Hardness Prediction for Object Detection Inspired by Human Vision

Yuwei Qiu, Huimin Ma and Lei Gao
EE Department, Tsinghua University

* All raw images in this slides are from ImageNet.

Outline

Image Complexity

- **Image complexity is important:** Tightly related to the performance of object detection algorithms
- **Inspiration from psychological experiments:** It is **easy** for **humans** to justify the complexity.
- Various definitions in different works:

  (-) There hasn't been a general accepted scheme.

  (-) None of previous definitions has built up connection between image complexity and human vision.

Motivation

- **Inspiration:** It is **easy** for **humans** to justify the complexity.
- **But,**
  - How do a person evaluate images?
  - Can we reasonably extract a kind of image complexity from the process of human observation?

Based on **eye tracking experiments**

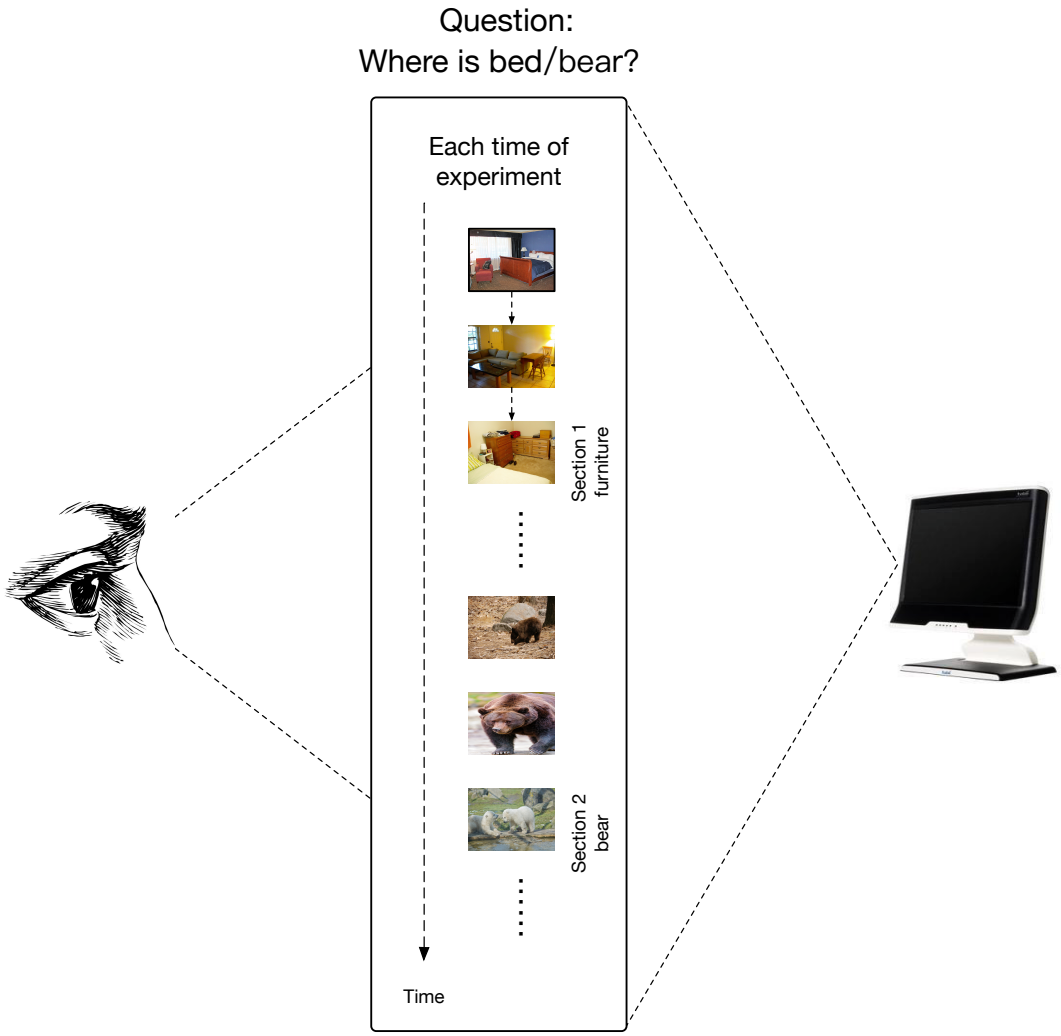
- **Goal**
  - We aim at quantifying the human vision and introducing it into the object detection task to predict the detection hardness.
  - *Make the computer like a human!*

Outline

Eye Tracking Experiments

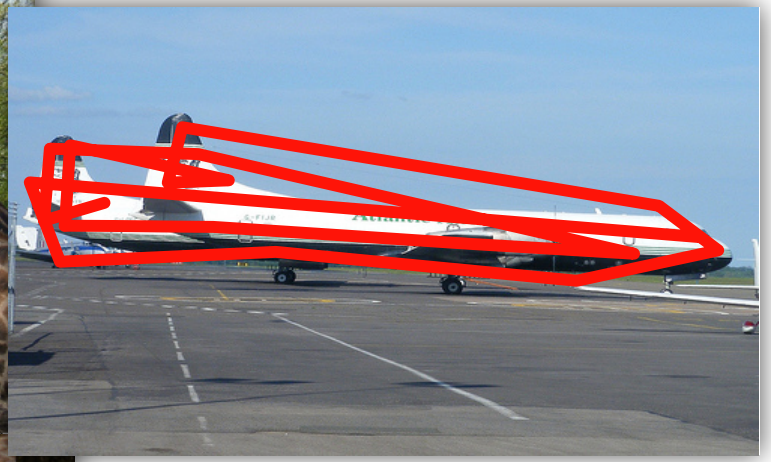
- **Participants:** Adults with an educational level above college and without any mental diseases or ophthalmic diseases.
- **Device:** *Tobii* eye trackers (record eye balls movements)
- **Task:** participants are asked to watch a series of images switching over the computer screen. Each of the participant has a different target object to seek out in one section of the experiment.
- **Image dataset:** 1280 natural scene images form ImageNet Database.

Eye Tracking Experiments



Scan Path

A broken line connecting every two gazes which have neighboring time stamps.

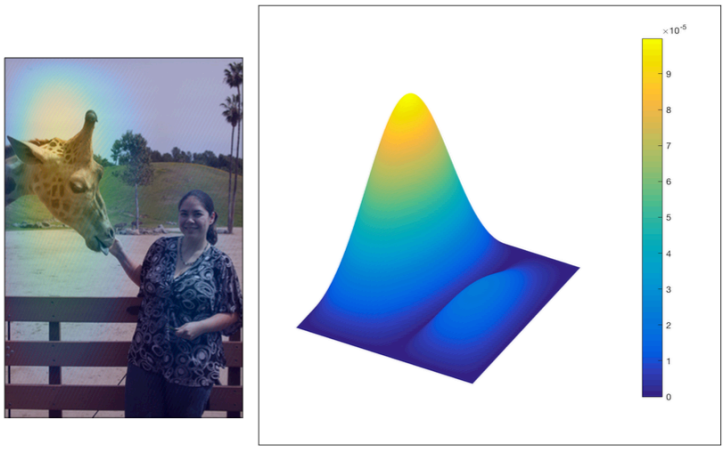


How to further describe the characteristics of scan paths?

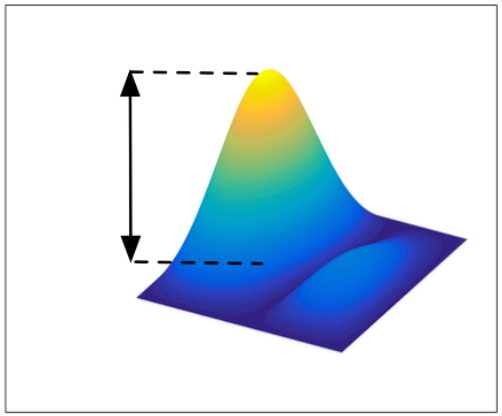
Heat Map

A map presenting **the density of gazes** over an image when people are watching it.

- A gray level form is chosen.



Heat Graph

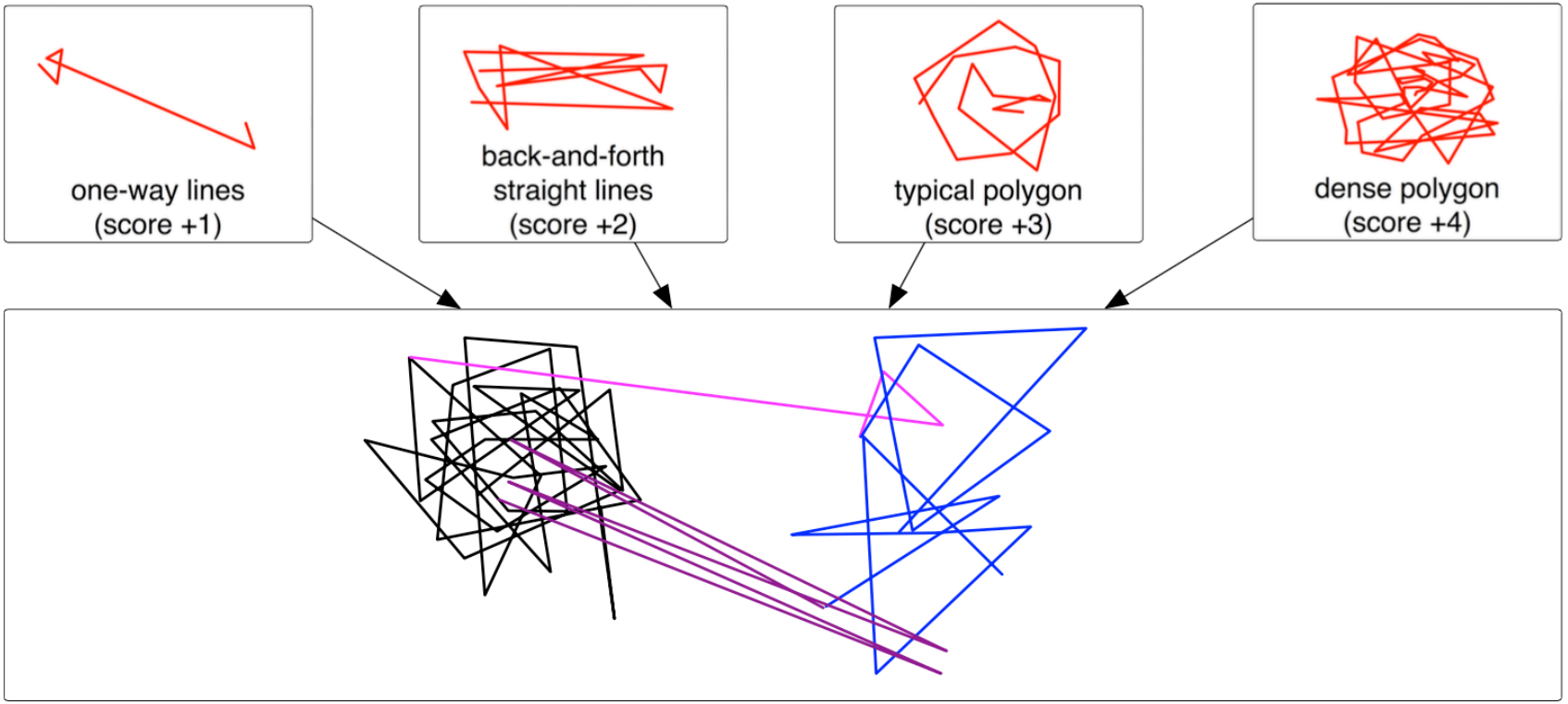


Peak of Heat

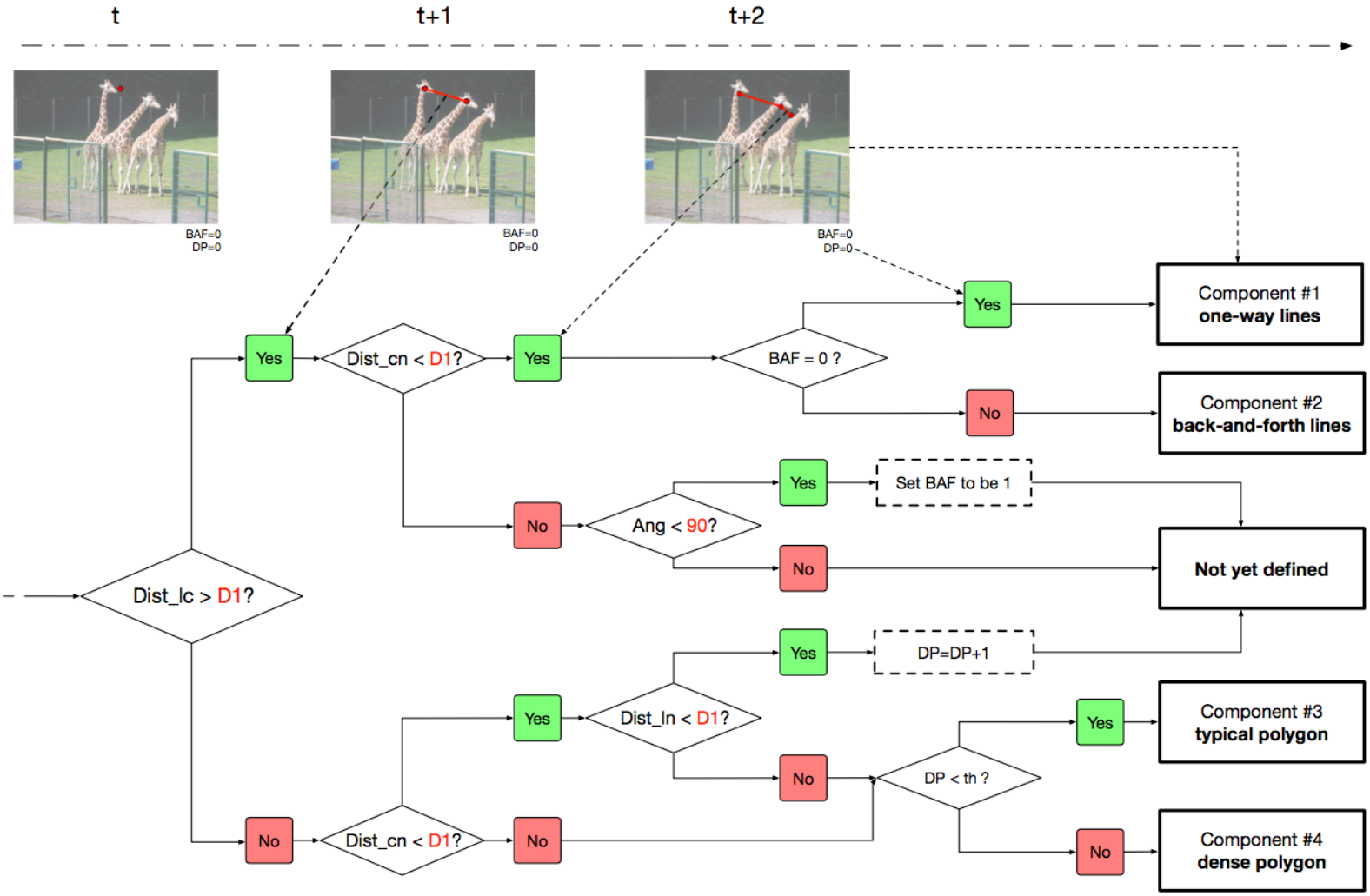
Local Minima

Scan Path Components

Massive gazes are connected in scan path, which is composed of basically four types of components.



Segmentation Pipeline



Outline

Definition

- More complex components means a higher complexity. The peak of heat describes the level of how much attention this person has paid to specific areas.

- A new image complexity metric: **Eye Tracking Complexity (ETC)**

Score of the component

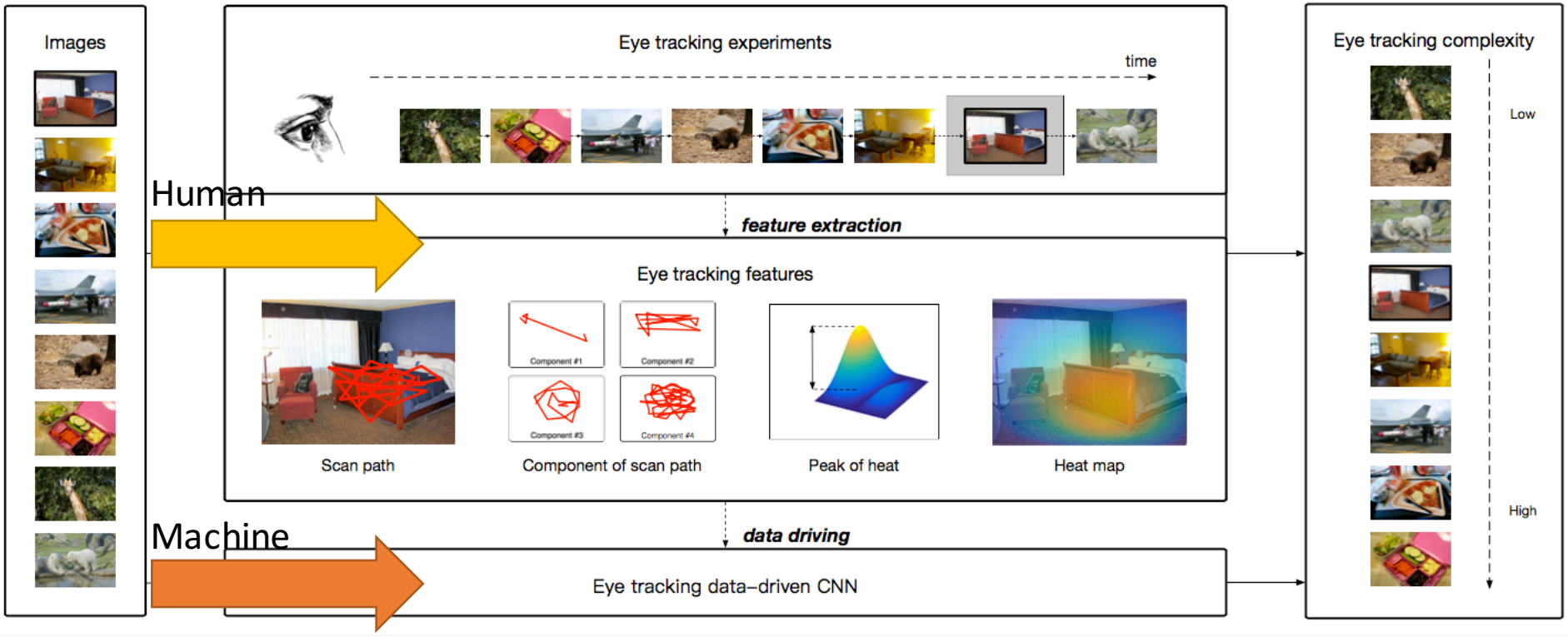
$$\mathrm{ETC} = \left\lfloor \sum_{c_i \in \mathcal{C}} s_i \times p_i + \frac{1}{2} \right\rfloor$$

Peak of heat in the local area

The set of scan path components

Outline

Pipeline


Images
Human
Machine
Eye tracking experiments
time
feature extraction
Eye tracking features
Component #1
Component #2
Component #3
Component #4
Scan path
Component of scan path
Peak of heat
Heat map
data driving
Eye tracking data–driven CNN
Eye tracking complexity
Low
High

CNN for ETC computation

We adjust the architecture of LeNet. This CNN contains two convolutional layers, two pooling layers and two fully connected layers.

It is used to classify *gray level images*, labeled by **eye tracking complexity extracted by eye tracking experiments**, into *categories*.

Testset

Why we construct these three testsets?

"Easiest Classes" in ILSVRC Bird, Dog, Tiger, Zebra

"Hardest Classes" in ILSVRC Back pack, Lamp, Ladle, Microphone

Classes	Bird	Dog	Tiger	Zebra	Back pack	lamp	Ladle	Micro phone	AETC
Set 1	12	12	12	12	0	0	0	0	
Set 2	6	6	6	6	6	6	6	6	
Set 3	0	0	0	0	12	12	12	12	

Table 3

Testset

Human Pathway:
Average eye tracking complexity (AETC) of three different sets. The number of each class are shown below.

Classes	Bird	Dog	Tiger	Zebra	Back pack	lamp	Ladle	Micro phone	AETC
Set 1	12	12	12	12	0	0	0	0	**4.8**
Set 2	6	6	6	6	6	6	6	6	**10.1**
Set 3	0	0	0	0	12	12	12	12	**14.8**

Table 3

Results

Images of the hardest classes (Backpack, Ladle, Lamp and Microphone) have in ETC 11 (with an ETC of 15) and 12 (with an ETC of 16) while the easiest classes (Bird, Dog, Tiger and Zebra) in ETC 1 (with an ETC of 5).

Classes	Bird	Dog	Tiger	Zebra	Back pack	lamp	Ladle	Micro phone
Output E.T.C.	9 imgs all with ETC = 1	9 imgs all with ETC = 1	9 imgs all with ETC = 1	9 imgs all with ETC = 1	9 imgs all with ETC = 12	9 imgs all with ETC = 12	8 imgs with ETC = 12; 1 img with ETC=11	9 imgs all with ETC = 12

Table 2

Outline

Contributions

(+) Human vision quantifying: Eye tracking features including scan path, heat map, components of scan path, and peaks of heat are either introduced or newly defined.

(+) New complexity definition: A new metric of image complexity is defined based on these eye tracking features. It also has been validated that the new defined complexity corresponds to the detection algorithms' average precision.

(+) Automatic computation: The new defined complexity can be computed either by carrying out eye tracking experiments and extracting eye tracking features, or just through a pre-trained CNN.

Thanks for your time!

Hierarchical algorithm design

Computer Pathway:
Three CNNs with different iterations (CNN-100, CNN-1000, CNN-10000) are trained for object detection. 50000 images for training and 10000 for validation.

Results:

mAP	Set 1	Set 2	Set 3
CNN-100	0.632	0.142	0.017
CNN-1000	0.637	0.254	0.086
CNN-10000	0.637	0.365	0.118

Table 4

All of three CNNs performed equally on precision over images in Category 1 (Bird, Tiger, Zebra) but quite different on Category 12 (Lamp).

Don't need extra computational resource.

It's better to have more iterations.

Potential Applications

- Hierarchical algorithm design
  - Too easy: simple structure/implementation
  - Too hard: complex structure, more iterations

Use the computational resource much more efficiently!