Advanced Robot Control and Learning

Prof. Sami Haddadin

Part 2

Machine Learning in Robotics

Chapter 2: Overview of Robot Perception

Agenda

1. Motivation
2. Robot Vision vs Computer Vision
3. Components of Robot Perception
4. Limits and Potential of Deep Learning for Robotics

Special-Purpose Robot Automation

custom-built robots

human expert programming

special-purpose behaviors

General-Purpose Robot Autonomy

general-purpose robots

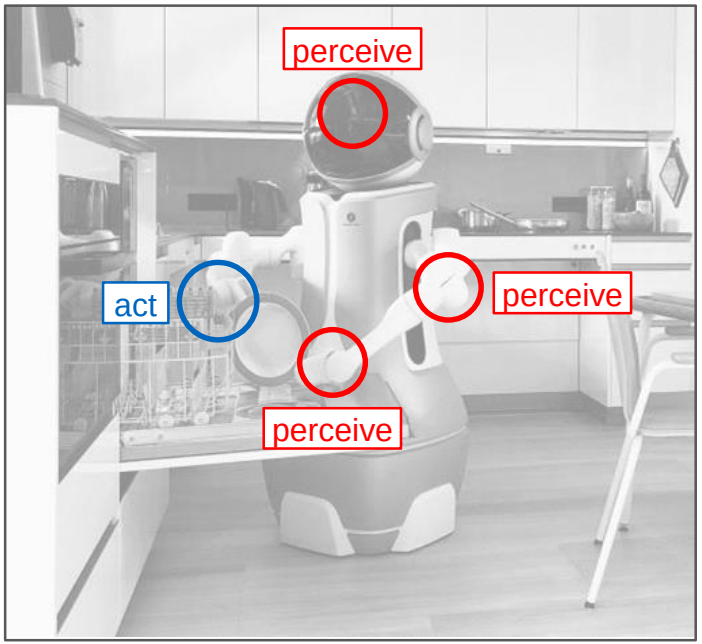
?

general-purpose behaviors

[Source: UT Austin CS391R]

A central challenge in Robot Autonomy is closing the perception-action loop.

- The real world is naturally multimodal
- Incomplete knowledge of the scene and objects
- Environment dynamics and other actors
- What reaction will our actions cause?



Robot Perception vs. Computer Vision

Robot perception is inherently different from computer vision. It is situated, embodied and active.

Situated: Robots are situated in the world. They do not deal with abstract descriptions, but with the here and now of the world directly influencing the behavior of the system.

Embodied: Robots have physical bodies and experience the world directly. Their actions are part of a dynamic with the world and have immediate feedback on their own sensation.

Active: Robots are active perceivers. It knows why it wishes to sense, and chooses what to perceive, and determines how, when and where to achieve that perception.

[Intelligence without Reason. Rodney Brooks (1991)]

Components of Robot Perception

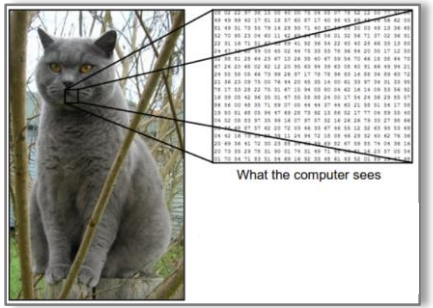
1. Modalities: different sensory modalities require different processing methods

2. Representation: building representations from sensor data for decision making and control

3. State estimation: using perception sensors for state estimation in robotics

4. Embodiment: active perception for embodied visual intelligence

Sensor Modalities

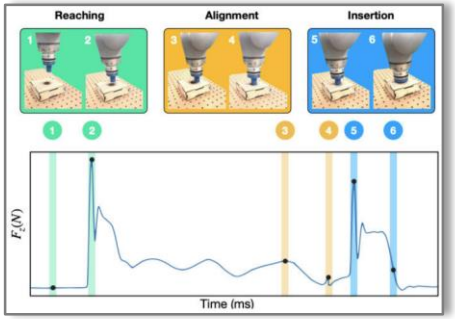


[Source: Stanford cs231n]

Pixels from RGB cameras

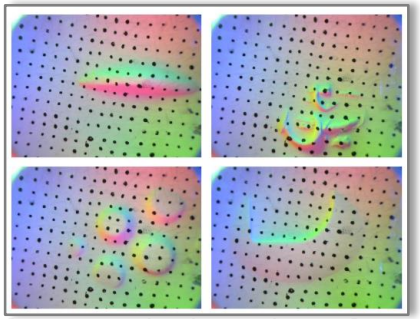
[Source: Qi et al. 2016]

Depth map/Point Cloud



[Source: Lee et al. 2018]

Time series (F/T sensing)



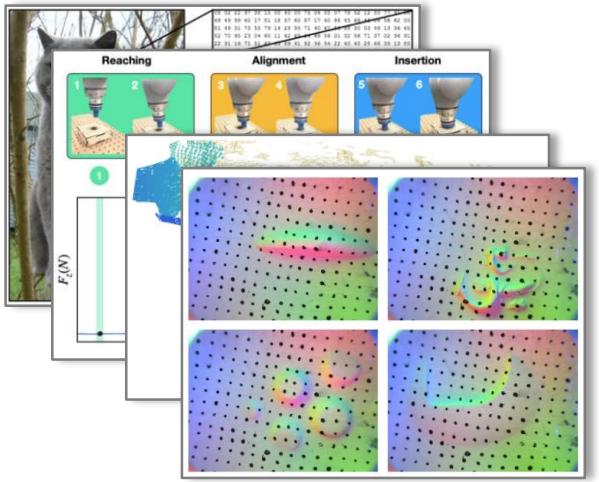
[Source: Calandra et al. 2018]

Tactile data

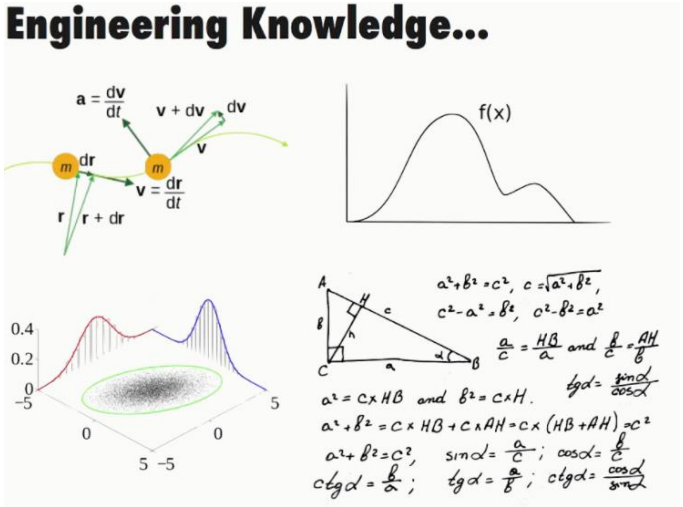
Representations

“Solving a problem simply means representing it so as to make the solution transparent.”

~ Herbert A. Simon, Sciences of the Artificial



Representation



[Source: Stanford CS331B]

Classical Representations

Using representations to solve problems is a very common pattern in engineering. Some examples:

- Laplace transform $F(s) = \int_0^\infty e^{-st} f(t) dt$ of the state-space representation of a dynamical system
- The Principle Component Analysis (PCA) finds a lower dimensional, latent representation of some data using the singular value decomposition $X = U \, \Sigma \, W^T$.
- Scale-Invariant Feature Transform (SIFT) features are another representation for pixel values of an image.

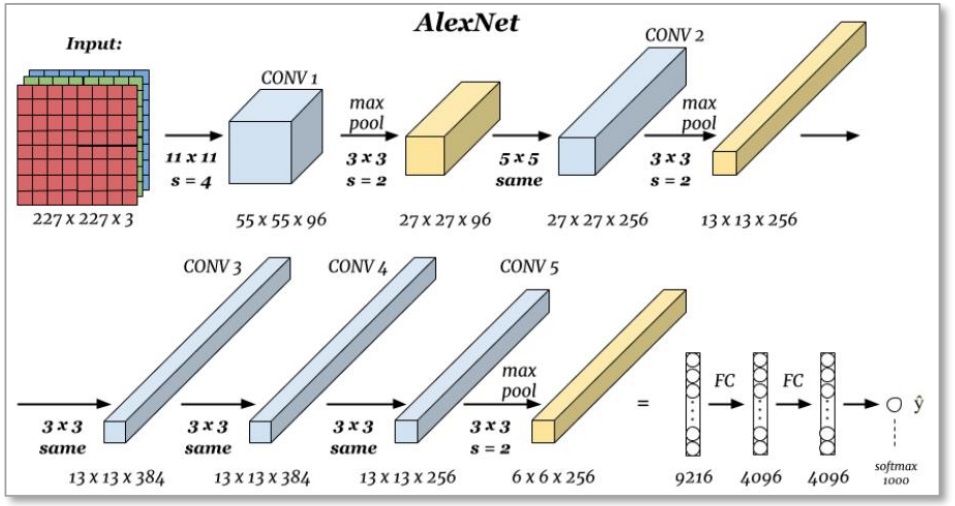
When are hand-crafted representations not enough?

Quick Review: Convolutional Neural Networks

Variants of the Convolutional Neural Network (CNN, or ConvNet) are the most widely used form for processing of visual information.

They consist of basic building blocks:

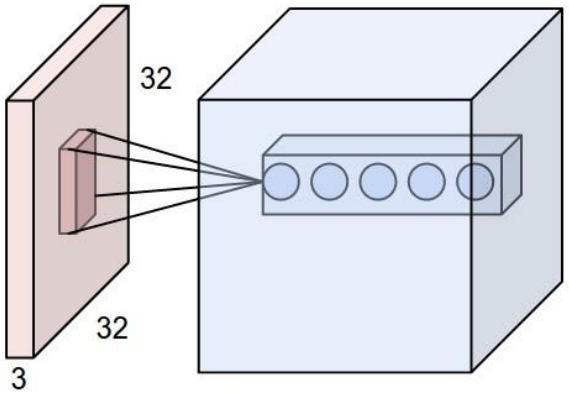
- an input,
- convolutional layers,
- activation functions,
- pooling layers, and
- fully connected layers.

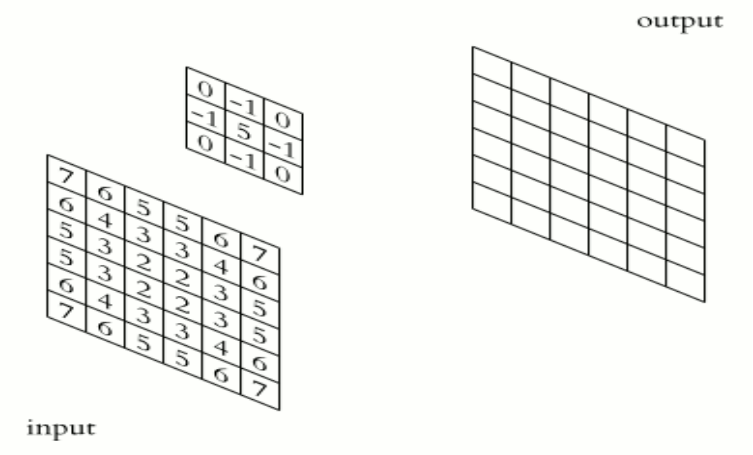


[Source: Stanford cs231n]

Quick Review: Convolutional Layers

Conv. layer will compute the output of neurons that are connected to local regions in the input, each computing a dot product between their weights and a small region they are connected to in the input volume.

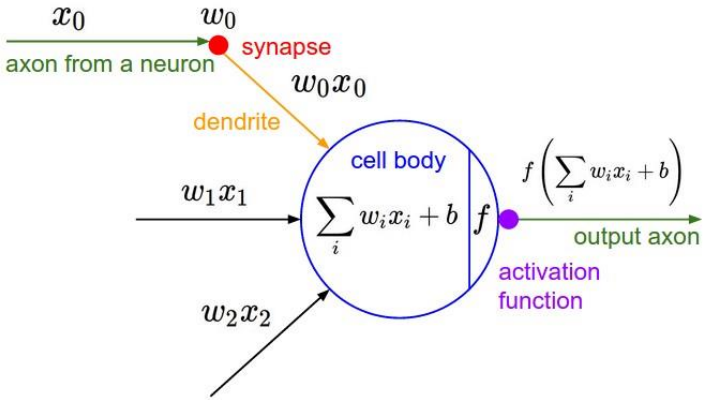




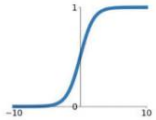
[Source: Stanford cs231n]

Quick Review: Activation Functions

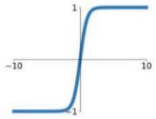
A non-linear activation function is applied elementwise to the result of the convolution. This leaves the size of the volume unchanged but enables the approximation of non-linear functions (e.g. XOR)



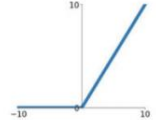
Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$



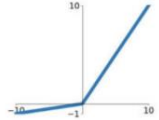
tanh
$\tanh(x)$



ReLU
$\max(0, x)$

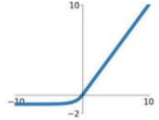


Leaky ReLU
$\max(0.1x, x)$



Maxout
$\max(w_1^T x + b_1, w_2^T x + b_2)$

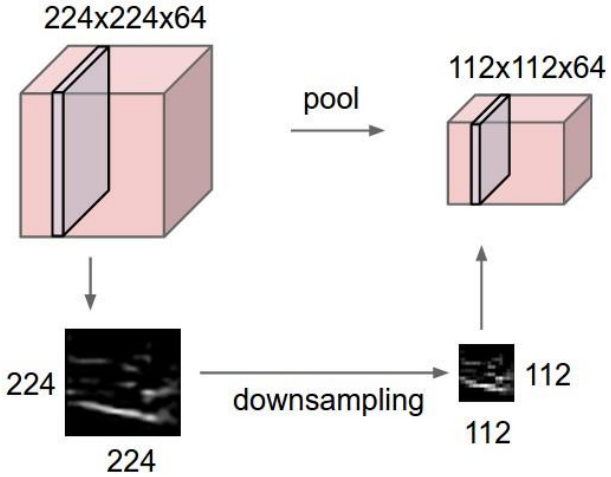
ELU
$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$

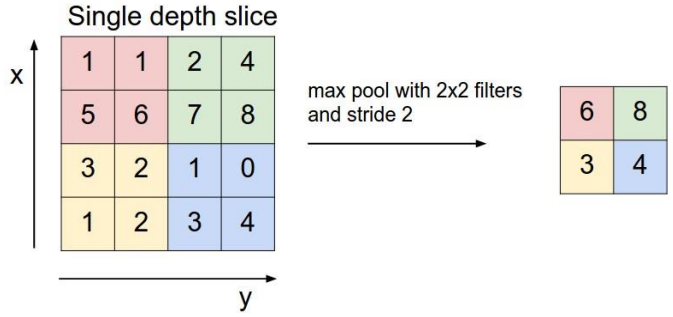


[Source: Stanford cs231n]

Quick Review: Pooling layers

Pooling layer will perform a downsampling operation along the spatial dimensions (width, height), resulting in smaller volume. Its function is to progressively reduce the spatial size of the representation to reduce the number of parameters and computation in the network, and hence to also control overfitting.

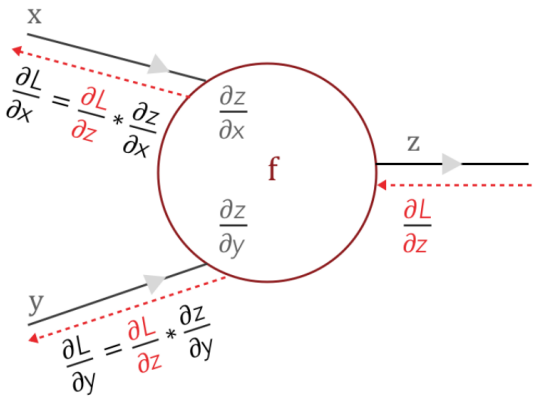




Note: Many architectures today don‘t contain pooling layers anymore but convolutional layers with a larger stride.

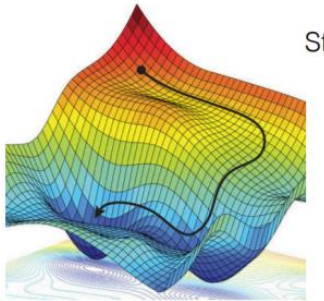
Quick Review: Optimization

Backpropagation is a way of computing gradients of expressions through recursive application of the chain rule. These gradients, given some loss function, allow updating the training parameters.



$\frac{\partial z}{\partial x}$ & $\frac{\partial z}{\partial y}$ *are local gradients*

$\frac{\partial L}{\partial z}$ *is the loss from the previous layer which has to be backpropagated to other layers*



Stochastic Gradient Descent (SGD)

learning rate

$$\theta = \theta - \eta \nabla_\theta J(\theta; x^{(i)}; y^{(i)})$$

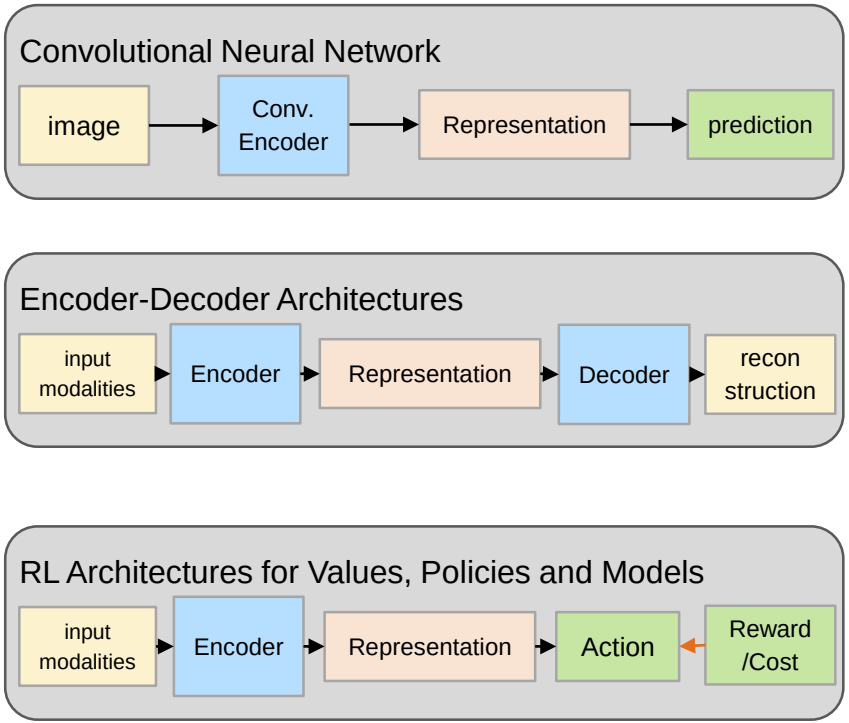
weights input label

Representation Learning

Although representation learning is an established field in computer vision, its application to robotics is still limited.

Applications and potential:

- Learn representations that fuse different modalities together
- Combine nature (physics-based inductive biases) with nurture (interaction-based learned models)



State estimation

A central task of perception is to estimate the state of the environment for control and decision making.

For linear systems and gaussian noise, a Bayes filter (e.g. Kalman filter) is the optimal state estimator.

For non-linear systems, there are non-linear Bayes filters that approximate the state belief.

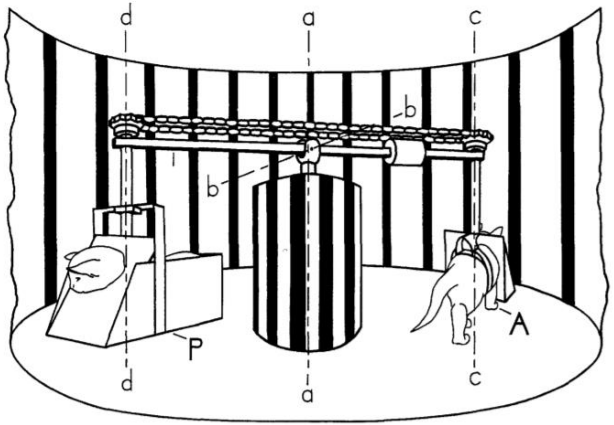
[Source: Toyota Research Institute]

Embodiment

Perception is facilitated by an embodied agent through actively exploring in the physical world.

It enables robots to perform active vision and targeted manipulation of its environment to further improve its perceptual and dynamics models.

However, the unification of physics and learning approaches to achieve such feats remains an open and challenging problem.



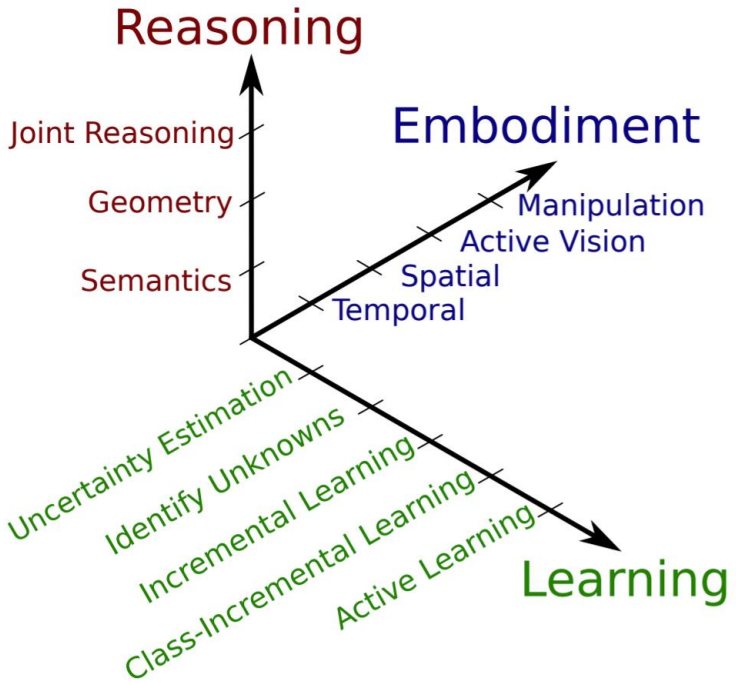
Kitten Carousel (Held and Hein, 1963)

Challenges for Deep Learning in Robotic Vision*

These fundamental differences motivate challenges along three orthogonal axes:

- learning,
- embodiment, and
- reasoning.

Individual challenges are positioned according to their increasing complexity.



*[Sünderhauf, N., Brock, O., Scheirer, W., Hadsell, R., Fox, D., Leitner, J., Upcroft, B., Abbeel, P., Burgard, W., Milford, M. and Corke, P. The International Journal of Robotics Research (2018)]

Learning Challenges for Robot Vision

0. Closed-Set Assumptions: The system can detect and classify objects of classes known during training. It provides uncalibrated confidence scores that are proportional to the system's belief of the label probabilities. State of the art computer vision methods, e.g. objects detectors, are at this level.

1. Uncertainty Estimation: The system can correctly estimate its uncertainty and returns calibrated confidence scores that can be used as probabilities in a Bayesian data fusion framework. Current work on Bayesian Deep Learning falls into this category

2. Identify Unknowns: In an open-set scenario, the robot can reliably identify instances of unknown classes and is not fooled by out of distribution data

Learning Challenges for Robot Vision

3. Incremental Learning: The system can learn off new instances of known classes to address domain adaptation or label shift. It requires the user to select these new training samples.

4. Class-Incremental Learning: The system can learn new classes, preferably using low-shot or one-shot learning techniques, without catastrophic forgetting. The system requires the user to provide these new training samples along with correct class labels.

5. Active Learning: The system is able to select the most informative samples for incremental learning on its own in a data-efficient way, e.g., by utilizing its estimated uncertainty in a prediction. It can ask the user to provide labels.

Embodiment Challenges for Robot Vision

0. None: The system has no understanding of any form of embodiment and treats every image as an independent from previously seen images.

1. Temporal Embodiment: The system learned that it is temporally embedded and consecutive are strongly correlated. The system can accumulate evidence over time to improve its predictions. Appearance changes over time can be coped with.

2. Spatial Embodiment: The system can exploit aspects of spatial coherency and incorporate views of objects taken from different viewpoints to improve its perception, while handling occlusions.

Embodiment Challenges for Robot Vision

3. Active Vision: The system has learned to actively control the camera movements in the world, for example it can move the camera to a better viewpoint to improve its perception confidence or better deal with occlusions.

4. Active Manipulation: As an extension of active vision, the system can manipulate the scene to aid perception. For example, it can move an occluding object to gain information about object hidden underneath.

Reasoning Challenges for Robot Vision

0. None: The system does not perform any sophisticated reasoning, e.g., it treats every detected object as independent from other objects or the overall scene. Estimates of semantics and geometry are treated as independent.

1. Object and Scene Semantics: The system can exploit prior semantic knowledge to improve its performance. It can utilize priors about which objects are more likely to occur together in a scene, or how objects and overall scene type are correlated.

2. Object and Scene Geometry: The system learned to reason about the geometry and shape of individual objects, and about the general scene geometry, such as absolute and relative object pose, support surfaces, and object continuity under occlusions and in clutter.

3. Joint Reasoning: The system jointly reasons about semantics and geometry in a tighly coupled way, allowing semantics and geometry to co-inform each other

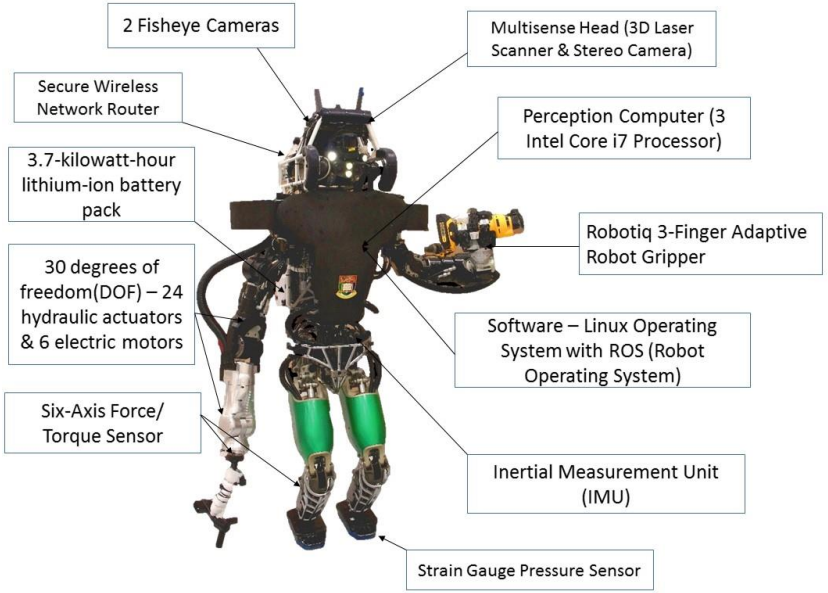
Archive

Components of Robot Perception

1. Modalities: different sensory modalities require different processing methods

2. Representations: learning representations with physical priors

3. Tasks:
4. Embodiment



[Source: **HKU Advanced Robotics Laboratory]**