A Neural Network Based Head Tracking System

Deepanshu(B20ME025)
https://iitj.ac.in/

Indian Institute of Technology, Jodhpur
Mechanical Engineering

With the proliferation of inexpensive multimedia computers and peripheral equipment, video conferencing finally appears ready to enter the mainstream. But personal video conferencing systems typically use a stationary camera, tying the user to a fixed location much as a corded telephone tethers one to the telephone jack. A solution to this is a head tracking system.

In this project we have an inexpensive, video-based , motorized head tracking system. It uses real time graphical user inputs or an auxiliary infrared detector as supervisory signals to train the Convolutional Neural Netowrk. There is a problem of controlling the movement of the camera when the person is communicating . The camera movements in this video conferencing system closely resemble the movements of human eyes.

Hardware Implementation

The figure show the whole system named “Marvin”. Marvin’s eye consists of a small CCD camera with a 65 degree field of view that is attached to a motorized platform. Two RC servo motors are used to rapidly pan and tilt over a wide range of viewing angles. The system also includes two microphones or ears that give Marvin the ability to locate auditory cues. Integrating auditory information with visual inputs allows the system to find salient objects better than with either sound or video alone.

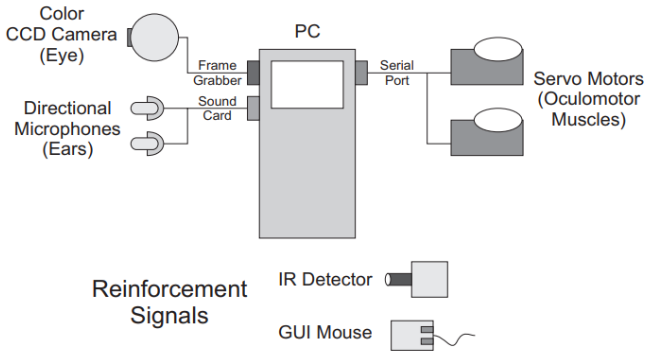

Figure 1: Schematic hardware of Marvin.

Neural Network Architecture

Marvin uses a convolutional neural network architecture to detect a head within its field of view. The video stream from the CCD camera is first digitized with a video capture board into a series of raw 120X160 RGB images. Each RGB color image is then converted into its YUV representation, and a difference (D) image is also computed as the absolute value of the difference from the preceding frame. Of the four resulting images, the Y component represents the luminance or grayscale information while the U and V channels contain the chromatic or color information. Motion information in the video stream is captured by the D image where moving objects appear highlighted. The four YUVD channels are then subsampled successively to yield representations at lower and lower resolutions. The resulting "image pyramids" allow the network to achieve recognition invariance across many different scales without having to train separate neural networks for each resolution.

Traning and Results

In typical batch learning applications of neural networks, the learning rate is set to be some small positive number. However in this case, it is desirable for Marvin to learn to track a head in a new environment as quickly as possible. Thus, rapid adaptation of the weights during even a single training example is needed. A natural way of doing this is to use a fairly large learning rate(=0.1).

Marvin is able to learn to track one of the authors as he moved around his office. The weights were first initialized to small random values, and Marvin was corrected in an online fashion using mouse inputs to look at the author’s head. After only a few seconds of training with a processing time loop of around 200 ms, the system was able to locate the head to within four pixels of accuracy, as determined by hand labelling the video data afterward.