Chapter 5: CV and HCI

Lecturer: Wei Liang

Content

- Computer Vision
- Recognition

Head and face, eyes, facial expression, hand and gesture, body and gesture

- Use Cases

People with disabilities, Entertainment, shopping, office, videoconference, virtual input devices, object-computer interaction, remote control, wearable visual interface

What is computer vision

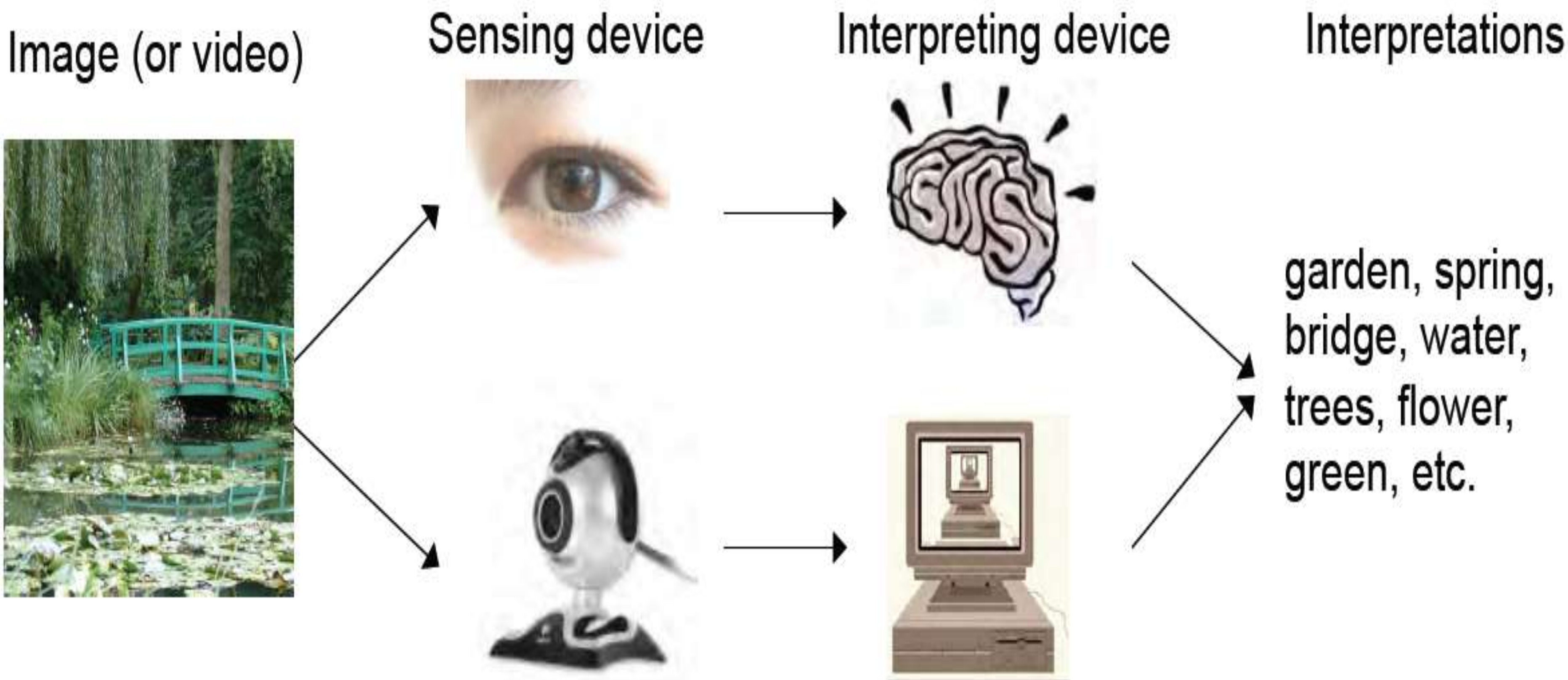


Image (or video)
Sensing device
Interpreting device
Interpretations
garden, spring,
bridge, water,
trees, flower,
green, etc.

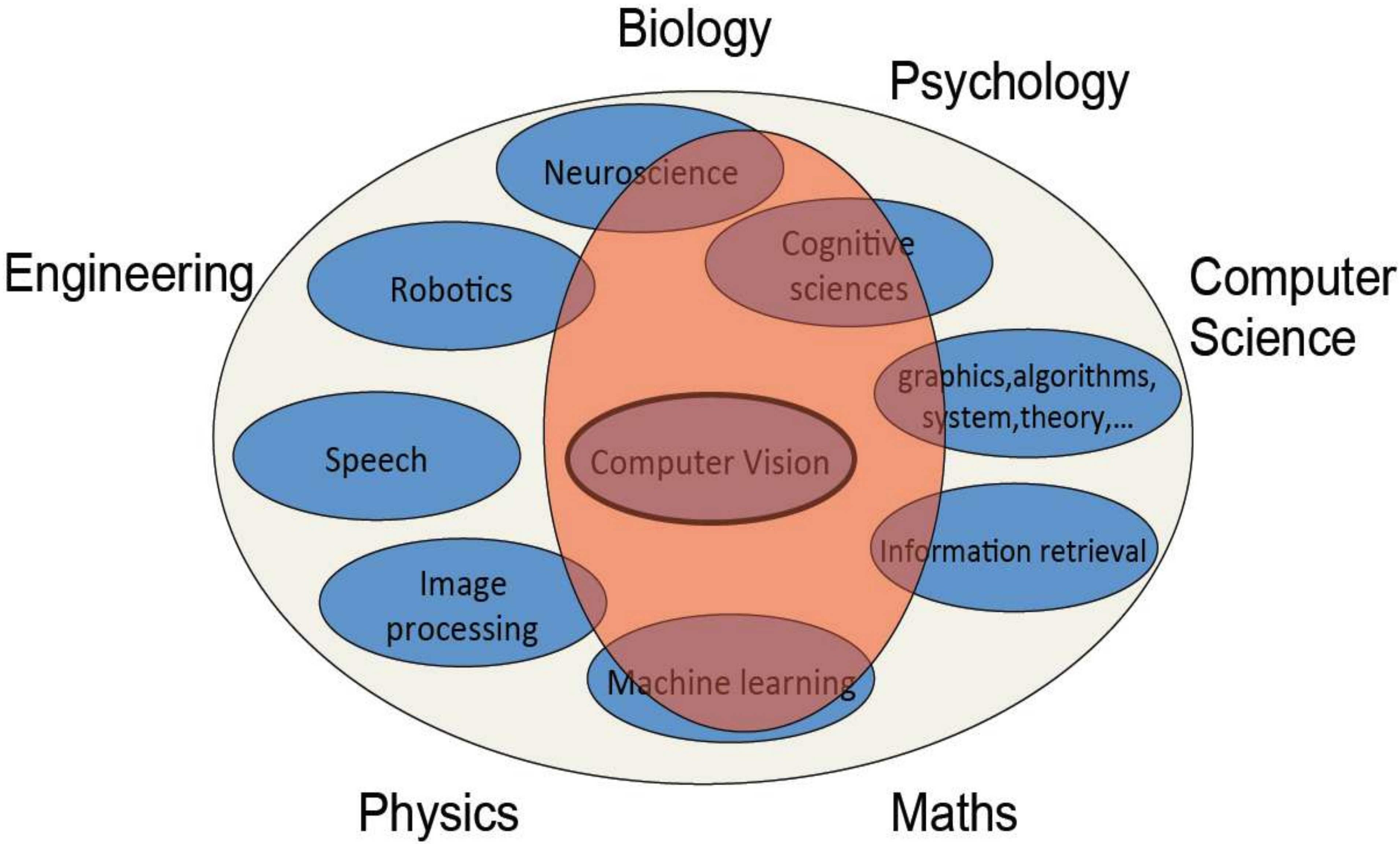
The goal of computer vision

To bridge the gap between pixels and "meaning"

What we see

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

What a computer sees


Biology
Psychology
Engineering
Computer Science
Physics
Maths
Neuroscience
Cognitive sciences
Robotics
graphics,algorithms, system,theory,...
Speech
Computer Vision
Information retrieval
Image processing
Machine learning

- Image Formation
- Early Vision
  - Linear Filters, Local Features, Texture
  - Stereopsis, Structure from Motion
- Middle-Level Vision
  - Segmentation, Model Fitting, Tracking
  - Registration, Range Data
- High-Level Vision
  - Detection, Classification, Recognition

Image Formation

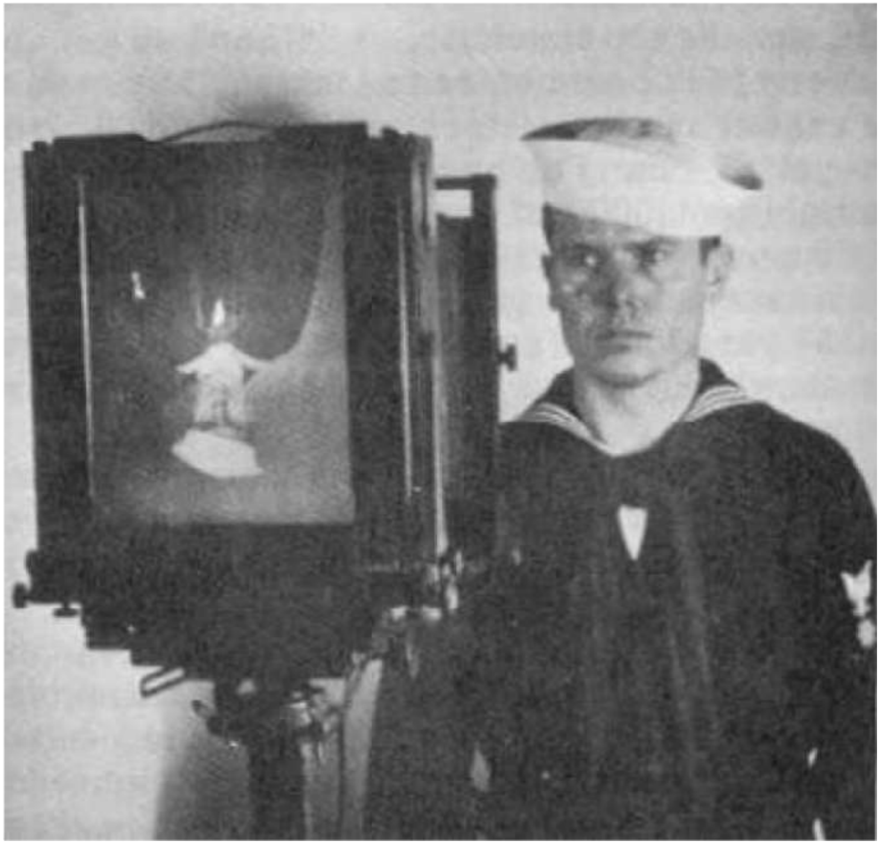


Image formation on the backplate of a photographic camera

Early Vision

edge

texture

region

contour and silhouette

local features

scale

Edge

- Sudden changes (discontinuities) in an image
- Where most shape information is encoded

Texture

flower

food

water

Local Features

Middle-Level Vision

edge

texture

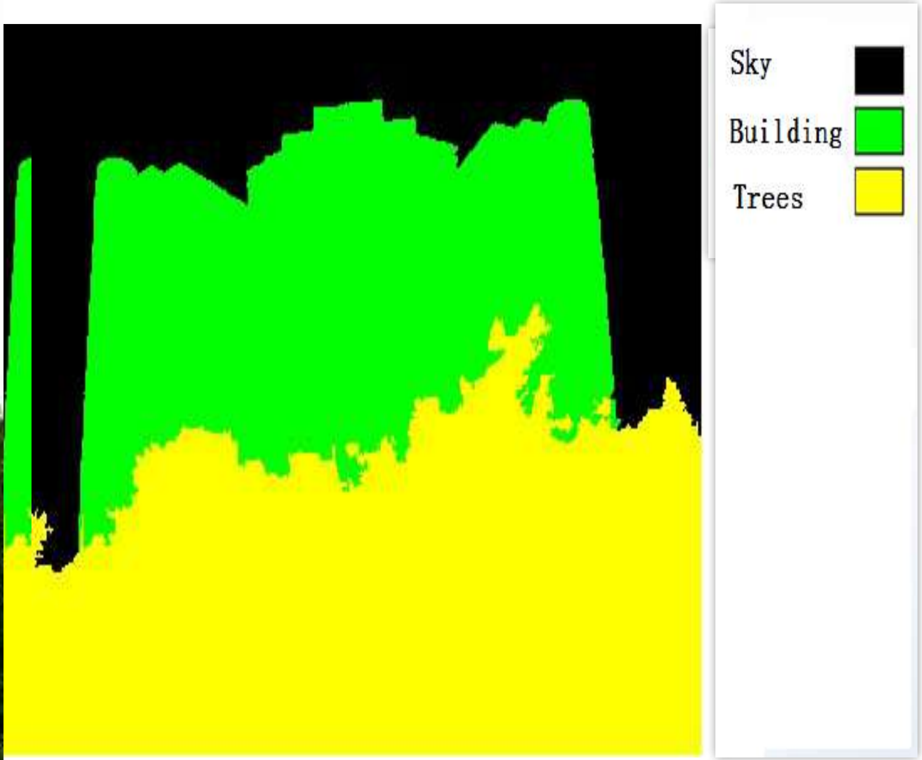
region

contour and silhouette

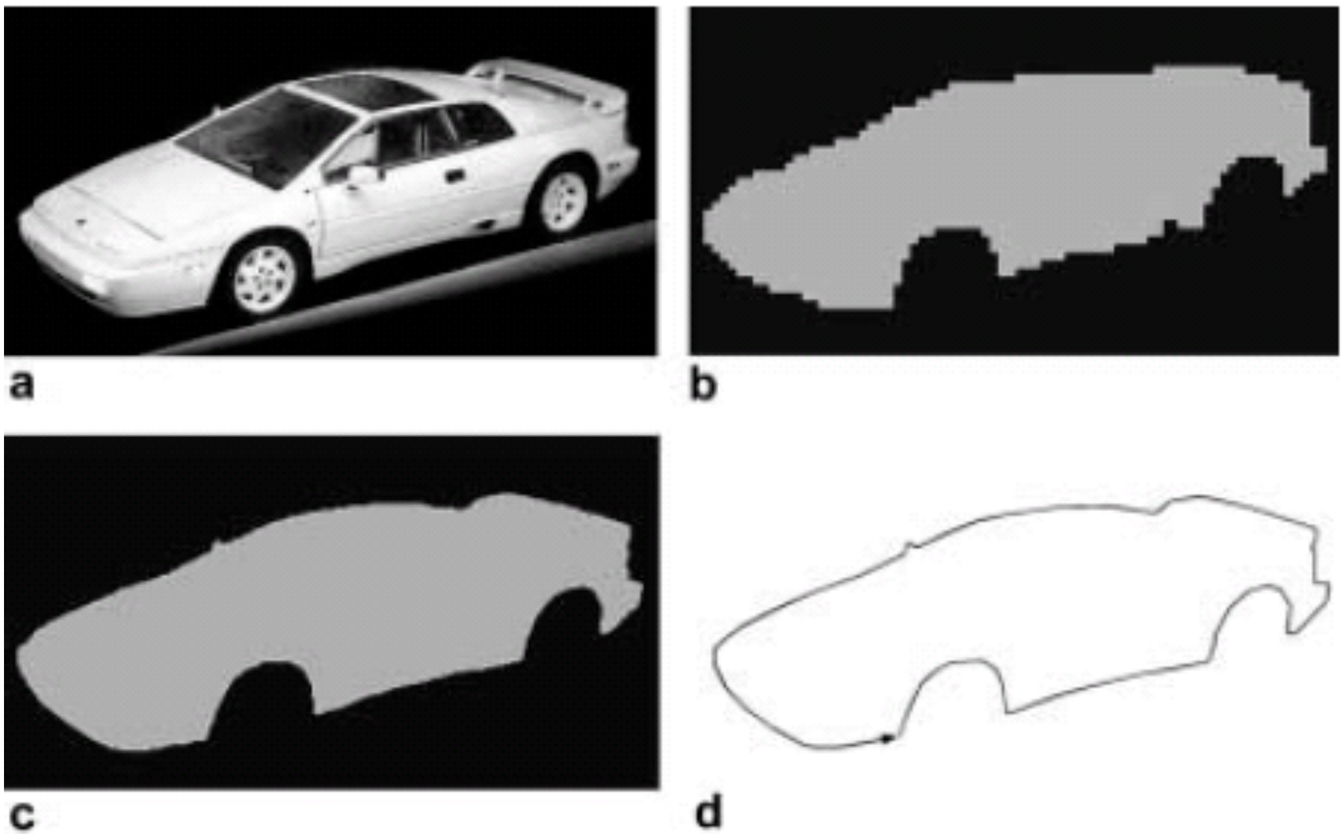
local features

scale

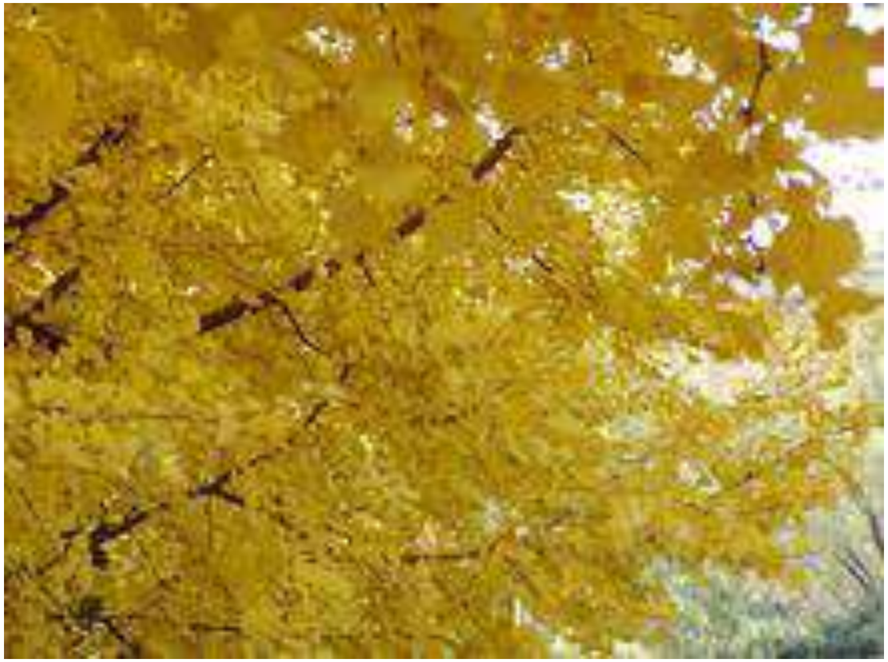
Region



Contour & Silhouette



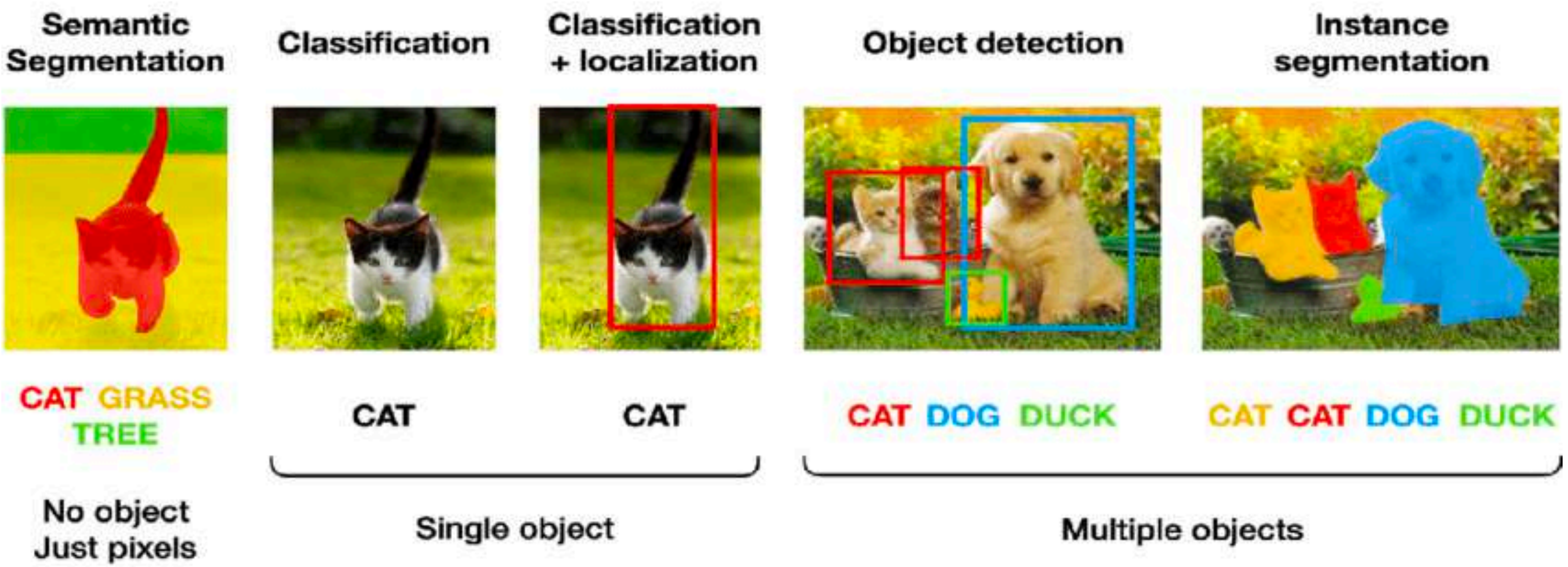
Scale



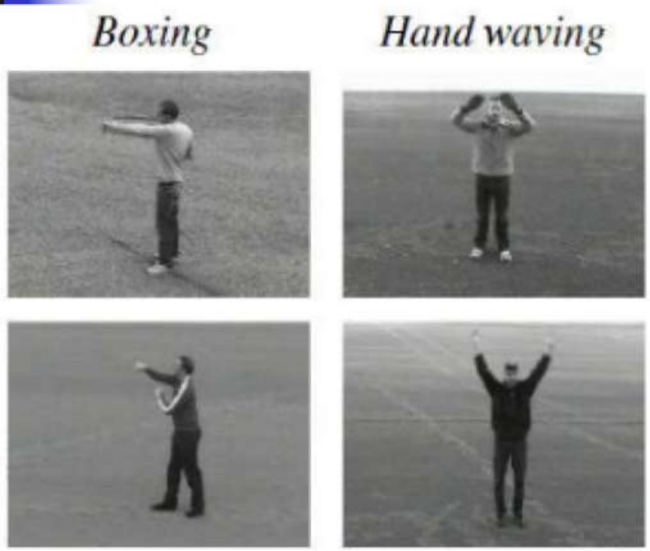
High-Level Vision

- Object Detection and Recognition
- Classification
- Activity Analysis

High-level task



High-level task



Action recognition

Group activity analysis

Event recognition

Beyond what is where

Functionality

What can you do with the tree trunk?

Physics

How likely is the stone balancing?

Intentionality

Why does the guy kick the door?

Causality

Who knocked down the domino?

Target of CV in HCI

- Give commands to a computer
- See with one or more cameras
- Stimulate by static or moving objects
- Understanding the information from images
- Have a feedback correctly

- Face
- Eyes
- Body

Face detection, recognition, verification

- Is a face present?/Where is the face?
- Who is that?
- Is the ID **?

the development stage of face recognition	related technologies	characteristics of different stages of face recognition
early algorithm stage	principal component analysis linear discriminate analysis	In the initial stage, face recognition is regarded as a general pattern recognition problem, and the mainstream technology is based on the geometric structure of the face.
artificial features and classifier stage	support vector machine adaboost small samples neural networks	In the climax stage, face recognition develops rapidly. At this stage, many classic methods and several commercialized face recognition systems appeared.
deep learning stage	neural networks deep learning	With the deepening of research on face recognition, researchers began to pay attention to the problem of face recognition facing real conditions.

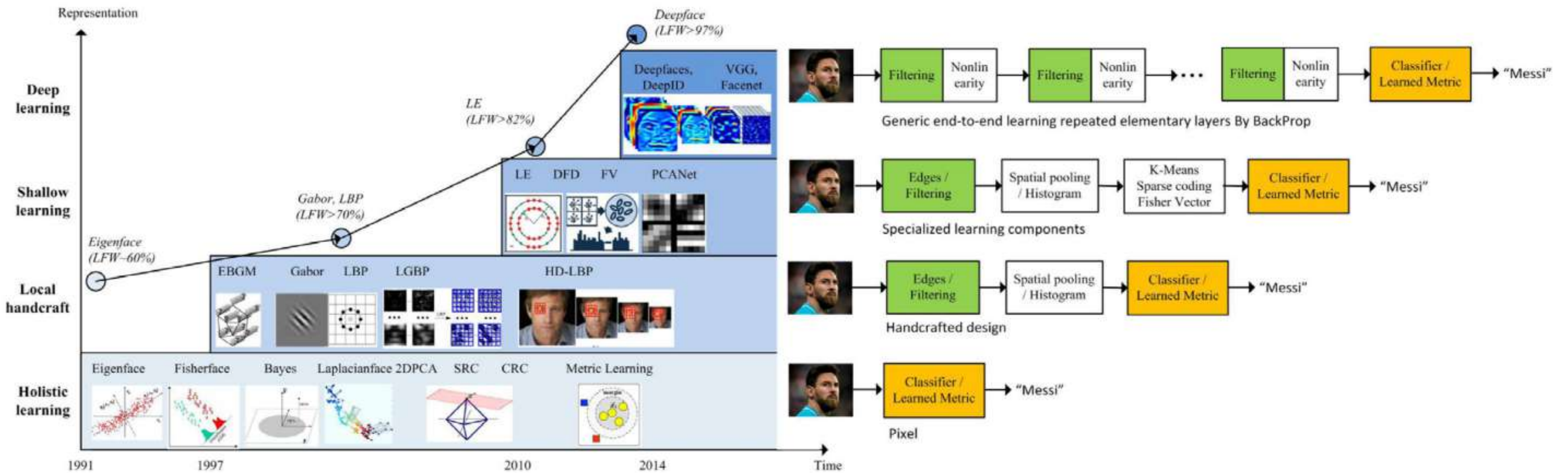
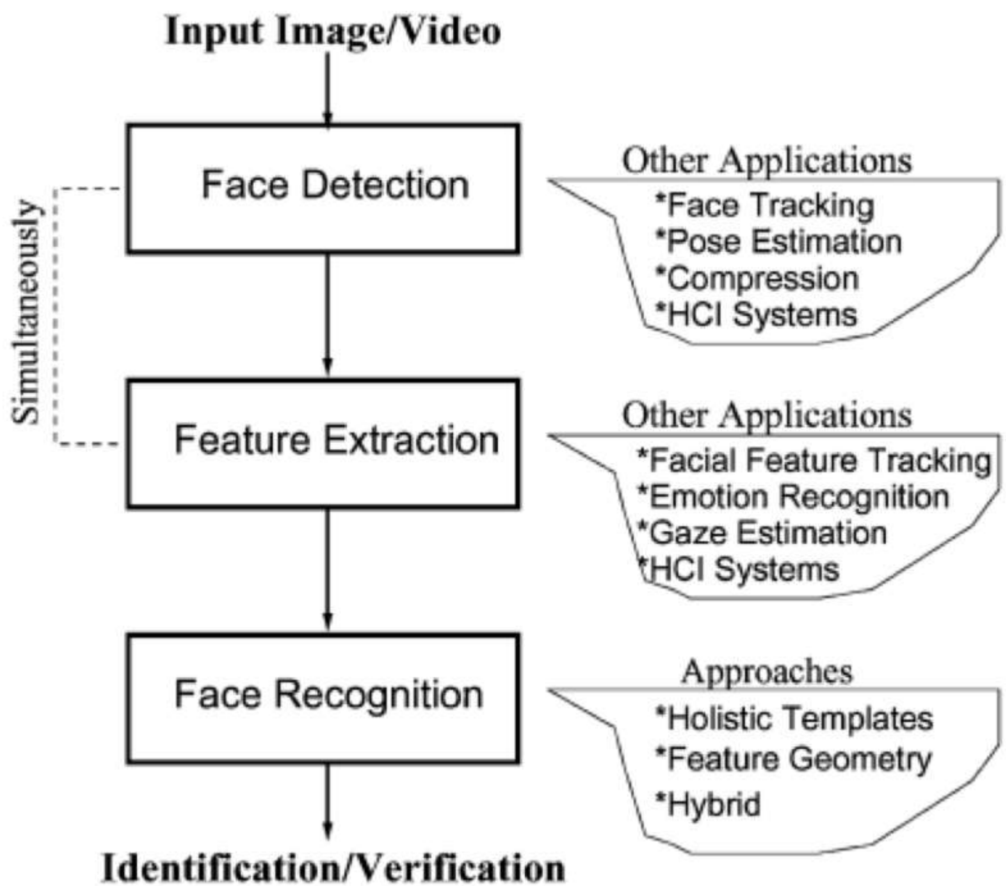
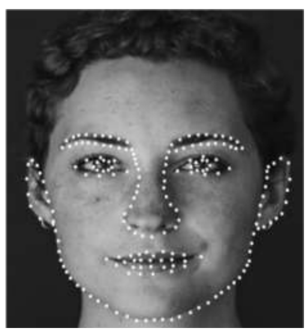
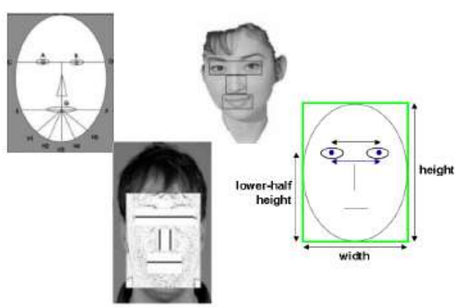


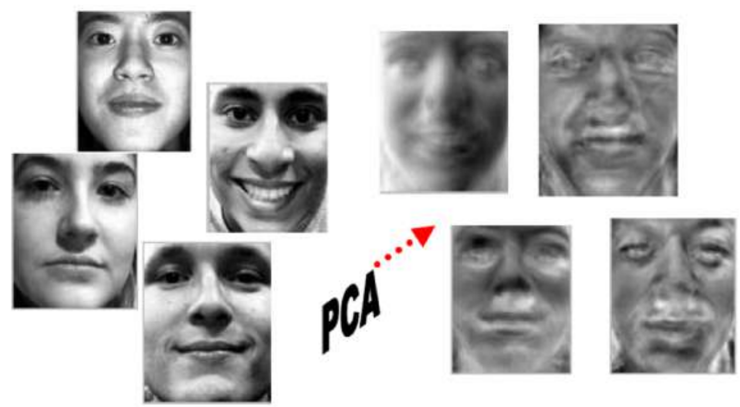
Fig. 1. Milestones of face representation for recognition. The holistic approaches dominated the face recognition community in the 1990s. In the early 2000s, handcrafted local descriptors became popular, and the local feature learning approaches were introduced in the late 2000s. In 2014, DeepFace [20] and DeepID [21] achieved a breakthrough on state-of-the-art (SOTA) performance, and research focus has shifted to deep-learning-based approaches. As the representation pipeline becomes deeper and deeper, the LFW (Labeled Face in-the-Wild) performance steadily improves from around 60% to above 90%, while deep learning boosts the performance to 99.80% in just three years.


Input Image/Video
Face Detection
Feature Extraction
Face Recognition
Identification/Verification
Simultaneously
Other Applications
*Face Tracking
*Pose Estimation
*Compression
*HCI Systems
Other Applications
*Facial Feature Tracking
*Emotion Recognition
*Gaze Estimation
*HCI Systems
Approaches
*Holistic Templates
*Feature Geometry
*Hybrid

- Feature-Based Approaches
  - Active Shape Model
    - Snakes [10–13]
    - Deformable Template Matching [14–16]
    - Deformable Part Model [17,18]
    - Point Distribution Model [19–21]
  - Low Level Analysis
    - Motion [22–26]
    - Color [27–30]
    - Gray Information [31]
    - Edges [32–37]
  - Feature Analysis
    - Feature Searching [38–40]
      - Viola-Jones [41–44]
        - Local Binary Pattern (LBP) [45–47]
        - AdaBoost [48]
      - Gabour Feature [49–51]
    - Constellation Analysis [52–56]







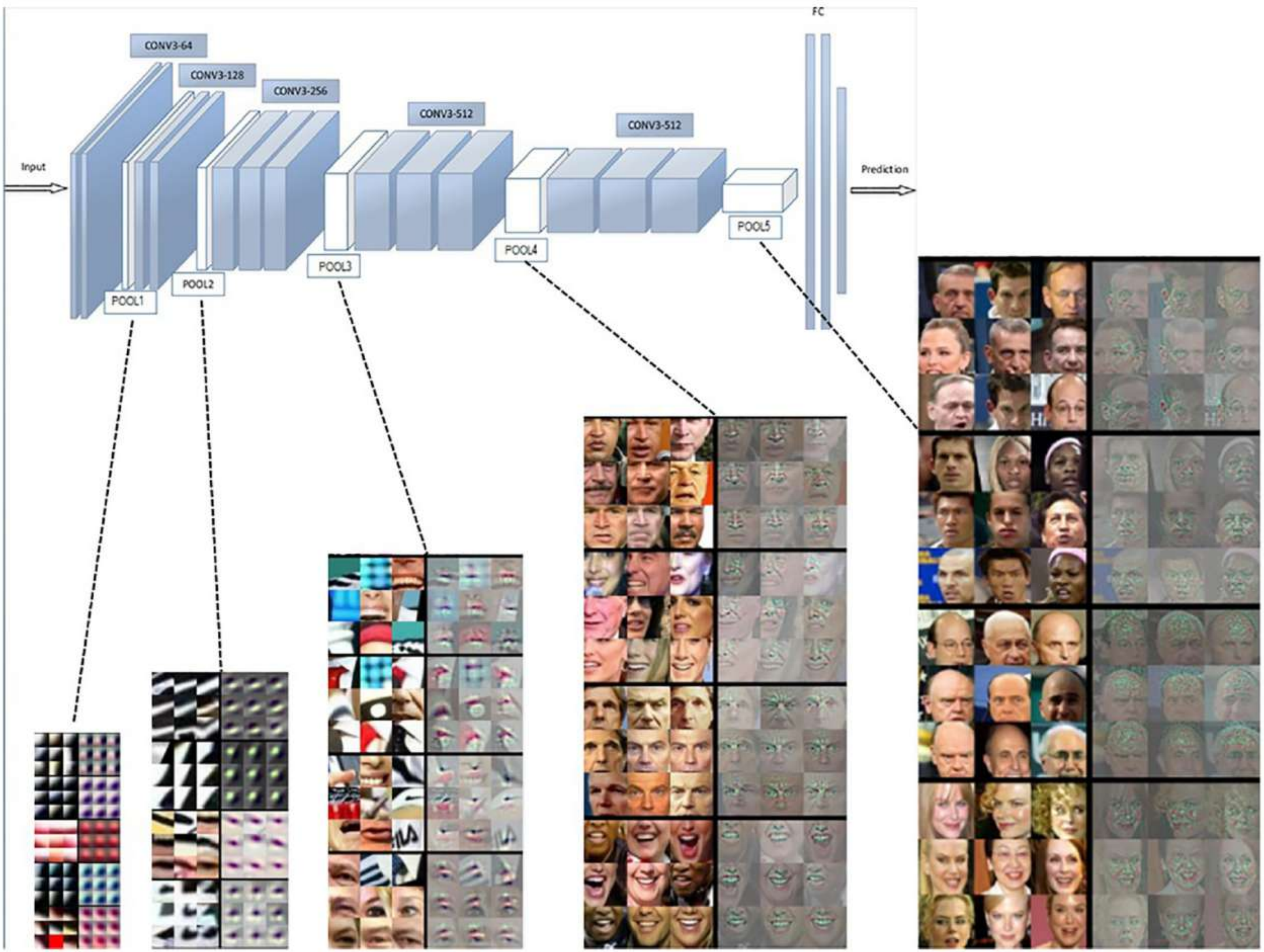
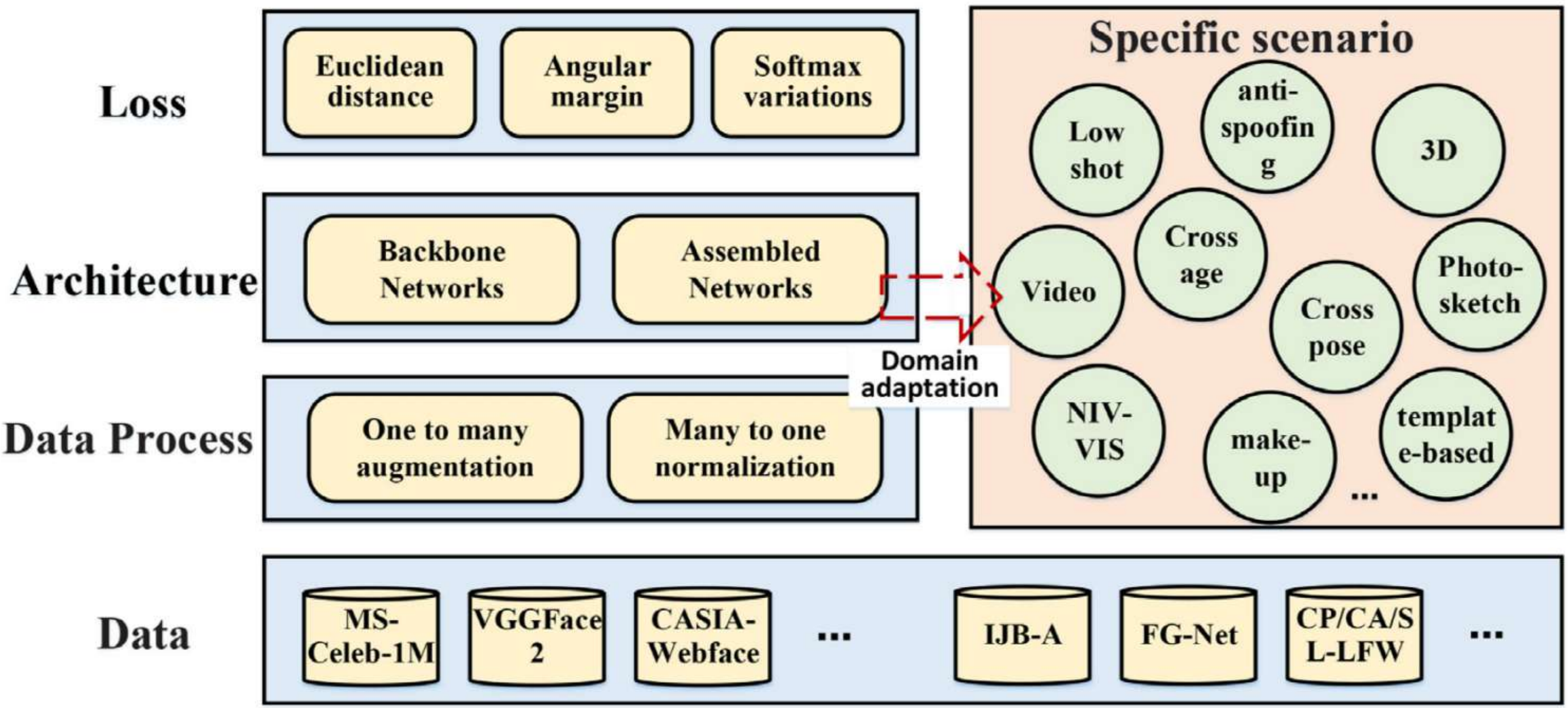
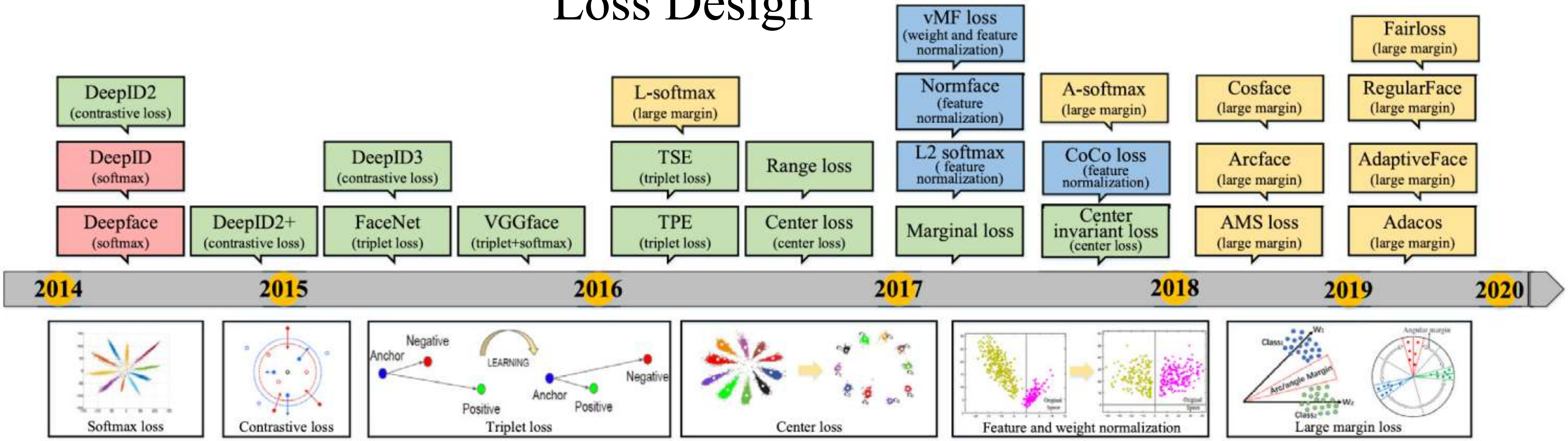


Fig. 2. The hierarchical architecture that stitches together pixels into invariant face representation. Deep model consists of multiple layers of simulated neurons that convolute and pool input, during which the receptive-field size of simulated neurons are continually enlarged to integrate the low-level primary elements into multifarious facial attributes, finally feeding the data forward to one or more fully connected layer at the top of the network. The output is a compressed feature vector that represent the face. Such deep representation is widely considered as the SOTA technique for face recognition.

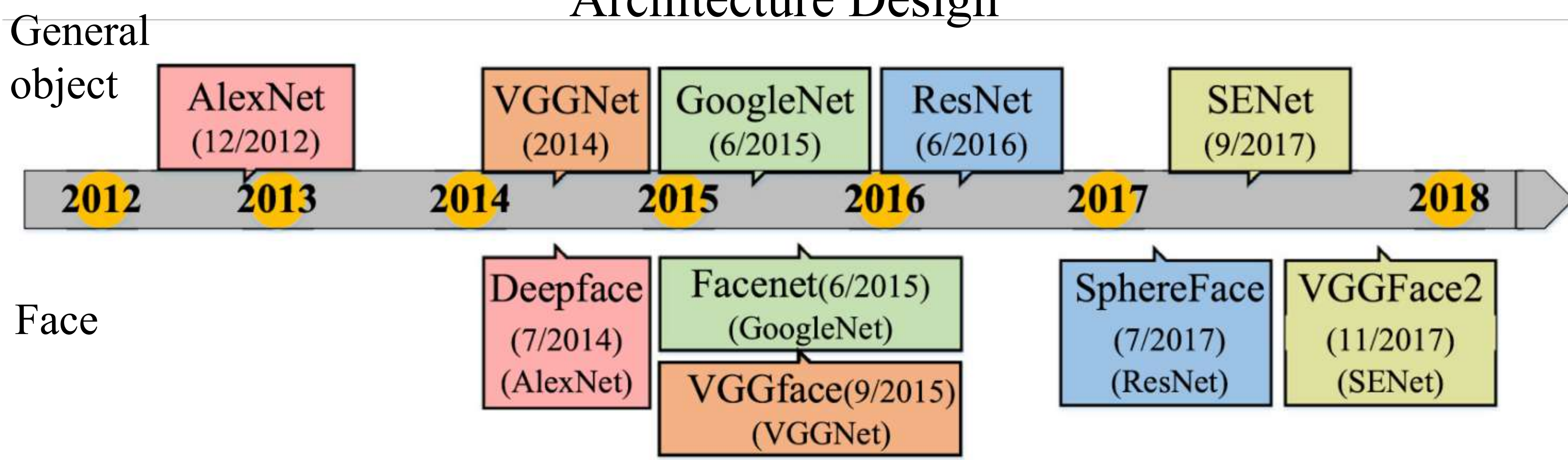
General->Specific



Loss Design



Architecture Design



Loss Design

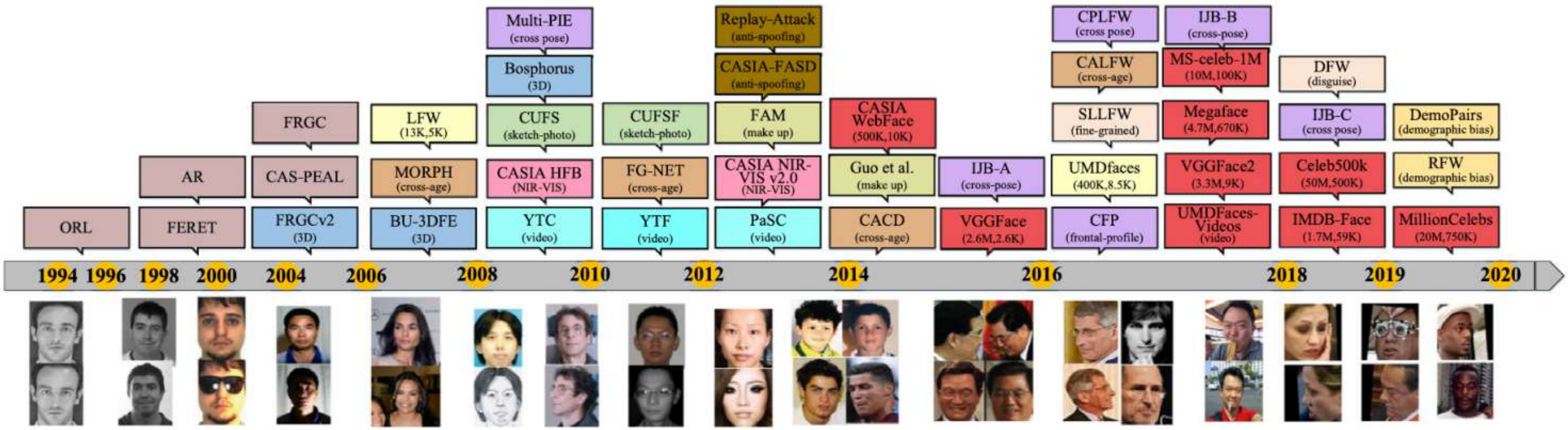


Fig. 17. The evolution of FR datasets. Before 2007, early works in FR focused on controlled and small-scale datasets. In 2007, LFW [23] dataset was introduced which marks the beginning of FR under unconstrained conditions. Since then, more testing databases designed for different tasks and scenes are constructed. And in 2014, CASIA-Webface [120] provided the first widely-used public training dataset, large-scale training datasets begun to be hot topic. Red rectangles represent training datasets, and other color rectangles represent different testing datasets.

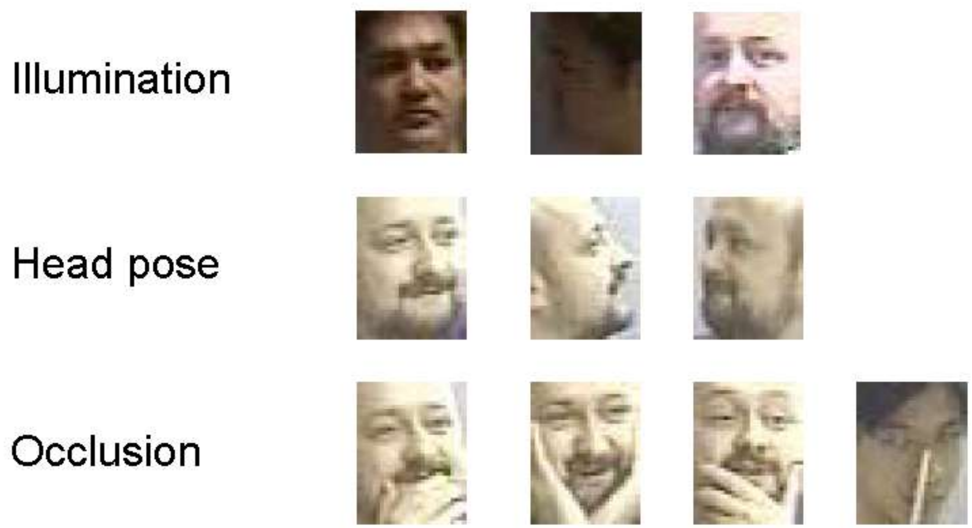
Face Recognition: Advantages

- Photos of faces are widely used in passports and driver's licenses where the possession authentication protocol is augmented with a photo for manual inspection purposes; there is wide public acceptance for this biometric identifier
- Face recognition systems are the least intrusive from a biometric sampling point of view, requiring no contact, nor even the awareness of the subject
- Face recognition can, at least in theory, be used for screening of unwanted individuals in a crowd, in real time
- It is a fairly good biometric identifier for small-scale verification applications

Face Recognition: Disadvantages

Suffer of:

- Pose
- Appearance
- Age
- Lighting
- Expression
- Ethics



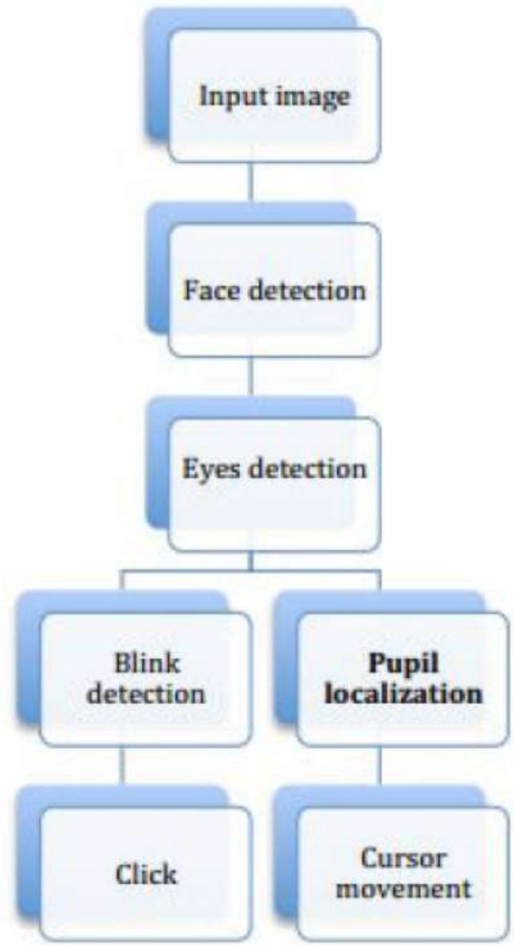
Eyes

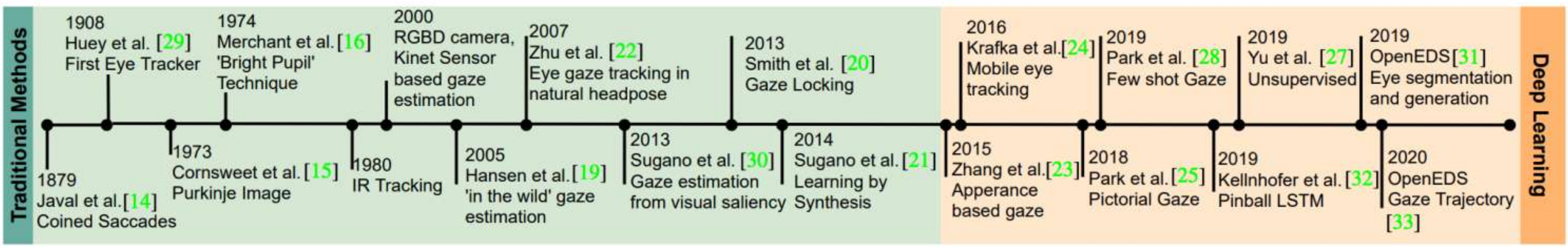
(a) Tobii Pro Glasses 2

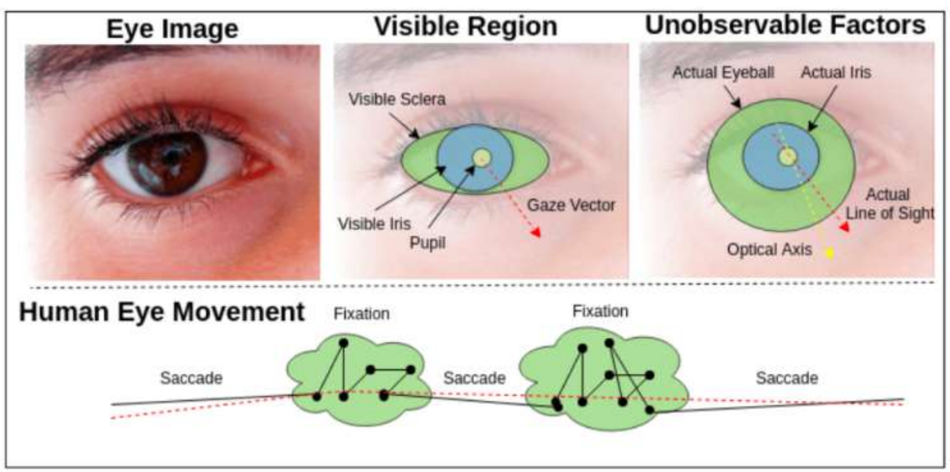
(b) Gaze Estimation obtained with Kinect

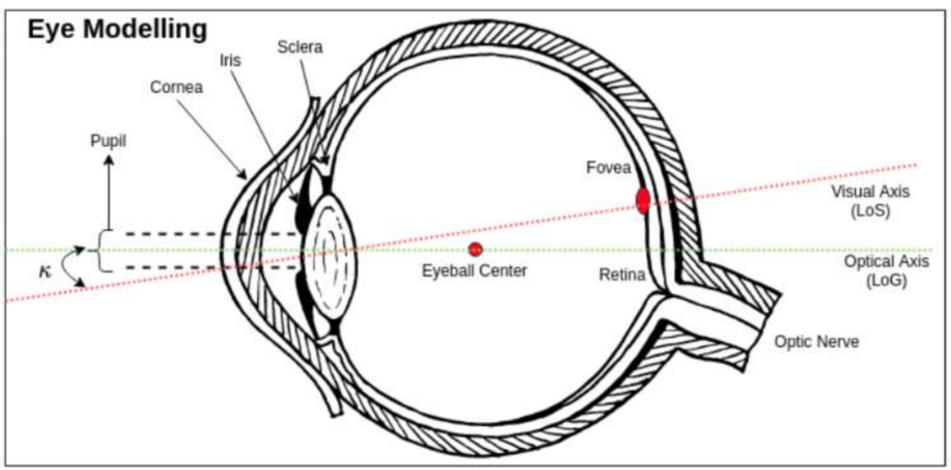
(c) EyeLink 1000 Plus

(d) EyeSee software

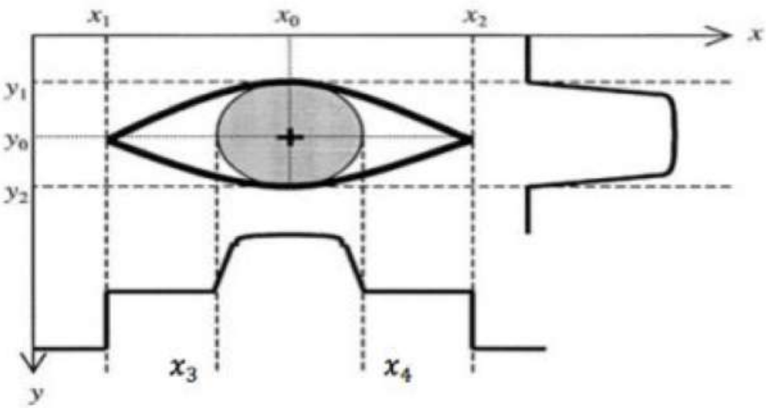


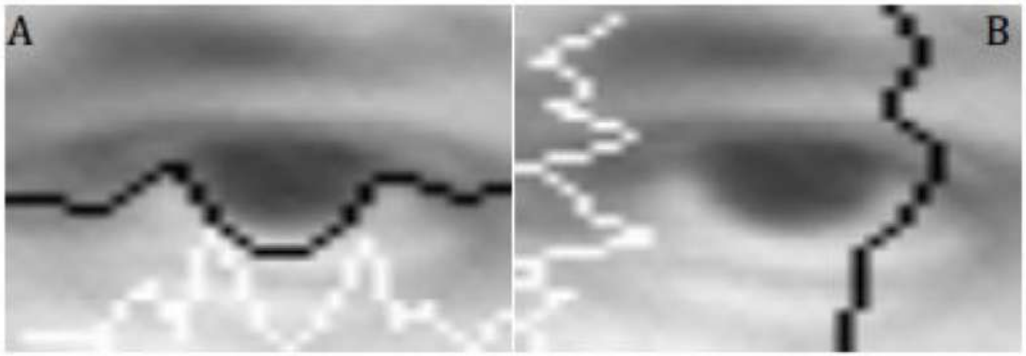

Traditional Methods
1908
Huey et al. [29]
First Eye Tracker
1974
Merchant et al. [16]
'Bright Pupil'
Technique
2000
RGBD camera,
Kinet Sensor
based gaze
estimation
2007
Zhu et al. [22]
Eye gaze tracking in
natural headpose
2013
Smith et al. [20]
Gaze Locking
1879
Javal et al. [14]
Coined Saccades
1973
Cornsweet et al. [15]
Purkinje Image
1980
IR Tracking
2005
Hansen et al. [19]
'in the wild' gaze
estimation
2013
Sugano et al. [30]
Gaze estimation
from visual saliency
2014
Sugano et al. [21]
Learning by
Synthesis
2016
Krafka et al. [24]
Mobile eye
tracking
2019
Park et al. [28]
Few shot Gaze
2019
Yu et al. [27]
Unsupervised
2019
OpenEDS [31]
Eye segmentation
and generation
2015
Zhang et al. [23]
Apperance
based gaze
2018
Park et al. [25]
Pictorial Gaze
2019
Kellnhofer et al. [32]
Pinball LSTM
2020
OpenEDS
Gaze Trajectory
[33]
Deep Learning

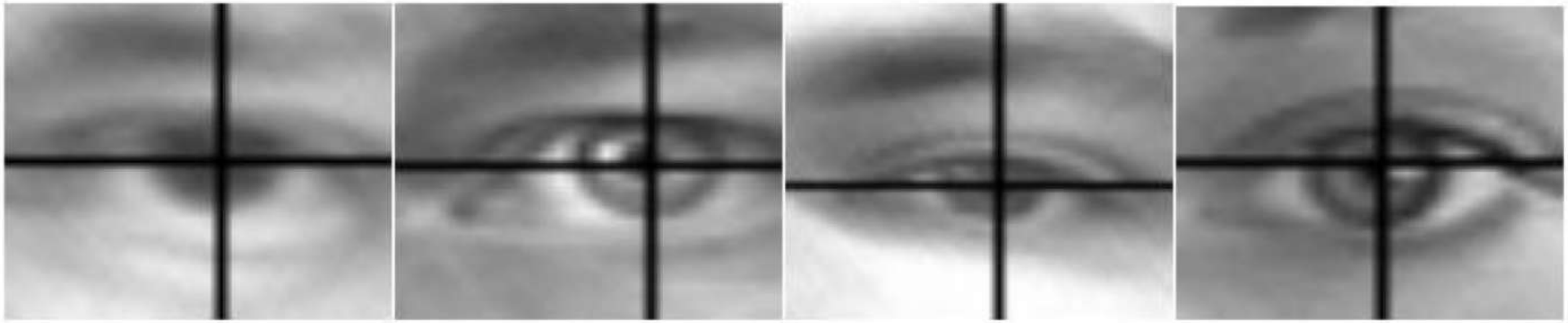

Eye Image
Visible Region
Unobservable Factors
Visible Sclera
Visible Iris
Pupil
Gaze Vector
Actual Eyeball
Actual Iris
Actual
Line of Sight
Optical Axis
Human Eye Movement
Fixation
Fixation
Saccade
Saccade
Saccade


Eye Modelling
Sclera
Iris
Cornea
Pupil
Fovea
Visual Axis
(LoS)
Optical Axis
(LoG)
κ
Eyeball Center
Retina
Optic Nerve

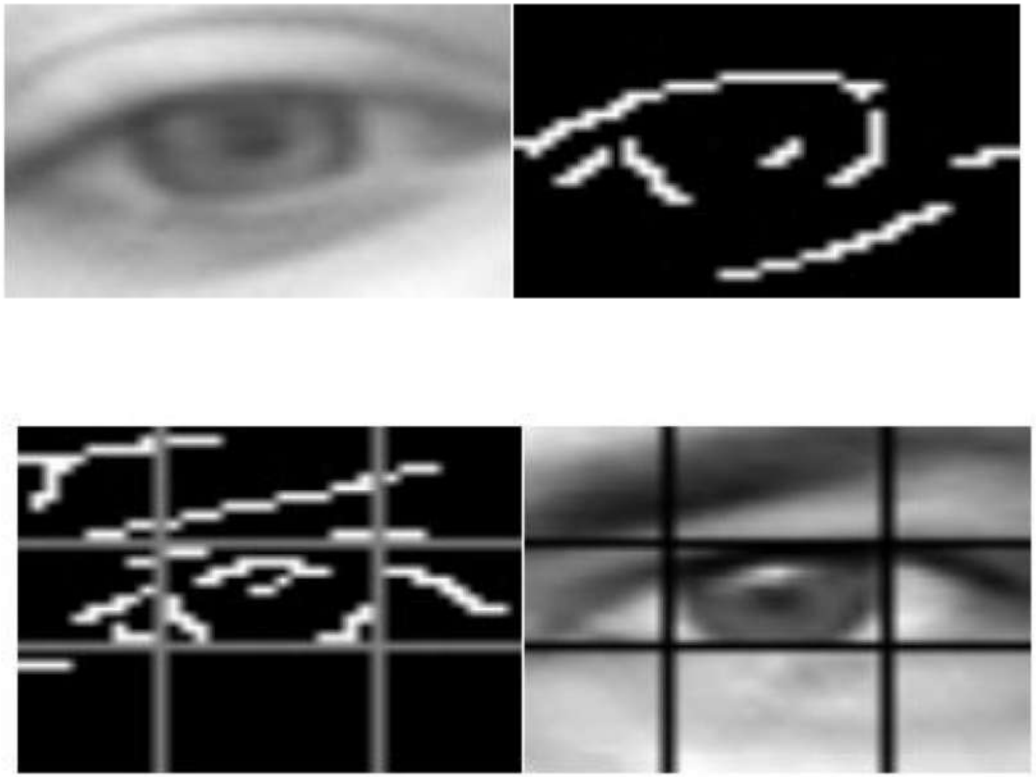
Projection Function

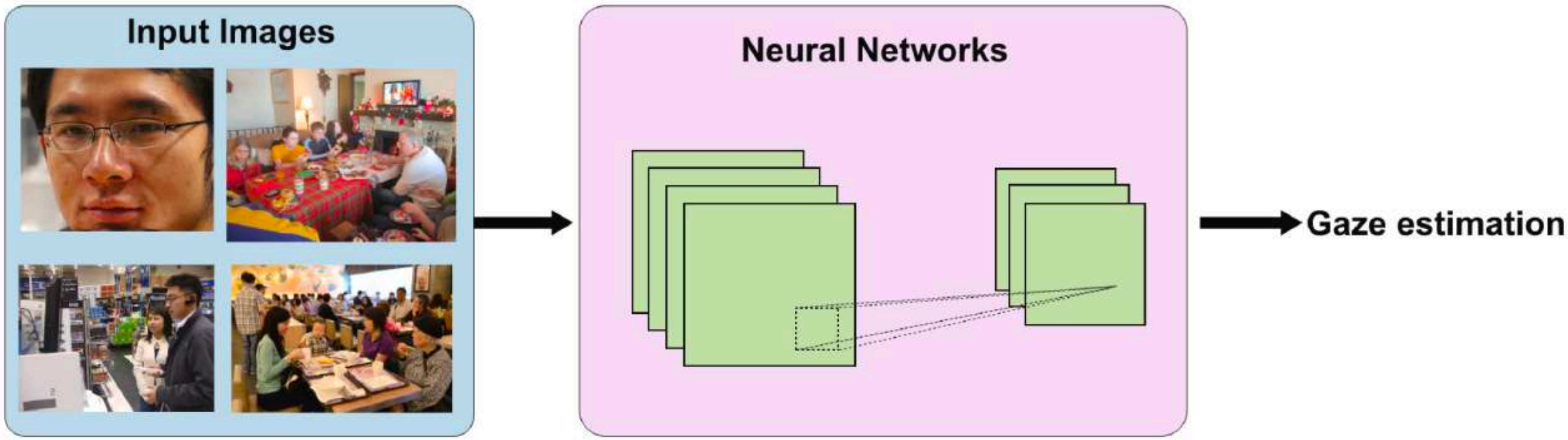






Edge Analysis

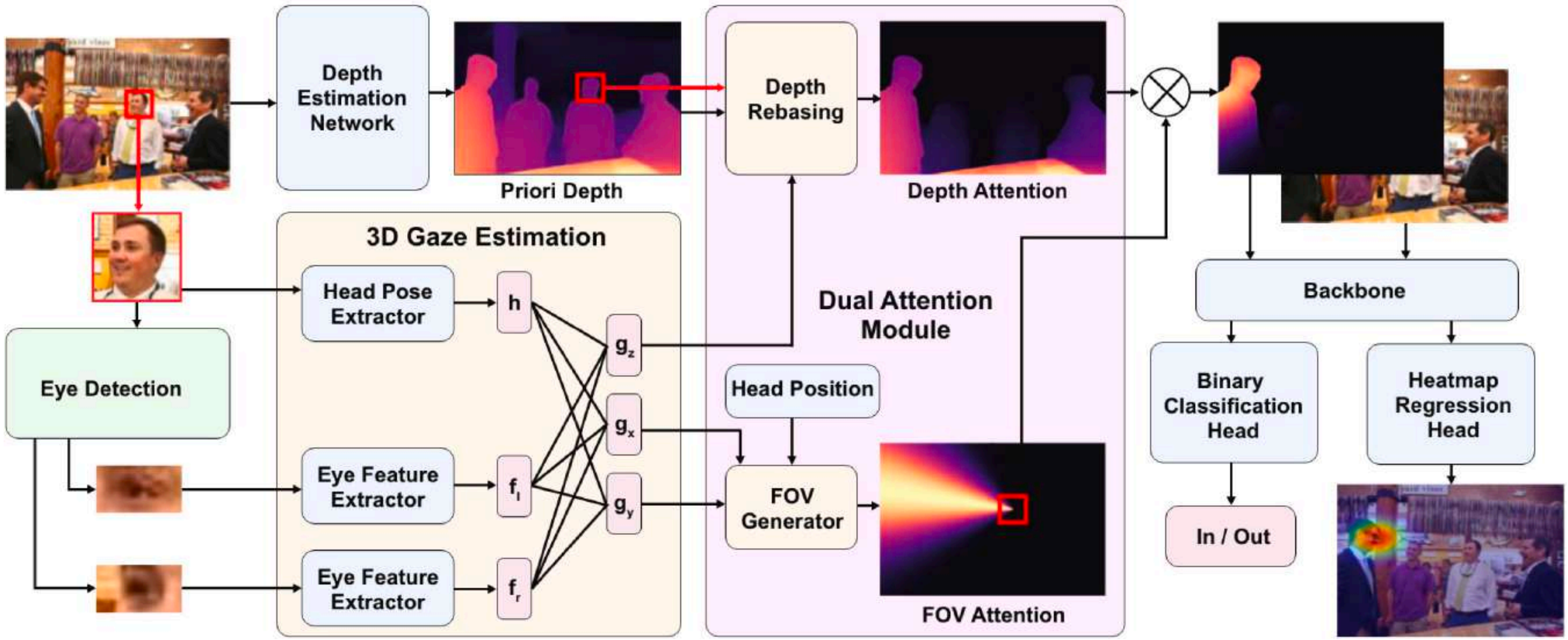




Framework

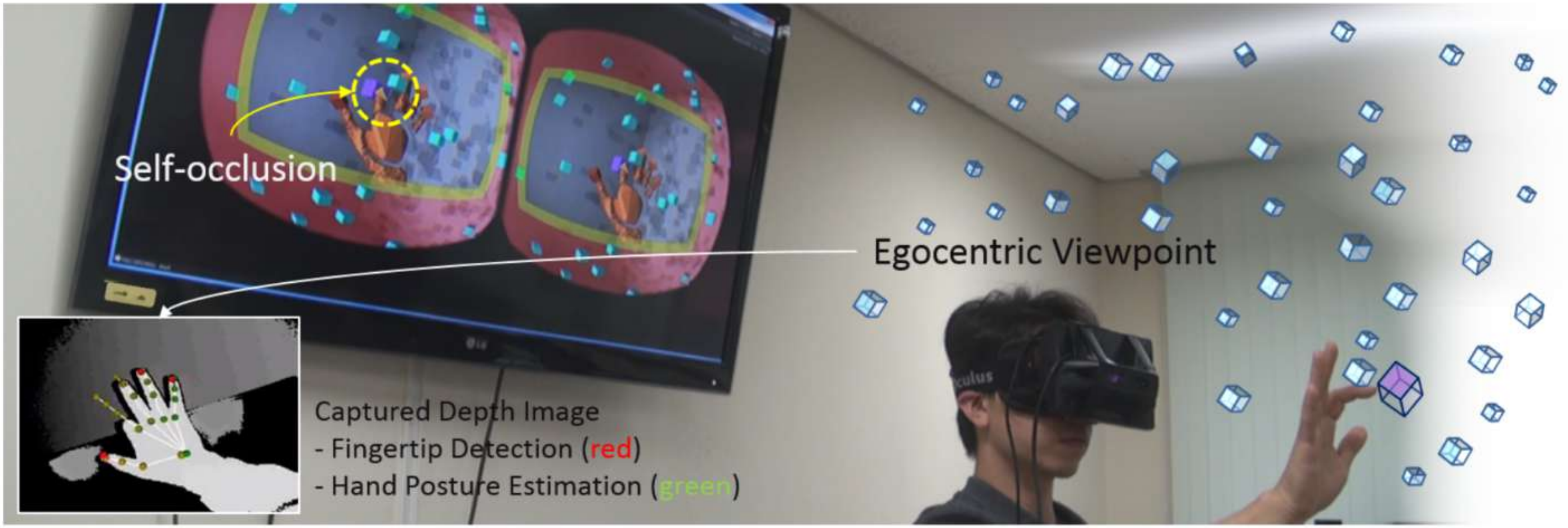
(a) MPIIGaze (b) Columbia Gaze (c) Gaze360 (d) GazeFollow (e) Gaze on Object

Dataset



Fang, Y., Tang, J., Shen, W., Shen, W., Gu, X., Song, L., & Zhai, G. (2021). Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11390–11399).

Hand and hand gesture

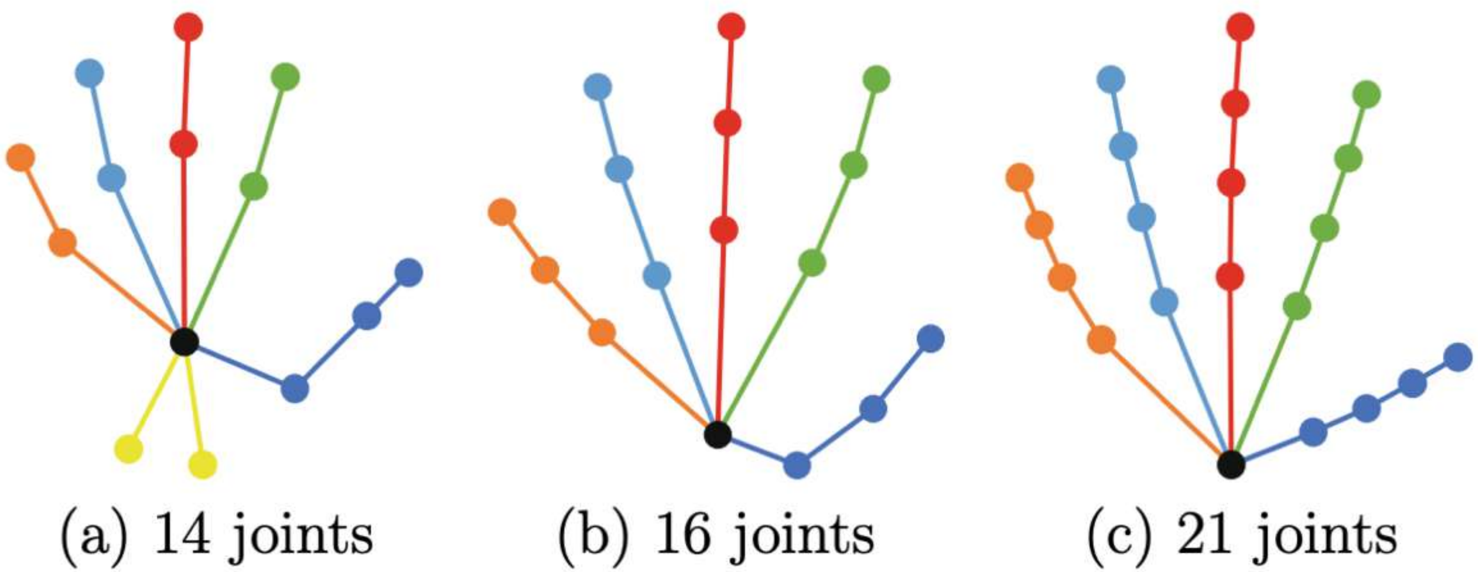


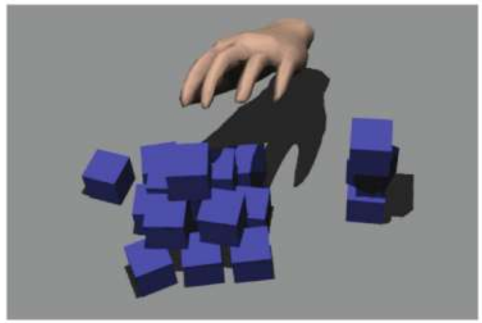
Hand and hand gesture

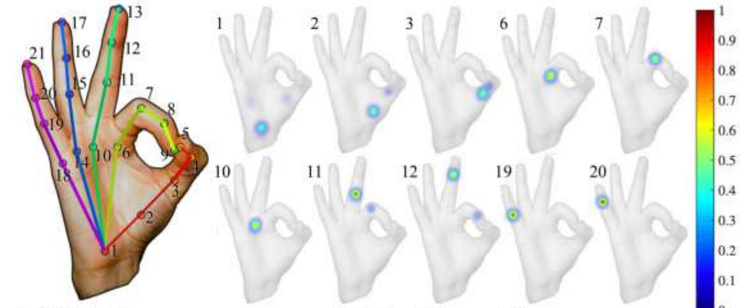
- Appearance based approach

  Finger tips, Silhouette

- Model based approach

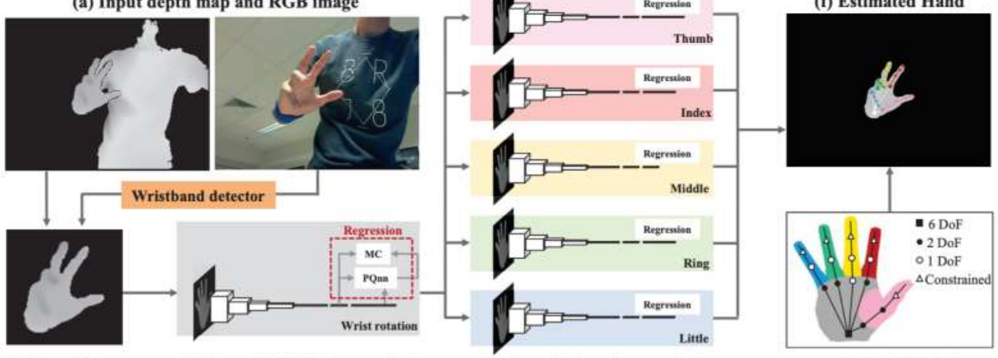




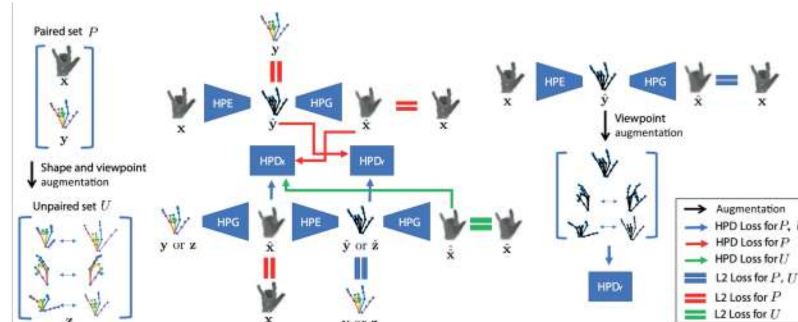


(a) Detections (b) Confidence Maps

(a) Input depth map and RGB image

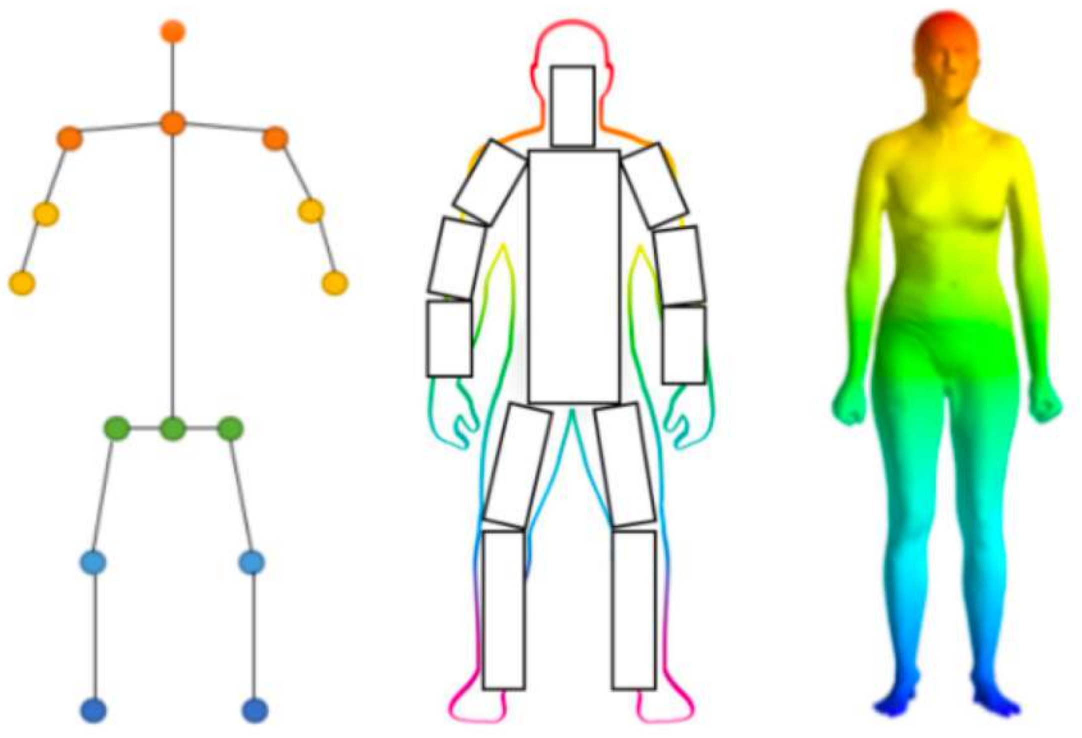


(b) Depth image (c) Stage 1 (Global regression) (d) Stage 2 (Local regression) (e) Hand Model

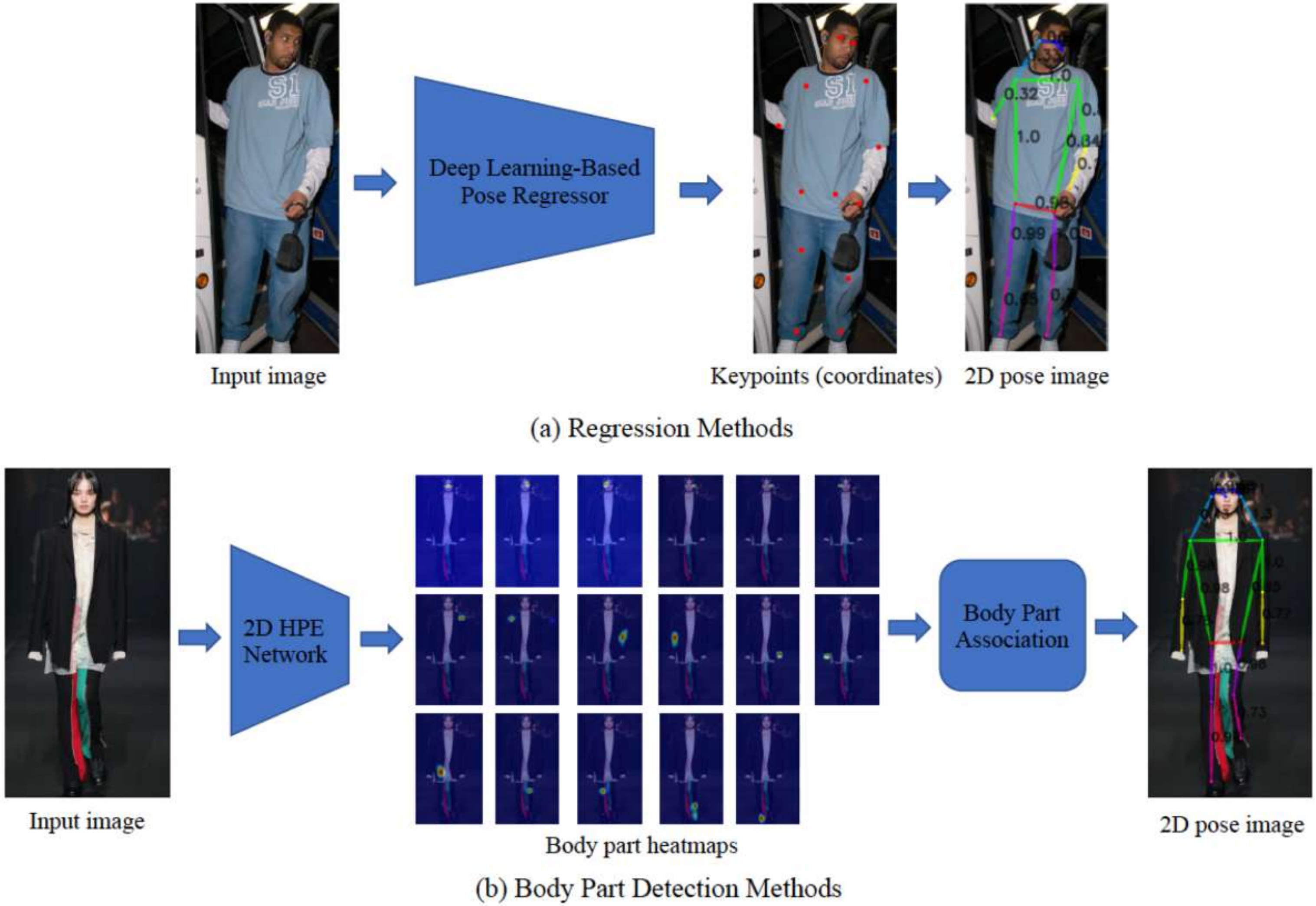


(a) Skeleton augmentation (b) Training stage (c) Testing stage

Body and body gesture



(a) Kinematic (b) Planar (c) Volumetric



Single-person 2D HPE frameworks. (a) Regression methods directly learn a mapping (via a deep neural network) from the original image to the kinematic body model and produce joint coordinates. (b) Body part detection methods predict body joint locations using the supervision of heatmaps.

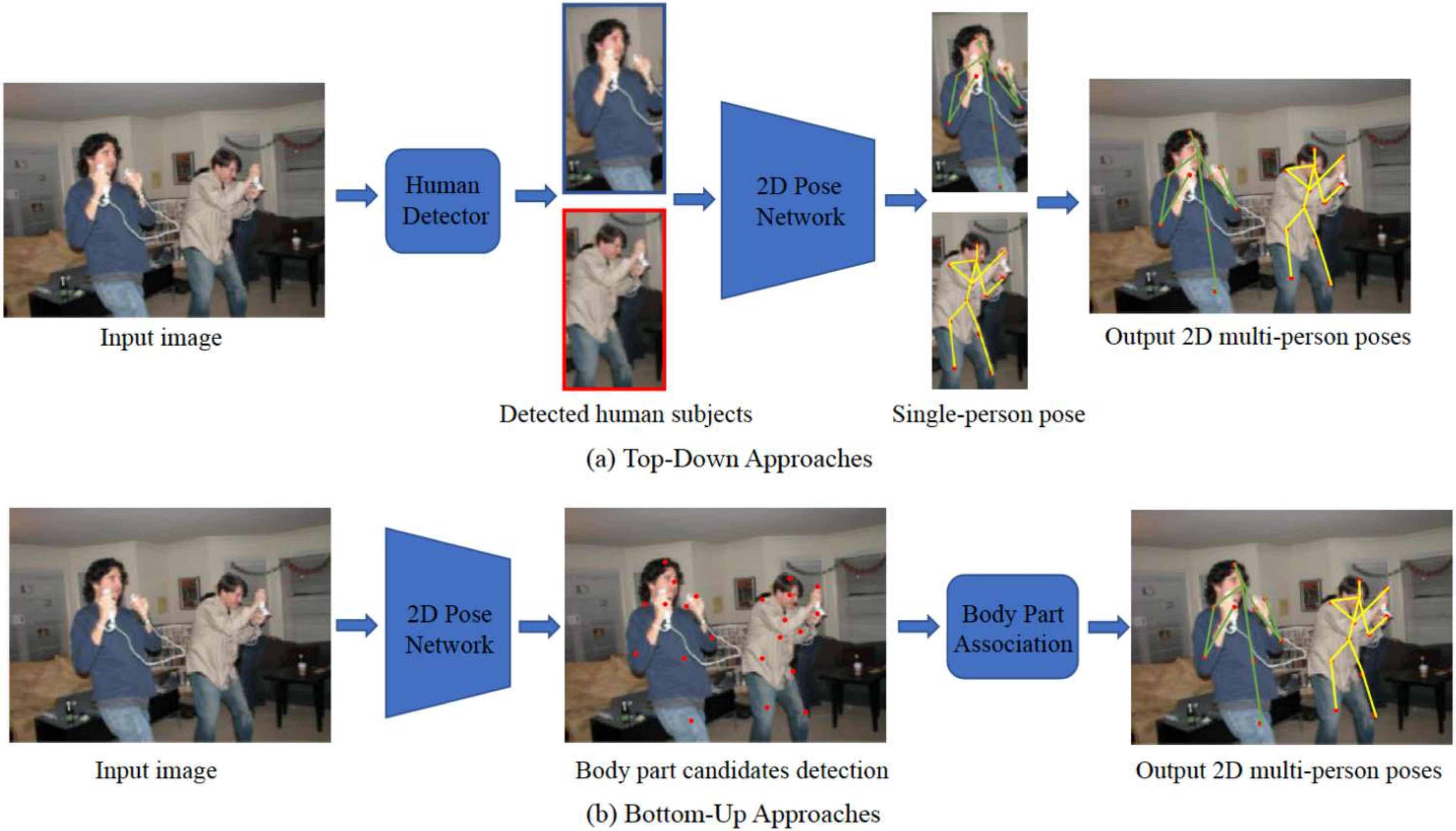
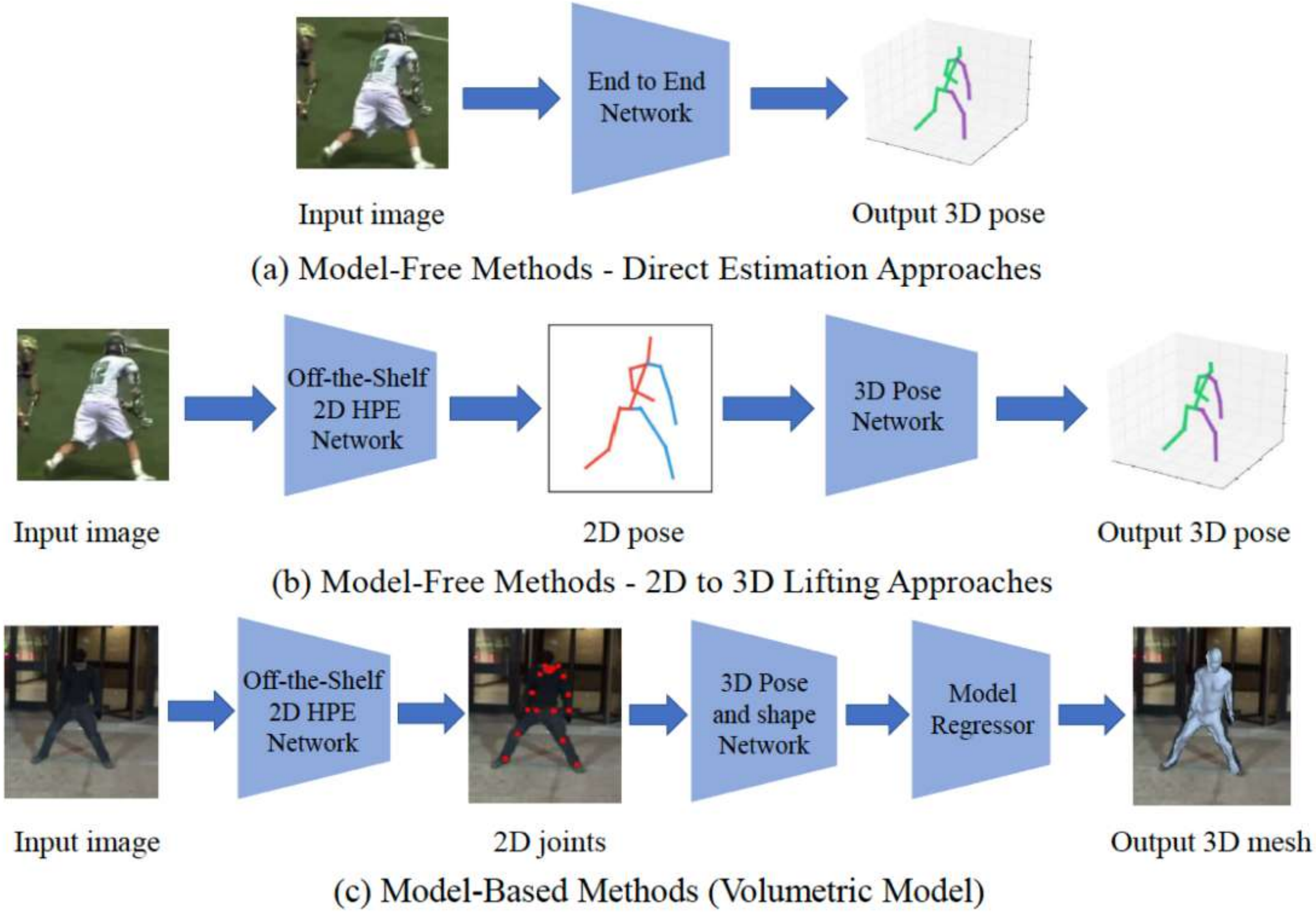


Illustration of the multi-person 2D HPE frameworks. (a) Top-down approaches have two sub-tasks: (1) humandetection and (2) pose estimation in the region of a singe human; (b) Bottom-up approaches also have two sub-tasks: (1) detect all keypoints candidates of body parts and (2) associate body parts in different human bodies and assemble them into individual pose representations.

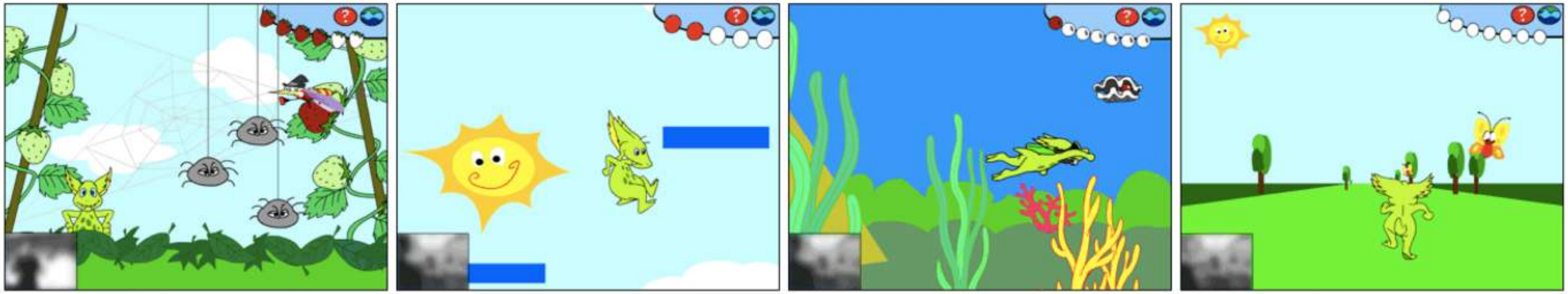


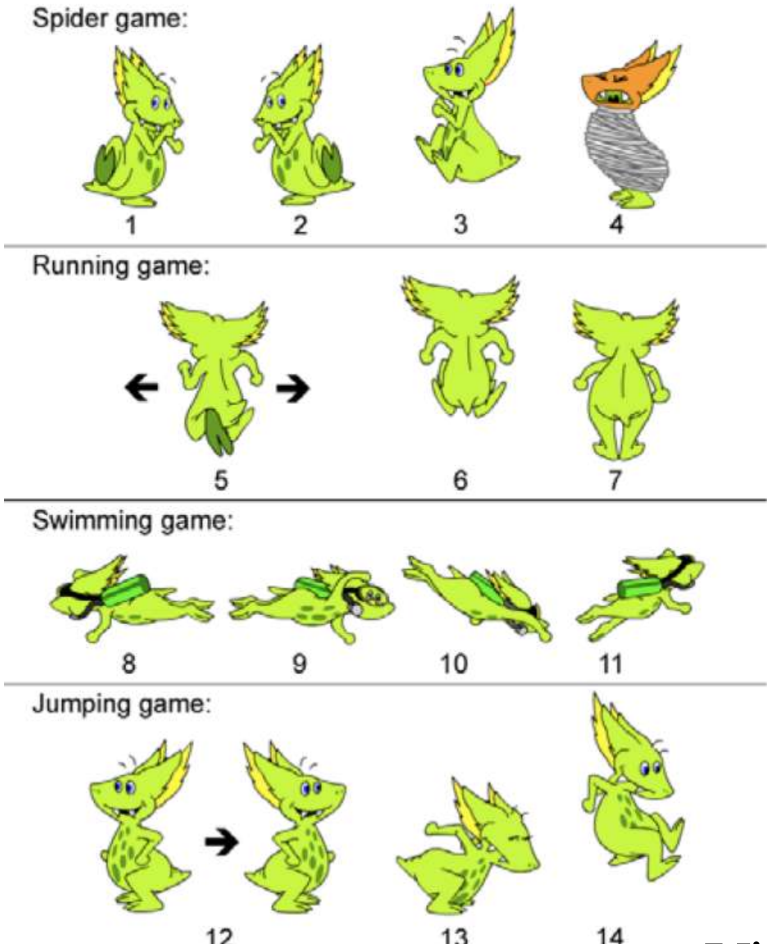
Single-person 3D HPE frameworks. (a) Direct estimation approaches directly estimate the 3D human pose from 2D images. (b) 2D to 3D lifting approaches leverage the predicted 2D human pose (intermediate representation) for 3D pose estimation. (c) Model-based methods incorporate parametric body models to recover high-quality 3D human mesh. The 3D pose and shape parameters inferred by the 3D pose and shape network are fed into the model regressor to reconstruct 3D human mesh.

Use Cases

People with disability

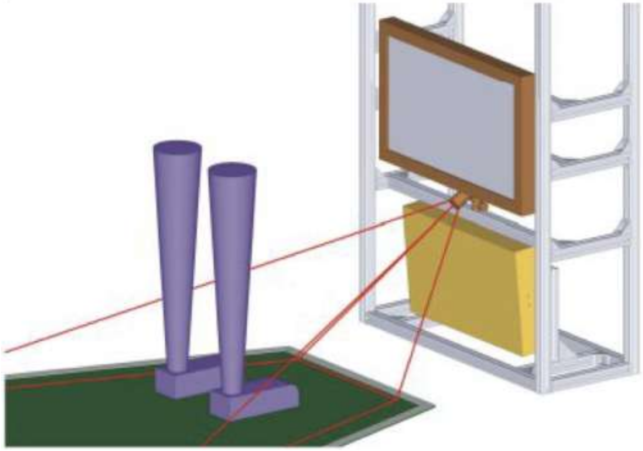
- Blind and visually impaired people
- Deaf and hearing impaired people
- Autism Spectrum Disorder



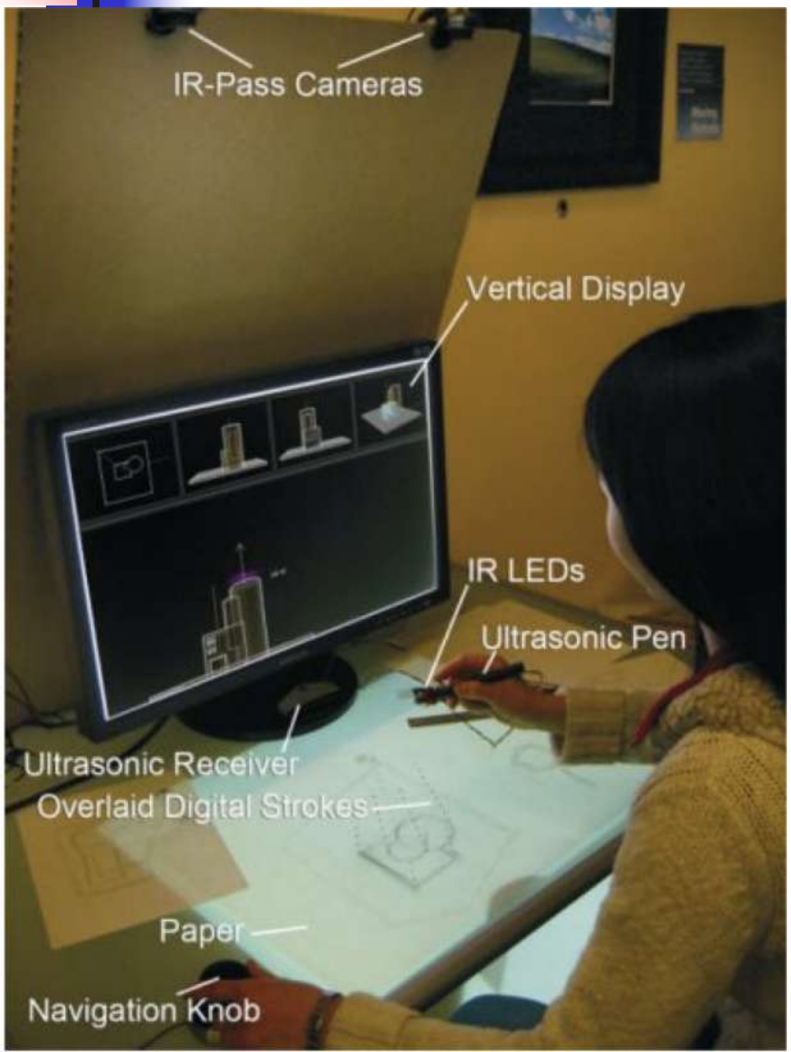

Spider game:
1
2
3
4
Running game:
5
6
7
Swimming game:
8
9
10
11
Jumping game:
12
13
14

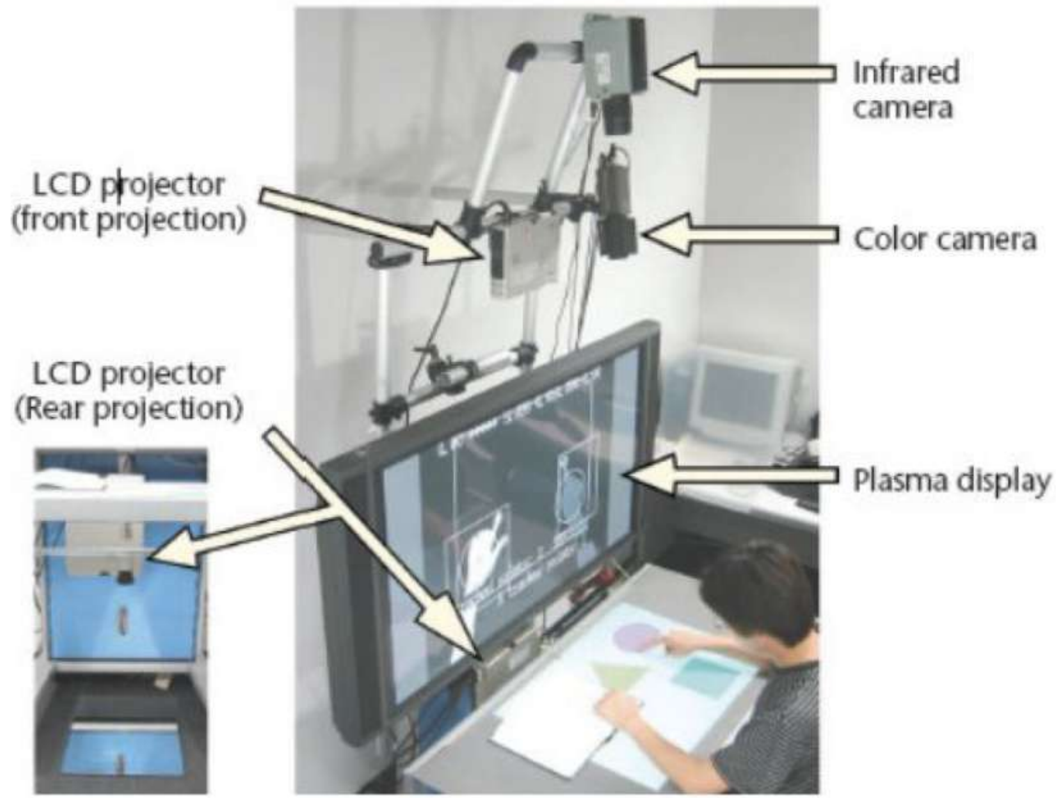
Video games

Shopping

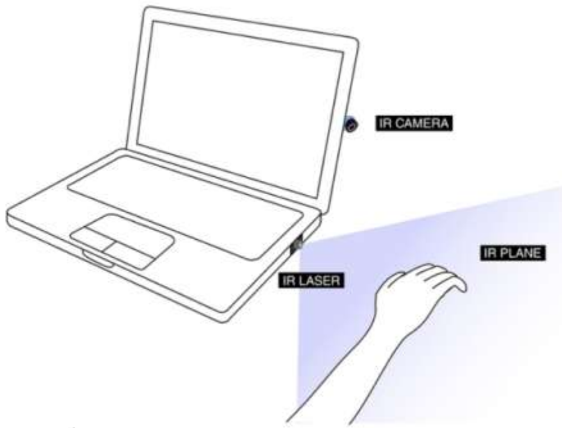


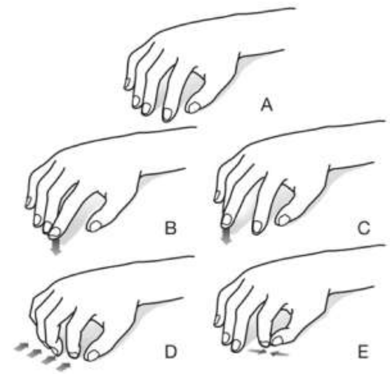
Office


IR-Pass Cameras
Vertical Display
IR LEDs
Ultrasonic Pen
Ultrasonic Receiver
Overlaid Digital Strokes
Paper
Navigation Knob


LCD projector
(front projection)
LCD projector
(Rear projection)
Infrared
camera
Color camera
Plasma display

Virtual input devices



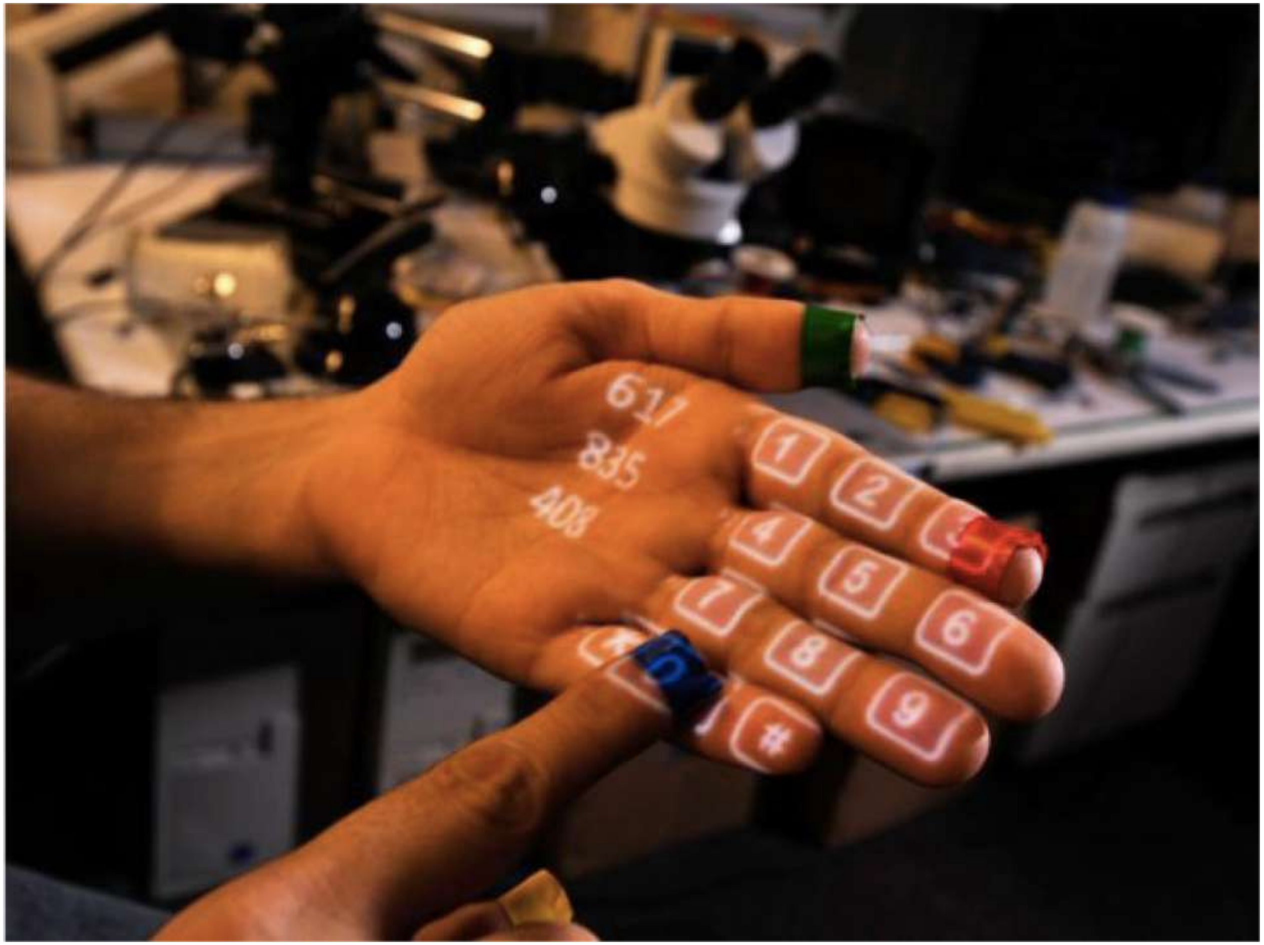


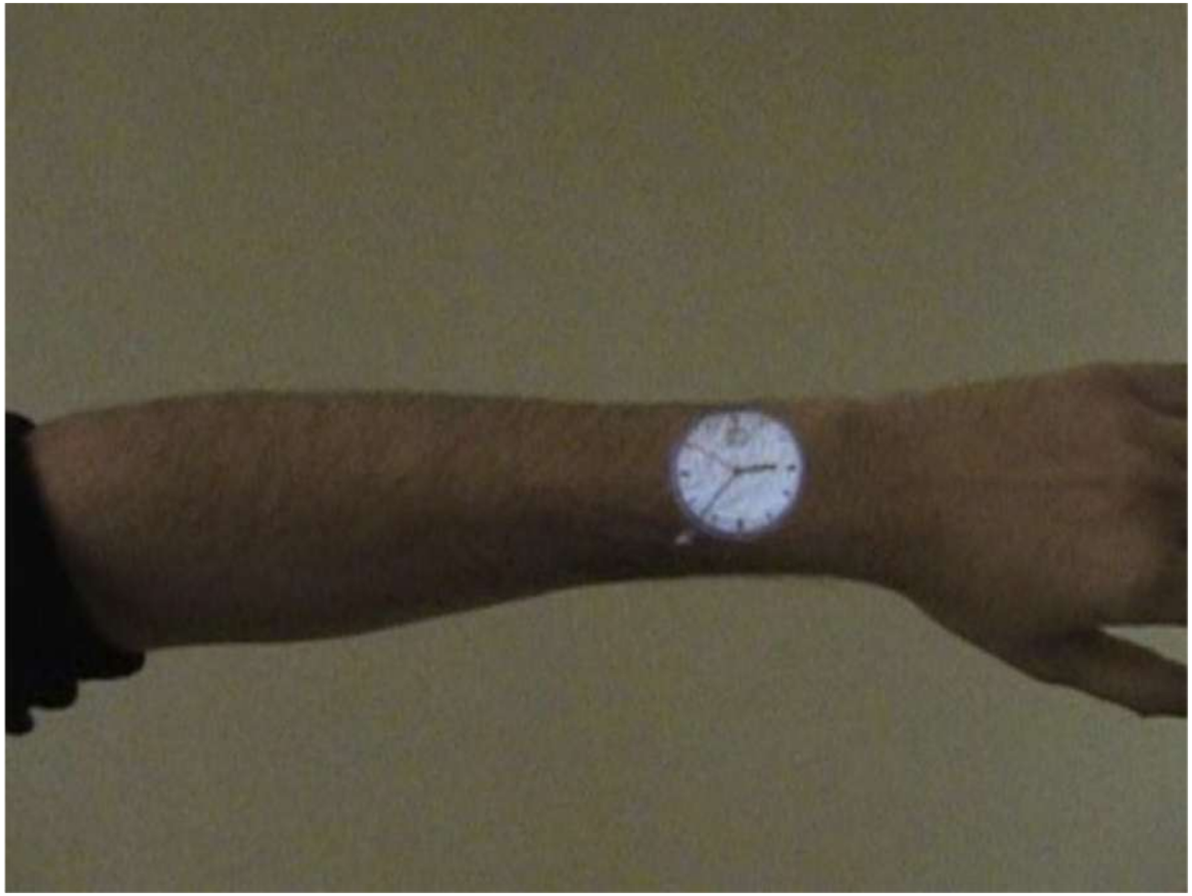
A. Move B. Left-click C. Right-click
D. Scroll E. Drag (Pinch and Move)

Object-computer interaction

Tapuma@MIT Media Lab

Wearable Visual Interface





Drawbacks of CV for HCI

- Noise
- Computing Cost
- Unstable Device
- Recognition Accuracy
- Ambiguity