Beyond Images

V. Michael Bove, Jr.

Abstract: The idea of the image - a still two-dimensional projection of a moving three-dimensional scene - enabled the creation of electronic visual communications. But now that the capture, transmission, and display of visual information is mediated by very powerful computational devices, perhaps designing systems around images is hindering advanced thinking about, and applications for, visual communications.

I note that for efficiency reasons, digital video coding is already moving in the direction of describing scenes rather than the images produced from them by cameras. This shift provides a rare opportunity for a dialogue between the technical and creative communities while a potentially radically different form of visual communication is still under development. In this paper I describe some of the implications of the shift from images to models and as a particular encouragement for the aforementioned dialogue offer intelligently interoperable video, which adapts to different viewing circumstances.

Introduction

All intuitive perception is intellectual, for without the understanding we could never achieve intuitive perception, observation, the apprehension of objects. On the contrary, we would stop short at the mere sensation that might possibly have meaning in reference to the will as pain or comfort; but for the rest it would be a succession of states devoid of meaning, and nothing like knowledge. Intuitive perception, that is, knowledge of an object, first comes about through the understanding that refers every impression received by the body to its cause. In space, which is intuited a priori, the understanding shifts this cause to the point from which the effect starts and thus recognises the cause as acting, as actual, i.e. as a representation of the same kind and class as the body.

Schopenhauer, in On Vision and Colours, addresses the important distinction between the perception of an image and the understanding of the objects that produced it. But in traditional forms of visual communication, the image itself is often the primal object. Increasingly, this imagery is handled in not just electronic but digital form, and is passing through computational devices as part of its capture/creation, distribution, and display.

The initial view - still widely held - of the digitisation of visual communication is that it will allow us to do all the things we'd always done with electronic images, but with more optimal and more adjustable trade-offs among efficiency, fidelity, and cost. It has become possible to distribute video over channels and media that didn't have enough capacity for analogue imagery, while existing video channels can carry higher-quality imagery or larger numbers of simultaneous programs. Costly, tedious, and often-unrepeatable darkroom artistry for image enhancement, retouching, or cinematic special effects have been replaced by a few button presses on a computer.

Suppose, though, we take into account that the information passing through these computational devices and digital channels isn't constrained (as were its analogue-voltage, silver-density, or ink-on-paper predecessors) to represent two-dimensional arrays of light intensities. Do there exist alternative representations that are a better match to the requirements of the creator or the viewer?

Throughout the discussion which follows, I will be taking the term 'image' in its most physical sense. I make the assumption that the cultural roles of images would continue similarly to be served - perhaps even in a richer way - by any future visual information representation.

In this paper I propose that in future, a significant portion of visual communication will involve not pictures of things, but rather models of things from which pictures can be created, as well as information describing how to process the models for display. Another way of looking at this change is to note that in the past, the engineering behind representations for electronic visual communications was largely concerned with the low-level characteristics of signals (e.g. analogue voltages or sequences of digital numbers); in future it will be more concerned with the properties of the scene - such as the shapes, arrangements, and motions of objects - in front of the camera.

The implications of the aforementioned shift in representation are vast, and in this paper I don't intend to try enumerating all of them. Rather, I regard the present situation as fortuitous. There briefly exists a rare opportunity for dialogue between the technical developers of systems and the creative users, before the technical components are standardised. We will find that scene modelling is one of those very infrequent axes in which both technical efficiency and creative freedom can be served simultaneously. The advantages will be significant and beneficial for both content creators and viewers.

In an effort to stimulate this discussion I will present considerations from a variety of disciplines. First, I will discuss the formation of images and certain difficulties (particularly in the area of interoperability) arising from the dimensional collapse that is part of their production. I will then present an alternative video representation based on scene understanding and segmentation into objects, and, as a case study of
possible benefits, discuss ‘meta-television,’ which, among other properties, compositionally adapts to viewing circumstances.

**Eyes and cameras**

Physical images form when light—probably having been emitted by and/or reflected from a three-dimensional scene—passes through a small aperture and falls onto a surface. A still image results when we integrate these light intensities over some finite interval of time. Both the dimensional collapse and the temporal integration discard potentially important information (though some of it might still be recoverable, given multiple two-dimensional views of a scene or a prior knowledge).

The first image-forming eyes (as opposed to simple light-sensitive spots, as in protozoa) were essentially pinholes facing arrays of photoreceptors. Such an arrangement persists in a few aquatic animals, most notably the chambered nautilus. Image formation, of course, doesn’t require an eye; gaps in the foliage of a tree can cast images of the sun onto the ground. During the fourth century BC, Aristotle in his *Problems* noted this phenomenon, while Chinese manuscripts from the same period indicate that image formation by apertures was already well known in the East. By the ninth century AD, both the Chinese and the Arabs were working with devices that we would equate with the camera obscura.

Focusing effects provide another example of the ways in which image formation discards information. The problem with image formation by pinholes, as the evolutionary process and then camera-makers discovered, is that the images are sharp only when the hole is small, in which case they are also dim. The addition of a lens to the camera obscura, which may have first been done in sixteenth-century Italy, allowed a larger aperture, and thus a brighter image, though objects were reproduced sharply only within a limited range of distances. Photographers refer to this effect as depth-of-field; those of a signal-processing turn of mind would describe the process as the convolution of a sharp image with a distance-dependent low-pass filter.

Depth-of-field effects are rarely seen in normal vision circumstances, thus it is remarkable that they quickly became part of the visual language used by artists. Detouching can, for instance, clearly be seen in certain paintings by Johannes Vermeer from the 1600s; it has been a matter of dispute among art historians whether these works were literally recording camera obscura images, or whether they were painted in a style imitating optical effects which were apparently fairly well-known by the ‘visually literate’ of that period. Interestingly, three hundred years later, some Photorealist painters claimed that they were painting not scenes but camera image planes, in an effort to justify their almost obsessive representational perfection in an era ruled by abstraction.

Because cameras and the eyes of vertebrate animals are image-forming devices of similar construction, it is tempting to think that there is something natural and fundamental about basing visual communication systems on camera images. But the images produced by eyes are only the preliminary steps in a perceptual process that seeks to model and interpret the three-dimensional world. ‘Seeing’ regards the models, not the images, and it is on this philosophical point that the eye and the camera (at least as the latter is commonly understood) part company.

The image formed by a camera obscura had to be copied or traced onto paper. Photography took the artist’s likeness of the camera obscura’s image plane and replaced it with a direct chemical copy; television, though, faced the problem of transmitting a time-varying, two-dimensional pattern using only a one-dimensional quantity (a voltage that varies with time).

Before we consider how to do that, we might examine the simpler case of cinema. It is a fortunate consequence of the development of the visual system that when the eyes are presented with a rapid enough series of snapshots of a moving scene the brain interprets the result as smooth motion. This effect is known to researchers as the phi phenomenon, and mechanical devices (initially toys) that exploit it have existed for many hundreds of years. ‘Rapid enough’ is a measure that depends on a number of factors including image brightness and the visual angle occupied by the images, but the important concept is that we can discretise the continuous temporal dimension, replacing the x-y pattern with a sequence of x-y images. In television, we can similarly discretise the vertical dimension by defining scan lines, finer detail than which we will discard, if we desired a discrete sequence of samples we could also discretise the horizontal dimension into pixels.

Beginning with the dimensional collapse performed by the lens onto the image plane, and ending with the definition of pixels, we have discarded a great deal of information in an effort to simplify the design of our video system. That simplicity may have cost us flexibility, since the camera’s scanning pattern imposes a requirement that the display either operates at the same spatiotemporal resolution, or else performs significant amounts of calculation to generate images at a different number of scan lines or frames-per-second. Losing the ability easily to decouple the parameters of image capture from those of display is certainly unfortunate from an applications perspective, and as we will see later, there may be coding-efficiency costs as well. The crucial point to note here is that frames, scan lines, and pixels aren’t fundamental parts of nature, they are artefacts imposed by cameras for convenience. If they start to become inconvenient, we should consider replacing them with a better representation.

The time-varying voltage that results from a television camera’s scanning process, or from applying similar electronic scanning to the frames of a strip of film, is often called a waveform. Even if the result is digitised, which is to say that the waveform is a sequence of discrete samples, permitted to assume one of only a limited number of values, the result is...
still effectively a waveform. Hereafter, I will use the phrase 'waveform coding' to describe any representation in which the data are segmented into units (frames, blocks of pixels, etc.) which are dictated by the spatiotemporal sampling format rather than by the scene content. Writers have used the adjectives 'structured,' 'model-based,' 'analysis-synthesis,' 'content-based,' and 'object-oriented' interchangeably [and one of the sure signs that this field is still wide-open is that there is no universal nomenclature] to describe representations in which the data are instead segmented into units based upon the physical or semantic properties of the scene being represented. Since the dimensional collapse of the camera has already thrown away a lot of information, this segmentation is not a trivial task, and relies on machine-vision techniques. Full discussion of the methods involved is beyond the scope of this article, but generally they involve some combination of a priori assumptions about the structure of real objects, and the interpretation of information across time (i.e. processing image sequences to find coherent regions with corresponding motion, colour, and texture) or space (i.e. combining information from multiple cameras). The resulting data units (which we will call 'objects') are likely accompanied by what Nicholas Negroponte has called 'the bits about the bits' and what I will call the 'scripting information,' which describes the relationships among the objects and how to manipulate and assemble them to synthesise a view of the scene. The scripting information may also encode how the presentation of the information might vary depending on user interaction, user preference, display type, or other state information.

We will in a later section consider the added flexibility created by better scene understanding; in order to see why it leads to more efficient data compression the reader must know a little about what is called predictive coding. Given signals with statistical redundancy, the decoder predicts the next unit of data (e.g. a video frame) based upon past data, while the encoder sends a correction signal to account for prediction errors. The more accurate the decoder's model, the less information that must be transmitted. It is for this reason that algorithms like MPEG-2 transmit motion vectors, \((x,y)\) values which tell the decoder to shift square blocks of a just-decoded image in order to create a prediction for the next image. But it is this author's opinion that 'motion vector' is a misnomer. First, the two-dimensional image-plane motion description isn't a very good model for the motion of real-world objects. Second, motion is described over an arbitrary block which might contain more than one scene object, having different motions. Finally, the motion vectors are chosen by the encoder not as the best approximations for motion in a small block of the image (even as describable in the meagre two-dimensional means provided) but rather as the correlation from one frame to another that minimises the error signal. This is fine if all one ever wants to do is minimise bitrate, but motion vectors as typically computed are not even optimal for such simple applications as interpolation into different frame rates, and though some remarkably clever work has been done on interpreting MPEG bitstreams to find scene changes and object or camera motions, ultimately decoupling production/acquisition from display or finding semantic content should be designed-in features of a video representation and not serendipitous afterthoughts.

Numerical interoperability

In the late 1930s, engineers in North America and Europe knew how to build a fully-electronic [as opposed to earlier methods, which used mechanical contrivances like spinning disks] television system. But they weren't yet satisfied with the resolutions available from the camera and display tubes, and with the frequency responses (related to the image detail) of inexpensive amplifier circuits. Development of a widespread service, though, called for standardising scanning rates and broadcast signal specifications. In a now-nearly-forgotten 1939 article, Allen DuMont proposed a way of transmitting image information along with scanning information which instructed a receiver how to display the images, creating a system in which the display had no single inherent resolution or frame rate and could thus provide better images as the cameras to produce them appeared. There were certain inefficiencies which doomed DuMont's scheme, but the fundamental property behind it continues to be appealing: we of the MIT Media laboratory later called it extensibility: An information representation easily extends to accommodate higher-quality information and higher-capacity channels as they become available.

A corollary to extensibility is scalability: A representation (e.g. a bit stream) can be parsed and decoded in part or in full to produce output at differing quality levels (to accommodate different channel capacities or decoder capabilities).

A third important principle is interoperability: The numerical parameters of source and display (i.e. resolution, frame rate) need not match.

Scalability, extensibility, and interoperability are useful properties for any sensory information representation to have. When we applied them to television, the result was called Scalable Open Architecture Television, where the intent was to produce a digital signal representation which allowed different displays to show images at a variety of resolutions and frame rates. In essence, receivers were not to reproduce the pixels taken by a camera at a particular spatial resolution and frame rate, but rather were to try to approximate the pixels that would have been taken had the continuous image been sampled on various different grids. Much more than in the analogue past, digital video will involve the creation of content for a heterogeneous display environment. As digital channels proliferate and consumer video viewing equipment expands beyond the traditional television receiver to include projectors, panels, portable communicators, resizable windows on personal computers, and
Object-based television

I have already alluded to a view of video in terms of objects and scripting information. It is indeed even possible to view a waveform coder like MPEG-2 as made up of objects: frames, motion vector arrays, error data; and a script: the bitstream syntax that identifies a group of bits as one or another type of object. More useful object-based coding of video, though, requires that the objects be not just data units but also meaningful parts of the scene. Such coding methods can range from image segmentation into regions or layers based upon motion (perhaps in combination with other cues like colour and texture),

The added compression efficiency of object-based video generally comes about because the more accurate transmitted model and the receiver’s greater computational ability permit better prediction. In layered or region-segmented two-dimensional coders, for example, the transmitted motion parameters are intended to be a more correct approximation for the image-plane projections of real-world objects than would result from (x,y) vectors computed on an arbitrary square grid. Added memory in the decoder also eliminates the need to retransmit information about occluded and revealed regions. Yet more computationally intensive algorithms can under certain circumstances estimate camera or object motion in three-dimensional space.

Decoding some of the more sophisticated representations is computationally equivalent to synthetic computer graphics rendering, with some additional steps for data decompression; the remaining major research challenge lies largely in the machine-vision methods used in encoding. But while scene-analysis methods continue to be developed, in many cases the needed information may already exist. Data on the approximate shapes and arrangements of objects and on camera motions might also be provided by computer previsualisation tools used in programme production; this might prove useful in coding or supporting viewer interaction or personalisation. Many video-editing systems now permit effects such as object compositing. And increasingly, video content is created on computer-graphics systems.

Even if the structured scene representation is used only within the studio, new production and post-production possibilities emerge. Research at the MIT Media Laboratory, to be described in a subsequent section, looks into transmitting the scene description to the receiver.

The intent of the research described herein is not to dictate a single algorithm for handling video, nor even to establish how to determine the ‘best’ description for a particular application or moving scene. Waveform-based encoders will soon be ubiquitous and inexpensive, and involve a short and deterministic encoding delay. It is likely that JPEG, MPEG, and the like will be with us for a long time. Thus, any system we consider designing has to be able to deal with these ‘traditional’ digital video representations as well as higher-level ones.

Given a sufficiently powerful and flexible decoder, the way in which a scene is described, and the forms of the constituent objects, represent an originator-specified trade-off among a number of considerations:

- interactivity/personalisation: Clearly, the more individually-described objects in the scene, and the more the representation is like a computer-graphics database, the more presentation options are enabled. Conversely, the originator of a programme can force all viewers to see precisely the same presentation by appropriate design of the script and selection of objects.
- compression efficiency: Exactly how the choice of representation affects the reconstructed image quality at a given bit rate remains an active research area in analysis/synthesis video. In any event, where image quality is the most important consideration, a flexible decoder permits an encoding process to select the most appropriate representation, even on a scene-by-scene or object-by-object basis.
- analysis ability: More sophisticated scene descriptions require more computation, and indeed some scenes simply may not lend themselves to high-level descriptions given current analysis methods and encoding hardware capabilities. It may be that ultimately there will be an agreed-upon hierarchy of modelling methods, each of which requires more analysis than those below it, in return for which it can provide more efficiency and features. If the desire for real-time encoding and transmission is foremost, unsophisticated representations of image sequences will be sufficient and appropriate.

Objects as described above are strictly data items, with all the procedural description residing in the script. Taking ‘object-oriented’ in
On being meta: What does it mean to carry interoperability beyond pixel resampling? Intelligent Weitzman, who has researched electronic documents that self-interoperability reconfigure under differing circumstances, notes:

By embedding knowledge of information structure and design constraints, information can maintain a sense of itself. This enables information presentation to be adaptive, taking into account the affordances and constraints of the delivery environment. Information sources can then present themselves appropriately under a wide variety of situations.35

An instructive precedent comes from the Western moveable-type tradition. It is not commonly known that differing sizes of a type font were designed separately to account for media and perceptual effects.26 One common design change is to reduce the contrast between thin and thick strokes (by making the thin ones relatively thicker) as the point size decreases; this is a response to a limitation of the ‘display’: in metal-type days there was a minimum stroke thickness that could be cast reliably. We now remember Baskerville, and even more so Bodoni, for pushing the technical limitations, creating fonts with more dynamic range than their predecessors. A second and even more subtle change is related instead to human perception. The ‘x-height’, or ratio of the height of an ascenderless, descenderless lowercase letter (an x, say) to that of an upper-case letter, increases with decreasing point size; this technique has long been known to improve legibility of very small type.

When phototypesetting came into vogue in the mid-twentieth century, all sizes of a letter came from one master film image, optically enlarged or reduced. Thus phototypesetting fonts manifested a compromise which was neither optimal nor unacceptable at any needed size. Digital typesetting added computational ability, and for scalability as well as data-compression reasons, digital systems almost immediately moved from bitmap images to mathematical outline descriptions in terms of lines, arcs, and parametric curves. Except for a few systems which had multiple font masters for different size ranges, though, the one-design-fits-all-sizes problem remained.

Knuth then introduced the idea of a ‘metafont’, a procedural mathematical description of letterforms that incorporated reactions to different parameters, enabling a designer to specify the design behaviour under different circumstances such as differing size and output resolution.27 But meta-design remains a largely unexplored field, for reasons we will note below.

Scene capture

If the action of a camera (or a group of cameras) will no longer be simply the formation of images, we need a new term. I propose ‘scene capture’. A potential dichotomy exists between technological abilities [what can our methods sense?] and creative requirements [what do we need to know about the scene to support a particular application]?

Resolving this issue requires a categorisation of the results of scene capture. The object-based video compression already discussed...
borrows heavily from what is called 'low-level' machine vision, whose literature can give us an exhaustive list of scene attributes that are easy to compute with unintelligent processes, but with minimal contextual reference. Cognitive science or 'high-level' machine vision literature provides views of perceptual organisation, but these may not connect well with either currently-practical analysis algorithms, or the creative/communicative motivations for this work. Theories of film and video production may present a categorisation that is too abstract to bear upon the technical issues. A recent attempt to bridge the gap from the production side is Zettl's dimensional theory of 'applied media aesthetics', in which the first dimension is light and colour, the second adds area, the third adds volume, and the fourth is time. Even that, though, doesn't concretely enough relate to the information we will be able to sense, interpret, or input explicitly in the act of creating video, and it doesn't say enough about distribution and consumption. I present the following categorisation of scene-capture information not from any desire to revamp cinematic theory, but only because it will make clearer the effects of object-based representations on production and post-production.

I suggest that we consider scene capture as resulting in three levels of information. The broadest category, and also the least directly visual, is what I will call contextual information. This tells how a 'shot' [if we can use that word in a broader sense to include a database merging the output of more than one camera's point of view] relates to others, and to the world at large. This information will include notions such as the location and heading of the camera or cameras, and the time at which the shot was made. It should be noted here that there are really two sorts of contextual information: production context information implies real time and real location, and in some cases may not persist with the rest of the scene data beyond the post-production stage. Information of this sort can be provided by equipping cameras with clocks, Global Positioning System (GPS) receivers, and sensors for acceleration, rotation, and level. Narrative context information connects together the elements that make up the presentation for the viewer, and is created/modified by the editing (to use the film/video term) or authoring (to use the multimedia/publishing term) process, though in an interactive viewing situation some narrative context information may also be created by the stored history of the viewer's actions. Another sort of production context information is somewhat more concerned with the image formation process, and includes lens characteristics and scene lighting; analysis algorithms show some progress toward recovering this information, while computer graphics rendering from a three-dimensional model can create or modify it. Traditional production and post-production has mostly entailed manipulating the contextual level.

What I will call semantic information captures the coarse-scale characteristics of, and relationships among, the objects that make up a scene. Scene analysis methods that recover three-dimensional shape, or that segment images, can provide some of the data at this level. Scriptwriting – whether cinematic or computer-graphic – has traditionally been about creating information at the semantic level and above, but given richer object descriptions and perhaps a priori information, post-production would be able to affect some of these properties as well.
Prototype programme material

MIT Media Laboratory’s Information and Entertainment Section has made two short productions called The Museum and Wallpaper. In both cases, three-dimensional models of the unoccupied sets were computed using a semi-automatic technique that takes ordinary camera images as input, enabling resynthesis of arbitrary points of view for regions of the scene that can be approximated as planar polygons, (Figure 1), and finding the three-dimensional locations of actors moving about in the sets. As we do not yet have methods for producing good three-dimensional models for people, the actors are represented as two-dimensional objects with multiple viewpoints provided by multiple cameras.

The first form of intelligent interoperability we explored was recropping the frame for differing sizes and shapes of screens. A more advanced behaviour that seems to offer great potential was changing the effective focal length of a simulated lens, (Figure 2), thus allowing the entire background to be visible in a small window while not unduly reducing the size of foreground elements. Doing the latter while maintaining good scene composition may also entail a shift in camera position, as shown in the illustration.

Figure 1: Still from the video Wallpaper, in which we used machine-vision methods to build a three-dimensional model of the set. Combined with multiple calibrated camera views of the actors, this modelling permits synthesising arbitrary views of the scene either in post-production or for interactive display.

Figure 2: In this example from the video The Museum, the focal length of a simulated lens changes when the video is viewed on a smaller screen; thus, the actors in the foreground remain recognisably large while much of the background stays visible.

For viewing, we have implemented a hardware-software computational pipeline for object-based video, capable of decoding a variety of representations. Scripting is currently done with a list-oriented interpreted language called ISIS, though the script parser is separable from the display pipeline, so various languages can be (and have been) used in our work. In any case, the scripting language, like PostScript, is human-readable but rarely edited directly, being rather the result of an analysis algorithm or interactions with post-production tools. Other research in the group aims toward making the great degree of computation needed for playback compact, inexpensive, and easily programmable. We felt that the degrees of freedom available for video display would be greatly enhanced by analogous audio functionality. Specifically, rather than channels corresponding to speakers, audio should ideally be represented as a set of localised sound sources and an acoustical environment in which they are placed. These sound sources would then be linked to the visual objects with which they are associated. As directed by the script, perhaps in conjunction with user interaction, the audio would be 'rendered' for the speakers associated with the video display. Also analogously with the visual case, the auditory synthesis methods are much better understood at this time than are the analysis methods. Until the analysis is better developed, changes in production methods and linkage with the video analysis can assist the process. In our production experiments we have used a separate wireless microphone for each actor, and have used our video analysis methods to provide the locations from which their voices emanate.

Some other considerations

Several current standardisation efforts at least partially echo the concerns and concepts I have explored above. While MPEG-4’s main area of concern is very highly compressed video at low spatiotemporal resolution, proposals have included support for multiple representations and object segmentation. The virtual reality language VRML (and more particularly some proposed extensions) likewise is of interest, though it is not optimised for efficiently representing dynamic real-world scenes.
Knuth's metafonts did more than just respond to the requirements of differing display sizes and resolutions; the language permitted a single, cleverly-programmed definition to generate — for instance — a sans-serif and a roman version of a font, given different parameters. Davenport has looked into visual storytelling systems with a broad range of responses to the viewer, and indeed the mechanisms discussed above can enable the creation of meta-content, not just support for meta-display.

Digital image processing has already enabled easier and more convincing modifications to supposedly veracious images. Object-based representations will certainly advance this trend.

More advanced display technologies can take greater advantage of scene modelling. Stereoscopic displays require at least two views, while holographic displays may require dozens or hundreds. Since the differences from image to image are quite small, it is possible to use a view-to-view predictive coder to represent them fairly efficiently, but under a model-based system, the same scene data given progressively more computation could drive a flat, stereoscopic, or volumetric display — a further sort of scalability.

Conclusions

The coming shift to more meaningful representations for video is likely to occur largely for efficiency reasons. In this paper, I have added intelligent interoperability as a motivation. I could with equal merit have discussed enhanced interactivity, creation of immersive virtual worlds, or assembly of content on the fly through semantic search capability. The preceding improvements would be noticeable to the viewer, while the creator would see a corresponding change in the methods of content capture and tools for editing. In any event, achieving the full potential for model-based scene representations will require dialogue — soon, before standardisation sets in — between the technical and creative communities.

Acknowledgements

The author wishes to thank many co-workers for their invaluable contributions to this work, and Glorianna Davenport, Judith Donath, Suzanne Niell, and the editors and reviewers of this journal for their thoughtful comments on an early draft of this article. This project has been sponsored by the Television of Tomorrow Consortium of the MIT Media Laboratory.

Notes

1 A. Schopenhauer, On Vision and Colours: an Essay, trans. E. F. J. Payne [Oxford UK: Berg Publishers, Inc., 1994], p. 10. Many of Schopenhauer's ideas on vision — especially on colour — were known to be incorrect even during his lifetime, but this quote remains relevant.
