Signal Processing: *Image Communication* 3 (1991) 259–274
Elsevier

1.15 Mbit/s coding of video signals including global motion compensation

Dirk Adolph and Ralf Buschmann

Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, Universität Hannover, Appelstr. 9a, W-3000 Hannover 1, Germany

Received 4 May 1990
Revised 16 January 1991

Abstract. A coder for television video signals working at a bit-rate of 1.15 Mbit/s is presented. It uses a special coding structure facilitating trick modes and a separated global and local motion compensation. Global motion is described by a central zoom and pan model. For the estimation of global motion a frame matching algorithm is explained. The advantages of using global motion parameters are given by the reduction of the amount of coded motion information as well as the reduction of regions to be replenished in the predicted pictures. Experimental results show that for the given bit-rate of 1.15 Mbit/s the quantization step size for prediction error coding can be reduced by the factor three in case of existing global motion. Also at the decoder side global motion parameters are used for a synthesis of omitted frames resulting in a higher resolution in time for global motions.

1. Introduction

A digital media like the compact disc allows to store a net bit-rate of 1.5 Mbit/s. Here a coding technique is presented that aims at coding arbitrary television material at a bit-rate of 1.15 Mbit/s. In combination with a stereo audio signal encoded at a bit-rate of 256 kbit/s, the capacity and overall bit-rate of a compact disc would be sufficient to store one hour of visual and audio information. The coder described has also been proposed to the Moving Picture Experts Group (MPEG) of the International Standardization Organization (ISO) in October 1989.

The coding of TV signals for digital storage medias has to consider the implications of random access, search modes and resolution improvement for single still pictures. On the other hand, real-time operation is mandatory for the decoder but only optional although desirable for the encoder.

Beginning the reconstruction of pictures out of the bit stream of a coded sequence in an arbitrary place requires a different coding strategy compared to known algorithms. Entry points are created in the bit stream where decoding can be started. For this reason the coded sequence is divided into groups of frames. The first picture to be read from a group of frames (GOF) by the decoder is coded independent of other pictures, while the remaining pictures depend on other pictures of the same GOF. The number of pictures in one GOF determines the decoding delay.

The basic concept of the presented coder is motion compensating hybrid coding. The prediction error of a Differential Pulse Code Modulation (DPCM) is encoded by a Discrete Cosine Transformation (DCT) [7]. The quantization in the frequency domain is adapted to the human visual perception [2, 3].

Motion compensation techniques are used for the improvement of prediction and for frame extrapolation. In this coder proposal a global and a local motion compensation are applied. In a first stage global motion, generated by zoom and pan of the camera, is estimated and described by three parameters. In many television sequences a pure zoom and pan model describes global motion almost exactly. In a second stage the local motion due to moving objects is described by a displacement vector field. This local motion is superimposed to the global motion. For global as well as for local motion estimation matching techniques are applied.

This paper describes the coding algorithm and also presents the results obtained by global motion compensation. The contents of the following sections are as follows. Section 2 explains the coding structure and discusses the complete block diagram. Analysis of global and local motion is described in Section 3. The estimation of global motion parameters is presented in Section 3.1, while segmentation and estimation of local motion follow in Section 3.2. In Section 4 the usage of global and local motion information for motion compensated prediction and motion compensated extrapolation is described. Experimental results obtained by computer simulations are discussed in Section 5.

2. Basic coder description

2.1. Input format conversion

To achieve a first data compression for coding CCIR601 television sequences at a bit-rate of 1.15 Mbit/s, input format conversion is performed. Instead of the CCIR Recommendation 601 format 4/2/2 a reduced input format (RIF)—very similar to Common Intermediate Format (CIF)—is used as source for coding. The RIF has different parameters for the 525/60 and 625/50 versions of CCIR format (see Table 1). The conversion from CCIR format to RIF and vice versa is performed by three-dimensional filtering, executed separately once in time and twice in space.

The conversion from CCIR to RIF is achieved by the following procedure. Temporal filtering is performed by omitting every second field. By this way also the vertical resolution is reduced by the factor 2 due to the interlace inherent in the CCIR format. Further the spatial resolution is reduced by sampling rate conversion including aliasing cancellation. In order to achieve compatibility with a 16 × 16 block raster, the picture width is reduced by symmetrically cutting off left and right columns in the pictures.

The reverse conversion from RIF to CCIR is executed at the decoder side. First the horizontal picture size is enlarged by repeating the outermost left and right columns replacing those columns which were cut off by forward conversion. Spatial enlargement by the factor 2 is realized by inserting zero picture elements (pels) and filtering with a symmetric FIR Filter. Noninterlaced–interlaced conversion is achieved by repeating each frame once and additional vertical down-filtering of each second field.

2.2. Coding structure

Besides the normal playback mode of a conventional coder, one goal of the presented coder is the support of the so called 'trick modes', e.g. random access, search modes and still mode. Random access denotes the ability of the decoder, to reach an arbitrary frame within a fixed time delay. Search modes are fast forward and fast reverse playback modes to reach interactively any frame in the encoded sequence. Still mode means that special frames can be shown in a higher resolution using additional information out of the bit stream.

All trick modes require the ability of synchronizing to the bit stream without starting from the beginning. This is reached by inserting entry points into the bit stream. Thereby the sequence of pictures is subdivided in groups of frames, henceforth called GOFs, containing a fixed number of frames each. In each GOF one frame is coded independently of other frames (intraframe), the remaining frames are coded depending on other frames (interframe). The number of frames belonging to one GOF determines the maximum random access delay of the decoder.

Table 1

Parameters of CCIR format and reduced input format (RIF)

Video system		Active lines		Number of active pixels per line		Fields per second
		LUM	CHR	LUM	CHR	
525/60	CCIR format	480	480	720	360	60 Hz interlaced
	RIF	240	120	352	176	30 Hz noninterlaced
625/50	CCIR format	576	576	720	360	50 Hz interlaced
	RIF	288	144	352	176	25 Hz noninterlaced

The grouping of eight frames to a GOF is resulting in the coding structure depicted in Fig. 1, showing nine successive frames in RIF at a frame rate of 25 Hz. The GOF is marked grey. At the top every frame is labeled with a combination of number and capital letter. The number indicates the position in time, the letter the type in which the frame is coded.

The 'A' frame in each GOF is intraframe coded, while 'B' and 'C' frames are interframe coded. The 'A' frame is coded like a still picture by means of the baseline system of JPEG in sequential mode [9]. So at the decoder side every 'A' frame can be decoded directly. 'B' frames are coded using motion compensated prediction. So decoding a 'B' frame is possible only if the previous 'A' or 'B' frame has been decoded. For example, to reconstruct frame 3B at the decoder side, frame 1A must also be decoded. In case of frame 7B, frames 1A, 3B and 5B are all necessary. 'C' frames are generated from the neighboring previous and following 'A' and 'B' frames by motion compensated interpolation [8]. Decoding frame 8C constitutes the worst case of random access. As the generation of frame 8C requires frames 7B and 9A, more than one GOF must be decoded to generate this frame. In extension to the present version of the codec, where 'C' frames are generated at the decoder side only, an additional error signal may be transmitted.

For motion compensated prediction of 'B' frames side information has to be transmitted. The side information about motion is separated into global and local motion information, which will be described in detail in Section 3. Global motion information is transmitted for each frame by a global motion parameter set. So either parameter sets are transmitted in a GOF (See Fig. 2). Local motion is transmitted for 'B' frames by means of vector fields. Therefore three local motion vector fields are coded in each GOF.

Motion compensated interpolation of 'C' frames is done by means of transmitted global motion parameter sets only. The parameter set describing

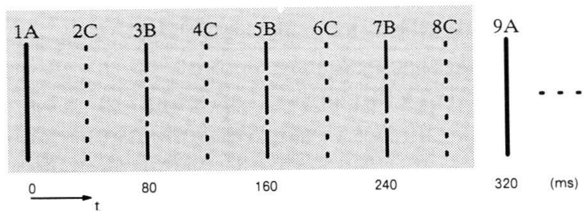

Fig. 1. Coding structure of eight frames belonging to a group of frames (GOF). Numbers denote the frame position, capital letters denote the type of coding: A: intraframe coding; B: predictive coding; C: motion compensated interpolation.

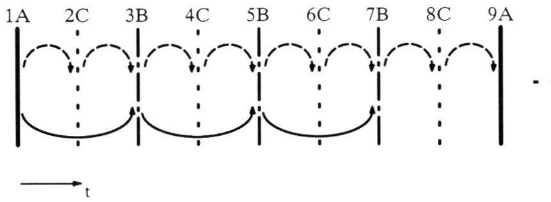

Fig. 2. Motion information transmitted for one group of frames (GOF). Dashed arrows denote parameters describing global motion. Solid arrows denote vector fields describing superimposed local motion: A: intraframe coding; B: predictive coding; C: motion compensated interpolation.

global motion from the previous to the actual 'C' frame, as well as the parameter set describing global motion from the actual 'C' frame to the following frame are necessary. Therefore the two global parameter sets between frames 7B, 8C and 9A also are mandatory.

The usage of the transmitted motion information for motion compensation of 'B' and 'C' frames is described in detail in Sections 4.2 and 4.3.

The occurrence of scene cuts in a sequence requires a special handling within a GOF. If the scene cut does not hit frame 'A', which is coded intraframe, motion compensated prediction or interpolation would be executed between pictures of different scenes. This would result in an increasing error signal and a loss of picture quality. In order to avoid this effect the last frame before a scene cut is frozen and displayed until the GOF has finished. The following GOF starts coding the new scene with an 'A' frame in intra frame code. The scene cut is delayed at the utmost by one GOF display time, if the last frame of the old scene happens to be an 'A' frame.

2.3. *Coder block diagram*

The proposed algorithm for coding moving images is based on a hybrid coding scheme involving more than one redundancy reduction method. The coder is working by using DPCM and DCT. A complete block diagram is depicted in Fig. 3. The basic structure of the encoder and decoder is a DPCM loop, where the previously reconstructed frame is kept in a frame memory and the difference between the current frame and the frame memory—the prediction error signal—is coded by the JPEG unit. Only if the switch controlled by the Inter/Intra Mode Control unit is opened, intra mode is selected and the original input frame (an 'A' frame) is coded. In this case the segmentation is not active and the whole frame is transmitted. The Intra/Inter Mode Control unit is acting frame wise and generates the coding structure of a GOF.

The units named JPEG and JPEG^{-1} are coding and decoding by using the sequential base line system proposed by JPEG [9]. In the base line system any input picture—compatible to an 8 × 8 pel block raster—is transformed blockwise by DCT. Transformation is followed by a quantization using a 'visibility matrix', which is adapted to the human visual perception (see Table 2). In case of interframe coding a homogeneous visibility matrix is used due to the visual perception of error signals, i.e., the visibility weights are the same for all coefficients. The numbers in Table 2 multiplied by an integer factor represent the step size of the quantization as a function of the coefficient index. The scaling factor is controlled by the bit amount which is necessary for coding the quantized coefficients in frequency. In order to guarantee a quantization with a prescribed bit amount, two-pass encoding is applied, generating picture adaptive Huffman codes.

For further redundancy reduction an analysis of motion and a segmentation is carried out. As shown in Fig. 3, two successive original frames are used for analysis and segmentation generating three streams of side information as output. The first stream describes the global motion parameters zoom and pan, by the second one a displacement vector field due to local motion is described. The third stream contains a segmentation mask, separating regions of global and local motion. These streams are encoded in the side information encoder. While the zoom and pan parameters are transmitted for each frame, the local motion vector field and segmentation mask are transmitted for 'B' frames only. In the same manner as at the decoder side, the analyzed global and local motion are compensated in the DPCM loop to generate the prediction frame.

Table 2

Visibility matrix for quantization of transform coefficients used after DCT on 8 × 8 pel blocks of intraframe coded pictures

+⟶ f_x	Luminance components								Chrominance components							
↓ f_y	16	11	10	16	24	40	51	61	17	18	24	47	66	99	99	99
	12	12	14	19	26	58	60	55	18	21	26	66	99	99	99	99
	14	13	16	24	40	57	69	56	24	26	65	99	99	99	99	99
	14	17	22	29	51	87	80	62	47	66	99	99	99	99	99	99
	18	22	37	58	68	109	103	77	99	99	99	99	99	99	99	99
	24	35	55	64	81	104	113	92	99	99	99	99	99	99	99	99
	49	64	78	87	103	121	120	101	99	99	99	99	99	99	99	99
	72	92	95	98	112	100	103	99	99	99	99	99	99	99	99	99

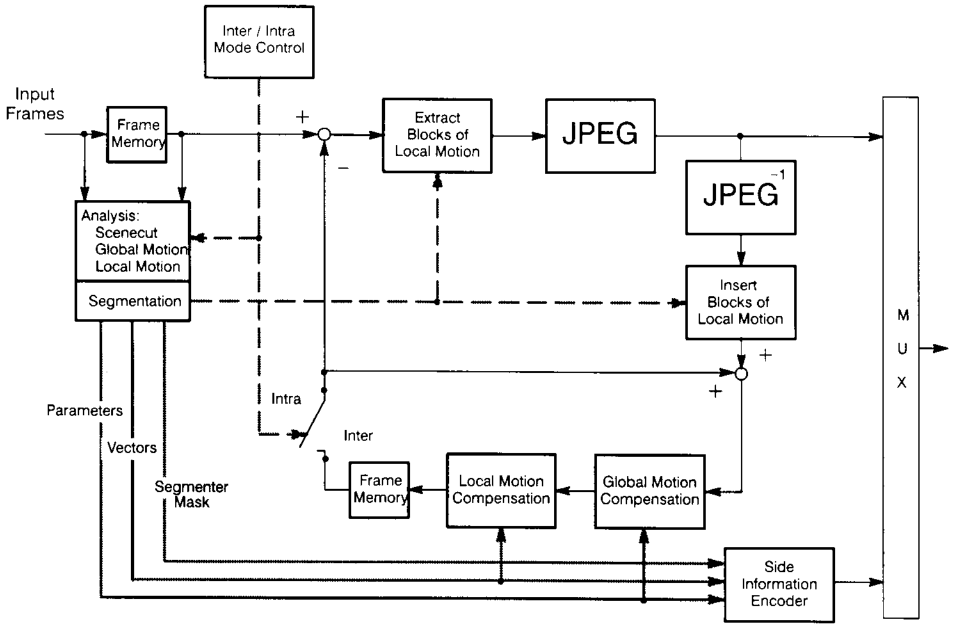

Fig. 3. Block diagram of the encoder.

In the block diagram the two units handling extraction and insertion of blocks of local motion are controlled by the segmentation mask. By this form of conditional replenishment only those blocks are processed by the JPEG coder, which the segmentation declares as locally moving. The remaining blocks are described by the previous frame and global motion parameters only. For each locally moving block of 8 × 8 pel one local motion vector is estimated and transmitted. After compensation of local motion in this block, the resulting prediction error passes the JPEG coder.

In contrast to the general principle of the coding structure, in this codec a motion compensated extrapolation instead of interpolation is used for synthesis and no error signal is coded for extrapolated frames. So the block diagram of the proposed decoder in Fig. 4 additionally contains motion compensated extrapolation. Two successively decoded frames and the zoom and pan parameters are used to generate 'C' frames as described in detail in Section 4.3.

3. Analysis of global and local motion

Many changes in contents of video sequences are due to motion. Motion in video sequences are either caused by single objects in front of the camera called *local motion*, or by motion of the camera itself resulting in *global motion* of the whole scene. Therefore motion estimation generally has to deal with a composite motion. This paper proposes an approach which separates motion into its global and local components. The advantage of

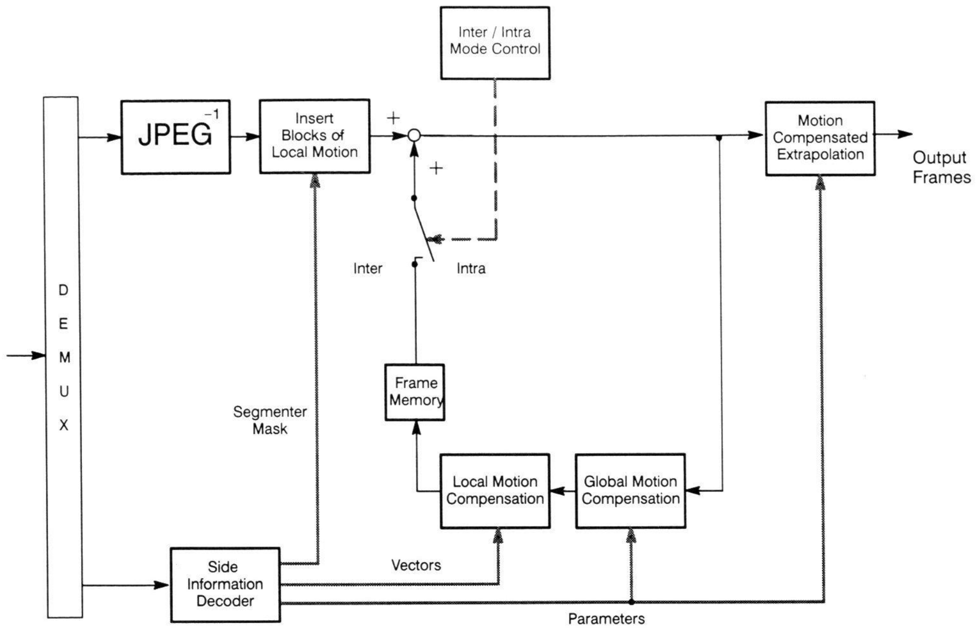
Inter / Intra Mode Control
JPEG^{-1}
Insert Blocks of Local Motion
+
+
Motion Compensated Extrapolation
Output Frames
D E M U X
Inter
Intra
Frame Memory
Segmenter Mask
Local Motion Compensation
Global Motion Compensation
Side Information Decoder
Vectors
Parameters

Fig. 4. Block diagram of the decoder.

this separation is twofold: on one hand a gain is reached by coding motion information in large picture regions by a few parameters only, and on the other hand knowledge about global motion is used at the decoder side for a synthesis of pictures without further information.

For the separation of the two components of motion a simple global motion model is introduced. The assumption of pure central zoom and pan for global motion is sufficient in many typical video sequences and can be described by

$$S_k(x; y) = S_{k-1}(z \cdot x + p_x; z \cdot y + p_y), \qquad (1)$$

where S_k is the luminance signal of the current frame, S_{k-1} that of the previous, z is the zoom factor, $(p_x; p_y)$ is the pan vector and $(x; y)$ are the pel coordinates related to the picture center. Due to camera geometry the zoom center is identical with the picture center.

The superimposed local motion of objects in front of the camera is described by a displacement vector field with integer pel accuracy.

3.1. Estimation of global motion by frame matching

In order to estimate global motion parameters several procedures have been developed. Kummerfeld et al. [6] proposed a two-stage algorithm. In the first stage a displacement vector field is calculated which is used in the second stage to evaluate the parameters zoom and pan by linear regression. Since exact matches of objects or blocks cannot be found, the vector fields exhibits systematic errors leading to systematic errors for the parameters of zoom and pan also. To avoid these errors Hötter [4] presented a differential technique for the estimation of global motion, which measures the parameters zoom and pan directly from the image signal. Using this technique Hötter achieves an absolute maximum estimation error in the displacement vector field less than 20% [4].

To increase the accuracy of the global motion estimation reached by Kummerfeld and Hötter, in the new algorithm zoom and pan are evaluated by using a frame matching technique.

The frame matching algorithm used to estimate a parameter set describing global motion with a very high accuracy can be described similar to known block matching techniques. While block matching algorithms estimate two vector components for each block, the new algorithm estimates three parameters describing the global motion of the whole image (see (1)). The luminance signals of two successive frames are matched using the mean absolute difference (MAD) as matching criterion. For a description of the algorithm the three parameters are united in a parameter vector $\boldsymbol{p}$, according to

$$\boldsymbol{p} = (p_x; p_y; z)^{\mathrm{T}}. \qquad (2)$$

Under the condition that the MAD criterion is sufficient for the detection of global motion parameters, frame matching can be used to detect the true parameter vector $\boldsymbol{p}_{\text{true}}$ with arbitrary accuracy. It is obvious that a full search strategy leads to a non-acceptable computational effort.

As a consequence of this, the space of possible parameter vectors has to be restricted. Assuming the existence of a start parameter vector $\boldsymbol{p}_0$, a 3-D parameter cube can be defined from which candidate parameter vectors are selected (see Fig. 5). The center of this cube is determined by the start parameter vector and its borders are determined by the accuracy of the start parameter vector. The assumed start parameter vector $\boldsymbol{p}_0$ must be very close to the true vector $\boldsymbol{p}_{\text{true}}$, and the borders of the cube with candidate parameter vectors has to be chosen in such a way that $\boldsymbol{p}_{\text{true}}$ is within the cube borders.

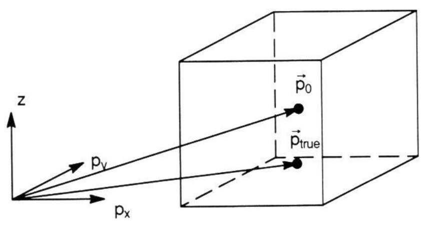

Fig. 5. Cube of candidate parameter vectors around the start parameter vector $\boldsymbol{p}_0$ including the true parameter vector $\boldsymbol{p}_{\text{true}}$.

To get the start parameter vector a hierarchical block matching algorithm is used generating a homogeneous vector field, which is close to the true displacements in the image sequence [1]. In order to eliminate the influence of uncertain regions for the calculation of the parameters by linear regression, a mask is generated afterwards marking those areas containing locally moving objects. For this purpose the vector field is divided into rectangular blocks and mean and variance of the vector components are measured in each block. Those blocks whose mean and variance exceed a threshold are eliminated to maintain only vectors describing global motion. Based on the selected displacement vectors a start parameter vector can be determined by linear regression.

Similar to usual block matching algorithms in this algorithm a three step search method is used, increasing the accuracy of the estimated parameters in each step. Due to the limitation of bit amount the accuracy of the final parameter vector is set to $\pm 0.2 \cdot 10^{-3}$ for zoom and $\pm 20.0 \cdot 10^{-3}$ for pan.

During the frame matching procedure global motion has to be compensated according to the values of the candidate parameter vectors. The global motion compensation must be done separately for each candidate parameter vector. The candidate minimizing MAD in the current step is used as a start parameter vector of the following step.

In order to limit the computational effort, the regions used for matching must be reduced. It is obvious that regions with high activity in the luminance signal—such as edges and textures—mainly contribute to the MAD criterion. For the determination of activity the frame is subdivided into blocks and the variance of the luminance signal is calculated. Selecting only blocks of high variance, computational effort can be reduced drastically. Blocks are therefore ordered according to their activity, and only the blocks of highest activity are analyzed during the matching algorithm. Using this strategy on the test sequence ‘Table-tennis’, the computation expense was reduced to 70%, 50% and 40% in the three steps compared to a matching of complete frames. The method does not decrease estimation accuracy.

3.2. Segmentation and local motion estimation

Global motion compensation is taking into account motions of the whole scene due to actions of the camera only. But if there are moving objects in the scene, an additional compensation of local motion is necessary. The separation of regions, which are generated using the global parameters only, from those which need replenishment is provided by the segmenter. Blocks not sufficiently described by global motion parameters are declared as 'local motion blocks' and are described additionally by a local motion vector and the resulting prediction error signal.

To generate the segmenter mask as well as the vector field describing local motion, two successive frames are processed as depicted in Fig. 6, where the older frame S_{k-1} has been global motion compensated beforehand. Change detection and a hierarchical displacement estimation is performed to frames S_k and S_{k-1} creating a binary change detection mask and an initial vector field, respectively. Using both the change detection mask and the initial vector field, segmentation and corona reduction generate a segmenter mask and a vector field describing local motion, respectively. As can be seen in the block diagram the segmenter mask and the local motion vector field are mutually dependent.

The change detection algorithm adapts to different video signals and is described in detail in [8]. A description of the hierarchical block matching algorithm can be found in [1]. The parameters of the matching procedure used in this coder are shown in Table 3. The block matching algorithm works with three hierarchy levels starting with large windows, a large measurement grid and a strong low-pass filter. In the following, hierarchy levels parameters become smaller, so that in the third level measurement grid and window size equal the block raster of DCT. Furthermore the last level is matching unfiltered luminance values. The maximum vector which can be generated by this hierarchical blockmatcher is ±17 pel taking into account the large velocities possible in television scenes.

Table 3

Parameters of hierarchical block matching in the three hierarchy levels. All numbers are in [pel] units. The parameters are employed horizontally and vertically generating a maximum displacement vector of ±17 pel.

	Hierarchy level		
	I	II	III
Distance of measurement points (measurement grid)	32	16	8
Measurement window size	64	16	8
Filter window size	5	3	1
Subsampling inside window	4	2	1
Maximum vector update	±7	±7	±3

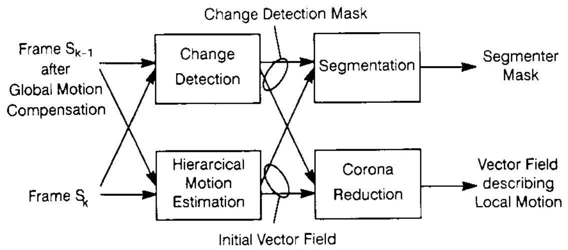
Change Detection Mask
Frame S_{k-1} after Global Motion Compensation
Change Detection
Segmentation
Segmenter Mask
Frame S_k
Hierarcical Motion Estimation
Corona Reduction
Vector Field describing Local Motion
Initial Vector Field

Fig. 6. Block diagram of the generation of segmenter mask and vector field describing local motion.

Change detection mask and initial motion vector field are created independently. The combination of these pieces of informations as depicted in the right half of Fig. 6 has advantages. Change detection mask and initial motion vector field can be combined to mutually reduce their inherent errors. On one hand the hierarchical displacement estimation creates a vector corona around moving objects due to the large matching windows in the first hierarchy levels. This vector corona can be reduced by erasing those vectors, which appear in regions declared as unchanged by the change detection. On the other hand the change detection mask is generated by thresholding the frame difference signal. Regions in which the frame difference exceeds the threshold due to noise are also assumed to be changed. Involving the motion information, those regions can also be reduced which were declared changed by the change detection due to noise. Therefore the change detection mask is set to unchanged, wherever no local motion occurs. So the combination of change information and motion information improves the robustness of the segmentation against noise and the reliability of the local motion estimation.

For coding the segmentation mask and the vector field describing local motion with a low data amount the following simplifications are carried out. The segmenter mask—available in pel accuracy—is converted to a block accuracy of 16 × 16 pel by setting those blocks to locally moved which contain less than 5% of changed area. For the purpose of coding the local motion vector field is simplified by selecting the center vector of each locally moving block in the final vector field. The single vectors are of integer pel accuracy. At encoder and decoder side a vector field valid for each pel is reconstructed by bilinear interpclation. Based on the interpolated vector field local motion compensation is executed.

4. Compensation of global and local motion

4.1. Compensation of global motion

Global motion compensation is necessary for motion compensated prediction ('B' frames) as well as for motion compensated extrapolation ('C' frames). The generation of compensated pictures is done by computing a mapping vector $\mathbf{v}$ for each pel of the new picture. The vector $\mathbf{v}$ describes the offset used to calculate the luminance value for each pel and is determined by the equation similar to (1)

$$\mathbf{v}(x;y)=(x\cdot(z-1)+p_x;\, y\cdot(z-1)+p_y). \quad (3)$$

In general, global motion compensation is not possible in all parts of a frame. New regions entering the picture cannot be described. For example panning the camera right generates new contents at the right border of the picture. In this coder unknown border regions in the frame are created by repeating the last known columns and lines.

If the calculated mapping vector does not hit a location on the sampling grid, filtering is necessary. This is done by a mapping filter in case of prediction and by a bilinear filter in case of extrapolation.

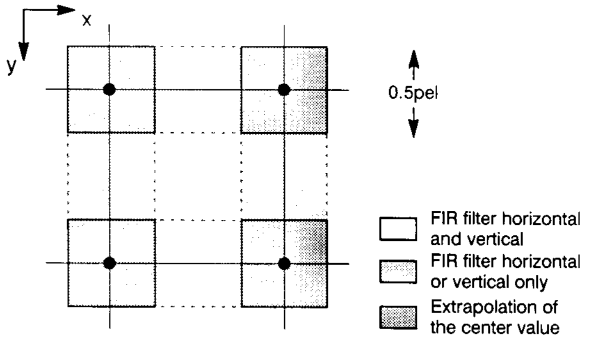
x
y
0.5pel
FIR filter horizontal and vertical
FIR filter horizontal or vertical only
Extrapolation of the center value

Fig. 7. Subdivision of the image plane employed by the mapping filter of half pel accuracy for prediction.

As shown in Fig. 7 the mapping filter subdivides the image plane into regions which are treated differently. If the mapping vector points to the central region between four neighboring pels, a symmetric FIR filter with the coefficients (2, −22, 148, 148, −22, 2) is used horizontally and vertically. If the mapping vector points to a location near to the sampling grid, the corresponding original gray value is selected. In the remaining other regions filtering is applied horizontally or vertically only. The mapping filter is of half-pel accuracy.

The bilinear filter works according to

$$S(x;y)=S_{11}\cdot(1-h)\cdot(1-v)+S_{12}\cdot h\cdot(1-v) + S_{21}\cdot(1-h)\cdot v+S_{22}\cdot h\cdot v. \quad (4)$$

As shown in Fig. 8, S_{11}, S_{12}, S_{21}, S_{22} are the luminance values at the spatial positions on the sampling grid nearest to $S(x;y)$, and h and v represent the fractional parts of the real coordinate $(x;y)$ starting from S_{11} as the respective origin of the coordinates diagram.

The advantage of the mapping filter is its low impact on the sharpness of the compensated frame. Therefore it is employed for prediction. In case of synthesis, i.e., extrapolation, the mapping filter has the disadvantage of leading to an inhomogeneous

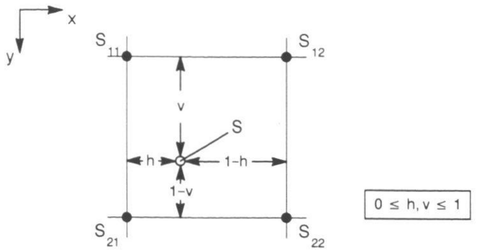

Fig. 8. Calculation of the luminance value S using a bilinear filter. S_{11}, S_{12}, S_{21}, S_{22} are the luminance values at the nearest sampling grid positions; h and v are the horizontal and vertical distances, respectively.

motion impression. This insufficient subjective impression of global motion can be avoided if compensation is done with the bilinear filter. In this case the loss of sharpness due to bilinear filtering must be accepted.

4.2. Motion compensated prediction of 'B' frames

As shown in Figs. 3 and 4 motion compensated prediction is executed by means of the global motion parameters and the local motion vector field. In a first step the global motion compensation and in a second step the local motion compensation are carried out.

For the global motion compensation in the first step two global motion parameter vectors are transmitted as can be seen in Fig. 2. These two global motion parameter vectors $(p_{x1}; p_{y1}; z_1)$ and $(p_{x2}; p_{y2}; z_2)$ are accumulated according to

$$p_x = z_2 \cdot p_{x1} + p_{x2}, \tag{5.1}$$

$$p_y = z_2 \cdot p_{y1} + p_{y2}, \tag{5.2}$$

$$z = z_1 \cdot z_2, \tag{5.3}$$

before (3) is employed as described in Section 4.1.

Hereafter in the second step the local motion vector field is applied. Superimposing global and local motions yield the resulting motion compensation for the prediction of 'B' frames.

4.3. Motion compensated extrapolation of 'C' frames

As shown in Fig. 4, motion compensated extrapolation is employed at the decoder side to synthesize the omitted 'C' frames generating a higher resolution in time for the decoded sequence. For motion compensated extrapolation global motion parameters are used only. Since the omitted 'C' frame S_k is to be extrapolated from the two neighboring coded frames S_{k-1} and S_{k+1}, global motion compensation must be executed twice. The global parameter vectors joining S_k with its neighbor frames are used for this purpose.

The principle of extrapolation is shown in Fig. 9 for the example of pure zoom-out. Zoom-out stands for a focal length reduction of the camera during a scene. Using the parameter vector describing the global motion from S_{k-1} to S_k, the previous frame S_{k-1} is compensated forward creating S_{k-1}^{com}. The frame S_{k+1} is compensated backward resulting in S_{k+1}^{com} by using the inverse of the parameter vector which describes the global motion from S_k to S_{k+1}.

The defined parts of S_{k-1}^{com} are inserted in frame S_k. Only those parts of S_k which cannot be extrapolated from S_{k-1}^{com} are inserted from S_{k+1}^{com}. The general assignment of regions to be inserted from the previous or the following frame is calculated by means of the global motion parameter vectors.

When extrapolation instead of interpolation is used for synthesis, double contours are avoided and sharpness is retained in the picture. This would be difficult in case of interpolation, where S_k would be averaged between S_{k-1}^{com} and S_{k+1}^{com}. It is obvious that the error signal between S_k and the corresponding original frame is larger for extrapolation than for interpolation. So in case of transmitting the error signal to the decoder, the interpolation technique will be preferred.

It is worth noting that the extrapolation of the omitted 'C' frames with global motion compensation leads to a smooth impression of motion although local motion is not compensated for these frames. This subjective masking effect takes place whenever global motion is present.

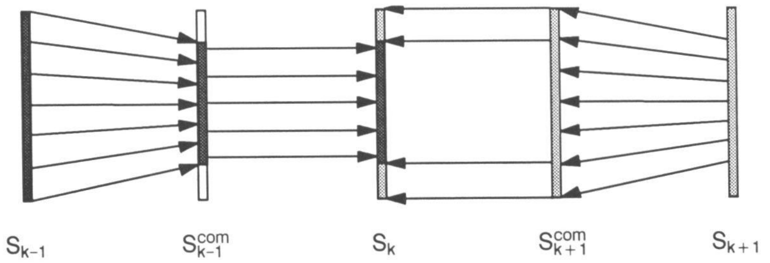

Fig. 9. Motion compensated extrapolation of an omitted frame S_k in case of pure zoom-out. The neighboring coded frames S_{k-1} and S_{k+1} are used.

5. Experimental results

The codec presented in this contribution has been developed and tested by computer simulations. The test sequence 'Table-tennis' consists of 250 frames in 625/50 CCIR format distributed by ISO Moving Picture Experts Group has been converted to RIF (described in Section 2) and used for the simulations. This test sequence consists of four parts of different motion as shown in Table 4. The simulation results without upconversion are depicted in the following example images. In order to check the causality of the simulated coder, a physical bit stream of 1.15 Mbit/s was produced as output of the encoder. The simulation results have been generated by decoding this bit stream with a separate decoder program. A real time image sequence display system was used to judge the quality of the decoded pictures.

The advantages of global motion compensation are demonstrated by several test pictures. In Fig. 10 two vector fields are shown generated by the hierarchical displacement estimation algorithm described in Section 3.2. The vector field in Fig. 10(a) has been estimated between two original frames. As can be seen global motion due to zoom-out is added to the local motions of ball, bat and arm. Figure 10(b) shows that after global motion compensation only vectors describing local motion remain. The motion description of the ball fails because of its high velocity.

Table 4

Scenes of test sequence 'Table-tennis' in RIF

Scene no.	Picture no.	Take and type of global motion
I	1–7	Take 1, no motion
II	8–73	Take 1, zoom out
III	74–120	Take 2, no motion
	121–143	Take 3, no motion
IV	143–250	Take 3, pan right

The vector fields in Fig. 10 also illustrate that global motion description significantly reduces the transmission effort for motion as large regions in the picture are sufficiently described by global motion. Global motion description also decreases the amount of bits used for coding the prediction error as illustrated in the following figures.

Corresponding to the vector fields in Fig. 10 prediction error signals can be found in Fig. 11. A comparison of Fig. 11(a, b) documents the gain obtained by a compensation of global motion. The error in the picture background has been reduced drastically. Furthermore the error signal area around the locally moving arm has grown smaller. Remaining errors at the borders of the frame results from the fact that due to the zoom-out new information enters the scene which cannot be predicted.

Replenishment is not necessary in those regions of Fig. 11 which are sufficiently compensated, i.e., where no local motion occurs. The decision on replenishment is done by the segmenter. A calculated segmenter mask is shown in Fig. 12. As blocks shown white have not been coded, the coding of the remaining blocks achieved a better quality.

Picture quality depends on the quantizer step size used for prediction error quantization. In the

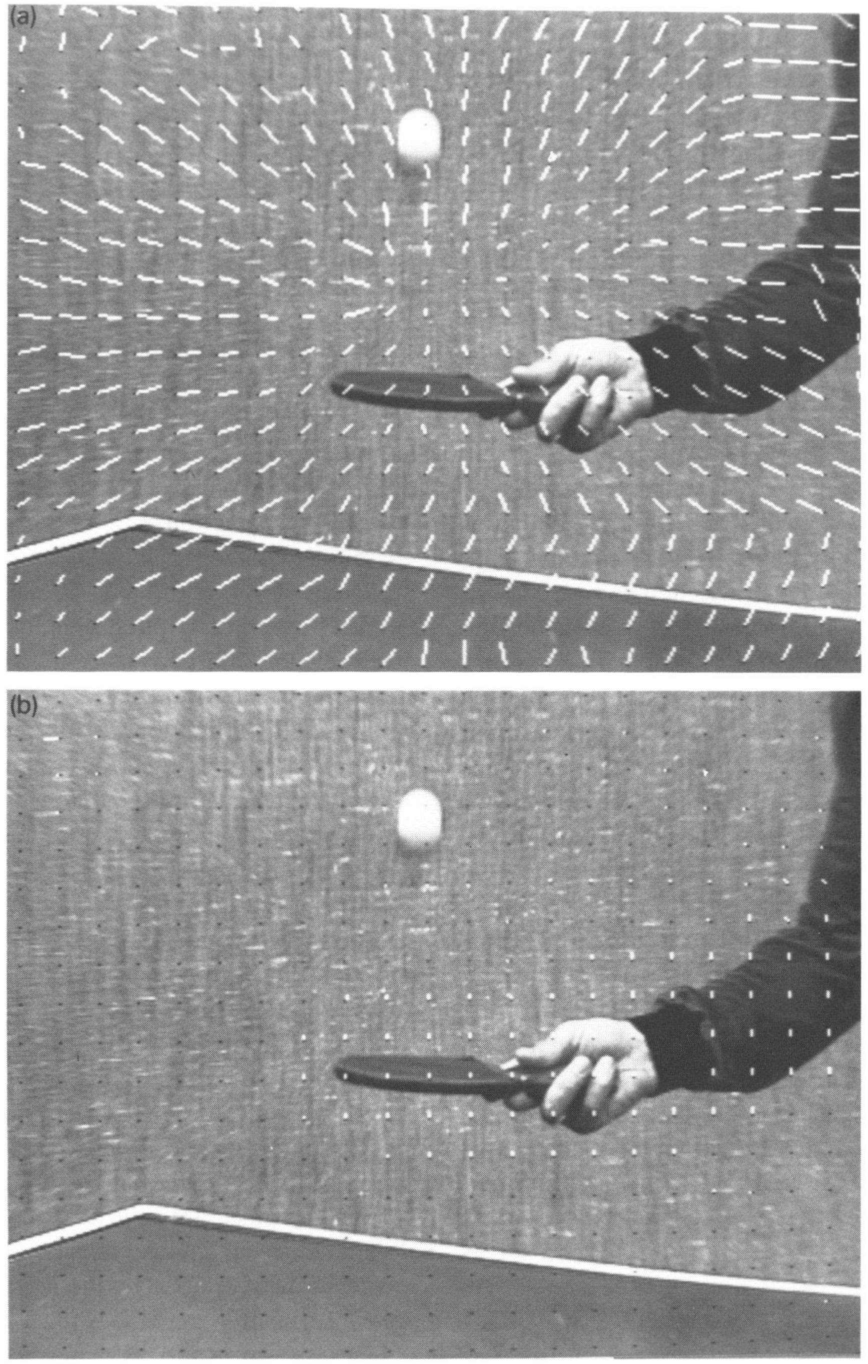

Fig. 10. Hierarchical displacement vector field estimated between frame 11 and (a) original frame 9, (b) original frame 9 after global motion compensation. Test sequence 'Table-tennis'; the vector field is underlaid by frame 11. Every 16th vector is depicted by a white line starting from a point marked black.

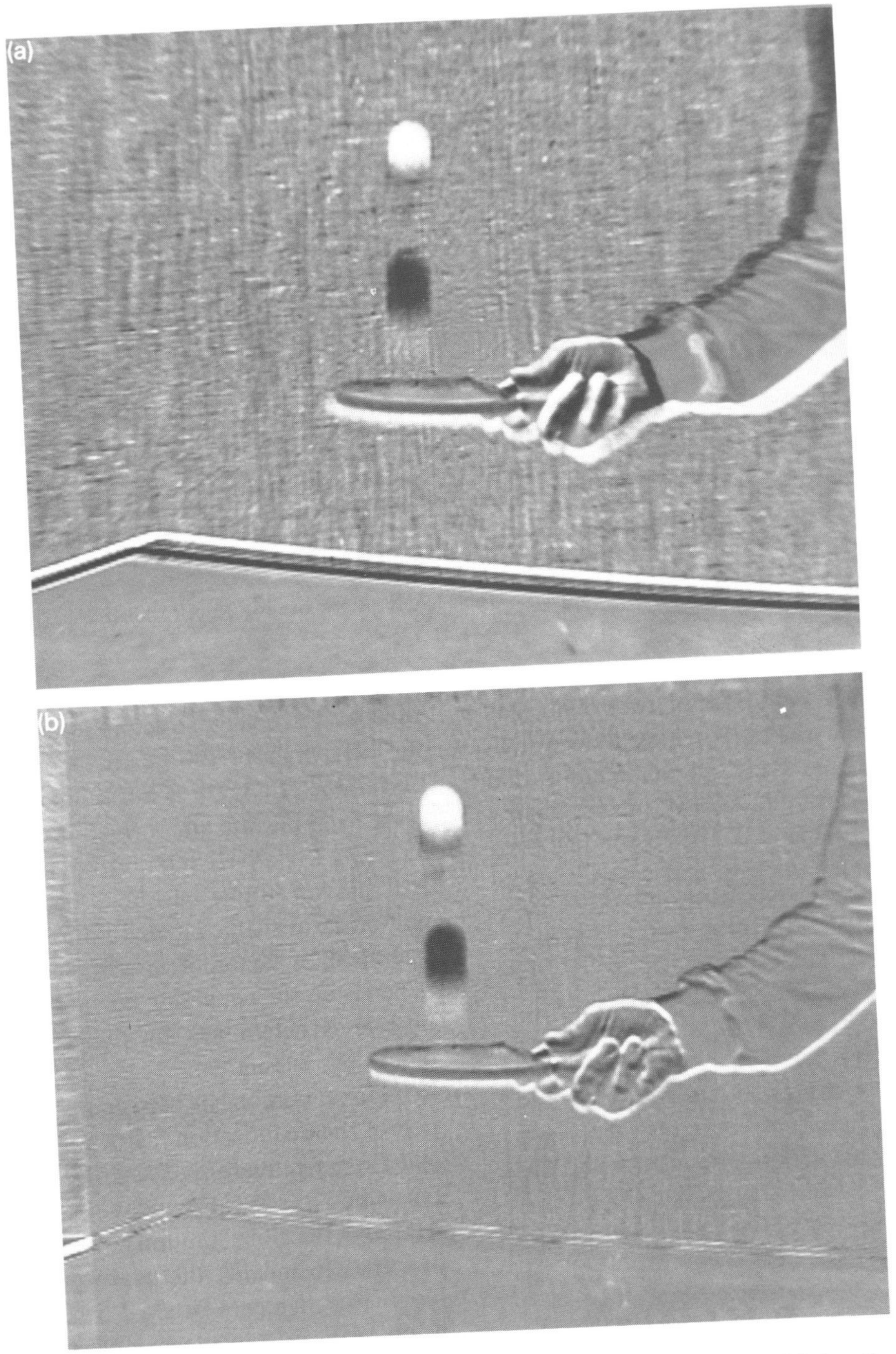

Fig. 11. Prediction error signal of frame 11 using (a) original frame 9, (b) original frame 9 after global motion compensation, for prediction; test sequence 'Table-tennis'.

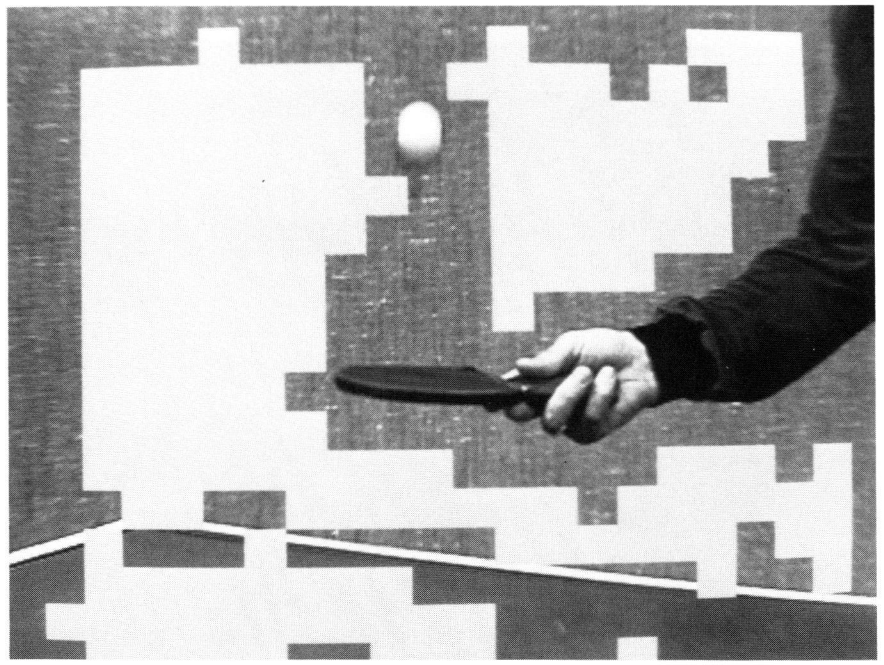

Fig. 12. Segmenter mask of frame 11 of test sequence 'Table-tennis'. Block size 16 × 16 pel; white blocks: no replishment.

JPEG coding algorithm exists a strong correlation between the integer scaling factor controlling step size of quantization and the resulting picture quality (see Section 2.3). Therefore this quantization factor can be used as a measure of picture quality and is depicted in Fig. 13. The factor has been measured in simulations with and without global motion compensation at the bit-rate of 1.15 Mbit/s. The coding structure can be recognized in the behaviour of the quantization factor. Every first frame of a GOF is intraframe coded with the same fixed bit-rate and a small quantizer step size to reach a relatively high intraframe picture quality. The intraframe bit-rate is determined by the half of the available bit amount for one GOF. The quantization factor of interframe coded pictures depends on the remaining bit-rate. The interframe bit-rate is determined by the number of coded blocks per picture.

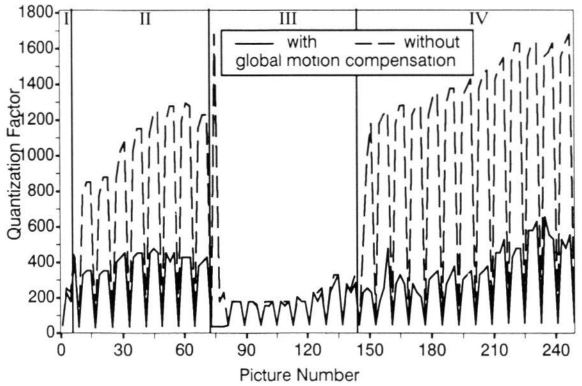

Fig. 13. Quantization factor of test sequence 'Table-tennis' measured in simulations with and without global motion compensation. I: scene without global motion; II: scene with central zoom out; III: scene without global motion; IV: scene with global pan.

If global motion has been compensated, a reduction of the quantization factor to 30% in scenes II and IV can be realized. Scene II contains a zoom-out, while in scene IV a global pan is dominant. Although the global motion has been compensated in scenes II and IV, the quantization factors are higher than in scenes I and III, where the camera is fixed. As the velocity of pan is growing during scene IV, the quantization factor is growing also. This can be explained by new regions entering into the image plane which also can be seen in Fig. 11(b).

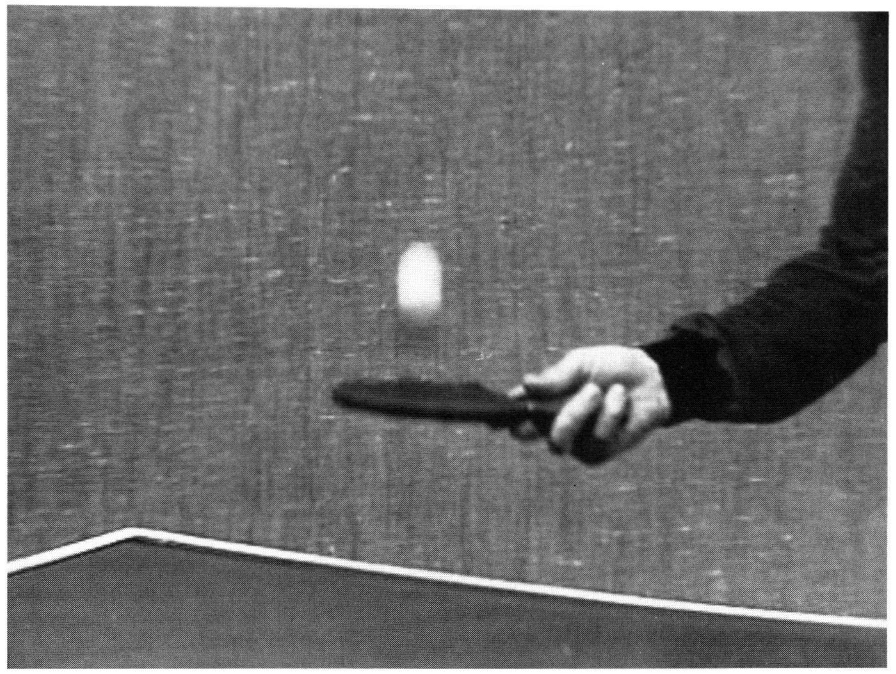

Fig. 14. Decoded frame 10 out of test sequence 'Table-tennis' generated at the decoder side by motion compensated extrapolation.

In Fig. 14 the performance and limits of extrapolation using global motion compensation are illustrated. The loss of sharpness in the picture is caused by the bilinear filter. The borders of the picture are generated from the following frame. Due to its local motion the arm is not synthesized well. The unbroken white table lines demonstrate that the global motion has been compensated accurately.

6. Conclusions

A coder for video signals working at a bit-rate of 1.15 Mbit/s has been presented. This coder uses a special coding structure facilitating trick modes such as random access with short decoding delay, search modes and high resolution of single still pictures.

The presented coder uses a separated global and local motion analysis and compensation. The global motion analysis and compensation is based upon a model of pure central zoom and pan of the camera. The estimation of global motion parameters is done by a linear regression on an estimated displacement vector field and improved by checking the prediction error signal with frame matching. This results in global motion parameters of high accuracy. Local motion estimation is done by a hierarchical displacement estimation resulting in an 8 × 8 pel vector field of integer accuracy.

The separation of regions described by global motion parameters only and of regions which need replenishment by additional local motion vectors and an error signal is provided by a segmentation algorithm. The algorithm uses change and motion information to generate the segmentation mask.

It was shown that the segmentation into regions of global and local motion has several advantages. The amount of coded motion information is reduced by coding only global motion parameters for large regions in the picture. A replenishment of regions sufficiently described by global motion is not necessary. That means that neither the local motion information nor the prediction error has to be transmitted in those regions. Vector fields describing local motion and the belonging prediction error only are coded in regions containing additional local motion. For the purpose of coding at the bit-rate of 1.15 Mbit/s experimental results show that the quantization step size for prediction error coding can be reduced by the factor three in case of existing global motion. This results in an improved picture quality.

The global motion parameters transmitted for every frame have also been used at the decoder side for a motion compensated extrapolation of omitted frames generating a higher resolution in time for global motion in the decoded sequences. The resolution in time for objects moving locally remained constant although this was masked by the global motion.

In very critical sequences global motion cannot be compensated by zoom and pan only. If additional to zoom and pan a translational camera motion is carried out, the central zoom and pan model fails although—and that has to be emphasized—the global motion parameter does not deteriorate picture quality. In further works a global motion model incorporating translational camera motion should be investigated. Using a global motion model of central zoom, pan and camera translation arbitrary camera actions can be described. First promising solutions in this field approximating arbitrary camera actions are given by models using more than three parameters as proposed in [5].

Acknowledgment

This work has been supported by Telenorma Frankfurt within the COMIS project of ESPRIT II. The authors want to thank Frank Sperling for implementing and tuning the global motion estimation algorithm. Furthermore the authors would like to express their gratitude to Dr. Kummerow and Professor Musmann for encouraging this work and for their suggestions and helpful discussions.

References

[1] M. Bierling, "Displacement estimation by hierarchical blockmatching", *3rd SPIE Symposium on Visual Communications and Image Processing*, Cambridge, USA, November 1988.

[2] R.J. Clarke, *Transformcoding*, University of Technology, Loughborough, United Kingdom, Academic Press, New York, 1985.

[3] J.D. Eggerton and M.D. Srinath, "A visually weighted quantization scheme for image bandwidth compression at low data rates", *IEEE Trans. Commun.*, Vol. COM-34, No. 8, August 1986.

[4] M. Hötter, "Differential estimation of the global motion parameters zoom and pan", *Signal Processing*, Vol. 16, No. 3, March 1989, pp. 249–265.

[5] M. Hötter and R. Thoma, "Image segmentation based on object oriented mapping parameter estimation", *Signal Processing*, Vol. 15, No. 3, October 1988, pp. 315–334.

[6] G. Kummerfeld, F. May and W. Wolf, "Coding television signals at 320 and 64 kbits/s", *2nd International Technical Symposium on Optical and Electro-Optical Applied Science and Engineering*, SPIE Conf., Image Coding, Cannes, France, December 1985.

[7] S. Okubo, "Video codec standardization in CCITT Study Group XV, *Signal Processing: Image Communication*, Vol. 1, No. 1, June 1989, pp. 45–54.

[8] R. Thoma and M. Bierling, "Motion compensating interpolation considering covered and uncovered background", *Signal Processing: Image Communication*, Vol. 1, No. 2, October 1989, pp. 191–212.

[9] JPEG, "Working Paper for Draft Proposal", *Joint Photographic Expert Group*, ISO/IEC, JTC1/SC2/WG8, CCITT SGVIII, Document N264, February 1989.