain and/or reduce the computational complexity in decoding. Some of the features of cartoons are:

- The camera movement is very simple, usually camera zooming and panning. In most cases, the camera remains still for one scene.
- There are fewer number of colors or shades of colors.
- The textural pattern is very simple. For example, one solid area is usually rendered with one single color.
- The boundaries of objects are very clear, so that the objects can be easily separated from the background.

A system according to the invention, specialized for encoding video of animated or cartoon content, encodes a video sequence. The system includes a background analyzer that removes moving objects from a series of video frames and generates a background definition for a static background used in a plurality of sequential video frames, a color clusterer that analyzes the colors contained in a video stream and creates a major color list of colors occurring in the video stream, an object identifier that identifies one or more objects that are constant within a series of video frames except for their position and rotational orientation within the series of video frames, and a hybrid encoder that encodes backgrounds and objects

derived from a video sequence according to one of a plurality of encoding techniques depending on the compression achieved by each of the plurality of encoding techniques.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of the system architecture of an exemplary embodiment of the invention.

Fig. 2A is an original cartoon frame before Intra-processing filtering.

Fig. 2B is the frame shown in Fig. 2A after filtering by the Intra-processing filter according to an embodiment of the invention.

Fig. 2C is the negative difference between the frames shown in Figs. 2A and 2B.

Figs. 3A and 3B show two consecutive frames in an example cartoon.

Fig. 3C shows the difference between the frames shown in Figs. 3A and 3B.

Fig. 3D shows the frame shown in Fig. 3C after sharpening.

Fig. 3E shows a filtered image of the frame shown in Fig. 3C after sharpening.

Fig. 4 is a histogram of the difference frame shown in Fig. 3C.

Fig. 5 is a video frame that exhibits a 3:2 pulldown artifact.

Fig. 6 is a block diagram of an embodiment of a modified encoder.

Fig. 7 is a graph showing the empirical results of measuring f_3 for all possible inter-frame luminance differences.

DETAILED DESCRIPTION OF THE INVENTION

A block diagram of the system architecture of an exemplary embodiment of the invention is shown in Fig. 1. The system 100 of Fig. 1 includes an encoder 102 that receives video 104 and produces an output to multiplexor 106. The output of multiplexor 106 is input into demulitplexor 108 which sends its output to decoder 110. Decoder 110 then outputs decoded video 112. In many embodiments, the encoder 102 and decoder 110 are implemented using a programmed general purpose computer. In other embodiments, the encoder 102 and decoder 110 are each implemented in one or more special function hardware units. In yet other embodiments, encoder 102 and decoder 110 each include a programmed general purpose computer that performs some of the functions of the encoder or decoder and one or more special function hardware units that perform other functions of the encoder or decoder. For example, encoder 102 may be implemented mostly on a programmed general purpose computer, but uses a dedicated H.264 encoder for performing H.264 encoding of specific portions of data, while decoder 110 may be implemented entirely using special function hardware units, such as an ASIC chip in a handheld video playback device.

Encoder 102 and decoder 110 are shown in Fig. 1 containing a number of blocks that represent a function or a device that performs a function. Each of the blocks, however,

represent both a function performed and a corresponding hardware element that performs that function, regardless of whether the block is labeled as a function or as a hardware device.

Cartoon footage is often stored in Betacam format. Due to the lossy compression techniques used by Betacam devices, the decoded video sequence slightly differs from the original one. This can be deemed as a kind of noise. Although the noise does not deteriorate the visual quality, it requires more bits and decreases the compression ratio. Therefore, if the source being compressed is from Betacam storage, the noise must be first removed before actual encoding in pre-pre-processing 114. The noise can be classified into two categories: Intra-noise (noise within one frame) and Inter-noise (noise between two frames).

The purpose of intra pre-processing is to remove the noise within one frame, such as an I-frame. Such a frame is usually the first frame in a video shot or scene, since it can be used as a reference for the subsequent consecutive frames in that video shot or scene.

During the procedure of producing animation, one solid area is usually filled with one single color, for example, in one frame, the entire sky is a particular shade of blue. However, after conversion from Betacam or other video storage, there are usually tiny differences in these areas. The Pre-Processor shown in FIG. 1 includes an Intra-processing filter (not shown). The Intra-processing filter is designed to map the colors with similar values into one color, and hence remove the tiny disturbances due to the lossy storage.

An example of the results of intra-noise and pre-processing is shown in Figs. 2A-2D. Fig. 2A is an original cartoon frame before filtering. Fig. 2B is the frame from FIG. 2A after filtering by the Intra-processing filter according to an embodiment of the invention. Fig. 2C is the negative difference between 2A and 2B (black indicates difference), sharpened and the contrast increased so that the differences are more easily human perceptible.

The purpose of inter pre-processing is to remove the noise in P and B-frames, usually the other frames besides I-frames within a video shot. An I-frame is used as a reference to remove the noise in P and B-frames.

Figs. 3A and 3B show two consecutive frames in an example cartoon. The difference between them is shown in Fig. 3C. After sharpening, the noise can be clearly seen from Fig. 3D.

By analyzing the noise distribution, we found that the norm of noise is usually very small, which sets itself apart from real signal, as shown in Fig. 4. A threshold is carefully selected based on the histogram shown in Fig. 4 to remove the noise. The filtered image is shown in Fig. 3E. The filtered image of Fig. 3E, after sharpening, is shown in Fig. 3F.

Besides the above two artifacts, if the original cartoon sequences have been processed by 3:2 pulldown and then de-interlaced, there will be the third artifact: interlacing. 3:2 pulldown is utilized to convert 24 fps source (typically film) into 30 fps output (typically NTSC video) where each frame in the 30 fps output consists of 2 sequential, interlaced fields. In other words, the 30 fps output comprises 60 interlaced fields per second. In a such an

output generated by 3:2 pulldown, the first frame from the source is used to generate 3 consecutive fields – the first two fields making up the first frame of the output with the last field making one half of the next frame. The second source frame is then used to generate the next 2 consecutive fields – the first field making up the second field of the second output frame and the second field making up the first field of the third output frame. With The third source frame we return to using it to generate 3 consecutive fields – the first field making up the second half of the third output frame and the second and third fields making up the fourth output frame. Note that this third output frame now has one field derived from the second source frame and one field derived from the third source field. This is not a problem as long as the output remains interlaced. Continuing with conversion, we return to the 3:2:3:2 cycle (hence 3:2 pulldown) and the fourth source frame is used to generate 2 output fields – both now used for the fifth frame of the output. Using this process repeatedly, every 4 frames of source are converted to 5 frames (10 fields) of output – a ratio of 24:30 – achieving the conversion from 24 fps to 30 fps (60 fields per second, interlaced).

The problem arises when a 30 fps interlaced source is converted into a 30 fps progressive (or non-interlaced) output. In this process the first and second fields for each frame are de-interlaced, yielding 30 non-interlaced frames per second. However, as described above if the 30 fps source was created using 3:2 pulldown, the third frame of the output contains the even lines of one source frame and the odd lines of a different source frame. The result is a frame that contains two half (interlaced) images of any objects that moved between the two frames of the original 24 fps source material. An example of such a frame in the cartoon context is shown in Fig. 5. In this circumstance, you would normally expect to see a frame with the interlace artifact every 5 frames of 30 fps progressive source. The pulldown interlacing artifact is often even more pronounced in cartoon based video than in live action video because the colors and edges of objects or more refined, yielding a striped artifact rather than a more blurred artifact typically seen in live action video.

In one embodiment, de-interlacing is performed by replacing each frame that contains the interlace artifact (every 5 frames) with either the preceding or following frame. In another embodiment, a reverse 3:2 pulldown is performed when converting from a 30 fps interlaced source to a 30 fps progressive output. Alternatively, if the animation is obtained before it is subjected to 3:2 pulldown (in 24 fps format) or in , in which case there will be no interlace artifacts.

Returning to Fig. 1, the encoder includes detecting scene boundaries and segmenting input video into shots 116, calculating the global motion vectors of video sequence 118; synthesizing background for each shot 120; comparing frames with background and extract moving objects 124; and encoding the background and video objects individually 126.

This process improves the compression ratio because the coding area is reduced from the whole frame to small area containing video objects, the background shared by frames

only needs to be encoded once, and by using global motion vectors, the bits needed for motion vectors of each macroblock can be reduced.

In the first step 114, the scene boundaries (start and end point of each scene in the video) are detected by segmenting the cartoon sequence into shots. Each shot then is processed and encoded individually. The scene change detection detects visual discontinuities along the time domain. During the process, it is required to extract the visual features that measure the degree of similarity between frames. The measure, denoted as $g(n, n+k)$, is related to the difference between frames n and $n+k$, where $k \geq 1$. Many methods have been proposed to calculate the difference.

In a many embodiments, one or both of two metrics are used to detect scene change: (1) directly calculate the pixelwise norm difference between frames; and (2) calculate the difference between histograms.

$$g(n,n+k)=\left[\sum_{x,y}\left(I_n(x,y)-I_{n+k}(x,y)\right)^2\right]^{1/2},$$

where $I(x,y)$ is the pixel value of the image at x and y position.

There are several types of transitions between video shots. One type of transition is the wipe: e.g., left-to-right, top-down, bottom-up, diagonal, iris round, center to edge, etc. Wipes are usually smooth transitions for both the pixel difference and histogram difference. Another type of transition is the cut. A cut immediately changes to next image, e.g., for making story points using close-up. Cuts typically involve sudden transitions for both pixel difference and histogram difference. Another type of transition is the fade. Fades are often used as metaphors for a complete change of scene. The last type of transition discussed here is the dissolve. In a dissolve, the current image distorts into an unrecognizable form before the next clear image appears, e.g., boxy dissolve, cross dissolve, etc.

In other embodiments, scene change is detected by analyzing the color sets of sequential frames. Scenes in many cartoons use only have a limited number of colors. Color data for sequential frames can be normalized to determine what colors (palette) are used in each frame and a significant change in the color set is a good indicator of a change between scenes.

Turning to scene change detection 118, given two images, their motion transformation can be modeled as

$$I_t(p)=I_{t-1}(p-u(p,\theta)),$$

where p is the image coordinates and $u(\theta)$ is the displacement vector at p described by the parameter vector θ. The motion transform can be modeled as a simple translational model of two parameters.

The unknown parameters are estimated by minimizing an objective function of the residual error. That is

$$\min_{\theta} \sum_{i} \rho(r_i, \sigma),$$

where r_i is the residual of the i'th image pixel.
$r_i = I_t(p_i) - I_{t-1}(p_i - u(pi, \theta))$.

Hence, the motion estimation task becomes a minimization problem for computing the parameter vector θ, which can be solved by Gauss-Newton (G-N) algorithm, etc.

Turning to background analysis 120, a static sprite is synthesized for each shot. The static sprite serves as a reference for the frames within a shot to extract the moving objects.

The static sprite generation is composed of three steps: common region detection, background dilation, moving object removal.

The frames of one video shot share one background. The common region can be easily extracted by analyzing the residual sequence. The residual image is calculated by calculating the difference between two adjacent frames. If one pixel is smaller than a pre-determined threshold in every frame of residual sequence, it is deemed as background pixel.

Once the common region is detected, it can be dilated to enlarge the background parts. If one pixel is adjacent to a background pixel and they have similar colors, then it is deemed as background pixel.

For the pixels obscured by moving objects and not dilated from the second step, their colors need to be discovered by eliminating moving objects. To detect moving objects, one frame is subtracted from its next frame.

Turning to color clustering 122, as mentioned before, the number of colors in cartoon is much smaller than that of natural video and a large area is filled with one single color. Therefore, a table, such as a master color list, is established in encoder side to record the major colors, which can be used to recover the original color in decoder side by color mapping.

Turning to object analysis 124, after the background image has been generated, the moving objects are achieved by simply subtracting the frames from the background,
$R_t(x,y) = I_t(x,y) - BG(x,y)$

where $I_t(x,y)$ is frame t, $BG(x,y)$ is the background, and $R_t(x,y)$ is the residual image of frame t. Compared with MEPG-4 content-based coding, an advantage of the present algorithm lies in combining the shape coding and texture coding together.

Assume the pixel value ranges [0, 255]. Then we have

$$\left.\begin{array}{l} 0 \le I_t(x,y) \le 255 \\ 0 \le BG(x,y) \le 255 \end{array}\right\} \Rightarrow -255 \le R_t(x,y) \le 255.$$

Then the residual image is mapped to [0,255] in order to make it compatible with video codec.

$$D_t(x,y) = round\left(\frac{I_t(x,y) + 255}{2}\right),$$

where *round(m)* returns the nearest integer towards *m*. After the conversion, both the background and residual image can be coded by generic codecs. However, the color differs from the original one due to rounding operation, called as color drifting. The artifact can be removed by color mapping, as discussed below with respect to post-processing.

Next, both the backgrounds and objects are encoded using traditional video encoding techniques 126. While this is indicated in Fig. 1 as H.264 encoding, to further improve the visual quality, in some embodiments, a hybrid video coding is used to switch between spatial and frequency domain. For example, for a block to be encoded, general video coding and shape coding are both applied and the one with higher compression ratio will be chosen for actual coding. Consider that the cartoon usually has very clear boundary, the hybrid coding method often produces better visual quality than general video coding method.

More particularly, in H.264 encoding, temporal redundancy is reduced by predictive coding. The coding efficiency of the transform highly depends on the correlation of prediction error. If the prediction error is correlated, the coding efficiency of the transform will be good, otherwise, it will not. In the case of cartoon, it is not uncommon for the prediction error not to be highly correlated for certain objects and/or backgrounds and thus H.264 performs poorly. Accordingly, each block is coded by the most efficient mode, DCT or no transform.

Turning to decoder 110, in general, decoding can be considered as an inverse process of encoding, including scene change synthesis 128, background synthesis 130, color mapping 132, object synthesis 134, H.264 decoder 136, shot concatenation 138, and post-processing 140.

After decoding through functions 128-138, there are often two types of artifacts: color drifting and residual shadow. As mentioned above, color drifting is caused by rounding operation when calculating residual images. It can be easily removed by color mapping. More particularly, using the major color list, as supplied by color mapper 132, post-processing 140 compares colors of the decoded image to the major color list and if the decoded image includes colors that are not on the major color list but close too a color on the major color list and significantly different from any other color on the major color list, the close major color is substituted for the decoded color.

Residual shadow arises from the lossy representation of residual image. As a result, the decoded residual image cannot match the background well, thus artifacts are generated.

The residual shadow can be removed by the following steps in post-processing 140: (1) The residual shadow only happens in the non-background area. Considering that the background of residual image is black it can serve as reference on which part should be filtered. (2) The edge map of the decoded frame is then detected. (3) Edge-preserving low-pass filtering is performed in the decoded frame.

In some embodiments, a further modification of H.264 encoding is used. The modification is based on the observation that human eyes cannot sense any changes below human perception model threshold, due to spatial/temporal sensitivity and masking effects. See e.g., J.Gu, "3D Wavelet-Based Video Codec with Human Perceptual Model", Master's Thesis, Univ. of Maryland, 1999, which is incorporated by reference as if set forth herein in its entirety. Therefore, the imperceptible information can be removed before transform coding.

The modification utilized three masking effects: (1) Background luminance masking: HVS (Human Visual System) is more sensitive to luminance contrast than to absolute luminance value. (2) Texture masking: The visibility for changes can be reduced by texture and textured regions can hide more error than smooth or edge areas. (3) Temporal masking: Usually bigger inter-frame difference (caused by motion) leads to larger temporal masking.

A block diagram of an embodiment of the modified encoder is shown in Fig. 6. The modified encoder integrates two additional modules to the framework of conventional video codec: skip mode determination 605 and residue pre-processing 610. Skip mode determination module expands the range of skip mode. Residue pre-processing module removes imperceptible information to improve coding gain, while not damaging subjective visual quality.

To remove perceptually insignificant components from video signals, the concept of JND profile See, e.g. X. Yang et al., "Motion-Compensated Residue Preprocessing in Video Coding Based on Just-Noticeable-Distortion Profile", *IEEE Trans on Circuits and Systems for Video Tech.*, vol. 15, no. 6, pp 742-752, June 2005, which is incorporated by reference as if set forth herein in its entirety, N. Jayant, J. Johnston and R. Safranek, "Signal compression based on models of human perception", *Proc. IEEE,* vol. 81, pp1385-1422, Oct. 1993, which is incorporated by reference as if set forth herein in its entirety. has been successfully applied to perceptual coding of video and image. JND provides each signal to be coded with a visibility threshold of distortion, below which reconstruction errors are rendered imperceptible.

In this section, the spatial part of JND is first calculated within frame. Spatial-temporal part is then obtained by integrating temporal masking.

At the first step, there are primarily two factors affecting spatial luminance JND in image domain: background luminance masking and texture masking. The spatial JND of each pixel can be described in the following equation

$JND_S(x,y) = f_1(bg(x,y)) + f_2(mg(x,y)) - C_{b,m} \cdot \min\{f_1(bg(x,y)), f_2(mg(x,y))\}$, for $0 \le x < H$, $0 \le y < W$, where f_1 represents the error visibility thresholds due to texture masking, f_2 is the visibility threshold due to average background luminance. $C_{b,m}$ ($0 < C_{b,m} < 1$) accounts for the overlapping effect of masking. H and W denote the height and width of

the image, respectively. *mg(x,y)* denotes the maximal weighted average of luminance gradients around the pixel at (x, y) and *bg(x,y)* is the average background luminance.

$$f_1(bg(x,y), mg(x,y)) = mg(x,y)\alpha(bg(x,y)) + \beta(bg(x,y))$$

$$f_2(bg(x,y)) = \begin{cases} T_0 \cdot (1-(bg(x,y)/127)^{1/2}) + 3 & bg(x,y) \le 127 \\ \gamma \cdot (bg(x,y) - 127) + 3 & bg(x,y) > 127 \end{cases}$$

$$\alpha(bg(x,y)) = bg(x,y) \cdot 0.0001 + 0.115$$

$$\beta(bg(x,y)) = \lambda - bg(x,y) \cdot 0.01, \text{ for } 0 \le x < H,\ 0 \le y < W,$$

where T0, γ and λare found to be 17, 3/128 and 1/2 through experiments. See, e.g., C. H. Chou and Y. C. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile", *IEEE Circuits and Systems for Video Tech.*, vol. 5, pp467-476, Dec. 1995, which is incorporated by reference as if set forth herein in its entirety.

The value of *mg(x,y)* across the pixel at *(x,y)* is determined by calculating the weighted average of luminance changes around the pixel in four directions. To avoid over-estimation of masking effect around the edge, the distinction of edge regions is taken into account. Therefore, *mg(x,y)* is calculated as

$$mg(x,y) = \eta \cdot we(x,y) \cdot \max_{k=1,2,3,4}\left\{\left|grad_k(x,y)\right|\right\}$$

$$grad_k(x,y) = \frac{1}{16}\sum_{i=1}^{5}\sum_{j=1}^{5} p(x-3+i, y-3+j) \cdot G_k(i,j),$$

where *p(x,y)* denotes the pixel at *(x,y)*.

The four operators $G_k(i,j)$ are:

$$G_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 3 & 8 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -3 & -8 & -3 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$G_2 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 8 & 3 & 0 & 0 \\ 1 & 3 & 0 & -3 & -1 \\ 0 & 0 & -3 & -8 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{bmatrix},$$

$$G_3 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 8 & 0 \\ -1 & -3 & 0 & 3 & 1 \\ 0 & -8 & -3 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{bmatrix},$$

$$G_4 = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 \\ 0 & 3 & 0 & -3 & 0 \\ 0 & 8 & 0 & -8 & 0 \\ 0 & 3 & 0 & -3 & 0 \\ 0 & 1 & 0 & -1 & 0 \end{bmatrix}.$$

we(x,y) is an edge-related weight of the pixel at *(x,y)*. Its corresponding matrix *we* is computed by edge detection followed with a Gaussian lowpass filter.

$we = e \otimes h$, where e is the edge map of the original video frame, with element values of 0.1 for edge pixels and 1 for nonedge pixels. h is a $k \times k$ Gaussian lowpass filter.

The average background luminance, *bg(x,y)*, is calculated by a weighted lowpass operator, *B(i,j)*, $i,j=1,\ldots,5$.

$$bg(x,y) = \frac{1}{32}\sum_{i=1}^{5}\sum_{j=1}^{5} p(x-3+i, y-3+j)\cdot B(i,j),$$

where $B = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 2 & 0 & 2 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$.

At the second step of JND model generation, the JND profile representing the error visibility threshold in the spatial-temporal domain is expressed as

$$JND(x,y,n) = f_3(ild(x,y,n))\cdot JND_S(x,y,n),$$

where *ild(x,y,n)* denotes the average interframe luminance difference between the *n*th and *(n-1)*th frame.

$$ild(x,y,n) = [p(x,y,n) - p(x,y,n-1) + bg(x,y,n) - bg(x,y,n-1)]/2.$$

f_3 represents the error visibility threshold due to motion. The empirical results of measuring f_3 for all possible inter-frame luminance differences are shown in Fig. 7.

In H.264, a macro-block is skipped if and only if it meets the following conditions all together (See, e.g., Advanced video coding for generic audiovisual services (H.264), ITU-T, March, 2005, which is incorporated by reference as if set forth herein in its entirety.):

The best motion compensation block size is 16x16;

Reference frame is just previous one;

Motion vector is (0,0) or the same as its PMV (Predicted Motion Vector); and

Its transform coefficients are all quantized to zero.

In fact, the above conditions are over strict for cartoon content. Even if the transform coefficients are not quantized to zero, the macro-block can still be skipped as long as the distortion is imperceptible.

Therefore, based on the basic concept of JND profile, in the modified encoder, in skip mode determination 605, the criteria to determine if a macro-block can be skipped. The minimally noticeable distortion (MND) of a macro-block can be expressed as

$$MND(i,j) = \sum_{x=0}^{15}\sum_{y=0}^{15} JND^2(x,y)\cdot\delta(i,j)$$

where $\delta(i,j)$ is the distortion index at point *(x,y)*, ranging from 1.0 to 4.0.

The mean square error (MSE) after motion estimation can be calculated as

$$MSE(i,j) = \sum_{x=0}^{15}\sum_{y=0}^{15}[p(x,y) - p'(x,y)]^2,$$

where *p(x,y)* denotes the pixel at *(x,y)* of original frame and $p'(x,y)$ is predicted pixel. If *MSE(i,j) < MND(i,j)*, the motion estimation distortion is imperceptible and the macro-block can be obtained by simply copying its reference block.

A byproduct is that the computational cost is reduced, since transform coding is not needed for a skipped macro-block.

The purpose of residue pre-processing 610 is to remove perceptually unimportant information before actual coding. The JND-adaptive residue preprocessor can be expressed as

$$\hat{R}(x,y) = \begin{cases} R(x,y) + \lambda\cdot JND(x,y) & R(x,y) - \overline{R}_B < -\lambda\cdot JND(x,y) \\ \overline{R}_B & \left|R(x,y) - \overline{R}_B\right| \le \lambda\cdot JND(x,y) \\ R(x,y) - \lambda\cdot JND(x,y) & R(x,y) - \overline{R}_B > \lambda\cdot JND(x,y) \end{cases},$$

where $\overline{R}_B$ is the average of residue in the block (the block size depends upon transform coding) around *(x,y)*. λ ($0 < \lambda < 1$) is used to avoid introducing perceptual distortion to motion compensated residues.

WHAT IS CLAIMED IS:

1. A system for encoding a video sequence, the system specialized for encoding video of animated or cartoon content, the system comprising:

a background analyzer that removes moving objects from a series of video frames and generates a background definition for a static background used in a plurality of sequential video frames;

a color clusterer that analyzes the colors contained in a video stream and creates a major color list of colors occurring in the video stream;

an object identifier that identifies one or more objects that are constant within a series of video frames except for their position and rotational orientation within the series of video frames; and

a hybrid encoder that encodes backgrounds and objects derived from a video sequence according to one of a plurality of encoding techniques depending on the compression achieved by each of the plurality of encoding techniques.

2. A method of encoding a video sequence, the method specialized for encoding video of animated or cartoon content, the method comprising:

removing moving objects from a series of video frames and generates a background definition for a static background used in a plurality of sequential video frames;

analyzing the colors contained in a video stream and creates a major color list of colors occurring in the video stream;

identifying one or more objects that are constant within a series of video frames except for their position and rotational orientation within the series of video frames; and

encoding backgrounds and objects derived from a video sequence according to one of a plurality of encoding techniques depending on the compression achieved by each of the plurality of encoding techniques.

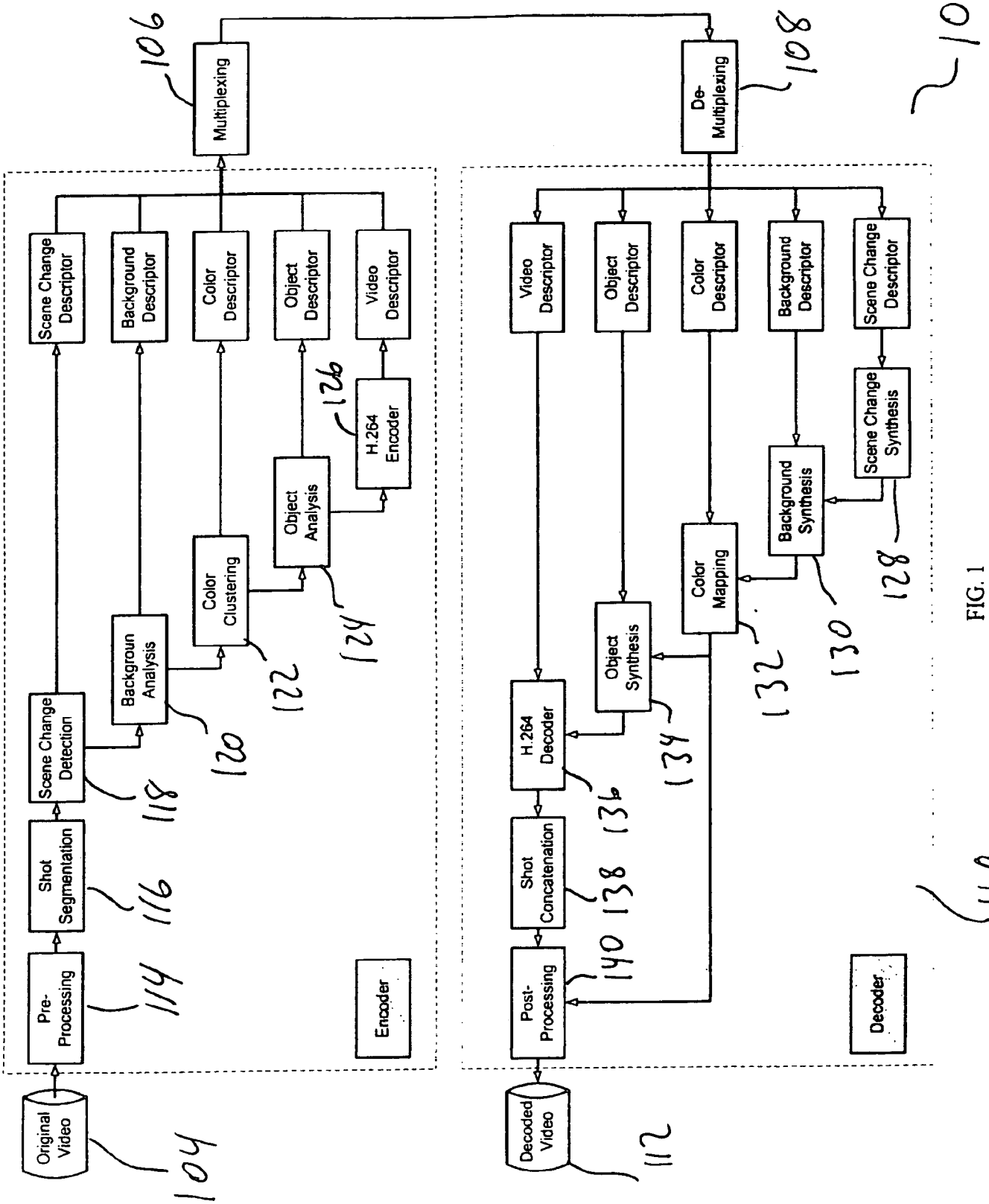

100
Multiplexing
106
De-Multiplexing
108
Scene Change Descriptor
Background Descriptor
Color Descriptor
Object Descriptor
Video Descriptor
H.264 Encoder
126
Object Analysis
124
Color Clustering
122
Backgroun Analysis
120
Scene Change Detection
118
Shot Segmentation
116
Pre-Processing
114
Encoder
Original Video
104
Video Descriptor
Object Descriptor
Color Descriptor
Background Descriptor
Scene Change Descriptor
Scene Change Synthesis
128
Background Synthesis
130
Color Mapping
132
Object Synthesis
134
H.264 Decoder
136
Shot Concatenation
138
Post-Processing
140
Decoder
110
Decoded Video
112

FIG. 1

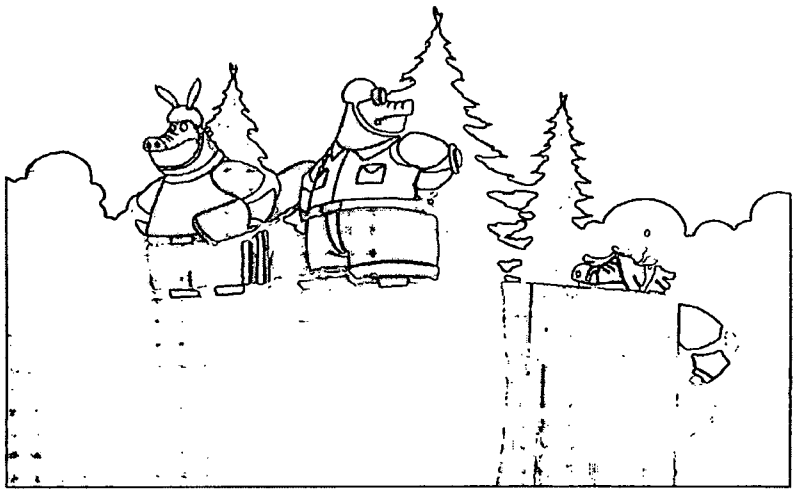


FIG. 2A

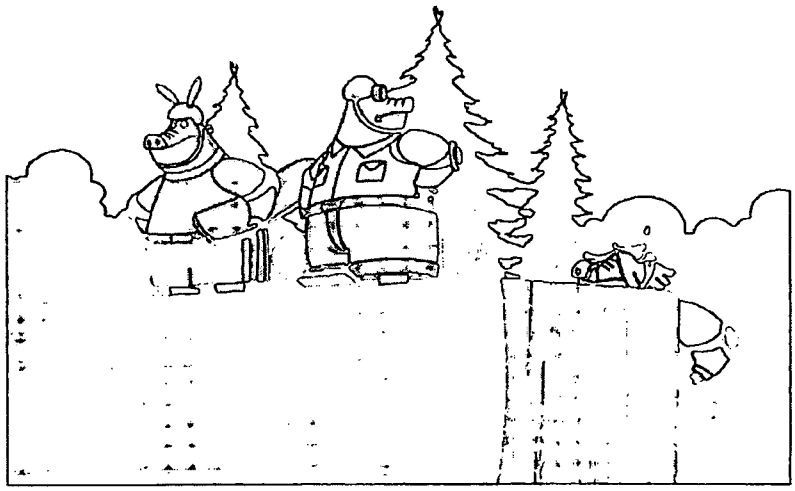


FIG. 2B

FIG. 2C

FIG. 3A

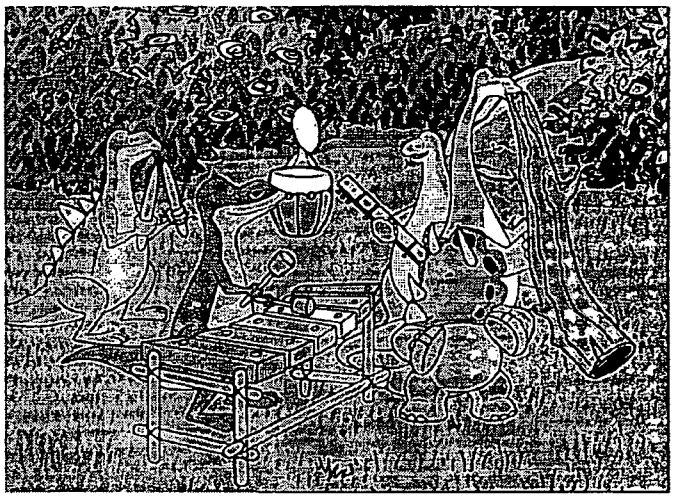


FIG. 3B

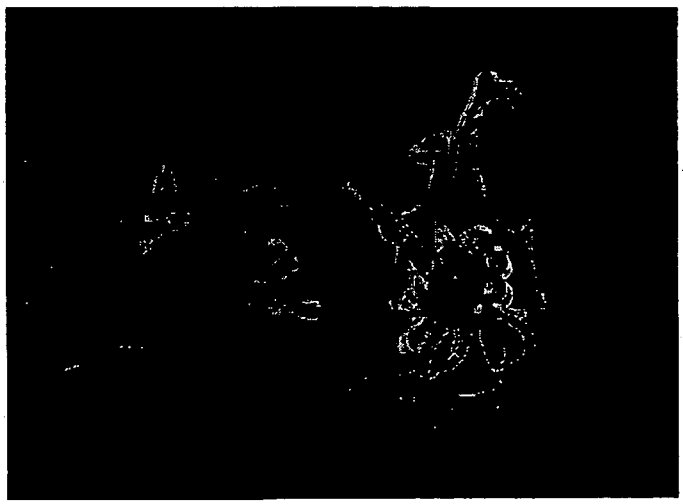


FIG. 3C

FIG. 3D

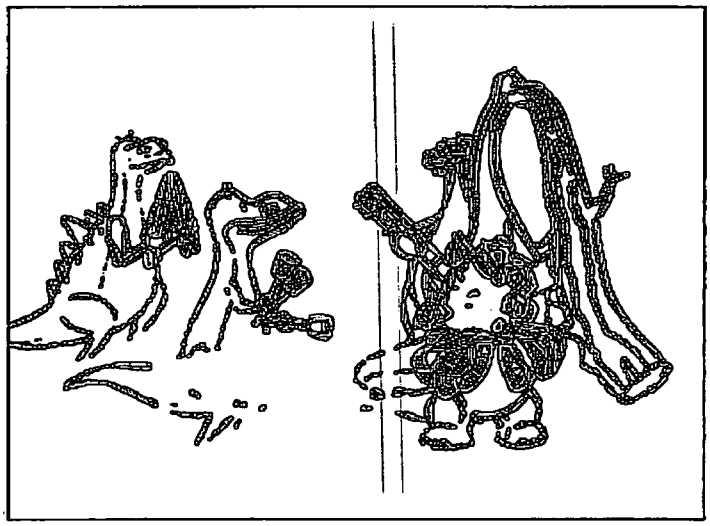


FIG. 3E

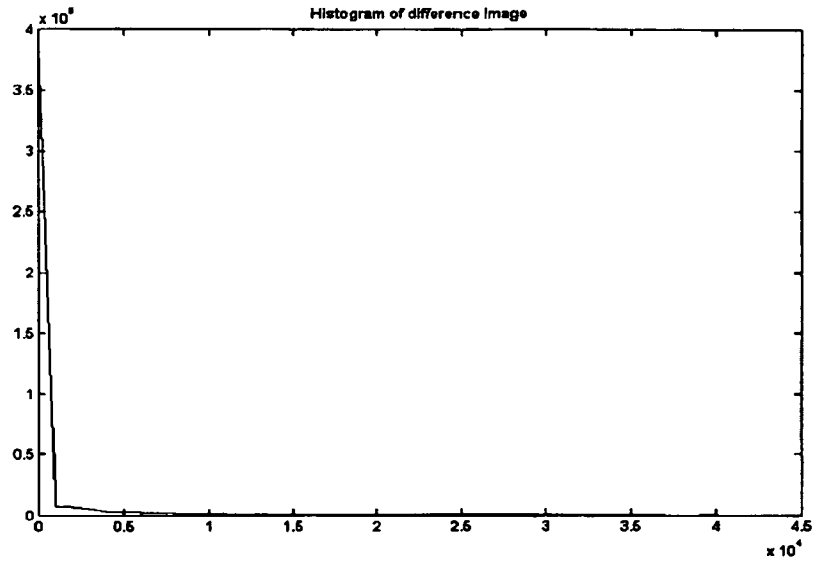

Histogram of difference image
x 10^4
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
x 10^4

FIG. 4

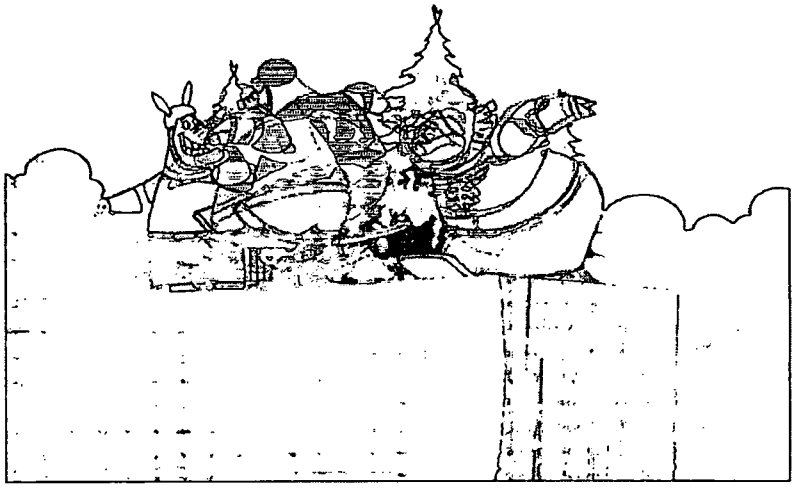


FIG. 5

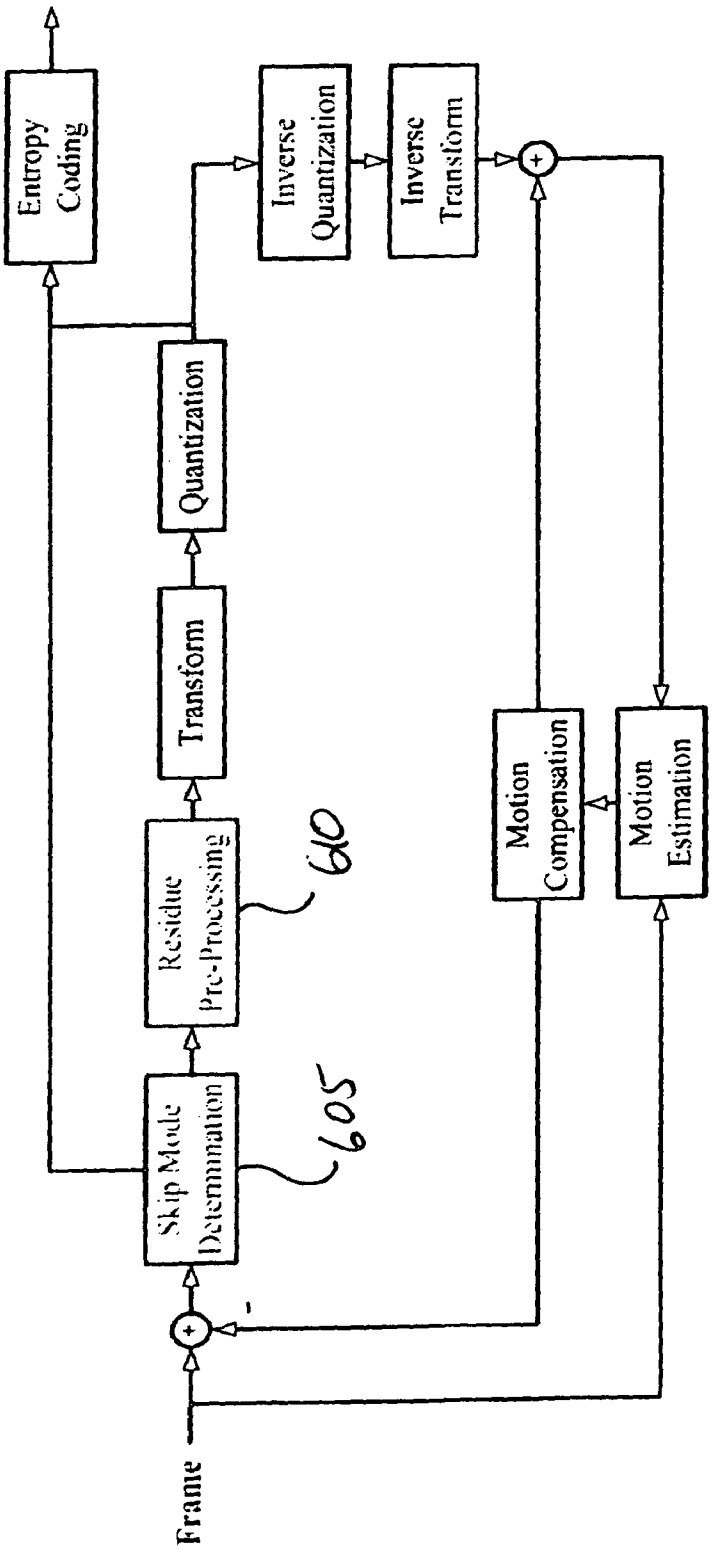

Entropy Coding
Inverse Quantization
Inverse Transform
Quantization
Transform
Residue Pre-Processing
60
Skip Mode Determination
605
Motion Compensation
Motion Estimation
Frame

FIG. 6

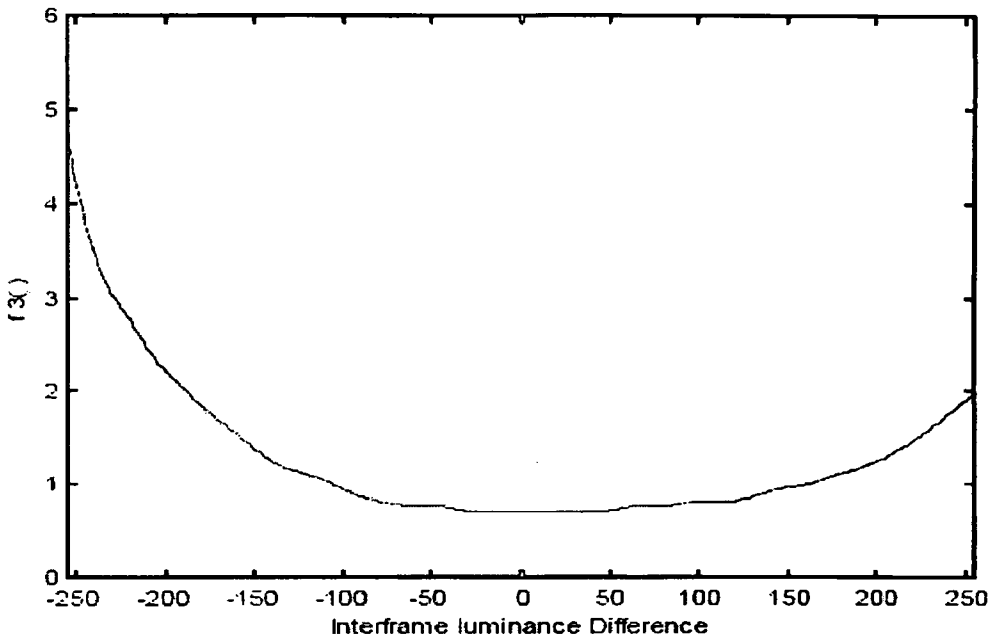

6
5
4
3
2
1
0
f(x)
-250
-200
-150
-100
-50
0
50
100
150
200
250
Interframe luminance Difference

FIG. 7