In *Proceedings of IS&T/SPIE's Symposium on Electronic Imaging*, February, 1995.

Structured Video and the Construction of Space

Judith S. Donath

November 1994

Abstract

Image sequences can now be synthesized by compositing elements - objects, backgrounds, people - from pre-existing source images. What are the rules that govern the placement of these elements in their new frames? If geometric projection is assumed to govern the composition, there is some freedom to move the elements about, but the range of placements is quite limited. Furthermore, projective geometry is not perceptually ideal – think of the distortions seen at the edges of wide angle photographs. These distortions are not found in perspective paintings: painters modify the portrayal of certain objects to make them perceptually correct. This paper first reviews projective geometry in the context of image compositing and then introduces perceptually based composition. Its basis in human vision is analyzed and its application to image compositing discussed.

1 Structured Video and Space

In traditional video the frame is indivisible and the sequence – a recording of events occurring in a single place and time – is the basic editing unit. In **structured video**, events in a frame are no longer limited to simultaneous, collocated actions. The basic units in structured video are 2 or 3D image components, rather than the rectangular frames used in traditional video. New image sequences can be made by compositing actors and objects taken from a variety of source images onto a new background. Sequences can be synthesized that depict real, but previously unfilmable events: a discussion among several distant people, or the comings and goings of the members of an electronic community.

Structured video is a way of bringing the versatility of computer graphics to video or conversely, of bringing video's detailed content and ease of creation to computer graphics. Video is a recording medium: it is optimized for input, for gathering detailed images of real scenes. Yet video is difficult to manipulate. The underlying model of the image (i.e., the real-world scene) is inaccessible. Editing and other interactions are limited to the frame level – the contents of the image are fixed at the time

the video is shot. Computer graphics images, on the other hand, are easily modified. The underlying model – the three-dimensional description of the scene and objects – is accessible. The observer's point of view and the location of objects can be easily changed. However, creating the scene is difficult. Every detail must be explicitly added: most computer graphics images are identifiable as such by their artificial simplicity.

Structured video research falls into two areas: image analysis and image synthesis. The analysis problem is one of machine vision: find the actors, find the background, segment out the objects, find the motion. The synthesis problem – putting the image back together, creating and manipulating a new image – is in the domain of computer graphics. The analysis and the synthesis are closely related: the information and structure provided by the former determines the level of control and the parameters that can be manipulated by the latter. If one knows the 3D shape of an object captured in the video, one can do far more with that object than if one has only a mask indicating its location.

So far, most work with structured video has been directed toward compression and very low bandwidth transmission. In these applications the reconstructed image is quite similar to the source image; any discrepancies are due to information discarded in order to maintain a low bit-rate, rather than to a deliberate re-arrangement of the image contents. This paper is concerned with a different type of reconstruction: the synthesis of a new image sequence by compositing images from a variety of sources. By recombining objects and backgrounds, structured video can be used to create realistic looking images of scenes that were never actually filmed. It can also be used to create more abstract images – ones in which the location and appearance of objects is used to convey information.

A synthesized scene need not necessarily look as though it had been recorded from real life. However, it should look like a coherent and unified scene, not a random collage of images. The objects in the scene must relate spatially to each other: an object should appear to be next to or behind of or pointing to its neighbors. Establishing coherent spatial relationships among the composited images is a key factor in ensuring that the resulting sequences are perceptually believable and can convey information effectively. The question of how to create these spatial relationships involves both the geometry of perspective projections and the psychology of visual perception.

2 Representing a 3D world on a 2D surface

Painting, photography and computer graphics are all means for creating 2D images of the 3D world, for transforming world information into a flat representation. With photography, the transformation is inherent in the optical system. With computer graphics, a model optical system is specified and the 2D points are calculated from the points in the 3D world model. With painting, perspective construction techniques are used to determine where objects should be placed and how they should appear.[1]

Photography. The camera shows the world in strict perspective. The optics of the camera ensure that there is a single vanishing point. Generally (though not necessarily) the picture plane is flat, and perpendicular to the principle viewing ray. One does not "choose" to transform the 3D world into a 2D perspective image when taking a photograph: this transformation is an inherent part of the process.

Computer Graphics. Strict perspective is almost universally used (oblique and other alternative projections are sometimes used, but generally for special purpose technical applications such as architectural or mechanical drawings). The standard "viewing transformations" allow more flexibility than a real camera – e.g. the picture plane can be tilted at an arbitrary angle relative to the view normal – but the basic eye/camera model still holds. A picture has a single viewpoint, aimed

1. Linear perspective is not the painter's only technique to represent a 3D scene, but it is the one that deals with the issues raised in this paper. Here, painting will refer to perspective representations.

along a single view vector (the principle viewing ray) and all objects in the image are drawn in accordance with that single point of view.

With computer graphics, perspective and the adoption of a single viewpoint per picture is a choice, rather than an inseparable part of the process. Choosing to render the background of an image from one point of view and each of the foreground objects from others is not much harder than rendering all from a single viewpoint. However, most standardized graphics packages are designed to implement a rendering pipeline that transforms all points in an image according to a single viewpoint. The analogy to camera optics is explicit: Pixar's *RenderMan* system, for example, calls its top-level view the "camera-coordinate system".

Painting. A painter *chooses* to use perspective. It is not inherent in the medium and many techniques for depicting a 3D landscape on the 2D canvas, from traditional Japanese landscapes to multi-faceted Cubist paintings, do not employ this technique. Formal geometric perspective in painting is most closely linked with Renaissance artists, and it is their approach to the problem of depicting depth on a flat surface that will be discussed in this paper.

A painter can choose to use perspective *selectively.* A painting is made iteratively and the mathematical directives of the rules of perspective are balanced against perceptual judgements. Deliberate violations of perspective rules by painters who clearly could have made a strictly perspectival painting but chose not to are especially interesting, for they indicate situations in which what "looks best" is at conflict with geometric projection.

Structured video. Structured video is based on the first two of these methods. Photography supplies the source images. Computer graphics techniques are used to create the synthesized image. Both of these methods assume a single camera (or eye) position: photography does so automatically, computer graphics by convention. Multi-source structured video, however, may combine several viewpoints. Each component image (subimage) may have been shot from a different point of view, with a different lens. In the reconstructed frame, an object may be placed in a location different from its location in the original image. However, it cannot appear just anywhere. Many factors, such as its size, shading, and the angle at which it was photographed, limit where an object can be placed and the relationship it will have with other objects in the synthesized scene.

What are the principles that guide where objects may be placed in the new scene? Projective geometry, as is used in computer graphics, provides one set of answers. To create a new, composite image, one places the source images so that all their viewpoints coincide with the viewpoint chosen for the final image. This alignment determines the placement of the component objects in the new scene. The process of geometrically based image composition is the topic of Section 3.

Projective geometry has two major drawbacks as a basis for image composition. First, the set of possible compositions is quite limited. Although some modifications can be made by applying affine transformations to the source images, the composition of the final image is primarily determined by the projective properties of the source images. Second, it does not necessarily yield perceptually pleasing pictures. It is here that the perspective representation techniques used in painting may be quite useful. Although painting techniques are seldom consulted in the context of digital video, they can provide clues about what makes images look "right" – even if they are geometrically incorrect. Using guidelines based on perceptual factors, as well as geometric rules, yields a different and more versatile set of possible object placements. Furthermore, it "corrects" some of the apparent distortion found in strict geometrical (or optical) perspective. These perceptual issues will be the focus of Section 4.

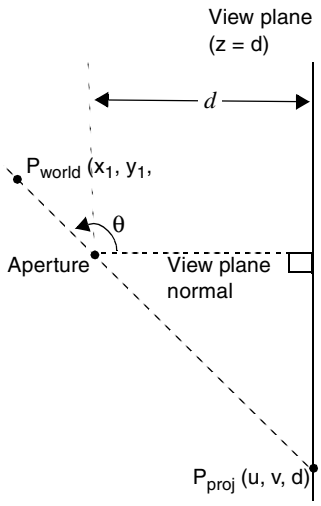


FIGURE 1. Perspective projection.

The perspective transformation

A perspective "camera" consists of an **aperture** and a **view plane** (VP). Light rays from the object pass through the aperture and are projected onto the view plane.[2] The characteristics of the camera determine where a point is projected.The camera can be described by giving the coordinates of the **view plane normal** (VPN), the normal vector from the aperture to the view plane. In Figure 1, point P_{world} is projected onto the view plane at P_{proj}, *d* is the length of the VPN and θ is the angle from the VPN to the vector going from the point to the aperture.

Camera movement

The various types of camera movement can be defined in terms of perspective projection. Camera movements that do not change the aperture, such as zooming and panning, do not change the content of the image, although they may scale and stretch it. **Zooming** changes *d*; it is the equivalent of moving the VP along the its normal vector[3]. In Figure 2, VP_1 and VP_2 are parallel planes with coincident VPNs. The image of the ball projected on VP_1 is a uniformly scaled, but otherwise identical, version of the image seen on VP_2. **Panning** (or **tilting**) changes θ; it is the equivalent of rotating the VP about the aperture. In Figure 2 the image projected on VP_3 is a scaled and skewed version of the projections on VP_1 and VP_2. Panning and zooming do not change the information content of the image: the proportions of the projected image are affected, but points are neither revealed nor hidden.[4]

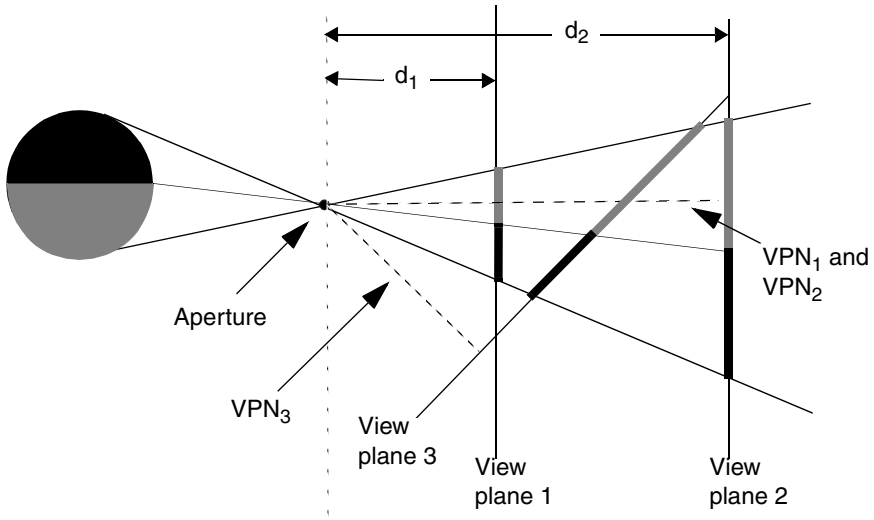


FIGURE 2. Zooming and panning. Only the projection plane moves. The resulting projections are scaled and skewed versions of each other, but they all contain the same information.

2. In practice, cameras and images do not behave quite like the ideal camera described above. The aperture of an ideal camera is an infinitely small point through which rays pass with perfect focus and no optical distortion. The computer graphics "camera" is an ideal one: its aperture is a mathematical point, infinitely small and the computer graphics image has infinite depth of focus. With real cameras, however, the aperture must be large enough to allow sufficient light to pass thru. As the aperture gets bigger the image gets brighter – and blurrier. Lenses make it possible to focus light through a larger aperture, but their depth of focus is still finite and the lenses themselves introduce some distortion.

3. Real cameras accomplish essentially the same thing using more complicated systems of lenses. A big difference between computer graphics and photography is that computer graphics has infinite depth of focus (since its aperture can be infinitely small). Moving the (virtual) projection plane back and forth simply changes the scale of the projection. In the real world, light must be focussed. Moving the projection plane about in a camera is used to focus the image, while the lens characteristics determine the magnification of the projection.

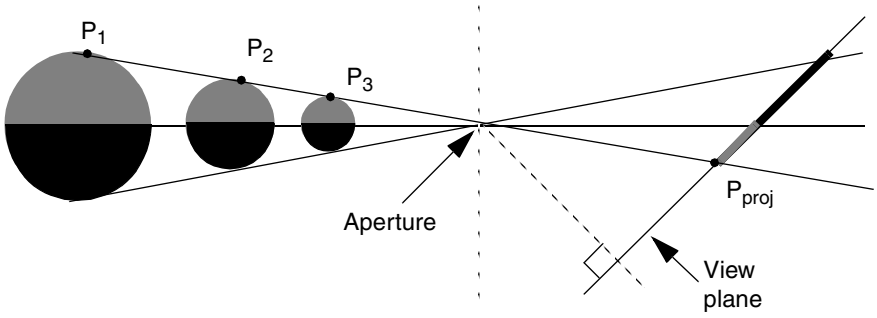


FIGURE 3. Ambiguity: An object will project the same image as will a larger, more distant version of itself.

Camera motions which do move the aperture, such as tracking shots, do change the content of the image. Sometimes it is not the aperture, but the object that moves. Any motion that changes the relationship between P_{world} and the aperture reveals new information and hides other information.

Ambiguity

For a given camera setup, all points that lie on the same point-to-aperture vector will have the same projected point. In Figure 3 P_1, P_2 and P_3 all lie on the same point-to-aperture line and all three points project as the same P_{proj}. If the aperture moves, however, this will no longer be true and they will project as three separate point.

For every projected image of a multi-point object, an infinite set of scaled and translated objects could have been the source. The projection of the shown in Figure 3 could have been made from any of the balls. There is no way of determining from the projection alone, which ball was the source object and where it was located.

For machine vision, this ambiguity is a serious difficulty. For structured video, it is a great advantage. It means that an image of an object can be reused in multiple scenes. Some reuses can be done with no change to the image: a large object at a great distance can substitute for a closer, smaller one. The ambiguity remains if the view plane moved: such changes do not alter the image

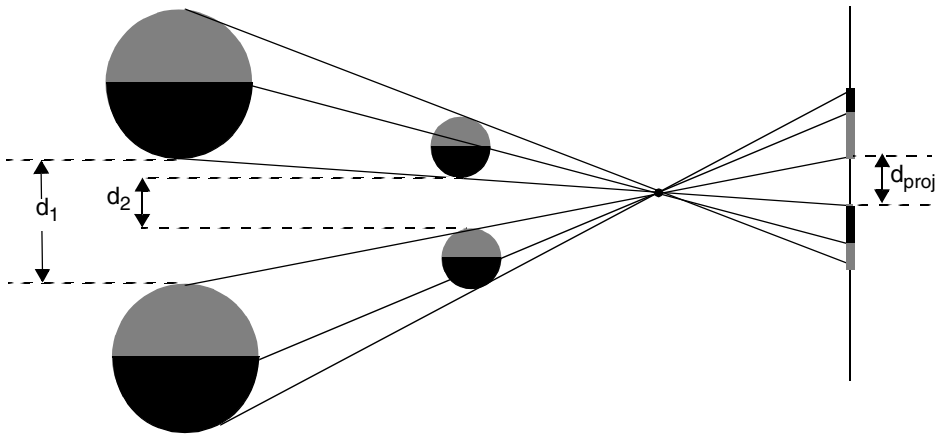


FIGURE 4. Motion. In order to travel the same projected distance the the larger, farther object must actually travel further.

4. Real images have edges. The real image plane is a clipped window onto the infinite view plane. Thus panning and zooming reveal new information at the edge of the images. (Even in the ideal case, the maximum angle of view is 180° and thus panning will always reveal new data.) These camera motions affect how much and what part of the image is within this clipped window; the full view plane image is unaffected.

Real images are sampled – their resolution is finite. The magnification of a small image will have less information than one that was photographed at the larger size.

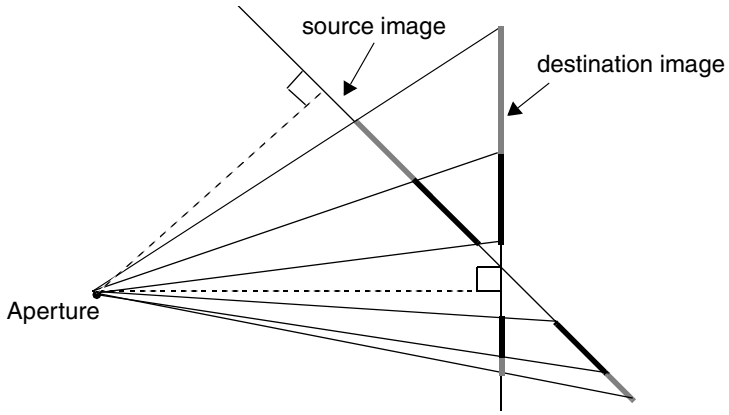


FIGURE 5. Reprojecting an image.

content. Thus, the substitutions (of a large object for a small one, etc.) can be made into scenes of varying magnification and VPN angle. (See McLean 91 for a discussion of the reprojection process).

Motion

Figure 4 shows a pair of objects that are scaled versions of each other. In both positions shown, they form the same projection. However, in order to travel the projected distance d_{proj} the larger, more distant object needs to move a greater distance. A large distant moving object may produce the same images as a smaller, closer, object moving parallel to it, but at a slower pace. (This is why far-away objects appear to move slowly, even if they are actually travelling at great speeds). Similarly, a nearby object moving away from the aperture will shrink in size rapidly, while a far away one would need to cover a much greater distance to show the same effect.

Compositing

If an object is photographed with the intention of using it as an element in a composited image, what geometrical data is needed so that it can be correctly placed in future images?

Minimal geometric description

The minimal amount of information that is needed to ensure that the resulting composition is geometrically consistent is the camera description. By definition, the projected image lies on the view plane. What is needed is a description of the camera: the coordinates of the aperture and the VPN. Given this information, the image can be used as a component of a new, composite image.

To make the composition, the aperture of the source and destination images must be made to coincide. The view normals, however, need not be colinear. The source image can be thought of as a new object in the scene, one that is subject to the same projective treatment as any fully 3D object. As long as the apertures are coincident, an image can be reprojected onto any other image, provided that second image is of sufficient extent[5]. Figure 5 shows this reprojection. If the destination image is to be constructed in strict accordance with the rules of projective geometry, the only variable is the rotation of the source VPN about the aperture. For any given rotation angle, there is a single reprojection of the source image: it cannot be arbitrarily located or scaled.

Other geometric data

As we have seen, an image is ambiguous: it can depict a small close object or a farther, larger one. It may be necessary to resolve this ambiguity in order to place the image in a destination scene. If the

5. Even with ideal, infinite images, if the view normal of the object and of the new image are not parallel, it is possible that the object's location is outside of the viewing hemisphere of the new image.

Not only does our ideal camera have infinite focus, our ideal view plane is of infinite extent and resolution. With a real image, enlarging the image will lower the resolution.

image is to share the destination picture with other object images, the depicted object's location in 3D space must be defined. This 3D location will determine what it occludes and what is occluded by it (or whether it has been impossibly co-located with another solid object).

Summary of geometrical composition

The first part of this paper has examined the geometry of structured object placement. We have seen so far that it is possible to remove an object from one image and place it in an infinite, though highly constrained, set of locations in another image. Using projective geometry as the basis for compositing images provides a clear set of functions for determining the placement of objects cut from one image into a new one. Images created with this technique are optically consistent: they *could* have been taken with a camera, given the (possibly scaled) original objects. Thus far, the underlying assumption has been that, when making realistic 3D images, the rules of projective geometry determine where objects may be placed. In the second part of the paper, we will examine that assumption more carefully.

4 Perceptually based image composition

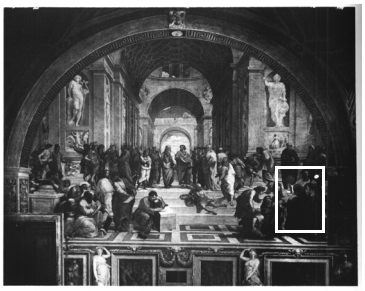
> Photography does give pictures which are in strict geometrical perspective. The results can be most disappointing. [Gregory 70]

What makes an image appear realistic? What makes it look "correct"? To answer these questions geometry alone is not sufficient: an image may be a true geometric projection of a scene, and yet appear strangely distorted and not at all realistic. And geometry is not always necessary: there are pictures in which the geometry of the projection has been altered (in particular, that use multiple viewpoints) and yet which look quite realistic. Structured video starts with recorded images which are geometrically projective. However, the synthesized sequences, the final output, need not be constructed in strict accordance with projective geometry. What is known about the way images are perceived that can provide guidance in constructing these sequences?

In Section 2 we mentioned that painters, even those painting in the realistic perspective tradition, do not produce paintings that are as geometrically consistent as a photograph. In particular, it is interesting to note their treatment of figures. A camera produces an image with a single viewpoint. Objects in the image near that point are shown with proportions similar to those the object has in real life. Those farther from the center, however, tend to appear stretched and distorted[6] – think of the unfortunate people portrayed near the edges of wide angle photographs! Figures in paintings, however, do not generally show this distortion. For example, in Raphael's *School of Athens* (Figure 6) all the figures are shown as if in their own center of projection. Those that are farther away are smaller, but there is no distortion of the proportions. Figure 6 shows a detail taken from an area away from the center of the image. In a photograph, the figures would be wider here than in the center, an effect of the projective geometry. In the painting, the figures' proportions are undistorted – they could be moved from one part of the painting to another without change. More importantly, the painting as a whole looks "correct". One does not feel that it would look better, or more realistic, if the figures were drawn in accordance with geometric projection.

Synthesized video images can be made using similar rules of construction. The use of camera-like geometry in their production is a convention, not a necessity – composited images need not inherit the marginal distortions of traditional photography. Geometric projection provides a very clear and unambiguous set of rules for how objects are projected and where they can be used. Yet the images produced by these rules may be perceptually unsatisfactory. Perceptually based composition is not as precise – it offers guidelines, not laws. What is known about the perception of images that can provide guidance in creating well-composed pictures and sequences?

6. "Distorted" refers to their perceptually qualities - as projections, they are perfectly correct and undistorted.



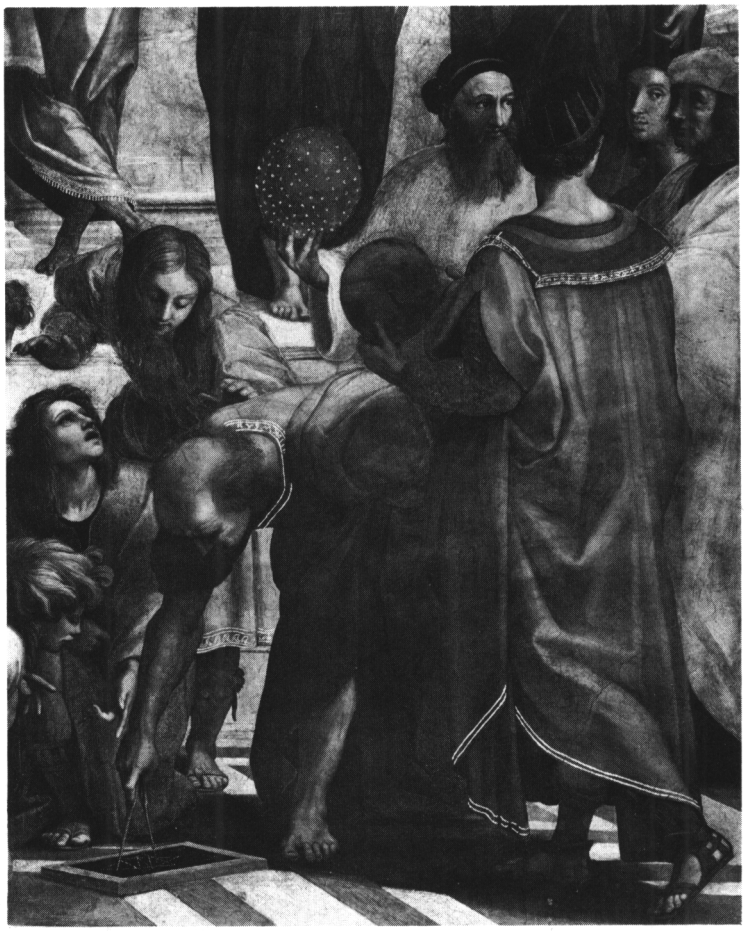


FIGURE 6. (Right) Detail from Raphael's School of Athens. *The detail is taken from an area away from the center of projection (shown by the white box above). In a true projection the sphere would be shown as ellipse; the human figures would evince similar geometry. Instead each figure is portrayed as if it were in the center.*

Seeing the image

A perfectly realistic image is an exact substitute for the real scene, one which (allowing for some variation in brightness and tonal range) presents to the eye precisely the same array of light and dark sensations as does the actual scene. A photograph or other such geometric projection can play this role – on the condition that the eye coincides exactly with the image's aperture or view-point. For every single view-point image, therefore, there is one and only one viewing location from which it presents a perfectly realistic view. From every other location, the image that falls upon the eye is a distorted version of the recorded scene. This fact was noted by Leonardo da Vinci, who wrote:

> If you want to represent an object near you which is to have the effect of nature, it is impossible that your perspective should not look wrong, with every false relation and disagreement of proportion that can be imagined in a wretched work, unless the spectator, when he looks at it, has his eye at the very distance and height and direction where the eye or the point of sight was placed in doing this perspective.[quoted in Kubovy 86, pg 52]

FIGURE 7. Image rotation with and without a frame.

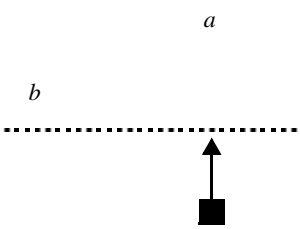
The argument, although sounding plausible, is, of course, wrong[7]. If it were true, one could look at a picture only from a single standpoint. Yet we are seldom constrained to view pictures from such a limited point of view. We are free to walk around them, to look closer and move away. From all these other vantage points, even though the eye is not receiving a simulated view of the world, the image still looks "correct". The picture does not serve as a direct substitute for the real scene; the aperture is not a surrogate for the eye. The image's realistic appearance is not due to exact mimicry of the optical sensation of the real scene, but to more complex cognitive processing.

The role of the frame

A perspective picture continues to represent the same scene as the viewer moves around it – a phenomenon that Kubovy refers to as the "robustness" of perspective. In order for the perceptual system to maintain this consistent interpretation, it is necessary for the surface of the image to be visible: the viewer must be able to infer the relative rotation of the picture plane. One way for the surface to be visible is for the rectangular frame to be visible. As we shall see, there is a strong tendency to interpret shapes that *could* be right-angled objects as such. Thus, a trapezoid is most likely to be seen as a rectangle rotated in space. A rectangular picture, seen from an angle, is trapezoidal in shape. It will be interpreted, correctly, as a rotated rectangle, giving the perceptual system the information needed to compensate for the rotation. If the frame is not visible, and there is no other means to determine the rotation of the image, the picture will appear distorted. In Figure 7, the top picture is the original and the other two are rotated versions of the same image. The middle image appears twisted; even knowing the extent and direction of the rotation does not make it possible to compensate for the displacement. Below it is the same image, with the frame visible. Here, it is possible to perceive and compensate for the rotation.

It is perhaps more exact to say that a perspective image is interpreted as a series of similar scenes as the viewer walks around it. The disparity between the images is usually not noticed. However, one striking manifestation of this is the phenomenon of "following". As you walk past a portrait that looks straight out, the eyes appear to follow you around the room. Similarly, roads that go from the front of the picture plane (that is, the bottom of the image) back towards the horizon appear to change course as one walks around them; pointing fingers (i.e. Uncle Sam) point at the viewer in any location. This phenomenon can be seen here in Figure 7, where the eyes always appear to face out from the page. The perception of the phenomenon is peculiar. It does not seem as if the actual scene is changing and the picture still appears to represent the same scene – yet there is a sensation of rotation. Kubovy ascribes it to the combination of physical movement about the image (which is irrelevant for pictures such as those shown here) and the comprehension of an unchanging scene. The diagram accompanying Figure 7 suggests that it may be a the result of a plausible mental model of the underlying formation. If the top, undistorted image was cut out and reprojected at position *a,* it would remain undistorted. If it were photographed at position be, it would appear as the middle and lower image do - *if* it were rotated to face the camera.

Finally, any lingering sense that an image appears realistic because it affects the eye in much the same way as the original scene is dispelled upon looking at it from different positions. Movement parallax, as Hochberg points out (Hochberg 78), is one of the strongest depth cues: and it entirely absent from a 2D images. Our perception of realism in pictures and films occurs in the context of a number of cues telling the perceptual system how different are the image and the real thing.



7. Leonardo was right, however, if by "an object near you" he was referring specifically to images that subtend a large angle – larger than approximately 35 degrees horizontally and 28 degrees vertically, the range encompassed by the human visual field [see Kubovy 86, pp 104-110 for a review of supporting experiments]. Images with such wide angles do show "every false relation and disagreement of proportion", until viewed from a location close to their original aperture.

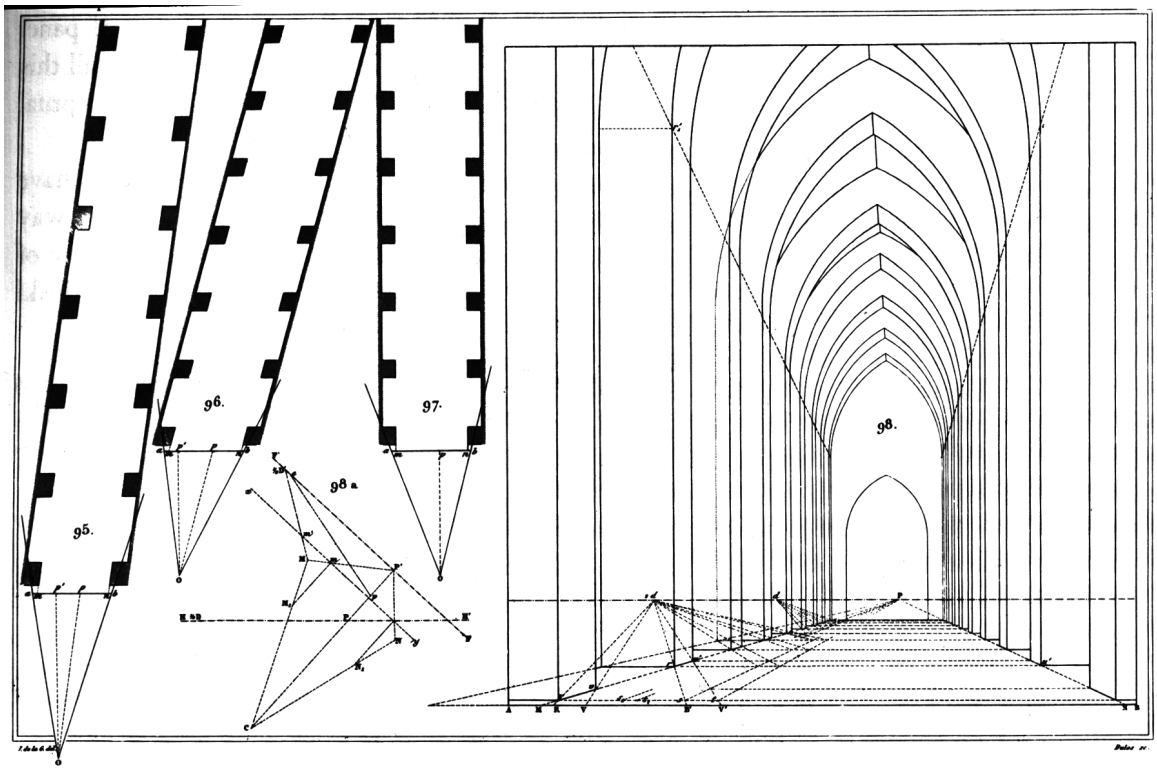


FIGURE 8. From La Gournerie, Traite de Perspecive Lineaire. All three plans are possible interprestations of the image to the right. Almost all viewers were interpret the image according to plan 97, with its regular and rectangular layout.

The resolution of ambiguity

As we have discussed in Section 3, a particular projective image could have been made by an infinite number of possible object and camera combinations. Yet for most images, only a single interpretation is reached. Part of the work of disambiguating an image is to infer the viewpoint from which the scene is shown. A photograph or a picture drawn in perspective has, as part of its structure, the point of view from which the scene was captured (what we have been calling the aperture) If the viewer of an image makes an incorrect assessment of where the viewpoint is located, interpretation of the entire scene will be distorted. The fact that we are able to see and understand pictures from a variety of viewing positions is evidence that finding the viewpoint is not an impossible task. However, it remains far from understood exactly how the human visual system achieves this –and it remains an unsolved problem in machine vision.To solve the problem of inverse perspective requires more than geometry; it requires some knowledge of the scene.Figure 8, from a19th century treatise on perspective, shows several possible interpretations of a perspective drawing. All are geometrically valid solutions, yet a viewer will interpret the image only in accordance with the plan labeled *97*.All the plans interpretations based on similar viewpoints, *95* a little to the right, *96* more to the left. The difference is that the viewpoint in *97* is consistent with the most perceptually believable ground plan, one that is based on rectangular forms and right angles.

Rectangular objects and the Ames experiments

In its effort to disambiguously interpret a visual image the visual system makes assumptions about the nature of the scene and the objects depicted in it. The depth cues, such as the assumption that occluding objects are in front and that objects that are decreasing in size are at increasing distances, are examples of such cues. A related phenomenon is the assumption of rectangularity: when confronted with an angular object that *could* be interpreted as a cube or other right-angled form, the visual system interprets it as such, even when that causes other inconsistencies in the scene interpretation

The Ames room (Figure 9) provides a striking example of this.It is a trapezoidal room designed to create the illusion that, from a certain vantage point, it is a regular right-angled space. Were the viewer to see it as it really is, an oddly shaped space, the people in would appear to be normal sized. Yet the visual system persists in seeing it as a regular rectangular room, even though means the people standing in it appear to be of wildly varying heights. If the people in the room move about, the room still seems rectangular – and the people appear to change size as they move.The Ames room shows how strongly the mind's assumptions about the shape of objects affect the interpretation of visual space. It is not known precisely what causes the Ames room illusion, but there does seem to be an "assumption of rectangularity". (see Arnheim 74, Hochberg 78, and Kubovy 86 for a range of interpretations).

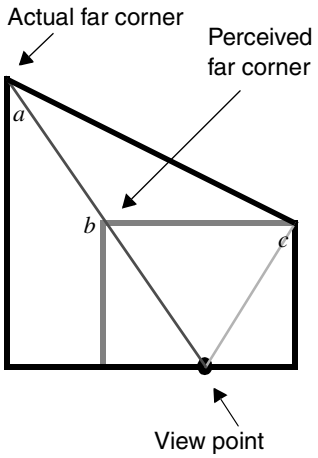


FIGURE 9. The Ames room.

A figure standing in the actual far corner (a) will be perceived as being in the closer illusory corner (b) and thus expects this figure to be similar in size to a figure standing at (c), thinking the two figures are equally far away. The one that is perceived to be at (b), however, is actually much farther away and thus appears to be very small.

Objects with right angles provide a powerful means for the visual system to disambiguate the source image, to decide upon a single 3D interpretation for the depicted scene. Once these objects are established, the location of the image's viewpoint can be surmised. It is the apprehension of this viewpoint that allows the viewer to look at a perspective image from a variety of locations,

Spherical objects and human forms

So far, our discussion of image perception has focussed on shapes with right angles: on the rectangular frames that indicate the tilt of the picture and the cubical objects that provide the cues necessary to disambiguate an 2D images. We will now turn to the depiction and perception of spherical objects, especially human figures.

Images of spherical and of rectangular objects are perceived differently. Painters, as we have mentioned, have long recognized this, and they often paint spheres (and more commonly, figures) from an individual center of projection. Looking again at Figure 6, notice that the architectural elements are portrayed from a single view-point. Only the sphere (which in a true projections would be a sphere) and the figures are placed in individual projection centers Why should this seemingly idiosyncratic approach, treating some objects one way, others another, work? What are the underlying perceptual factors?

As we have seen, the perception of a seemingly rectangular object helps the viewer determine the 3D source of a 2D image. Rectangular objects contain geometric information: the projected angles can correspond only to a single object-camera relationship, assuming that the real angles are right angles. Spherical objects do not have these visible structural cues. The projection of an off-center sphere is an ellipse and it is perceived as such. With a rectangular figure, the viewer's perception of the scene as a whole changes to maintain the cubical shape; with the spherical figure, the viewer's interpretation of the object changes.

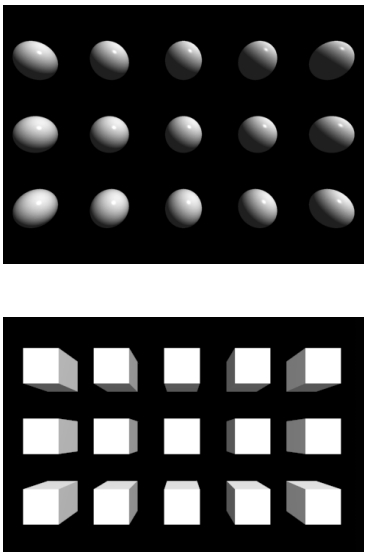


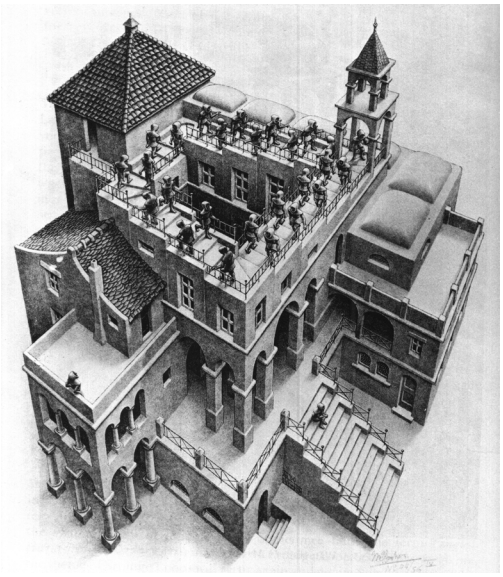
FIGURE 10. Perspective images of spheres and cubes

Compare the images in Figure 10. As a cube moves further from the central axis, more of its side face is visible. The viewer is aware that this added region is a different facet, and that the object has not changed shape. With the sphere, an analogous projection effect occurs. However, there is no demarcation to indicate that the added area is a side view and the object itself appears to be elongating.

As we shall see in the next section, we do not really "see" things at off-center projections: the limited angle of the visual ensures that we are looking at things when we see them. Thus, none of the extreme projections in Figure 10, neither cubes nor spheres, shown in are actually familiar sights. The cubes, thought, closely resemble cubes we have seen: those that fall within the visual field and are at an oblique angle. The elongated spheres, however, are completely outside our viewing experience: to the eye, a sphere always appears as a sphere.

Seeing the whole picture

In the previous section we discussed creating images with internally inconsistent projections. If images were viewed as a single, united entity, this might cause problems. However, images are seen in parts. An image may be quite inconsistent internally, and still retain its three dimension



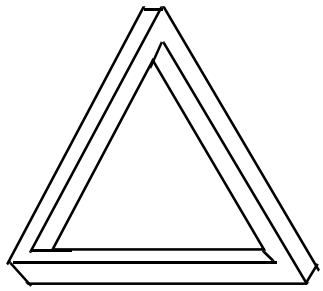


FIGURE 11. Inconsistent images that retain their appearance of depth, by Escher and after Penrose.

appearance. The images in Figure 11 are good examples of this. Although it is clear to the viewer that the depicted objects are "impossible", the image still seems to be a realistic depiction. Hochberg explained this with a hypothesis about perception and memory:

> the inconsistent regions of the picture are not normally compared to each other directly… any object is usually examined by a succession of multiple glimpses, and the various regions that are looked at each fall in turn on the same place in the eye. That is the separate parts of the figure all have to be brought at different times to the central part of the retina, the *fovea,* if they are to be seen in full clarity of detail… What we perceive of the world is determined therefore both by the processes that guide fixation and by those that determine what we retain from a sequence of fixations… it is now evident that we cannot make a full accounting of pictorial representation in terms of… [any] analysis that restricts itself to discussions of the stimulation of the visual system. [Hochberg 72]

Inconsistent images highlight the fractured way we look at images. All images, not only the inconsistent ones, are seen as a combination of parts. Looking at an image is an active process. Gaze and attention are driven by the both interests of the viewer and the structure of the picture.The purely projective image is made for viewing as a whole; it is the reproduction, perfectly made, of a single glance. The perceptually based construction is made for viewing successively. The the architectural elements, the rectangles, the cubes, are for the big overall view, for the glance that takes in the scene as a whole. Then the eye wanders about, glancing at the figures and faces. The rendering of these forms "undistorted" by projection lets them be viewed individually, as closeups within the larger scene.

Motion

Although the topic of the paper is structured *video*, the primary focus has been on still images. Do the same principles hold true for moving images as for stills?

It seems reasonable that they should. A single frame of a movie is a still photograph[8]. Certainly the principles of projective geometry are the same for both. But what about the perceptual techniques? Will they also work in motion sequences or be the cause of unpleasant artifacts?

8. For this paper we will ignore the differences between film and video, since the focus is on geometry rather than actual rendering.

For still images, there are painted examples of perceptually based projection. There is no motion equivalent that I know of. Without actually creating some sequences we cannot say for sure that it would work, but there do not seem to be obvious reasons why it should not. The most likely source of problems would come from size inconsistencies. With the purely projective techniques, once the equations for translating one projection to another are set, the size of the object should remain consistent throughout its motions. For perceptually projected images, the desired height (since projected width is one of the "artifacts" we are trying to eliminate) must be determined for the object in each location and the projection scaled appropriately - this scaling factor may vary with each location.

5 Object based image construction

One of the basic themes that has emerged in the preceding discussion of image perception is that we see different types of objects differently. Rectangular objects (and any thing the visual system suspects of being a rectangular object) provide important cues about the orientation and that structure of the image. Rectangular objects in the image are used to determine the viewpoint from which the scene was create; an image that is itself rectangular provides the viewer with perceptually important information about the orientation of the physical image. The visual system can interpret a skewed right angle: it adjusts its perception of the image. Spherical objects, on the other hand, are subject to apparent distortion whenever they are projected away from the center of projection. Their structure provides them with no internal clues as to their actual shape, and thus their 3D projection is quite ambiguous. Painters have handled this dichotomy in object perception by rendering rectangular and spherical objects differently: the former according to the rules of geometric projection and the latter as if the view normal was centered upon each of them.

As an approach to rendering structured video, this has several interesting ramifications. First, it allows composited images to avoid the distortions common to all photographic images (such as people near the edges appearing rather squat and stretched). Second, it can simplify the process of shooting images for use in later compositions.

A basic structured video technique is to divide the image into foreground and background – a division that generally parallels the perceptual division of people (spherical objects) and architecture (rectangular objects). Using a perceptually based approach one would choose the rendering technique based on object description. Angular objects, and the whole back ground, are rendered according to the images main view vector. Figures and other non-rectangular forms are rendered with their own center of projection. Object shape classification –rectangular or spherical – needs to be stored with each image.

When shooting footage of people for use with this technique, the subject is always kept in the center of the image. (This is a reasonable thing to do even if shooting footage for use with strictly geometric rendering, if it is not known in advance exactly where the image will be placed on a frame. Central shooting will require some transformations to achieve geometric correctness, but the process will be simpler, and the resulting image better, than a translation from one off-center projection to another). The only transformation that will be needed is uniform scaling. The calculation of what view of the object should be presented is the same as with geometric projection. (See Figure 12) The only difference is the orientation of plane onto which the image is projected

What is a realistic image?

The geometry of representational images ranges from the mathematical projections of photography to the multifaceted figures of Cubism. Simply defining what is a "realistic" image has occupied philosophers (and more recently, psychologists) for centuries in debates about whether realism is an objective quality or a cultural phenomenon (see Black 72, Gombrich 69, Hagen 86). Were the ancient Egyptians painting realistic pictures in a culture that viewed images differently and for

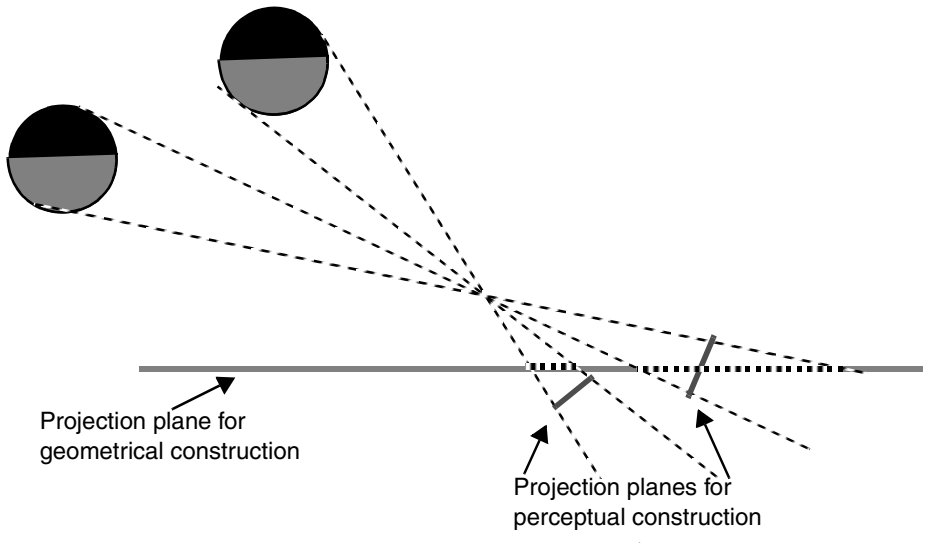


FIGURE 12. Geometrical and perceptual projections. The projected information is the same - the difference is in the orientation of the projection plane.

whom Western perspective drawings would thus have appeared distorted? or was realistic depiction simply not their goal? Was the development of perspective in 15th century Italy the *invention* of another technique – or was it the *discovery* of a true way of reproducing the visible landscape, "unique in giving a true account of the world" [Willats, pg. 236]. The eye is often described as being similar to a camera. Is there a close analogy, thus conferring upon the photographic image the designation of truly "realistic" or is it a misleading characterization and the photograph is also "one system among many" for depicting the external world?

View-based and object-based representations

One way of approaching the question of "what is a realistic image" is to make a distinction between object-based and view-based methods of depiction [Willats 90, Arnheim 74 on Form]. Object-based systems convey the actual size and shape of the depicted object. Architectural drawings, which use parallel projection to maintain true relative edge lengths, are an obvious example; traditional Chinese and Japanese paintings use similar techniques. View-based systems, including perspective painting and photography, show how an object or scene would look from a particular viewpoint; they illustrate the act of seeing the object, rather than the object itself. The question of realism thus depends upon what the image means to depict: the form of the object or its appearance from a specified place.

Distinguishing between viewer- and object-based representations is quite useful in the context of structured video. The structuring data may be object-based, such as a 3D model (Holtzman 91) The information encoded in this model is a description of the object itself. It does not include a particular viewpoint – that information is indeterminate until the object is rendered. Or the structuring data may be view-based, such as the 2D pictures of objects we have been discussing. The information about the depicted object is in relationship to a particular viewpoint (specified by the aperture and VPN). A view-based representation alone does not provides absolute information about the object, such as its size or shape.

Traditional video provides only view-based, perspective images. Structured video can, however, be rendered in either a view- or an object-based style (given the necessary additional data). A view-based rendering is well suited for showing objects in relationship to each other, within the context of a scene. However, perspective rendering does "distort" the shape of objects, and makes visible size an ambiguous combination of actual size and location. For certain applications, such as some visualizations, an object-based rendering technique that clearly shows the objects' properties may be

a better choice. The trade-off is that it cannot also convey the sense of depth that a perspective image does.

Beyond realism

The video frame is no longer indivisible. With the ability to manipulate objects within the frame, the spatial dimension can be used to express more than just the physical relationship between objects in a scene. The size and location of an object's representation in an image can have meaning beyond its physical extent: they can be dimensions indicating importance, activity level, number of frozen dinners bought. Yet the 2D image is limited in how much data it can express. Unambiguously representing three dimensions is already beyond its ability; each symbolic dimension undermines its ability to display physical reality.

Giving up the representation of physical reality need not entail also losing the spatial cohesion of a three dimensional image. This paper has explored the nature of projective images, discussing both their geometry and their perceptual qualities. The goal has been to learn how to use them to construct believable spaces, regardless of whether the reality they represent is visual or symbolic.

Bibliography

Adelson 91 — Adelson, Edward H. “Layered representations for image coding.” MIT Media Laboratory Vision and Modeling Group Technical Report No. 181, Dec. 1991.

Arnheim 74 — Arnheim, Rudolf. *Art and Visual Perception.* Berkeley: The University of California Press, 1954; revised and expanded edition, 1974.

Black 72 — Black, Max. “How do Pictures Represent.”. In *Art, Perception and Reality*, pp. 95-130. Baltimore: The Johns Hopkins University Press, 1972.

Blake 90 — Blake, E.I. “The Natural Flow of Perspective: Reformulating Perspective Projection for Computer Animation”. *Leonardo*, Vol. 23, No. 4 1990.

Descargues 76 — Descargues, Pierre (intro. & commentary); Allison, Ellyn Childs (editor); Paris, Mark I. (trans.); *Perspective.* New York: Harry N. Abrams,Inc.1976.

Foley 90 — Foley, James D.; van Dam, Andries; Feiner, Steven K.; and Hughes, John F. *Computer Graphics Principles and Practice.* Reading, MA: Addison Wesley Publishing Co., Inc., 1990.

Gregory 70 — Gregory, R.L. *The Intelligent Eye.* London: Weidenfeld & Nicholson, 1970.

Gregory 78 — Gregory, R.L. *Eye and Brain*. 3rd ed. New York: McGraw-Hill Book Co., 1978.

Gregory 90 — Gregory, R.L. “How do we interpret images?”. In *Images and Understanding,* pp. 310-330. Edited by Horace Barlow, Colin Blakemore and Miranda Weston-Smith. Cambridge: Cambridge University Press, 1990.

Gombrich 69 — Gombrich, E.H. *Art and Illusion*. 2nd ed. Princeton, NJ: Princeton University Press, 1969.

Gombrich 72 — Gombrich, E.H. “The Mask and the Face: The Perception of Physiognomic Likeness in Life and in Art.”. In *Art, Perception and Reality*, pp. 1-46. Baltimore: The Johns Hopkins University Press, 1972.

Hagen 86 — Hagen, Margaret A. *Varieties of Realism: Geometries of Representational Art.* Cambridge: Cambridge University Press, 1986.

Hochberg 72 — Hochberg, Julian. “The Representation of Things and People.”. In *Art, Perception and Reality*, pp. 47-94. Baltimore: The Johns Hopkins University Press, 1972.

Hochberg 78 — Hochberg, Julian 1978. Perception. 2nd ed. Englewood Cliffs: Prentice Hall.

Hochberg 86 — Hochberg, Julian. “Representation of motion and space in video and cinematic displays.” In *Handbook of perception and Human Performance,* vol. 1, pp. 22.1 - 22.64. Edited by Boff, K; Kaufman, L.; and Thomas, J. New York: Wiley, 1986.

Holtzman 91 — Holtzman, Henry Neil. “Three-Dimensional Representations of Video Using Knowledge Based Estimation.”, Master’s Thesis, MIT, 1991.

Kubovy 86 — Kubovy, Michael. *The Psychology of Perspective and Renaissance Art*. Cambridge: Cambridge University Press, 1986.

McLean 91 — McLean, Patrick Campbell. “Structured video coding.” Master’s thesis, MIT, 1991.

Pirenne 70 — Pirenne, M.H. *Optics, Painting and Photography.* Cambridge: Cambridge University Press, 1970.

Stephenson 76 — Stephenson, Ralph and Debrix, J.R. *The Cinema as Art.* 2nd Edition. Middlesex: Penguin Books Ltd., 1976.

Watlington 89 — Watlington, John A. “Synthetic Movies.” Master’s Thesis, MIT, 1989.

Willats 90 — Willats, J. “The draughtsman’s contract: how an artist creates an image.” In *Images and Understanding*, pp. 235-255. Edited by H. Barlow, C. Blakemore and M. Weston-Smith. Cambridge: Cambridge University Press, 1990.

Vries 68 — Vries, Jan Vredeman de. *Perspective.* New York: Dover Publications, 1968. New intro by Adolf K. Placzek. Unabridged and unaltered republication of work originally published by Henricus Hondius in Leiden in 1604.