A Modular Visual Tracking System

Mike Wessler*
MIT Artificial Intelligence Laboratory
545 Technology Sq. NE43-803
Cambridge, MA 02139
wessler@ai.mit.edu

I am currently building an active visual tracking system for a real world robot. The hardware is being built at MIT under the supervision of Professors Rod Brooks and Lynn Andrea Stein, and is humanoid in form. The software is also humanoid: I am basing its organization on models of early vision in the human brain. Most of the software is still in the design phase; what I describe here is the part of the system that is already up and running.

The robot, named Cog, has roughly the same degrees of freedom in the waist, neck, arms and eyes as a human and is designed with similar proportions in mind. The eyes sport a simulated fovea – each eye consists of two cameras mounted in the same plate, one with a wide field of view, and one with a much narrower view. Both cameras produce 128 × 128 gray scale images. The narrow one is used for tracking, and can be used for object recognition, while the wider view will be used for motion detection and peripheral vision.

I have written a visual tracking system that will eventually be hooked up to the motors in Cog's neck and eyes. For now, tracking is simulated by moving a small 16 × 16 "attention window" around the larger image from one of the cameras. On startup, the system memorizes the 16 × 16 segment in the center of the full image as a "reference" image. For the rest of the run, the system moves the window around to maintain whatever was initially present within its view.

The tracking system runs as follows. Once a new frame is grabbed, the system makes a guess about where the new window location should be, based on its previous velocity. Next, a portion of the image slightly larger than the attention window is selected, and the derivative of this region is computed.[1] Finally, a simple correlation is performed between the memorized reference image and the nine 16 × 16 windows centered around the pixel position nearest the guess. The one with the best correlation becomes the new center of the attention window. The entire sequence of grab, process, and correlate runs at 15 Hz on one of the Motorola 68332 processors that make up Cog's "brain".

The system can track anything that accelerates less than one pixel (0.15 degrees for the narrow-angle camera) per frame per frame. This becomes a problem only for objects that jerk suddenly; the system does very well with objects like hands and faces moving at normal speeds. Furthermore, by examining the result of the best correlation, the tracker knows exactly when it has "lost" the image, and should request the coordinates of a new region to track.

The tracking routines form the heart of a much larger system that models Stephen Kosslyn and Olivier Koenig's view of early vision in human brains. Instead of tracking whatever happened to be in the center of the camera view on startup, the full architecture will have an attention shift module to direct the gaze from one object to another. This module receives inputs from other modules that detect motion, faces or other "popouts" in the field of view. An opposing forces model, described by Marcel Kinsbourne, implements the decision of when to switch attention: if one of the "watchers" detects a very strong stimulus, or when the tracking system has lost the image or has gotten "bored" with the image, the attentional system will prompt a tracking change.

Further up the line, information from the tracking window will be fed to a system that keeps a representation of location. This system gradually builds up a map of the world as attention is yanked from one object to another. Furthermore, by watching its hands and arms, Cog can learn to associate certain arm positions with certain locations in space, in effect learning hand-eye coordination.

Finally, because the image within the tracking window is relatively constant, a moving background can be segmented out simply by averaging one image with the next. This yields extremely useful data for an object recognition system, which can then put "labels" on the locations that the location system is recording. Information from the object recognition system can also be used to guide attention around the image, for example from a face to the eyes or mouth.

*This research is supported by an NSF Graduate Fellowship and by ARPA ONR contract N00014–91–J–4038. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

[1] Originally, I had taken two derivatives in the x and y directions, but it turns out that a single derivative runs twice as quickly, with very little decrease in reliability.