Perceptual-components architecture for digital video

Andrew B. Watson

NASA Ames Research Center, Moffett Field, California 94035-1000

Received February 13, 1990; accepted June 9, 1990

A perceptual-components architecture for digital video partitions the image stream into signal components in a manner analogous to that used in the human visual system. These components consist of achromatic and opponent color channels, divided into static and motion channels, further divided into bands of particular spatial frequency and orientation. Bits are allocated to an individual band in accord with visual sensitivity to that band and in accord with the properties of visual masking. This architecture is argued to have desirable features such as efficiency, error tolerance, scalability, device independence, and extensibility.

INTRODUCTION

At this time there is great interest in new and extended standards and architectures for electronic transmission of visual information. These include extensions of conventional analog broadcast TV (HDTV), digital packet-switched video, and so-called open architecture TV.[1] Packet video is a scheme in which the video signal is digitized and broken down into small (100–1000 bits) packets for transmission over an asynchronous packet-switched network. An excellent introduction to the current status of the subject is available in a recent journal special issue.[2] The principal virtues of packet video are that (1) it permits integration of diverse information sources (video, speech, data), (2) it exploits the variable-bit-rate network to provide constant quality, despite the bursty nature of video signals, and (3) it permits simple multiplexing of multiple video sources, which in turn yields improved channel utilization and transmission efficiency.

Whatever the details of its implementation, packet video will require that the image stream be coded in an efficient and robust manner. The purpose of this paper is to argue that this code should be designed to match the perceptual apparatus of the human viewer, because this approach leads naturally to desirable attributes such as efficiency, error tolerance, scalability, extensibility, and device independence.

The plan of the paper is as follows. I begin with a description of a general method of coding the image stream in a manner that mimics the coding employed in the early stages of the human visual system. I call this a perceptual-components architecture (PCA). In the second section I describe the advantages of a PCA in a packet-video environment. This paper does not describe a completed project but rather suggests a profitable direction for further research and development.

PERCEPTUAL-COMPONENTS ARCHITECTURE

While the deeper mysteries of human vision remain unsolved, two centuries of research have yielded a picture, in some detail, of the early stages of the visual process. In recent years the picture has undoubtedly been clarified by the concurrent development of computer and imaging technology. These kinds of technology have provided metaphors, mathematics, and algorithms with which to understand and describe biological vision. Consequently much of early vision can be cast in signal-processing terms, such as filtering, sampling, and coding. More specifically, early vision has been characterized as a branching stream, in which separate modules divert from the common flood the information of their specific concern, such as color, motion, and shape. This partition of information into separate components is key to the signal architecture described below. I begin by specifying some useful terminology. I then propose a partition of the visual signal into various components. Each aspect of the partition is motivated by a brief review of relevant evidence. My purpose is to illustrate the general features of this architecture rather than to recommend detailed specifications.

Terminology

I characterize video input and output as a digital image stream, $a(x, y, t, c)$, where the indices refer to space, time, and color. For this discussion we ignore the analog processes of capture and display. The input is to be coded, packed, transmitted, received, unpacked, and decoded into an output stream. We are concerned primarily with the design of the code, although this cannot be entirely divorced from the issues of categorization into packets. The goal is to code the image stream in an efficient and robust manner. The general type of coding proposed here is usually called transform coding. A linear transform is applied to the image, yielding a set of coefficients, describing the amplitudes of the transform basis functions. The coefficients are quantized and subjected to some lossless coding. At the receiver, each step is inverted.

The linear transform that we consider here can be viewed as a set of transforms, each of which consists of a filtering and a subsampling operation. Each individual transform is associated with a filter, which selects a particular band, or region of the spatiotemporal frequency domain, and a particular direction in the trivariant color space. Each filter is associated with an image component, which is its inverse Fourier transform and which represents the elementary signal selected by the filter. Each coefficient, when inverse quantized and inverse transformed (rendered), will result in

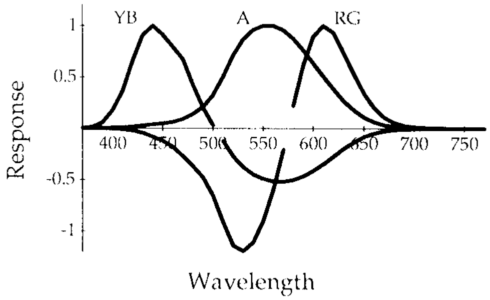


Fig. 1. Spectral responses of three opponent color channels as described by Guth *et al.*[7] The PCA first partitions the image stream into three comparable chromatic bands.

the addition of a suitably amplified and displaced component to the output stream. For purposes of exposition, a group of bands that share some aspect may be referred to as a channel.

Color

The current model of human visual color processing begins with image capture by three distinct cone types, colloquially denoted R, G, and B. The cone signal may be computed by a linear transformation from monitor or camera primaries, r, g, and b. Capture (and some adaptive gain control) is followed by a linear transformation into a so-called opponent scheme[3–5] comprising an achromatic channel, a channel representing red–green differences, and a channel representing yellow–blue differences. To a first approximation, these channels may be computed as R + G, R − G, and R + G − B. We do not attempt here to specify the precise color directions of these three channels but note that this is an area of intense activity.[6–10] Figure 1 illustrates the spectral response functions for one current opponent model.[7] Accordingly, in the PCA the first partition of the image stream is into three opponent chromatic channels, which we designate achromatic (A), red–green (RG), and yellow–blue (YB). For the moment, we treat this partition as separable from the spatial and temporal content of the signal, but as we shall see these three bands undergo rather different subsequent processing.

Space

The cones perform a spatial sampling of the image, and at the retina and the lateral geniculate nucleus these samples are transformed by a difference-of-Gaussians type of operator, which appears to serve primarily in adaptive gain control, signal decorrelation, and multiplexing of chromatic channels.[11] At the level of the primary visual cortex, electrophysiological measurements show that individual cells are tuned for specific bands of spatial frequency and orientation.[12–19] Likewise many psychophysical results are consistent with the existence of channels selective along these dimensions.[20–26]

In primate visual cortex, individual neurons have an average spatial frequency bandwidth of ~1.4 octaves, and an orientation bandwidth of ~40 deg,[13,27] suggesting perhaps four to eight frequency bands and approximately as many orientation bands.

In the PCA the spatial frequency dimension is subdivided into a number of bands of spatial frequency or resolution. These are depicted in Fig. 2 by circular concentric bands in the two-dimensional spatial frequency domain. The width of each band (difference between inner and outer diameters) is proportional to its center frequency (midpoint between inner and outer diameters). This reflects the approximately constant logarithmic bandwidth of cortical neurons and also corresponds to a self-similarity among the resolution bands.

Each resolution band is further divided into oriented components, represented by the wedge-shaped regions separated by straight lines in Fig. 2. Since the spatial image is real, each frequency-orientation band actually consists of a pair of segments on either side of the origin. As noted above, to each band there corresponds an image component, which may be regarded as the elementary signals into which the image stream is decomposed. An example component corresponding to one of the bands in Fig. 2 is pictured in Fig. 3. It has a certain spatial frequency and orientation and is localized in both space and frequency.

I emphasize that these illustrations are only schematics of the partition of the image stream into perceptual components. The illustration in Fig. 2, for example, does not show that spectral bands of the various components may overlap and that their borders may be gradual rather than sharp. Many aspects of the partition, such as the precise width, number, and shape of resolution and orientation bands, are subjects for research and debate.

Since we intend that the PCA be device independent, we must specify a unit for the spatial dimension that is independent of device-specific quantities such as pixels and picture heights. A unit such as centimeters is unsatisfactory because, independently of viewing distance, it cannot be related to human visual acuity. The unit adopted in the PCA is degrees of visual angle. This has the drawback that, in order

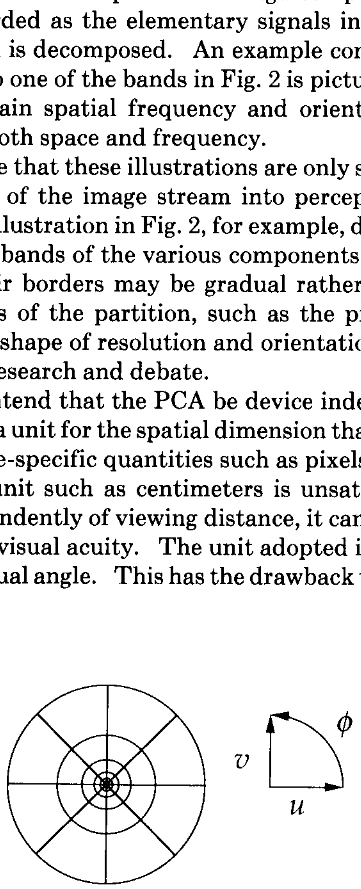


Fig. 2. Subdivision of spatial frequency domain in a PCA. The axes are horizontal (u) and vertical (v) spatial frequency. Orientation (ϕ) is given by angle relative to the u axis.

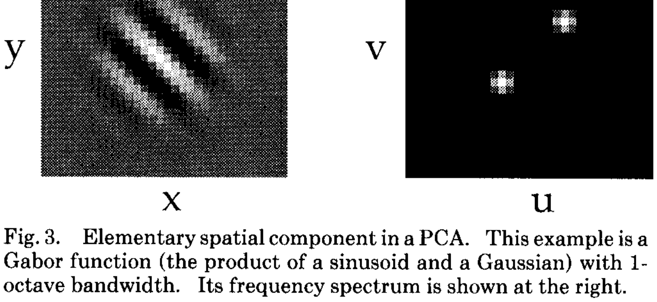
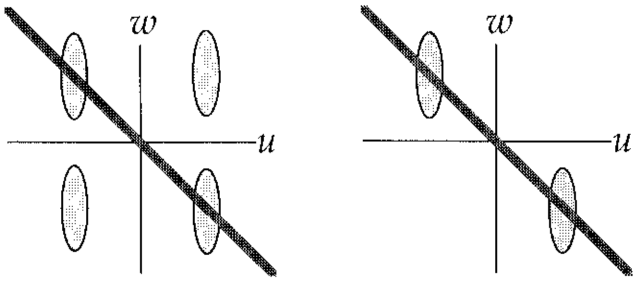


Fig. 3. Elementary spatial component in a PCA. This example is a Gabor function (the product of a sinusoid and a Gaussian) with 1-octave bandwidth. Its frequency spectrum is shown at the right.



Separable spatiotemporal band Rightward motion band

Fig. 4. Comparison of separable spatiotemporal band and motion band. In each figure the diagonal line indicates the spectrum of an image in rightward motion with unit velocity. The velocity is equal to the slope of the line in this frequency diagram. Separable real components (left) would occupy all four shaded regions, whereas an inseparable motion component (right) occupies only two of the four regions. The axes represent temporal frequency (w) and one spatial frequency dimension (u).

actually to render the image, one must assume a particular viewing distance from the display surface. But a design viewing distance is always implicit in displays designed for human viewing.

Time

There is considerable psychophysical evidence for two separate visual channels selective for low and high temporal frequencies.[28–31] These have often been called sustained and transient, on account of their differing impulse responses, but I will call them static and motion channels. These channels are probably related to the so-called M and P pathways in the primate retina, geniculate, and cortex, but this has yet to be established with confidence.[32]

Applying this notion to the PCA, we divide the image stream into high- and low-pass temporal bands. The low-pass band represents the essentially stationary components of the stream, while the high-pass band represents moving elements, whence the names static and motion channels.

Space and Time (Motion)

When one is using real components, there are essentially two ways to represent a band of spatiotemporal frequency (Fig. 4). In one, separable space–time signals (standing waves) are used, whereas in the other, inseparable motion signals (traveling waves) are used. Various psychophysical and electrophysiological results suggest that, for the high-pass temporal channel, the visual system uses direction-selective filters and therefore codes inseparable motion signals.[27,29,31,33–36] Specific forms for these human motion filters have been proposed.[37–42]

In accord with these observations, in the PCA each band in the motion channel consists of a pair of regions arranged diametrically about the spatiotemporal frequency origin. An example is pictured in Fig. 5. The space–time component to which the region in this example corresponds is a small patch of sinusoid oriented and moving at an angle of 45 deg.

It is well established that the human observer is insensitive to high spatiotemporal frequencies.[43] This may be a consequence of the low-pass temporal character of the parvocellular-driven cortical pathways.[44] For this reason, in our architecture the motion channel contains only low to medium spatial resolution bands. The notion that video codes may discard high spatiotemporal components has a considerable history[45,46] and is part of the justification for the use of interlacing in current video standards. However, it has also been argued that the smooth-pursuit eye movements that track moving images reduce this advantage.[47,48] However, in those experiments, the entire image moved predictably, making tracking almost certain, so further studies with more natural sequences are warranted.

Space and Color

Although there is debate about precisely how to quantify the difference, there is general agreement that spatial resolution is markedly lower in the chromatic channels than in the achromatic channel.[49–55] This fact is of course exploited in current TV broadcast systems. In the PCA it permits us to omit the uppermost spatial resolution bands from the two chromatic channels.

Space, Time, and Color

There is also evidence that the chromatic pathways contribute little to the motion sense.[56] Likewise, chromatic temporal sensitivity is considerably lower than achromatic.[57] Consequently we omit the motion channel from the two chromatic channels, RG and YB.

An additional difference between achromatic and chromatic channels is in the treatment of low-pass components. Achromatic sensitivity declines markedly at low spatial and temporal frequencies,[43] whereas chromatic sensitivity remains high.[52,54,57] It may therefore be possible to discard the static spatially low-pass achromatic component altogether, whereas the component must be retained in the chromatic channels. However, the information needed to represent the static spatial low-pass component is small, and it may be essential to images with intense low spatial frequencies; so, unless there is a need, it may be wiser to preserve that component.

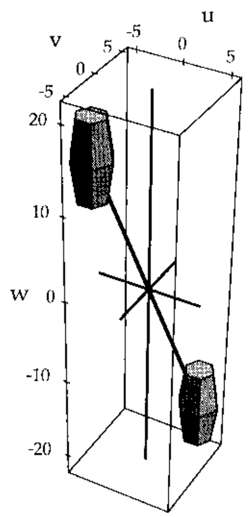


Fig. 5. Single motion band. The line connecting the two lobes is only a visual guide. The axes indicate spatial frequencies (u and v) in cycles/degree and temporal frequency (w) in hertz.

Summary

The PCA partition of the image stream into spatial, temporal, and chromatic bands is summarized in Fig. 6. Each panel is a representation of the spatiotemporal frequency domain for one opponent color channel. Only the positive half of the temporal frequency domain is shown, since for real signals the negative half is redundant. Each spheroid represents a particular band. Together the bands make up something like a wedding cake, which may be divided into layers (static versus motion), slices (various orientations), and concentric rings (various resolutions or spatial frequencies).

Several features of the PCA introduced in the previous subsections are evident in this figure. First, only the achromatic channel contains a motion channel, or second layer to the cake. Second, this motion layer does not extend to as high spatial frequencies as does the static layer. Third, both red–green and yellow–blue channels contain only lower-resolution bands.

We have not spelled out the precise algorithm needed to implement this architecture, and we note that there are many technical challenges in the design of efficient implementation. The general scheme proposed, however, is consistent with recent developments in image coding. First we note the development of pyramid image codes,[58–60] which subdivide the image by resolution and subsample each band in proportion to resolution. More recently, oriented pyramids have been developed, and several of these codes have been designed as analogs to image coding in human visual cortex.[61–66] The transforms used in these codes are quite similar to the finite prolate-spheroidal wave forms explored by Wilson[67] and Wilson and Spann[68] and to the wavelet codes studied by Mallat.[69] Subband codes, an area of intense activity, are a close relative of pyramid and wavelet codes, which use quadrature mirror filter techniques to ensure lossless coding.[70,71] Typical subband codes differ from the proposed architecture in that they employ three bands: horizontal, vertical, and diagonals. The diagonal band contains components at two orthogonal orientations. The more recently devised hexagonal subband codes[61,62,64,65,72] are more in keeping with the PCA. The subband concept has also been extended to the time dimension,[72,73] and a number of authors have explored the use of subband codes in packet video.[74,75]

Sampling

Because the PCA outlined above adheres to the general form of hierarchical, subband, or pyramid coding, it is amenable to the sampling principles employed in those techniques. Typically sampling is proportional to two-dimensional bandwidth for spatial imagery or to three-dimensional bandwidth for an image stream. The total number of samples (coefficients) in the code of a given segment of the stream should be near the number of pixels in the input. In the example pictured in Fig. 6, with three resolution bands each an octave wide, spatial samples would be 16 times more numerous in the highest-resolution band than in the lowest. Likewise each motion band, which we envision as having twice the temporal bandwidth of the static band, would contain twice as many samples as the static band at the corresponding spatial frequency.

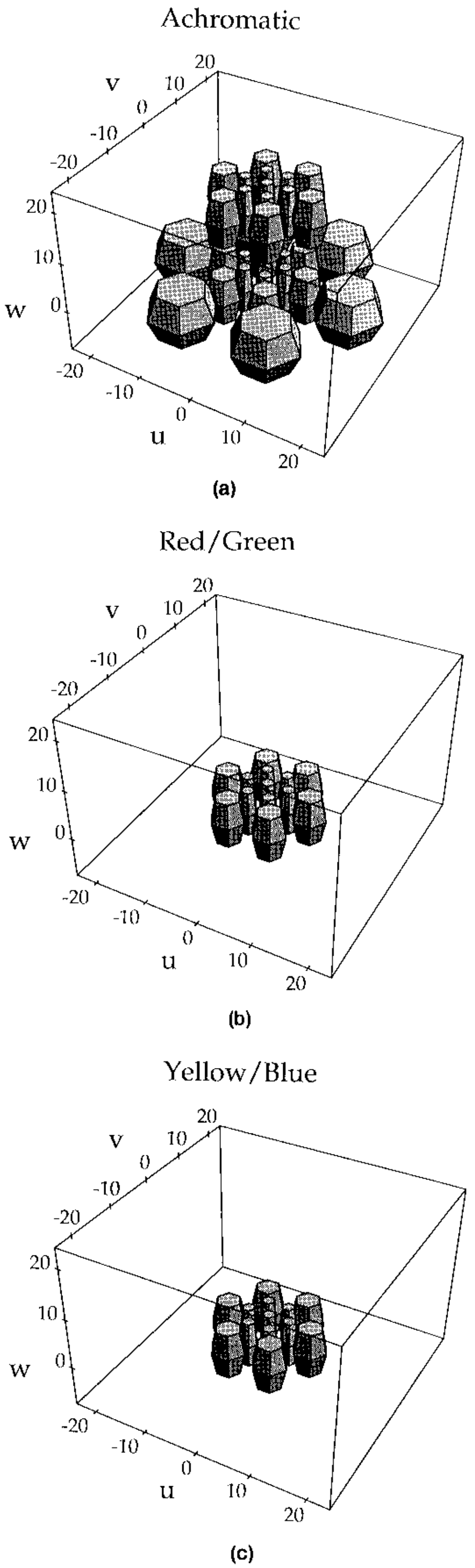


Fig. 6. Subdivision of spatial, temporal, and chromatic domains in the PCA. The axes are labeled in hertz (w) and cycles/degree (u and w). Each panel is for a separate chromatic channel.

Quantization

In traditional image coding, bit allocation is typically based on a constraint such as minimization of rms error. In the PCA, bit allocation is based on the visibility of quantization errors. This has two consequences. The first is that quantization levels are allocated to a particular band in proportion to visual sensitivity to that band. In particular, fewer levels are required for high spatial frequencies than for low. The precise number of levels required for the motion and chromatic bands can also be predicted with some accuracy from the literature on visual sensitivity, but the most accurate measures would be obtained for signals closely matched to the individual components.

The second consequence is that visual masking, wherein quantization errors become less visible when superimposed upon large signals, calls for the use of a nonuniform quantizer. An example of how visual sensitivity can be used directly to construct a quantizer that in principle eliminates visible quantization errors is pictured in Fig. 7.[76,77] In general, such a quantizer will allocate fewer levels to larger transform coefficient values.

Such a quantizer can have a large effect on the number of required levels. The masking exponent (w in Fig. 7) describes the power-law rise in detectable contrast increment as a function of component contrast. Figure 8 shows how various masking exponents yield various numbers of levels. A typical value of the exponent derived from psychophysical experiment is 0.7, which in this example yields 11 levels, compared with 101 required for a uniform quantizer.

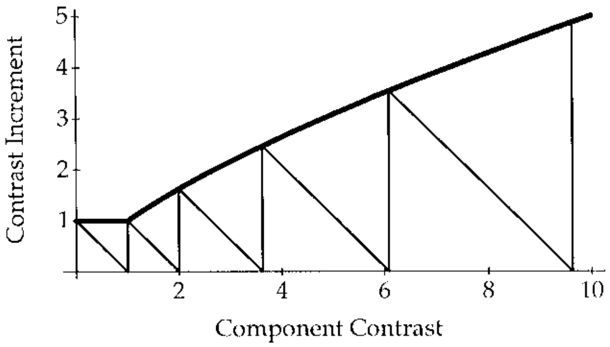


Fig. 7. Design of a perceptually lossless quantizer. The heavy line describes the just-detectable increment in the contrast of a component as a function of the contrast of the component: $\delta(c, t, w) = t\ \mathrm{Max}[1, (c/t)^{w}]$, where c is component contrast, t is contrast threshold, and w is the masking exponent. Here contrasts are expressed in units of t, and the masking exponent is 0.7. To ensure that quantization errors are invisible, successive thresholds and levels are placed at the corners of abuting triangles whose height (and base) equals the permissible quantization error. See Ref. 77 for more details.

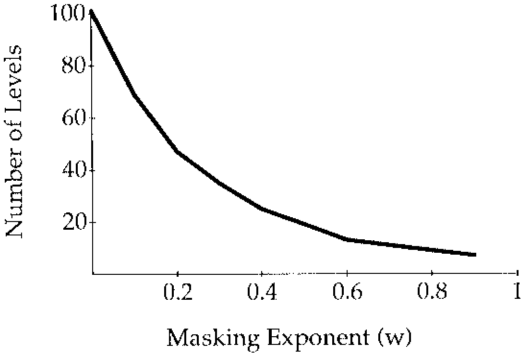


Fig. 8. Number of quantization levels as a function of the masking exponent w. In this example, the input range is {0, 100}, and the threshold is 1. The case of exponent = 0 corresponds to a uniform quantizer, which in this example would require 101 levels.

Unknowns

Although the broad outline of a PCA is clear, there are many questions that remain to be answered. Among these we note the following: What are the contrast thresholds in the various bands? What are the masking exponents in various bands? What are the spatial bandwidths of each band?[78] Do they vary between static and motion channels[30] or between achromatic and chromatic channels? What is the orientation bandwidth of each band?[79,80] Does it vary between static and motion channels[81] or between achromatic and chromatic channels?[82-84] What are the temporal bandwidths of static and motion channels?[30,85] What are the precise directions in color space of the three chromatic channels?[6-10]

Apart from the quantization process, the architecture described above is essentially linear. There are, however, a number of strong nonlinearities in early vision that may provide profitable opportunities for additional compression. These include adaptation to local mean luminance, contrast adaptation,[86] cross-orientation inhibition,[87,88] and contrast gain control[89] or contrast normalization.[90]

There are in addition many questions regarding the specific implementation of a PCA, such as the design of efficient filters and sampling patterns, lossless coding algorithms, and schemes for categorization into packets. Finally, the ultimate value of this approach must be assessed by evaluation of complete prototype systems.

An Example

As an example of a scheme that comes close to the spatial and chromatic subdivision illustrated in Fig. 6, I briefly describe the hexagonal orthogonal pyramid (HOP) transform.[62,64,65] Figure 9 shows a set of seven transform kernels. Each high-pass kernel (outer ring in Fig. 9) corresponds to one of the bands in the outermost ring of Fig. 2 or 6. Each kernel is applied to each nonoverlapping hexagonal neighborhood of seven image pixels on a hexagonal image raster, yielding one coefficient per neighborhood per kernel. Thus each kernel effectively subsamples the image by a factor of 7 (in two dimensions). The center, low-pass, kernel in Fig. 9 generates an image that is again subdivided by reapplication of the seven kernels, to generate the next-lower-resolution set of bands. The process is repeated to a desired number of levels. This particular code is lossless, and each individual component is orthogonal to all others. The number of coefficients is equal to the number of input pixels.

The original {r, g, b} image is linearly transformed to an opponent color space {A, RG, YB}, and the HOP transform is applied separately to each band. The nonuniform quantizer described above is applied separately to each resolution band in each chromatic image. The contrast threshold pa-

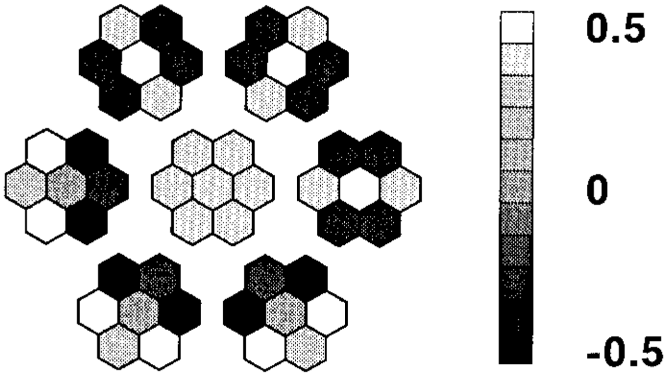


Fig. 9. Transform kernels for the HOP transform. The seven orthogonal kernels consist of three edge detectors, three bar detectors, and a blob detector.

rameter of this quantizer is adjusted to minimize bit rate while retaining acceptable visual quality. Each chromatic channel is then inverse HOP transformed, and inverse color transformed. Plate I shows the three separate chromatic channels and the combined results. The bit rates for the three bands are 0.86, 0.041, and 0.054; and for the complete image, 0.96 bit/pixel.

ADVANTAGES OF A PERCEPTUAL-COMPONENTS ARCHITECTURE

Compression

Compression of the image stream is an essential feature of packet video. Without any form of compression, digital transmission of broadcast quality video would require a bit rate of the order of 175 Mbits/sec (512 pixels × 480 lines × 30 frames × 8 bits × 3 colors), excluding any packet overhead.

It is well known that certain components of the image signal convey little information to the observer, either because little information is present or because the observer is insensitive to those components. These components should clearly be transmitted at a lower bit rate, with lower priority, or not at all. Examples are high spatiotemporal frequencies and high spatiochromatic frequencies. Again, separation into perceptual components makes this information triage a relatively simple matter.

Perceptual components coding has shown promise in compressing static monochrome and color images.[59,62,65,66,77,91,92] Perceptually lossless coding is approached at bit rates of approximately 1 bit/pixel for color images, representing a compression factor of 24 relative to uncompressed video. This figure approaches that of more established techniques, and, given the preliminary nature of this work, it is likely that this figure can be substantially improved.

Motion

Motion is arguably the most challenging aspect of coding and rendering the image stream because it generally leads to a high and variable bit rate. However, in those portions of the image that lack motion, successive frames are highly correlated, and the bit rate is low. In portions in which motion does occur, there is a high correlation between appropriately offset frames. Schemes such as conditional replenishment and motion compensation have been devised to exploit these redundancies.[93–95]

As part of a PCA, motion components provide several advantages. First, they provide a form of automatic motion compensation. The coefficients in the motion band, when rendered into the output stream, are themselves moving image components. Thus, rather than displacing large blocks of the image, as is conventionally done, with typical problems at block borders, this technique moves elementary visual components in a smooth and continuous way.

Second, motion components provide a form of conditional replenishment. While the image is in motion, the static bands are silent, and only half of the motion bands are active (those within 90 deg of the direction of motion). Because motion bands are absent from the two chromatic channels, as well as at the higher spatial frequencies of the achromatic channel, motion components can be conveyed with few coefficients and consequently a low bit rate. When the image is stationary, motion channels are silent. The static channels are active, but, because they are quite low pass in time, samples are infrequent, and again a low bit rate results. A still lower bit rate can be obtained if coefficients are transmitted only when they change.

Perceptual-components coding has not yet been applied to the temporal dimension. But, given the considerable temporal redundancy of video imagery, it is likely to provide a large amount of additional compression over that obtained with static images. Subband coding has been applied in the time domain, with apparent success, but published results do not quantify the advantage gained.[73,74] As noted above, PCA uses motion components, whereas subband coding typically uses separable space–time components (Fig. 5). When separable space–time components are used, the band mixes two opposing directions, only one of which typically contains useful information. Thus I suspect that use of motion components will provide better compression than will the use of separable components.

It has long been argued that moving scenes do not require high spatial resolution, and that stationary scenes do not require high temporal resolution, and that this permits bandwidth reduction.[45,46,96,97] This feature is implemented in the PCA in the absence of high-spatial-frequency bands in the motion channel. Furthermore each static band has half the bandwidth of the motion band at the same resolution and orientation, and there are twice as many motion bands as static bands at a given resolution (each static band has orientation but no direction), so that at a given resolution motion bands require four times as many samples.

A further virtue of the segregation of static and motion components is that only static components are required in the two chromatic channels, as discussed further below.

Color Coding

The PCA partitions the image stream into three channels: luminance or white–black (WB), red–green (RG), and yellow–blue (YB). It is widely appreciated that chromatic information is coded at much lower resolution than the luminance information, and indeed this is the basis for the bandwidth allocations of luminance and color signals in conventional broadcast TV. Image coding experiments show the remarkable difference in bit-rate requirements of luminance and color. The example in Plate I shows that an

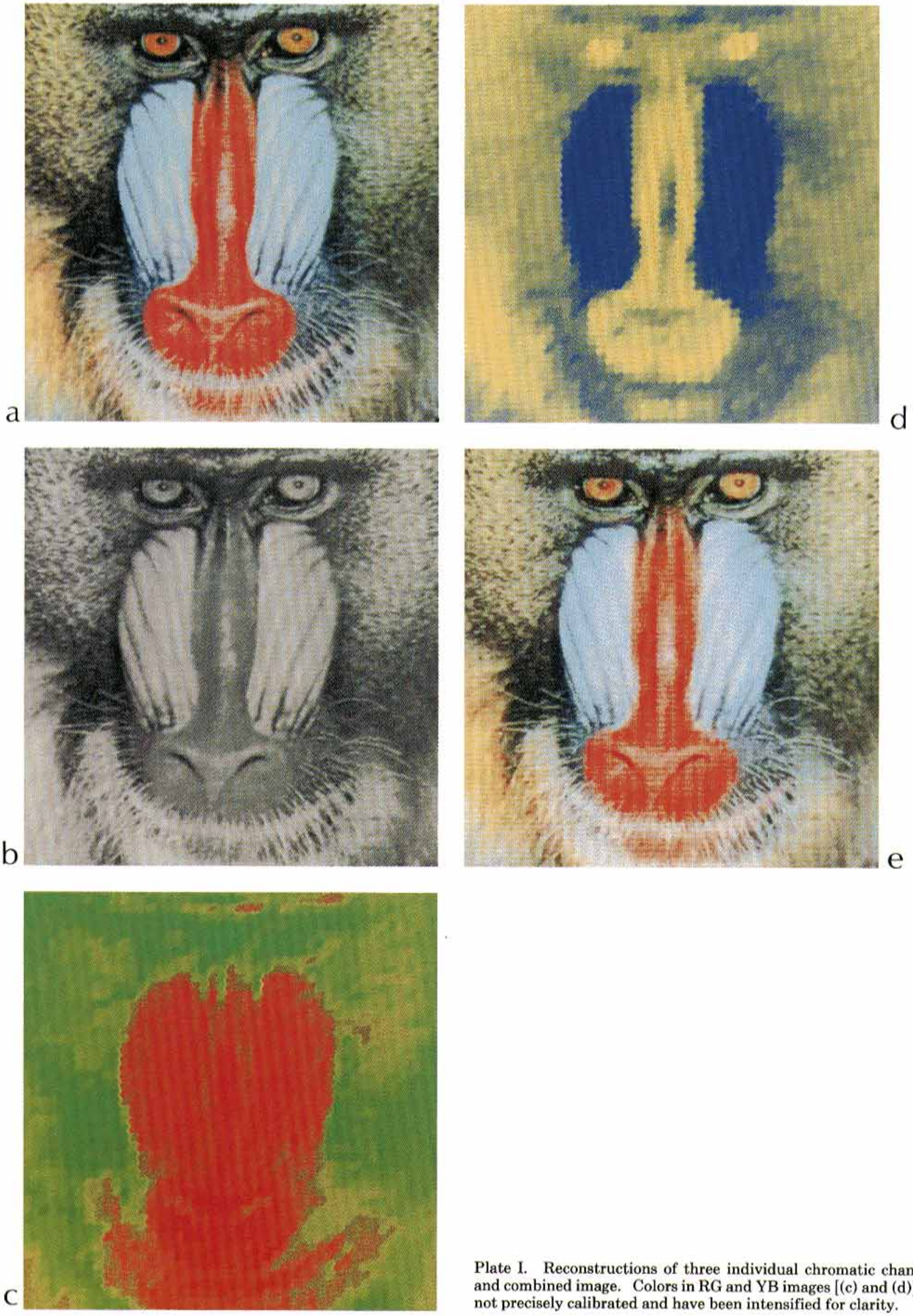


Plate I. Reconstructions of three individual chromatic channels and combined image. Colors in RG and YB images [(c) and (d)] are not precisely calibrated and have been intensified for clarity.

image coded with the HOP transform requires 0.86 bit/pixel for the achromatic channel and only an additional 0.095 bit/pixel for the full-color image. Note that the color image requires only ~11% more bits. It is clear that efficient coding of the image stream must exploit this feature, and this can be done by means of a PCA.

A second argument for segregation of achromatic and chromatic signals is that in the human visual system the latter appear to contribute little to perception of motion. This suggests that chromatic components at high temporal frequencies (motion signals) need not be included in the PCA or, if included, require only modest bandwidth. This is illustrated in Fig. 6, wherein Figs. 6(b) and 6(c), depicting chromatic components, lack the second layer of the cake corresponding to motion signals.

Error Tolerance

Packet networks are characterized by the possibility of packet loss. In a broadcast context, there is no possibility of retransmission.[74] It is therefore important that the loss of a single random packet produce only small diminution in quality. Unlike for recursive or block-organized codes, errors or omissions in perceptual codes do not propagate and tend to be visually benign.[77] This is presumably due to the essentially linear decomposition of the image into components of approximately equal visibility and to the residual redundancy in the components. Furthermore, errors are confined to the band in question and to the locations corresponding to the lost components. As noted above, loss of some components is more injurious than loss of others (e.g., low-resolution static bands), but these components tend to be those with the lowest bit rate, suggesting the option of additional error-correction coding or deliberately injected redundancy.

Scalability

Beyond the case of random packet loss, there is also the possibility of a more systematic triage of the packet stream. This may arise because of network overload or because the user or the provider, for reasons of cost, selects less than complete delivery of information. The bit allocations discussed above are based on visibility of errors and are designed to maintain near-perfect (perceptually lossless) transmission. To allow for a graded delivery of information the architecture should permit simple segregation of signals into various priority levels. This idea has been presented elsewhere in terms of "guaranteed packets" and "enhancement packets,"[98] "droppable" and "nondroppable" bits,[99] and "most significant" and "least significant parts."[100]

The advantage of perceptual coding in this context is that it divides the image sequence into components of known (or knowable) perceptual significance. In the prototype scheme described above, low spatial frequencies are arguably more important than high, low temporal frequencies are more important than high, and luminance information is more important than chromatic information. These valuations are of course dependent on the user and are more difficult when we compare across dimensions (high temporal frequencies versus chromatic information), but the perceptual components nevertheless allow these valuations to be made on perceptually meaningful dimensions.

It should also be noted that various other services, such as transmission of still images, progressive transmission, and transmission of miniaturized versions of a video stream (for browsing or channel selection), are also accommodated as scaled transmissions.

Device Independence

Many of the advantages of a device-independent video standard have been described by Lippman.[1] An important aspect of device independence is resolution independence. The lack of such independence is the basis for much of the incompatibility among the existing broadcast standards and proposed HDTV standards. The systems differ in line, pixel, and frame rates. While digital schemes in general permit conversion from any standard to any other standard, provided that the information is there, the computation involved may be prohibitive.

The PCA is attractive in this respect in that it allows for simple rendering into an arbitrary resolution format. There are two reasons for this. The first is that the components are specified in device-independent dimensions. For example, a given spatial resolution band is specified in cycles/degree. It is the responsibility of the rendering engine to ensure that the component has the proper dimensions on the final display, based on a particular design viewing distance. The second reason is that, because the perceptual components are easily represented in the frequency domain, they allow for a particularly simple method of decimation and interpolation. In effect, the components are assembled in a frequency-domain buffer that is padded or clipped before inverse transformation.[60,63,77]

Extensibility

Existing broadcast TV is based on a number of incompatible regional standards (NTSC, PAL, SECAM). Much of the current debate about HDTV centers on what new standard to adopt. The reason that a new standard is necessary, and that the debate on its form is so heated, is that neither old nor new proposed standards are easily or fully extensible. The concept of an extensible standard is that it permits additions and improvements. The perceptual components architecture is extensible in two important ways.

First, because the components are specified in a device-independent way, additional components can be added at any time. For example, additional resolution may be added by including higher-resolution bands. Additional field of view may be added by including more samples within each band. If the receiver is incapable of rendering this information, the information is ignored.

This extensibility permits the quality of service to improve in a graded fashion (this is also an aspect of scalability discussed above). As coding algorithms and channel bandwidths improve, suppliers may add quality to the broadcast. Customers willing to pay for the additional quality will purchase receivers capable of rendering the additional components, and perhaps they will pay a higher fee for the premium service. But no abrupt change in the transmission standards will be required.

The second extensibility of the PCA is the following. The perceptual components that have been described are those characteristic of the early stages in the human coding of the image stream. As vision science progresses, the later stages will be revealed. These later stages are likely to address a more object-oriented rather than pixel-oriented representation. There is reason to believe that this higher-level coding will provide still more powerful leverage for compression of the image stream. It is therefore important that the architecture accommodate these higher-level components. The device independence of the PCA satisfies this requirement.

Although the emphasis here is on transmission of imagery, it should be noted that the notion of a device-independent image description language can also handle graphics elements. Thus the image rendering engine might also be capable of graphics rendering.

CONCLUSION

I have sketched the outlines of a perceptual-components architecture for digital video and have attempted to show how this architecture will yield desirable attributes such as efficiency, error tolerance, scalability, device independence, and extensibility. Many of the ideas in the PCA are not new, and they have been implemented in partial form in existing TV systems or image coding systems. What is new is the principle of adopting human visual coding as the model in all aspects of the architecture. This radical proposal must no doubt be moderated by practical considerations, but the digital aspect of future video provides a great freedom in which to achieve this ideal.

ACKNOWLEDGMENTS

I thank the members of the NASA Vision Group for their advice and encouragement. This study was supported by NASA grant RTOP 506-47.

REFERENCES

1. A. Lippman, "Open architecture television," in *Digest of Topical Meeting on Applied Vision* (Optical Society of America, Washington, D.C., 1989), pp. 122–126.
2. L. Turner, T. Aoyama, D. Pearson, D. Anastassiou, and T. Minami, "Packet speech and video," IEEE J. Select. Topics Commun. **7,** 629–869 (1989).
3. L. M. Hurvich and D. Jameson, "An opponent-process theory of color vision," Psychol. Rev. **64,** 384–404 (1957).
4. J. Larimer, D. H. Krantz, and C. M. Cicerone, "Opponent process additivity I: red/green equilibria," Vision Res. **14,** 1127–1140 (1974).
5. J. Larimer, D. H. Krantz, and C. M. Cicerone, "Opponent process additivity II: yellow/blue equilibria and non-linear models," Vision Res. **15,** 723–731 (1975).
6. C. R. Ingling, "The spectral sensitivity of the opponent-colors channels," Vision Res. **17,** 1083–1090 (1977).
7. S. L. Guth, R. W. Massof, and T. Benzschawel, "Vector model for normal and dichromatic color vision," J. Opt. Soc. Am. **70,** 197–211 (1980).
8. J. Krauskopf, D. R. Williams, and D. W. Heeley, "Cardinal direction of color space," Vision Res. **22,** 1123–1131 (1982).
9. A. M. Derrington, J. Krauskopf, and P. Lennie, "Chromatic mechanisms in lateral geniculate nucleus of macaque," J. Physiol. (London) **357,** 241–265 (1984).
10. G. Buchsbaum and A. Gottschalk, "Trichromacy, opponent colours coding and optimum colour information transmission in the retina," Proc. R. Soc. London Ser. B **220,** 89–113 (1983).
11. C. R. Ingling and E. Martinez, "The spatiochromatic signal of the r-g channel," in *Colour Vision*, J. D. Mollon and L. T. Sharpe, eds. (Academic, London, 1983).
12. F. W. Campbell, G. F. Cooper, and C. Enroth-Cugell, "The spatial selectivity of the visual cells of the cat," J. Physiol. (London) **203,** 223–235 (1969).
13. R. L. De Valois, D. G. Albrecht, and L. G. Thorell, "Spatial frequency selectivity of cells in macaque visual cortex," Vision Res. **22,** 545–559 (1982).
14. J. G. Robson, "Neural images: the physiological basis of spatial vision," in *Visual Coding and Adaptability*, C. S. Harris, ed. (Erlbaum, Hillsdale, N.J., 1980).
15. R. L. De Valois and K. K. De Valois, *Spatial Vision* (Oxford U. Press, Oxford, 1988).
16. J. P. Jones, A. Stepnoski, and L. A. Palmer, "The two-dimensional spectral structure of simple receptive fields in cat striate cortex," J. Neurophysiol. **58,** 1212–1232 (1987).
17. M. J. Hawken and A. J. Parker, "Contrast sensitivity and orientation selectivity in lamina IV of the striate cortex of old world monkeys," Exp. Brain Res. **54,** 367–372 (1984).
18. M. J. Hawken and A. J. Parker, "Spatial properties of neurons in the monkey striate cortex," Proc. R. Soc. London Ser. B **231,** 251–288 (1987).
19. A. B. Watson, "Cortical algotecture," in *Vision: Coding and efficiency*, C. B. Blakemore, ed. (Cambridge U. Press, Cambridge, 1990).
20. F. W. Campbell and J. J. Kulikowski, "Orientation selectivity of the human visual system," J. Physiol. (London) **187,** 437–445 (1966).
21. F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," J. Physiol. (London) **197,** 551–566 (1968).
22. C. Blakemore and F. W. Campbell, "On the existance of neurones in the human visual system selectivity sensitive to the orientation and size of retinal images," J. Physiol. (London) **203,** 237–260 (1969).
23. C. B. Blakemore, J. Nachmias, and P. Sutton, "The perceived spatial frequency shift: evidence for frequency-selective neurones in the human brain," J. Physiol. (London) **210,** 727–750 (1970).
24. A. B. Watson, "Summation of grating patches indicates many types of detector at one retinal location," Vision Res. **22,** 17–25 (1982).
25. A. B. Watson, "Detection and recognition of simple spatial forms," in *Physical and Biological Processing of Images*, O. J. Braddick and A. C. Sleigh, eds. (Springer-Verlag, Berlin, 1983).
26. J. G. Daugman, "Spatial visual channels in the Fourier plane," Vision Res. **24,** 891–910 (1984).
27. R. L. De Valois, E. W. Yund, and H. Hepler, "The orientation and direction selectivity of cells in macaque visual cortex," Vision Res. **22,** 531–544 (1982).
28. J. J. Kulikowski and D. J. Tolhurst, "Psychophysical evidence for sustained and transient mechanisms in human vision," J. Physiol. (London) **232,** 149–163 (1973).
29. D. J. Tolhurst, "Separate channels for the analysis of the shape and the movement of a moving visual stimulus," J. Physiol. **231,** 385–402 (1973).
30. A. B. Watson and J. G. Robson, "Discrimination at threshold: labelled detectors in human vision," Vision Res. **21,** 1115–1122 (1981).
31. A. B. Watson, "Temporal sensitivity," in *Handbook of Perception and Human Performance*, K. Boff, L. Kaufman, and J. Thomas, eds. (Wiley, New York, 1986).
32. W. H. Merigan and T. A. Eskin, "Spatio-temporal vision of macaques with severe loss of P_β retinal ganglion cells," Vision Res. **26,** 1751–1761 (1986).
33. E. Levinson and R. Sekuler, "The independence of channels in human vision selective for direction of movement," J. Physiol. (London) **250,** 347–366 (1975).
34. A. B. Watson, P. G. Thompson, B. J. Murphy, and J. Nachmias, "Summation and discrimination of gratings moving in opposite directions," Vision Res. **20,** 341–347 (1980).
35. A. Pantle and R. W. Sekuler, "Contrast response of human visual mechanisms sensitive to orientation and direction of motion," Vision Res. **9,** 397–406 (1969).
36. D. B. Hamilton, D. G. Albrecht, and W. S. Geisler, "Visual cortical receptive fields in monkey and cat: spatial and temporal phase transfer function," Vision Res. **29,** 1285–1308 (1989).
37. A. B. Watson and A. J. Ahumada, Jr., "A look at motion in the frequency domain," in *Motion: Perception and Representa-*

tion, J. K. Tsotsos, ed., NASA Tech. Mem. 84352 (Association for Computing Machinery, New York, 1983).

38. A. B. Watson and A. J. Ahumada, Jr., "Model of human visual-motion sensing," J. Opt. Soc. Am. A **2**, 322–342 (1985).
39. E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," J. Opt. Soc. Am. A **2**, 284–299 (1985).
40. J. P. H. van Santen and G. Sperling, "Elaborated Reichardt detectors," J. Opt. Soc. Am. A **2**, 300–321 (1985).
41. D. J. Heeger, "Model for the extraction of image flow," J. Opt. Soc. Am. A **4**, 1455–1471 (1987).
42. D. J. Fleet and A. D. Jepson, "Hierarchical construction of orientation and velocity selective filters," IEEE Trans. Pattern Anal. Mach. Intell. **11**, 315–325 (1989).
43. J. G. Robson, "Spatial and temporal contrast sensitivity functions of the visual system," J. Opt. Soc. Am. **56**, 1141–1142 (1966).
44. K. H. Foster, J. P. Gaska, M. Nagler, and D. A. Pollen, "Spatial and temporal frequency selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey," J. Physiol. (London) **365**, 331–363 (1985).
45. W. E. Glenn and K. G. Glenn, "Discrimination of sharpness in a televised moving image," Displays (October 1985), pp. 202–206.
46. W. E. Glenn, K. G. Glenn, and C. J. Bastian, "Imaging system design based on psychophysical data," Proc. SID **26**, 71–78 (1985).
47. J. H. D. M. Westerink and C. Teunissen, "Perceived sharpness in moving images," in *Human Vision and Electronic Imaging: Models, Methods, and Applications*, J. Allebach and B. Rogowitz, eds., Proc. Soc. Photo-Opt. Instrum. Eng. **1249** (to be published).
48. B. Girod, "Eye movements and coding of video sequences," in *Visual Communications and Image Processing '88: Third in a Series*, T. R. Hsing, ed., Proc. Soc. Photo-Opt. Instrum. Eng. **1001**, 398–405 (1988).
49. D. H. Kelly, "Pattern detection and the two-dimensional Fourier transform: flickering checkerboards and chromatic mechanisms," Vision Res. **16**, 277–287 (1976).
50. C. Noorlander, M. J. G. Heuts, and J. J. Koenderink, "Influence of the target size on the detection threshold for luminance and chromaticity contrast," J. Opt. Soc. Am. **70**, 1116–1121 (1980).
51. C. Noorlander, M. J. G. Heuts, and J. J. Koenderink, "Sensitivity to spatiotemporal combined luminance and chromaticity contrast," J. Opt. Soc. Am. **71**, 453–459 (1981).
52. C. Noorlander and J. J. Koenderink, "Spatial and temporal discrimination ellipsoids in color space," J. Opt. Soc. Am. **73**, 1533–1543 (1983).
53. K. T. Mullen, "The contrast sensitivity of human color vision to red/green and blue/yellow chromatic gratings," J. Physiol. (London) **359**, 381–400 (1985).
54. K. T. Mullen, "Spatial influence on color opponent contributions to pattern detection," Vision Res. **27**, 829–839 (1988).
55. A. B. Poirson and B. A. Wandell, "Sensitivity to Gabor patches modulated in color directions," in *Digest of 1989 Annual Meeting* (Optical Society of America, Washington, D.C., 1989), p. 174.
56. P. Cavanagh and S. Anstis, "The contribution of color to motion in normal and color-deficient observers," Vision Res. (to be published).
57. D. H. Kelly, "Luminous and chromatic flickering patterns have opposite effects," Science **188**, 371–372 (1975).
58. S. Tanimoto and T. Pavlidis, "A hierarchical data structure for picture processing," Comput. Graphics Image Process. **4**, 104–119 (1975).
59. P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," IEEE Trans. Commun. **COM-31**, 532–540 (1983).
60. A. B. Watson, "Ideal shrinking and expansion of discrete sequences," NASA Tech. Memo. 88202 (National Aeronautics and Space Administration, Washington, D.C., 1986).
61. E. H. Adelson, E. Simoncelli, and R. Hingorani, "Orthogonal pyramid transforms for image coding," in *Visual Communications and Image Processing II*, T. R. Hsing, ed., Proc. Soc. Photo-Opt. Instrum. Eng. **845**, 50–58 (1987).
62. A. B. Watson and A. J. Ahumada, Jr., "An orthogonal oriented quadrature hexagonal image pyramid," NASA Tech. Memo. 100054 (National Aeronautics and Space Administration, Washington, D.C., 1987).
63. A. B. Watson, "The cortex transform: rapid computation of simulated neural images," Comput. Vision Graphics Image Process. **39**, 311–327 (1987).
64. A. B. Watson, "Recursive, in-place algorithm for the hexagonal orthogonal oriented quadrature image pyramid," in *Surface Measurement and Characterization*, J. M. Bennett, ed., Proc. Soc. Photo-Opt. Instrum Eng. **1099**, 194–200 (1989).
65. A. B. Watson and A. J. Ahumada, Jr., "A hexagonal orthogonal oriented pyramid as a model of image representation in visual cortex," IEEE Trans. Biomed. Eng. **36**, 97–106 (1989).
66. J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," IEEE Trans. Acoust. Speech Signal Process. **36**, 1169–1179 (1988).
67. R. Wilson, "Finite prolate spheroidal sequences and their applications I: Generation and properties," IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-9**, 787–794 (1987).
68. R. Wilson and M. Spann, "Finite prolate spheroidal sequences and their applications I: Image feature description and segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **10**, 193–203 (1988).
69. S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," IEEE Trans. Pattern Anal. Mach. Intell. **11**, 674–693 (1989).
70. M. Vetterli, "Multi-dimensional sub-band coding: some theory and algorithms," Signal Process. **6**, 97–112 (1984).
71. J. W. Woods and S. D. O'Neil, "Subband coding of images," IEEE Trans. Acoust. Speech Signal Process. **ASSP-34**, 1278–1288 (1986).
72. E. Simoncelli and E. H. Adelson, "Non-separable extensions of quadrature mirror filters to multiple dimensions," MIT Media Lab. Vision Science Tech. Rep. **119** (Massachusetts Institute of Technology, Cambridge, Mass., 1989).
73. G. Karlsson and M. Vetterli, "Three dimensional subband coding of video," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (Institute of Electrical and Electronics Engineers, New York, 1988), pp. 1100–1103.
74. G. Karlsson and M. Vetterli, "Packet video and its integration into the network architecture," IEEE J. Select. Topics Commun. **7**, 739–751 (1989).
75. J. C. Darragh and R. L. Baker, "Fixed distortion subband coding of images for packet-switched networks," IEEE J. Select. Topics Commun. **7**, 789–800 (1989).
76. D. J. Sharma and A. N. Netravali, "Design of quantizers for dpcm coding of picture signals," IEEE Trans. Commun. **COM-25**, 1267–1274 (1977).
77. A. B. Watson, "Efficiency of an image code based on human vision," J. Opt. Soc. Am. A **4**, 2401–2417 (1987).
78. H. R. Wilson, D. K. McFarlane, and G. C. Phillips, "Spatial frequency tuning of orientation selective units extimated by oblique masking," Vision Res. **23**, 873–882 (1983).
79. G. C. Phillips and H. R. Wilson, "Orientation bandwidths of spatial mechanisms measured by masking," J. Opt. Soc. Am. A **1**, 226–232 (1984).
80. R. Blake and K. Holopigian, "Orientation selectivity in cats and humans assessed by masking," Vision Res. **25**, 1459–1467 (1985).
81. A. M. Derrington and G. B. Henning, "Pattern discrimination with flickering stimuli," Vision Res. **21**, 597–602 (1981).
82. R. F. Quick and R. N. Lucas, "Orientation selectivity in detection of chromatic gratings," Opt. Lett. **4**, 306–308 (1979).
83. M. S. Livingstone and D. H. Hubel, "Anatomy and physiology of a color system in the primate visual cortex," J. Neurosci. **4**, 309–356 (1984).
84. A. Bradley, E. Switkes, and K. De Valois, "Orientation and spatial frequency selectivity of adaptation to color and luminance gratings," Vision Res. **28**, 841–856 (1988).
85. R. F. Hess and G. T. Plant, "Temporal frequency discrimination in human vision: evidence for an additional mechanism in the low spatial and high temporal frequency region," Vision Res. **25**, 1493–1500 (1985).

86. G. Sclar, P. Lennie, and D. D. DePriest, "Contrast adaptation in striate cortex of macaque," Vision Res. **29**, 747–755 (1989).
87. A. M. Derrington and G. B. Henning, "Some observations on the masking effects of two-dimensional stimuli," Vision Res. **29**, 241–246 (1989).
88. A. B. Bonds, "Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex," Vis. Neurosci. **2**, 41–55 (1989).
89. I. Ohzawa and R. D. Freeman, "Contrast gain control in the cat visual system," J. Neurophysiol. **54**, 651–665 (1985).
90. D. J. Heeger, "Computational model of cat striate physiology," in *Computational Models of Visual Perception*, J. A. Movshon and M. Landy, eds. (MIT Press, Cambridge, Mass., 1990).
91. A. B. Watson, "Receptive fields and visual representations," in *Human Vision, Visual Processing, and Digital Display*, B. E. Rogowitz, ed., Proc. Soc. Photo-Opt. Instrum. Eng. **1077**, 190–197 (1989).
92. J. G. Daugman, "Entropy reduction and decorrelation in visual coding by oriented neural receptive fields," IEEE Trans. Biomed. Eng. **36**, 107–114 (1989).
93. F. W. Mounts, "A video encoding system with conditional picture element replenishment," Bell Syst. Tech. J. **48**, 2545–2554 (1969).
94. F. Rocca and S. Zanoletti, "Bandwidth reduction via movement compensation on a model of the random video process," IEEE Trans. Commun. **COM-20**, 960–965 (1972).
95. A. N. Netravali and J. D. Robbins, "Motion-compensated television coding—Part I," Bell Syst. Tech. J. **58**, 631–670 (1979).
96. R. F. W. Pease and J. O. Limb, "Exchange of spatial and temporal resolution in television coding," Bell Syst. Tech. J. **50**, 191–200 (1971).
97. W. E. Glenn and K. G. Glenn, "HDTV compatible transmission system," J. Soc. Motion Pict. TV Eng. **96**, 242–246 (1987).
98. M. Ghanbari, "Two-layer coding of video signals for VBR networks," IEEE J. Select. Areas Commun. **7**, 771–781 (1989).
99. C. Chamzas and D. L. Duttweiler, "Encoding facsimile images for packet-switched networks," IEEE J. Select. Areas Commun. **7**, 857–864 (1989).
100. F. Kishino, K. Manabe, and Y. Hayashi, "Variable bit-rate coding of video signals for atm networks," IEEE J. Select. Areas Commun. **7**, 801–806 (1989).