End-to-End Learning of Vision-Based Obstacle Avoidance for Off-Road Robots

Yann LeCun
The Courant Institute, New York University
http://yann.lecun.com
Eric Cosatto, Jan Ben, Urs Muller, Beat Flepp
Net-Scale Technologies, Morganville, NJ
http://www.net-scale.com

What is End-to-End Learning?

- Partitioning a system into individual modules is a time-honored engineering practice.
- The partition is performed so that each module can be individually specified, designed, and tested. function
- However, there have been several recent demonstration that, with enough training data, learning algorithms are much better at building complex systems than humans: speech and handwriting recognition.
- Relying more on learning and less on hand-crafting can yield better results.

Why End-to-End Learning?

Never solve a more difficult problem than the one at hand

Partitioning a task into independent sub-tasks helps engineering the system, but those sub-tasks might be fundamentally more ill-posed than the large task.

- Example 1: you might think that shape recognition requires solving explicit models of how shapes can be distorted.
- Example 2: you might think that scene classification requires explicit object detection and recognition.
- Example 3: you might think that visual object detection and recognition requires explicit segmentation from the background, and explicit provisions for invariance to pose and lighting.
- Example 4: you might think that vision-based robot navigation requires constructing a 3D model of the world.

Maybe not....

Successful Examples of Visual End-to-End Learning

- handwriting recognition (a few years ago)
- face detection (this morning)
- object recognition (wednesday morning)

Vision-Based Robot Navigation

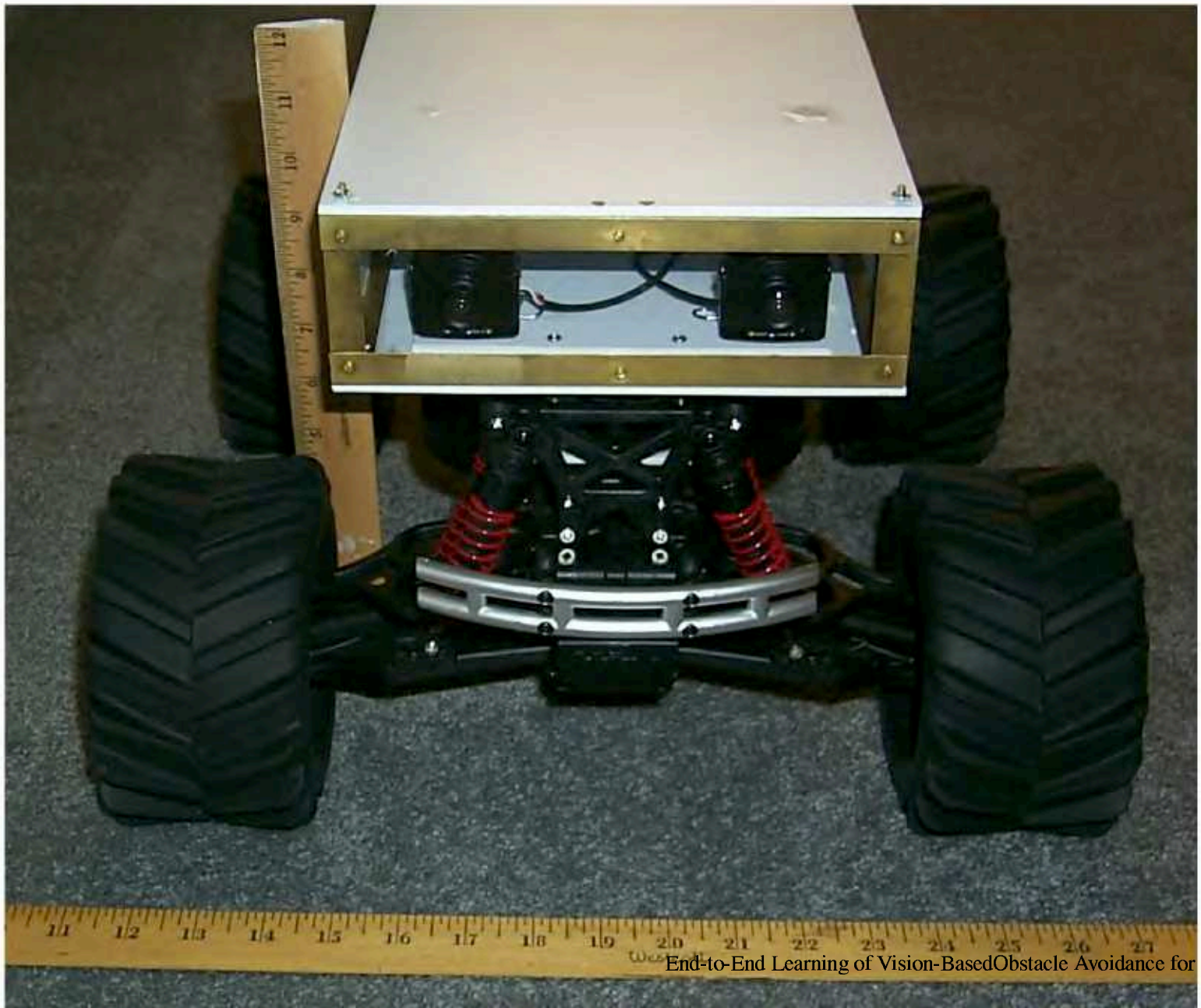
- Traditional robot navigation involves a wide variety of sensors (laser range finders, radar, sonar hardware stereo-vision systems....)
- Those sensors are individually unreliable.
- Recognizing obstacles from those sensor outputs is a largely unsolved problem (is it a bush or a rock?).
- Active sensors are bulky, power-hungry, and easily detectable.
- What about applying the end-to-end learning concept to vision-based robot navigation?

Learning Vision-Based Navigation for Mobile Robots

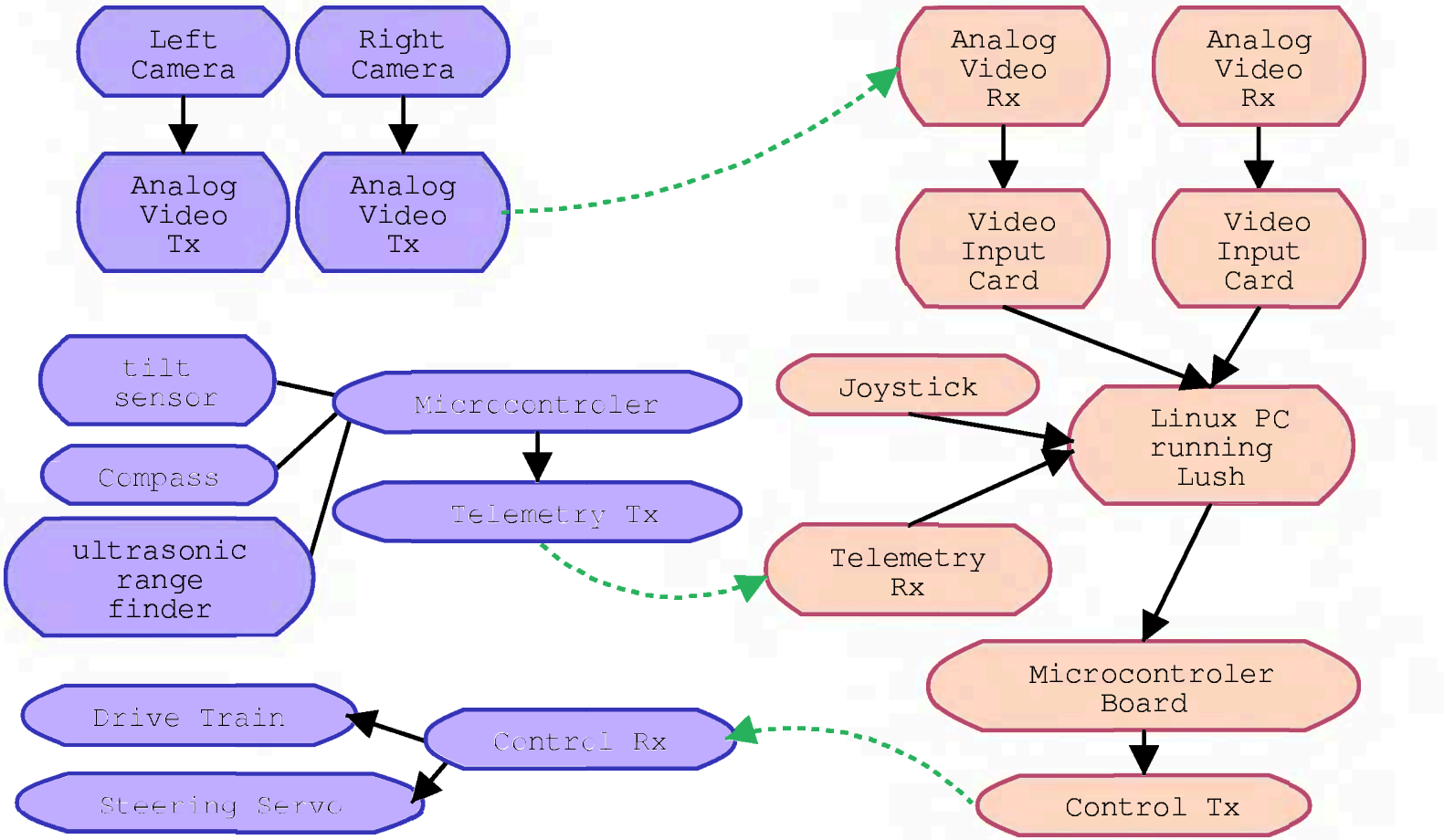
- We builts a truck-type mobile robot where the only sensors are two video cameras.
- We want to train the robot to map raw visual inputs to steering angles **directly**.
- The simpleest approach is supervised training.
- We collected several hours of “supervised” data: stereo video + human steering control

Robot

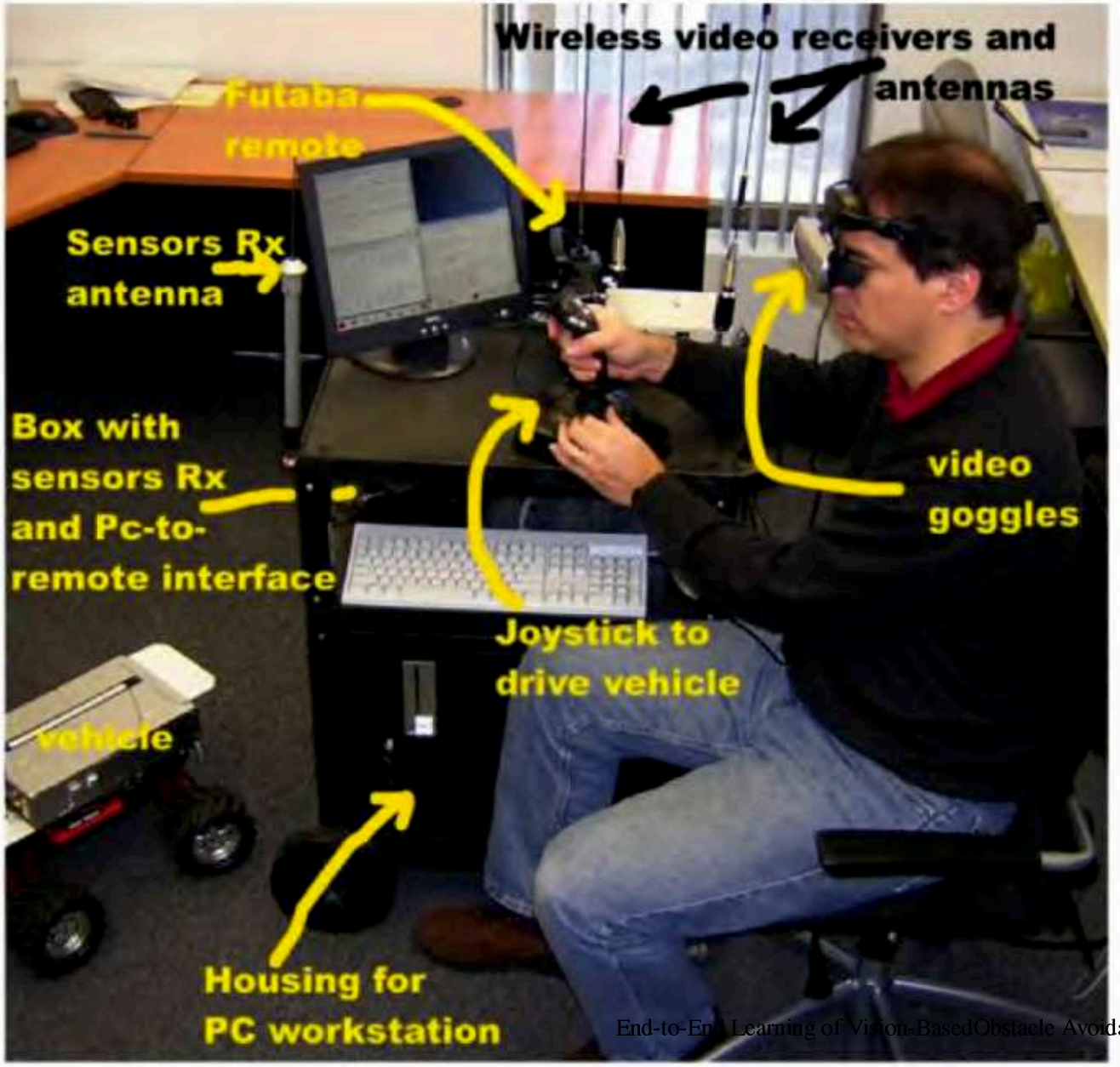
Robot



Block Diagram



Block Diagram



Video Frame Samples

Video Frame Samples

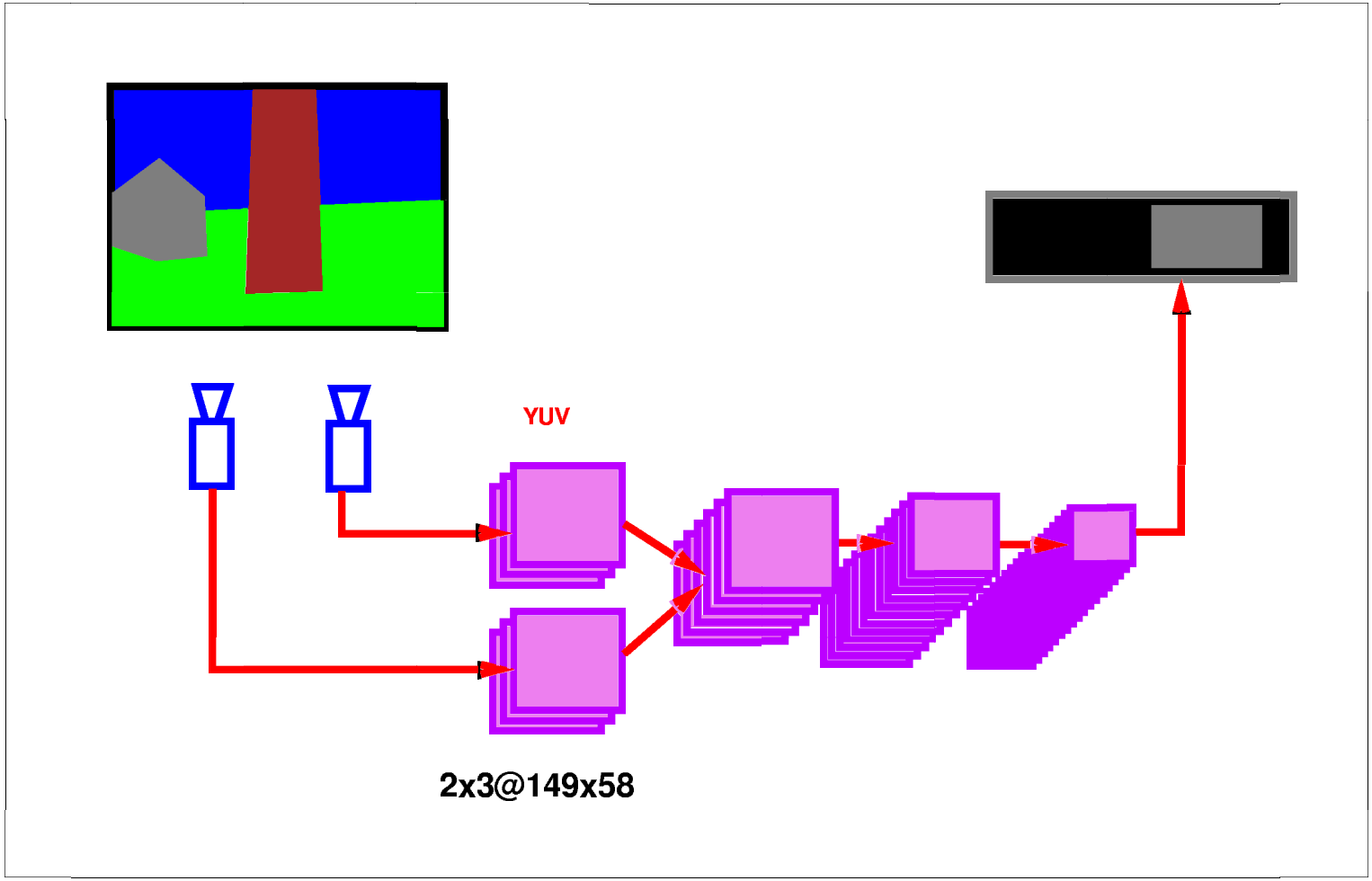
Video Frame Samples

Video Frame Samples

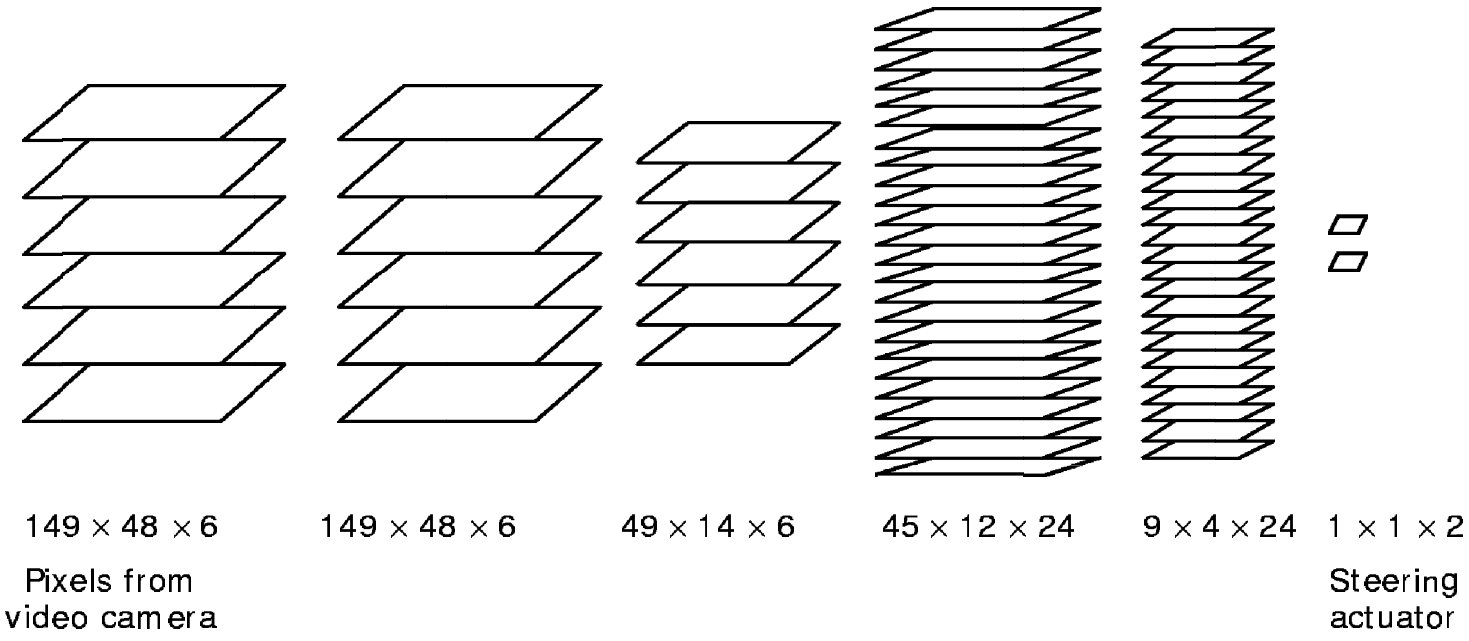
Convolutional Net

- Architecture: A convolutional net fed with raw video (left and right YUV images at 150x60 pixel resolution).
- The convolutional net has about 3 million connections. It runs at about 15 frames per second on a standard PC.

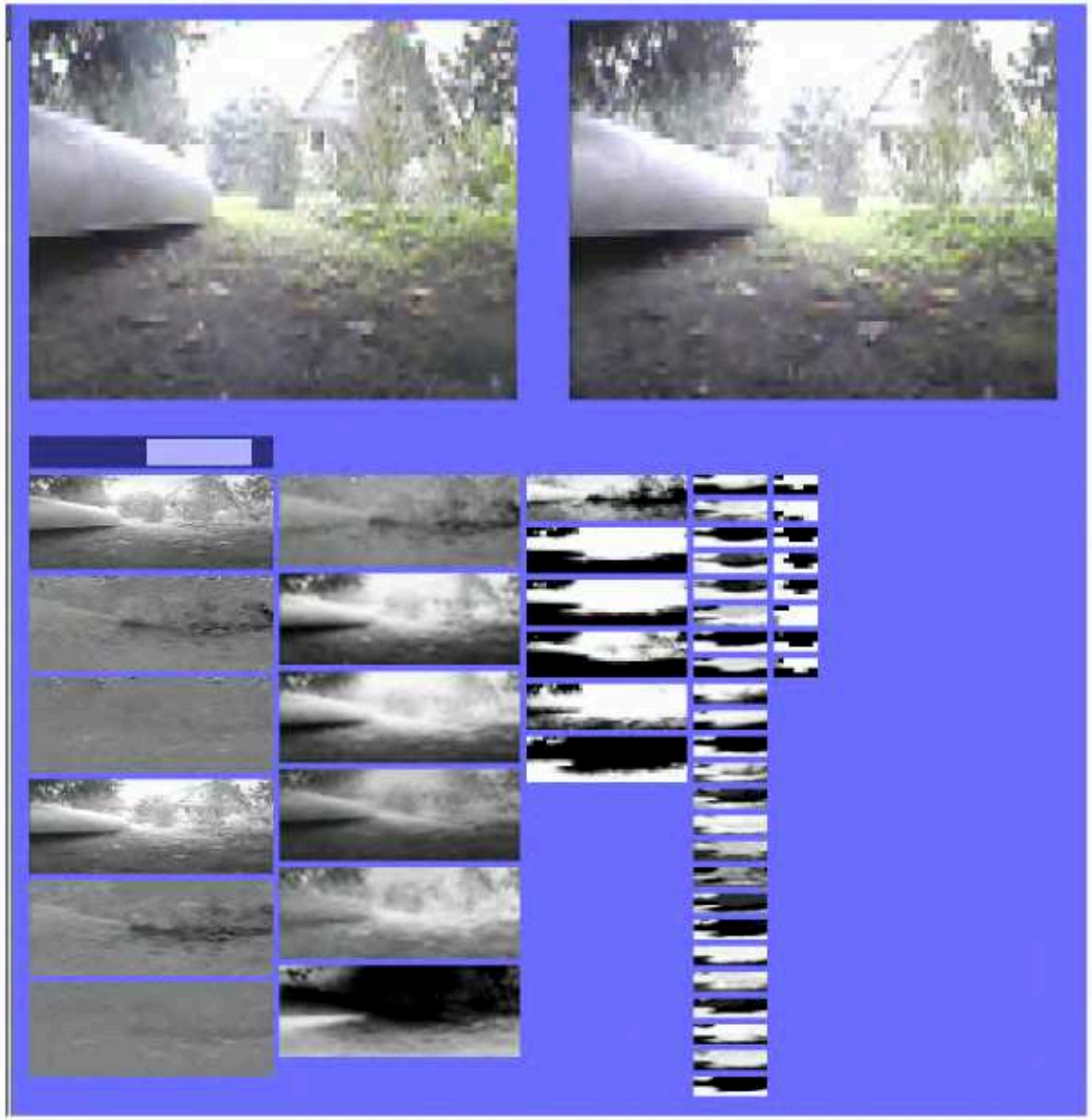
Convolutional Net



Convolutional Net

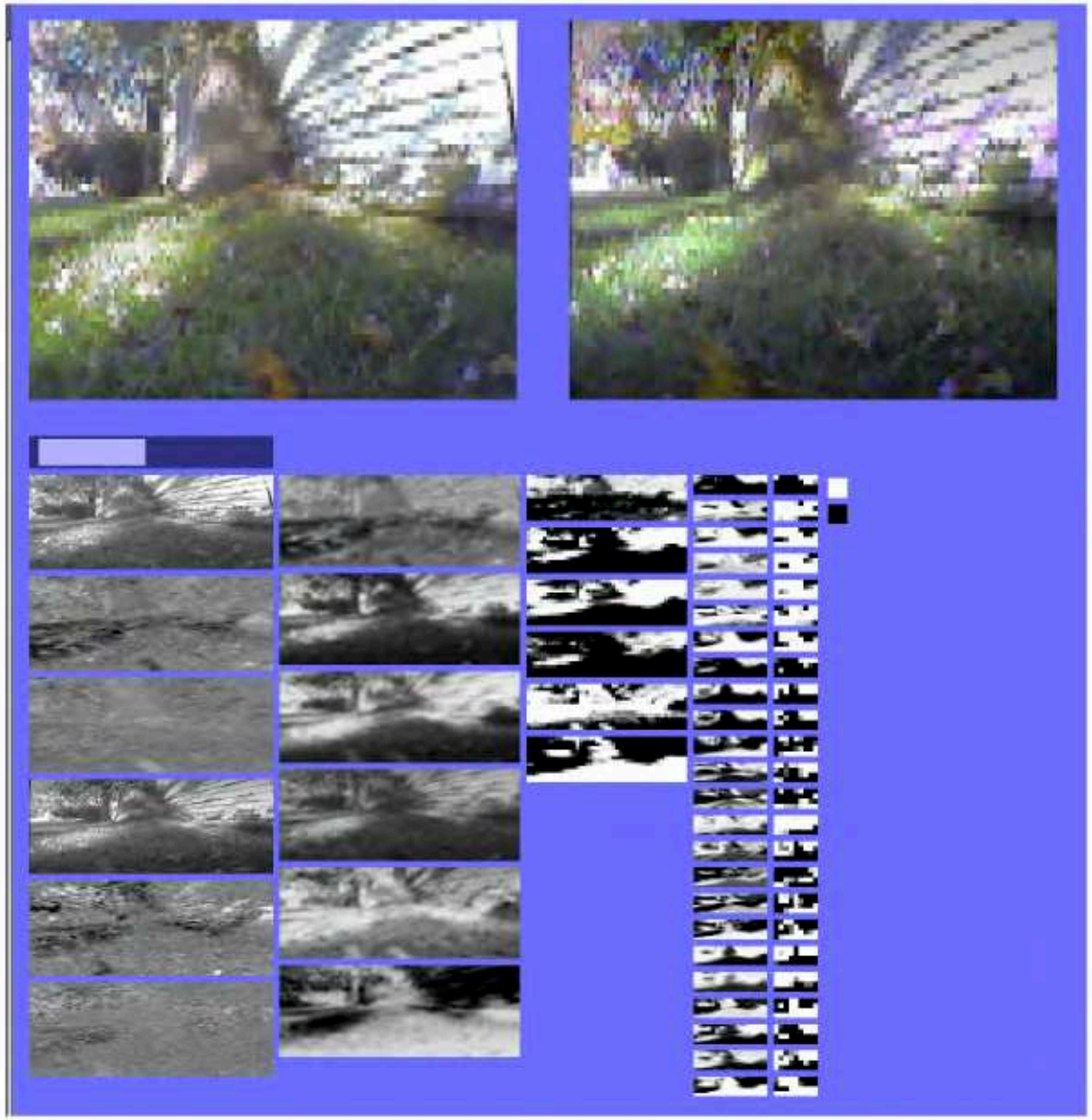


Convolutional net In Action



Convolutional net In Action

Convolutional net In Action



Shameless Plug:

Lush, the Programming Language for Research

- We have been developing **Lush**, a programming language for research.
- It uses a Lisp-like syntax, and has a huge library for numerical computation, statistics, graphics, image and video processing, machine learning....
- It has a very efficient compiler to C and allows users to freely mix Lush and C/C++ within a single function.
- Have a look at **http://lush.sf.net**.