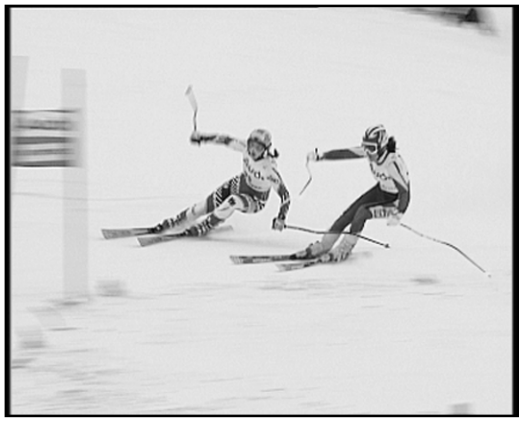
from *SIAM News*, Volume 35, Number 5

Skiing Ghosts

By James Case

Next to women's skating, alpine skiing is usually the toughest ticket at the Winter Olympics. Further evidence of skiing's popularity can be found in the TV ratings.

This year, in its broadcasts of the Salt Lake Olympics, NBC surprised and delighted fans with a remarkable viewing aid. Soon after each new name was posted atop the leader board in a particular race, those watching at home witnessed a rerun in which both past and present leaders appeared to negotiate the course at the same time, as if they had emerged simultaneously from a single starting gate. The actual starts were never shown, and the eventual winner was usually well ahead by the time the viewing audience joined the (virtual) race in progress. On occasion, however, the leader was overtaken. This happened quickly, when it happened at all, usually as the result of some specific mistake by the eventual loser. Such mistakes provide "color commentators" with welcome opportunities to chime in with error analysis.



Virtual races: A new technology shows fans of downhill skiing the race they most want to see—the two fastest racers in an event matched head-to-head, on the same course at the same time. Image courtesy of Dartfish.

The new technology enabled the network to show the race fans most wanted to see—the one in which the two fastest racers were matched head-to-head, on the same course at the same time. For obvious safety reasons, such races can only be virtual.

When one racer overtook another, their images became indistinct as they drew closer, merged, and finally separated with order interchanged. If there were instances in which the images merged, only to separate again in the original order, I don't recall them. As the trailing image approached the leader, both assumed a somewhat ghostly "see-through" appearance, which lasted until they were again comfortably separated. Technically inclined viewers wondered how the reruns were created so quick-ly, and why the proximate images appeared ghostly. A trip to the library turned up the necessary background information.

According to [1], modern digital television images are composed of pixels, arranged in rows (a.k.a. lines) and columns. The pixels are briefly illuminated by electrons shot from "electron guns," at voltages that vary (digitally, of course) from 0 (darkest) to 7 (brightest). In the days of black-and-white TV, a single (analog) electron gun sufficed. In the age of color, every pixel is subdivided into regions painted with red, blue, and green phosphors, each color being activated by a distinct beam of electrons shot from a separate electron gun. The instructions for illuminating a single pixel therefore constitute 3 bytes of information. Standard VHS broadcasts consist of 30 frames per second, with 720 pixels per row and 480 rows per frame. The raw signal thus delivers more than 30 megabytes of information per second, a figure reducible by a factor of at least thirty by standard data-compression techniques. HDTV signals, which few homes are equipped to receive, deliver roughly five times as much information.

Creating the Virtual Races

The process by which the exciting virtual races were created is protected by a U.S. patent—number 6,320,624—issued to Serge Ayer and Martin Vetterli (inventors) on November 20, 2001. The patent was promptly assigned to l'Ecole Polytechnique Fédérale in Lausanne, Switzerland, which reassigned it to Dartfish, a start-up in which Ayer and Vetterli are heavily involved. The patent enumerates 24 novel capabilities, more than half of which were utilized in the Salt Lake broadcasts. It also describes in some detail the multistage process by which the virtual races were created.

The first stage merely synchronizes the image sequences that record the contestants' individual runs, which typically took place several minutes—and several racers—apart. Ski racers are allowed to decide for themselves (within limits) when to pass through the electronic starting gate that activates the official timepiece. The world's most famous clockmakers bid for the privilege of furnishing this device, which stops only when an "electric eye" reports that the finish line has been crossed.

After the two image sequences have been synchronized, each is passed through a "signal splitter," so that exact copies can be processed simultaneously. One copy of each is then examined to determine the camera setting that produced each individual frame. These settings include the direction in which the camera is pointing—as specified by its angle of elevation and (angular) deviation from due north—along with the focal length of the lens and the (variable) aperture angle. These parameters are easily recovered from the images themselves, if the backgrounds contain suitably many unique features, such as trees and protruding rocks. In the absence of such features, it is better to use "instrumented cameras," which record these measurements with great precision, along with the action of interest. The patent covers both techniques. Because snowy hillsides tend to contain few unmistakable landmarks, the Olympic races were shot with two instrumented cameras per racecourse.

While these camera settings are being extracted from one copy of the split signal obtained from each image sequence, the other copy is undergoing background/ foreground separation, followed by "weight-mask" computation. In the present application, background/foreground separation yields one image sequence of the (rapidly moving) skier in the foreground, and another of the (slowly moving or stationary) ski slope behind. The slope is stationary unless the camera pans to follow the skier, as it must when he or she passes close to the camera. The two sequences are obtained with "motion-estimation" hardware, which uses dedicated VLSI circuitry of a sort now readily available.

Such hardware works [1] by moving the current frame relative to an $N \times N$-pixel image of the foreground object (in this case the skier) obtained from the preceding frame, in such a way that the object remains at all times within a designated "search window" in the current frame. The distortion of the $N \times N$ preceding image r by the current image c, for all displacement vectors (i,j) such that the (larger, mobile) search window contains the (smaller, fixed) preceding image, is measured by

$$D(i,j) = \sum_m \sum_n |r_{m,n} - c_{m+i,n+j}|. \quad (1)$$

Commonly used implementations move a (mobile) 31×31-pixel search window relative to a (fixed) 16×16-pixel object. Here, $|\cdot|$ measures the (scalar) distance between 3-byte, 3-component excitation vectors r and c, while $D(\cdot,\cdot)$ is a sum of such distances. If (i^*,j^*) minimizes $D(i,j)$ among the finitely many admissible pairs (i,j), then (i^*,j^*) constitutes a reasonable estimate of the skier's motion vector between the past and present frames.

Such computations are ideally suited to parallel implementation, with $D(i,j)$ computed simultaneously for each admissible (i,j). One standard implementation uses an $N \times N$-processor array with N data ports to determine (i^*,j^*) after only $(2p) \times (N + 2p - 1)$ computation cycles; the search window is specified by the requirement that $i,j \in [-p, p - 1]$. A cheaper implementation uses only two data ports to determine (i^*,j^*) after $(N + 2p - 1)^2 + 2 + \log_2 N$ such cycles. The patent stipulates only that the motion-detection proceedure should be robust. After the background/foreground separation is complete, the resulting images must be resized and reoriented for projection into the chosen focal plane, before being reassembled in any of several possible ways.

Custom Blends

It is possible, among other things, to project the images of two particular ski racers onto an image of the slope over which they actually raced, but recorded during the previous summer, when covered with grass instead of snow. The result would appear ridiculous, in part because each skier would be surrounded by a small rectangle of snow, which no amount of processing could blend convincingly into the grassy background. Even when superimposed on the appropriate background, the blending process can be complicated. If, for example, the sun changes position appreciably between one run and the next, or merely ducks behind a cloud, the light in the various image sequences is difficult to reconcile.

If the sun changes position appreciably between one run and the next, or merely ducks behind a cloud, the light in the various image sequences is difficult to reconcile.

In practice, the blending is accomplished by creating an image sequence s_{mn} that is a convex combination $w_{mn}f_{mn} + (1 - w_{mn})b_{mn}$ of the foreground sequence f_{mn} and a background sequence b_{mn}. The quantity w_{mn}, called the "weight-mask sequence," must be constructed with some care if the foreground sequence f_{mn}—which is defined only on a subset of the index pairs (m,n)—is to blend smoothly into the background. This is done by choosing $w_{mn} = 1$ near the center of the window containing the superimposed object (skier), while forcing w_{mn} to approach zero as the pixel location (m,n) approaches the boundary of that window, so that the boundaries of this "soft window" will be more or less invisible to the casual observer. Alternatively, two foreground sequences f_{mn} and g_{mn} can be mixed with a common background b_{mn}.

All this explains why the Olympic skiers' images appeared ghostly when one was overtaking another. As the superimposed images grew closer, the windows containing them began to overlap, causing their respective images to be mixed with the snowy background, giving each a shadowy appearance. When they actually occupied the same piece of real estate, each was about equally weighted, while the background received little or no weight, causing the two to meld into a single seemingly substantial (if indistinct) composite. A comparable sequence, showing two high-jumpers surmounting a common bar, is on display at www.dartfish.com.

Abundant Applications

When quick turnaround is not of the essence, the image separation and recombination can be accomplished with relatively low-tech components. For athletes who wish to compare their latest efforts to their "personal bests," as digitally recorded on videotape, an effective system can be constructed from a high-end laptop/PC and an off-the-shelf digital video camera. When the system was first demonstrated in 1998, it took twenty-four hours to produce a single virtual rerun of a race between two downhill skiers. At the 2002 Olympics, that time had been reduced to a single second!

The patent specifically mentions that the method can be used in private sports training, as well as public broadcasting, and might be applied to such industrial problems as car crash analysis. It also points out that audio and video sequences can be combined, so that a video could feature, for instance, actor A lip-synching soundtracks laid down by performer B. Three-dimensional stereo sequences (of the sort required for many medical purposes) can be combined as well; this would permit doctors and researchers to track a patient's recovery from surgery in unprecedented detail. The lip-synching application requires more than the mere scaling

of time, to synchronize the relevant signals, but can be accomplished by means of dynamic programming.

Finally, certain licensees have begun to explore the application of the new techniques to situations in which the advertisements visible in the background at major sporting events pertain to products sold in some but not all of the markets into which a particular broadcast will be sent. It is now possible to superimpose different ads on the affected billboards, allowing the same space to be sold more than once.

References

[1] V. Bhaskaran and K. Konstantinides, *Image and Video Compression Standards: Algorithms and Architectures*, 2nd ed., Kluwer Academic Publishers, Dordrecht, the Netherlands, 1997.

James Case writes from Baltimore, Maryland.