Chapter 7

COMPUTER VISION IN VISUAL EFFECTS

Doug Roble

7.1 Introduction

Computer vision has changed the way movies are made. Sure, computer graphics gets all the press, but computer vision techniques have made a significant impact on the way films with visual effects are envisioned, planned, and executed. This chapter examines the state of the practice of computer vision techniques in visual effects. We also examine the changes computer vision has brought to the industry, the new capabilities directors and visual effects creators have available, and what is desired in the future.

Let us examine the concept of visual effects in terms of computer vision. The main task of visual effects is to manipulate or add things to an image. Artists are very skilled, but the more information they have about the original image and the more tools the computer presents to them, the more effective they can be.

Computer vision has made a huge impact in the field of visual effects over the last decade. In the early 1990s, we saw the shift away from using physical models and photographic techniques to create and add fantastic things to an image. It has became commonplace to scan each frame of film into the computer and composite computer-generated elements with the filmed background image. Digitizing the filmed image is so common now that whole films are digitized, manipulated in the digital realm, and filmed out to a negative for printing. (*Oh Brother, Where Art Thou?* is a good

example of this. The color of each frame of the film was digitally adjusted to give the film its "look.") And, of course, digital cameras are starting to make inroads into Hollywood. *Star Wars, Attack of the Clones* was shot completely with a digital camera.

Once a sequence of frames is digitized, it becomes possible to apply standard computer vision algorithms to the digitized sequence in order to extract as much information as possible from the image. Structure from motion, feature tracking, and optical flow are just some of the techniques that we examine in relation to filmmaking.

7.2 Computer Vision Problems Unique to Film

7.2.1 Welcome to the set

A film set is a unique work environment. It is a place of high creativity, incredible tension, sometimes numbing boredom. The set is a place of conflict in the creation of art. The producers and their staff are responsible for keeping the show on budget and on time, and they fear and respect the set, for filming on a set is very expensive indeed. The director and crew are trying to get the best possible images and performances on film. These two forces are at odds with each other on the set, and often the computer vision people are caught in the middle!

The set is where it all begins for the digital artists responsible for what we call *data integration.* Data integration is acquiring, analyzing, and managing all the data recorded from the set. It is common for a visual effects facility to send a team of one or more people to the set with the standard crew. Before, during, and after filming, the data integration crew pop in and out of the set, recording everything they can about the makeup of the set. This is where the trouble starts: the producers are not crazy about the data integration crew; they are an added expense for a not so obvious result. The directors can get irritated with the data integration team because it is just one more disturbance on the set, slowing them down from getting one more take before the light fades or the producers start glaring again.

At least that is the way it was in the beginning. Movie makers are getting much more savvy and accustomed to the power that computer vision provides them in the filmmaking process. While a director might be somewhat annoyed that a vision person is slowing the process slightly, he or she is infinitely more annoyed when the visual effects supervisor tells him or her how to film a scene because of the inability of the effects people to deal with a particular camera move.

Watch some movies from the 1980s. Special effects were booming, and there were elaborate, effects-laden movies in that decade. But effects shots always telegraphed themselves to the audience because the camera stopped moving! If a stop-motion creature effect or some elaborate set extension was planned, the director could either lock the camera down or move it in very limited ways (a rotation about the lens' nodal axis or a simple zoom could usually be handled, but nothing more). Now directors can film however they want. The camera can be on a giant crane, a plane, a Steadicam, or even a handheld shot. If there is not enough information in the scene for the computer vision algorithms to work, we drop objects in the scene and paint them out later! (A couple of problem scenes for the ambitious researcher out there: a camera mounted in a plane flying over the ocean on a clear day—computing the camera location of each frame is tricky, because all your reference points are moving! And yet, if the camera track is not correct, inserting a boat on the surface of the ocean will look wrong. Or, a camera carried by a skier filming on a smooth glacier—this is where adding objects makes the problem go away, but without them, there is no hope!)

For a major production with many effects shots, a team of two to three people is sent to the set to record information. They typically stay with the film crew throughout the production, traveling to outdoor locations and working on the indoor sets that have been constructed. This data integration crew records everything it can to make the digital artists' lives easier. Here is a list of things that the crew tries to capture:

- **Environment maps:** Environment maps are spherical (or semispherical) photographs of the set, generally taken from near center of the set or near where the digital object is going to be inserted into the set. Digital artists began running into the center of sets with a chrome ball in the late 1980s and early 1990s. The quickest, cheapest way to acquire an environment map is to go to your garden supply center, buy a decorative chrome ball, and mount it on a tripod (Figure 7.1). Then, when the chance presents itself, run into the center of the set and snap two quick photographs of the chrome ball from both sides. It is relatively easy to map the pixels of the image to the area on the ball, and from that, it is possible to project the images onto a spherical environment map.

  This environment map is then used in the renderer to render the digital objects as if they are in the environment of the set. Shiny objects reflect portions of the set, and even diffuse objects change their color depending on the incident light from the environment maps.

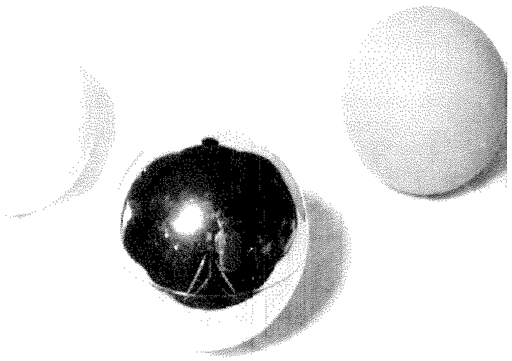

Figure 7.1. Chrome and diffuse balls used to capture environment maps.

  This ability to light digital objects with the light that existed on the set has caused a flurry of development over the last decade. Paul Debevec was a pioneer in the use of high dynamic range (HDR) imagery in conjunction with environment maps, showing just how powerful the combination can be.

  A snapshot of a chrome ball is only so useful. It creates a picture of a 360-degree environment, but since film (be it digital or photochemical) has a limited dynamic range, much of the information in the image is inaccurate. All photographers have experienced this to some extent. Setting the exposure controls to capture detail in the dark areas of a scene means that the bright areas will be overexposed and clipped. Alternatively, setting the exposure for accurately capturing the bright areas of the scene means that the dark areas will have no information in them.

The concept behind HDR photography is simple: take multiple images of a scene each with a different exposure. Then, combine the images into one image by using only the nonclipped pixels of the images and an appropriate mapping function. The results are amazing. Instead of an image with 8-bit pixels and values that go from 0 to 256, HDR images store floating-point numbers that go from zero to a very large number indeed. It is even possible (and sometimes very useful) to have negative color values!

There are many advantages to HDR images, but in the context of environment maps, the main advantage is that they are an accurate way of capturing the light energy of the environment, not just a simple snapshot. These environment maps can be used as elaborate lights in a rendering package and to create stunningly convincing images. A good introduction to these concepts can be found in [11] and [3].

Camera and object motion: Computing camera location and motion is where computer vision first made inroads into the film industry. Directors acknowledged, but did not like, the constraint that the camera had to stop moving or had to move in a very specific way if an effect was going to work.

The first method used to record camera motion did not rely on visual effects at all. Encoding devices were attached to the moving parts of a crane or a dolly. The measurements from the encoding devices were sent to a computer and, combined with an exact knowledge of the rig's geometry, the location and orientation of the camera were determined.

These measurements were then sent to a motion-control camera rig, and the motion of the camera was replicated. A motion-control camera rig is basically a large robotic camera mover. Very precisely manufactured and with very accurate stepper motors, the best motion-control rigs could move the camera starting at one end of a long set, perform a complicated move across the set, return to the starting point, and be only mere millimeters off from the original location.

Of course, the encoding process is rife with errors. It is well-nigh impossible to accurately encode the motion of the camera. The measurement and modeling of the rig is difficult to do exactly, it is impossible to encode any flex or looseness in the rig, and it is impossible to encode the inner workings of the camera. Given these problems, it is rare that a camera move is encoded on one camera system and replicated on another. More often, both the original camera move and the replication of the move was done on the same rig. This minimized the errors of the system and often worked well.

But, there is still a problem. Most motion-control rigs are big, heavy machines. They have to be, given the weight of the camera they are carrying and the need for stiffness in the rig. So, they could never move very quickly—certainly not as quickly as a handheld camera or a camera mounted on a dolly or a crane.

Given all these limitations, motion-control rigs were only used for shots with models or very limited shots with actors.

Recently, "real-time" motion-control rigs have been used on sets. The Bulldog Motion Control Camera rig from ImageG received a Technical Academy Award in 2001. It is extremely fast and can be used to reproduce very fast camera moves.

What does this all have to do with computer vision? Quite a bit, actually. Motion-control rigs are still used in visual effects because it is often much more cost effective to build a detailed model of a ship or location than to construct it all as a digital set. Industrial Light and Magic built a large model of a pirate ship and sailed it in a pond behind its facility for a large number of shots in the film *Pirates of the Caribbean* (2003). It was just more cost effective.

There are two problems associated with motion-control rigs and computer vision. The first problem is relatively easy. If a model is filmed with motion control, the images are often enhanced with digital effects. Unfortunately, even though the camera path is known from the motion-control move, it cannot be trusted. The path is very repeatable, but it is not subpixel accurate in the digital realm—the internal camera motion and flex of the devices is not recorded. However, the camera pose estimation is a very easy subset of the general problem. Once features in the image are identified, the motion of the features is easy to predict because of the known motion of the camera. In fact, while the overall path of the camera may have errors, the incremental motion from one frame to the next is quite accurate. So the camera motion is used to predict the feature motion, and a quick template match in the local area is performed to find the features to subpixel accuracy. Standard structure from motion techniques are used to refine the motion. Convergence is very fast, since the starting path is quite good.

The second problem with motion-control cameras is quite difficult and, to some extent, impossible to solve. Often it is required to replicate the motion of a handheld camera with the motion-control rig. Often, a director films the actors with a "wild" camera in front of a bluescreen. Then, the camera move is later reproduced on a motion-control rig filming an elaborate model. The actors are composited with the filmed model for the final shot. Tracking the camera on the stage is not terribly difficult. The data integration team places markers or known geometry on the bluescreen, and pose estimation techniques are used to recover the camera path. The problem comes when transferring the camera motion to the motion-control rig.

Motion-control rigs do not stop before a frame of film is shot. It is a common misconception, but they do not move, stop, shoot, move, stop.... Rather, they shoot the sequence of frames in one continuous motion. There are many advantages to continuous motion, but the two overriding benefits are speed of shooting and motion blur. Stopping before each frame slows down the motion-control rig terribly, and stage time is expensive. If the camera is moving while the shutter is open, one gets motion blur for free. Model shots look very unconvincing without motion blur.

This continuous motion of the rig leads to the problem. Because of the size of the motion-control rig, there are physical limits to the torque of the stepper motors and how fast they can move the rig. Therefore, at the speed that is needed for the shot, the camera may not be able to exactly reproduce the desired motion. When the foreground and background are merged, they may not line up, and high-frequency motions of the camera move will not be preserved in the background plate.

This problem is difficult to fix. If the problem is extreme enough, the motion-control camera motion is tracked (using methods described above) and compared with the original camera motion. Three-dimensional points in the scene are then projected onto the image plane of the real camera and the motion-control rig. The discrepancy of these points is then used to build a general warping field that is applied to the background images. Naturally, this is a 2D fix to what is a 3D problem, and it cannot correct all the problems, but it usually suffices. Of course, a generalized warp will filter the image and effectively reduce the resolution of the background image so the images are typically digitized at a higher resolution.

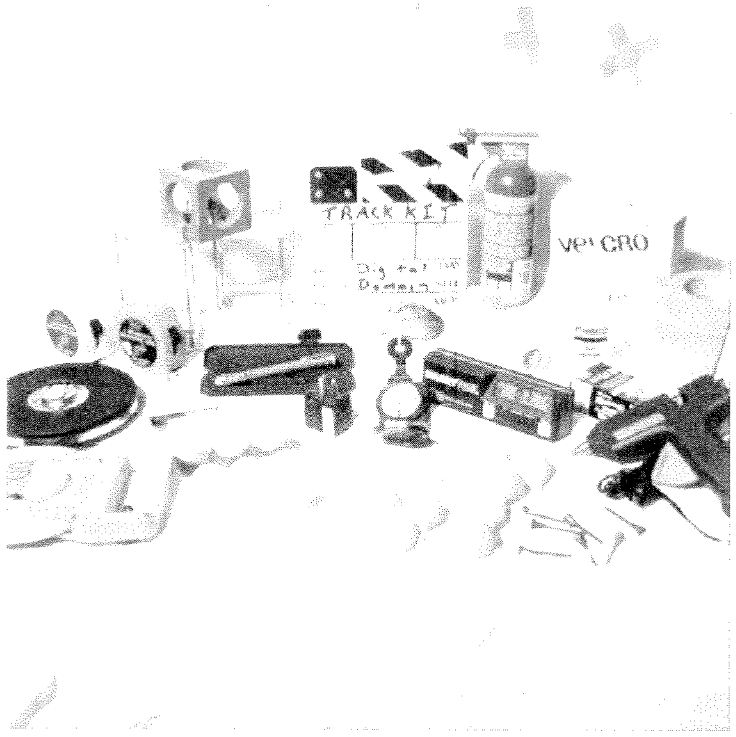

Figure 7.2. Typical gear in a data integration tracking kit.

Set and object geometry: The data integration team is responsible for capturing geometric information about the set. This can range from some simple measurements of the height of a locked-off camera or where a light is in relation to the camera to full 3D reconstruction of the set with textures and BRDF (bidirectional reflectance distribution function) analysis of the surfaces.

Most data integration teams have assembled a "tracking kit" (Figure 7.2) that is brought to the set and (hopefully) has all the devices needed to capture the required information. The following items are included in most kits:

- **Tape measures:** Tape measures are still the fastest way to get at least some measurements from the set. Often, a data integration team member dashes onto the set and measures the distance between two notable landmarks. This data is used to give a scale to the geometry reconstructed using photogrammetry.
- **Tripods:** A good tripod is essential for the work the data integration team does. Being able to lock down a camera and have it stay locked down is very important for HDR image recording.
- **Chrome and diffuse spheres:** As discussed earlier, the data integration team is responsible for capturing environment data in the scene. These spheres are mounted on a tripod, set in the center of the scene, and photographed from various points of view. These images are then stitched together for a complete environment map.
- **Camera calibration charts:** High-quality checkerboard grids are taken on the set to determine the intrinsic parameters of the cameras being used. These charts are typically mounted on a rigid metal plate to insure accuracy. Sometimes, because of the lens being used or the situation, a larger calibration chart is needed. In this case, the grid is printed on large-format, plasticized paper. This large grid—typically 2 or 3 meters on a side—is attached to a flat wall.
- **Tracking geometry:** Visual effects companies typically have a machine shop to build necessary items for model shoots. The machine shop is also used to build highly accurate reference geometry for computer vision applications (Figure 7.3).

  Metal cubes, from a couple of centimeters to as large as a meter on a side, are placed in a set to provide some known geometry. The cubes help provide scale for photogrammetry and are also a way to measure and control error.

  These reference devices are not left in the set during actual filming, but are often filmed by the main camera and removed before main filming.
- **Brightly colored tape and Ping-Pong balls:** A huge problem, from a computer vision perspective, of shooting on a bluescreen set is that there is nothing to track! The background is a uniform blue, and the foreground is typically moving people. Quite challenging for camera tracking!

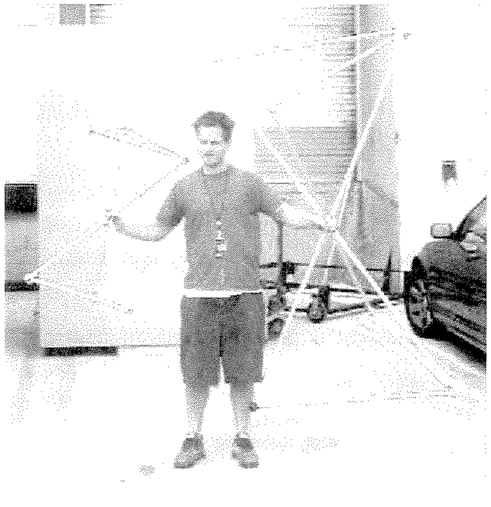

Figure 7.3. Large, rigid geometry used for photogrammetry and camera calibration.

  Hence the need for tracking objects. Pieces of tape or colored Ping-Pong balls are taped to the bluescreen background. These provide handy reference points for the tracking algorithms but do not cause too many problems for the matte extraction algorithms.

  As an aside, bluescreen shots can be quite challenging for a good camera track. While being able to place reference markers may seem ideal, there is a delicate balance between placing enough reference points and placing too many for an easy matte extraction. (The reason that the crew was filming in front of a blue screen was to make it easy!) But often, more of a problem is that the blue screen is usually quite a distance away from the actors and is typically a flat screen. Computing the pose of a camera with points that are all coplanar can be difficult sometimes, and the fact that the reference points are far away from a camera that probably does not move all that much makes accuracy even more challenging. Finally, the actors are always a factor. Depending on the number of actors, all the tracking markers may be obscured.

  What happens in the worst case, if the camera is moving but the actors have obscured all the reference points? We rely on the data integration artist to solve the problem by hand. Camera positions can be interpolated from known points, but in worst-case sections, there is no alternative to solving it using the artist's skill at knowing "what looks good."

- **A total station survey device:** Back in the early 1990s, when computer vision techniques were starting to make an impact on filming and the set, it was quite common to hire a survey crew to create a survey of reference points on the set. These points were used in a camera pose estimation algorithm to compute the location of a moving camera for each frame of film.

  As computer vision techniques became commonplace, visual effects studios purchased their own survey devices and trained people in their use. Now, almost every studio owns a "total station." Total station is the generic name for the survey devices used by burly government workers often seen blocking traffic. A total station is a tripod-mounted, highly accurate, laser distance and angle measure. A total station has an accurate telescoping sight and a laser pointer. Once the station is set up, the survey person simply sights or points the laser at a spot in the terrain and presses a button. The distance to the spot is measured using phase measurement. The angle and distance provide the 3D location of the point. Key points in the set can be digitized quite quickly.

  The problem with total stations is that they are time consuming to set up. And if the station is moved, there is a realignment process to compute the location of the total station in relation to the first setup point. Also, the only points that are digitized are the points that are selected during the session. If the data integration artist needs to compute a point that the person on the total station did not compute, different photogrammetry-based techniques must be used.

  The huge advantage of a total station is that a highly accurate reference geometry of key points in the set is built. This can be enormously helpful for camera or object tracking later. Relying solely on structure from motion techniques to compute the location of the camera is unwise—those techniques are notoriously bad if the camera focal length changes during the shot. This certainly happens during a zoom and often happens during a simple change of focus. With reference geometry, the camera's intrinsic parameters can be calculated from frame to frame. This gives the director the ultimate freedom on a set.

- **Digital cameras with different kinds of lenses, including a fisheye:** Of course, the data integration team carries a digital camera on the set. This is used to record the layout of the set, reconstruct the geometry of the set, and build environment maps. The cameras are usually high resolution and have calibrated lenses.

  Data integration teams usually carry a high-quality fisheye lens to make environment maps, but there are issues with these lenses. Fisheye lenses can be used to quickly build a 360-degree panorama of the set, but they also typically have poor optic qualities, particularly large bloom around bright light sources or reflections. These can be difficult to deal with when assembling the HDR images—they are typically painted out by hand!

  To deal with this problem, there are special-purpose 2D slit cameras mounted on a motorized tripod head. These cameras slowly spin around, grabbing a slice of the environment at a time, usually changing the exposure as they go. They do not exhibit the bloom problem that fisheye lenses do, but they are much slower and much more expensive.

- **Video camera (progressive scan):** A video camera is also brought to some sets or location shoots. One or more can be used to augment what the main film (or digital video) camera is shooting. We have built stereo rigs that mount a video camera a fixed distance away from the main film rig so that a stereo correspondence can be built on every frame. This is useful for closeup shots of actors where the moving geometry of the face needs to be digitized.

  Of course, all the problems associated with stereo are present in this setup, including changing light sources, nonideal subjects, and changing lens parameters.

- **LIDAR (light detection and ranging):** Since the late 1990s, another option for building set geometry has been LIDAR. A LIDAR unit looks similar to a total station and accomplishes the same kind of task, just at a much finer detail and automatically. A LIDAR device is a time-of-flight laser mounted on a mechanical base that rotates the laser up and down and left and right. In a certain amount of time, a swath of the set can be scanned to high accuracy and great detail.

LIDAR produces a cloud of 3D points that represent the surfaces that the unit can see from a given position. By moving the device and scanning from different positions, the multiple clouds of points can be aligned and joined to provide a remarkably complete point-based representation of the set. The range of a professional LIDAR device is around 100 to 200 meters, so it is effective both inside and outside.

The density of points that LIDAR produces is both a strength and a weakness. Just a couple of scans produces hundreds of thousands of points which, though amazing to look at, are useless to a digital artist who is trying to recreate a set digitally. Some sort of surface representation is much more useful to the artist.

Computing the appropriate surface from a cloud of points is an ongoing research project. Some successful techniques have been developed based on the use of radial basis functions. And recently, point-based representations of geometry have started to make an impact on this problem [8, 9].

Textures and beyond: Finally, the data integration team is responsible for recording the textures of the set. It is becoming quite popular to create completely digital representations of sets or locations. Director David Fincher has been a strong proponent of using photogrammetry techniques to completely recreate a set inside the computer. The visual effects company Buf Films recreated the interior of an entire house for him for the film *Panic Room*. But getting the geometry to mimic the real world is only part of the battle. A considerable amount of the "reality" of the real world is how light reflects off an object.

It is the data integration team's job to help the modelers who are building the digital set in any way they can. The most important thing to record is the diffuse textures of an object. Lighting the set with diffuse lights or using cross-polarizing filters on the light and the camera lens results in images with very few specular highlights. (Having a specular highlight baked into a texture is a bit of a pain. Part of the modeling process is cleaning up textures and painting out specular highlights. There are typically quite a few people working on this.)

Of course, as anyone who has played a video game can attest, textures are only one step on the way to reality. Next, shaders are written that use the textures and add things like bump mapping or displacement maps to give the surface of an object a feeling of reality. This is where the data integration team comes into play again. By photographing images of an object from different calibrated vantage points, the geometry of the object can be constructed, and the diffuse texture taken from one point of view can be compared with the diffuse texture from another point of view. Any disparity in the texture must be due to some displacement of the surface not modeled by the underlying geometry [4]. It becomes a simple stereo problem again. The disparity map produced by template matching along the epipolar lines intersecting the polygons of the surface can be used as a bump or displacement map in a shader. This helps the artists achieve a higher sense of reality.

Beyond displacement maps, changing the lighting model of the surface helps improve the reality of the digital model. Light rarely reflects off a surface in a purely diffuse model. Rather, it scatters according to a BRDF. This can be measured in the lab using a goniospectrometer, and a couple of big visual effects houses have purchased one of these devices. It consists of a moving camera that accurately measures the scattered light in a hemispherical area, producing an accurate distribution function for lighting models. Recently, there has been some exciting work on measuring approximate BRDF models without one of these cumbersome devices. It is now possible to measure a BRDF with a video camera and a flashlight moving in an approximate hemispherical arc [7]. Of course, the results of such measurements are quite crude, but they can be put to use by shader writers to make a surface look even better.

Finally, many objects do not simply reflect light using a diffuse and specular lighting model but rather allow the light to penetrate into the object where it scatters and bounces around before emerging again. This phenomena resulted in a modification to the BRDF model to produce the BSSRDF model in which subsurface scattering is taken into account. Marble exhibits this behavior, but even more notably, human skin is very translucent. Shine a laser pointer through your hand to see exactly how translucent it is. The data integration team often does just that. By shining a light into an object and photographing the glow produced by the subsurface scatter, a reasonable model of the scatter term can be built.

Obviously, the data integration team needs to be equipped to capture all aspects of geometry and light on the set. It is a fun job, the set is an exciting place to be, and there is a certain thrill with working alongside celebrities. But that thrill quickly fades as the reality of a high-pressure job with long hours and often a long schedule sets in. Once the film is exposed and the sets are struck, the team comes back to the office, and the processing of the images and data can begin in earnest.

7.3 Feature Tracking

Digital artists use one computer vision technique more often than all others: pattern tracking, or template matching. Digital artists who specialize in 2D effects (rotoscoping, wire removal, compositing) use pattern tracking nearly constantly. Being able to accurately follow a feature in an image is vital for quick, accurate effects.

As an example of how this is used, consider the film *Interview with the Vampire*, released in 1994. In that film, the vampire Lestat, played by Tom Cruise, is cut with scissors by another vampire. These cuts are on both cheeks. The director wanted the cuts to appear on the cheeks, bleed slightly, then fade as the vampire's healing power took hold. All within one continuous shot.

Of course, one really cannot cut Tom Cruise's cheek, and even if it was possible with makeup, getting it to heal convincingly on camera would be very difficult. Luckily, this kind of effect was fairly commonplace even back in the early 1990s. Digital artists modeled two pieces of "cut" geometry and animated it with standard modeling tools. These elements were rendered out as images, assuming no motion of Tom's head. They simply animated the cuts using the first frame of the sequence as a guide.

The challenge then was to track the cut elements to Tom's moving head throughout the shot. This is where computer vision techniques came in to play. At that time, the major digital image manipulation and compositing packages all had some method of pattern tracking. They are usually based on the minimization of the cross-correlation of the original image area in a search image area. J. P. Lewis wrote a seminal paper on template matching that many visual effects companies still use today [6].

In the case of *Interview with the Vampire*, two spots on Tom's right cheek were pattern tracked, and the resulting positions were used to scale and rotate the original images of the rendered cut to track along with the motion of the cheek. Because Tom did not move his head much, the artists were able to get away with only using a two-point technique. If the movement of the cheek was more complicated, the artists could use four points and a much more complicated warp of the image to track it exactly. (Of course, there are times where even a complicated deformation of a 2D image will not hold up: in that case, we need to do a more complicated 3D track of the camera and object move so that the effect can be rendered in the right position. This is discussed later.)

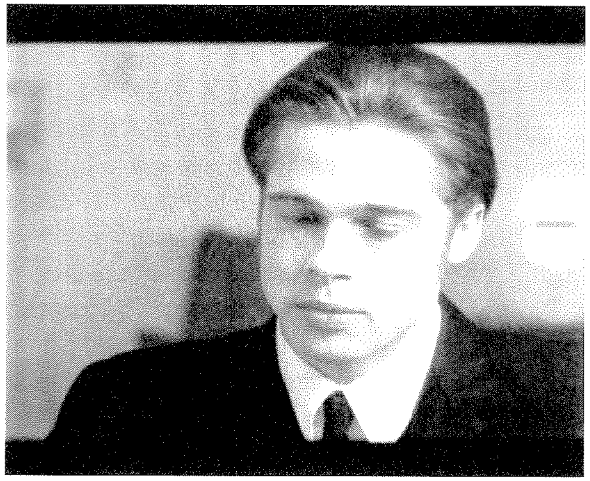

Figure 7.4. Brad Pitt's makeup was toned down using feature tracking in *Interview with the Vampire*.

Pattern tracking has become such an integral part of a digital artist's toolbox that it is used all the time without audiences even noticing. In the same film, Brad Pitt's vampire makeup was too noticeable on screen. The director asked if it was possible to tone the makeup down. Feature tracking solved the problem again. The digital artist created a 2D digital makeup element—basically some pancake makeup—and this was tracked to the offending vampire veins. A quick composite later and the veins had been masked to a certain extent. See Figure 7.4.

The effect was quite successful, but it was not exactly easy. Pattern tracking in a film environment faces many challenges. First, it must be robust in the presence of noise. All vision techniques must deal with noise, and pattern tracking is no exception. A recurring theme in this chapter is the audience's ability to notice even the smallest error when it is magnified on the big screen. Some films are even being transferred to large-format film and projected on an Imax screen sure to expose any flaws in the graphics or vision!

In dealing with noise, there are a couple of techniques that work with pattern tracking. One is relatively simple: the median filter. A small median filter does a nice job of removing the worst noise without destroying the underlying geometry of the image. More elaborate, the noise in the image can be characterized by analyzing a sequence of images over a still subject. By applying an adaptive low-pass filter over the image that follows the characteristics of the individual camera, the noise can be knocked down quite a bit.

Extreme lighting changes still stymie most pattern tracking algorithms. A high degree of specular reflection blows out most underlying information in the image and is quite difficult to deal with in a pattern track. Flashes of lightning or other extreme lighting changes also produce bloom on the image that change the geometry of the pattern being tracked. At this point, there is no recourse but to track the offending frames by hand.

7.4 Optical Flow

Recently, over the last five years, optical flow techniques have really started to make a big impact on the film industry. In fact, there are a couple of companies that do fairly well selling software based on optical flow techniques.

Optical flow is the 2D motion of the features in an image sequence, from one frame to the next [1, 10]. Consider a pattern track centered on every pixel. After an optical flow analysis of an image sequence, a secondary frame is generated for every original image in the sequence. This new image does not contain colors, but rather a 2D vector for each pixel. This vector represents the motion of the feature seen through the pixel from one frame to the next.

In an ideal case, if the pixels of the image n were moved all along their individual motion vectors, an exact replica of image $n+1$ will be created. The pixels in image n would be pushed around to make image $n+1$.

Of course, things never work out that perfectly, but it is possible to see the power of such information. The first application of optical flow follows naturally: change the timing of a sequence of images. If the director shoots a sequence using a standard film camera at normal speed, 24 frames are exposed per second. It happens occasionally that the director wants to change the speed of the film, either slow it down or speed it up. Optical flow helps in either case.

Slowing down an image sequence is relatively easy. If the director wanted to see the shot as if it were filmed at 48 frames per second, one simply duplicated each frame of film. This is what was done before the advent of digital image manipulation. A single film negative was printed onto two frames of print film. Playing the resulting film back resulted in slow motion. However, it looked a little jerky. The film pauses for 2/24 second, and it is a noticeable effect. The solution was to generate in-between frames that were blends of image n and image $n+1$. Again, this was done during the printing process: the in-between frames were generated by exposing the print film with the n and $n+1$ negative frames for only half the exposure time each. This created an in-between frame that looked different than the original frames, but also looked pretty bad. It was a double exposure of two different images and exhibited quite a bit of strobing and streaking.

Optical flow came to the rescue. In the explanation above, it was shown that by moving the pixels along the optical flow motion vectors, a new image can be generated. By scaling the motion vectors by half, the new image is exactly between the two original images. This new image typically looks quite a bit better than a blend of the neighboring images, and it does not exhibit any of the blurred or double exposed edges. It also has the advantage that it can be used to slow the original sequence down by any factor. New images can be generated by scaling the motion vectors by any value from 0.0 to 1.0.

But pause a moment. It was stated earlier that if the pixels are pushed along their motion vectors, the image at $n+1$ can be created from the pixels that make up image n. Unfortunately, it is not that easy. The motion vectors typically go in all directions, and sometimes neighboring motion vectors diverge or converge. There is no guarantee that the pixels in image $n+1$ will be all filled by simply following the motion of the pixels. Instead, the operation of creating the in-between image is a "pixel pull" operation. For each pixel in image $n+1$, a search is done to find the motion vector that will land in the center of the pixel. This motion vector will probably be the motion vector located at a subpixel location that is the result of an interpolation between the four neighboring pixels. A search is made in the local neighborhood of the destination pixel to find a likely starting pixel location in image n. Then, a gradient descent is used to compute the exact subpixel location that, when added to the interpolated motion vector, lands on the center of the desired pixel.

But there is more! Interpolating the color of the subpixel leads to sampling and aliasing artifacts. In fact, if the motion vectors converge—that is, if the colors of a group of pixels all end up at one pixel in image $n+1$—the original pixels should be filtered to generate an accurate pixel color. Typically, the previous backwards tracing is done from the four corners of the pixel. This creates a four-sided polygonal area that is sampled and filtered. These results create a superior in-between image.

Speeding up a sequence of images is just as straightforward, with a slight twist. Consider the problem of speeding up the action by a factor of two. In that case, it is desired to make it look like the action was filmed with a camera that ran at 12 fps. It should be easy, right? Simply throw away every other frame from the original sequence! This works fine, with one problem: motion blur. The objects in the images are now supposed to be moving twice as fast, and the motion blur recorded on the images looks like it was filmed with a 24-fps camera. This produces the strobing, harsh effect normally associated with sped-up film.

Optical flow comes to the rescue in this case as well. The motion vectors associated with each pixel can be used to streak the images so that the motion blur appears correct. Like the sampling issue before, blurring an image according to optical flow motion has some hidden issues. The easy, but incorrect, way to add motion blur to the image is to use the motion vector to create a convolution kernel. The entries in the convolution kernel that are within a certain distance from the motion vector are given a weight based on a Gaussian falloff. The final kernel is normalized and convolved with the image to produce the resultant pixel. This is done for each pixel in the image. This technique certainly blurs the image in a complex directional sense, but it does not quite produce correct motion blur. (It is reasonably fast—it does require a different kernel computation at every pixel—and quite acceptable in some situations.)

This method, as mentioned, is incorrect. Consider this example: a model of a spaceship has been filmed on the set with a black background behind it. After the shoot, it is decided to add more motion blur to the shot—maybe it makes the shot more dynamic or something. So, somehow, motion vectors for each pixel are computed. It need not be optical flow in this case. Since the model is rigid, it is possible to use the motion of the motion-control rig or track the motion of the model to generate a reproduction of the model shot in the computer. This can then be rendered so that instead of writing out the color of every pixel, the pixels contain the motion vector. Whether created by optical flow or some rendering technique, assume the pixels contain the 2D motion vector of the feature under the pixel. Now, consider what is happening at the edge of the model. A pixel covering the edge of the model will have the appropriate motion vector, but the neighboring pixel that only covers background contains no motion at all! The background is black, and even if the camera is moving, it is impossible to detect the motion. The problem becomes evident at this point. Using the previous technique to add motion blur if the spaceship is moving fast, the pixels covering the spaceship will be heavily blurred. The pixels right next to the spaceship will have no blurring at all. This causes a massive visual discontinuity that audiences (and visual effects supervisors) are sure to notice.

There is a (mostly) correct solution to this: instead of blurring the pixels, distribute the energy of the pixels along the motion vector. This gets a little more complicated in that the motion vectors are converging and diverging in places. This caused problems in interpolating images, and it causes problems in adding motion blur. Basically, the color of a pixel is distributed along the motion vector path from the pixel. Doing the naive approach of simply adding a fraction of the pixel's color to the pixels that intersect the path produces bright and dark streaks in the motion blur where the vectors converge and diverge. Rather, the interpolated motion vectors must be used at each corner to create a motion polygon of converge. Then, the color is added to each pixel under the motion polygon, using a weighting that is the area of coverage of the individual covered pixel divided by the total area of the motion polygon. This produces lovely motion blur at quite a cost in computation time. In the reality of visual effects, we try to get away with the cheap solution for motion blur, knowing that we can always pay the price for the more accurate version of motion blur.

The previous paragraphs seem to indicate that optical flow is a robust, trustworthy algorithm. This is not the case. Optical flow gets terribly confused with pixels that contain features that disappear (or appear) in the space of one frame. This is commonly referred to as foldover. Also, specular highlights and reflections in objects can fool even the best optical flow algorithm. And most optical flow algorithms assume a certain spatial continuity—one pixel moves in pretty much the same direction as its neighbor—which makes computing optical flow on complicated moving objects like water or blowing tree leaves quite difficult.

There are many solutions to these problems. If the foldover areas can be found, they can be dealt with by predicting the motion of the pixels from previous or following frames. Sequences of images are often analyzed both forwards and backwards; where the pixel motion vectors do not match up is usually a good indication of foldover. Most optical flow algorithms also compute an indication of confidence in the flow for a pixel or an area. Where foldover areas are indicated, the discontinuity in the vectors is encouraged rather than avoided. And interpolating the pixels for in-between frames is done by recognizing that the pixel will either be disappearing or appearing and interpolating in the appropriate direction.

Optical flow is not just for image manipulation. It is increasingly being used to extract information about the image. In 2002, Yung-Yu Chung et al. published an influential paper titled "Video Matting of Complex Scenes" [2]. The techniques in this paper have made an impact on the film industry: it is now quite a bit easier to extract a matte from an object that was not filmed in front of a bluescreen. The paper describes a technique for computing the matte based on Bayesian likelihood estimates. Optical flow comes into play by pushing the initial estimate of the matte (or in this case, the trimap—a segmenting of the image into foreground, background, and unknown elements) around as the contents of the image move. It uses the same technique as described above for computing an in-between image, but this time applied to the trimap image. Many visual effects companies have implemented compositing techniques based on this paper.

Additionally, ESC Entertainment used optical flow to help reconstruct the facial motion of the actors in the movies *Matrix Reloaded* and *Matrix Revolutions*. They filmed an actor in front of a bluescreen with three (or more) synced cameras. Instead of using traditional stereo techniques, which have problems with stereo correspondence, the artists placed a 3D facial mask on the face of the actor in each film sequence. This was relatively easy to do for the first frame of the film—the actor's face had not deformed from the digitized mask. The sequences of film were analyzed for optical flow. Then the vertices of the mask were moved in 2D according to the optical flow. Since each vertex could be seen in two or more camera views, the new 3D position of the vertex was calculated by triangulation. Doing this for each vertex on every frame produced quite clean-moving, deforming masks of face's of the actors. This avoided any stereo correspondence problem!

7.5 Camera Tracking and Structure from Motion

With the advent of computer vision techniques in visual effects, directors found that they could move the camera all over the place and the effects crew could still insert digital effects into the image. So, naturally, film directors ran with the ability and never looked back. Now all the large effects shops have dedicated teams of computer vision artists. Almost every shot is tracked in one way or another.

In the computer vision community, the world space location of the camera is typically called the camera's extrinsic parameters. Computing these parameters from information in an image or images is called *camera tracking* in the visual effects industry. In computer vision, this task has different names depending on how you approach the solution to the problem: camera calibration, pose estimation, structure from motion, and more. All the latest techniques are available to the artists computing the camera motion.

The overriding concern of the artists is accuracy. Once a shot gets assigned to an artist, the artist produces versions of the camera track that are rendered as wireframe over the background image. Until the wireframe objects line up exactly, the artist must continually revisit the sequence, adjusting and tweaking the motion curves of the camera or the objects.

A camera track starts its life on the set. As mentioned before, the data integration team gathers as much information as is practical. At the minimum, a camera calibration chart is photographed, but often a survey of key points is created or the set is photographed with a reference object so that photogrammetry can be used to determine the 3D location of points on the set. This data is used to create a polygonal model of the points on the set. (This model is often quite hard to decipher. It typically has only key points in the set, and they are connected in a rough manner. But it is usually enough to guide the tracking artist.)

The images from the shot are "brought online"—either scanned from the original film or transferred from the high-definition digital camera. Also, any notes from the set are transcribed to a text file and associated with the images.

Before the tracking begins, the lens distortion must be dealt with. While the lenses used in filmmaking are quite extraordinary, they still contain a certain amount of radial distortion on the image. Fixed lenses are quite good, but zoom lenses and especially anamorphic lens setups are notorious for the distortion they contain. In fact, for an anamorphic zoom lens, the distortion is quite hard to characterize. For the most part, the intrinsic parameters of the lens are easily computed from the calibration charts on set. For more complicated lenses, where the standard radial lens distortion equations do not accurately model the behavior of the lens, the distortion is modeled by hand. A correspondence between the grid shot through the lens and an ideal grid is built and used to warp the image into the ideal state.

However the lens distortion is computed, the images are warped to produce "straightened images," which approximate what would be seen through a pinhole camera. These images are used online throughout the tracking, modeling, and animation process. The original images are used only when the final results are composited with the rendered components of the shot. This seems like an unnecessary step; surely the intrinsic parameters of the lens can be incorporated into the camera model. The problem is that these images are not just used for tracking but also as background images in the animation packages that the artists use. And in those packages, a simple OpenGL pinhole camera model is the only one available. So, if the original images were used to track the camera motion and the tracking package uses a complicated camera model that includes radial (or arbitrary) lens distortion, the lineup would look incorrect when viewed through a standard animation package that does not support a more complicated camera model.

The implications of this are twofold. First, two different sets of images must be stored on disk, the original and the straightened version. Second, the artists must render their CG elements at a higher resolution because they will eventually be warped to fit the original images and this will soften (blur) the rendered image. Surprisingly, this does not impact the pipeline as much as one would expect; CG elements are typically blurred slightly to blend better with the original filmed plate.

The artist assigned to the shot uses either an inhouse computer vision program or one of many commercially available packages. For the most part, these programs have similar capabilities. They can perform pattern tracking and feature detection, and they can set up a constraint system and then solve for the location of the camera or moving rigid objects in the scene. Recently, most of these program have added the ability to compute structure from motion without the aid of any surveyed information.

In the most likely scenario, the artists begin with the sequence of images and the surveyed geometry. Correspondences between the points on the geometry and the points in the scene are made, and an initial camera location is determined. Typically, the data integration team has computed the intrinsic parameters of the camera based on the reference grid shot with the lens.

Once the initial pose is determined, the points associated with each point on the reference geometry are tracked in 2D over time. Pattern trackers are used to do this, though when the pattern tracks fail, the artist often tracks the images by hand.

Why would pattern tracks fail? One of the most common reasons is that feature points in the scene—the ones that are easy to digitize with a total station—are often not the best points to camera-track: corners of objects. Corners can be difficult to pattern-track because of the large amount of change that the pattern undergoes as the camera rotates around the corner. And often, the background behind the corner contributes to confusion of the pattern tracking algorithm. Corner detection algorithms can pick them out, but the subpixel precision of these algorithms is not as high as those based on cross-correlation template matching.

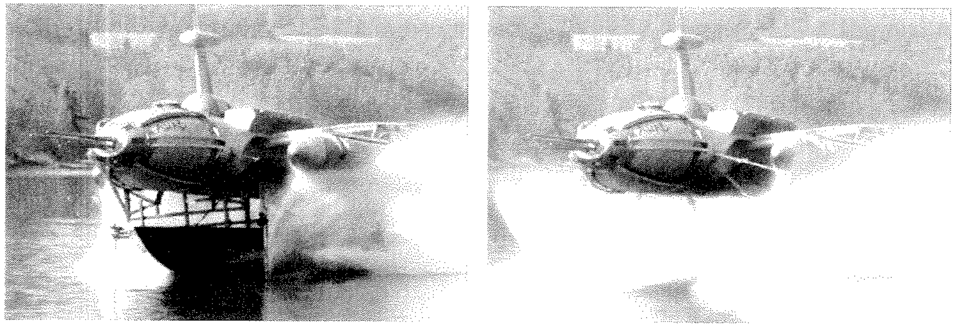

Figure 7.5. Snapshot of a boat/car chase in the movie *xXx*. Left: Original. Right: Final composite image with digital hydrofoils inserted.

Motion blur poses many problems for the tracking artist. Pattern trackers often fail in the presence of motion blur. And even if the artist is forced to track the pattern by hand, it can be difficult to determine where a feature lies on the image if it is blurred over many pixels. Recently, the artists at Digital Domain were tracking a shot for a movie called *xXx*. In the shot, the camera was mounted on a speeding car traveling down a road parallel to a river. On the river, a hydrofoil boat was traveling alongside the car. The effect was to replace the top of the boat with a digitally created doomsday device and make it look like it was screaming down the river on some nefarious mission (Figure 7.5). The first task was to track the camera motion of the car with respect to the background landscape. Then the boat was tracked as a rigid, moving object with respect to the motion of the camera. Naturally, the motion blur for both cases was extreme. The automatic tracking program had a very difficult time due to the extreme motion of the camera. Feature points would travel nearly halfway across the image in the space of one or two frames. Optical flow algorithms failed for the same reason. Tracking the points proved quite difficult because of the extreme streaking of the information in the image. The artists placed the point in the center of the streak and hoped for the best.

This worked to some extent. In this extreme example, the motion blur of the boat was at times so extreme that the streaks caused by the motion blur were not linear at all. The motion of tracked camera and boat were represented as a linear interpolation between keyframe positions at every frame. The motion blur generated by the renderer did not match the motion blur of the background frame. It is surprising how much small detail the eye can pick up. When the artists used spline interpolation to interpolate the camera motion and added subframe keyframes to correctly account for sharp camera/model motion, the digital imagery fit with the background noticeably better.

Beyond tracking points, digital tracking artists have many tools available for a good track. The points can be weighted based on the certainty of the points—points that are obscured by objects can still be used: even though they cannot be seen, their positions are interpolated (or guessed) and their uncertainty set very high. Camera pose estimation is still an optimization problem, and often the position computed is the position that minimizes the error in the correspondence of points. That being the case, removing a point from the equation can have quite an effect on the computed location of the camera. This results in the dreaded "pop" in motion after a point leaves the image. This pop can be managed by ramping the point's uncertainty as the point approaches leaving the image. Alternatively, the point can be used after it leaves the image: the artist pretends that the point's location is known, and this can minimize any popping.

Other constraints beyond points are available in many packages. Linear constraints are often useful: an edge on the surveyed geometry is constrained to stay on the edge of an object in the image. Pattern trackers can follow the edge of the object as easily as a point on the object.

Of course, it is not necessary to have any surveyed information. Structure from motion techniques can solve for the location of both the camera and the 3D points in the scene without any surveyed information at all. If there is no absolute measurement of anything on the set, there is no way for structure from motion algorithms to determine the absolute scale of the 3D points and camera translation. Scale is important to digital artists: animated characters and physical effects are built at a specific scale, and having to play with the scale to get things to look right is not something that artists like to do. So, even when relying on structure from motion, some information—like the measurement of a distance between two points on the set—is quite useful for establishing the scale of the scene.

The convergence of optical flow techniques, following good features for tracks, and structure from motion solvers has caused large ripples through the film and video production facilities. Now, with no a priori knowledge of the set, a reasonable track and reconstruction of the set can be created—as long as the camera moves enough and enough points can be followed. Some small visual effects facilities rely on almost all structure from motion solutions. Larger facilities are finding it to be an enormous help, but rely on more traditional camera tracking techniques for quite a few shots. Automatic tracking is automatic when it is easy.

Consider a shot where the camera is chasing the hero of the film down a street. At some point, the camera points up to the sky and then back down to center on the main character. Typical effects might be to add some digital buildings to the set (or at least enhance the buildings that are already in the scene) and perhaps add some flying machines or creatures chasing our hero. This makes the camera tracking particularly challenging: the camera, when it points to the sky, will not be able to view any trackable points. But, if buildings and flying objects are being added to the scene, the camera move during that time must at least be "plausible" and certainly be smooth and not exhibit any popping motion as tracking points are let go. Automatic tracking techniques are generally not used on this kind of shot because the artist will want complete control of the points used to compute the camera motion. The camera will be tracked for the beginning of the shot and the end of the shot, and then the camera motion for the area where the camera is only pointing at sky or nonvalid tracking points (like the top of the actor's head) is created by clever interpolation and gut feeling for what the camera was doing. Often, something can be tracked, even if it is a far-off cloud, and that can at least be used to determine the rotation of the camera.

7.6 The Future

Camera tracking and photogrammetry are the main applications of computer vision in visual effects. But new techniques are making it into visual effects all the time.

Space carving techniques are starting to show up in visual effects. In the movie *Minority Report*, a 3D image of a person was projected in Tom Cruise's apartment. The camera roamed around the projection, and it needed to look like a true 3D display. The projected actor was filmed on a special greenscreen stage, surrounded by many synchronized video cameras. Each camera was calibrated and fixed. Since the actor was surrounded by greenscreen, it was relatively simple to determine what part of each image was actor and what was background. From this, a view volume was formed. The intersection of all the view volumes of all the cameras produced a convex hull of the actor. Projecting the images onto the resulting geometry produced a remarkably realistic model of the person. This is just the beginning for this technique. At recent conferences, similar techniques are being developed to extract amazingly detailed deforming models of humans with concavities and everything.

It is a terribly exciting time to be involved in computer vision and visual effects. Over the years, the advantages of robust computer vision tools have become evident to the artists and the management. Computer vision techniques allow the artists to do things that previously were not possible and to do them faster and more efficiently, which makes management happy.

Bibliography

[1] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, 1996.

[2] Y.-Y. Chuang, A. Agarwala, B. Curless, D. H. Salesin, and R. Szeliski. Video matting of complex scenes. *ACM Transactions on Graphics*, 21(3):243–248, July 2002.

[3] P. E. Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. *Siggraph98, Annual Conference Series*, pages 189–198, 1998.

[4] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. *Computer Graphics, Annual Conference Series*, pages 11–20, 1996.

[5] H. W. Jensen, S. R. Marschner, M. Levoy, and P. Hanrahan. A practical model for subsurface light transport. *Siggraph01, Annual Conference Series*, pages 511–518, 2001.

[6] J. Lewis. Fast normalized cross-correlation. *Vision Interface*, 1995.

[7] V. Masselus, P. Dutre, F. Anrys. The free-form light stage. *Proceedings of the 13th Eurographics Workshop on Rendering*, pages 247–256, June 2002.

[8] M. Pauly, R. Keiser, L. P. Kobbelt, and M. Gross. Shape modeling with point-sampled geometry. *ACM Transactions on Graphics*, 22(3):641–650, July 2003.

[9] H. Pfister, M. Zwicker, J. van Baar, and M. Gross. Surfels: Surface elements as rendering primitives. *Siggraph00, Annual Conference Series*, pages 335–342, 2000.

[10] R. Szeliski and J. Coughlin. Spline-based image registration. *Technical Report No. CRL-94-1*, Cambridge Research Lab., DEC, April 1994.

[11] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. *Siggraph99, Annual Conference Series*, pages 215–224, Los Angeles, 1999.

Chapter 8

CONTENT-BASED IMAGE RETRIEVAL: AN OVERVIEW

Theo Gevers

and Arnold W. M. Smeulders

In this chapter, we present an overview on the theory, techniques, and applications of content-based image retrieval. We choose patterns of use, image domains, and computation as the pivotal building blocks of our survey. A graphical overview of the content-based image retrieval scheme is given in Figure 8.1. Derived from this scheme, we follow the data as they flow through the computational process (see Figure 8.3), with the conventions indicated in Figure 8.2. In all of this chapter, we follow the review in [155] closely.

We focus on still images and leave video retrieval as a separate topic. Video retrieval could be considered a broader topic than image retrieval, as video is more than a set of isolated images. However, video retrieval could also be considered simpler than image retrieval, since, in addition to pictorial information, video contains supplementary information such as motion, spatial constraints, and time constraints (e.g., video discloses its objects more easily, as many points corresponding to one object move together and are spatially coherent in time). In still pictures, the user's narrative expression of intention is in image selection, object description, and composition. Video, in addition, has the linear timeline as an important information cue to assist the narrative structure.