JOURNAL OF VISUAL COMMUNICATION AND IMAGE REPRESENTATION
Vol. 4, No. 1, March, pp. 79–91, 1993

Experimentally Obtained Thresholds for a Conditional-Replenishment Image-Sequence Coder

THOMAS V. PAPATHOMAS[1] AND DAVID MALAH[2]

AT&T Bell Laboratories, Murray Hill, New Jersey 07974

Received March 15, 1991

We present the results of several psychophysical experiments that we conducted to determine appropriate values for threshold parameters to be used for block classification in an image-sequence conditional-replenishment encoding algorithm. The algorithm partitions the image into square blocks of pixels and, following motion compensation, classifies them into two categories, "nonsmooth" and "nontextured" ones. Nontextured blocks are tested with criteria different from those used for testing nonsmooth blocks for deciding whether to replenish or to copy the block. Since the end product of such algorithms is judged by human observers, the objective of the experiments that we conducted was to obtain pertinent characteristics of the human visual system (HVS) for incorporating them into the threshold selection part of the encoding process. In three separate major experiments we obtained near-optimal values for three threshold parameters by testing the detectability of stationary as well as moving targets under various conditions. In applying our observations to conditional replenishment coding, target detectability will refer to the visibility of motion-compensated prediction error and a consequent need to replenish the corresponding part of the image. The three specific experiments involved (a) uniform targets against a uniform background of different intensity, (b) textured targets against a uniform background of the same intensity as the average target intensity, and (c) textured targets against a textured background with the same standard deviation but different average intensities. The experiments reveal how to adjust the values of the thresholds to suit local conditions and they give rise to an adaptive threshold modification technique, based on the HVS characteristics. This approach achieved significant reductions in the bit rate needed to encode image sequences, without affecting the perceived image quality.

1. INTRODUCTION

There is currently a great deal of activity in the standardization of transmission rates and algorithms for encoded image sequences [1–3]. The activity of the CCITT (International Telegraph and Telephone Consultative Committee) is directed to subprimary rates at multiples of 384 kbps (kilobits per second) and 64 kbps for videoconferencing and videophone applications at reduced spatiotemporal resolution; encoding/decoding must be performed at video rates in this case [1]. On the other hand, the ISO (International Standards Organization) considers the storage and retrieval of more complex "moving" images on digital storage media (such as CD-ROM) at approximately 1 Mbps [2, 3]. The experiments conducted in this work were motivated by the encoding algorithm developed in [4], which is a modified version of the baseline CCITT Reference Model (RM) coder [1], aimed at matching it to the ISO specifications of rate and type of moving images [3].

A block diagram of a generic motion-compensated conditional-replenishment coder is shown in Fig. 1, simplified from [1] by not showing the switching between "inter"- and "intra"-coding, which could be controlled by the block-classification box, as shown in the figure. For each incoming block in the current frame, an attempt is made to determine whether this block's position has changed from the previous (reconstructed) frame; if so, the displacement between the corresponding blocks in the two frames is estimated and its value $D = (D_x, D_y)$ is to be transmitted to the decoder as side information. Subsequently, some measure of the difference between the block in the current frame and the motion-compensated block in the previous frame is computed and, based on this comparison, the current block is classified into one of two major categories: either it matches well the corresponding block in the previous frame or it does not; in the former case the block is "copied" from the previous (reconstructed) frame, utilizing D, whereas in the latter case the block is encoded using discrete-cosine transform (DCT) coefficients, which are quantized (Q) and transmitted as output. The classification code C assigned to the current block is also transmitted as side information. The blocks marked Q^{-1} and IDCT perform the inverse operations of the quantizer Q and the DCT, respectively, and they are used to reconstruct the pixel values of a

[1] Present address: Department of Biomedical Engineering and the Laboratory of Vision Research, Rutgers University, New Brunswick, NJ 08903.

[2] Present address: Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel. This work was performed at AT&T Bell Laboratories while on sabbatical leave at the Department of Signal Processing Research.

1047-3203/93 $5.00

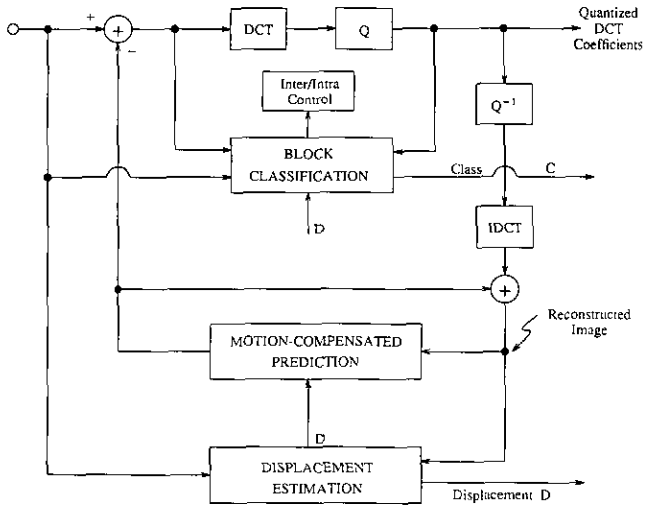

FIG. 1. Block diagram of the generic motion-compensated conditional-replenishment coder. DCT and IDCT refer to the discrete cosine transform and its inverse, respectively. Q is the quantizer and D is the estimated displacement vector of the block under consideration.

block that has already been encoded and quantized. Section 2 is devoted to a brief outline of the particular conditional-replenishment encoding algorithm that we worked with. In Section 3 we discuss the automatic-gain control (AGC) characteristics of television monitors that are relevant for the analysis of the experimental results. The three major groups of psychophysical experiments and their results are presented in Section 4. Finally, in Section 5, we discuss the importance of the experimental data and how they can be applied to the encoding process.

2. THE BLOCK-CLASSIFICATION ALGORITHM

The decision tree, which is part of the block-classification stage of the encoder developed in [4], is shown in Fig. 2a. In this section, a brief explanation of this decision tree is given. Our discussion clarifies that, in order to improve the encoding efficiency of the algorithm, the setting of the threshold values in the decision tree must be based on the characteristics of the human visual system (HVS). We mention parenthetically that several researchers have incorporated properties of the HVS into their encoders [5, 11–14].

With reference to Fig. 1, assume that the displacement estimator indicates that the (reconstructed) block of the previous frame that best matches the current block **B** is displaced by $D = (D_x, D_y)$ pixels. The value of the displacement is chosen to minimize the mean absolute difference (MAD)

$$\text{MAD} \triangleq \frac{1}{|\mathbf{B}|} \sum_{i,j \in \mathbf{B}} |s_n(i, j) - r_{n-1}(i + D_x, j + D_y)|, \quad (1)$$

where $|\mathbf{B}|$ denotes the size (in pixels) of the block **B**, s_n is the current frame, and r_{n-1} is the previous reconstructed frame. We refer to the block of the previous reconstructed frame that best matches the current block as the motion-compensated (MC) block. Having obtained D, we now return to the flowchart of Fig. 2a.

The first decision box, labeled A, examines whether B'_{d_v}, defined in Eq. (2), which is a measure of the variance of the difference between the current block and the previous MC block, is below a threshold T_2:

$$B'_{d_v} \triangleq \frac{1}{|\mathbf{B}|} \sum_{i,j \in \mathbf{B}} |(s_n(i, j) - \bar{s}_n) - (r_{n-1}(i + D_x, j + D_y) - \bar{r}'_{n-1})| < T_2, \quad (2)$$

where **B** denotes the original block in the current input frame s_n, the subscript indices d and v in B'_{d_v} denote "block-difference" and "variance," respectively, and $\bar{s}_n$ and $\bar{r}'_{n-1}$ are the mean values of the current block and the

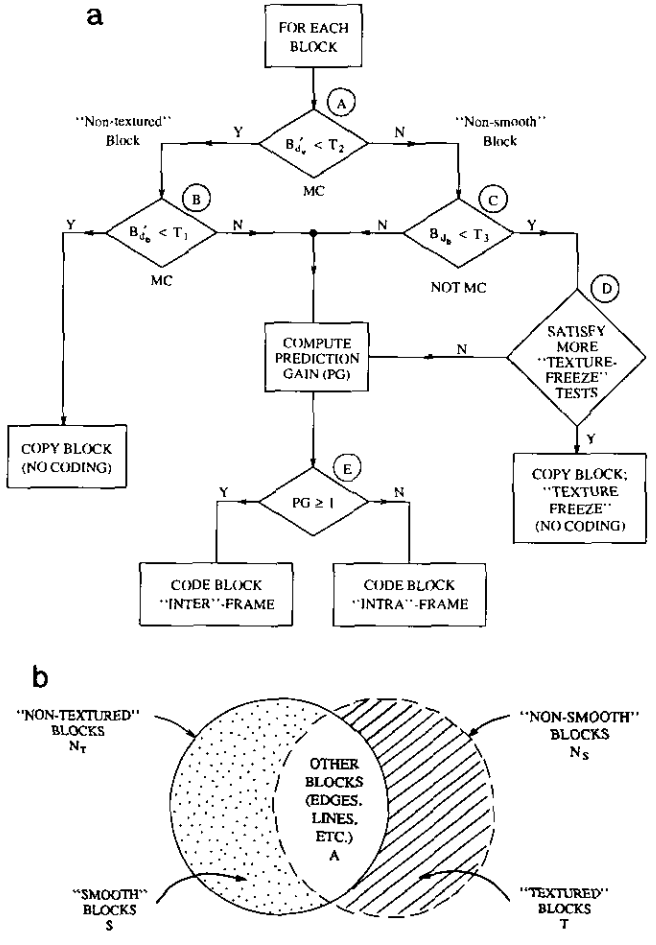

FIG. 2. (a) Simplified flowchart of the decision tree for the block classification stage of the algorithm. The symbols used are explained in Section 2; in particular, MC stands for "motion-compensated." (b) A Venn diagram illustrating the block classification.

previous MC reconstructed block, respectively; the prime in B'_{d_v} and $\bar{r}'_{n-1}$ is meant to denote that we are dealing with the *motion-compensated* block. If $B'_{d_v} < T_2$, the motion-compensated block difference is nearly uniform, which means that both corresponding blocks are most probably "nontextured." If not, it means that the motion-compensated block difference has a high variance and hence there is a high probability that at least one of the blocks is a "textured" block. The block is therefore classified at this stage as a "nonsmooth" block. For correct classification, the threshold T_2 in (2) must be determined experimentally since, in general, T_2 is a function of both image parameters and the HVS characteristics.

If the inequality in (2) is satisfied, a second test is applied on the block that has just been characterized as nontextured. We check whether the absolute difference between the mean intensities (brightness) of the current and MC reconstructed blocks is below a threshold T_1 (see decision box B in Fig. 2a):

$$B'_{d_b} \triangleq |\bar{s}_n - \bar{r}'_{n-1}| = \frac{1}{|\mathbf{B}|} \left| \sum_{i,j \in \mathbf{B}} (s_n(i,j) - r_{n-1}(i + D_x, j + D_y)) \right| < T_1. \quad (3)$$

If this test is passed, it means that the current block is almost indistinguishable from the corresponding one in the previous frame and the decision is made to "copy" it, rather than replenish (encode) it. If the test fails, the block must be encoded; a further test is conducted on the prediction gain, the outcome of which determines whether "intraframe" or "interframe" encoding will be used. As with T_2, experiments with human observers helped us determine how T_1 varies with the image parameters.

If a nonsmooth block is involved, i.e., the test in (2) was not passed, then we would like to check whether both corresponding blocks are textured blocks, typically belonging to the background, in which case the current block can be "frozen," i.e., not replenished. The first test (box C in Fig. 2a), which checks if the mean values of the corresponding blocks are sufficiently close, is similar to that of inequality (3), but without motion compensation (since the blocks are assumed to belong to the background), i.e.,

$$B_{d_b} \triangleq |\bar{s}_n - \bar{r}_{n-1}| = \frac{1}{|\mathbf{B}|} \left| \sum_{i,j \in \mathbf{B}} (s_n(i,j) - r_{n-1}(i,j)) \right| < T_3, \quad (4)$$

where the selection of the value of T_3 and its variation with image parameters is based on visual experiments. If the above condition is satisfied, a series of additional tests for freezing the texture block are conducted in tandem [4]. If they all pass, then the block is copied. However, if inequality (4) fails or if any of the tests for freezing the block fail in the above series, the decision is made to encode the block. The block is then subjected to the same prediction-gain test that was applied to the nontextured blocks that had to be replenished. The encoding is then performed based on data either from the current frame or from the previous frame, depending on the value of the prediction gain, and these details are covered in a companion paper [4].

The Venn diagram of Fig. 2b is meant to illustrate the block types that are identified by the classification process. A "smooth" block is characterized by a negligibly small variance in the intensity values of its pixels; i.e., all of its pixels have a nearly uniform intensity. A textured block, as the name implies, is one in which the pixel intensities vary widely around the average value and, at the same time, the spatial distribution of intensity within the block is irregular or random. Thus a nonsmooth block is not necessarily a textured block; for example, a block with a uniformly dark left half and a uniformly light right half contains a vertical edge and is neither smooth nor textured. Such a block belongs to the category "other" in Fig. 2b. A block with a fine texture that cannot be resolved by the HVS (thus appearing as uniform) is another example that will be classified in the category labeled other. We next follow the steps of the classification algorithm with reference to the diagram of Fig. 2b.

Decision box A of Fig. 2a partitions the universal set Ω of all blocks into two major sets: N_T (nontextured) and N_S (nonsmooth), enclosed by the solid and dashed circles of Fig. 2b, respectively. The blocks in N_T satisfy inequality (2); i.e., they pass the test of box A in Fig. 2a; the blocks in N_S do not. It should be clear that smooth blocks are only a subset of those which satisfy inequality (2). This is because the test is performed on the variance of the difference of the current block and the previous MC block; thus a block with an edge in it which is just translated from one frame to the next could be compensated by MC and still pass the test (an example of the class other in Fig. 2b). However, a textured block of high-intensity variance will certainly fail the test.

Again, it must be made clear that textured blocks are not the only class of blocks that will fail the test of decision box A of Fig. 2a. For example, a block with an edge in it, which has changed in the next frame so as to have the edge rotated, will also probably fail the test (another example of a block in the class other of Fig. 2b). However, a smooth block will certainly pass the test. In summary, it is for the above reasons that blocks which satisfy inequality (2), i.e., pass the test of decision block A in Fig. 2a, belong to the set N_T and those that do not pass the test belong to the set N_S.

By definition, $N_T \cup N_S = \Omega$, the set of all blocks. Also,

as we saw, the set $A \triangleq N_T \cap N_S$ is not null, but contains blocks with edges, lines, etc. Since $N_T \cup N_S = \Omega$, the sets of smooth blocks (S) and textured blocks (T) are $S = \overline{N}_S$ and $T = \overline{N}_T$, where $\overline{X}$ denotes the complement of a set X. Decision box B of Fig. 2a tests whether nontextured blocks can be copied or not. The purpose of decision boxes C and D is to determine whether the nonsmooth block is textured and whether "texture freeze" is possible; the latter requires that the corresponding blocks in both the current and the previous frames be textured and that they have similar characteristics.

The emphasis in this paper is on the rationale, the design, and the results of psychophysical experiments for obtaining performance characteristics of the HVS. These results can be used in the above three key tests of the algorithm, to derive appropriate values of thresholds T_1, T_2, and T_3, for the purpose of achieving near-optimal compromises between bit rate in the encoded sequence and image quality. Before presenting the rationale and the design of the psychophysical experiments, we briefly discuss the AGC characteristics of television monitors that must be taken into account for the purpose of interpreting the experimental data.

3. TELEVISION MONITOR AGC CHARACTERISTICS

The relationship between the local luminance, L, produced on a monitor's screen by the deflected electron beam and the voltage, V, applied at the plate is known to be governed by the so called "gamma law"

$$L(V) = c_0 V^{\gamma_d} + c_1, \tag{5}$$

where c_0, c_1, and γ_d are parameters that vary from monitor to monitor [6, 7]. In particular, the value of γ_d usually varies in the range 1.5 to 3.0.

In general, c_1 is negligible for most values of V, so it is omitted from Eq. (5) in our discussion. Furthermore, the voltage V is obtained from a digital-to-analog converter (DAC), whose input is the digital value y of the pixel's gray level intensity (usually, y is in the range 0 to 255). Thus, a good approximation for L is given by

$$L(y) = K_0 y^{\gamma_d} \quad \text{or} \quad \ln(L) = \ln(K_0) + \gamma_d \ln(y). \tag{6}$$

The most important fact to bear in mind about Eqs. (5) and (6) is that the vast majority of monitors are equipped with an AGC circuit, which has the effect of varying the values of c_0 and K_0 according to the average value of the applied voltage over the entire frame. Thus, Eq. (6) can be rewritten as

$$L(y, Y) = K_0(Y) y^{\gamma_d} = (\alpha Y + \beta) y^{\gamma_d}. \tag{7}$$

In Eq. (7) above we make it explicit that L depends on both the local pixel intensity $y(i, j)$ and Y, which is the average value of y over the entire image (i and j are the row and column indices for a particular pixel). The dependence of L on Y comes from $K_0(Y)$. Based on measurements with several monitors, a linear relation, $K_0(Y) = \alpha Y + \beta$, is assumed, where the parameters $\alpha < 0$ and $\beta > 0$ are monitor specific. For simplicity, the spatial arguments i and j of y were omitted in Eq. (7). It is obvious that, as the average input intensity Y increases, the value of K_0 decreases; as a consequence, a fixed local value of y in two different frames with averages Y_1 and Y_2 will result in a lower local luminance for frame 1 if $Y_1 > Y_2$.

To determine the values of α, β, and γ_d, we developed the following two-step procedure (in the first step we estimate γ_d, and in the second we estimate the values of α and β).

Step 1. Partition the screen into six large rectangles (arranged in two rows and three columns) and assign a uniform value for the intensity y within each rectangle, such that the average value over the entire image is Y_0. For each rectangle with $y = y_k$ use a photometer to measure the luminance L_k, $k = 1, 2, \ldots, 6$. Repeat the above process for different sets of six values of y whose average is also Y_0. In this manner, we can obtain pairs (y_k, L_k), $k = 1, 2, \ldots, 6m$, all measured with $Y = Y_0$, where m is the number of different partitioned images that were generated.

It follows from Eq. (7) that, since Y is fixed at Y_0, K_0 is also kept constant. Thus, if we plot the points (y_k, L_k) on a log–log scale, the slope of the straight line that best fits these data points will give us a good estimate of γ_d. When the procedure of step 1 was applied to a Sony PVM-1271Q monitor (which was later used in our psychovisual experiments), the plot of Fig. 3a was obtained, from which the value of γ_d was estimated to be 2.21.

Step 2. Here we fill the entire screen with a constant value of y, say Y_l, and we measure the luminance L_l. We repeat this procedure over a wide range of y values, and we thus obtain pairs (Y_l, L_l). We next observe that, since a value for γ_d has already been obtained in Step 1, we can now solve Eq. (7) for K_0 from the known values of L_l and Y_l; we also note that we can now use $Y = Y_l$ in Eq. (7) for all the images used in Step 2. Thus we obtain $K_{0,l} = L_l / Y_l^{\gamma_d}$ and then plot the resulting pairs $(Y_l, K_{0,l})$ with Y_l on the horizontal axis and $K_{0,l}$ on the vertical axis, both on linear scales. Since $K_0 = \alpha Y + \beta = \alpha Y_l + \beta$, the slope and vertical-axis intercept of the line that best fits the data will give us good estimates for the values of α and β. The procedure of Step 2, when applied to the Sony PVM-1271Q monitor, produces the plot of Fig. 3b, from which the values of α and β were estimated to be -2.24×10^{-6} and 1.34×10^{-3}, respectively.

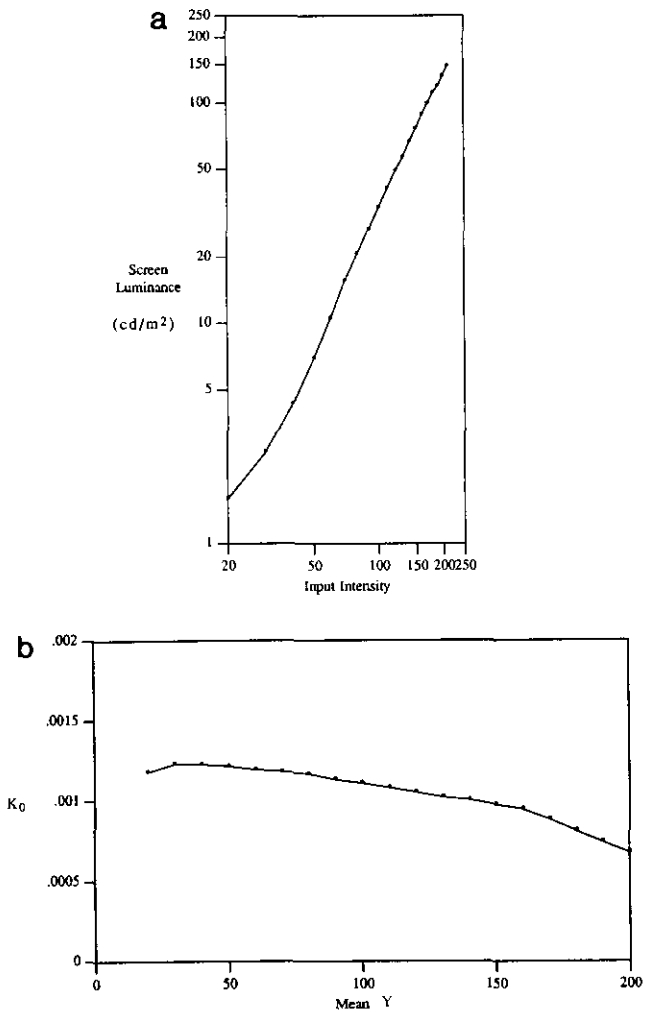

FIG. 3. (a) The luminance L plotted as a function of input intensity y on a log–log scale, with the average screen intensity kept fixed. The slope gives an estimate for the parameter γ_d of Eq. (7). In (b) the value of K_0 as a function of Y (see Eq. (7)) is plotted. The values of α and β are obtained from the slope and vertical intercept, respectively.

The importance of characterizing the behavior of a television monitor with a model, as expressed by Eq. (7), becomes evident when one experiments with a variety of such display devices. The values of α, β, and γ_d may vary widely from one unit to the next, but they allow us to express the value of a universal measurable quantity, i.e., the luminance L, in terms of the input intensity, y, for a particular unit. This, in turn, makes it possible to conduct the psychophysical experiments on one monitor only and express the resulting data in terms of the universal variable L. For different monitors, we can then cast these data in terms of the specific monitor's intensity variable y. Thus, we need only conduct the psychophysical experiments on one monitor for obtaining appropriate thresholds for the encoding algorithm. The values of these thresholds can then be fine-tuned for any other monitor. Our experience has shown that one set of threshold values is usually adequate for a wide variety of monitors.

4. PSYCHOPHYSICAL EXPERIMENTS FOR OBTAINING THRESHOLDS

In the three subsections that follow we present the rationale, the stimuli, the methods, and the results of three types of experiments that we conducted to help in setting values for the thresholds T_1, T_2, and T_3 of Fig. 2a. We also attempt to illustrate how one can fine-tune these values for any display monitor without having to repeat these experiments, as mentioned at the end of the previous section.

However, before we go into the details on each of the three types of experiments, we discuss the common properties and conditions that were shared by all of the three types. The vast majority of the experiments was performed on a Sony PVM-1271Q television monitor, whose screen measured 12 in. diagonally. We also conducted the same experiments with a subset of the observers using another PVM-1271Q display unit and a Sony Trinitron PVM-1910 monitor (20 in. diagonal). The images were viewed from a distance of 34.5 in., which is six times the image height, a factor normally used in similar situations. Thus, the picture height always subtended a visual angle of 9.53°. There were 720 and 480 pixels along the x (horizontal) and y (vertical) directions, respectively, resulting in an image of 8.625 by 5.75 in. on the Sony PVM-1271Q unit. The target blocks were squares whose sides b were of three different sizes: 8, 16, and 24 pixels. We note here for reference that an 8-pixel (horizontal or vertical) line segment subtended an angle of 9.55 min of arc. From this point on, we use the terms "target" and "block" interchangeably.

Images were generated on an Abekas A60 digital disk recorder and were displayed at a rate of 30 frames per second (interlaced). The integer intensity values used in the imaging system that we employed ranged from 1 to 254 inclusive. The values 0 and 255 are reserved for system synchronization signals, following the CCIR-601 international standard for digital video.

The generic form of the stimuli is shown in Fig. 4. The observer is asked to count the number of visible square targets, whose discriminability from the background becomes more difficult as we move from target 1 to target $r + G$. The manner in which the target blocks are visually segregated from their background box varies from one type of experiment to the next. In experiments of type 1, each target has a uniform intensity whose value is different from the uniform intensity of the background. In experiments of type 2 the targets are textured, but their

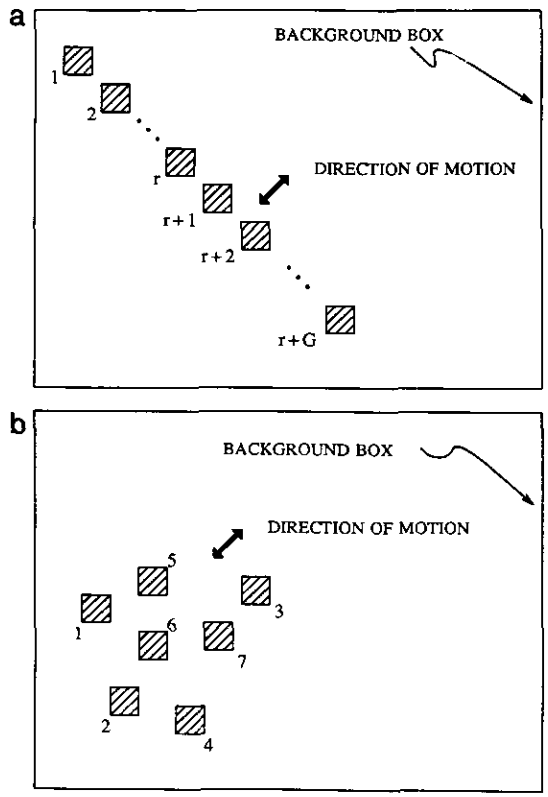

FIG. 4. The generic form of the stimuli used in the experiments: (a) Linear arrangement, (b) random arrangement.

average intensity is the same as that of the uniform background. Finally, in experiments of type 3 both the targets and the background are textured and they have the same standard deviation; a target is visually segregated from the background because its average intensity differs from that of the background. There are $r + G$ targets, arranged as shown in Fig. 4a, and are assigned discriminant values (such as intensity or standard deviation) such that they become progressively more difficult to detect as we move to higher-numbered targets. The integer r is a random variable, uniformly distributed in the typical range [0, 3]; since the first r targets are easy to detect, we call them "decoys." G is an integer (typically 7) which is fixed for a given type of experiment. In our experiments G ranged from a minimum of 4 to a maximum of 9. The task of the observer is to report the number of visible targets, which gives us a measure of the detectability of the targets. The variability in the number of the decoys from session to session, which was explained to the observers before the experiment, is meant to discourage them from expecting a fixed number of targets. Parenthetically, we also used another arrangement of targets, in which they occupied random positions within a lattice, as shown in Fig. 4b; the same scheme of having a random number r of visible decoys was used in the random arrangement as well.

To study the dependence of the thresholds on the speed of the targets, we created and displayed animation sequences in which the targets moved as a cluster in the direction shown by the arrows of Fig. 4, along the main diagonal with unity slope. The speed of motion was one of the parameters that we varied in all the experiments.

In a typical experiment, the screen was partitioned into an array of 2 by 2 large rectangular (360 by 240 pixels) background boxes, each with a different value for the background. We also used an arrangement of 2 × 3 boxes, each 240 × 240 pixels. Targets were placed within the boxes, but each box was assigned its own random number of decoys. The linear arrangement of Fig. 4a or the random one of Fig. 4b was used consistently throughout the screen in a given session of the experiment. We used multiple boxes in each screen to speed up the experimental process without sacrificing accuracy in the results.

To obtain a good representation, we selected the majority of the observers (roughly 90%) from people who use television mostly for home viewing. The rest were people who work on image processing. There were no systematic differences in the performances of these two groups. Nine observers, on average, participated in each of the experiments discussed in the following subsections. All observers had normal, or corrected to normal, vision and were naive as to the purposes of the experiments.

4.1. *Experiments of Type 1: Thresholds for Smooth Blocks*

Rationale. This group of experiments, collectively referred to as experiments of type 1, was conducted for guiding the setting of the value for threshold T_1 of expression (3). Its purpose, as shown in decision box B of Fig. 2a, is to test whether a nontextured block, such as one having an almost uniform intensity (a smooth block), can be detected by the observer on a *uniform* background having a different mean intensity. We assume that the block has already been classified as nontextured in the previous step by decision box A of the decision tree in Fig. 2a. The task of Experiment 1 is to find how T_1 varies with the following parameters: block size, background intensity y (or luminance L), and block speed. Since T_1 is the threshold for the difference between mean intensities of the block and the background, the experiment must test the ability of the HVS to detect a smooth block near threshold.

Stimuli. On the basis of the above considerations, we utilized the stimuli shown in their generic form in Fig. 4, in which the background was of uniform intensity y. Target q was assigned intensity $y + (r + G + 1) - q$, $q = 1, 2, \ldots, r + G$. Observers were shown these stimuli and

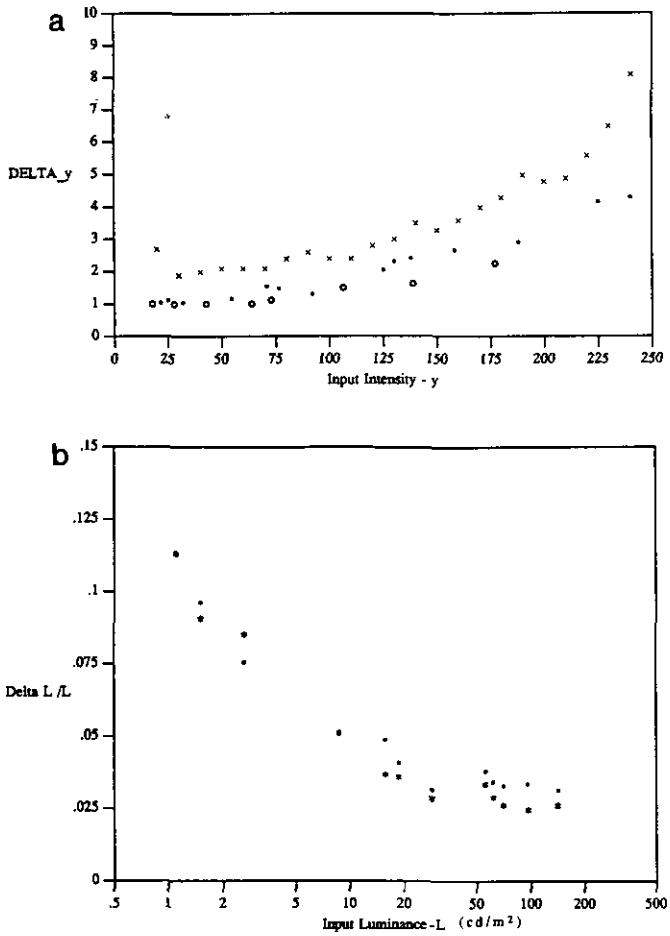

FIG. 5. (a) The jnd (just noticeable difference) $\Delta y = T_1$ as a function of background intensity y with the block size as a parameter. Sizes of 8 × 8, 16 × 16, and 24 × 24 are shown by x's, solid circles, and open circles, respectively. (b) The universal curve of $\Delta L/L$ versus L, with target speed v as a parameter. Solid circles and asterisks are used to show results for stationary and moving blocks (speed = 101.2 min of arc per second), respectively.

were asked to report the number, w, of targets that they could detect. The difference $(r + G + 1) - w$ is a good measure of the threshold incremental intensity, or *just-noticeable difference* (jnd), Δy, that is barely visible against a uniform background y. This value of $\Delta y = (r + G + 1) - w$, which is a function of y, was indeed the value used for threshold T_1 in the encoding algorithm in [4]. Three different block sizes were used in separate groups of experiments: squares with side length of $b = 8$, 16, and 24 pixels. For the 16- and 24-pixel targets we conducted experiments in which the blocks were moving along the 45° diagonal with equal horizontal and vertical velocity components. The values of these components were set at 1, 2, or 4 pixels per frame. Since the image display system displayed frames at a rate of 30 Hz, these rates correspond to angular speeds of 50.6, 101.2, and 202.4 min of arc per second along the diagonal.

Methods. The number of observers that participated in the groups of experiments with block size 8, 16, and 24 pixels were 10, 10, and 8, respectively. Observers were given adequate time to count the detectable blocks. Although there were no time limits for responding, they were asked to complete the task as soon as they could. This is a reasonable approach, since the end product of the encoding/decoding scheme is a sequence of frames which viewers watch casually, without having the inclination or luxury of time to scrutinize for minute details. As for the values of the background intensity y, at least 12 values were used, spanning the allowable range.

Results. What is needed for decision box **B** of the algorithm of Fig. 2a is a value $T_1 = \Delta y$ for the threshold as a function of the background intensity y, the block size b, and the target speed v, based on the observers' responses. This is plotted as a function of background intensity y with the block size as a parameter in Fig. 5a, averaged across all observers for each condition. The graph was obtained from experimental results with the Sony PVM-1271Q monitor with stationary targets. As expected, the threshold $T_1 = \Delta y$ decreases with increasing target size. Although it is not obvious from Fig. 5a, the Weber-like ratio $\Delta y/y$ seems to have a fixed value for a wide range of input intensities. To study the Weber ratio $\Delta L/L$ [8, 9] we must work with the universal variable L, the *luminance* (see Section 3).

The effect of the background luminance L and speed v on the Weber ratio $\Delta L/L$ is shown in Fig. 5b, which displays the average performance of the observers for 16 × 16-pixel targets. Results for stationary and moving targets (speed = 101.2 min of arc/s) are shown by solid dots and asterisks, respectively. The Weber ratio for each point was estimated from Eq. (6), from which $dL/L = \gamma_d\, dy/y$, hence

$$\frac{\Delta L}{L} \approx \gamma_d \frac{\Delta y}{y}. \tag{8}$$

Thus ΔL/L is obtained from the known values of γ_d, Δy, and y and the value of the luminance L, which is needed in Fig. 5b, is, of course, computed from Eq. (7). As can be seen from Fig. 5b, $\Delta L/L$ decreases as the speed increases, verifying that moving targets are easier to detect. The Weber ratio $\Delta L/L$ decreases as L increases, as expected, and remains nearly constant for a large range of L values (18.6–144 cd/m²), corresponding to intensity values (y) in the range 70 to 210. The reasons for conducting experiments of type 1, even though data for $\Delta L/L$ vs L exist in the literature, are explained under **Discussion**. This universal graph ($\Delta L/L$ vs L) can serve as the starting point toward obtaining $T_1 = \Delta y$ for any set of parameters and for any monitor. The process is the reverse of that followed for obtaining $\Delta L/L$ vs L from Δy vs

y. In other words, first $\Delta y/y$ is computed from γ_d and $\Delta L/L$ using Eq. (8). Then y is found, as a function of average image intensity Y, from L, α, β, and γ_d using Eq. (7). These values of $\Delta y = T_1$ can be employed in decision box **B** of the decision tree in Fig. 2a.

4.2. *Experiments of Type 2: Textured versus Smooth Blocks*

Rationale. This group of experiments was designed to determine an appropriate value for the threshold T_2 of decision box A in Fig. 2a. As shown in expression (2), this test attempts to classify each block into one of two major categories (nonsmooth and nontextured), by examining the intensity variations within the block about the average intensity. In fact, a convenient way to test how easily this variation can be detected by the HVS is to make the average intensity of the *textured* targets equal to that of a *uniform* background and to vary the standard deviation of the targets until they become barely visible. The standard deviation at which the target becomes visible is the value below which the textured block blends with the uniform background of the same luminance. This is, indeed, the value above which textured blocks are segregated from uniform blocks by the HVS. Subsequently, we can conduct additional experiments by varying the speed of the targets, as well as their spatial frequency characteristics by convolving the images with two-dimensional (2-D) low-pass filters at various cutoff frequencies.

Stimuli. The stimuli were very similar in most respects to those employed in the experiments of type 1, as shown in their generic form in Fig. 4. The background in the box is of uniform intensity y. Targets were formed by assigning to each of their pixels a random intensity integer y_r with a uniform distribution in $[y - c, y + c]$, thus assuring the same average intensity y as that of the background. The value of c is assigned to the targets so that their visibility decreases monotonically as we go from target 1 to target $r + G$; this is accomplished by simply assigning to target q a value of $c = 2(r + G + 2 - q)$, for $q = 1, 2, \ldots, r + G$. Hence $4 \leq c \leq 2(r + G + 1)$, and of course, the standard deviation σ is $c/\sqrt{3}$. In this experiment there were seven regular targets ($G = 7$) and up to 2 decoys ($0 \leq r \leq 2$). All the blocks were of the same size (24 × 24 pixels). In addition to displaying static targets, we also used moving targets with the same set of speeds as in experiment 1. To study the effect of visual detail (i.e., of spatial frequency content) on detectability, we conducted a full set of experiments for three different types of targets: (a) the ones just described, which we refer to as the *unfiltered* stimuli; (b) blocks which were filtered using a separable finite-impulse response (FIR) low-pass 2-D filter whose cutoff frequency was 12.56 cycles/degree; this frequency corresponds to one-quarter of the sampling frequency $f_s = 50.27$ pixels/degree, i.e., half the effective signal band; we call these the *half-band-filtered* stimuli; (c) blocks filtered as in (b) above, but with a cutoff frequency of 6.28 cycles/degree, which we term *quarter-band-filtered* blocks.

Methods. Similar instructions were given to the eight observers of this type of experiments as given to those who participated in experiments of type 1. Sixteen different values of background intensity y were tried in the range 10 to 240, corresponding to luminance values in the range 0.9 to 357 cd/m^2. Each observer viewed targets on 16 different background intensities; for each intensity, he/she reported on detectability of stationary and moving targets (three different speeds, v). For each combination of y and v, all three types of targets were displayed (unfiltered, half-band- and quarter-band-filtered). Thus, each observer viewed 192 (16 × 4 × 3) combinations of conditions.

Results. The observer's response w of the number of detectable targets was used to compute σ_t, which is the standard deviation $c/\sqrt{3}$ corresponding to the $(r + G - w + 1)$th block; σ_t is a measure of the standard deviation threshold at which textured blocks are segregated from the uniform background. The ratio σ_t/L, averaged across observers, is plotted in Fig. 6a against L, the background luminance, with the speed v as a parameter. Results for stationary blocks ($v = 0$) and blocks moving at $v = 101.2$ min of arc per second are shown using solid circles and open triangles, respectively. Generally, moving targets are easier to detect, as expected. The effect of filtering on detectability can be seen from the data displayed on Fig. 6b. Here σ_t is plotted versus y for unfiltered (solid dots) and quarter-band-filtered images (open circles). The results for half-band-filtered blocks are somewhere in between, but are omitted to avoid cluttering of the figure. It is evident that blocks with low spatial frequency components are easier to detect than blocks with high frequencies for the same intensity variance.

4.3. *Experiments of Type 3: Thresholds for Textured Blocks*

Rationale. The objective of this set of psychophysical experiments was to come up with a reasonable value of threshold T_3 used in the test of decision box C of Fig. 2a and in expression (4). The test in expression (4) is very similar to that of expression (3), which was applied to nontextured blocks; the difference is that motion compensation is not employed here (see Section 2). The task implied by (4) is to find how sensitive the HVS is in distinguishing between two nonsmooth blocks which differ in average intensity. Our objective is to find an optimal value for T_3 to resolve the conflicting requirements of

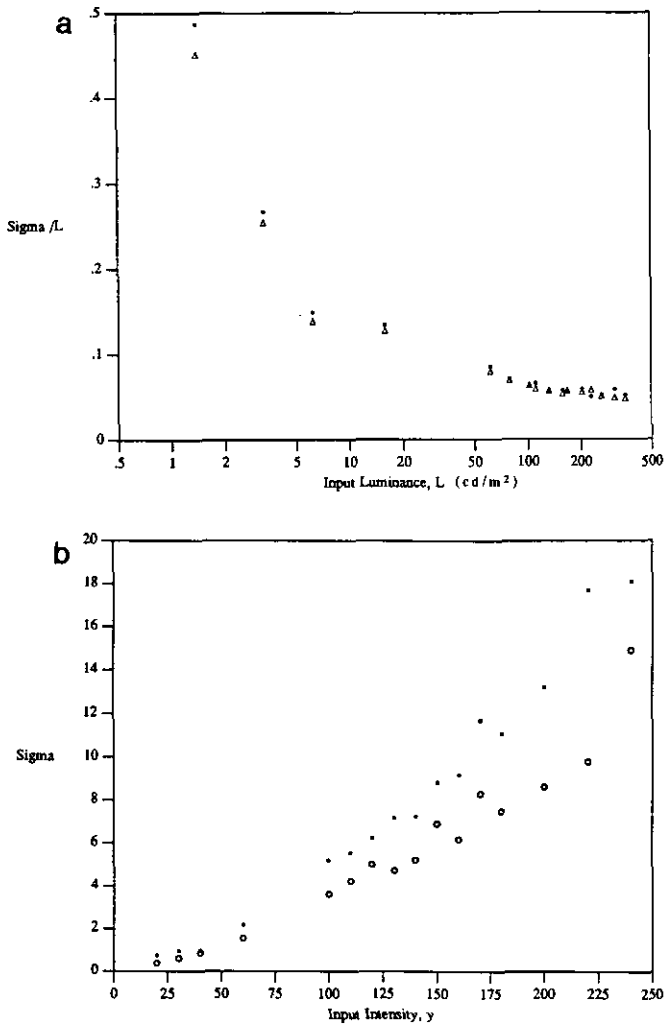

FIG. 6. (a) The ratio σ_t/L plotted versus background luminance L with speed as a parameter. Results for stationary and moving targets (speed = 101.2 min of arc per second) are shown by solid circles and open triangles, respectively. (b) The threshold standard deviation σ_t shown as a function of background intensity y. Solid and open circles are used for unfiltered and quarter-band-filtered targets, respectively.

low bit rate and high image quality in the motion sequence. This is achieved by obtaining the maximum value $T_{3,\max}$ for which the image quality is barely distinguishable from sequences obtained with lower values of T_3, while sequences utilizing values higher than $T_{3,\max}$ show significant image degradation to the HVS. To facilitate such a study, it was limited to the important case where the blocks are *textured* blocks and the experiments were designed to study the dependence of $T_{3,\max}$ on average intensity, y, the standard deviation of the texture, σ, and the spatial frequency content of the image. In addition to static blocks, we also employed animated sequences to study the effect of speed on the detectability of the targets.

Stimuli. Since the objective is to test the ability of the HVS to distinguish between textured blocks having different average intensity, we employed stimuli in which both the background and the targets were formed by noise distributions which differed only in their mean intensity. In other words, pixels in the background were assigned random intensity integers with a Gaussian distribution of mean y and standard deviation σ. Targets were positioned as shown in Fig. 4a. Pixels in target block q were assigned random integers with a Gaussian distribution of mean $y + y_q$ and standard deviation σ, where $y_q = (r + G + 1) - q$, $q = 1, 2, \ldots, r + G$. Thus, the targets have the same σ as the background and they only differ from it with respect to the average intensity by y_q. Just as in Experiment 1, the difference $\Delta y = (r + G + 1) - w$, where w is the number of visible targets, is a good estimate for the threshold T_3 to be used in the algorithm. Indeed, this value of Δy represents the just-noticeable difference in average intensity. All targets were square with 24-pixel sides. Five different values of standard deviation were used: $\sigma = 0$ (uniform intensity), 4, 8, 12, and 16. Eight different average intensity values were tried in the range 20 to 200. Stationary as well as moving targets (speed = 101.2 min of arc/s) were employed. Finally, in addition to the original images described above, we also used a set of quarter-band-filtered images, obtained by convolving the original with the same low-pass FIR filter as the one used in Experiment 2 with a cutoff frequency of 6.28 cycles/degree.

Methods. Sixteen observers participated in two groups of experiments of type 3. In the first experiment group each of the eight observers was shown targets with all *five* values of σ (see *Stimuli* above) for each of *eight* average intensities; for each of these conditions we used both *unfiltered* and *filtered* images. Thus each observer viewed 80 ($8 \times 5 \times 2$) combinations of conditions. In the second group of experiments, σ was held fixed at 12. Each of the eight observers viewed 32 ($8 \times 2 \times 2$) combinations of conditions: for each of eight average intensities we used both unfiltered and quarter-band-filtered images.

Results. The threshold incremental mean intensity $T_3 = \Delta y$ was computed, based on the observers' responses. These values of T_3, averaged across all the observers, are plotted in Fig. 7a as a function of the background mean intensity y with the texture's standard deviation σ as a parameter. The results were obtained on the Sony PVM-1271Q monitor with 24 × 24-pixel targets. Since the curve for $\sigma = 0$ corresponds to uniform intensity (smooth blocks on uniform background), the bottom curve of Fig. 7a is almost identical to the bottom curve of Fig. 5a of Experiment 1, as expected. We also observe that as σ increases, i.e., as the texture becomes "rougher," the value of T_3 increases, which means that it is more difficult

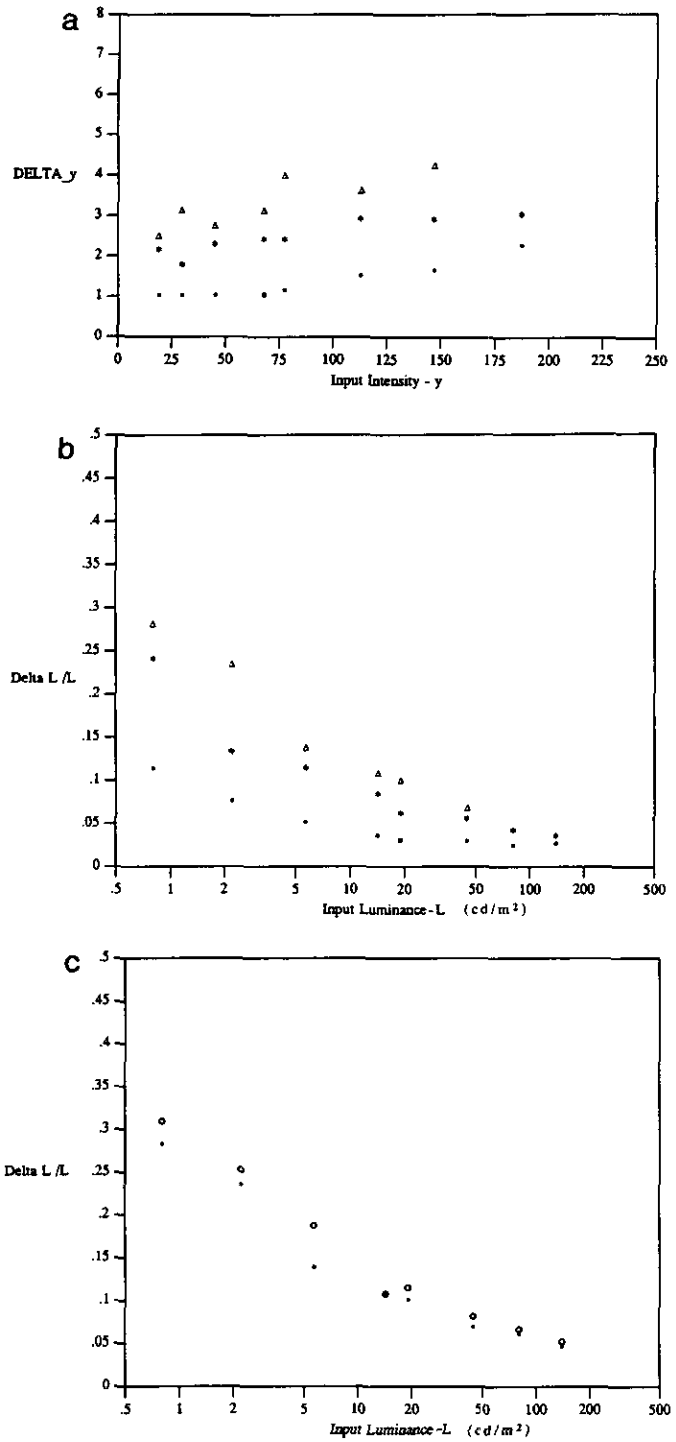

FIG. 7. (a) The jnd $\Delta y = T_3$ plotted against the average background intensity y, with the standard deviation of the texture (noise) as a parameter. Triangles, asterisks, and solid circles are used to denote values for σ of 12, 8, and 0, respectively. Data were obtained with unfiltered images and stationary targets of 24 × 24 pixels. (b) The variation of $\Delta L/L$ vs L is shown by these three curves, using as a parameter the standard deviation σ. Data were obtained as in (a) and the same notational convention for the values of σ was used as in (a). (c) The universal curve of $\Delta L/L$ versus L, with solid and open circles denoting unfiltered and quarter-band-filtered images, respectively.

for the HVS to detect targets that have the same texture as the background but differ from it only in average intensity; the degree of difficulty increases with σ. The data of Figure 7a are recast in Fig. 7b in the universal form of the Weber ratio $\Delta L/L$ as a function of L (see also Fig. 5b), with σ as a parameter.

The effect of spatial-frequency content on T_3 is shown in Fig. 7c, in which the universal Weber ratio $\Delta L/L$ is plotted against L, with data for unfiltered and quarter-band-filtered images denoted by solid and open circles, respectively. The value of σ was fixed at 12 for all the data of Fig. 7c. These plots reveal that T_3 increases only slightly for filtered images. By contrast, when one compares the data for unfiltered texture images of Fig. 7c to the data for uniform (stationary) targets of Fig. 5b (solid circles), it is evident that uniform targets are much easier to discriminate against a uniform background than are textured targets against a textured background. As a result, T_3 is two to three times as large as T_1. Thus, even though the tests defined by expressions (3) and (4) involve the same computational steps (except for motion compensation), the actual threshold values that should be used in their implementation are quite different. The above results were all obtained from the first group of experiments (see **Methods** above). The objective of the second group of experiments was to observe the effect of target motion on detectability. It was found that, when the targets are moving at the speed of 101.2 min of arc/s, the value of T_3 is scaled by a factor of approximately 0.86 and 1.05 for unfiltered and quarter-band-filtered images, respectively.

5. DISCUSSION—ALGORITHM PERFORMANCE

We emphasize that, even though the experiments described in this paper were motivated by the particular algorithm at hand, their results are of general interest and can be applied to a wide variety of encoding schemes. For example, the present version of the algorithm in [4] does not adapt the weights based on the speed of the blocks. Nevertheless, we conducted experiments to study the dependence of the thresholds on the speed of the blocks. These experimental results can be used in future versions of the algorithm or they can be applied to other algorithms that utilize the information on the speed of the blocks to adapt their thresholds.

The data for some of our experimental conditions are available in standard psychophysical reference manuals. In particular, plots of the Weber ratio $\Delta L/L$ versus L (or ΔL versus L) are included over wide ranges of L in most standard texts on vision [8, 9] and picture coding [10]. It is thus reasonable of the reader to ask why we conducted experiments of type 1. There are two reasons for that:

First, in addition to data for stationary targets, we also obtained data for moving targets at various speeds, which are not easy to find in the literature. Since stimuli for the latter type of experiments were generated anyway, it was easy to test the observers with stationary blocks during the same period, to compare the effect of speed on detectability. Second, the algorithm's tests require the threshold values in terms of the monitor's intensity values y and not in terms of the luminance L. The mapping from the universal variable L to the particular variable y can be carried out as outlined in Section 3 but, since this is only an approximation, it is best to test the particular television monitor directly in order to match its characteristics. As far as experiments of types 2 and 3 are concerned, the desired plots are generally not readily available and this was another reason, in addition to the two just mentioned, why we conducted our own experiments in those cases. It is obvious from the graphs in Figs. 5, 6, and 7 that the values of the thresholds vary in a well-behaved manner with the various parameters (block size, speed, local background intensity, etc.). Piecewise linear functions can be used to adjust the value of the threshold to suit the local conditions for each target block under examination. This is a prime example of an adaptive threshold modification technique, which is based on the characteristics of the HVS.

Turning to the results from each of the three types of experiments conducted, the general observation is that there was very little interobserver variation for any given task. The performance in particular tasks and its relevance to the encoding algorithm is discussed in the corresponding subsections of Section 4. Some observations that are common to all three types of experiments are discussed next.

As expected from spatial integration considerations, thresholds were lower for larger target sizes in all experimental conditions. Similarly, moving targets proved much easier to detect than stationary ones, which was also expected. Pilot studies indicate that detectability is easiest over a range of speeds centered around $v_0 = 50.6$ min of arc per second, but that it deteriorates for very slow and for very fast moving targets relative to v_0. Note that in the context of conditional replenishment based on the visibility of motion-compensated prediction error, ease of target detection translates to greater expenditure of bits for coding. Finally, when comparing the effect of filtering on the value of thresholds, the reader should bear in mind that uniform and Gaussian noise distributions were used in the unfiltered images for experiments of types 2 and 3, respectively (of course, following filtering, the uniform noise distribution of the original image tends to become Gaussian).

Values of thresholds T_1, T_2, and T_3 (see expressions (3), (2), and (4), respectively) which were based on the above experimental results were incorporated in the encoding algorithm used in [4] and outlined in Section 2. The variation of the threshold values with neighborhood conditions (background intensity, texture characteristics, etc.) was implemented. Several standard and locally available sequences were tried with the algorithm, each testing particular attributes and conditions. Examples are "Miss America," the least demanding sequence, depicting a sitting woman who speaks in a newscaster's posture; "Beach and Flowers," which includes rotating objects and "panning"; "Table Tennis" featuring a ping-pong game with panning, zooming, a rough-textured background, and a black-and-white cartoon poster; and "Flower Garden," portraying a flowery field with a windmill, shot from a moving car. The latter two sequences are part of the standard ISO test sequences [3]. We mention parenthetically that the algorithm estimates the pan and/or zoom parameters and compensates for them, before the block classification step of Fig. 2a is applied.

The encoding process was performed on an Alliant FX/8 parallel computer, which is a vector concurrent machine with a peak performance of 188 MFLOPS (million floating-point operations per second). The encoded frame sequence was displayed either side by side with the original sequence, each occupying half of the screen, or immediately followed it in time, each filling the entire screen. The encoded sequences were comparable in quality to the original ones, with imperfections and artifacts that were barely visible. In a typical sequence the bit rate is reduced from about 100 Mbits/s for the original to roughly 1 Mbit/s for the encoded series. Additional techniques were employed such as multiple reference frames, zoom and pan compensation, heuristics to compute D_x and D_y of Eq. (1) efficiently, to mention just a few, which are discussed in detail in a companion paper, dedicated to the computational aspects of the encoding algorithm [4].

To give the reader a concrete idea of the value of utilizing the HVS data from the psychophysical experiments described in this paper, we present some of the results obtained with the image sequence coder developed in [4] in which the HVS data relevant to the decision tree in Fig. 2a were applied. The graphs in Fig. 8 show the bit rate per frame as a function of the frame number for a sequence of 100 frames from the ISO image-sequence Table Tennis. The improvements introduced in the CCITT RM baseline coder [1, 15] are best demonstrated when the quantization step q_s is kept constant (in Fig. 8 it is $q_s = 8$). The top solid line corresponds to the bit rate obtained when the coder is operated without global-motion compensation and without the application of the HVS data. The coder is then similar to the baseline RM coder. The first frame, as well as frame 90 (at which a scene cut occurs), are coded intra.

In the first 25 frames of the sequence there is only local

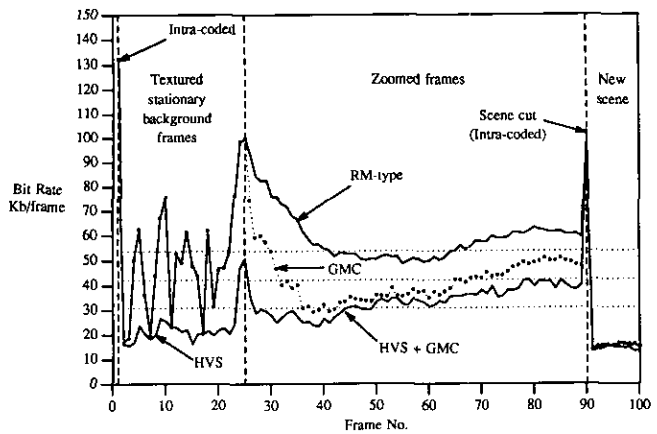

FIG. 8. Bit rate as a function of frame number for 100 frames of the ISO image sequence "Table Tennis," using a fixed quantization step size $q_s = 8$. Upper solid line, RM-type coder; dotted line with solid circles, result of adding global-motion compensation (GMC); lower solid line, result of adding both GMC and HVS-based conditional replenishment ("texture freeze").

motion of the table tennis player on a rough textured background which (in combination with apparent slight camera jitter) causes the relatively large fluctuations in bit rate (for the fixed quantization step size used). In the next group of frames—until the scene cut in frame No. 90—*zooming* is present. The dotted line with solid circles shows the bit rate obtained when global-motion compensation (GMC) is applied by using an additional reference frame which contains a zoomed version of the last reconstructed frame. Clearly, there is a significant reduction in rate for the group of zoomed frames. However, GMC does not affect the first group of frames. The need to reduce the rate for this group of textured-background frames motivated the psychophysical experiments reported in this paper, enabling the efficient implementation of the decision tree in Fig. 2a. The result of using the HVS data in this decision tree is shown by the lower solid line in Fig. 8 (denoted by HVS, for the first group of frames, and GMC + HVS for the zoomed frames). The reduction in bit rate for the textured frames, due to the texture freeze, is very significant, while the quality of the reconstructed frames remains basically unchanged. Some reduction in rate is also obtained for the zoomed frames, as texture freeze can be applied here too—with respect to the zoomed reference frame mentioned above. If GMC is *not* applied, texture freeze in frames with global motion (e.g., zoom and/or pan) should *not* be applied. The corresponding average bit rates, over the whole 100 frames, are marked in Fig. 8 by the light dotted horizontal lines (being approximately 54, 42, and 31 kbit/frame—for the RM, GMC, and GMC + HVS coders, respectively). Clearly, the improvement obtained by using the HVS data in the conditional replenishment decision tree well justifies its application.

ACKNOWLEDGMENTS

The authors thank Mr. Krishna Mackay who, under their supervision at AT&T Bell Laboratories, developed much of the software for generating and displaying stimuli for experiments of type 1 and 2 and administered several experiments. We also thank Dr. N. S. Jayant, who offered us guidance and encouragement and provided the facilities and resources for this project. We thank the people in his group, particularly Thrassos Pappas, Bob Safranek, and Christine Podilchuk, for valuable discussions, feedback, help, and support. Finally, we appreciate the patient participation of numerous colleagues and co-workers as observers in the experiments.

REFERENCES

1. S. Okubo, Video codec standardization in CCITT study group XV, *Signal Process. Image Comm.* **1**(1), June 1989, 45–54.
2. H. Yasuda, Standardization activities on multimedia coding in ISO, *Signal Process. Image Comm.* **1**(1), June 1989, 3–16.
3. L. Chiariglione, Standardization of moving picture coding for interactive applications. *IEEE Global Telecommunications Conference (GLOBECOM)*, pp. 559–563, Nov. 1989.
4. D. Malah, Improvements in hybrid image-sequence coders for storage applications, in preparation.
5. J. Yogeshwar, *A New Perceptual Model for Video Sequence Encoding*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 1990.
6. E. Catmull, A tutorial on compensation tables, *Comput. Graphics* **13**, 1979, 1–7.
7. J. D. Foley and A. vanDam, *Fundamentals of Interactive Computer Graphics*, Addison–Wesley, Reading, MA, 1983.
8. T. N. Cornsweet, *Visual Perception*, Academic Press, New York, 1970.
9. H. B. Barlow and J. D. Mollon (Eds.), *The Senses*, Cambridge Univ. Press, Cambridge, England, 1982.
10. A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation and Compression*, Plenum, New York, 1988.
11. J. Yogeshuar and R. J. Mammone, A new perceptual model for video sequence encoding, *International Conference on Pattern Recognition, Atlantic City, NJ, 1990*, pp. 188–193.
12. A. Wong *et al.*, MCPIC: A video coding algorithm for transmission and storage applications, *IEEE Comm. Mag.*, Nov. 1990, 24–32.
13. C. Herpel *et al.*, Adaptation and improvement of CCITT reference model 8 video coding for digital storage media applications, *Signal Process. Image Comm.* **2**(2), Aug. 1990, 171–185.
14. D. Adolph and R. Buschmann, 1.15 Mbit/sec coding of video signals including global motion compensation, *Signal Process. Image Comm.* **3**(2–3), June 1991, 259–274.
15. Description of Reference Model 8 (RM8), *Document 525, CCITT SGXV*, Working party XV/4, Specialist Group on coding of visual telephony, June 1989.

THOMAS V. PAPATHOMAS was born in Kastoria, Greece, in 1949. He received his B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Columbia University in 1971, 1972, and 1977, respectively. He worked as a member of the technical staff and as a supervisor at AT&T Bell Laboratories from 1977 to 1982, designing microprocessor-based controllers for DC and AC telecommunication power plants, and then as a researcher in the Department of Vision Research at Bell Laboratories from 1982 to 1989. He is presently an associate professor of biomedical engineering and the Assistant Director of the Laboratory of Vision Research at Rutgers University. He taught as a lecturer at Columbia University from 1976 to 1983. He is a member of Tau Beta Pi, Eta Kappa Nu, IEEE, SPIE, and ARVO. His interests are in the areas of motion, stereo and texture perception, image processing, and scientific visualization.

DAVID MALAH received the B.Sc. and M.Sc. degrees in 1964 and 1967, respectively, from the Technion–Israel Institute of Technology, Haifa, Israel, and the Ph.D. degree in 1971 from the University of Minnesota, Minneapolis, Minnesota, all in electrical engineering. During 1971–1972 he was an assistant professor at the Electrical Engineering Department of the University of New Brunswick, Fredericton, New Brunswick, Canada. In 1972 he joined the Electrical Engineering Department of the Technion, where he is presently a professor. From 1979 to 1981 he was on sabbatical leave at the Acoustic Research Department of AT&T Bell Laboratories, Murray Hill, New Jersey, and a consultant at Bell Labs during the summers of 1983, 1986, and 1988. During 1988/1989 he was on sabbatical leave at the Signal Processing Research Department of AT&T Bell Laboratories in Murray Hill. Since 1975 (except during 1979–1981 and 1988–1989) he has been in charge of the Signal Processing Laboratory, at the EE Department, which is active in speech and image communication research and real-time hardware development. His main research interests are in image and speech coding, image and speech enhancement, and digital processing techniques. Since 1987 he has been a Fellow of the IEEE.