Chapter 1

Dialogue 2: The State of the Art

1.1 Eyes

Arc: Welcome, Melambus. I will tell you a few secrets I know about measuring the world from images. Secrets about light.

Mel: Light?

Arc: Light! Light! It's everywhere you look. Every piece of our world, no matter how small, is constantly bombarded by photons, little particles traveling in straight lines at the highest speed the universe allows. They bump on the objects around us, dancing through holes and cracks in a continuous straight line movement. These straight lines are the *light rays*. Eyes and cameras capture some of these rays and make images, one of the most fundamental bases of our knowledge.

Mel: What does it mean to capture rays? What do eyes and cameras do exactly?

Arc: To capture a ray means to measure a ray's properties. These are physical and geometrical properties. Photons are absorbed by some materials more than others and reflected by some materials more than others. So the ray that hits the surface of an object and subsequently comes to an eye or a camera where it is captured, brings with it some property of the object's surface.

The ray has intensity (few photons, many photons), color, polarization. It also comes from a direction in space, and this is a geometric property. Thus, to make an eye or a camera you must have some elements that can capture the rays. These are the photocells or photoreceptors. Evolution has come up with many ingenious designs for photocells and engineers have designed interesting artificial ones. The next step in making an eye is to place the photoreceptors on a surface. This surface will be the retina. Advanced eyes like our own have millions of photocells densely packed like pile in a carpet. Other eyes do not have so many, but if you observe the eyes in our world you will find that photocells are not uniformly distributed on the retina; instead, there are areas where there is a high number of them. These are the foveas. To simplify things for the moment, let's distribute the photocells regularly over the retina; and since we will mostly talk about geometric properties of the rays, let's consider that at each point of the retina there is a photocell, and we will not mention them again.

Mel: But there seems to be a problem here. From every point of an object there are many rays coming out, one in each possible direction. So, if I take a retina and face it toward the object, at every retinal point there are many rays coming from different parts of the object. It is as if many images are formed on top of each other. I wonder then, what can be measured.

Arc: The trick is to eliminate all these images except one. That's the next step in making an eye. Enclose the retina and leave only a small opening, a pinhole. Now, each retinal point captures a single ray coming from a single object point. We can measure the ray's properties and also find out from which direction it came. That's the pinhole eye, glorious and simple. Actually, in reality it is somewhat more complicated. If the pinhole is too small, not many photons reach the retina and the image is poor, fuzzy and dark. The trick is to make the hole larger; now, from this pupil more light goes through, but we must make sure that all rays coming out of the same point in the world are captured by the same retinal photocell. This way we get a bright, sharp image and at the same time we retain an understanding of the rays' geometric properties. This is accomplished with the help of a lens, appropriately placed at the aperture. Now you have the modern eye of land vertebrates, like our own. The introduction of the lens, of course, gives rise to many features in advanced eyes, like the machinery for changing focus and the size of the pupil. To focus rays coming from a distant target you need a weaker lens than to focus rays coming from a close target. The same lens can achieve both when it can change its focal length, but putting aside these complications, from a geometrical point of view this eye or camera is equivalent to a pinhole eye.

Mel: If I understand you then, your point is that in making an eye, the fundamental thing is to first pick a point and consider all the light rays coming to this point. Then you consider an imaging surface (the retina containing the photocells) and you cut the rays with that surface. The intersection of the rays with the surface gives rise to the image.

Arc: Exactly! If you cut the rays with a plane you get the so-called camera-type eye or simply planar eye. You can only cut a part of the rays, so this eye has a limited field of view, but you could also consider a surface that cuts all the rays. An example would be a sphere with center the point where all rays come together. This is the so-called spherical eye or full field of view eye. It is of course very difficult to build an eye like this, but nature has come close to it with approximations in the eyes of many insects and crustaceans, the so-called compound eyes. For now let's concentrate on camera-type eyes. Pick a point. This is the point at which you will consider all incoming rays. We call it the eye center or camera center, or simply the center of projection. Now pick a plane that cuts some of the rays. This will be the imaging plane or image. You can now start studying the geometry of visual space-time.

Mel: Hmm! Don't I need to know the relationship between the camera center and the image plane? It seems to me that if I eventually need to measure things in the world I would need to know the angle between different rays; for this, I would need to know the distance between the center and the image plane. I would also need some way to measure distances on the image and that in turn would depend on how the photocells are placed.

Arc: This is true! If you don't have any of this information then you have what is known as an uncalibrated eye or camera. But if you know the relationship of the camera center to the image plane and you also have a way to denote different locations on the image precisely, then you have a calibrated eye, i.e., you know the eye's calibration parameters. Simply put, you have a calibrated eye when you know the ray defined by the center and any image point.

Mel: But aren't all eyes calibrated?

Arc: Eyes in nature are more or less calibrated, because a calibrated eye is more powerful as you can easily imagine. But suppose that I give you a video taken by a camera whose calibration parameters are not known. When you begin analyzing the images in the video, you start with an uncalibrated eye. There is, however, a more important reason why we should examine these kinds of eyes. These eyes are pure projective devices. Being so general and simple they have the potential of giving us some general principles that will be present in any eye, calibrated or not.

Mel: I see. Even if I don't know the angles between the rays, I know where they cut the image plane, and this could be enough information. Besides, this process of imaging is equivalent to the central projection about which much has been known since the time of the ancient Greeks.

1.2 Central Projection

Arc: Exactly. Let's look at the properties of this projection as they relate to our problem. Remember our basic goal. As we look at the world from different views, we want to find maps that relate the images taken from different viewpoints and use these maps to find models of the scene. The world looks different from different viewpoints. All we have is the images. If we ever hope to find model of the world, we must understand how the different images are related to each other. Let's do something simple first. Keep the camera center fixed and move the image plane to another position. Each of the planes cuts every ray at a point which is the image of the point in the world from where the ray is coming. As you see, the images that you get are different, but they are related in an interesting way.

Mel: They appear quite different indeed. Lengths are different ($AB \neq A'B'$), even ratios are different ($AC/AB \neq A'C'/C'B'$). The angles are different as well. Is there something that remains the same in both images?

Arc: Yes, there is. It is the ratio of ratios! It is known as the cross ratio. Notice that, because of the central projection, if three or more points lie on a straight line in the first image, the corresponding points will also lie on a straight line in the second image. Now take four points A, B, C, D and A', B', C', D'. The lines $l_1 = OAA'$, $l_2 = OBB'$, $l_3 = OCC'$, and $l_4 = ODD'$ all lie on the same plane. The cross ratio is $\frac{AC}{CB} \div \frac{AD}{DB}$ and it is constant, i.e., $\frac{AC}{CB} \div \frac{AD}{DB} = \frac{A'C'}{C'B'} \div \frac{A'D'}{D'B'}$. Its value does not depend on the image plane, but only on the lines l_1, l_2, l_3, l_4.[1]

Mel: So, this cross ratio is really the cross ratio of a pencil of four lines lying on the same plane and going through the same point O. You can cut these lines with any line l and the cross ratio you find will be the same, no matter what l is. It only depends on the angles the four concurrent lines make.

Arc: Exactly, as long as l does not pass through O.

Mel: I guess you can also talk about the cross ratio of a pencil of planes $\Pi_1, \Pi_2, \Pi_3, \Pi_4$ as the cross ratio of the pencil of their four lines l_1, l_2, l_3, l_4 of intersection with any plane Π. And again, this cross ratio will be independent of the choice of Π.

Arc: I am glad you understand the concept. Now it is easy to grasp the relationship between the two images. The images formed by cutting the rays with different planes can be mapped to each other with a very easy map. It is a special map that depends only on how you map four points not lying on a straight line.

Mel: You mean if I know that points A, B, C and D in the first image map to points A', B', C' and D' in the second, then I know where every other point of the first image maps to in the second image? That is, this map that sends the points of the first image to points of the second image depends only on how you map four points?

Arc: Yes. This kind of map is called *homography*. Let's call our map H. So, if x (a point in the first image) is mapped to x' (a point in the second image), we write $Hx = x'$. If you apply two homographies H_1, H_2, one after the other you would have the composition of two maps $H_2H_1x = H_2x' = x''$. H^{-1} is the inverse of H, that is, if $Hx = x'$, then $H^{-1}x' = x$. Obviously HH^{-1} or $H^{-1}H$ maps a point to itself; it is the identity map $HH^{-1}x' = x'$ or $H^{-1}Hx = x$. It's easy to see that four points, with three not collinear, define this map. If $HA = A'$, $HB = B'$, $HC = C'$, and $HD = D'$, then the map is defined for all other points. Consider the intersection of AB and CD. Call this point E. Call

[1] This has been known for more than 2000 years. For a proof see any high school geometry book or [REF].

E' the intersection of $A'B'$ and $C'D'$. Since H sends the lines AB and CD to $A'B'$ and $C'D'$, it must send E to E'. So, $HE = E'$. Now, take any point F on AB (different from A, B, and E). The cross ratio of the four points A, B, E, F has some value r. But since the cross ratio is preserved, if you pick a point F' on $A'B'$ so that the cross ratio of A', B', E', F' is also r, it follows that H maps F to F'. You can do this for all points of the lines AB, BC, and CA. Now pick a point G not lying on AB, BC or CA, and draw line from G that cuts AB, AC and BC at points P_1, P_2 and P_3. But then, according to what we said before H should map P_1, P_2, P_3 to three collinear points P_1', P_2', P_3'. Since the cross ratio is preserved, point G should be mapped to G' so that the cross ratio of P_1', P_2', P_3', G' is the same as the one for P_1, P_2, P_3, G.[2]

So, there you have it. If you know how four points map from one image to the other, you know how to map any point. If you give me the first image and only four points on the second along with how they map to four points of the first, then I can make the second image!

Mel: This seems very powerful. Let me summarize it. If you keep the camera center fixed and you change the image plane, then the images you get are all related to each other by a homography. That's a simple mapping and it's completely determined if you know how four points map from one image to the other. Collinear points map to collinear points. I must admit, this light geometry appears very interesting. But you have mostly been talking about points and lines. How about parallel lines? And what happens to curves? For example, as I am moving the image plane, if I get parallel lines in one image, how do they appear in the other images for different locations of the image plane? And if I get a circle in one image, how does that circle appear in the other images? In other words, which figure will the homography map the circle to?

Arc: Let's look at parallel lines first. You must have seen parallel lines in the world, like train rails. How do they appear on your eye.

Mel: They appear to converge at a point.

Arc: Exactly. When you observe a line and the line moves away to infinity, in your image it stops somewhere. That point is called the vanishing point of the line, it's the image of the point at infinity. You can easily find that point by bringing from the camera center a parallel to the line. The point where this line cuts the image is the vanishing point. Obviously parallel lines have the same vanishing point on the image. In the same way, you have the vanishing line of a plane, that's what people call the horizon. Now let's talk about curves, what happens when you map circles with homographies. There is an easy way to find this. The camera center and the circle define a cone. To see how the circle will appear in another image as you move the image plane, you just have to cut the cone with that image plane. You get the so-called conic sections: circles, ellipses, parabolas or hyperbolas. So, a homography sends rectilinear figures

[2]foot here:

into rectilinear figures, and conic sections into conic sections. Images can differ greatly, figures of finite extent sometimes going into figures of infinite extent, parallel lines sometimes going into intersecting lines, and vice versa and conics of one type (parabolic, elliptic, hyperbolic) going into conics of another type.

1.3 Conics

Mel: Fascinating indeed. It seems that the conics have a special property. Although they look different, they are probably pretty similar.

Arc: Oh yes, they are! It's easy to see why. Take a circle and choose four points on it, A, B, C, and D. Now pick any other point on the circle, P, and form the pencil PA, PB, PC, PD. This pencil has a cross ratio that depends only on the angles the four lines are making with each other. So, if you take another point Q on the circle and you make a new pencil QA, QB, QC, QD it will have the same cross ratio since the angles formed in the new pencil are the same as the corresponding ones in the old pencil ($A\hat{P}B = A\hat{Q}B$, etc.). In different words, if you fix points A, B, C, D and point P moves on the circle, the pencil PA, PB, PC, PD has a fixed cross ratio. Now, let's say that you have the circle on your image plane, and you move the image plane to another location. We know that the circle will map to a conic—and by appropriately placing the plane you can get any conic. Furthermore, points A, B, C, D, P, Q will map to A', B', C', D', P', Q' on the conic. Since, however, the cross ratio is preserved, the cross ratio of pencil $P'A', P'B', P'C', P'D'$ is the same as the cross ratio of PA, PB, PC, PD. It follows that the cross ratio of $P'A', P'B', P'C', P'D'$ is the same as the one of $Q'A', Q'B', Q'C', Q'D'$. There you have it. If you fix four points on a conic then for any point P on the conic the resulting pencil with P as center has a constant cross ratio!

Mel: Fascinating! The reciprocal is true as well, right? If you pick four points on the plane and you consider all the pencils PA, PB, PC, PD for different points P so that all these pencils have the same cross ratio, then the points A, B, C, D and any of the points P lie on a conic.

Arc: Obviously. There is, furthermore, an interesting twist to this that is very useful. Take a conic and pick two points on it, O and O'. Now, consider two pencils centered at O and O', that is, connect O with every point on the conic (one pencil) and similarly O' with each point on the conic (second pencil). Of course one line of each pencil will be tangent to the conic at O and O', respectively. If you pick any four points A, B, C, D the two pencils OA, OB, OC, OD and $O'A', O'B', O'C', O'D'$ have the same cross ratio. Clearly, you can make a correspondence between the two pencils: for any point X of the conic, correspond OX to $O'X$. So OA corresponds to $O'A$, OB to $O'B$, etc. Finally, correspond the tangent at O to OO' and the tangent at O' to $O'O$. As you see, the two pencils are in a one-to-one correspondence and they have the property that any four lines in one pencil have the same cross ratio as the corresponding four lines in the other. Right?

Mel: That's clear. So what?

Arc: To see what this means, now erase the conic. You get two pencils that are in a one-to-one correspondence and any four lines of the first have the same cross ratio as the corresponding lines in the second.

Mel: Oh, I get it. So. if I have this situation, then the corresponding lines of the two pencils intersect at points that lie on a conic, right?

Arc: Right. Now let's begin our study of visual space.

1.4 Uncalibrated eyes

Mel: Before you go on, I am somewhat bothered by what you have called uncalibrated cameras. Do I know anything when I use such a camera? For example, to get the image of a point in the world I must connect it with the camera center. The resulting ray will cut the image plane at a point, which is the image of the world point. But since I do not know where the image plane is, I cannot make this construction. Similarly, if I have the image of a point, it seems I cannot connect the camera center with this point to get the ray. Is there anything I can do?

Arc: It's not so bad! Besides, the only reason we even look at this kind of eye is to gain intuition into the process. Indeed, for these eyes you don't know the rays. If you have an image, the situation could be very different because the camera center could be close or far from the image plane. But if you know six points in the world and their images, then you can do the operations you mentioned. Here is why. The trick is to do everything with regard to a number of reference points. Take, for example, six points A, B, C, D, E, F in the world and consider their images $I_A, I_B, I_C, I_D, I_E, I_F$. If you know these points you can find the camera center O and back-project an image point. You can also find the image of any other point.

Take the pencil of planes $\Pi_B, \Pi_C, \Pi_D, \Pi_E, \Pi_F$ that contain OA and B, C, D, E, F, respectively. As you cut these planes with the image plane you find the intersection lines $I_AI_B, I_AI_C, I_AI_D, I_AI_E$ and I_AI_F. Now take a known plane Π and put it in the scene. Ray OA cuts this plane at $I_{A'}$. Lines AB, AC, AD, AE, AF intersect plane Π at $I_{B'}, I_{C'}, I_{D'}, I_{E'}, I_{F'}$, respectively. We can clearly measure the cross ratio of any pencil of planes made with four of the planes $\Pi_B, \ldots, \Pi_F$. This can be done by measuring the cross ratio using the intersections $I_AI_B, I_AI_C, \ldots, I_AI_F$ with the image plane. As the cross ratio does not change, this means that we know the cross ratio of the line pencil $I_{A'}I_{B'}, I_{A'}I_{C'}, I_{A'}I_{D'}, I_{A'}I_{E'}$. This, however, implies that $I_{A'}$ lies on a conic passing from $I_{B'}, I_{C'}, I_{D'}, I_{E'}$ on plane Π. Similarly, we know the cross ratio $I_{A'}I_{B'}, I_{A'}I_{C'}, I_{A'}I_{D'}, I_{A'}I_{F'}$, which means that $I_{A'}$ lies on a conic passing from $I_{B'}, I_{C'}, I_{D'}, I_{E'}, I_{F'}$. So, $I_{A'}$ lies on two conics which have points $I_{B'}, I_{C'}, I_{D'}$

in common. $I_{A'}$ is the fourth common point. This means that we can find $I_{A'}$ and since we know A, we know $AI_{A'}$, which passes from the camera center O. In a similar way we can find the other lines and their intersection will give us O! It now becomes quite easy to back-project an image point or to find the image of a point in space. Consider a point X in the world and its image I_x. Given I_x we want to find the ray OX, or, given X we want to find I_x. If you are given I_x, then you can easily determine the cross ratio $I_AI_B, I_AI_C, I_AI_D, I_AI_X$. This, however, determines the plane OXA in the scene. If you now take another reference point, B, instead of A, you can determine the plane OXB. These planes intersect at OX, the ray defined by the center O and the image point I_x. Inversely, given X, we can determine I_x. Since O, A, B, C, D, X are known points, we know the planes OAB, OAC, OAD and OAX, that is, we know their cross ratio. Intersecting this pencil of planes with the image plane, we obtain that this cross ratio is the cross ratio of the lines $I_AI_B, I_AI_C, I_AI_D, L_1$, where L_1 is the intersection of plane OAX with the image. But since the cross ratio of this line pencil is known, this means that we know L_1, and I_x lies on L_1. Similarly, we can get another line L_2 and I_x is at the intersection of L_1 and L_2.

Mel: Very nice indeed. You did everything with the reference points you chose. If you know those six points in the world and their images, you can do everything by relating any new points to the old points. And you are doing everything without knowing where the image plane is. But I still don't understand the real essence of this. What is the jump from uncalibrated eyes to calibrated eyes? What does it take to go from one kind of eye to the other?

Arc: To go from one kind of eye to the other, you need just one homography! Given the center, out of all possible locations of the image plane there is one that corresponds to your specific camera, and as we said you are just one homography away from it. Let me explain this in detail. Let's start with a calibrated eye. Take a point O, your camera center, and a plane, your retina. Consider a line from the center perpendicular to the image plane. It intersects the plane at P, the so-called principal point; the line is the principal axis or ray. The distance OP is the focal length. Now you need to measure things on the image. You do that by placing the traditional coordinate system with the x and y axes (horizontal and vertical, respectively) and with origin at P. Any image point is characterized by two numbers, the x- and y-coordinates. You can easily form images by joining O with a point A in the scene. Intersection with the image plane provides the image I_A of A. If the involved parameters are unit valued, that's a normalized camera.

Mel: OK, that's a calibrated camera. You know the focal length and you know how to measure locations on the image, right? But you also need to measure locations in space. How do you do that?

Arc: You need another coordinate system. Put its origin at O, have the x and y axes parallel to the x, y axes of the system on the image and the z axis on the principal ray. Now you are all set. You can measure things in space, in this camera coordinate system.

Mel: So, if I don't know the focal length and/or if I cannot measure things on the retina in this nice manner, I get an uncalibrated eye?

Arc: Exactly. Let's say you don't know the focal length. That means that you will have to move the retina to another, unknown location, but when you move the retina, keeping the projection center fixed, how are the images before and after the movement related?

Mel: They are related by a homography! Actually, it's a simple one because each image point is going to move to a new position along the straight line connecting the point to the principal point. Isn't this called zooming?

Arc: Yes. Now let's look at the retina. Retinas can be weird things, because of the way they are constructed to contain photocells. The origin of coordinates need not be the principal point. On top of this, because of sampling rates and photocell distribution, the units of measurement need not be the same along the two axes, i.e., the pixels may not be square. Finally, the two axes need not be perpendicular to each other.

Mel: I get it. To see how the final image will look you must map the "normalized" image to the actual image, the black squares to the red parallelograms. And that's a homography!

Arc: Exactly. To be more precise, this map is even easier than a general homography, since you only need three (non-collinear) points to completely define it. That is because the total number of calibration parameters is five, two for the image center, two for the units of measurement along each axis, and one for the angle that the axes make with each other.

Mel: That answers my questions. You can get to the image captured by any planar eye starting with a normalized eye that has the same center. After you have the normalized image you can map it with a homography and get the image of any camera with the same center. The catch is that if your camera is uncalibrated then you won't know this homography.

Arc: I couldn't have said it better. Start with the normalized situation. Take your coordinate system, put the camera center at the origin and put the image plane perpendicular to the z axis at distance 1 from O. You can clearly see that there is a simple map that brings any point X in space to its image x_n. Call this map P_n and you can say $x_n = P_n X$. You get this map by connecting X with O. Now you have to account for the calibration parameters, so the final image point corresponding to X will be $x = Hx_n = HP_n X$. It is the homography containing the calibration information. You can then say that the image x of any point X in space is found by applying a map P to X, that is, $x = PX$, with $P = HP_n$.

Mel: But P is actually the composition of two maps. You first apply P_n to X, then you apply H to what you find, right?

Arc: Obviously. The thing to remember is that you have a camera or an eye when you know the map P. For a calibrated eye you know exactly P and from this you can get everything else, the camera center and the calibration parameters. For an uncalibrated eye, you know a P.

Mel: Hmm, you mean that if I have an uncalibrated image, I could have many different maps P making it?

Arc: You can actually have infinite maps. It's easy to see why. Recall that a homography is mapping images to images. It's a mapping that preserves collinearity, and it is fully defined if you know how four points are mapped (three of them not collinear). You can also have a homography in space, but now you need eight points to define it. So, for every point X in space, you can apply a homography H and move the point to a new position $X' = HX$. Now, map this new point not with map P but with map PH^{-1}. H^{-1} is simply the inverse of H. What do you get?

Mel: Well, $PH^{-1}X' = PH^{-1}HX = PX = x$. I get the same image as before!

Arc: Exactly. So, if you have an uncalibrated image, it could have been the result of different projections of different worlds! And all these worlds are related to each other by a homography.

Mel: I get it. It is as if you can move the eye center anywhere. This is all so nice and clean. I wonder, is the geometry of visual space that clean?

1.5 The essence of making 3D models

Arc: Are you anxious to know? Let's start! Take two eyes and put them at two locations looking at the world. Consider them calibrated. Our goal is to figure out how the two images are related and then develop a model of the scene.

Mel: Why are you placing the eyes far apart? If I move around the world, then my eyes take successive positions that are very close.

Arc: That is true, but to take the eyes in a general position is better because it represents a more general situation. Obviously, it includes the case where the eyes are very close to each other. So, it makes sense to study the general case first. But there is an additional reason why this is a good thing to do. Remember, our ultimate goal is to make a model of the scene. As you can understand, that will happen *first through finding where the eyes are.* Then, by extending the rays from each camera center to image points representing the images of the same feature in space, *the point in space* where these rays intersect gives us the space location of scene points and hence a *model.* Now, if we make an error in our calculations, the rays won't intersect and so we'll have an uncertainty in the location of the space point. The closer the cameras are to each other, the larger that uncertainty. So, it makes sense to put the cameras far apart.

Mel: What you say makes sense, but there is a bit of magic in it. You are assuming that in the two images you can find the points that are the images of the same point in space. Then, using this you will find where the cameras are by relating the two images somehow. After that the 3D model pops out like magic. But how are you going to find the corresponding points?

Arc: I would suggest that we take things in order. Your question is a good one but its time has not yet come. We will address it later. Don't you want a nice, clean geometry for visual space? That's what I am about to give you. Technicalities with images can be dealt with when the time comes.

Mel: But visual space is based on images. I thought it would be a good idea to stick to images. But, please, go on.

Arc: Take the two eyes looking at the world. These can be either one eye that moves between two locations taking pictures, or two different eyes. It doesn't matter. Let's say they look at a point X. The rays OX and $O'X$ cut the two image planes at x and x', the images of X. We call OO' the *baseline* and the plane defined by OO', OX and $O'X$ an *epipolar* plane. The images x and x' lie on this plane. OO' cuts the image planes at points e and e', which we call *epipoles*. As you see, the epipole is the image in one view of the camera center in the other view. One has to be careful in interpreting the epipoles for calibrated or uncalibrated cameras, but let's worry about that when the time comes. The epipolar plane will cut the image planes along two lines l and l' that contain x and x'. We call these points corresponding points, and the process of finding corresponding points in the two images we call the visual correspondence process. Now, if you only know x, what can you say about its corresponding point x'?

Mel: The baseline and the ray OX determine the epipolar plane that intersects the second image along the epipolar line l', which contains x'. So, if I know x, then I know that its corresponding x' lies on the epipolar line, which, by the way, is the image of the line OX.

Arc: Very good. Clearly there is a map emerging from this epipolar geometry. This map sends points in one view to the corresponding epipolar lines in the other view, i.e., $x \to l'$. Call this map E, and you have $Ex = l'$. This map relates the two images. If you understand this map then you understand how the two images are related and you have a chance of finding how the cameras are related, that is placed.

Mel: Hmm, before you go on with these maps, I need to understand how the two cameras are related to each other. I start with a camera at the first position. Then I move it to the second position, right? What is involved in that movement?

Arc: That is what people call rigid movement. Pick up that book and stand it on the table. That's a rigid movement. The object changed position but the object itself did not change. All distances between different object points remained the same. Mathematicians have figured out what is involved in a rigid movement, and it's pretty simple. Every rigid movement is the sum of a rotation plus a translation.

Mel: This rotation and translation, are they really what their intuitive meaning tells me they are?

Arc: Yes, purely translational movement amounts to a movement where all the object points appear to slide along parallel lines, all of them by the same amount. Purely rotational movement is slightly more involved. To define it you need an axis that we call the rotation axis. Now, an object is rotating around an axis when every object point is moving along a circle, with the rotation axis perpendicular to each circle and passing from its center. As you can see, the closer the object point is to the axis, the smaller the circle becomes; if the axis passes through the object, the object points that are shared with the axis are not moving as part of the rotational movement. So, for pure rotational movement, you need an axis around which the rotation will take place, and you need an angle that signifies the amount of rotation, how much each object point moved along its corresponding circle.

Mel: OK, let's get back to our cameras. You say that the movement from the first position to the second is rigid, and so it is the sum of a rotation and a translation. But which translation and rotation? It seems to me that I can do this in myriad ways, depending on where I put the rotation axis.

Arc: You are very observant indeed. That's correct, but if you pick one point from which the rotation passes, then everything becomes fixed. Let's take the rotation axis passing from the camera center. Rotate the first camera so that its axes, i.e., its coordinate system axes become parallel to the axes of the second camera; then translate it so that it coincides with the second. That's the movement. Now we can check how the images in the two locations are related. But let's do it in stages. First let's see what happens if the eye only rotates; then let's examine the situation when the eye only translates. After that we can put the two things together. As we do this, sometimes matters will be simplified if we move the scene instead of the camera. What I mean is, if you rotate the camera around an axis passing through its center by some amount, the images you will get will be the same as if you rotated the scene around the same axis by the same angle in the opposite direction. Similarly, if you translate the camera the new image would be the same if you instead translated the scene with exactly the opposite translation.

Mel: OK then, let's start with translation only, and take an object translating in front of the camera. As the object translates all its points move along lines parallel to the translation vector. Now, how is the camera going to observe these parallel lines? They will all converge at a point, as you said, the vanishing point. All the image points will move along straight lines converging at that point, which obviously represents the translation, right?

Arc: Yes, you can find this point by bringing the line from the camera center parallel to the translation vector. Where this line cuts the image is where the vanishing point is. So it represents the direction of the translation, not the exact translation. That's the epipole. So, a point x before the motion will move to a point x' after the motion and points x, x' and e all like on a straight line. If the object goes away from the camera, then x' is closer to e than x. The opposite is true when the object comes closer.

Mel: I see. If I move the camera instead of the object I get the same situation. The epipoles in the two cameras are the same and the epipolar lines are all radiating from e and e'.

Arc: Obviously. The epipoles are the same, i.e., when you superimpose the two images they fall on the same point. Also notice that when an object point is close to the camera its image moves faster than the image of a point which is further away, and if an object is infinitely far away, its image doesn't move at all under translational movement! Being at infinity, its image will be at the vanishing point and will stay there even if it moves. Now let's look at the map from points in the first image to the corresponding epipolar lines in the second image.

Mel: That's pretty easy. You just have to connect the epipole with the point. If I superimpose the two images, points x, x' and $e = e'$ are collinear.

Arc: Good. Since we have calibrated cameras you can represent the direction of translation $\mathbf{t}$ with the image point where $\mathbf{t}$ cuts the image. Let's agree on some symbolism. If I have a point A on the image and I want the line connecting A with a point x, let's write this as $[A]_\times x$. That is, $[A]_\times$ is a map that if we apply to a point, it will map it to the line from A to the point. So, if $\mathbf{t}$ represents the translation direction, then the epipolar line in the second image corresponding to x is $[\mathbf{t}]_\times x$. Now we are ready to study the purely rotational case, which is even easier. Do you see why?

Mel: Yes. When a camera rotates around an axis passing through its center the camera center remains fixed. Only the camera plane changes. We discussed this already! The mapping between the two images before and after the motion is a homography!

Arc: Great! Let's call this homography R. Then a point x in the first image will map to a point x' in the second and you have $x' = Rx$. This is, of course, a special kind of homography because you don't need four points to define it. You need only two. Do you have an understanding of this mapping?

Mel: Hmm, let me do the trick you suggested. Instead of rotating the camera, let me rotate the scene around the rotation axis. Every point is then moving along a circle perpendicular to the axis at the circle's center. So, to see how the image points will move I just need to see how these circles will appear on the image, but we know that as well. They will be conics!

Arc: Exactly. The rotation axis defines a family of cones with apex at the camera center. These cones will cut the image plane at conic sections. Under pure rotation the image points move along these conics, except the point where the rotation axis cuts the image. That point remains still.

Mel: So, that means the image change under rotational movement does not depend on the scene at all, is that right?

Arc: Of course it's right. It depends only on the rotation itself.

Mel: But this means that if we only have rotation (around an axis passing through the camera center), then we cannot say anything about the scene. I cannot build models of the world.

Arc: Right again! Only the translation contains information about the scene. Can you put the two together now, and find the map from points in one view to the corresponding epipolar lines in the other view?

Mel: Well, let's say I have two eyes looking at the world. I can go from the first to the second with a rotation followed by a translation. First, I will rotate the first eye so that the image planes become parallel to each other. Then I will translate the eye until it coincides with the second. Let's take a point x in the first image. The rotation will map it to a new point $x' = Rx$ where R is the rotation homography. Then the translation will map x' to the epipolar line $[\mathbf{t}]_\times x'$, that is, $[\mathbf{t}]_\times Rx$. So, a point x in the first image maps to the epipolar line $l' = Ex$, where the map $E = [\mathbf{t}]_\times R$.

Arc: Very well done! That's the mapping between any two views of the world.

Mel: Thanks. Now I can imagine how you can make a model if you are given two images taken by two eyes in unknown positions, and if you have corresponding points in the two views. I guess you will find the map E and from that, somehow, you can get the rotation and the direction of translation. You start from the first camera, apply the rotation, then slide it along the translation axis. You will not know where to put it exactly because you will not be able to find the exact translation, but it doesn't matter. Wherever you put it along the translation axis, you can reconstruct the points of the object or the scene, since you must respect the angles between the rays captured on the retina. You will get the correct object in space. It may be large and far away or small and close by. You can reconstruct the world up to a scale! But let me come back to the mapping from points in one image to the corresponding epipolar lines in the second. I understand it in a local sense, but not in a global sense. If you give me a point x, the map will give me its epipolar line in the other view, but how does the whole thing look globally?

1.6 A global view

Arc: You hit the right question. Let's look at all the epipolar lines in both images. As you can see, there is a pencil of epipolar lines in each image centered at the epipole. Now pick a point C on the baseline and it follows that if you consider this point as the center of projection, the two pencils of epipolar lines are related by a homography. Actually, you can put the lines of each pencil in a one-to-one correspondence.

Mel: Hold on, there is more to this. Since the two pencils lie on the rays passing through C and cut by the two image planes, the cross ratio is preserved. If I pick any four lines of the first pencil, their cross ratio is equal to the one of the corresponding lines of the second pencil.

Arc: Perfect. Now, if you have two pencils that are in a one-to-one correspondence and the cross ratio of any four lines in the one is equal to the cross ratio of the corresponding lines in the other, then what does that mean?

Mel: Oh, you talked about it. It means that if I put both pencils on the same plane (i.e., superimpose the two image planes), then corresponding lines in the two pencils will intersect at points lying on a conic!

Arc: This is it! That's the global property you were asking about. The two epipoles lie obviously on the conic. e' is the translation direction $\mathbf{t}$. As for e, since we first rotate and then translate, it is also $\mathbf{t}$, but before the rotation is added. So, $e = R^{-1}\mathbf{t}$, with R^{-1} the inverse mapping of the rotation homography.

Mel: This looks cute, but what does this conic really mean? How is it related to the mapping between the two images?

Arc: This conic is the locus of the intersections of corresponding epipolar lines. That means that a point on this conic, if you consider it as a point in the first image, has a corresponding point in the second image, which is the same! In other words, if despite the movement of the camera there are points that don't move, these points must lie on this conic!

Mel: Hmm, you mean that points which lie on this conic will not move between the two images?

Arc: Not exactly. It simply means that if there are points that do not move, those points will lie on this conic. You see, how points move under the rigid transformation is dependent on where they are in space. If you give me two cameras separated by a rigid motion, there are points in space that will project onto the same point in the two images. These points lie on a curve that is called a horopter. The images x, x' of any point on the horopter are such that $x = x'$. Now, the image of the horopter is our conic. The horopter does not depend on the scene in view; it only depends on the rigid motion between the cameras. If the horopter intersects the scene in view at a number of points, then it is these points that will not move from the first image to the second.

Mel: So, I may have the situation where there are no points with zero movement because the horopter does not cut any of the surfaces in view. Or I may have many points which have zero movement and those points will all lie on this conic.

Arc: Exactly. Let's call this conic the zero motion contour. It contains the epipoles and it also contains the point where the rotation axis cuts the image. If you call this point Ω, then obviously $R\Omega = \Omega$, because this point won't move under rotation. What is its corresponding epipolar line?

Mel: Well, let's apply the map E. It is $E\Omega = [t]_\times R\Omega = [t]_\times \Omega$.

Arc: But $[t]_\times \Omega$ is the line connecting t to Ω, and so Ω lies on its corresponding epipolar line, and so lies on the zero motion contour. Now that you have this situation, you can easily see the map E, mapping a point to its corresponding epipolar line. Take any point x. Connect e, the epipole in the first view, with x; this straight line cuts the zero motion contour at point x_H. Now connect x_H with e' and you get l', the epipolar line of x. Do you see why?

Mel: I do. Connecting e with x, you get an epipolar line in the first view. This defines an epipolar plane with center O which intersects the horopter at a point, say X_H. The image of X_H in the first view is obviously x_H, but because of the zero motion contour property, the image of X_H in the second view is also x_H. So l' is the epipolar line in the second view corresponding to x. And that's the mapping!

Arc: Are you now satisfied regarding the global structure? This conic has on it everything you need to know.

Mel: I appreciate the insight with the zero motion contour but I still have no sense of how the whole image moves. The epipolar geometry tells me how a point in one view is related to the other view. I would like to see how this geometry relates everything at the same time. But anyway, how do you find the rotation and the translation? How do you place the cameras?

1.7 Finding the epipoles

Arc: We'll get to it. But first let me show you how you find the epipoles, even for uncalibrated cameras.

Mel: The epipoles are like the translation, right?

Arc: Not exactly. The epipole e' in the second image is the translation $\mathbf{t}$ but transformed by the calibration homography H_2, so e' is really $H_2\mathbf{t}$. e is also the translation but before the rotation. So, you have to map $\mathbf{t}$ first by the inverse rotation R^{-1} and then with the corresponding calibration homography in order to get e, i.e., $H_1R^{-1}\mathbf{t}$. Now, recall what a normalized camera is.

Mel: I remember that. A normalized camera is a calibrated camera with focal length one and the origin of the image coordinate system at the principal point; x and y are perpendicular and have the same units. I can now go from the image of a normalized camera to the image of any new camera with the same center by warping the image with a homography which depends on the calibration parameters of the new camera. If x is the normalized image of a point, its image in the new camera will be $\hat{x} = Hx$, where H is the homography map. I can go from the image of any camera to the normalized one by applying the inverse homography, $x = H^{-1}\hat{x}$.

Arc: Very good. Now take two uncalibrated cameras looking at the world. Let's say that H_1 and H_2 are the homographies that bring a normalized image to the image of camera 1 and camera 2, respectively. You know how to get the map between the two images if they are calibrated. Start with the first camera. Transform the image to a normalized one. This means that a point x in the first image will move to $H_1^{-1}x$. Now apply a rotation. This means that a new homography R will be applied to each point and our point $H_1^{-1}x$ will move to $RH_1^{-1}x$. Now bring the normalized image $RH_1^{-1}x$ to a new image as if the camera was the same as the second camera. You simply must apply homography H_2, and the point becomes $H_2RH_1^{-1}x$. What did we do?

Mel: We applied three homographies, H_1^{-1}, R and H_2, one after the other to the first image. We basically rotated the first camera and corrected the image so that we now have the same two cameras separated only by translation.

Arc: That's it. Call $H_2RH_1^{-1} = H$, since applying homographies one after the other amounts to applying a new homography. So, x moves first to Hx and now I have pure translation, which means that the epipolar line l' corresponding to x will be found by connecting e' to Hx, i.e., $l' = [e']_\times Hx$. So, the new map is $F = [e']_\times H$ and $l' = Fx$.

Mel: That's pretty simple. Basically now the rotation gets mixed up with the unknown calibration parameters and instead of the translation vector you have the epipole. In the calibrated case we had the map $E = [t]_\times R$ and the hope was that, if we could find this map, we would eventually find t and R, and thus place the cameras so that we can get a model for the scene. Now the map is $F = [e']_\times H$. Could we eventually find e' and H?

Arc: Not really. You see, there are infinite, different homographies that can give rise to the same map F.

Mel: Why?

Arc: Let me show you how you get them. Take your two uncalibrated cameras looking at the world. Point X in space gives rise to images x and x'. Now put a plane P in space. The ray OX cuts plane P at X_P. Notice that the image of X_P on the second camera is a point x'_P that obviously lies on the epipolar line of x. So, if I know x_P, I know the epipolar line of x, by connecting e' with x_P.

Mel: But x_P is not the corresponding point of x.

Arc: It is not, but it lies on x's corresponding epipolar line, and that's what matters. Now, notice something interesting. Consider the map between the first image and plane P; the rays passing through O cut both planes, so you get a map sending points x to points X_P.

Mel: Oh, OK. That map is a homography.

Arc: Good. Call this homography H. Now look at the similar map between the second image and plane P, the map that sends points x' to points X_P.

Mel: A homography again.

Arc: Let's call it H'. How about the map from plane P to the second image?

Mel: Obviously a homography, the inverse of H', and we usually write it ${H'}^{-1}$.

Arc: Now let's follow the map from x to x_P via plane P. From x we go to X_P with homography H and from X_P to x_P with homography ${H'}^{-1}$. So, ...

Mel: So, from x we go to x_P with homography $H{H'}^{-1}$.

Arc: Great. Call this homography H_P and you have $x_P = H_P x$. But then the epipolar line l' corresponding to x is $l' = [e']_\times H_P x$ and our map is $F = [e']_\times H_P$.

Mel: Beautiful! So, if you change plane P, you get a different homography H_P but map F remains the same!

Arc: Exactly. You see, the homography H_P contains information about rotation, the two cameras' calibration parameters, and translation and location in space too.

Mel: But what about the homography you talked about before, the map $H = H_2 R H_1^{-1}$? You showed that the map $F = [e']_\times H$. But H depends on H_1, H_2 and R. $H_{1,2}$ have to do with the cameras' calibration parameters, R with the rotation. H_1, H_2 and R do not depend on translation or the location of scene points in space. What's wrong?

Arc: Oh! H is special. H is the same as the H_P's we discussed before, but this time you put the plane P at infinity! That's why the name of this homography is the infinite homography. We write it $H_\infty = H_2 R H_1^{-1}$. Since plane P is at infinity, H_∞ doesn't depend on the translation or space location of points.

Mel: I see. For example, if the eyes look at an object with parallel lines, these lines in each image converge at vanishing points, right? These vanishing points are images of points at infinity. So H_∞ maps vanishing points in the other image since translation doesn't influence those points' locations.

Arc: Exactly. Moreover, any of these homographies H_P, including H_∞, map

e to e', since in general the line ee' cuts that plane somewhere. So, if you have three vanishing points and the epipoles in the two images, H_∞ is defined. Or if you know four corresponding points in the two images that lie on plane P, any of these H_P's is also defined.

Mel: Well, what can you do now if you find the map F for your uncalibrated eyes? Can you find a model of the scene?

Arc: You can, but it is not that useful. Remember that when you have an uncalibrated camera and you get an image, it could be the image of different worlds projected on different cameras.

Mel: Yes, for any camera, a world point X projects to its image x with a map P, so that $PX = x$. If I move the world with a homography H, each point X moves to position $X' = HX$. If I now project this point with the map $P' = PH^{-1}$, that is, if I make its image with a new camera, I get $P'X' = x$, the same image.

Arc: Very well said. Now, take your two uncalibrated eyes and look at the world. Both images could obviously be due to the projection of different worlds to different eyes. All we have is corresponding points x and x' between the two images and from these we can find the map F that relates the two views. If the two eyes are characterized by a map P and a map P' looking at the world X, the situation doesn't change if you instead have camera $PH^{-1}, P'H^{-1}$ looking at a new world HX. In the calibrated case you could find X up to some scale. That is, you know the object in your field of view, you just don't know its size. Now, the best you can hope for is to find HX, for some H that is unknown.

Mel: Hmm, is that of any use? I can see that X and HX have something in common, but they could look very different.

Arc: That's correct. There are still a few questions you can answer with this model; it's a pretty poor model, but it contains information about space. From this model, for example, you know which points in space are in a straight line, and you know which points are between any pairs. That's something.

Mel: Well, it is indeed something. It is as if you take the correct model of what is in your view and stretch it in ways so that you preserve collinearity. It will look quite distorted but we have some knowledge of how it is distorted. How do you find it anyway? In the calibrated case it was easy. If I had map E, I could somehow find the rotation and translation, and I could place the cameras. That way I could find the points in the scene up to a scale. Now I cannot find the rotation. How would I do it?

Arc: It's really easy. When you know the map F, you basically have a correspondence between the epipolar lines in the two images. Because the cross ratio is preserved, if I take four epipolar lines in the first image they have the same cross ratio with their four corresponding epipolar lines in the other image. So, if you know the correspondence between three epipolar lines in the two images then you know the epipolar geometry (because then you know the correspondence of any other epipolar line). Remember, if you know how three epipolar lines correspond to each other in the two images, and I give you a fourth one in the first image and I ask, what is the corresponding epipolar line in the other image? You know how to find it.

Mel: Yes, because the fourth one and its corresponding line make the same cross ratios with the other three, in the two images.

Arc: Good. So, our problem becomes easy: you have two triples of lines (l_1, l_2, l_3) in the first and (l'_1, l'_2, l'_3) in the second image, and you know that l_i corresponds to l'_i. That is, there are three planes that cut the two image planes along (l_1, l_2, l_3) and (l'_1, l'_2, l'_3). Let's say the same thing in different words. Assume that the angle between l_1, l_2 is ϕ and the one between l_2, l_3 is θ. Similarly, ϕ' and θ' for the other lines. We then have to find three planes that give angles ϕ, θ and ϕ', θ' at the intersections when we cut them with the image planes.

Mel: Is that possible?

Arc: Yes, you just need to understand what happens when you cut two intersecting planes (an open book) with another plane. Depending on how that plane is oriented with respect to the other two, the angle you get at the intersection will be different. There are many different planes that will give rise to the same angle.

Mel: I don't see this clearly.

Arc: Here is an example. Take two planes, Π_1, Π_2 that are perpendicular to each other, they make a 90-degree angle. Let's find the planes that also make a 90-degree angle when they cut our two planes. A good way to think of planes is to think of the vector perpendicular to them, what we call normals. Now, take a sphere and place yourself at its center. Take the normal of a plane and bring it to the center of the sphere (don't rotate it as you move it). At its new position, the normal cuts the sphere at a point. For example, normals n_1, n_2 represent our two planes Π_1, Π_2. If you bring a plane Π perpendicular to both Π_1, Π_2 (and thus to their intersection ϵ), it will make a 90-degree angle between the lines ϵ_1, ϵ_2. But now you can start changing that plane by specifying the lines ϵ_1 and ϵ_2 on the two original planes—remember ϵ_1, ϵ_2 are the intersections with the two original planes. You can move ϵ_1 towards ϵ and ϵ_2 toward ϵ on their corresponding planes and still get a 90-degree angle. As you move ϵ_1 it is as if you spin the plane containing n_2 and as you move ϵ_2 it is as you spin the plane containing n_1. These planes cut the sphere along great circles and you want those points at which the great circles passing from n_1 and n_2 are perpendicular to each other. These are obviously curves passing from n_1 and n_2. This is the case, no matter what the angle is. They change according to how the original planes with normals n_1 and n_2 change and according to how the desired angle θ changes. Now it's easy to solve our original problem. You have three epipolar lines on an image plane and their corresponding ones in the other plane. You want to place those planes in space and the respective camera centers so that the geometry between the two images is satisfied.

Mel: Yes, but we agreed that I have no chance of finding one solution. I can find a solution.

Arc: Yes, of course. What you will find is a model of the scene distorted by a homography. Take then the first camera and put it anywhere. Then, you immediately have three planes defined by the camera's center and each of the epipolar lines. Now you must cut this pencil of planes with a plane so that you make angles θ' and ϕ'. You know there are many planes that will make angle θ' when they cut Π_1, Π_2. All their normals lie on a curve passing from n_1 and n_2. Similarly, there is another curve representing all the planes that will make angle ϕ' with planes Π_2, Π_3. Where these curves intersect each other, that's the plane you want. After you choose the first image plane, since you know the geometry of the two views, you know the epipole and so you know the straight line on which the second camera center lies. Take it anywhere, and you are done. You chose the two eyes and you are consistent with the epipolar geometry. But the model you will make by extending the rays will be one homography away from the real one.

Mel: So, that's it? Is that all with the geometry of multiple views?

1.8 Three views

Arc: Basically yes. A few more things are possible. You can, for example, take three views now and study the problem again. Something interesting happens here. Lets say you have three corresponding points x, x', x'' in three views and you know the geometry between any two views. So there is a map E mapping x to its epipolar line l in the second view $x \to Ex = l$. Similarly, $x'' \to E'x'' = l'$ maps a point in the third image to its corresponding epipolar line l in the second. So, x'' lies on both epipolar lines, and it is known. So there is a map T taking two corresponding points x, x' and produces the third point $x'' = T(x_1, x_2)$. This map of course depends on the motion between the different views.

Mel: I wonder why this happens.

Arc: From view one and view two, you can get a model of the scene. Let's say you have calibrated eyes. This model is then the correct one up to some scale, but you can do the same for frames two and three. So the two models you get must be the same!

Mel: I get it. That's what gives the additional knowledge.

Arc: And the same thing happens if you have corresponding lines l, l', l'' in three views, instead of points. If lines correspond, then from two views you cannot find any model. The planes defined by the camera centers and the image lines will always intersect, no matter where the lines are, but if you have a third view, then all three planes must intersect at a line in space. Similarly, there is a map mapping two lines to their corresponding one in the third view, $l'' = T(l, l')$. Again, this map depends on the motion between the different views.

Mel: Are these maps for points and lines in the three views the same?

Arc: Yes, you just need to be a bit careful, because in the first case you map two points to a third, and in the second case you map two lines to a third. Actually, you can mix them and study the most general case: you have three views of a point on a line. In each view you could have a point, or a line (on which the point lies), or both. These three things are related by some form of the map T. For example, take the triad of point, point and line. Clearly, the line in the third image defines a plane where the 3D point lies. But this means that the map between the two first images is homography, H, so $x = Hx'$. It is depending on the line l'' and this means $H(l'')x' = x$. This is map T in this case. As before, from this map you can eventually find the geometry between the views. After that, you can build a model for the scene.

Mel: And you can go on like this with four views, and five views, etc.? When are you going to stop?

Arc: I will stop now, because more views don't really help. So, this is it. You have two or three views of points and lines, and you know the correspondence among them. Then, these entities are related to each other via various maps. For example, for two views, the map E maps a point x to its corresponding epipolar line, and map E is $[t]_\times R$, with t the translation and R the rotation homography. If I know E then I can find the translation and rotation, that is, the rigid transformation. For three frames, there is a map, T, let's say, that maps two lines l', l'' to their corresponding one $l = T(l', l'')$. If I know T, then I know the rigid transformations between the views, as before. In this case, it is a bit more complicated because you have several transformations $(1 \to 2, 2 \to 3)$. If the cameras are not calibrated, then you only find the epipoles, and establish a correspondence between epipolar lines. In the former case you can build a correct model for the scene up to a scale. In the second case, you can build a model of the scene distorted by an unknown homography.

Mel: But it's clear that you can use many views, as many as possible.

Arc: Certainly. From every two or three views you can make a model of the scene. Then you should put them all in the same coordinate system and you will have a very accurate model.

Mel: I must say I already understood a lot, but there are still a few things boggling my mind. First, how to solve the problem of correspondence in the two views and, second, how to actually find the maps that relate the multiple views, and get the rotation and translation out of them. And something more: I am not so clear about models of the world. I understand what it means to make an exact model of the scene up to a scale, but it seems there are descriptions of space that contain less information, like the ones you get when you have an uncalibrated eye. It's not clear to me how they all fit together.

1.9 Finding the maps and placing cameras

Arc: Regarding you second question, we'll talk about it later in my lab. Let's concentrate on finding the map E. As you see, the map E depends on the translation and the rotation: The translation is a line and the rotation is another line (axis) and an angle. These are not many things. Now, you can imagine, that E cannot be anything. If you have two corresponding points x and x', then Ex is a line passing from x'. That constrains E. The more corresponding pairs you have, the more you constrain E. Actually you need at least seven points to get it. After that, you need to split the map E into a rotation and a translation and you are done.

Mel: You still haven't told me how to find the correspondence between the views. You need this for all your theories to make sense.

Arc: First of all, you can do cool things even if you get the correspondence by hand.

Mel: You mean, sit down and look at individual pixels in images and pick them by hand with their corresponding ones?

Arc: Yes, a manual task.

Mel: Wouldn't that need lots of people?

Arc: Yes, lots of them. Actually, this is happening today in the entertainment industry. It's called special effects.

Mel: You mean special effects in movies?

Arc: Yes, exactly. It is with the aid of the knowledge I described to you that such interesting effects are possible. Eventually it all amounts to different coordinate systems and their relationships. If you can find, for example, the camera path given a video acquired while the camera was moving, you are in business.

Mel: You mean, for every two locations of the camera along its path, to know the rigid transformation relating the respective viewpoints?

Arc: Yes, because then you know how to put things in the scene that were not there before, or you can take things out of the video, although this is a much harder problem.

Mel: That's some form of altering reality, right?

Arc: Sure, since we value vision so much, we are used to believing pretty strongly in what an image tells us.

Mel: You still haven't told me about solving the correspondence problem.

Arc: And I will not tell you, because I don't know and frankly I don't care either.

Mel: How can you not care? Without it, how are we going to get to the visual space-time geometry and the models of the world that we can build?

Arc: You are interested in this problem, I am not. I am an engineer and I like to play with geometry. I looked at the problem and it is very hard, because when you change viewpoint many changes can happen to the images. We have tools to pick points in the images, but when we try to match them, we don't even know what is occluded in each view. The images around corresponding points quite often look very different, but despite that we have a few tools to help with the correspondence. A useful one is to try many combinations. Pick a correspondence, from features in the first frame to features in the second. Then find the map E. You need a few correspondences, for example, with eight is easy (with seven it's possible, but harder), but you can always use more. If for most groups of correspondences you pick, you find the same E, you are basically done. You just keep trying out different possibilities, many of them, until for most groups of correspondences you get the same map E. You can also use various heuristics that people used over the years. Statisticians and computer scientists have figured out ways to perform these techniques of trying out many things very fast, and so, that's nice. This doesn't always work. It is some form of art. You still have to do manual things.[3]

Mel: But isn't it easier to match the images if they are closer to each other?

Arc: Yes, it is, but still not easy. And don't forget that if you have a small baseline this will result in some uncertainty in the model that you will eventually recover. So, not much difference. But what you can do is start with video and track the features in the video across frames. People use this sort of technique today. This tracking will eventually give you correspondences in frames that are far apart.[4]

Mel: I see a difference. I am interested in visual space-time geometry and I have a strong desire to learn how to build descriptions of the world as we move an eye around. But I really want to build them and look at them, and study them more. If I start from images close to each other I will have a chance to achieve this goal, but I will only be studying a motion problem. Your problem seems harder.

Arc: I am glad you realize that. It is harder because it is not only related to moving around and understanding your immediate extra-personal space; it also has to do with recognizing the world, what is there. Because if you know an object, you recognize it by matching that knowledge—a model-to the image

[3]Yiannis will add a note
[4]Yiannis will add a note

or images. So, whatever form this model takes, recognition is the problem of deciding whether two views (one is the image and the other is a view of a model) are separated by a rigid transformation. It's the same problem.

Mel: I see that, but again you don't know what models to extract to use for recognition. That's why I need to get to the case of two close views. We need to know whether we can automatically develop models of the scene. Then, at least we will know what object representations we will ever be able to get.

Arc: I like your enthusiasm but I can still study this space of rays. It's so well defined and so neat. It may have surprises. As for objects, I don't even know what they are, and please don't involve me in cognitive discussions because I have problems understanding cognitive arguments. I simply tell you that I don't know what an object is, or how it is defined.[5] I like well-defined problems. These are problems I may be able to solve; the others, who knows? Consider it this way. Take a look at this picture and tell me what objects you see. You will see some, but not all. Different people may not see the same objects. Clearly, I not only recognize the objects but I find them as well. So, I recognize and segment probably at the same time.[6] But enough said. For the time being, go to Apollonios. He will tell you about frames close to each other. He will also tell you about some patterns that he discovered in moving images. I bet you like them. I like your problem, and you are on a good path. I told you all I know about the geometry of multiple views. Think of it as a tool, not as the answer to the geometry of visual space-time. When you use this tool wisely, you can only profit.

Mel: But points and lines? The world has so much more than that.

Arc: Yes, but with points and lines I know what I am doing. Don't forget that many applications have points and lines. How about pictures of buildings?

Mel: True, but what about correspondence?

Arc: The problem of visual correspondence is like a nightmare. Here is why: We are stating our problem by having two views of the world. That's the beginning. Given these views, we will match points in the views and we'll start our geometry.

Mel: So far, so good.

Arc: But you see, we are silently assuming that the two cameras are looking at the same scene.

Mel: Oh yes! Otherwise, what would it mean to correspond? It would make no sense.

Arc: Right. I could give you two views of different scenes and your program that attempts to solve the correspondence problem would just have to trust me! Look at it this way. If we have one camera on the earth and another one on

[5] Yiannis will add a note.
[6] Yiannis will add a note.

Venus looking at terrain, we wouldn't expect to correspond anything. But if we turned both cameras towards the polar star, we might be able to correspond. Whether you can potentially correspond depends on the camera geometry in relation to how distant the scene is.

Mel: Hmm! It seems to me that the requirement that both cameras look at the same scene is not explicitly incorporated into the definition of the problem. That's a reason why the problem is hard.

Arc: Exactly. Now, here is the weirdest thing. To ensure that the cameras are looking at the same scene, you should know something about the camera geometry and the scene geometry.

Mel: Obviously. But wait a minute. That's what I am trying to find.

Arc: Perfect. I think you got my point. To even think about finding correspondence, you should know something about camera placement and a scene model. But to find this, you need the correspondence. It's like a chicken-egg problem! That's why I stay away from it.

Mel: But what about my problem?

Arc: Your problem is your problem. "Euclid gave a definition for a point. A point is that which has no part. Sometimes, when we study in astrophysics the Milky Way galaxy, we use a point to denote it. Sometimes we use a point to denote the earth, and at other times to denote a molecule. When you study vision and images the question is, what is a point?"[7] We have been considering points in images as intersections of lines. That is, our image points are like Euclid's ideal points. You now can ask yourself some questions. What are the structures in images that have no parts, that is, what are the things that are the seeds of perception? What is a quantum in an image? I don't know what they are. Maybe from this work these things will come out, but for now I have no idea. Whatever they are, they obey the geometry I explained to you. Most of it has been known for a long time by a group of people known as photogrammetrists.[8] In recent years it has been understood that linear maps can be found[9], and so it is feasible and easy to find the camera geometry, and subsequently 3D scene models, given correspondences. The essence of your problem is to find two transformations. The 2D transformation that matches the images taken from different viewpoints and the 3D transformation (rotation-translation) that relates the two viewpoints. Go to Apollonios. I hope you find what a point is in an image. It was a pleasure speaking with you.

[7]footnote here
[8]Yiannis will add a note.
[9]Yiannis will add a note.