THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 51 April 1972 Number 4

Conditional Vertical Subsampling– A Technique to Assist in the Coding of Television Signals

By R. F. W. PEASE

(Manuscript received July 23, 1971)

In an interlaced scan television system, the vertical sampling rate of an image can be halved by sampling every other field. Picture elements in the missing fields are replaced in the display by both temporal and vertical interpolation, but the resulting pictures show some visible defects. This paper describes how these defects can be eliminated at the extra cost of fully sampling in selected areas of the picture. For a typical Picturephone® *scene with active movement the selected areas make up about 6 percent of the picture elements in the unsampled field. The technique can be combined with a wide variety of interframe-coding techniques. In one particular example in which the television signal is specified as clusters of frame-to-frame differences, the cost of specifying "active" frames (14,000 significant frame differences per frame) is reduced from 68,000 bits to 42,500 bits. This corresponds to a reduction in bit rate from 2 Mbits sec^{-1} to 1.3 Mbits sec^{-1}.*

I. INTRODUCTION

The technique of reducing the horizontal-sampling frequency ("subsampling") in the moving parts of the television image has been pre-

viously described.[1] It was found that the frequency could be halved without visible degradation for most object speeds. For slow speeds the degradation was visible but not objectionable. Combining this technique with that of conditional replenishment[2] has proved particularly effective,[3] because in periods of fast movement, conditional replenishment by itself becomes uneconomic because so many picture elements (pels) change significantly from frame to frame.

An obvious question to ask is whether a similar advantage results from subsampling vertically.

In an interlaced television scan format, the idea of halving the vertical-sampling frequency can be confusing because we must distinguish between fields and frames. A diagram showing the vertical position of lines for successive fields is shown in Fig. 1. One method of halving the vertical-sampling frequency is to sample every second line in each field and to replace the unsampled lines by interpolating the values of vertically adjacent elements in the same field [e.g., an unsampled line with coordinates $y = 2$ and $t = 2$ is replaced by an average of lines with y, t coordinates (0, 2) and (4, 2)]. This has been tried and the degradation is subjectively objectionable.

The second method is to sample alternate fields so that in stationary pictures the vertical-sampling frequency is halved. The unsampled fields are replaced by an average of the four nearest neighbors in Fig. 1 (e.g., an unsampled line with coordinates $y = 2$ and $t = 2$ is replaced by an average of lines with y, t coordinates 1 , 1; 1 , 3; 3 , 1; and 3 , 3). With this method the resulting pictures are subjectively satisfactory except for fast moving contrasty edges which appear blurred and somewhat jerky, and in dark areas with little horizontal but much vertical detail;[4] in these areas an aliasing pattern is sometimes visible.

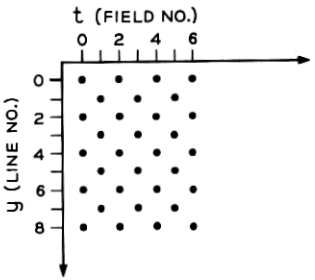

Fig. 1—Temporal position (or Field No., t) and vertical position (or Line No., y) of lines in a 2:1 interlace scan.

Henceforth, we will use the term "vertical subsampling" to refer to this second method.

If we vertically subsample over most of the moving regions except where there are visible defects, the entire picture should be subjectively satisfactory. We call this technique, conditional vertical subsampling (CVSS). The question is: How much extra channel capacity is needed to bring about the required improvement? This paper describes some experiments designed to answer this question.

Because the extra information is generated at an irregular rate, the use of conditional vertical subsampling requires a buffer memory and hence is probably most useful in conjunction with a buffered coder. In this paper we have in mind the eventual use of conditional vertical subsampling in a buffered interframe coder similar to those described in Refs. 2 and 3.

II. EXPERIMENTAL ARRANGEMENT

The basic apparatus is shown in Fig. 2. The output of the television camera (a silicon diode array vidicon) is sampled at 2.02 MHz and is digitized to 8-bit accuracy. The scan format is similar to that used in the *Picturephone* service in that there are 271 lines per frame and 60 fields per second. Each frame is made up of two interlaced fields.

Consider first only the odd-numbered fields. The coded version of a previous odd field appears at the output of field delay 1 and is compared, picture element by picture element, with the digitized input to determine whether the stored value is an adequate representation of the current value. For any pel, if the difference between the input signal and the stored signal exceeds a value ($T1$) of four levels out of 255, then $t_1 = 1$ and S1 is switched to the "one" position so that the value of the input replaces the previous value in the field delay. This new value is circulated twice in the field delay (because S1 is held at 0 during even fields) and after exactly one frame time is compared with the new input signal. Usually, in the absence of movement, the comparison is good enough so that no further updating need take place (i.e., no new information need be sent to a hypothetical receiver). Thus, each odd field is conditionally replenished as described in Ref. 2.

When the replenished version of line 3 , 3 (i.e., line No. 3, field No. 3) emerges first from switch S1, the replenished version of line 1 , 3 is appearing at the output of the line delay. Similarly, the replenished version of line 3 , 1 is appearing at the output of field delay 1 and the line 1 , 1 is appearing at the output of the second half-line delay. Thus,

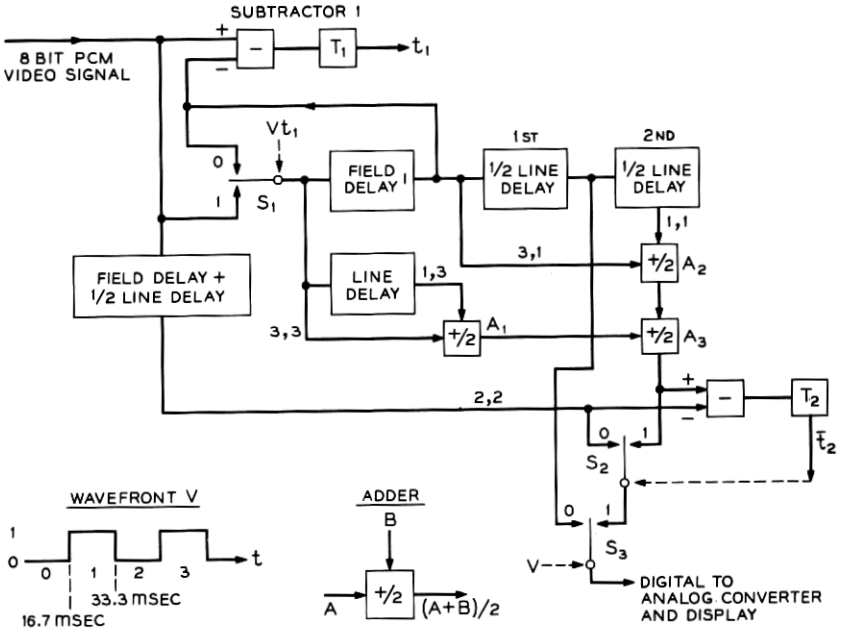

Fig. 2—Basic diagram of apparatus used to evaluate conditional vertical subsampling. Odd fields are coded by conditional (frame) replenishment using the loop formed by switch S1 and field delay 1; the output for these fields appears at 0 on switch S3. The output for even fields appears at 1 on switch S3 and is derived either from the vertically subsampled signal at the output of adder A3 or from the original output delayed by 1 field and 1/2 line.

lines 3 , 3; 1 , 3; 3 , 1; and 1 , 1 are simultaneously available and so, using adders A1, A2, and A3, we can form the spatially and temporally interpolated value, which as described in the previous section, is displayed instead of line 2, 2 when there is subsampling.

To maintain the correct temporal and spatial sequence of displayed fields, the replenished version of line 3 , 3 is not displayed until one field time plus one-half line time after first appearing at switch S1. Thus, line 3, 3 and other lines in odd fields are taken from the output of the first one-half line delay, and if we wish to vertically subsample continuously, the output is taken from adder A3 for even fields (i.e., S2 is kept in the 1 position and S3 is switched by waveform V).

However, we want to investigate how much the picture quality is improved by replacing the interpolated value with the original signal in selected parts of the even fields. To do this in the proper time sequence, the original signal is delayed by one field time plus one-half line time

and then compared with the interpolated value in subtractor 2. When the magnitude of the difference equals or exceeds the threshold $T2$, then $t_2 = 1$ and the original is displayed. Otherwise the interpolated value is displayed. In a practical coder the position and amplitude of those parts in the original signal used to replace the interpolated value must be coded and transmitted to the receiver. It is the cost of transmitting this extra information which must be balanced against the resulting improvement in picture quality.

The output of switch S3, which corresponds to the output of the hypothetical receiver, is converted to an analog signal and displayed on a television monitor with a 5-1/2 inch by 5-inch raster and a polarizing faceplate. The picture was viewed at 36 inches in a room with average illumination for an office (70 foot-candles).

Three series of experiments were carried out. In the first series, the subjective effect of varying the threshold $T2$ was investigated. In the second series, we measured the frequency of occurrence of picture elements for which the difference between the interpolated value and the delayed signal equals or exceeds $T2$ (henceforth we shall refer to such events as VSS differences). In the third series, we investigated the subjective and numerical effects of grouping the VSS differences into clusters so that a separate address word need not be used for each VSS difference. For most experiments requiring numerical results, the scene was the swinging model head shown in Fig. 3. The period of the swing was 2.7 seconds and the amplitude corresponded to 44 picture elements. The maximum speed of the head corresponded to 3-1/2 pels per frame interval. From time to time, some experiments were also carried out with live subjects or with very contrasty material.

III. RESULTS AND DISCUSSIONS

3.1 *Subjective Effects of Varying the Threshold T2*

The results of the two extremes is already known; for $T2 = 1$, we have the original picture and for $T2 = 255$, we have the vertically subsampled picture with the aforementioned defects. With a threshold of $T2 = 4$ (out of 255 levels), close scrutiny (15 inches) of the hair of the model when stationary revealed a just detectable difference from the original only if the original was compared instantaneously; otherwise there was no difference between the two pictures. With $T2 = 8$, there was a slight loss in vertical detail in the hair of the model which again was only noticeable at the normal viewing distance by switching instantaneously to the original. In some areas containing a lot of dark

Fig. 3—Swinging model head scene. The head is swung with a peak amplitude shown by the distance between adjacent vertical pencils.

vertical detail but little horizontal detail, the low contrast aliasing pattern could be made out. For $T2 = 12$, there was a more visible loss of vertical detail of stationary pictures and fast movement (4 pels per frame interval and above) of contrasty edges (64 levels per pel) showed some smearing or occasional raggedness. For $T2 = 16$, the above effects were more pronounced. As the threshold was increased beyond 16, the above effects became progressively more serious. For $T2 = 32$, the picture quality was virtually the same as for $T = 255$.

3.2 *Frequency of Vertical Subsampling Differences For Even Fields*

In even fields we counted the number of VSS differences for various values of $T2$. The numbers were recorded on a digital printer with a sampling frequency of about 5 Hz; the scene was the swinging model head. Simultaneous counts were also made of the frame differences exceeding a threshold of 4 (out of 255 levels) occurring in the odd fields. This count served as a check for the repeatability of the scene and also gave a measure of the amount of activity in the scene.

Some of the results are shown in Fig. 4. The dotted lines indicate the frequency of frame differences and show the characteristic periodic

pattern as the head swings. The variability in the depth of the trough is probably due to the relatively low sampling frequency. The maximum activity generates more than 7000 significant frame differences in the odd field which, in comparison with other *Picturephone* scenes is considered "active."[3] The corresponding curves for the vertical subsampling differences show the same form of activity but are much less variable. For instance, for $T2 = 8$ the variation is from 1100 per field at the end of the swing to 1900 per field at the bottom of the swing. This is understandable as those vertical subsampling differences which arise from the spatial interpolation tend to remain unchanged as the movement in the picture decreases.

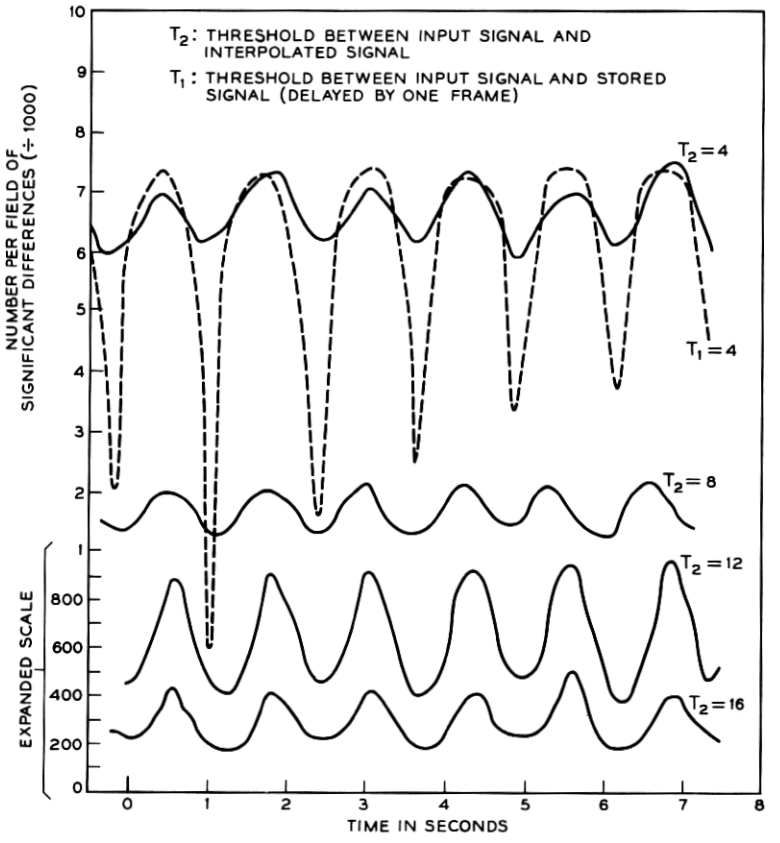

Fig. 4—Frequency of significant differences as a function of the scene (swinging model head).

Note the big decrease in counts of VSS differences as the threshold increases. For $T2 = 8$, which for a scene with movement gives a picture virtually indistinguishable from the fully sampled picture, there are only about 2000 significant VSS differences compared with more than 7000 significant frame differences. For $T2 = 16$, which still gives a substantial improvement in picture quality over that with $T2 = 255$, the frequency of significant VSS differences is less than 500 which is negligible compared with the frequency of significant frame differences.

If the threshold $T1$ (defining a significant frame difference) is raised to 8, then the picture quality is visibly degraded due to the "dirty window effect."[2] One other effect is to reduce the frequency of significant frame differences from about 7000 per field to about 5000 per field when the model head is at the bottom of the swing (Fig. 5). A third effect of increasing $T1$ from 4 to 8 is to increase the counts of significant VSS differences when $T2 = 8$ from about 2000 to 2600 per field. This increase can be seen by comparing the relevant graphs in Figs. 4 and 5. The increase is not noticeable for values of $T2$ greater than 10. This third effect can be easily explained by the increased difference between the respective odd fields and the original signal being added to the difference introduced by interpolation. Note however, that as both thresholds increase from 4 to 8, the counts of VSS differences decreased more than the corresponding counts of significant frame differences in spite of the fact that the subjective degradation of increasing $T1$ is more severe than that due to increasing $T2$.

3.3 *Cost of Transmission*

One way to transmit the extra information would be to use 8 bits to specify the horizontal position and another 8 bits to specify the revised amplitudes of each picture element showing a significant VSS difference. However, this is probably wasteful because: (*i*) the significant VSS differences tend to occur in clusters so that usually one address word can serve several VSS differences; the cost of this method is that either the length of the cluster or the end of the cluster must also be sent to the receiver. (*ii*) 8 bits are probably unnecessary for the amplitude information as the extra information can probably be adequately specified as a relatively coarsely quantized difference signal.

3.3.1 *Effect of Cluster Coding*

Direct observation of the spatial distribution of significant VSS differences (Fig. 6) shows that they, like frame differences, tend to occur in clusters and so cluster coding is probably advantageous.

To accentuate this clustering and to reduce further the number of bits required for addressing, it is usually advantageous to run two closely separated clusters into one larger cluster.[3] This is particularly important because the VSS differences due to temporal interpolation near a horizontally moving edge occur as two separate clusters (with differences of opposite sign) on either side of the edge but at the edge itself the differences are usually less than significant. This effect can be seen in Fig. 6 around the edges of the cheek of the model head; the reader may wish to verify exactly how this comes about by drawing out the video signal corresponding to a moving step on a target of a storage type television camera for three successive fields or by referring to the Appendix.

It also follows that isolated significant VSS differences are relatively

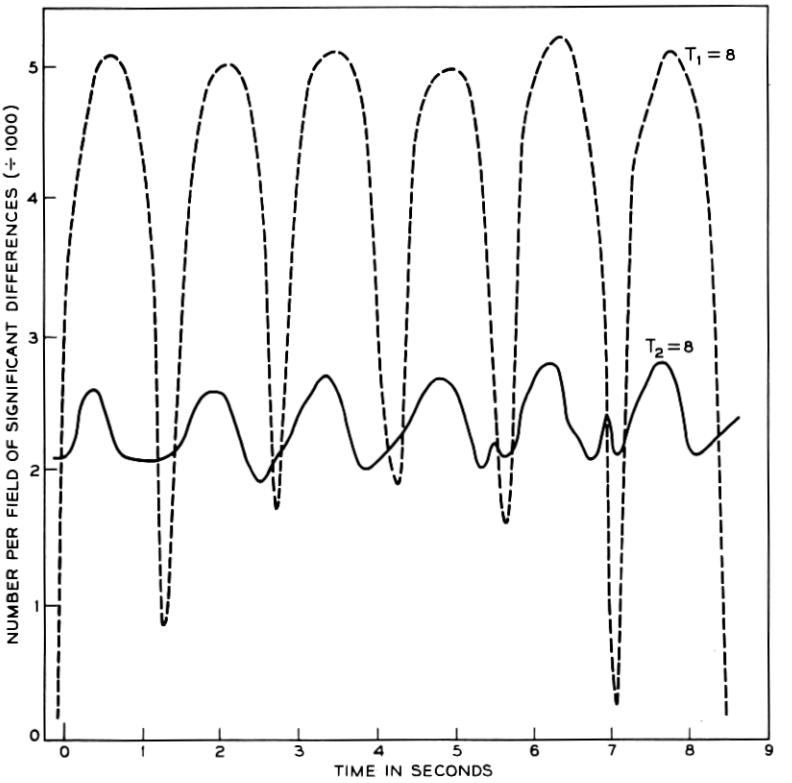

Fig. 5—Frequency of significant differences as a function of time. Scene is a swinging model head.

Fig. 6—Flags representing VSS differences for $T2 = 8$ while the head is swinging.

expensive to transmit, and they should not be transmitted if such rejection does not cause subjective degradation of the picture.[3] We investigated this effect by leaving uncorrected single VSS differences with no neighboring VSS differences within 2 pels horizontally. The subjective effect was negligible except for values of $T2$ of 12 and above, when occasionally the loss in vertical detail could be made out or a moving edge appeared somewhat more ragged. To control switch S2 in Fig. 2, the subtractor and threshold circuit were followed by logic to reject isolated VSS errors and to run together clusters of VSS differences separated by one or two pels. We also rejected VSS differences in the first and last lines of even fields because vertical interpolation here will be very frequent but rejecting these values is subjectively negligible. With these restrictions, we measured the frequency of clusters and of pels contained in clusters. Some of the results are summarized in Table I and indicate that the average length of clusters remained at about 5 irrespective of picture motion.

For comparison, equivalent data are shown for frame differences (using the same logic in recording the counts) and show that when the

scene is active the cluster length is longer for frame differences. These data correspond very well with equivalent data reported for scenes with live subjects.[3]

3.3.2 *The Effect of Quantizing the Signal*

Amplitude information in television signals is usually most economically specified by quantizing a prediction error. For example, the amplitude of the previously scanned pel is often used as a prediction of the current pel; the difference is then quantized and coded for transmission. This technique could be used here but we decided instead to quantize the prediction error generated by the interpolated value and the input. Such errors tend to be smaller than either element-to-element or frame-to-frame differences. In the Appendix, we show how this comes about in one particular case.

In Fig. 7 we have replotted some of the data of Fig. 4 and some extra data to show the frequency distribution of amplitudes of VSS differences for the swinging head. The curve is more peaked than comparable curves of element differences or frame differences and shows very few differences of amplitude greater then 10 percent of the peak amplitude. Thus quantizing the VSS difference directly is one attractive approach and should cost no more than 3 or 4 bits per pel if fixed length codes are used and perhaps much less if variable lengths are used.

As a first step, we divided the magnitude of the difference amplitude scale up into seven classes as shown in Table II and then assigned weights as shown to each class. The pictures that resulted from correcting the VSS errors with these quantized signals showed little difference from those in which the VSS errors are corrected with an 8-bit signal when the scene of the swinging head and the threshold 8 was used. Thus, a fixed code length of 3 bits would be sufficient to describe the amplitude information for each pel and still allow one word to denote the end of a cluster. However, when the scene contained a very

TABLE I—FREQUENCY OF FRAME DIFFERENCES, VSS DIFFERENCES, AND CLUSTERS FOR DIFFERENT THRESHOLDS FOR THE SWINGING HEAD SCENE (FIG. 2)

	End of Swing				Bottom of Swing				
Type of Difference	Frame	VSS	VSS	VSS	Frame	VSS	VSS	VSS	
Threshold	$T1 = 4$	8	12	16	4	8	12	16	
No. of Differences	123	750	350	230	7000	1400	650	400	Per Field
No. of Clusters	47	125	60	35	440	260	140	80	Per Field

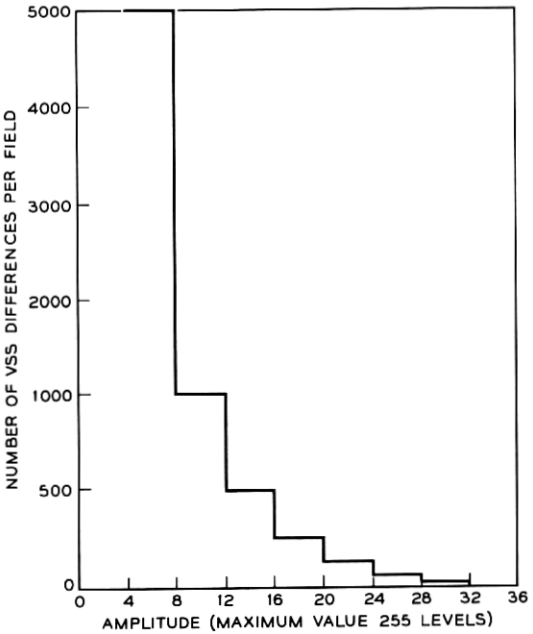

Fig. 7—Histogram of number of VSS differences as a function of amplitude values of 4 and above.

contrasty (50 percent of peak amplitude) edge moving at 4 pels per frame interval or faster, there was visible degradation of the edge. Therefore, a 15-level quantizer with the decision levels and weights shown in Table III was used so that a 4-bit vocabulary would suffice for the amplitude information and one word would be left to denote the end of a cluster. Resulting pictures were indistinguishable from those using an 8-bit correction signal.

If for each cluster we allow 4 bits per pel, 8 bits for the horizontal address, and 4 bits to denote the end of run, then we can plot the data which was summarized in Table I to show the total number of bits per field (excluding synchronizing bits) needed to specify the extra

TABLE II—7-LEVEL QUANTIZING SCALE

Decision Level	0	±8	±15	±25	out of 255
Representative Weight	0	±10	±19	±28	out of 255

TABLE III—15-LEVEL QUANTIZING SCALE

Decision Level	0	±3	±6	±10	±15	±23	±33	±44	} out of 255
Representative Weight	0	±4	±8	±12	±18	±28	±38	±50	} out of 255

information for correcting significant VSS differences (Fig. 8). For $T2 = 8$, which gives very satisfactory correction of the VSS differences, and for a frame with 14,000 significant frame differences, the even field can be sent at a cost of only 8500 bits. For $T2 = 12$ the cost comes down to 4500 bits and for $T2 = 16$ there is a further drop to 1600 bits.

For comparison, the cost, allowing 4 bits per frame difference and 12 bits per cluster of transmitting the frame differences, is also shown.

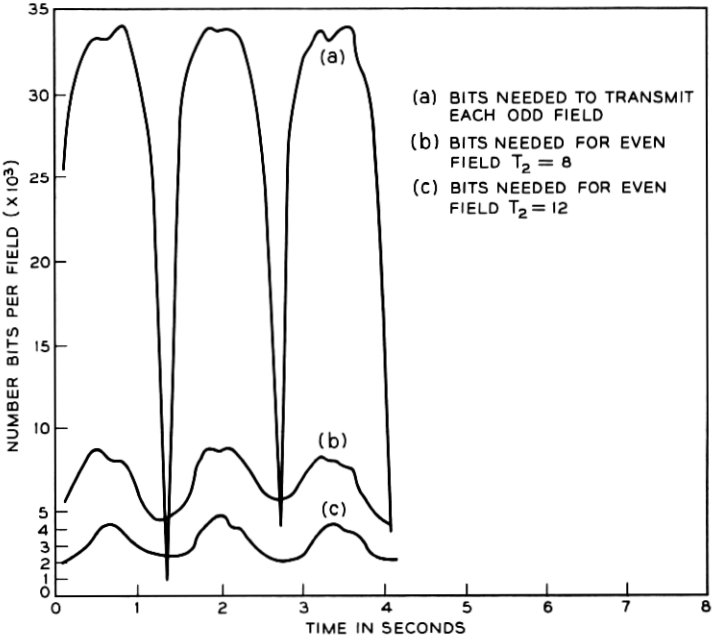

Fig. 8—Number of bits required to transmit each odd field by conditional replenishment and each even field by conditional vertical subsampling. For two values of $T2$. Scene: Swinging model head. In each field the pels requiring transmission are grouped in clusters by first rejecting isolated changes and then running together clusters separated by 1 or 2 pels. Twelve bits are assigned to each cluster and 4 bits to each pel containing a cluster. Synchronizing bits have been excluded.

Thus even though the frame differences occur in longer clusters in active frames, the cost of transmitting the odd field is much greater than the cost of transmitting the even field.

IV. CONCLUSIONS

For the scene used we can now answer the original question of comparing the advantages of horizontal subsampling and vertical subsampling by comparing the total cost of transmitting one frame (excluding synchronizing bits and "forced replenishment"[2,3] bits). For vertical subsampling we need, for an active frame, 34,000 bits for the odd field and 8,500 bits in the even to give a subjectively pleasing picture; this gives a total of 42,500 bits. When the moving parts of the picture are horizontally subsampled, the number of bits used to specify amplitude is halved but the number required for addressing is unchanged. For two fields, therefore, there are 28,000 bits for the amplitude information and 12,000 bits for the address information to give a total of 40,000 bits. Thus, the numerical advantage of the two techniques is very similar.

It should be pointed out that in the experiments described the vertical subsampling was applied to both the moving and stationary part of the picture. With a buffered interframe coder, the main problem is created by the moving parts of the picture. If the vertical subsampling is applied only to the moving part then the number of bits needed during the even fields is reduced still further. However, as can be seen from Fig. 6, there are relatively few VSS differences in the background and so in this case the savings would be marginal. The maximum savings so gained can be estimated by assuming that the only VSS differences that need correcting are those due to movement. The number of such differences can be estimated by assuming that they account for the difference between the number of VSS differences occurring for stationary scenes and for moving scenes. Thus in our experiment and for $T2 = 8$, the number of bits in the even field is approximately halved, and the total number of bits required to code the frame is reduced from 42,500 to 38,000.

One advantage of the technique of conditional vertical subsampling is that it can be combined with a wide variety of interframe-coding techniques, because the odd fields are not affected. One exception is that it is no longer straightforward to express amplitude information in the odd fields as field-to-field differences.[5] Otherwise the amplitude information in the odd fields can be expressed as element-to-element differences,[6] frame differences,[3] or even two-dimensional spatial differences.[7] How well the techniques of horizontal and vertical sub-

sampling can be simultaneously applied remains to be investigated. Horizontal subsampling in the odd field will certainly introduce extra VSS differences, but the numerical effect of these can probably be minimized by horizontally subsampling in the even fields as well.

V. ACKNOWLEDGMENT

I would like to acknowledge the many stimulating discussions with J. C. Candy, D. J. Connor, C. C. Cutler, L. H. Enloe, B. G. Haskell, J. O. Limb, and F. W. Mounts. The technical assistance of W. G. Scholes is also greatly appreciated.

APPENDIX

Vertical Subsampling Differences Arising From the Horizontal Movement of a Vertical Edge

Consider a vertical edge, consisting of a black to white transition of 128 levels, being moved horizontally across the field of view of a television camera at a speed of p pels per frame interval when referred to the target of the camera. Assume that the video-signal level for each pel is a direct measure of the quantity of light which has fallen on the appropriate area of the target for 1/30 second just prior to being scanned. If the target is stationary, then a plot of signal level as a function of horizontal position is shown by the line ABCD in Fig. 9. If immediately after pel 0 is scanned, the edge moves horizontally to the right at 8 pels per frame interval, then in the next frame the signal level will be a measure of the amount of time during the intervening 1/30th of a second that each point on the target received light from the bright side of the edge. For uniform motion therefore, the line ABED represents the signal level in the second frame. In the third frame the signal level is represented by the line AHD which is simply a translation of ABED by 8 pels. The signal level for the adjacent lines in the interlaced field is shown by the line AFGD and is simply a translation of ABED by 4 pels (there being no vertical difference). When vertically subsampling however, we substitute for the interlaced field, an interpolated signal ABD and the VSS differences are represented by the vertical distances between lines BD and BFGD. Note that at the halfway point this difference is zero and on either side of the edge we have relatively large differences with a maximum value of 32 levels. This is in agreement with the "flags" shown in Fig. 6, which occur in pairs of clusters straddling the edge of the moving cheeks.

We can also use Fig. 9 to compare certain differences in level which

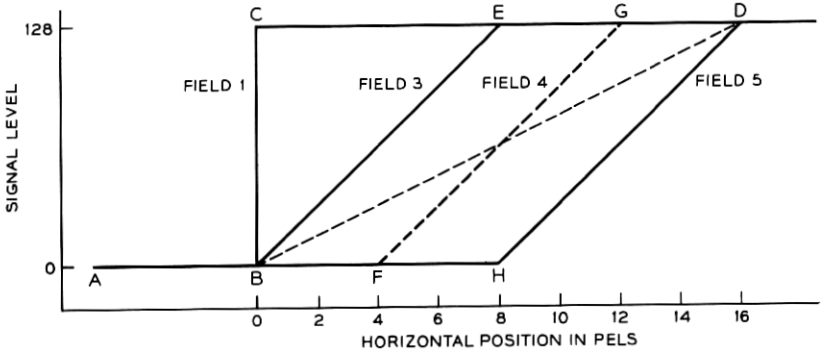

Fig. 9—Signal levels as a function of horizontal position for different fields.

may be quantized or coded for transmission. First of all, element-to-element differences: For a stationary position we have a maximum step height of 128 levels which decreases as the speed increases.

Frame-to-frame differences for a steadily moving edge are represented by the vertical distances between lines ABED and lines ABHD. The differences here have values up to 128 levels (a more accurate analysis shows that this level is only approached for values of $p \gg 1$).

Field-to-field differences are represented by the vertical distance between ABED and AFGD and can take values up to 64 levels.

Thus, VSS differences have, for the model, the smallest maximum value and would seem to be most economical to quantize and code.

REFERENCES

1. Limb, J. O., and Pease, R. F. W., "Exchange of Spatial and Temporal Resolution in Television," B.S.T.J., *50*, No. 1 (January 1971), pp. 191–201.
2. Mounts, F. W., "Video Encoding System with Conditional Picture Element Replenishment," B.S.T.J., *48*, No. 7 (September 1969), pp. 2545–2553.
3. Candy, J. C., Franke, M. A., Haskell, B. G., and Mounts, F. W., "Transmitting Television as Clusters of Frame-to-Frame Differences," B.S.T.J., *50*, No. 6 (July/August 1971), pp. 1889–1917.
4. Limb, J. O., and Pease, R. F. W., "A Simple Interframe Coder for Video Telephony," B.S.T.J., *50*, No. 1 (July/August 1971), pp. 1877–1888.
5. Pease, R. F. W., and Scholes, W. G., "Field Difference Quantization," unpublished work.
6. Limb, J. O., and Mounts, F. W., "Digital Differential Quantizer for Television," B.S.T.J., *48*, No. 7 (September 1969), pp. 2583–2599.
7. Connor, D. J., Pease, R. F. W., and Scholes, W. G., "Television Coding Using Two-Dimensional Spatial Prediction," B.S.T.J., *50*, No. 3 (March 1971), pp. 1049–1061.