JOURNAL OF VISUAL COMMUNICATION AND IMAGE REPRESENTATION
Vol. 1, No. 2, November, pp. 189–198, 1990

Attentive Transmission

HAGIT ZABRODSKY AND SHMUEL PELEG

Department of Computer Science, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

Received March 12, 1990; accepted September 12, 1990

A simple *attention-based* model is proposed for efficient transmission of visual information using multiresolution structures. Images are sampled nonuniformly in space and time, such that sampling is dense at the focus of attention and sparse in the periphery (retinal-like). Assuming that the focus of attention usually corresponds to eye position while scanning an image, image features which are "eye catching" (such as sharp edges, motion, and high flicker rate) are used to drive the dense center of sampling. The transmitted image is reconstructed by combining each new sample with previous samplings to give progressive transmission. The selection of sampling points depends only on previously transmitted information, and only sampled values without their location need to be transmitted.

1. INTRODUCTION

A simple model of *visual attention* is used to efficiently transmit visual information. The human attention mechanism is a means of selecting a subset of a large amount of stimuli, enabling concentration of processing power on this limited subset, while processing the rest of the stimuli less efficiently. Reading this page, for example, the reader selectively attends to a few words at a time out of the many on this page.

The process of efficient transmission of visual information can be considered as a subsampling problem. To transmit an image efficiently means to subsample the image so that for any transmission level the image could be reconstructed as visually acceptable as possible. In analogy to human visual attention mechanisms, the input image represents the vast amount of input stimuli. The sampling points are analogous to the limited processing power available. Incorporating attention involves the concentration of processing power and sampling points on subareas of the image while leaving other areas sparsely sampled.

In the visual attention model we assume that there is a high correlation between center of attention and *point of foveation* (area in visual world appearing in the center of the visual field) [20, 14]. Studies of eye movement and point of foveation define features and characteristics that can be attributed to the visual attention mechanism. Using these features we define an *attentive transmission system,* which transmits, at each step, a *sampling* of the input, analogous to the information obtained while foveating. The sampling is nonuniform and retinal-like, dense in the center and sparse in the periphery. The center location and resolution of the samplings change according to the attention model.

To implement attentive transmission two types of image pyramids are used: a *Laplacian pyramid* which represents multiresolution spatial information of images and a *Difference Laplacian energy pyramid* (DoL pyramid) representing temporal information in image sequences. A *quad tree* is embedded in these pyramids by mapping each node in the quad tree to a pixel of the image pyramid. A "sampling" of an image is defined as a branch of the quad tree in the image pyramid. These data structures are defined in Section 3.

Transmitting several samplings of the Laplacian pyramid allows the receiver to partially reconstruct the pyramid, and from it an approximation to the imput image. For image sequences, additional transmission of temporal information from the DoL pyramid enables the focusing of attention on areas of temporal change. Similar images in sequences, as in motion sequences, need not therefore be reconstructed from scratch.

Choosing the location and resolution of the samplings is analogous to choosing a new *focus of attention* (FOA), and we used an *attention function* to model human visual attention. The criteria in the attention function are selected so as to produce two important advantages to the system.:

1. With attentive transmission, the limit on the number of samplings transmitted is not known apriori. The attention function chooses the samplings so that, at any moment, the reconstruction is as visually recognizable as possible. The first samplings include low spatial frequency information in the image, giving general outlines of the image contents. As transmission progresses the samplings include information of higher spatial frequency, with more details of the image. The higher spatial frequency samples are concentrated mainly in areas of "interest" in the image. The attention function gives

1047-3203/90 $3.00

high priority to sampling in areas of high temporal frequencies enabling better reconstruction of moving objects rather than background.

2. Selection of new sampling points by the attention function depends only on previously transmitted information. The receiver can therefore predict the location of the subsequent sampling from the available data, and there is no need to transmit the sampling locations and resolution.

Smart sensing, introduced by Burt [2], is a class of systems that incorporate attention to process efficiently sensed information, where the task at hand directs the processing power to most needed parts of the input. Smart sensing, as a method for image compression, requires that processing power be concentrated upon the more salient or interesting features of the input image. Burt [4] developed a tracking system based on smart sensing, where the interesting features of the spatiotemporal input are the gradients in the temporal domain (changes introduced by motion). The tracking can be regarded as part of effective compression of the temporal difference between images in a sequence.

In this paper an approach for efficient resampling of a sequence of 2-D images is developed, modeled as efficient image transmission. This approach follows the smart sensing methodology, using a model of visual attention. Modeling of the visual attention system was introduced by Koch and Ullman [10], where a possible underlying neuronal circuitry was described. Previous work on nonuniform sampling which is directed by the input image appeared in [12, 15, 21].

2. FEATURES OF VISUAL ATTENTION

The brain directs the eyes so that the central area of the retina—the *fovea*—receives visual information from a selected and limited part of the visual world. *Foveation* is the process of directing the fovea to a specific location for a cetain duration. We assume (following [20] and recently [14]) a correlation between foveation and attention and that focusing of attention can be approximated and studied by measuring the points of foveation. Following such studies a model of human attention can be assumed.

We use a visual attention model, similar to the model assumed in [10], having the following features:

Single focus—Physical properties of the retina imply a single fixation point at a time, and *focus of attention* (FoA) is therefore assumed to be at a single location. Some experiments [19] show a possibility of divided attention, but it has not yet been shown whether there is actual double FOA or a widening of the single FOA's field to include several details.

Zoom lens—It has been shown psychophysically [7] and physiologically [11] that the spatial extent of the FOA may vary (consider, as an example, focusing attention on a person walking toward you versus focusing attention on the face and expression of the person). When FOA is spatially spread out, there is less sensitivity to "attention catching" features [7], as if spreading of attention incorporates spreading of limited processing power.

Peripheral information—Attention enhances the selected location rather than blocks irrelevant information [8]. This ensures that information is still gathered and processed in the periphery, yet less sensitively.

Refocusing—Attention, as foveation, shifts from one location to another. Yarbus [20] studied and measured the criteria for foveation on images from which he inferred attraction criteria for FOA localization:

- sharp spatial gradients (edges) introduced by contrast,
- sharp spatial gradients (edges) introduced by color differences,
- temporal gradients introduced by high velocity motion or sudden appearance and disappearance of an object,
- high-level criteria according to task at hand, as attending to a certain precued location.

The relationship and relative influence of each of the above criteria is yet unknown, but it can be shown that changes in the temporal dimension are more prominent than those in the spatial domain. A bright flash of light, even in the periphery, will draw our attention away from any static object currently being attended to.

Proximity—There is some experimental evidence [16, 5, 6] that when shifting the FOA to a new location, there is a preference for closer locations.

Inhibition of return—Psychophysical evidence [13. 9] shows that there is a tendency of delaying the return of FOA to a position recently attended to.

3. DATA STRUCTURES FOR ATTENTIVE TRANSMISSION

Attentive transmission is a smart sensing system for sampling and transmitting 2-D image sequences. The images are densely sampled in the sampling center and sparsely sampled in the periphery (retinal-like). Location of the sampling center (the FOA) depends on visual attention features similar to those described in the previous section. The attentive transmission model uses *image pyramids* as follows:

Gaussian pyramid is a multiresolution image representation, where an input image has a set of reduced-resolution images associated with it. These images are generated by repeatedly blurring the image followed by subsampling by a factor of 2. We blurred the images by a convolution with a 4×4 normalized, separable, and sym-

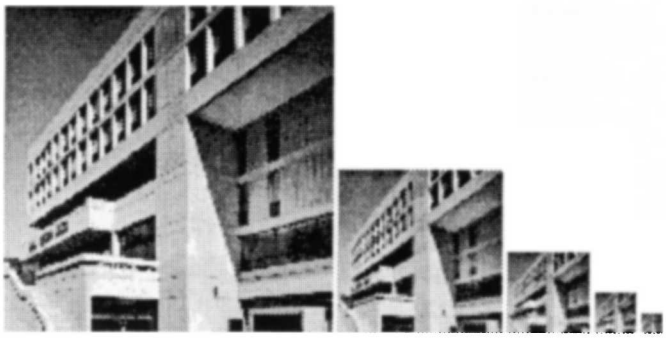

FIG. 1. A five-level Gaussian pyramid (original image size is 128 × 128).

metric kernel resembling a Gaussian probability function (see description in [1, 3]). Each image is therefore represented at several levels of resolution. For an input image of size 1024 × 1024, for example, reduced resolution images of sizes 512 × 512, 256 × 256, 128 × 128, 64 × 64, and 32 × 32 are generated. An example of a Gaussian pyramid is given in Fig. 1.

Laplacian pyramid is a set of multiresolution images, each one being the difference between successive levels in the Gaussian pyramid. The difference is performed after magnifying the smaller image to the size of the larger image. The Laplacian pyramid is a set of bandpass filtered copies of the input image, and the original image can be fully reconstructed from it. The Laplacian image pyramid (denoted L) is used to represent the spatial information in an image. An example of a Laplacian pyramid is shown in Fig. 2.

Difference of Laplacian energy (DoL) pyramid: Given two images in a sequence, a difference image, on a pixel by pixel basis, is computed. The Laplace operator ($\Delta^2 G$) is used to extract high frequency information from the difference image, which is then smoothed and squared to obtain the energy of temporal changes. The DoL pyramid is obtained by constructing a Gaussian pyramid from the

FIG. 2. A five-level Laplacian pyramid (original image size is 128 × 128).

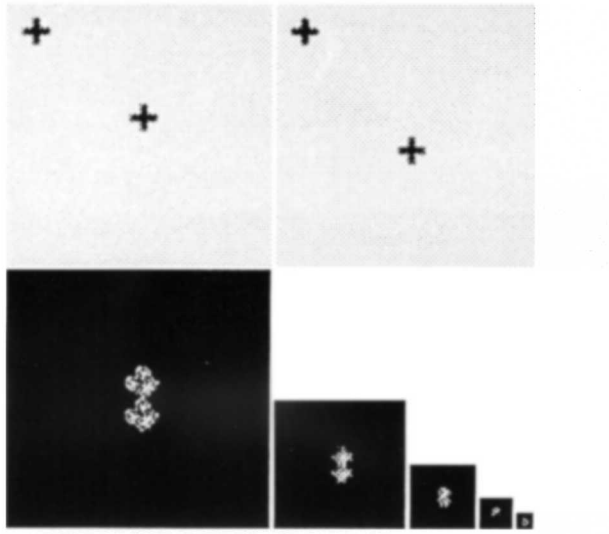

FIG. 3. A five-level DoL pyramid. The two original iamges are in the top row, where the right cross moved from one image to the next. The constructed DoL pyramid is in the bottom row, where only the moving object has any energy.

temporal change energy image. This DoL pyramid represents band-limited versions of the temporal changes between the two input images. The DoL pyramid (which is denoted by M) of two images in a sequence is used to represent the temporal change between the images. A DoL pyramid is displayed in Fig. 3.

Quad trees are hierarchical representations of an image, based on recursive subdivisions of the image array into quadrants [18, 17] as is shown in Fig. 4. Every node in the quad tree represents a square region in the image

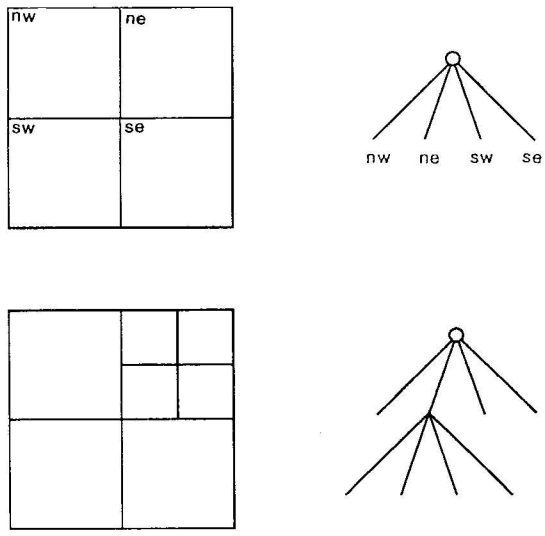

FIG. 4. Examples of quad trees. The tessellation is displayed on the left, and the tree representation is on the right.

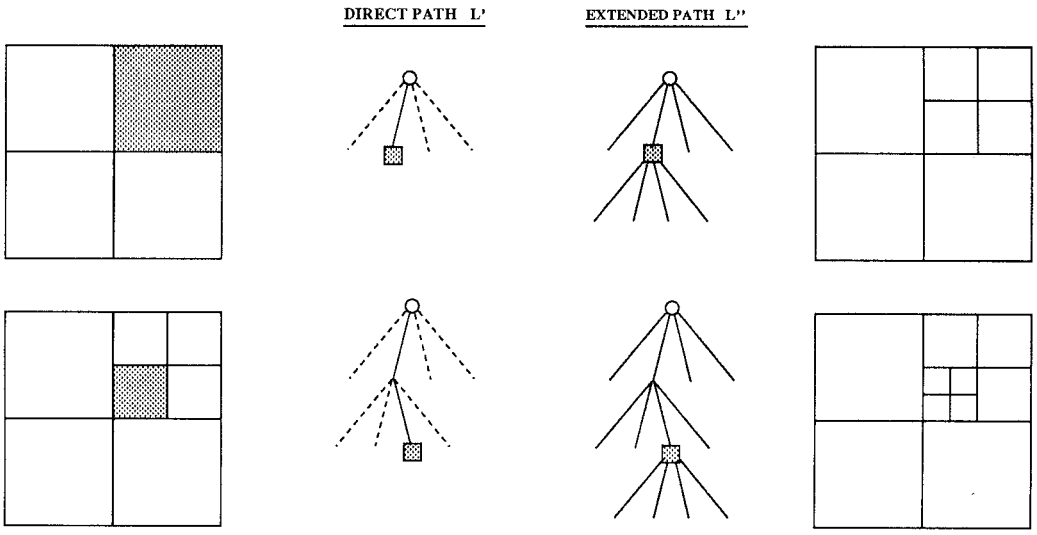

FIG. 5. Examples of direct and extended paths. The FOA nodes and their direct paths are on the left, shown by solid lines. The extended paths and their corresponding tessellations are on the right.

obtained by recursive quadration as follows: the root node of the tree corresponds to the entire image; the four nodes at the next level (if they exist) correspond to the four quadrants of the image and so on; the leaf nodes of the full quad tree correspond to single pixels in the image.

Given an image array and its representing quad tree, the *direct path* to the FOA node in the quad tree is defined as the branch in the quad tree from the root node to the FOA node. The *extended path* is defined as all nodes in the direct path with all their immediate son nodes. These definitions are illustrated in Fig. 5.

A sequence of quad tree paths can be used to traverse an image pyramid. This is done by defining the natural mapping between nodes in a quad tree and pixels in the image pyramid. The root node of the quad tree corresponds to the single pixel of the smallest image in the pyramid, the four quad tree nodes in the next level correspond to the four pixels in the next to highest pyramid level, etc. In this manner a path in the quad tree corresponds to a path in the pyramid.

4. THE ATTENTION FUNCTION

The *attention function* assigns to every pyramid node a value corresponding to the "attention catching" degree of that node, which conforms as much as possible with the features described in Section 2. Nodes with highest attention values are chosen as the FOA. (In Koch and Ullman [10], *saliency* is a spatially uniform version of the attention function.)

Let L'' be an extended path in the Laplacian pyramid L, and let M'' be the corresponding extended path in the DoL pyramid M. Let $s(P)$ denote the *level* of node P in the pyramid L, which is the number of nodes in the direct path from the root node to node P in the quad tree corresponding to L. Let $L[P]$ and $M[P]$ be the respective pyramid pixel values corresponding to node P.

We use the following attention function F for a node P:

$$F(P) = \lambda^L_{s(P)} \cdot L[P] + \lambda^M_{s(P)} \cdot M[P] - i(P),$$

where the first term is the contribution of spatial gradients in P, the second term is the contribution of temporal gradients in P, and the third term is an inhibition term.

$\lambda^L_{s(P)}$ and $\lambda^M_{s(P)}$ are normalization functions which equalize the influence of nodes at different pyramid levels. $\lambda^L_{s(P)}$ and $\lambda^M_{s(P)}$ are inversely proportional to $2^{s(P)}$, ensuring that nodes at levels closer to the root have greater "attention catching" values than nodes at lower levels. (Specifically we used $\lambda^L_{s(P)} = 128/(6 \cdot 2^{2|P|})$ and $\lambda^M_{s(P)} = 128/(2^{s|P|}$.)

The third term $i(P)$ is the inhibition function which serves to prevent the return of the FOA continuously to the same node and enables nodes of lower "attention catching" values to be attended to. When a node P has been transmitted at time t, but not at time $t + 1$, it is assigned an inhibition value. If P was a leaf node in the extended path when transmitted, its inhibition is set to

$$i(P) = C \cdot 2^{s(P)},$$

where C is a constant (we used $C = 2$). Otherwise, its inhibition value is the average of the inhibition values of its four son nodes. The inhibition is thus smaller for nodes closer to the root, preventing the FOA from being chosen continuously at low levels. Following this initial

setting, the inhibition value is divided by a constant after every transmission step, causing an exponential decay.

5. THE ATTENTIVE TRANSMISSION PROCESS

Attentive transmission proceeds between two processes, the *transmitter* and the *receiver*. The transmitter's input is a sequence of 2-D images. The aim is to use a limited transmission channel so that during the transmission the receiver can construct a representation of the input sequence as visually acceptable as possible.

5.1 *Transmission*

Transmission will be performed in discrete *transmission steps*. The current input image may be replaced by the successive image in the input sequence after a single or several transmission steps.

At every given moment, the transmitter has a representation of:

L, a Laplacian pyramid of the current image in the input sequence;

L',L'', a direct path and its corresponding extended path of the Laplacian pyramid L;

M, a DoL pyramid created from the current and previous input images. M is all zero if the input does not change;

M',M'', a direct path and its corresponding extended path of pyramid M. Nodes correspond to those of L',L'';

$\hat{L}$, an estimated reconstruction of pyramid L which is based on all previously transmitted information.

Following is the quad tree traversal for transmission using the "attention function" as described in the previous section.

1. Initialize the FOA and L' to be the root node (the top level of the pyramid).
2. Extend L' to L'' by including all sons of nodes in L'.
3. Visit all Laplacian pyramid nodes of the extended path L'' and transmit their values. Two cases are possible:

- Single images: In this case only the four new sons of the FOA node should be transmitted, as all other nodes in the path have already been transmitted.

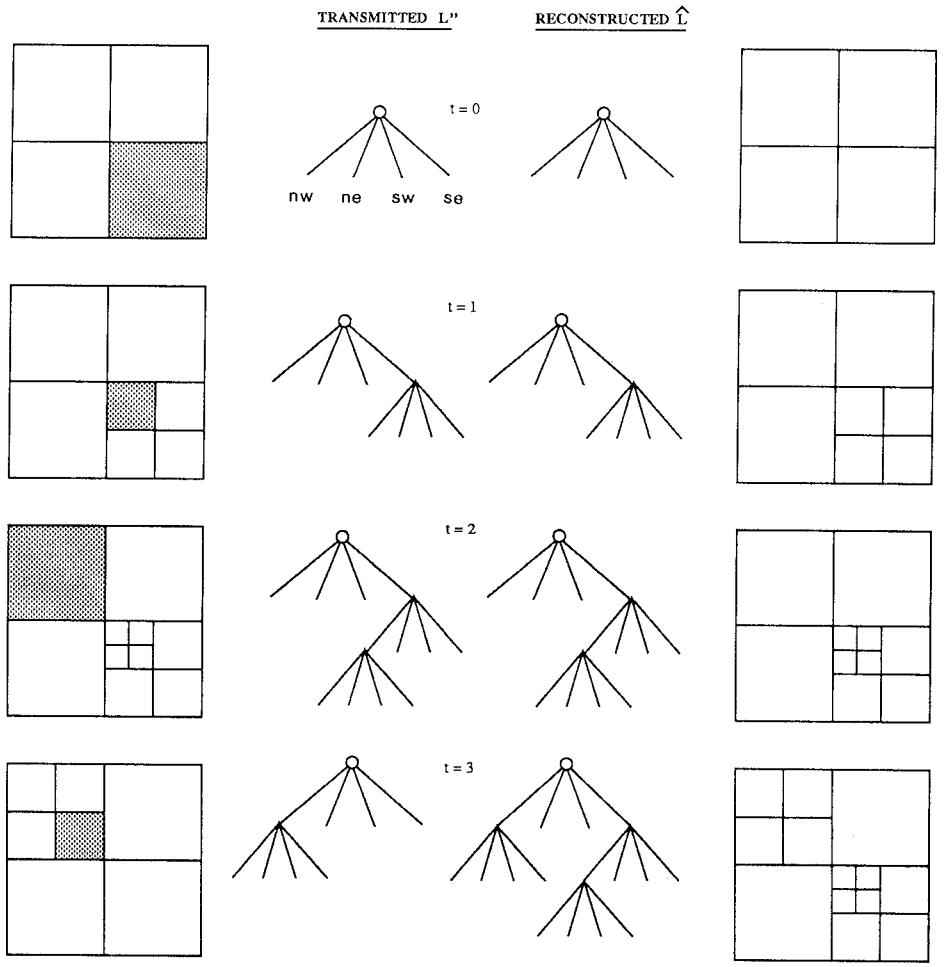

FIG. 6. Steps in the pyramid traversal algorithm. The selected FOA is depicted by the shaded area.

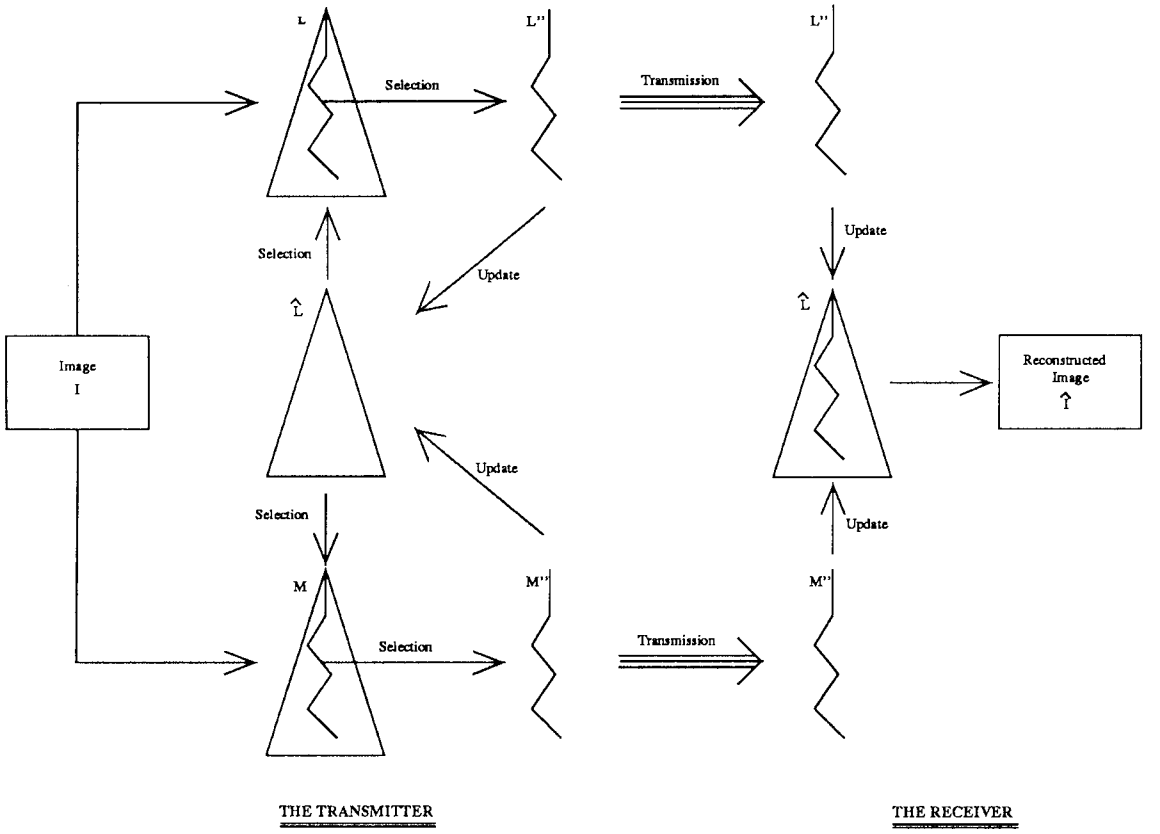

FIG. 7. The transmission process.

• Image sequences: In this case the DoL pyramid, having energy of temporal changes, is also being used. For all nodes in the extended path both values of the Gaussian and DoL pyramids are transmitted.

4. Select a single leaf node P of L'' having highest "attention value." Node P will be regarded as the new FOA.

5. L' will now be assigned the direct path to the selected node P.

6. Repeat from Step 2.

The nodes transmitted in Step 3 are used to reconstruct the Laplacian pyramid. Several steps in the pyramid traversal and reconstruction are shown in Fig. 6.

At every transmission step, the transmitter selects new paths L'' and M'' of the current pyramids L and M based on the current reconstruction $\hat{L}$. The receiver has the same reconstruction $\hat{L}$ of the pyramid L since it is based only on transmitted information. The receiver can therefore select the same paths L'' and M'' as the transmitter. Having selected the same paths, only their values (without their locations) need to be transmitted.

The pyramid values along the paths L'' (a path of the spatial pyramid) and M'' (a path of the temporal pyramid) are transmitted to the receiver. These values are used to update the reconstructed pyramid $\hat{L}$ in both the transmitter and the receiver. The process of transmitting and receiving is shown in Fig. 7.

5.2 *Reconstruction*

The updating of $\hat{L}$ with the newly transmitted L'' is performed as follows: For every node P in L'', if P does not exist in $\hat{L}$, add that node to $\hat{L}$ and set its value to $L[P]$. If node P exists in both L'' and $\hat{L}$, replace the value of $\hat{L}[P]$ by $L[P]$. If the difference of these two values exceeds a certain threshold, then eliminate all subtree nodes. This subtree elimination takes place when there is substantial change between frames, which makes the previous information about the specific region obsolete. The above threshold determines the sensitivity of the system to changes in the images. This updating method ensures that when a change in input occurs, only the area of change is erased from the reconstruction, and areas with little change are preserved. The receiver reconstructs the current input image from the reconstructed Laplacian pyramid $\hat{L}$ [4, 2] where any nonexisting node is set to zero value.

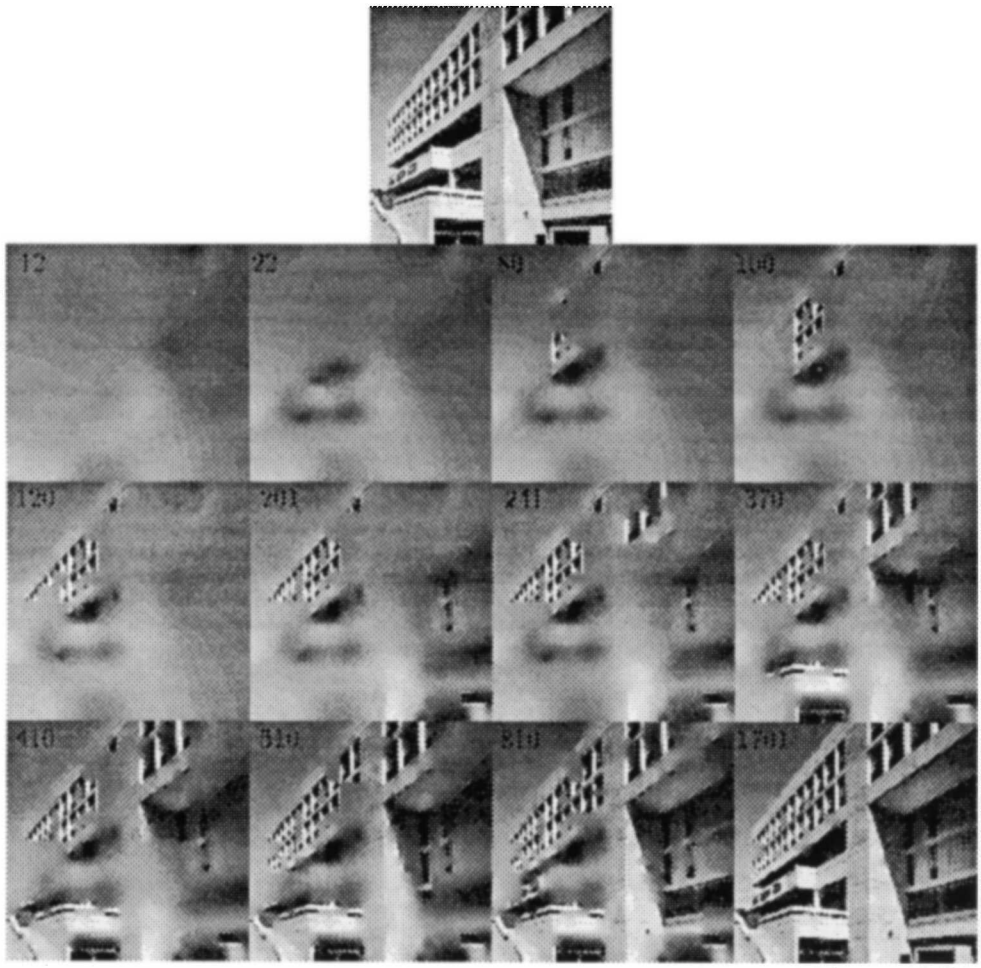

FIG. 8. Attentive transmission of a static image. Original (top) and several stages in the transmission process. Image size is 128 × 128. Numbers represent the transmission step.

6. EXAMPLES

6.1. *Static Images*

Figures 8 and 9 are examples of the reconstruction of static images using attentive transmission. It can be noted that:

- The early reconstructions include only low spatial frequency information, which give a general visual perception of image contents.
- Image elements are reconstructed in the order of their saliency as defined by the attention function. The face outline, the eyes, and the sharp boundary of the building are the first to be transmitted.
- Even when a salient feature is being transmitted, not all details are immediately collected. There is a tendency to collect some details from several salient features rather than all details from a single feature.
- In Fig. 8, the last transmission step at $t = 1701$ gives an average transmission rate of 2.9 bits/pixel. In Fig. 9 when $t = 387$ the transmission rate is 0.66 bit/pixel and at $t = 2089$ the transmission rate is 3.6 bits/pixel. Although these rates may not be optimal, it should be noted that attentive transmission is an incremental transmission and is advantageous when transmission time is nonuniform or unknown apriori.

6.2. *Image Sequence*

Figure 10 is an example of the transmission process of a sequence of images. In this example, temporal changes are created by moving the right cross. All the comments concerning static images are relevant for dynamic inputs. In addition, note the importance of sharp temporal changes at catching the FOA and the processing power. In more complex image sequences, there will be some trade-off between attraction of FOA by temporal changes and attraction by spatial changes.

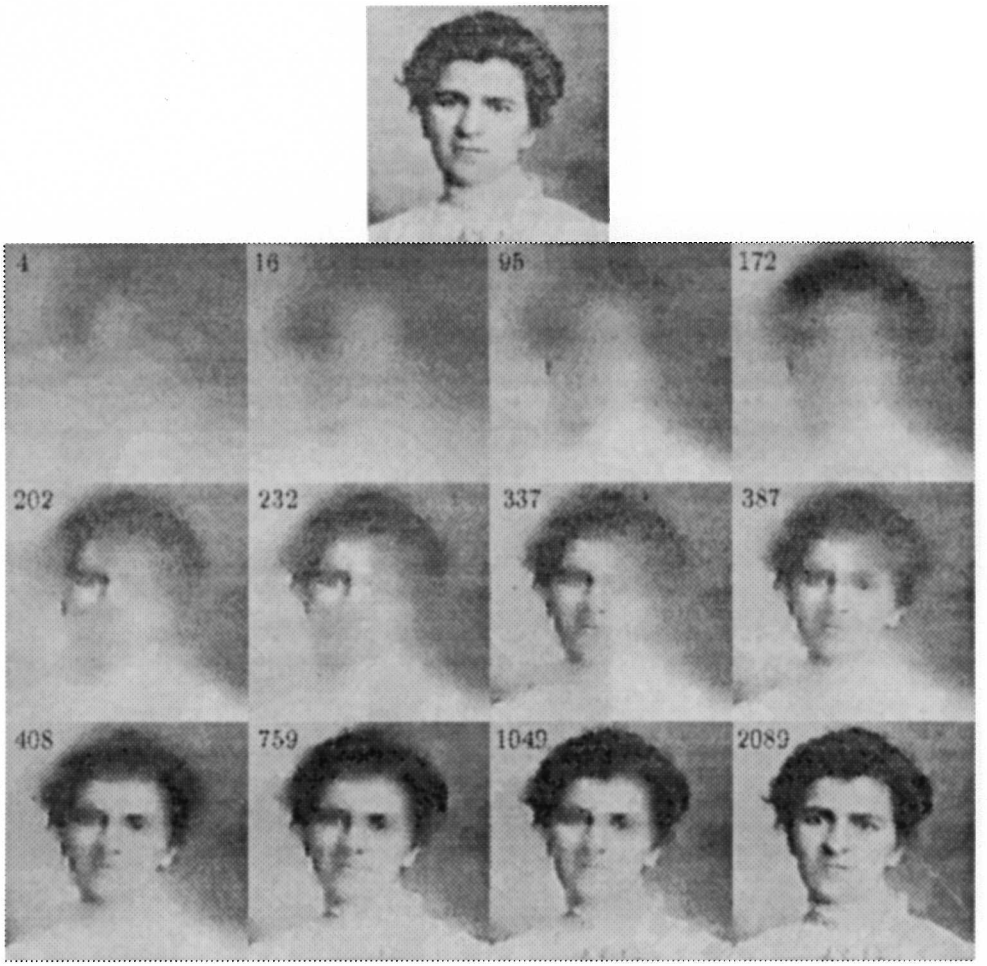

FIG. 9. Attentive transmission of a static image. Original (top) and several stages in the transmission process. Image size is 128 × 128. Numbers represent the transmission step.

7. CONCLUSION

In this correspondence we described a method for progressive transmission of digital images which uses a simple model of visual attention, namely attentive transmission.

In analogy to the focusing of visual attention on a limited spatial area, the method we propose concentrates processing power (data transmission) on a limited part of the image. This is done by transmitting sampling points which are nonuniformly distributed, dense in the center and sparse in the periphery (retinal-like). The center location and resolution of the samplings change according to the attention model.

Attentive transmission has the following advantages:

- Given that the limit on the number of sampling points to be transmitted is not known apriori, attentive transmission as a progressive transmission method makes an attempt so that at every step the reconstruction of the input images is as visually recognizable as possible. The initial transmitted data include low spatial frequency information giving general outlines of the image contents. As transmission progresses the samplings include information of higher spatial frequency, with more details.
- Concentrating sampling points on areas of temporal change in image encourages reconstruction of similar images in sequences, as in motion sequences, to take the form of updating in areas of change rather than of reconstructing the whole image from scratch.
- Using a specifically defined model of attention, and basing selection on previously transmitted data, both transmitter and receiver predict the same pattern of sampling points to be transmitted. This promises that only sampled values without their location need be transmitted.

The idea of attentive processing is promising not only in the area of image transmission. As a general paradigm for increasing efficiency, this idea could be incorporated

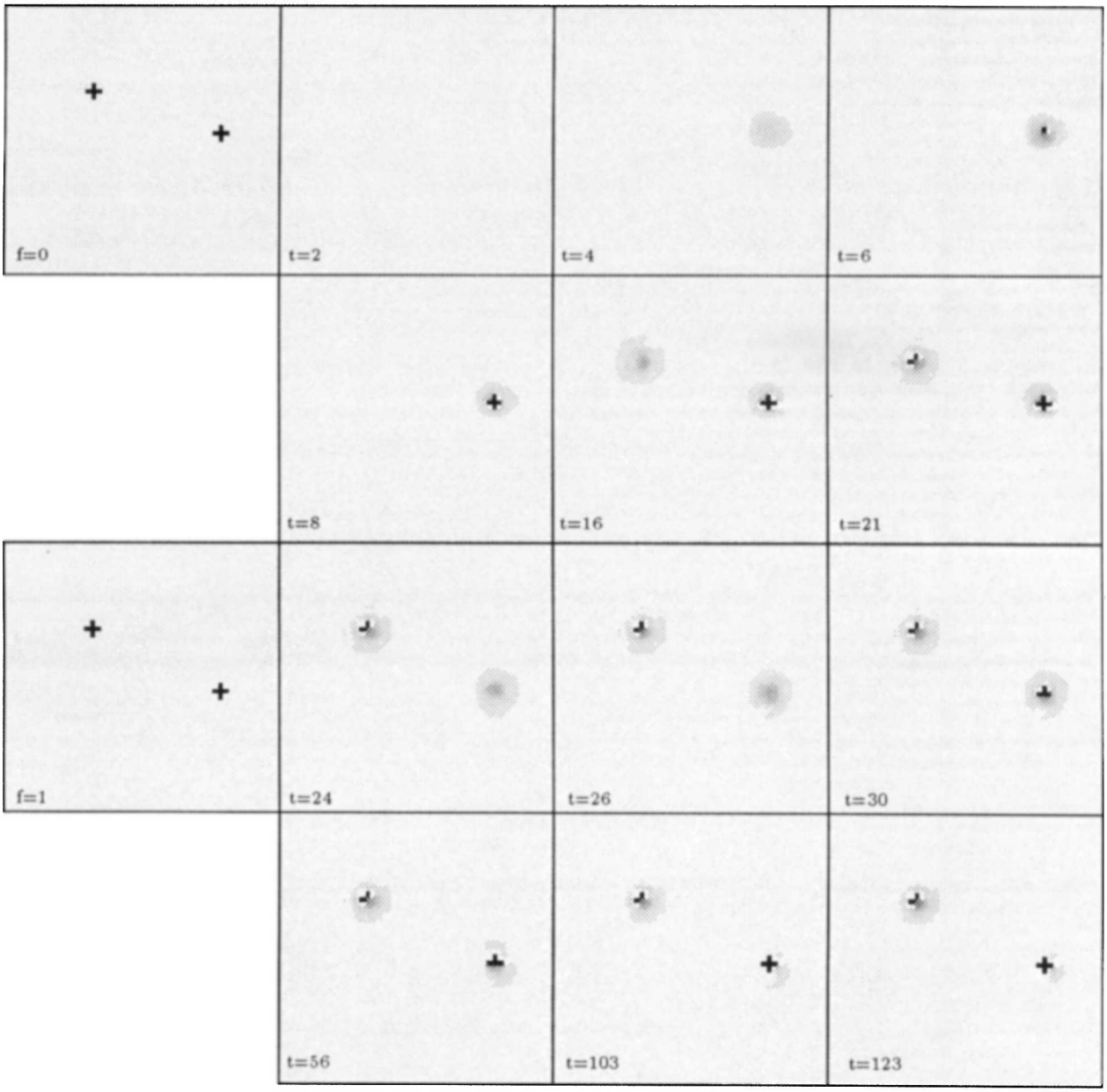

FIG. 10. Attentive transmission of an image sequence. Selected time steps are shown. Two frames of the image sequence to be transmitted (left) and the reconstruction by the receiver at six selected time steps. Original images are marked $f = 0, 1$, and reconstructions are marked by transmission steps $t = 2, 4, \ldots$. Input images change at $t = 0, 22$. The images are of size 128×128.

in other expensive (in terms of processing power) applications which are based on local processing.

REFERENCES

1. P. Burt and E. H. Adelson, The Laplacian pyramid as a compact image code, *IEEE Trans. Comm.* **COM-31,** 1983, 532–540.
2. P. J. Burt, Algorithms and architecture for smart sensing, In *Image Understanding Workshop,* DARPA, Cambridge, 1988.
3. P J. Burt, Fast filter transforms for image processing, *Comput. Graphics Image Process.* **16,** 1981, 20–51.
4. P. J. Burt, Smart sensing with a pyramid vision machine, *Proc. IEEE* **76,** 1988, 1006–1015.
5. F. L. Engel, Visual conspicuity and selective background interference in eccentric vision, *Vision Res.* **14,** 1974, 459–471.
6. F. L. Engel, Visual conspicuity directed attention and retinal locus, *Vision Res.* **11,** 1971, 563–576.
7. C. W. Eriksen and J. D. St. James, Visual attention within and around the field of focal attention: A zoom lens model, *Perception Psychophys.* **40,**4, 1986, 225–240.
8. G. C. Grindley and V. Townsend, Voluntary attention in peripheral vision and its effects on acuity and differential thresholds, *Quart. J. Exp. Psychol.* **20,** 1968, 11–19.

9. R. Klein, Inhibitory tagging system facilitates visual search, *Nature* **334,**4, 1988, 430–431.
10. C. Koch and S. Ullman, Shifts in selective visual attention: Towards the underlying neural circuitry, *Human Neurobiol.* **4,** 1985, 219–227.
11. J. Moran and R. Desimone, Selective attention gates visual processing in the extrastriate cortex, *Science* **229,** 1985, 782–784.
12. S. Peleg, O. Federbush, and R. Hummel, Custom made pyramids, In *Parallel Computer Vision,* (L. Uhr, Ed.), Academic Press, New York, 1978, 125–146.
13. M. I. Posner and Y. Cohen, Control of language processes, In *Attention and Performance,* (H. Bouma and D. Bouwhuis, Eds.), Erlbaum, Hillsdale, NJ, 1984.
14. M. I. Posner and S. E. Peterson, The attention system of the human brain, *Annual Rev. Neurosci.* **13,** 1990, 25–42.
15. H. Rom and S. Peleg, Image representation using Voronoi tesselation: Adaptive and secure, In *Proceedings of the CVPR, Computer Vision and Pattern Recognition,* Ann Arbor, MI, June 1988, 282–285.
16. D. Sagi and B. Julesz, Enhanced detection in the aperture of focal attention during simple discrimination tasks, *Nature* **321,** 1986, 693–695.
17. H. Samet, Hierarchical representations of small rectangles, *ACM Comput. Surveys* **20,**4, Dec. 1988, 271–309.
18. H. Samet, The quadtree and related hierarchical data structures, *ACM Comput. Surveys* **16,**2, June 1984, 187–260.
19. C. P. Wikens, The effects of divided attention on information processing in manual tracking, *J. Exp. Psychol.* **105,** 1978, 1–17.
20. A. L. Yarbus, *Eye Movements and Vision,* Plenum, New York, 1967.
21. Y. Yeshurun and E. L. Schwartz, Shape description with a space-variant sensor: Algorithms for scan-path, fusion and convergence over multiple scans, *IEEE Trans. Pattern Anal. Mach. Intelligence* **11,**11, Nov, 1989, 1217–1222.