Session 3

HRI Technology

Intelligent connection from perception to action

Introduction

“Intelligent connection of perception to action”

Michael Brady
Artificial intelligence and robotics
1985

Introduction

“Intelligent connection of perception to action”

Michael Brady
Artificial intelligence and robotics
1985

Introduction

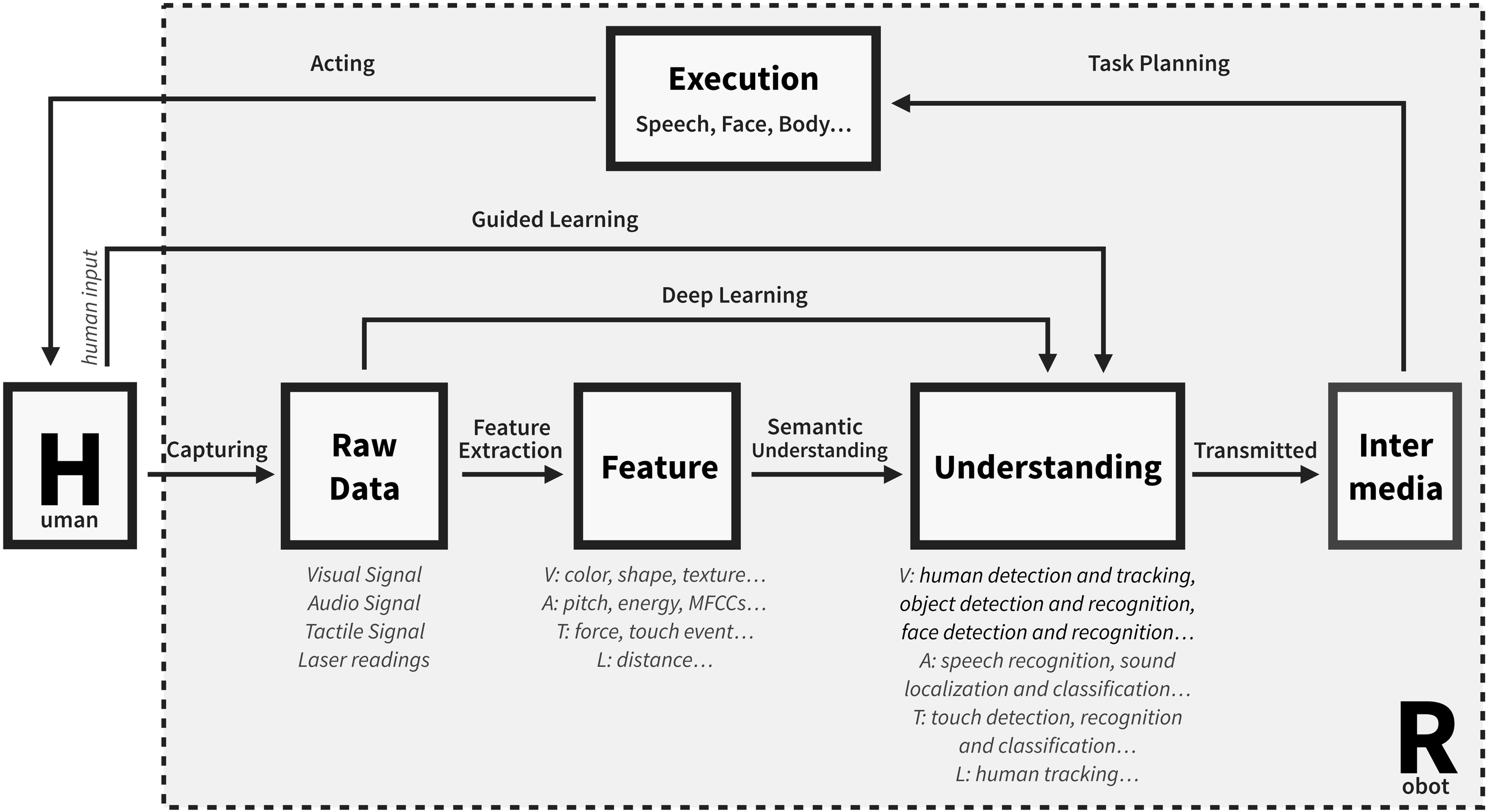
“Intelligent connection of perception to action”

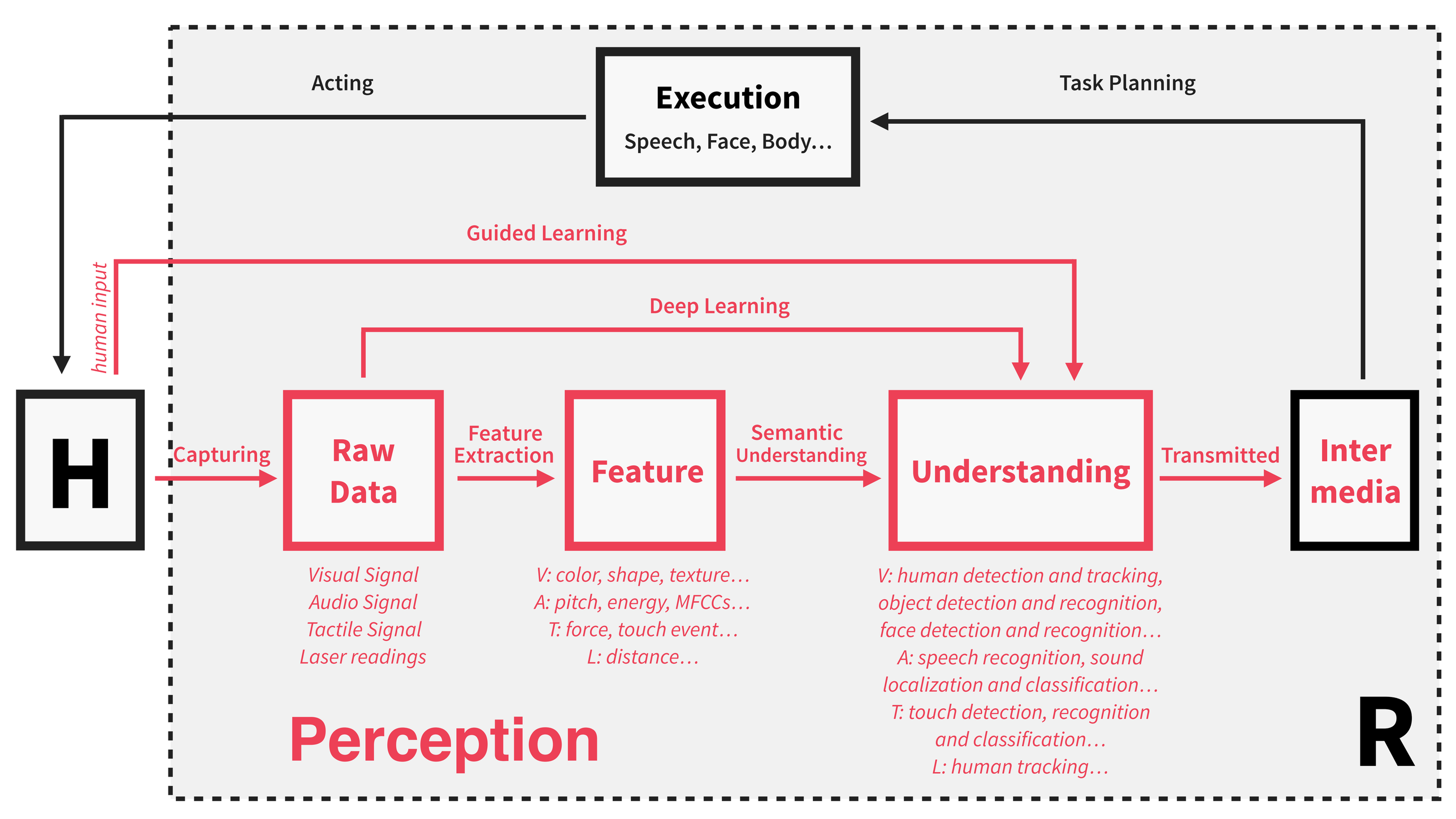
Michael Brady
Artificial intelligence and robotics
1985

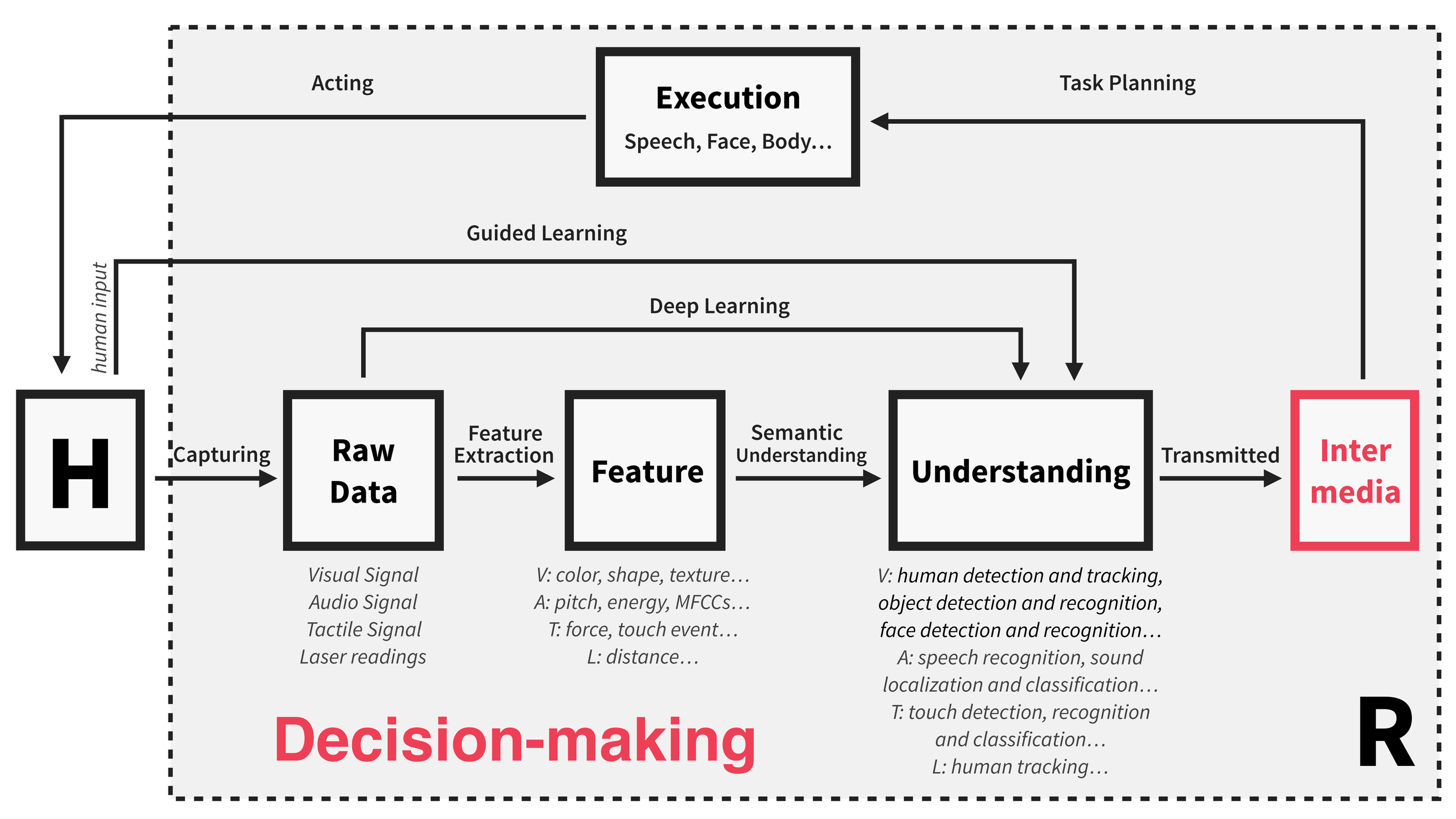
Introduction

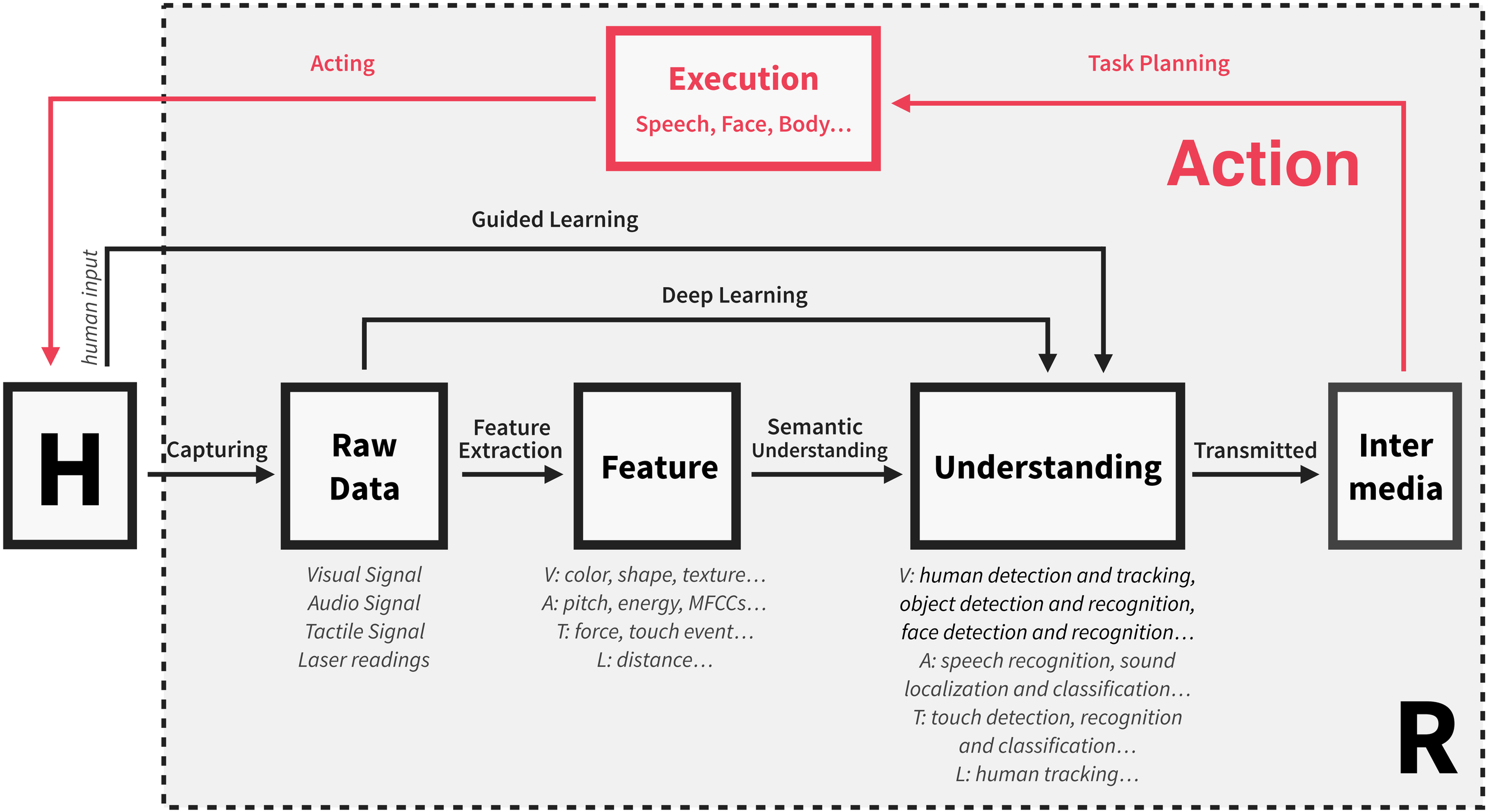
“Intelligent connection of perception to action”

Michael Brady
Artificial intelligence and robotics
1985


Acting
Execution
Speech, Face, Body…
Task Planning
Guided Learning
human input
Deep Learning
Human
Capturing
Raw Data
Feature Extraction
Feature
Semantic Understanding
Understanding
Transmitted
Inter media
Visual Signal
Audio Signal
Tactile Signal
Laser readings
V: color, shape, texture…
A: pitch, energy, MFCCs…
T: force, touch event…
L: distance…
V: human detection and tracking, object detection and recognition, face detection and recognition…
A: speech recognition, sound localization and classification…
T: touch detection, recognition and classification…
L: human tracking…
Robot


Acting
Execution
Speech, Face, Body…
Task Planning
Guided Learning
human input
Deep Learning
H
Capturing
Raw Data
Feature Extraction
Feature
Semantic Understanding
Understanding
Transmitted
Inter media
Visual Signal
Audio Signal
Tactile Signal
Laser readings
V: color, shape, texture…
A: pitch, energy, MFCCs…
T: force, touch event…
L: distance…
V: human detection and tracking, object detection and recognition, face detection and recognition…
A: speech recognition, sound localization and classification…
T: touch detection, recognition and classification…
L: human tracking…
Perception
R


Acting
Execution
Speech, Face, Body…
Task Planning
Guided Learning
human input
Deep Learning
H
Capturing
Raw Data
Feature Extraction
Feature
Semantic Understanding
Understanding
Transmitted
Inter media
Visual Signal
Audio Signal
Tactile Signal
Laser readings
V: color, shape, texture…
A: pitch, energy, MFCCs…
T: force, touch event…
L: distance…
V: human detection and tracking, object detection and recognition, face detection and recognition…
A: speech recognition, sound localization and classification…
T: touch detection, recognition and classification…
L: human tracking…
Decision-making
R


Acting
Execution
Speech, Face, Body…
Task Planning
Action
Guided Learning
human input
Deep Learning
H
Capturing
Raw Data
Feature Extraction
Feature
Semantic Understanding
Understanding
Transmitted
Inter media
Visual Signal
Audio Signal
Tactile Signal
Laser readings
V: color, shape, texture…
A: pitch, energy, MFCCs…
T: force, touch event…
L: distance…
V: human detection and tracking, object detection and recognition, face detection and recognition…
A: speech recognition, sound localization and classification…
T: touch detection, recognition and classification…
L: human tracking…
R

Perception

Signals → **Methods**

Visual-based

Audio-based

Tactile-based

Laser sensors

Data Capturing

Feature Extraction

Semantic Understanding

Perception

Signals	→	Methods
Visual-based		Data Capturing
Audio-based		Feature Extraction
Tactile-based		Semantic Understanding
Laser sensors		

tip: most information (~75 %) received everyday for human beings are visual signals

Visual-based -> Data Capturing

Sensors

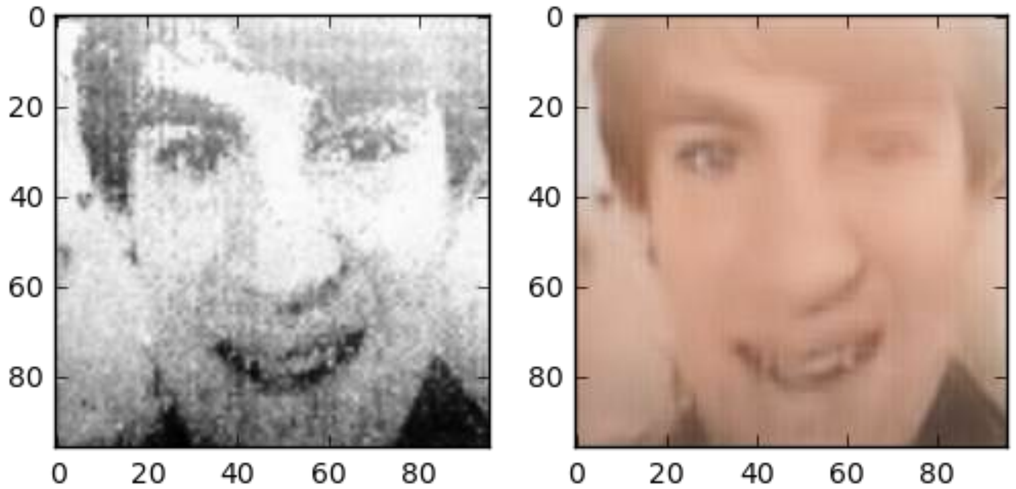
2D camera
3D camera

Visual-based -> Data Capturing

Sensors

2D camera

Grayscale Camera
RGB Color Camera



Visual-based -> Data Capturing

Sensors

3D camera

Depth Camera
- Kinect
- Intel RealSense SR300

Stereo Camera
- Intel RealSense F200
- Bumblebee® Stereo Vision camera

Depth Image

The value of each pixel in depth image refers to the distance between the object and camera

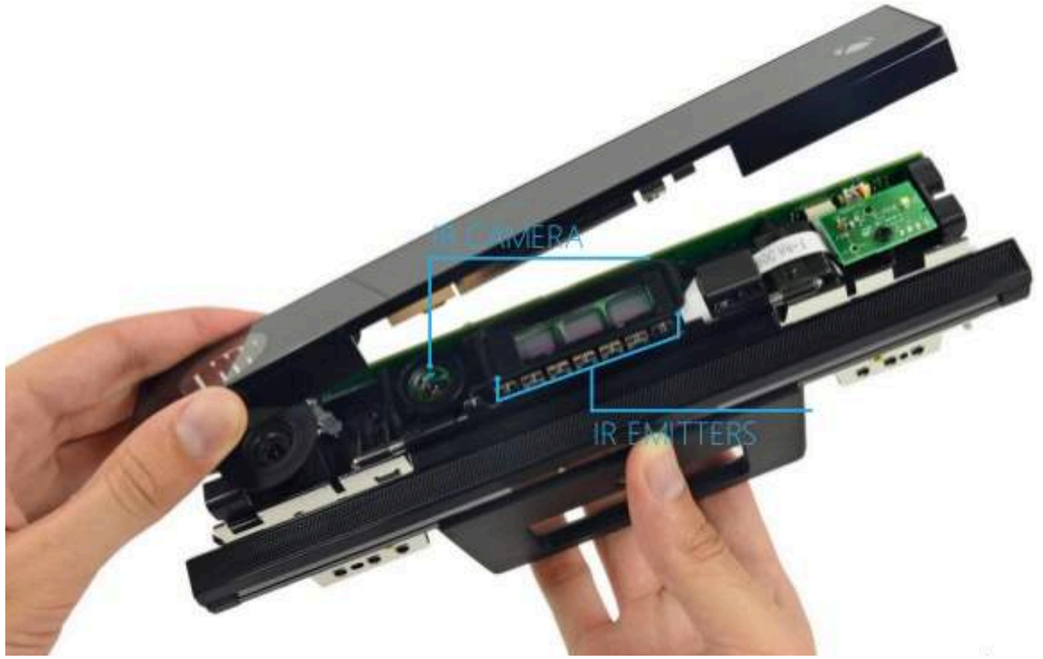
Visual-based -> Data Capturing

Sensors

3D camera

Depth Camera
- Kinect
- Intel RealSense SR300

Stereo Camera
- Intel RealSense F200
- Bumblebee® Stereo Vision camera



Kinect V2 for Windows

IR Emitter/Projector and IR Camera

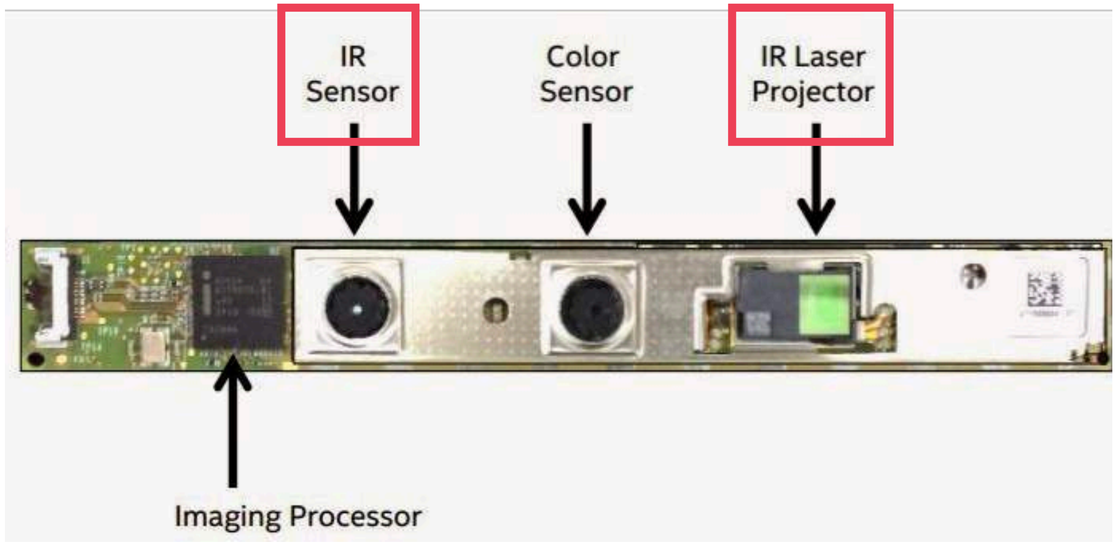
Visual-based -> Data Capturing

Sensors

3D camera

Depth Camera
- Kinect
- Intel RealSense SR300

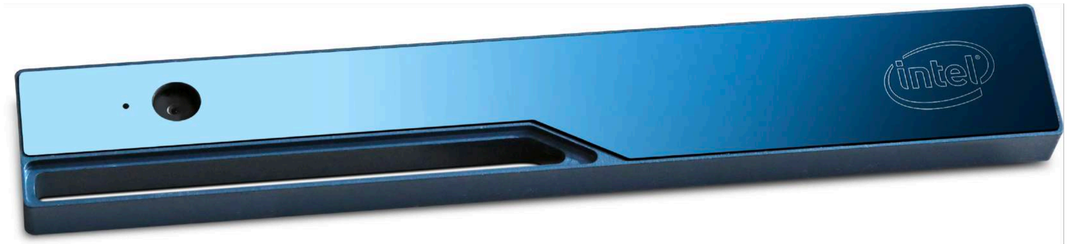
Stereo Camera
- Intel RealSense F200
- Bumblebee® Stereo Vision camera



Visual-based -> Data Capturing

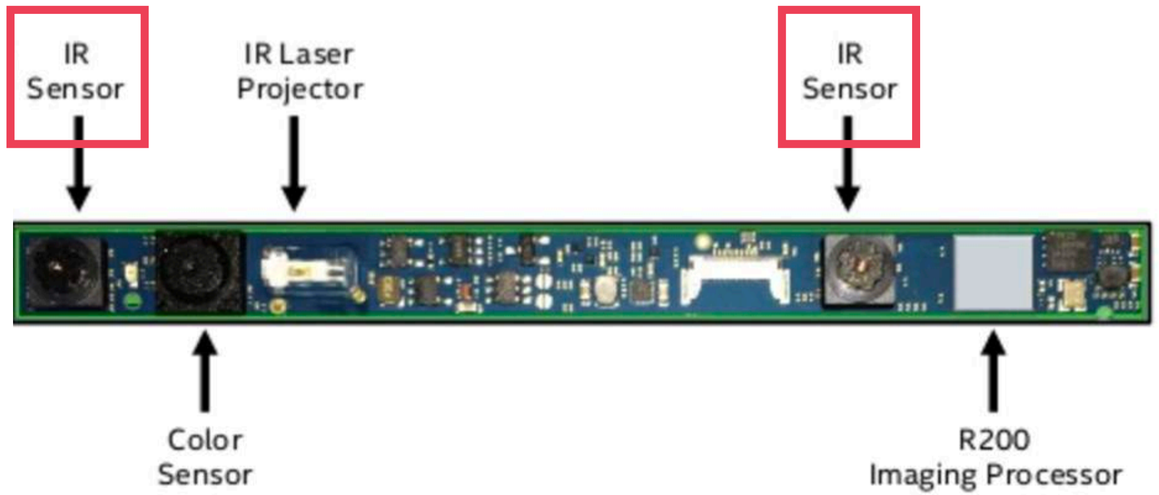
Sensors

3D camera



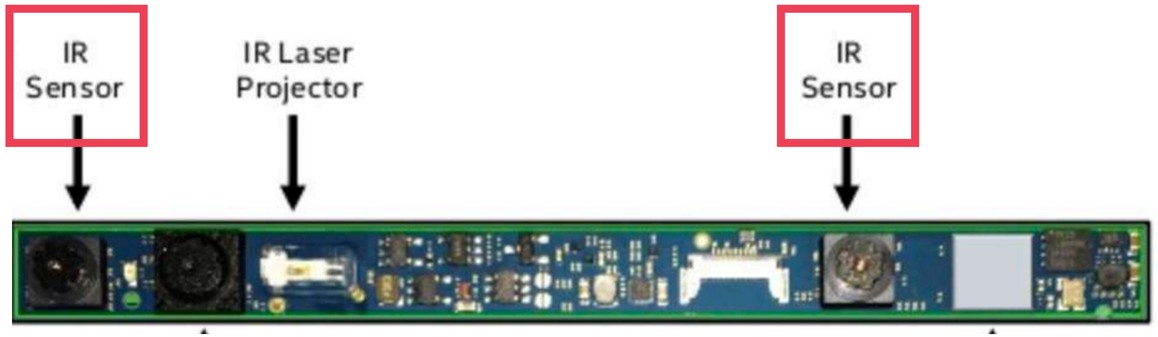
Depth Camera
- Kinect
- Intel RealSense SR300

Stereo Camera
- Intel RealSense F200
- Bumblebee® Stereo Vision camera



Visual-based -> Data Capturing

Sensors



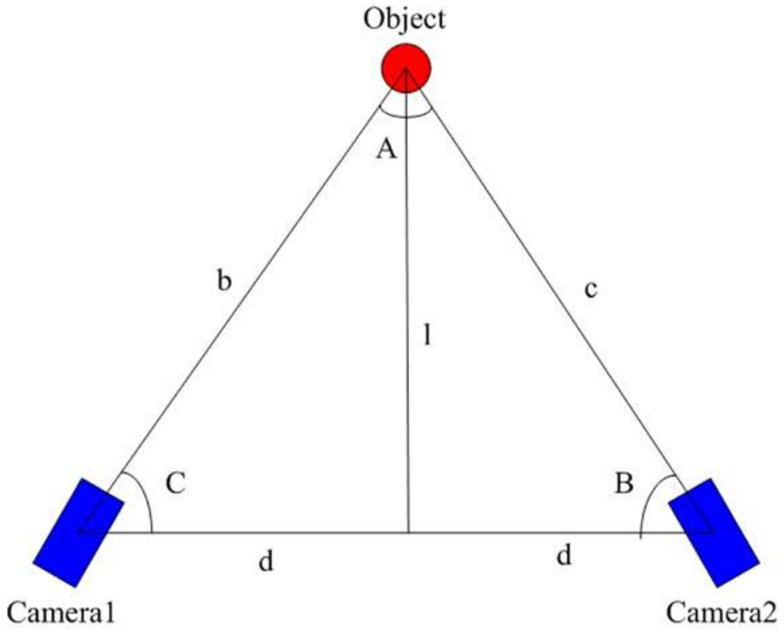
3D camera

Depth Camera

- Kinect
- Intel RealSense SR300

Stereo Camera

- Intel RealSense F200
- Bumblebee® Stereo Vision camera



Visual-based -> Data Capturing

Sensors

3D camera

Depth Camera
- Kinect
- Intel RealSense SR300

Stereo Camera
- Intel RealSense F200
- Bumblebee® Stereo Vision camera

Visual-based -> Feature Extraction

Feature

2D-based

- Color
- Intensity
- Texture
- Shape
- Motion
- Skelton

3D-based

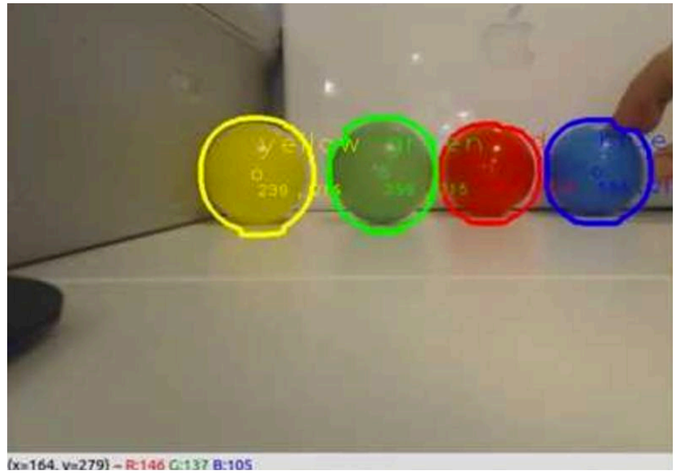
- RGB-D signals
- Stereo-Based Depth Signals
- Skeleton
- 3D Point Clouds

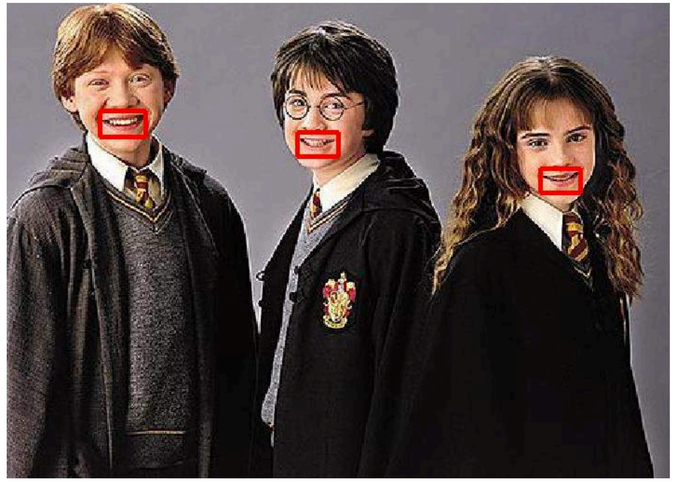
Visual-based -> Feature Extraction

2D-based

Color

- an effective feature to represent **objects**
- Extracting color **saliency** features to detect object and face
- easily affected by changing illumination



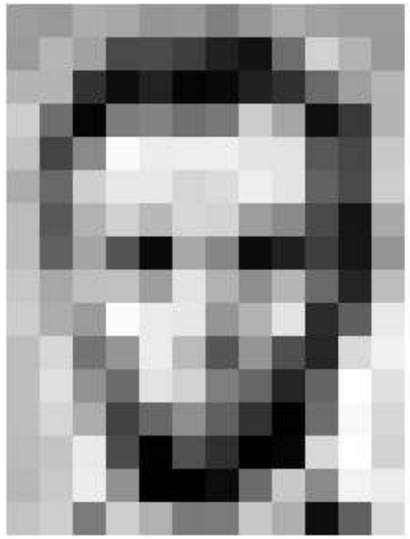


Visual-based -> Feature Extraction

2D-based

Intensity

- the gray-scale value of each pixel in an image
- intensity can be employed for feature representation
- extract **conspicuous regions**



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Visual-based -> Feature Extraction

2D-based

Visual Texture

- Visual texture is an important property for visual signals, and different objects usually demonstrate different texture characteristics
- texture representation methods (LBP, ILBP..) for texture analysis and classification
- usually used for **facial expression recognition**

Visual-based -> Feature Extraction

2D-based

Shape

- especially for facial image analysis and human detection
- lip, eyes, brow, cheek, furrow, cheeks and chins as human face features
- methods: Adaptive hough transform (AHT), edge orientation histograms (EOH)
- **Edge** is a useful technique to describe the shape information of objects

Sobel operator

Visual-based -> Feature Extraction

2D-based

Motion

- Motion features have been widely used for object detection, **tracking** and recognition
- Optical flow is a typical motion feature, which is the distribution of velocities of brightness patterns movement in an image
- Anther threshold function is applied to compute the absolute difference between continuous frames in a video sequences, and the positions with high intensity values in a binary map were represented as motion features

Optical flow

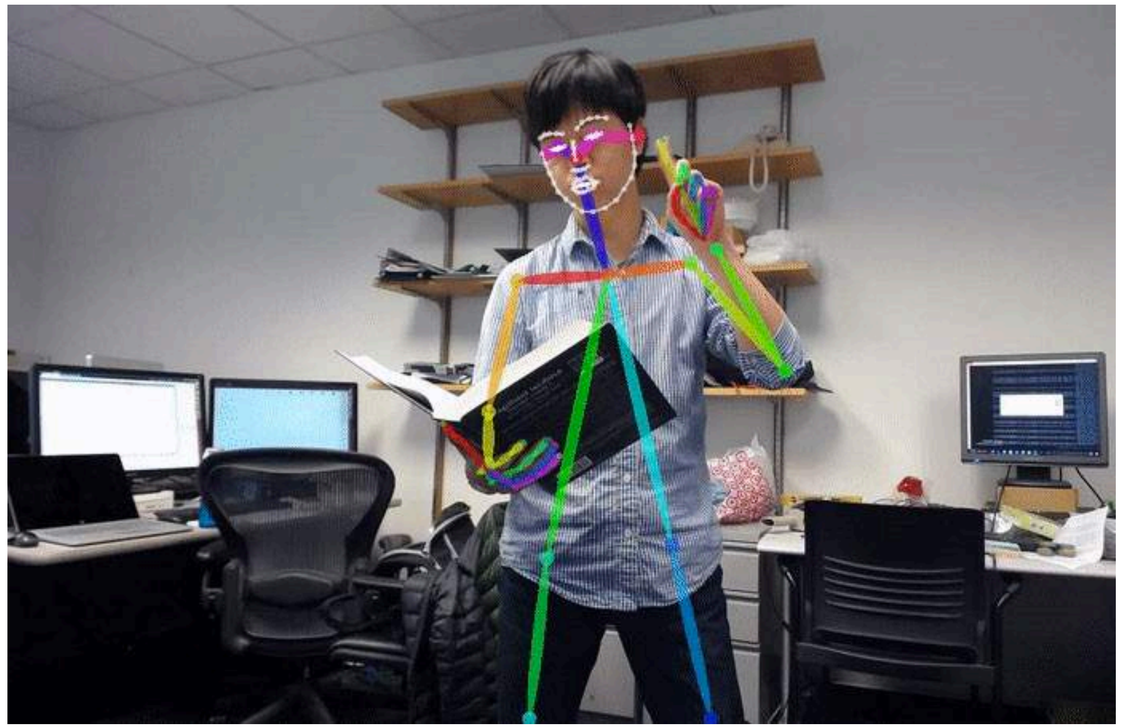
Visual-based -> Feature Extraction

2D-based

Skelton

Real-time multi-person keypoint detection

- 15 or 18-keypoint body estimation
- 2x21-keypoint hand estimation
- 70-keypoint face estimation

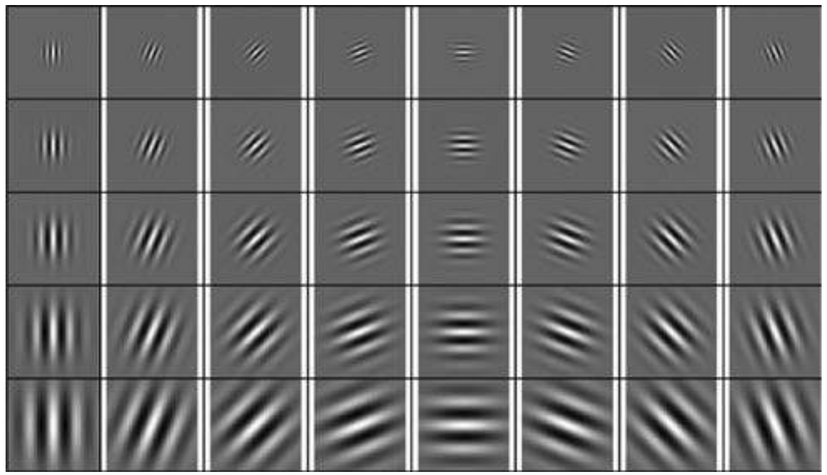


Visual-based -> Feature Extraction

2D-based

Filtered Feature

- discrete cosine transform (DCT)
- Gabor wavelet
- DCT features have been used in local feature-based face recognition
- Gabor-based features have shown good performance in face recognition and facial expression recognition



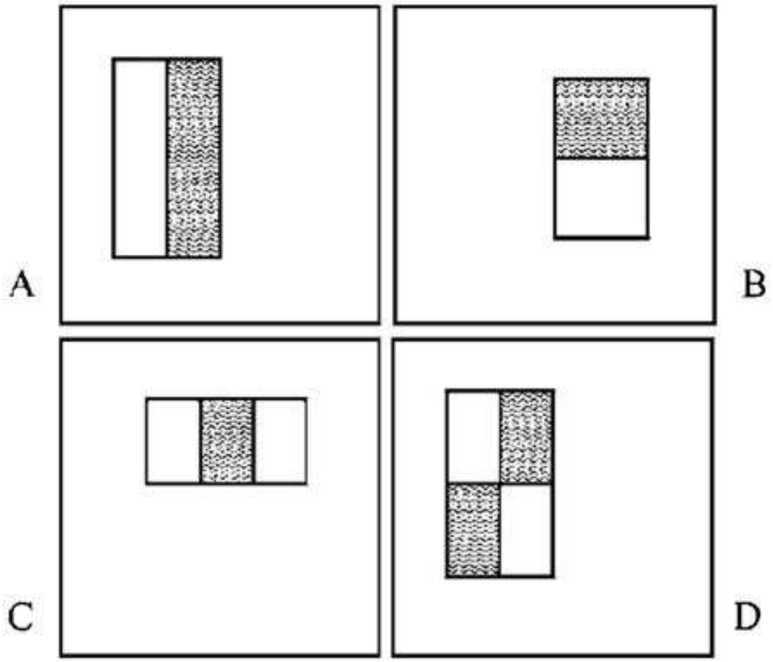
Gabor kernels
at eight orientations and five scales

Visual-based -> Feature Extraction

2D-based

Haar-Like Feature

- originally defined as the difference of the sum of pixels in two rectangular areas
- The famous success is the Haar-like features based real-time face detection



Visual-based -> Feature Extraction

3D-based

RGB-D signals
Stereo-Based Depth Signals

- 3D visual signals can provide **distance** information of objects which the robot is manipulating or interacting
- Kinect and stereo vision systems can obtain **3D information of objects**.
- Depth images also can be converted to **3D cloud points** by mapping each pixel into the corresponding 3D coordinates

Visual-based -> Feature Extraction

3D-based

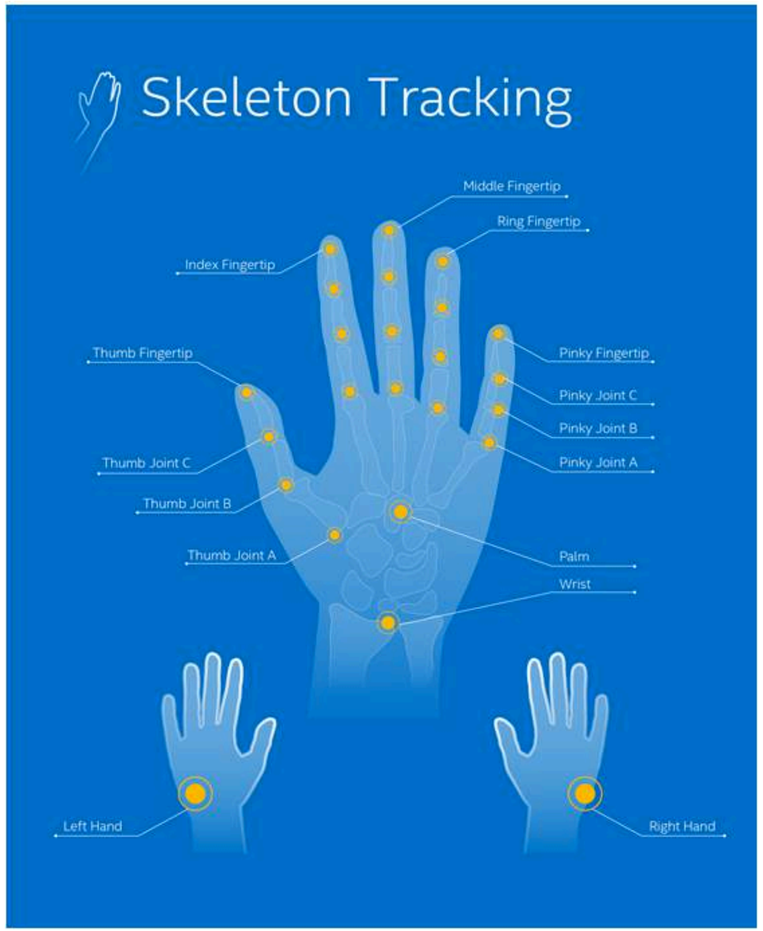
RGB-D signals
Stereo-Based Depth Signals

- Kinect camera can only work over a **range of several meters**
- Stereo vision, many commercially available systems have been developed due to mature techniques. These systems are of different sizes and can be used for both **indoor and outdoor** environments

Visual-based -> Feature Extraction

3D-based

Skelton



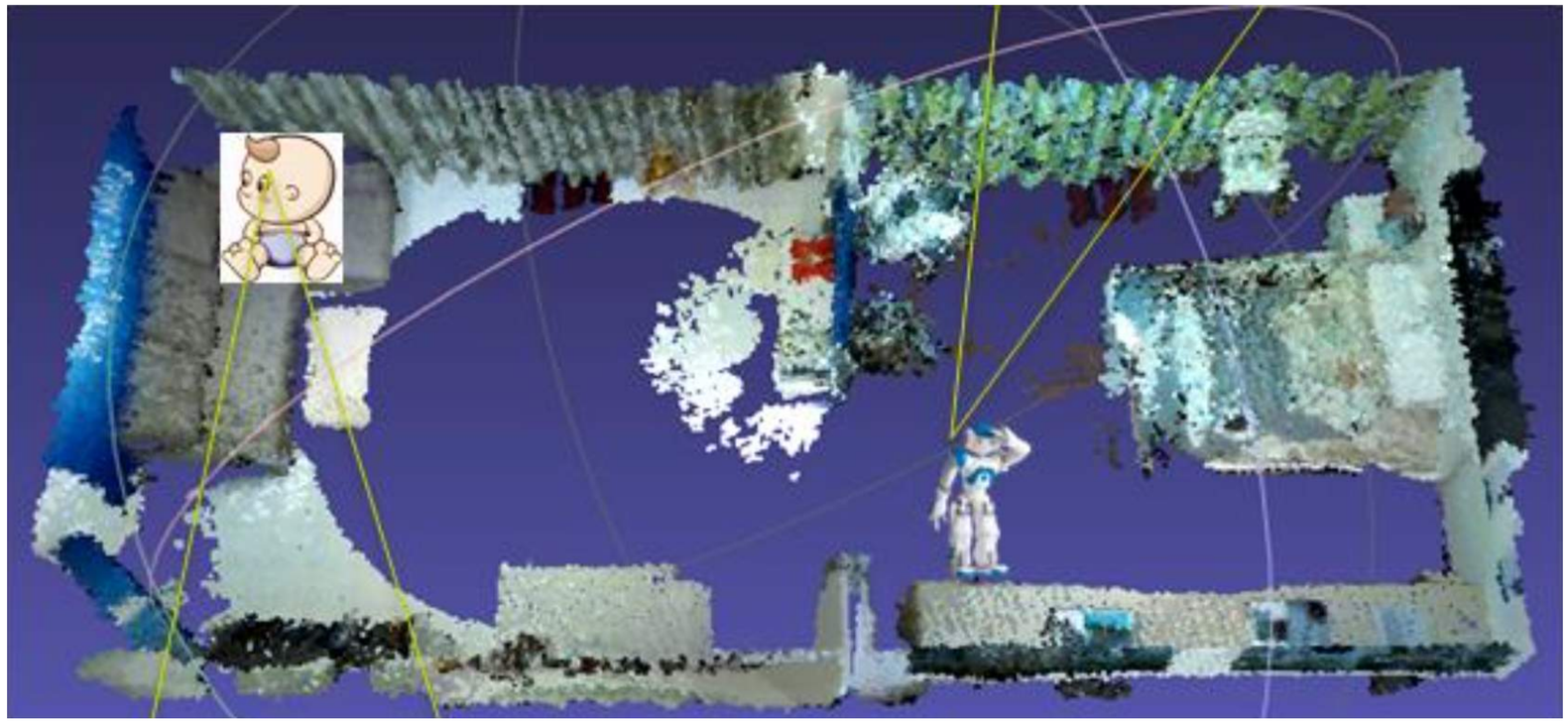
Visual-based -> Feature Extraction

3D-based

3d point cloud

A **point cloud** is a set of data points in some coordinate system.

In a three-dimensional coordinate system, these points are usually defined by *X*, *Y*, and *Z* coordinates, and often are intended to represent the external surface of an object.



Visual-based -> Semantic Understanding

Methods

Object Detection & Recognition

- Template Matching
- Clustering
- Nearest Neighbor
- Boosting
- Gaussian Mixture Models

Object Tracking

- Gradient descent
- mean shift
- Kalman filter
- Particle filter

Object: human, face, body, object…

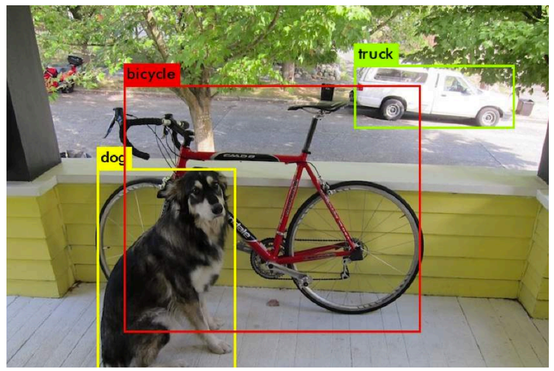
Visual-based -> Semantic Understanding

Methods

Object Detection & Recognition

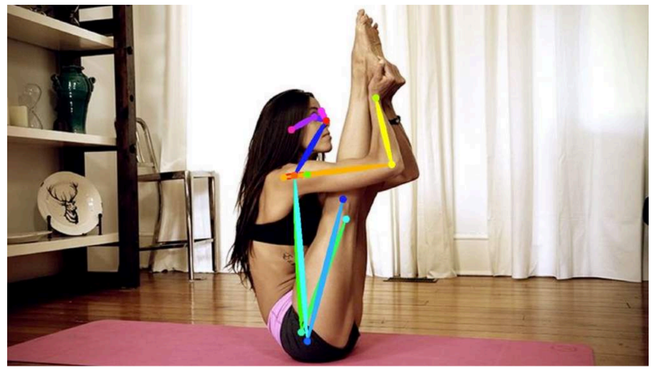
YOLO

link



Open Pose

link



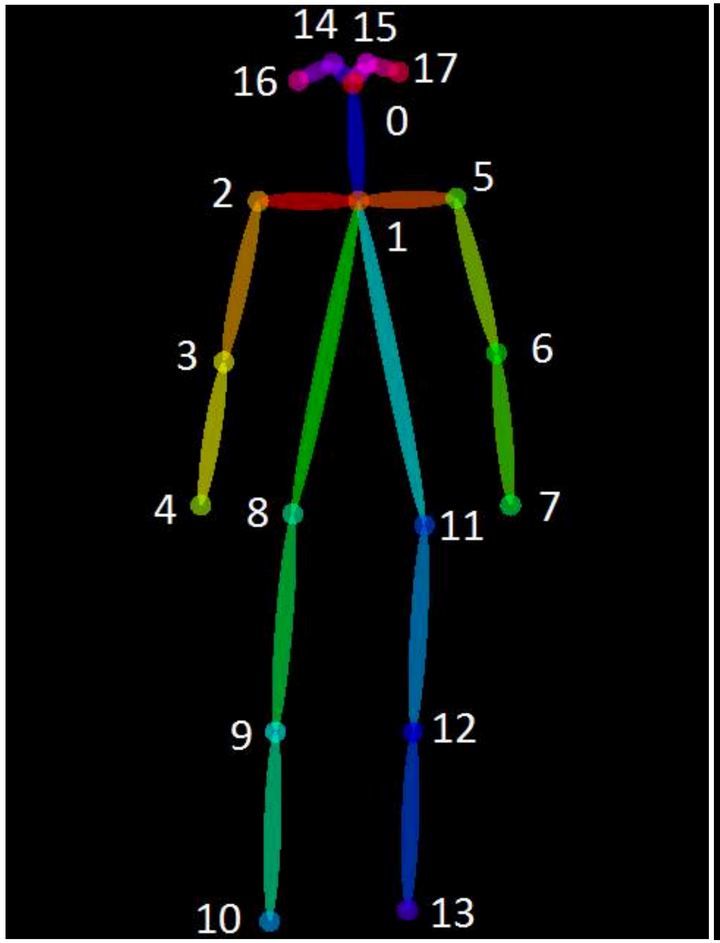
Dense Pose

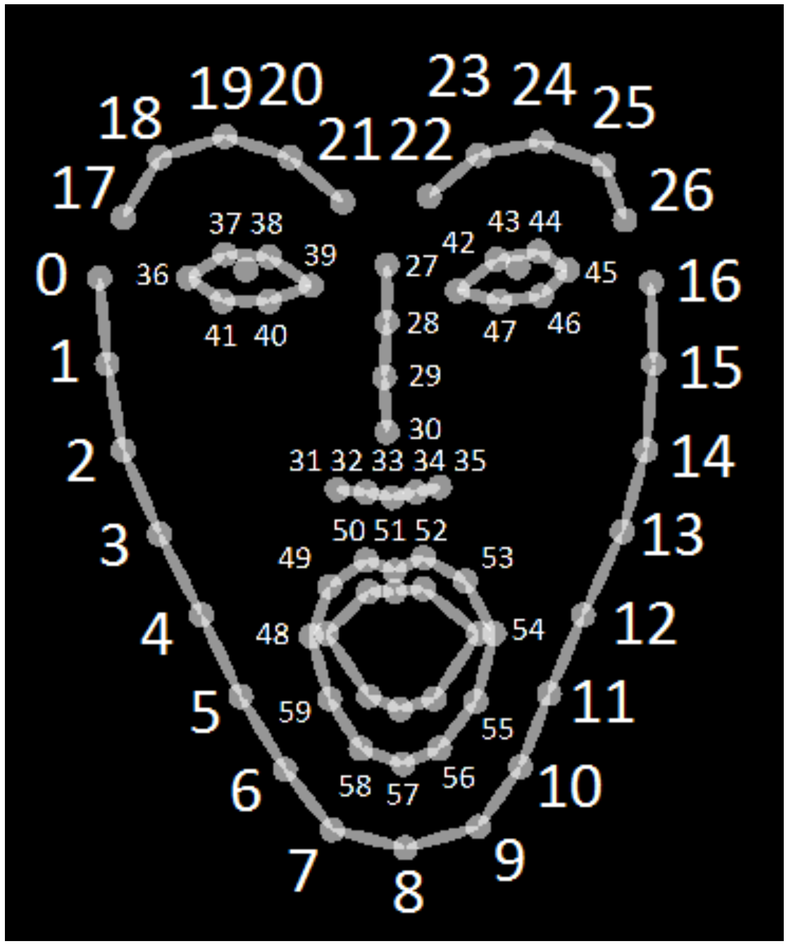
link

Visual-based -> Feature Extraction

2D-based

Skelton - OpenPose





OpenPose represents the first real-time multi-person system to jointly detect human body, hand, and facial keypoints (in total 130 keypoints) on single images.

Input: Image, video, webcam, and IP camera. Included C++ demos to add your custom input.

Output: Basic image + keypoint display/saving (PNG, JPG, AVI, ...), keypoint saving (**JSON**, XML, YML, ...), and/or keypoints as array class.

link

Visual-based -> Feature Extraction

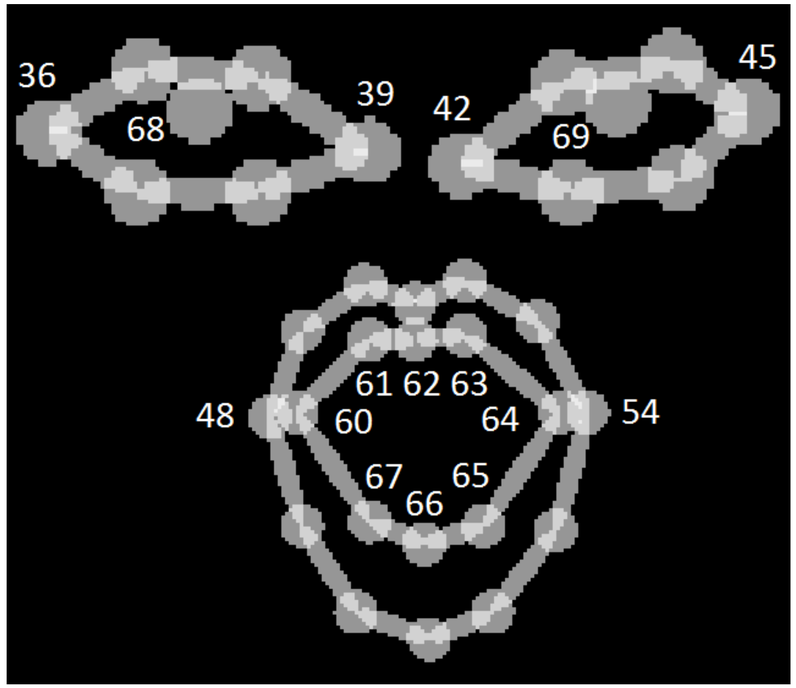
2D-based

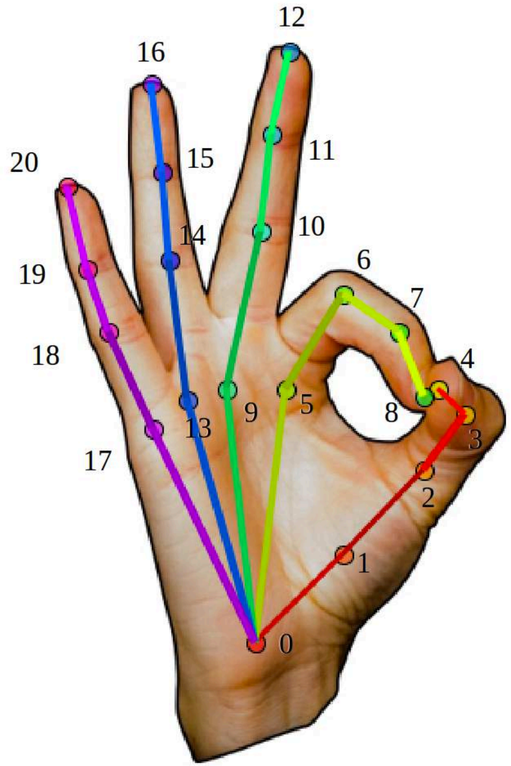
Skelton - OpenPose

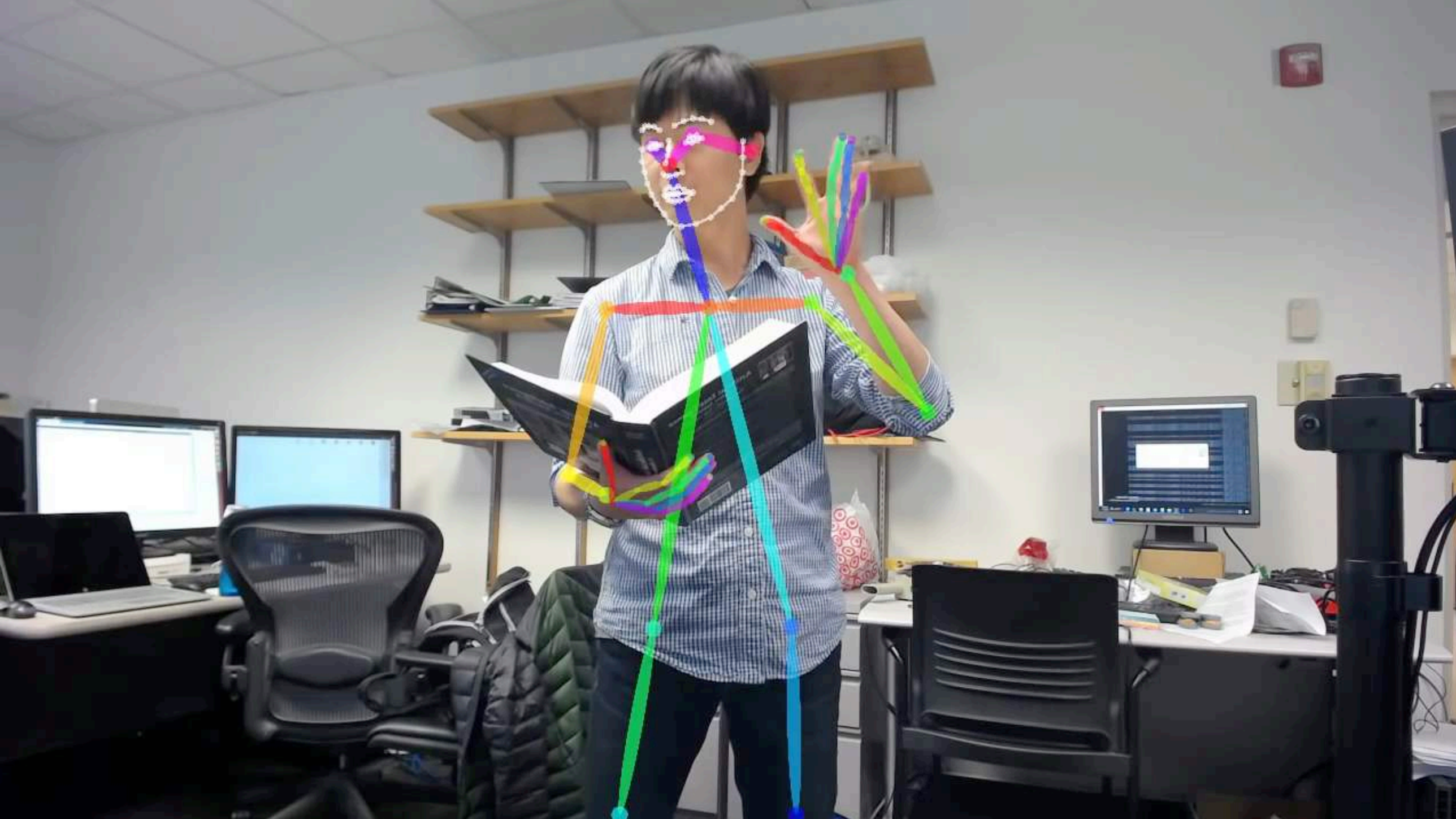
OpenPose represents the first real-time multi-person system to jointly detect human body, hand, and facial keypoints (in total 130 keypoints) on single images.

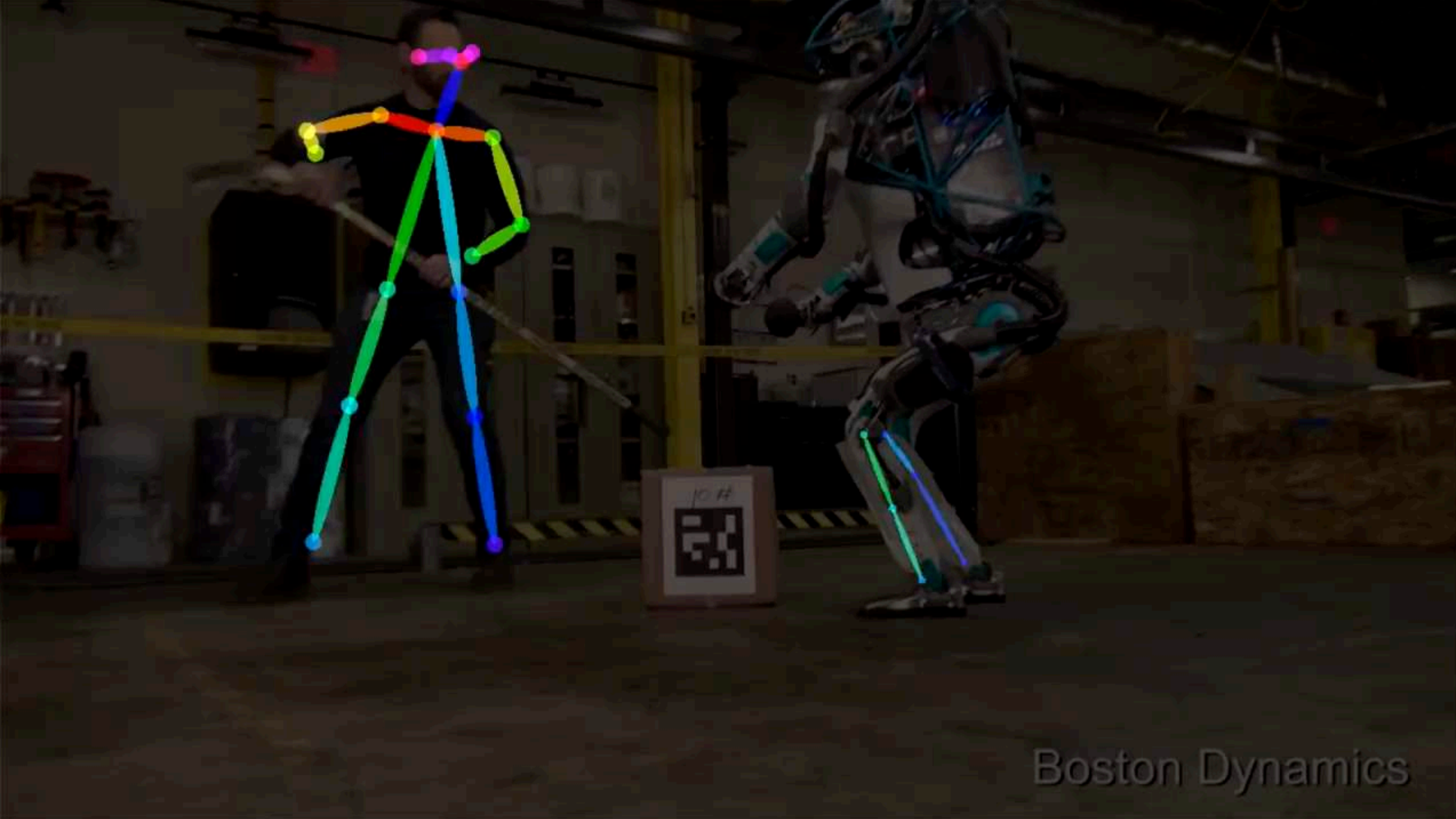
Input: Image, video, webcam, and IP camera. Included C++ demos to add your custom input.

Output: Basic image + keypoint display/saving (PNG, JPG, AVI, ...), keypoint saving (**JSON**, XML, YML, ...), and/or keypoints as array class.








Boston Dynamics

Visual-based -> Feature Extraction

2D-based

Skelton - OpenPose

OpenPose represents the first real-time multi-person system to jointly detect human body, hand, and facial keypoints (in total 130 keypoints) on single images.

Input: Image, video, webcam, and IP camera. Included C++ demos to add your custom input.

Output: Basic image + keypoint display/saving (PNG, JPG, AVI, ...), keypoint saving (**JSON**, XML, YML, ...), and/or keypoints as array class.

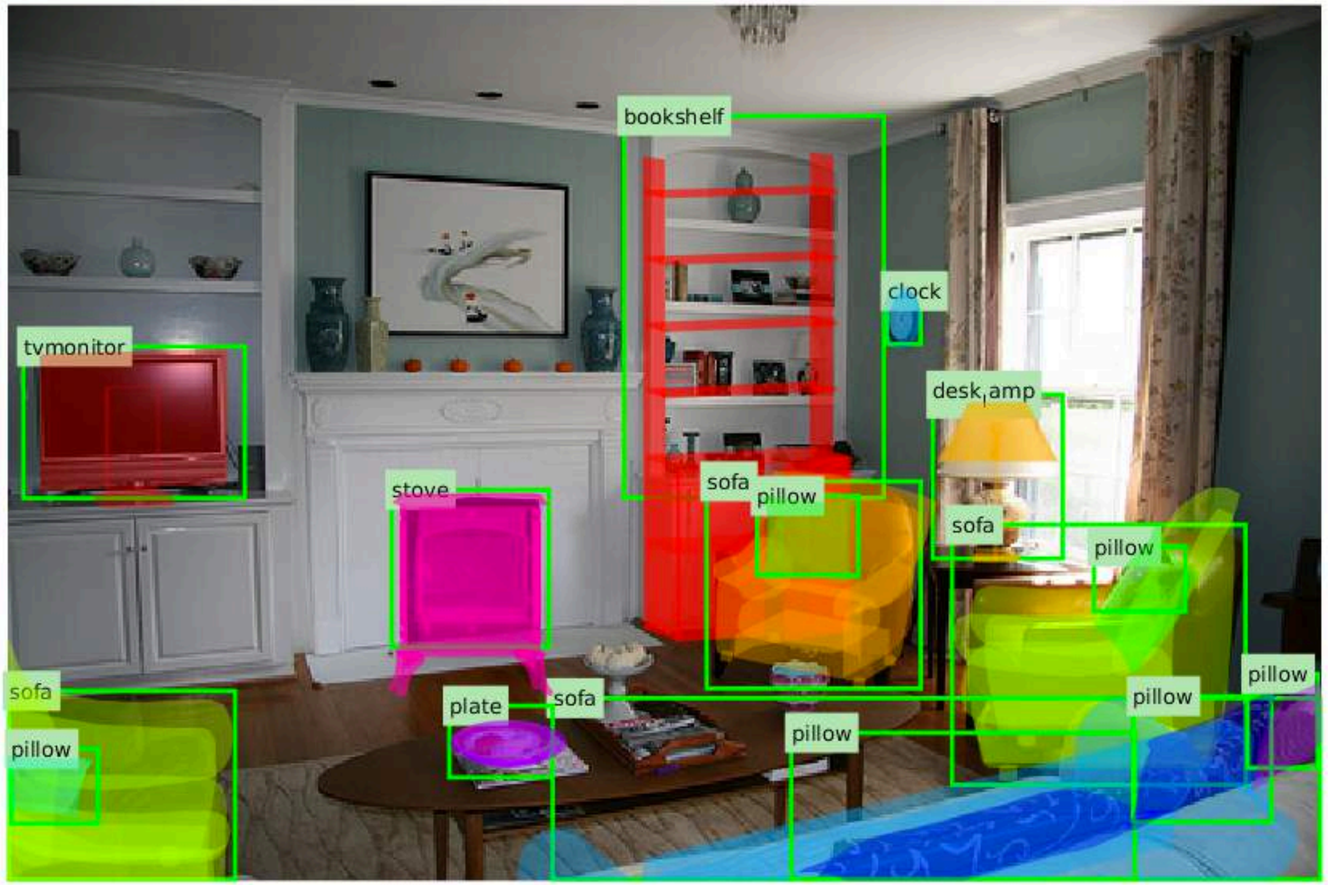

```
{
    "version":1.1,
    "people":[
        {
            "pose_keypoints_2d":[582.349,507.866,0.845918,746.975,631.307,0.587007,...],
            "face_keypoints_2d":[468.725,715.636,0.189116,554.963,652.863,0.665039,...],
            "hand_left_keypoints_2d":[746.975,631.307,0.587007,615.659,617.567,0.377899,...],
            "hand_right_keypoints_2d":[617.581,472.65,0.797508,0,0,0,723.431,462.783,0.88765,...]
            "pose_keypoints_3d":[582.349,507.866,507.866,0.845918,507.866,746.975,631.307,0.587007,...],
            "face_keypoints_3d":[468.725,715.636,715.636,0.189116,715.636,554.963,652.863,0.665039,...],
            "hand_left_keypoints_3d":[746.975,631.307,631.307,0.587007,631.307,615.659,617.567,0.377899,...
            "hand_right_keypoints_3d":[617.581,472.65,472.65,0.797508,472.65,0,0,0,723.431,462.783,0.88765,
        }
    ],
    // If `--part_candidates` enabled
    "part_candidates":[
        {
            "0":[296.994,258.976,0.845918,238.996,365.027,0.189116],
            "1":[381.024,321.984,0.587007],
            "2":[313.996,314.97,0.377899],
            "3":[238.996,365.027,0.189116],
            "4":[283.015,332.986,0.665039],
            "5":[457.987,324.003,0.430488,283.015,332.986,0.665039],
            "6":[],
            "7":[],
            "8":[],
            "9":[],
            "10":[],
            "11":[],
            "12":[],
            "13":[],
            "14":[293.001,242.991,0.674305],
            "15":[314.978,241,0.797508],
            "16":[],
            "17":[369.007,235.964,0.88765]
        }
    ]
}
```

Visual-based -> Feature Extraction

2D-based

Object

Object can be served as one type of feature extracted from image or video in multiply scenarios such as home, roads, workplace, public places, etc. Extraction is generally represent with label, bounding box and confidence.



Visual-based -> Feature Extraction

2D-based

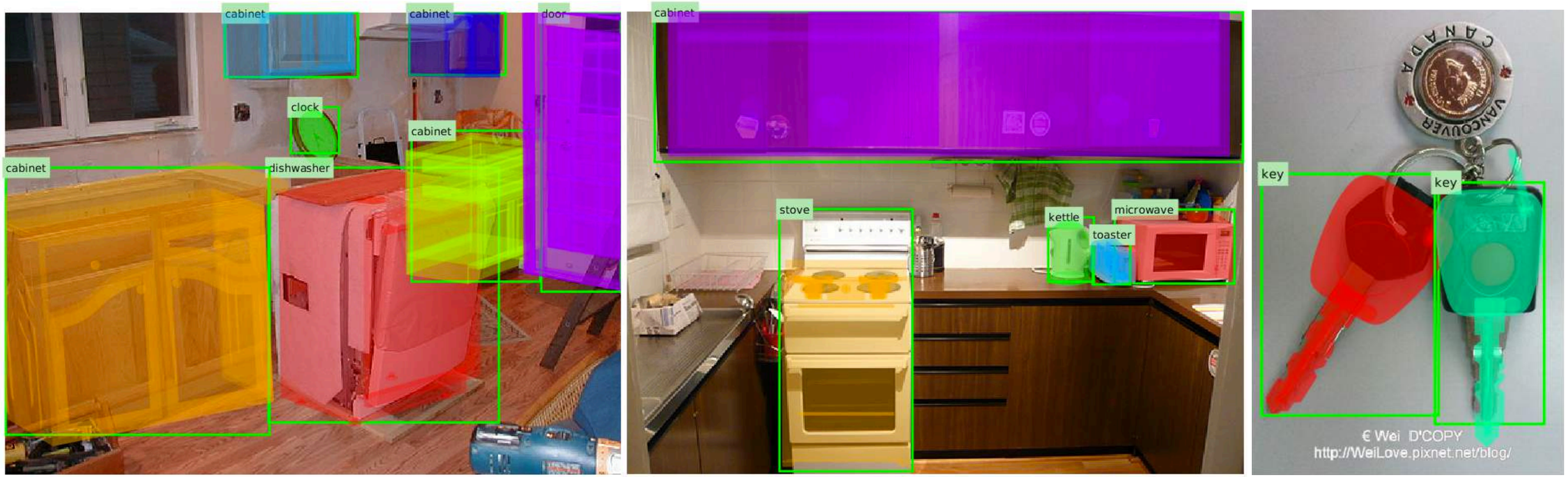
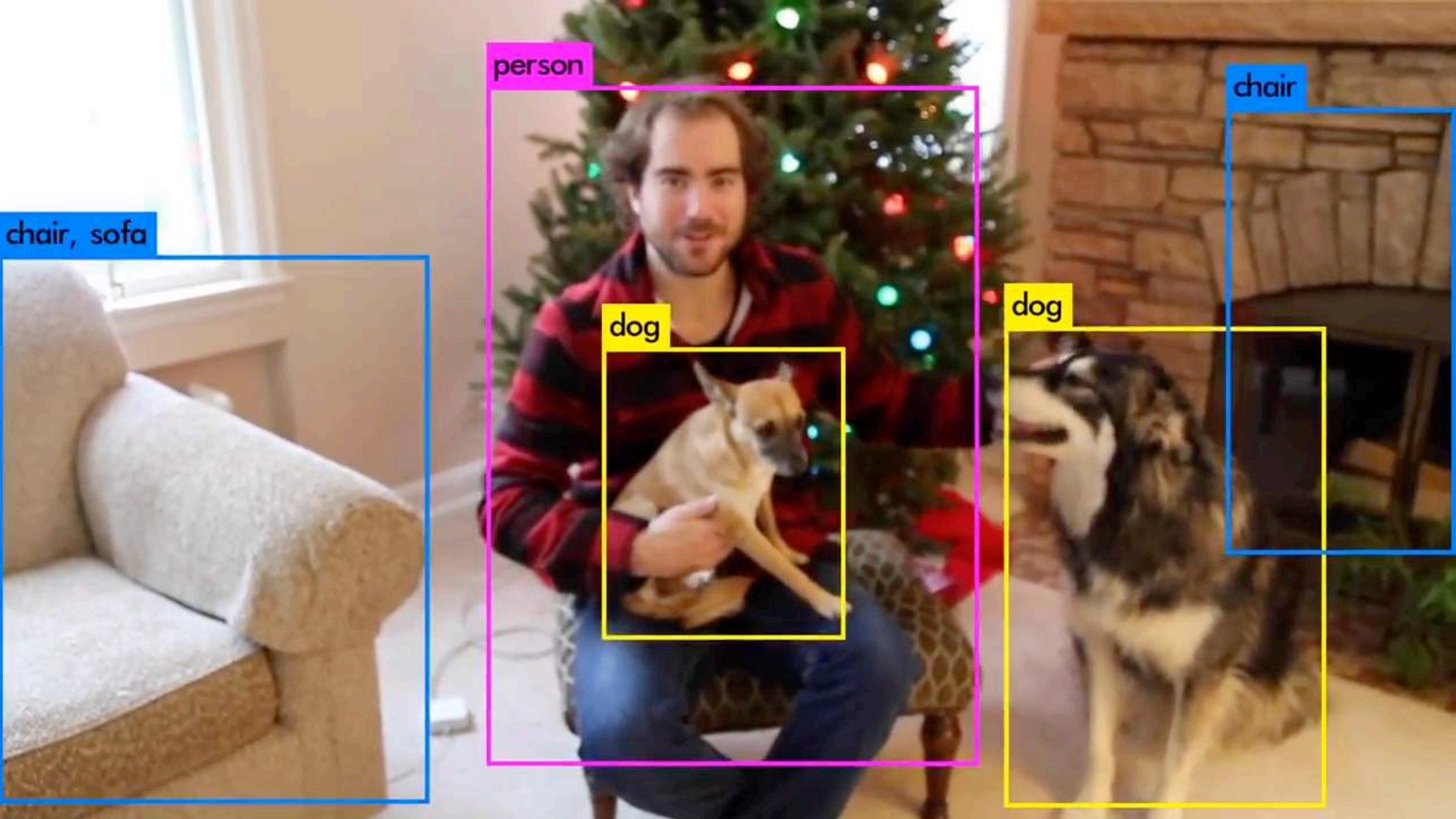
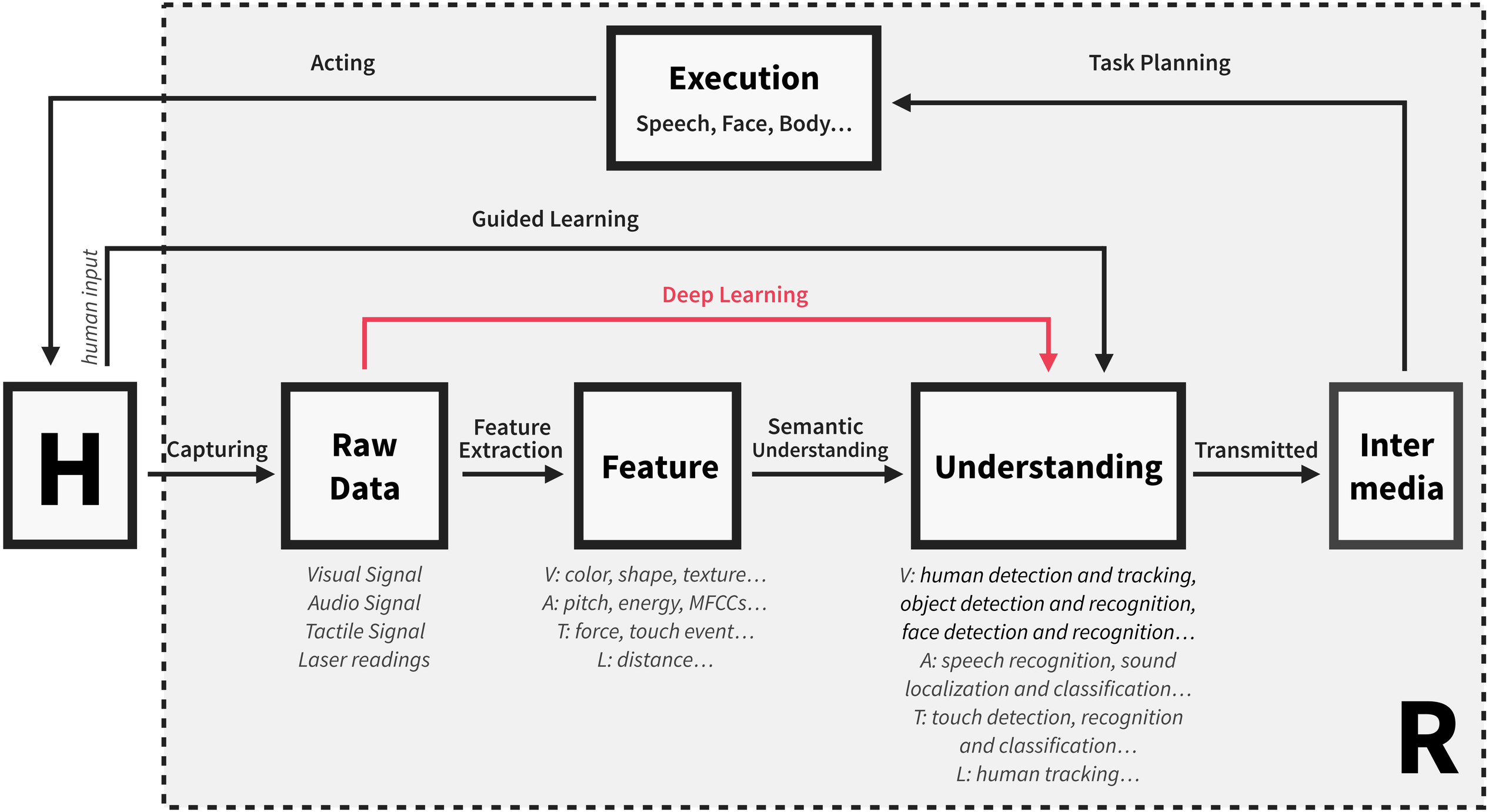


image source: ObjectNet


person
chair
chair, sofa
dog
dog


Acting
Execution
Speech, Face, Body…
Task Planning
Guided Learning
human input
Deep Learning
H
Capturing
Raw Data
Feature Extraction
Feature
Semantic Understanding
Understanding
Transmitted
Inter media
Visual Signal
Audio Signal
Tactile Signal
Laser readings
V: color, shape, texture…
A: pitch, energy, MFCCs…
T: force, touch event…
L: distance…
V: human detection and tracking, object detection and recognition, face detection and recognition…
A: speech recognition, sound localization and classification…
T: touch detection, recognition and classification…
L: human tracking…
R

Visual-based -> Neural Network

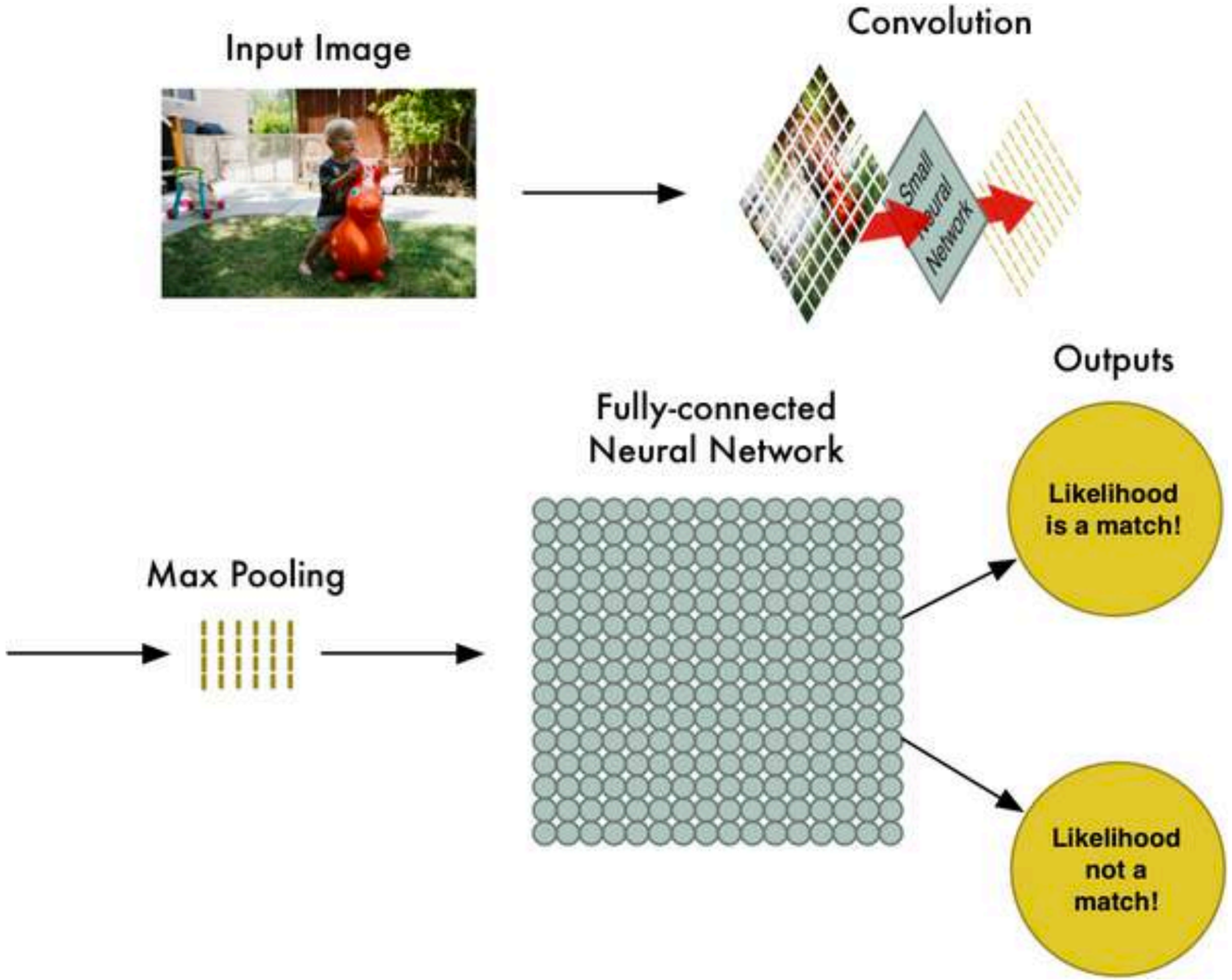
Methods

CNN（卷积神经网络）
Convolutional Neural Networks

Convolution -> feature extraction

Max Pooling -> reduce the over-fitting

Fully-connected NN -> classification



Visual-based -> Neural Network

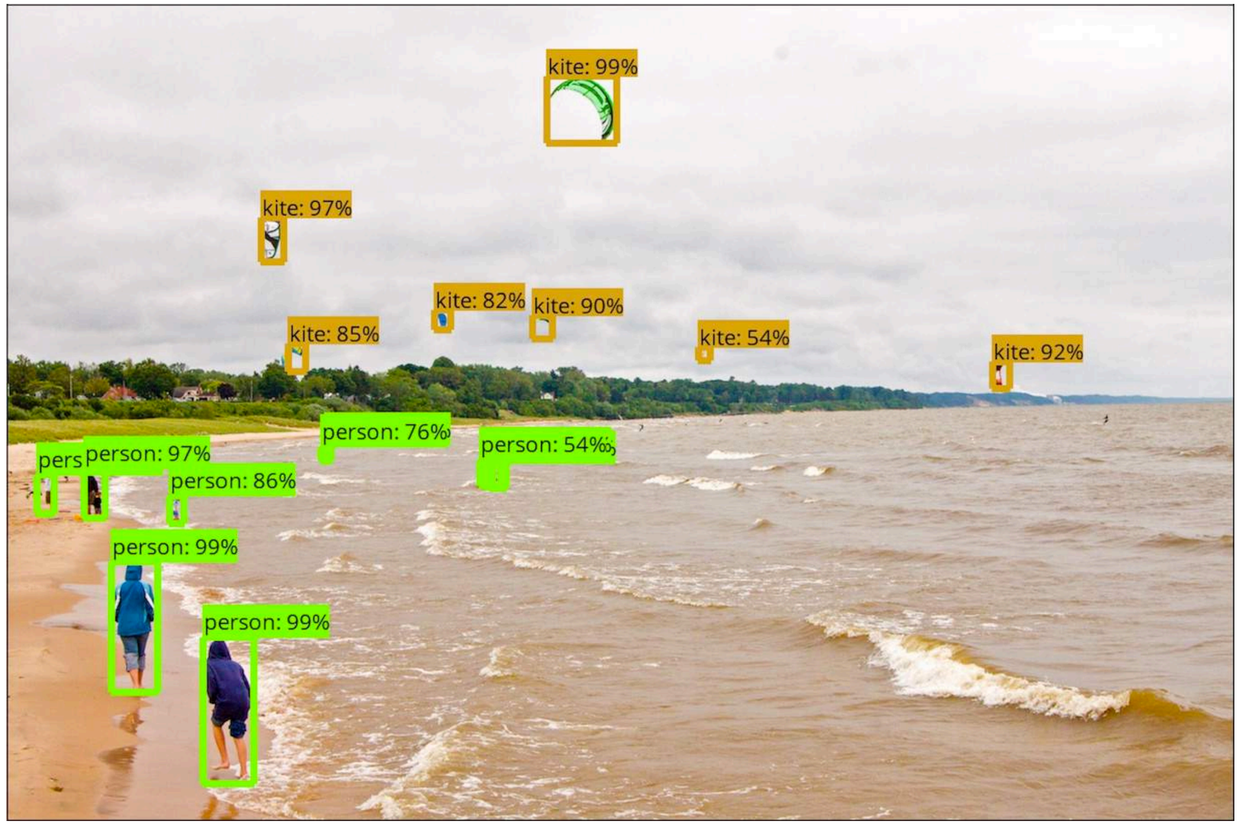
Methods

CNN（卷积神经网络）
Convolutional Neural Networks

Google TenserFlow Object Detection API

link

The TensorFlow Object Detection API is an open source framework built on top of TensorFlow that makes it easy to construct, train and deploy object detection models. It creates accurate machine learning models capable of localizing and identifying multiple objects in a single image.



Perception

Signals → **Methods**

Visual-based

Audio-based

Tactile-based

Laser sensors

Data Capturing

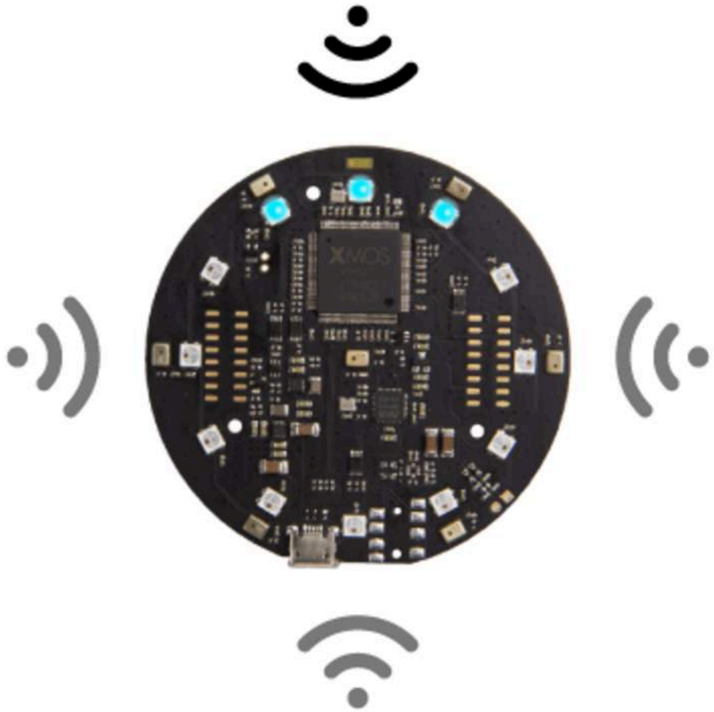
Feature Extraction

Semantic Understanding

Audio-based -> Data Capturing

Sensors

Microphone
Microphone Array



ReSpeaker Mic Array
7 PDM digital microphones
10m detective distance

Visual-based -> Feature Extraction

Feature

Pitch
Energy
MFCCs

Audio-based -> Feature Extraction

Pitch

- Pitch refers to the fundamental frequency
- Pitch contour, range, mean, median, inflection range, and rate of change
- Methods: simplified inverse filter tracking (SIFT), the pitch constraints and dynamic programming search, and the statistical method based on cep-strum

Audio-based -> Feature Extraction

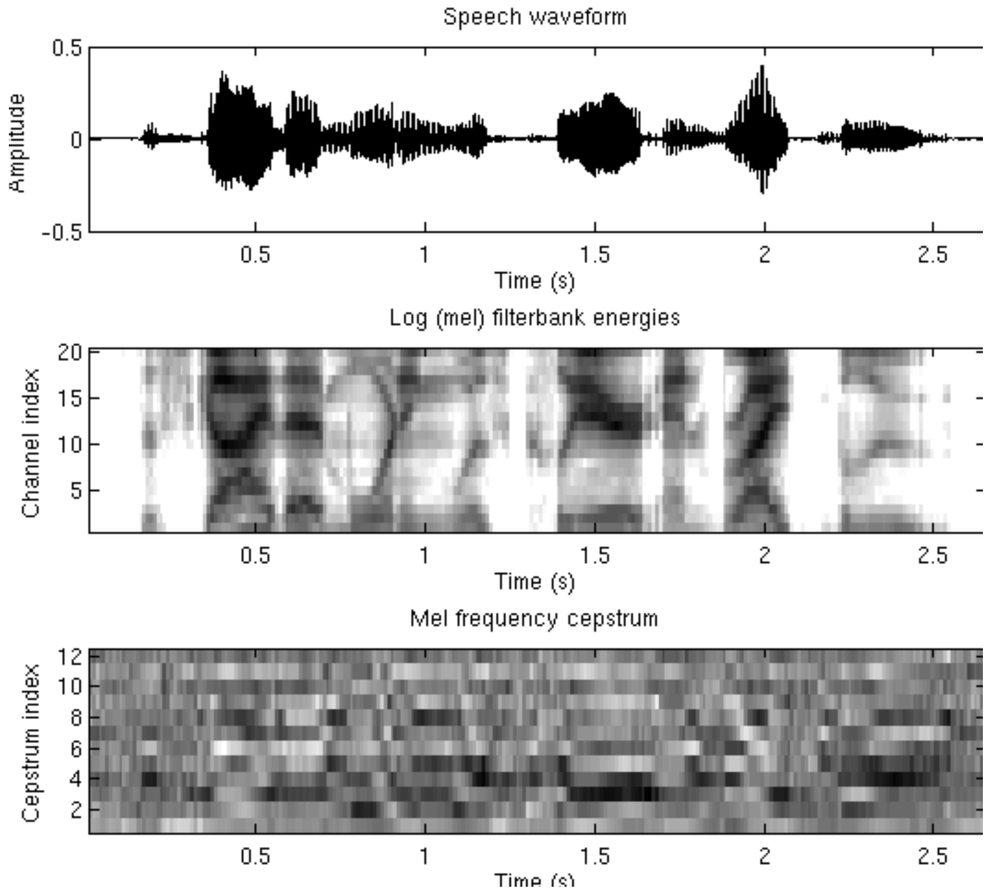
Energy

- volume or intensity of speech which aims to measure the variations of speech signals' amplitude

Audio-based -> Feature Extraction

MFCCs (频率倒谱系数)

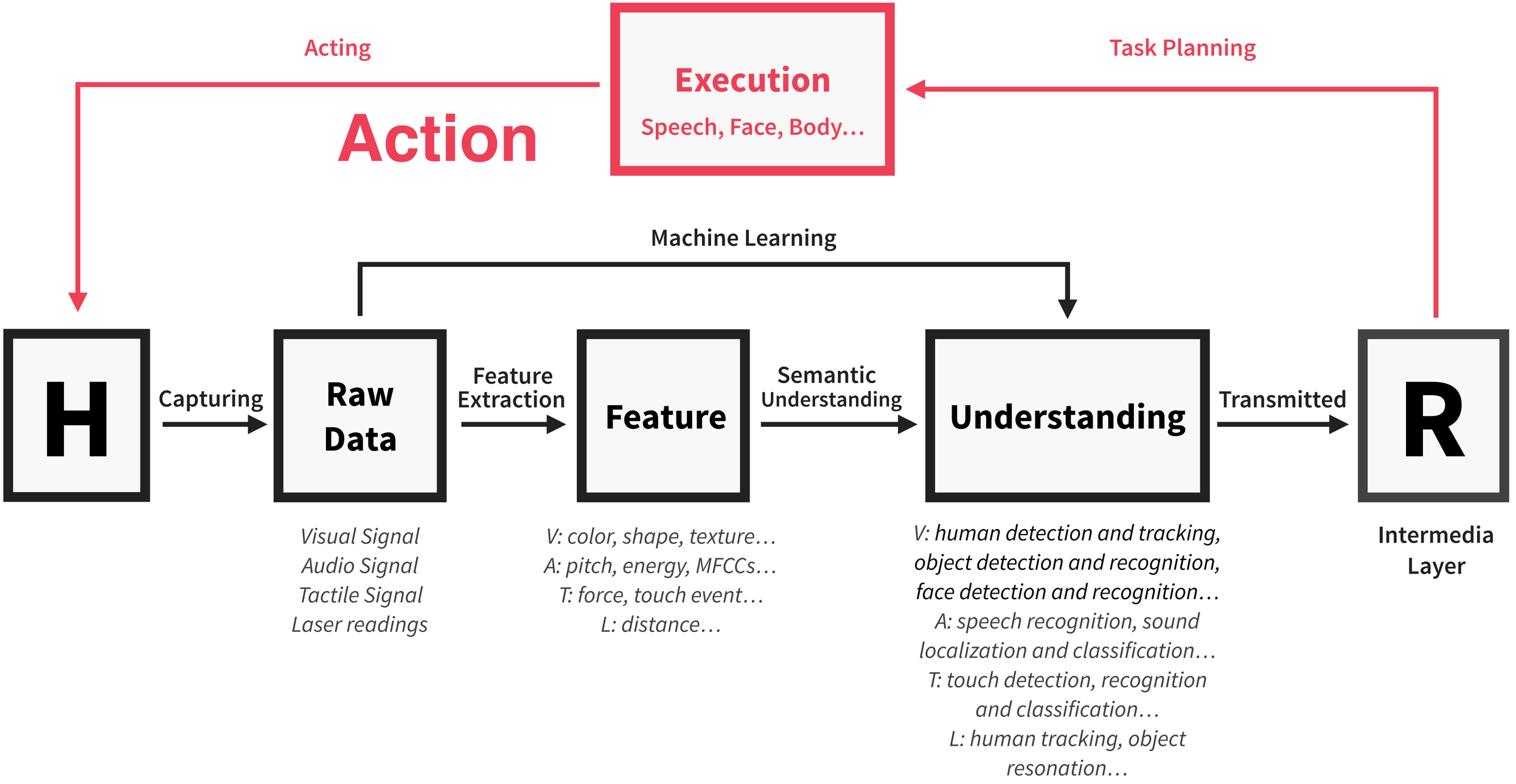
- MFCCs are coefficients which collectively make up an mel frequency cepstrum (MFC) to represent a short-term power spectrum of a sound

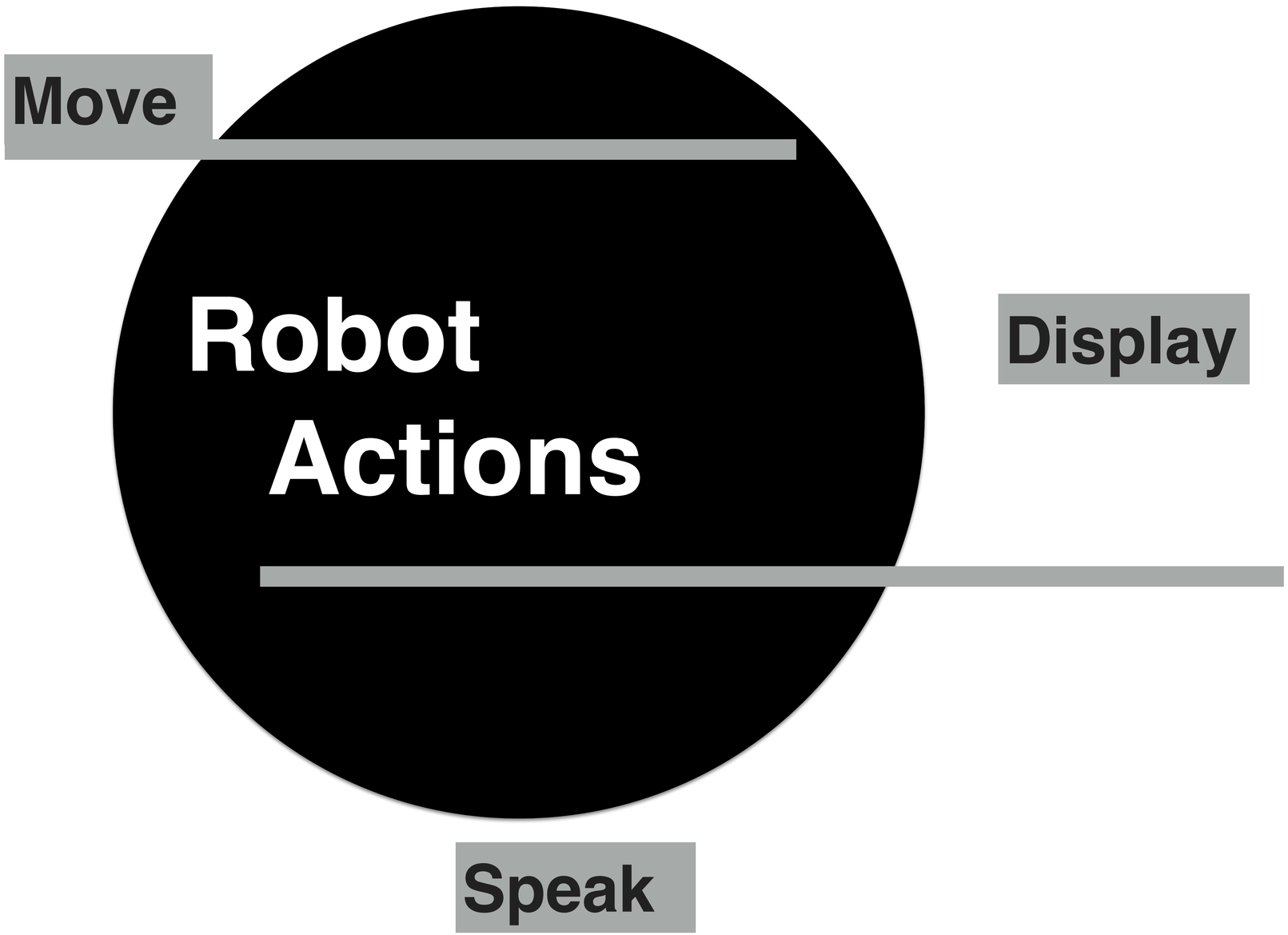


Audio-based -> Semantic Understanding

Methods

Speaker Localization
Speech recognition
Sound event classification
Emotion recognition
Rhythms recognition

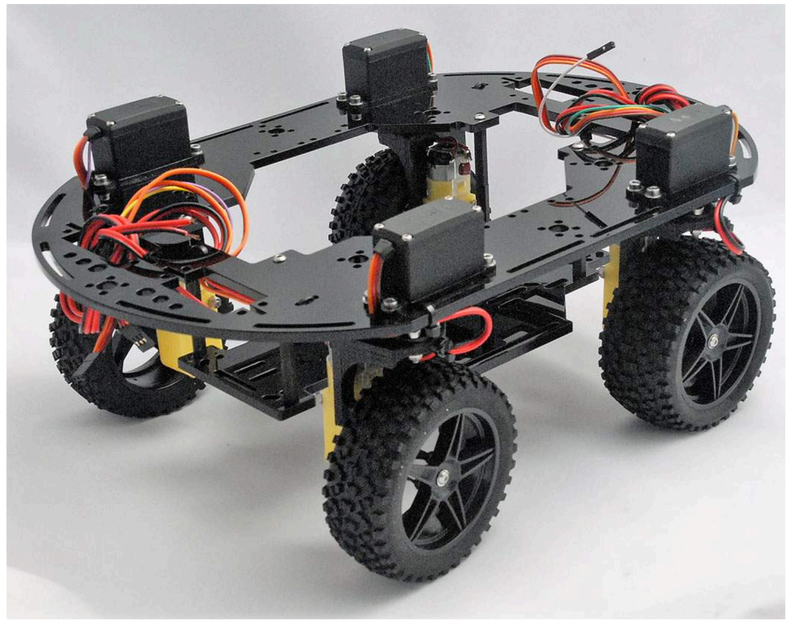

Acting
Execution
Speech, Face, Body…
Task Planning
Action
Machine Learning
H
Capturing
Raw Data
Feature Extraction
Feature
Semantic Understanding
Understanding
Transmitted
R
Visual Signal
Audio Signal
Tactile Signal
Laser readings
V: color, shape, texture…
A: pitch, energy, MFCCs…
T: force, touch event…
L: distance…
V: human detection and tracking, object detection and recognition, face detection and recognition…
A: speech recognition, sound localization and classification…
T: touch detection, recognition and classification…
L: human tracking, object resonation…
Intermedia Layer


Move
Robot Actions
Display
Speak

Action

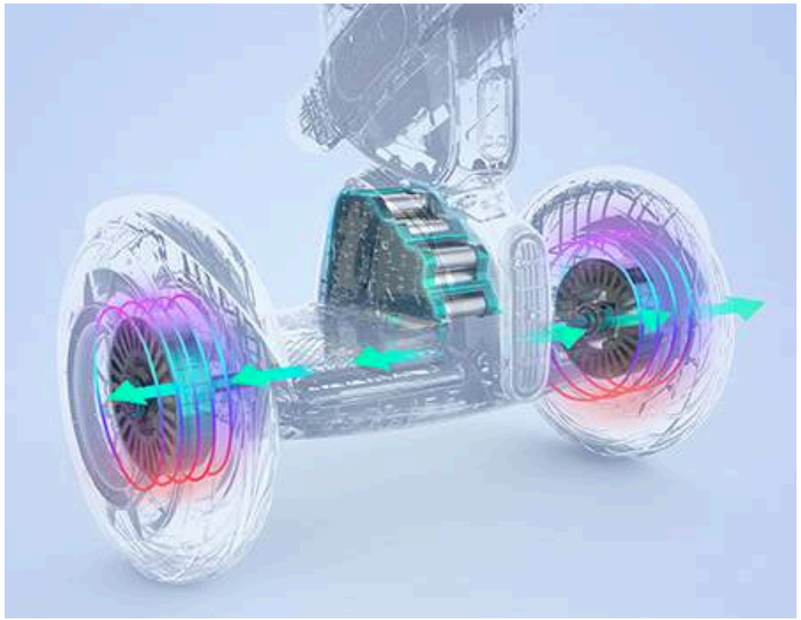
Move walk

supported by mobile robot platform



four-wheel car

two-wheels car



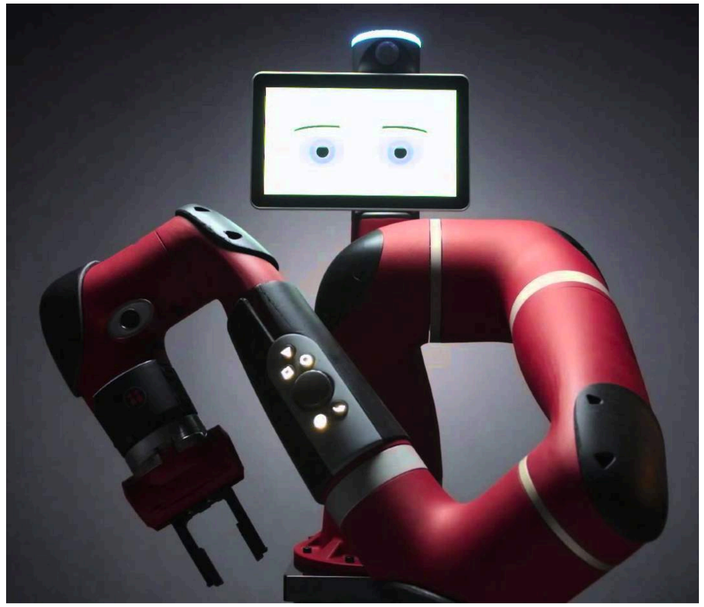
segway

Action

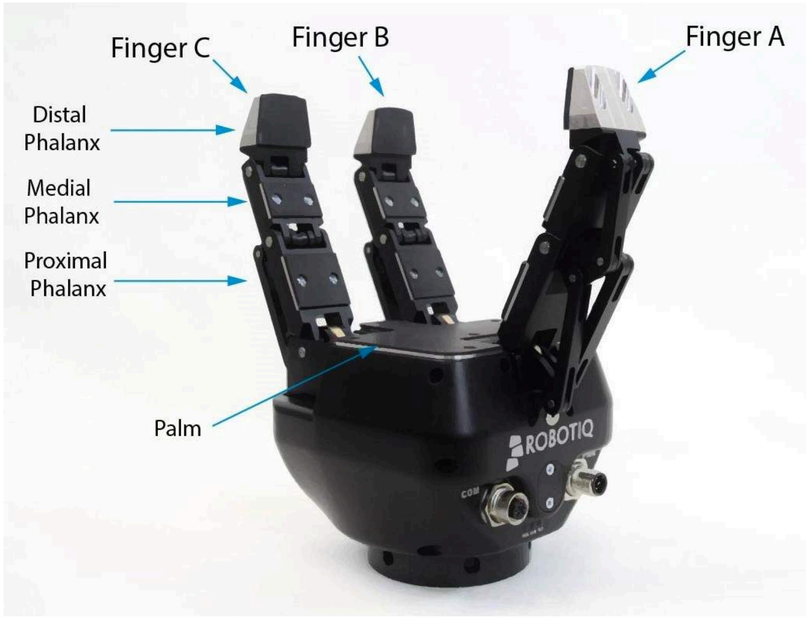
Move move arms and hands

supported by robot arms and hands kit

embedded double arms



Single arm

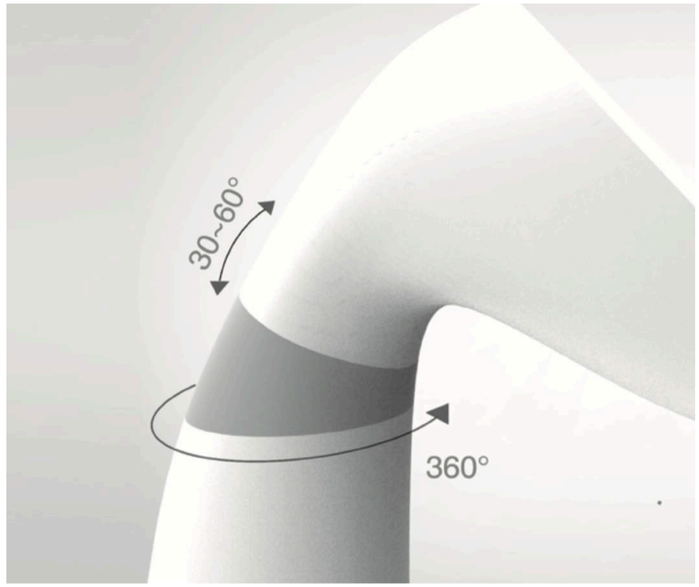


Three-finger robot hand

Action

Move rotate head

supported by mechanical construction

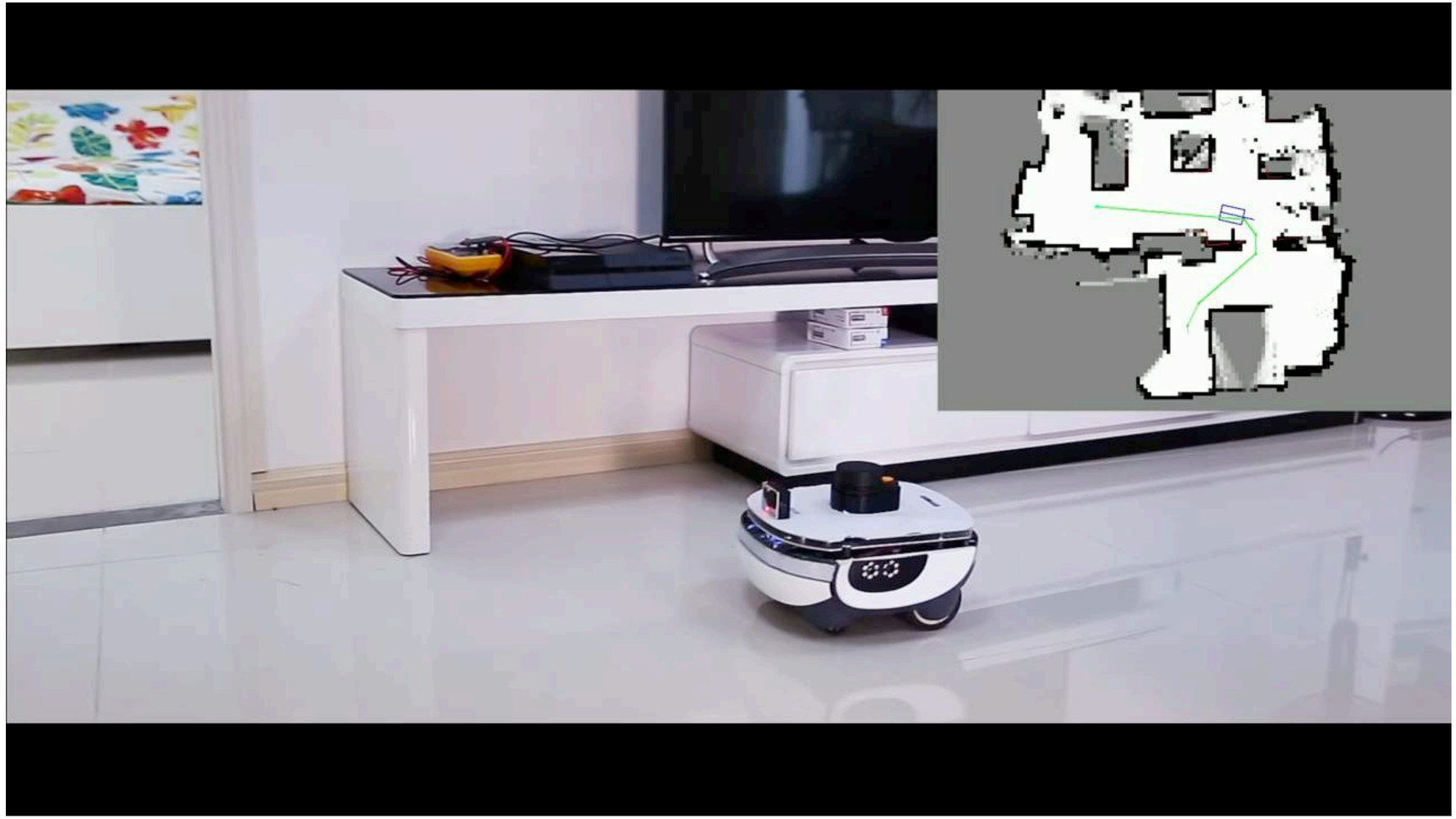


2 degrees of freedom

Action

Move localize and navigate

SLAM (Simultaneous localization and mapping) technology

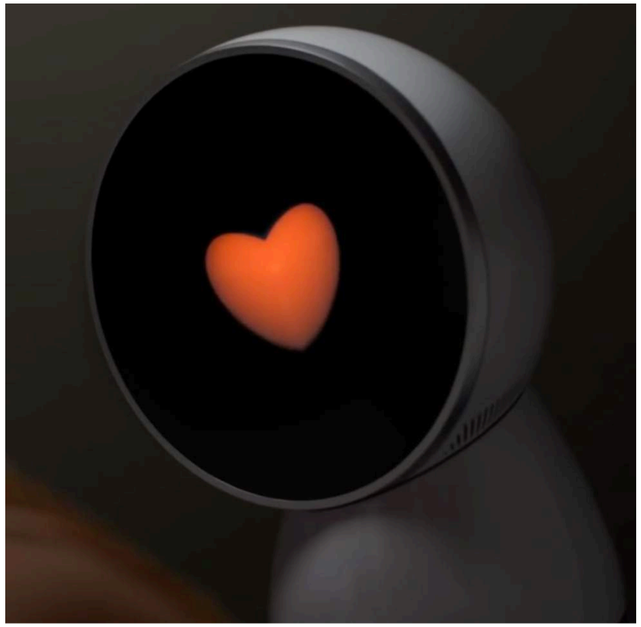


Action

Display

supported by multi-touch LCD screen

user interface



iconic emotion

image/video

Action

Speak

supported by speech synthesis technology

Configuration

- multi language
- pitch
- volume
- gender
- speed

Content

- word
- number/address/name
- sentence
- dialogue
- question & answer

Case