هوش مصنوعی

فصل ۲۵

ادراک: بینایی ماشینی

Perception: Machine Vision

کاظم فولادی قلعه
دانشکده مهندسی، دانشکدگان فارابی
دانشگاه تهران

http://courses.fouladi.ir/ai

هوش مصنوعی

ادراک: بینایی ماشینی

◆

مقدمه

ادراک و عامل

ادراک از **حسگرها** شروع می‌شود.

حسگر: هر چیزی که بتواند جنبه‌هایی از محیط را ثبت کرده و به صورت یک ورودی به برنامه‌ی عامل ارسال کند.

مثال: حسگر یک‌بیتی ساده: روشن/خاموش، شبکیه‌ی چشم انسان

بینایی: مفیدترین حس در برخورد با دنیای فیزیکی

روی‌کردهای استفاده از ادراک‌ها توسط عامل		
استخراج ویژگی *Feature Extraction*	**بازشناسی** *Recognition*	**بازسازی** *Reconstruction*
عامل تعداد اندکی از ویژگی‌ها را از ورودی‌های حسی خود شناسایی کرده و آن‌ها را مستقیماً به برنامه‌ی عامل خود می‌فرستد. با ویژگی‌ها به‌صورت واکنشی برخورد می‌شود.	عامل بر اساس اطلاعات دیداری و غیره میان اشیایی که با آنها مواجه می‌شود، تمایز قایل می‌شود. بازشناسی می‌تواند تصویر را برچسب‌گذاری کند.	عامل از روی یک تصویر یا مجموعه‌ای از تصاویر، یک مدل هندسی از دنیا می‌سازد.

ادراک: بینایی

گرافیک کامپیوتری

از محرک حسی برای بازسازی دنیا استفاده می‌کنیم:

$$S = g(W)$$

ادراک (تحریک)
Percept (Stimulus)

دنیا
World

تابع گرافیک کامپیوتری
Computer Graphics

وضعیت دنیا (واقعی یا تخیلی) W را به محرک تولید شده از دنیا S نگاشت می‌دهد.

ادراک: بینایی

بینایی وارون گرافیک است

از محرک حسی برای بازسازی دنیا استفاده می‌کنیم:

$$W = g^{-1}(S)$$

دنیا
World

ادراک (تحریک)
Percept (Stimulus)

تابع بینایی کامپیوتری
Computer Vision

محرک تولید شده از دنیا S را به وضعیت دنیا W نگاشت می‌دهد.

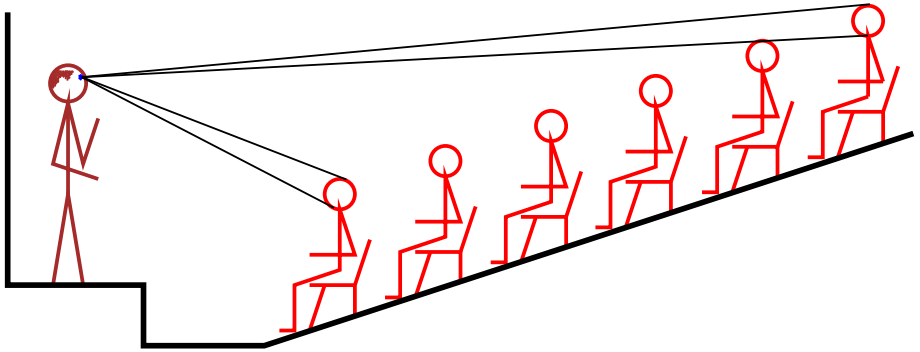
اما متأسفانه g وارون مناسبی ندارد!

ادراک: بینایی

مثال (۱ از ۳)

آیا می‌توانیم به بینایی به‌عنوان وارون گرافیک نگاه کنیم؟

$$W = g^{-1}(S)$$

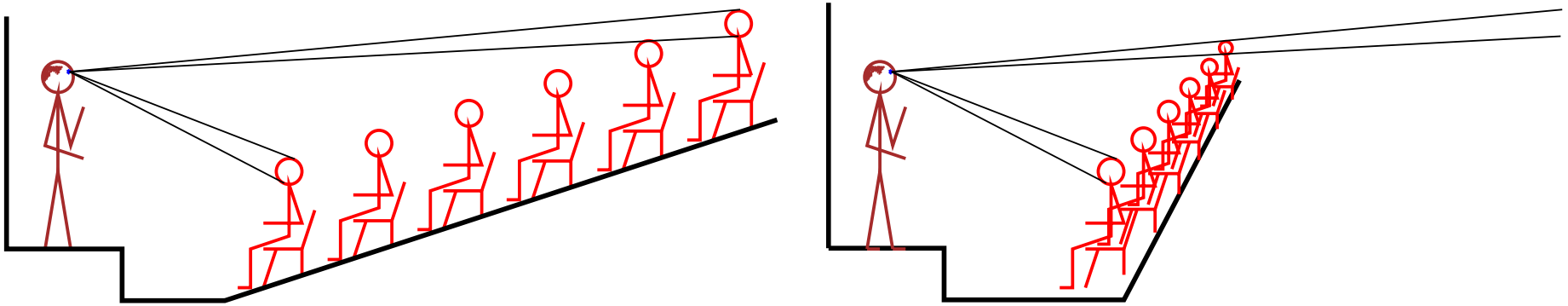


ادراک: بینایی

مثال (۲ از ۳)

آیا می‌توانیم به بینایی به‌عنوان وارون گرافیک نگاه کنیم؟

$$W = g^{-1}(S)$$



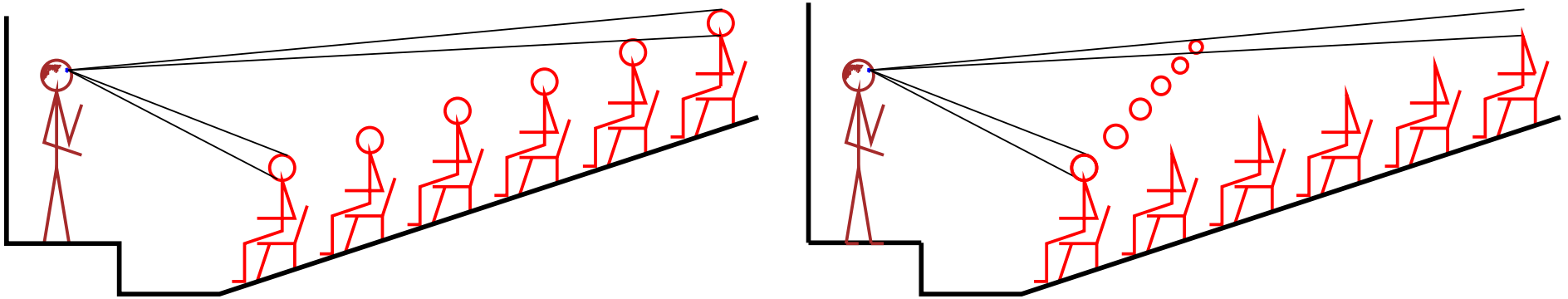
مشکل: ابهام انبوه!

ادراک: بینایی

مثال (۳ از ۳)

آیا می‌توانیم به بینایی به‌عنوان وارون گرافیک نگاه کنیم؟

$$W = g^{-1}(S)$$



مشکل: ابهام انبوه!

ادراک: بینایی

رویکرد بهتر: روی‌کرد بیزی

روی‌کرد بیزی برای استنتاج پیکربندی دنیا

$$P(W \mid S) = \alpha P(S \mid W) P(W)$$

دانایی پیشین	گرافیک
Prior Knowledge	*Graphics*

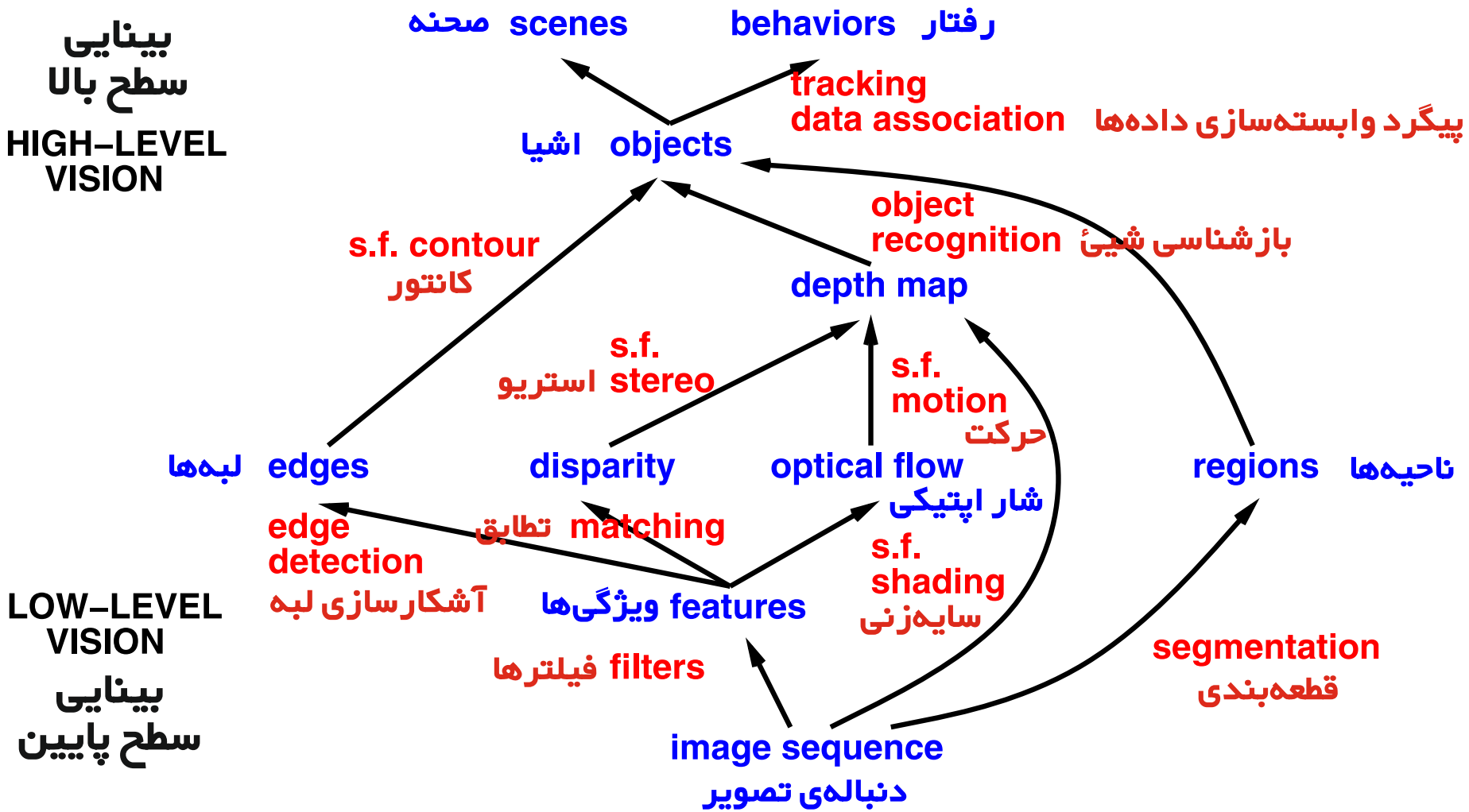
روی‌کرد بازهم بهتر: نیازی نداریم که صحنه‌ی دقیق را بازیابی کنیم
فقط اطلاعاتی که برای موارد زیر نیاز داریم را استخراج می‌کنیم:

ناوبری	دست‌کاری	بازشناسی	شناسایی
Navigation	*Manipulation*	*Recognition*	*Identification*

بینایی کامپیوتری

زیرسیستم‌ها

COMPUTER VISION: SUBSYSTEMS



بینایی نیاز دارد اشاره‌های مختلف را با هم ترکیب کند.

هوش مصنوعی

ادراک: بینایی ماشینی

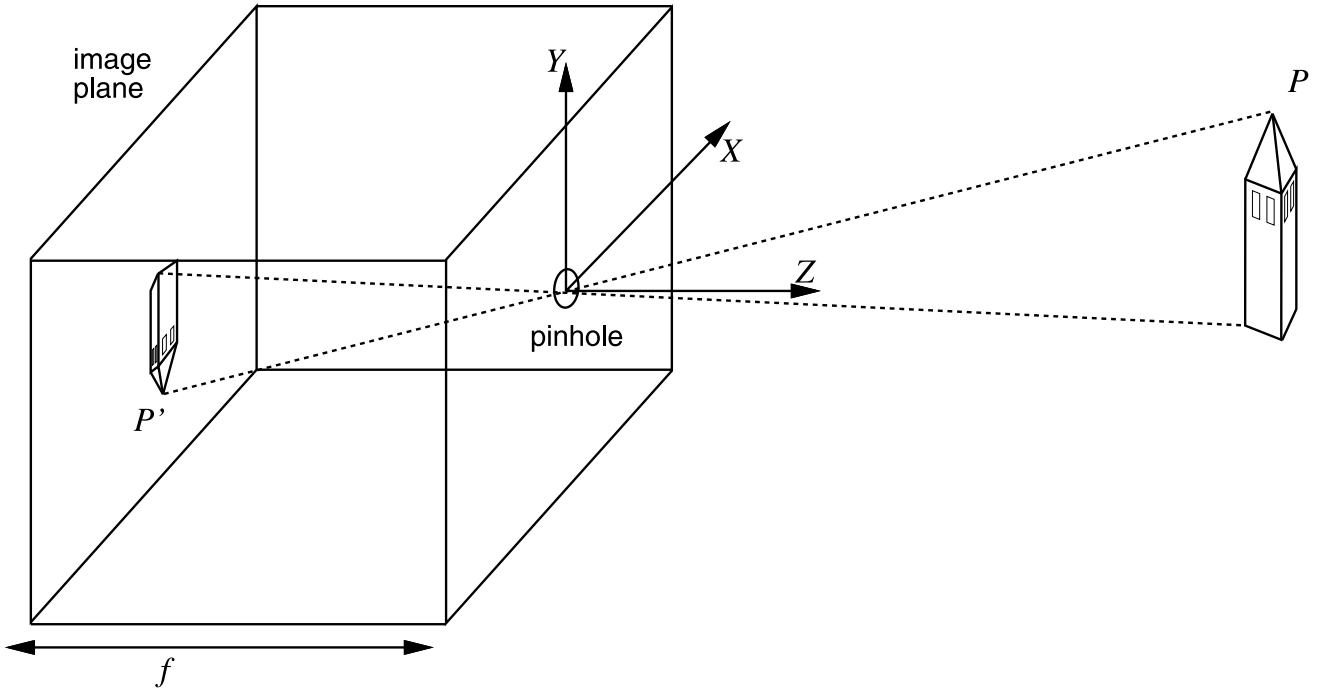
۱

تشکیل تصویر

تشکیل تصویر

دوربین «سوراخ‌سوزنی»

IMAGE FORMATION



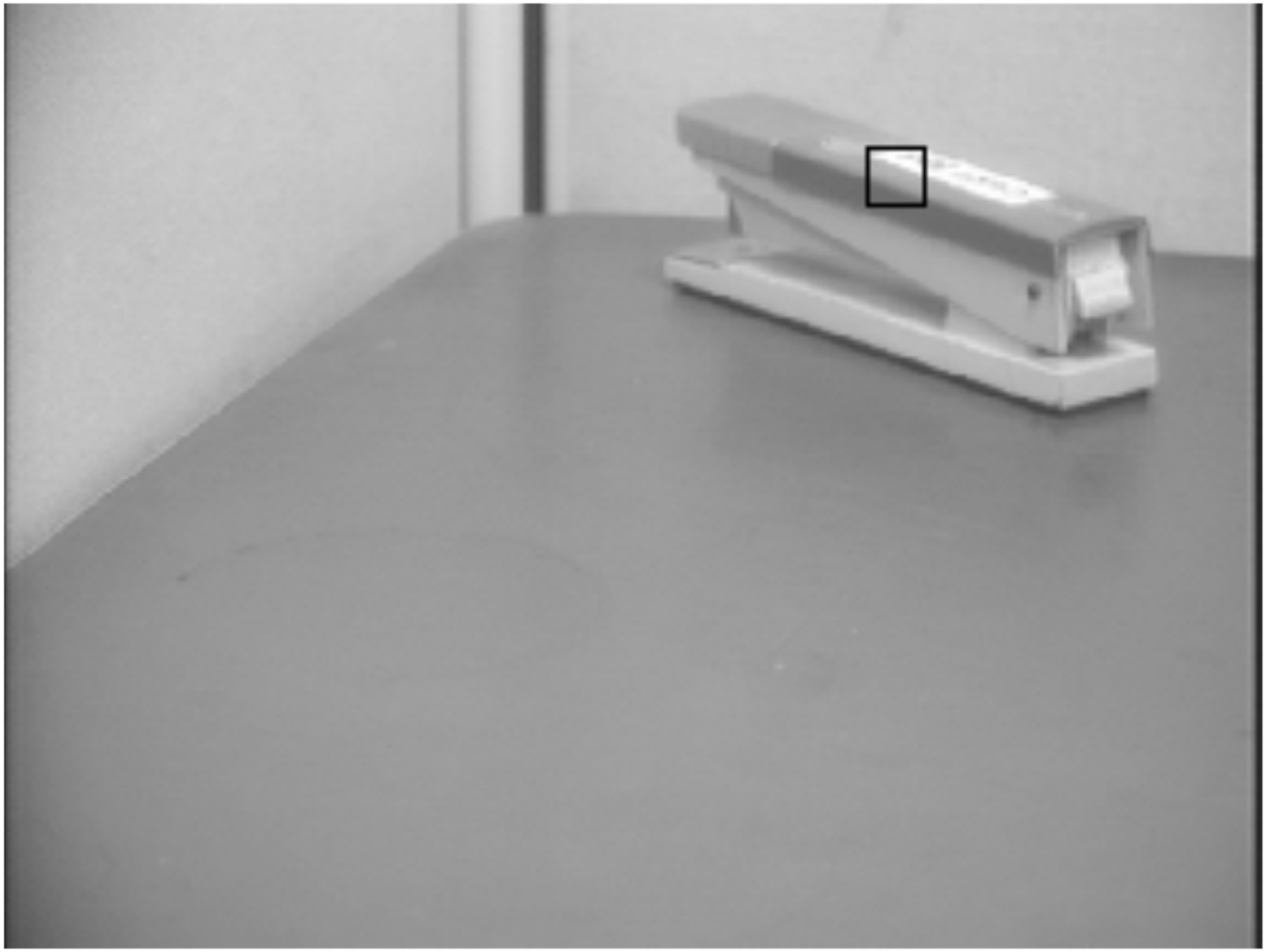
P یک نقطه در صحنه با مختصات (X,Y,Z) است.
P' تصویر آن در صفحه‌ی تصویر با مختصات (x,y,z) است.

$$x = \frac{-fX}{Z},\ y = \frac{-fY}{Z}$$

(از روی مثلث‌های متشابه)

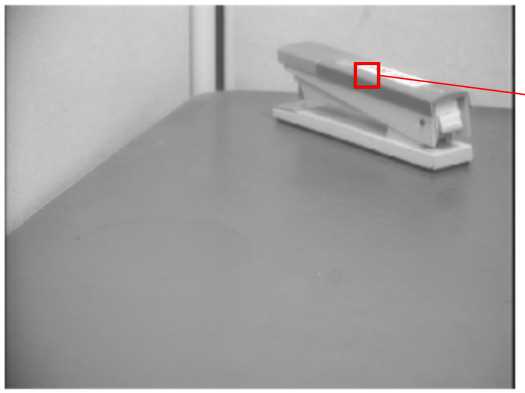
مقیاس/فاصله نامعین است.

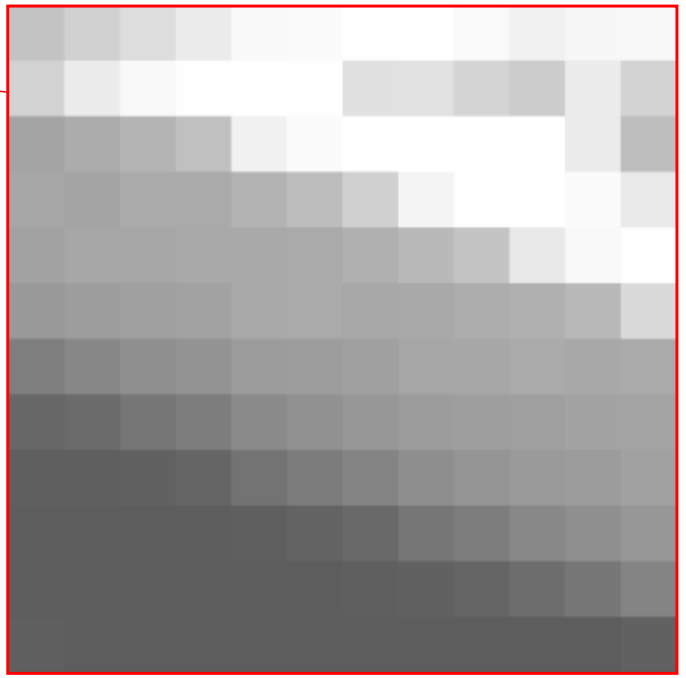
تصویر



تصویر

مقادیر پیکسل‌ها





195	209	221	235	249	251	254	255	250	241	247	248
210	236	249	254	255	254	225	226	212	204	236	211
164	172	180	192	241	251	255	255	255	255	235	190
167	164	171	170	179	189	208	244	254	255	251	234
162	167	166	169	169	170	176	185	196	232	249	254
153	157	160	162	169	170	168	169	171	176	185	218
126	135	143	147	156	157	160	166	167	171	168	170
103	107	118	125	133	145	151	156	158	159	163	164
095	095	097	101	115	124	132	142	117	122	124	161
093	093	093	093	095	099	105	118	125	135	143	119
093	093	093	093	093	093	095	097	101	109	119	132
095	093	093	093	093	093	093	093	093	093	093	119

pixel = **pic**ture + **el**ement

$I(x, y, t)$ is the intensity at (x, y) at time t

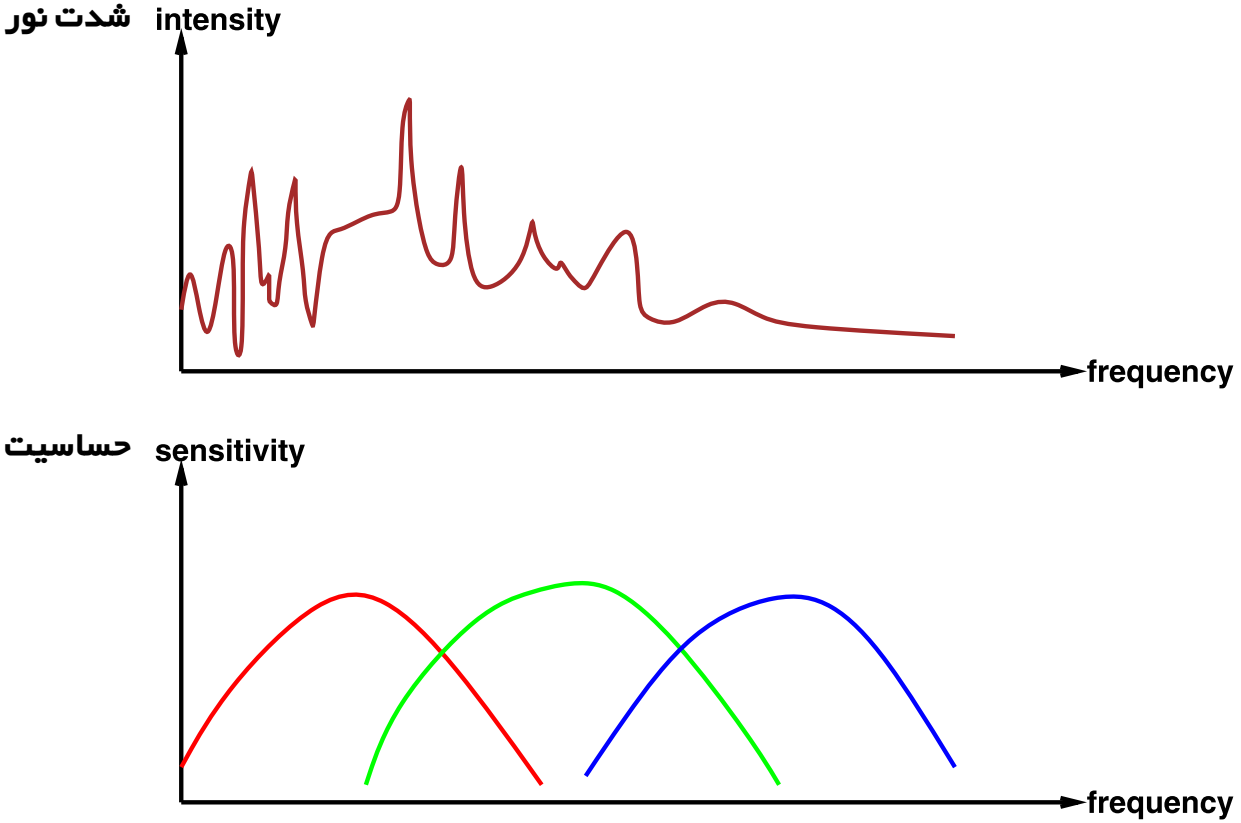
دوربین CCD $\approx$ ۱۲ میلیون پیکسل

چشم انسان $\approx$ ۲۴۰ میلیون پیکسل، 0.25 ترابیت بر ثانیه

بینایی رنگی

COLOR VISION

شدت نور با فرکانس تغییر می‌کند ⇐ سیگنال بی‌نهایت بعدی



چشم انسان سه نوع سلول حساس به رنگ دارد؛

هر یک از کل سیگنال بی‌نهایت بعدی انتگرال می‌گیرد ⇐ بردار شدت نور سه‌عنصری

هوش مصنوعی

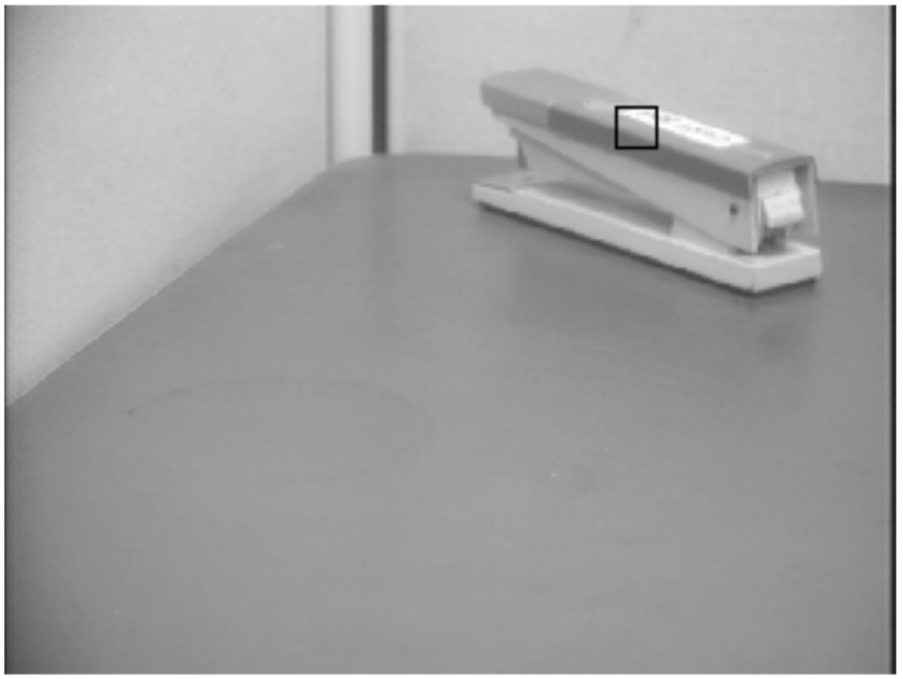
ادراک: بینایی ماشینی

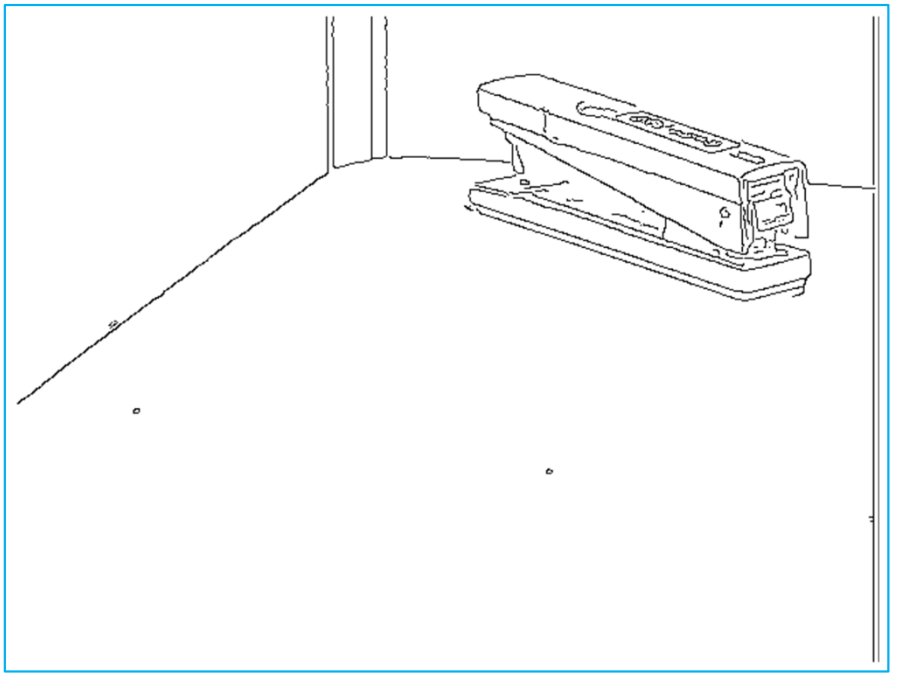
۲

عملیات پردازش تصویر اولیه

آشکارسازی لبه

EDGE DETECTION





Edges in image ⇐ discontinuities in scene:

1) depth
2) surface orientation
3) reflectance (surface markings)
4) illumination (shadows, etc.)

لبه‌ها در تصویر ⇒ ناپیوستگی‌ها در صحنه:

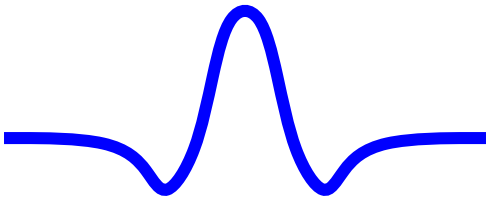
۱) عمق
۲) جهت رویه
۳) بازتاب (علامت‌گذاری‌های رویه)
۴) تابش (سایه‌ها و ...)

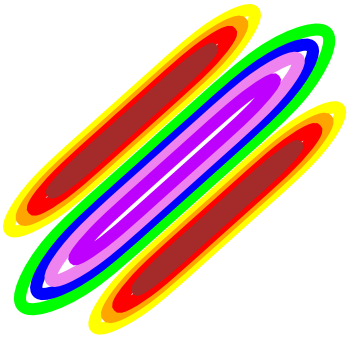
آشکارسازی لبه

الگوریتم

EDGE DETECTION

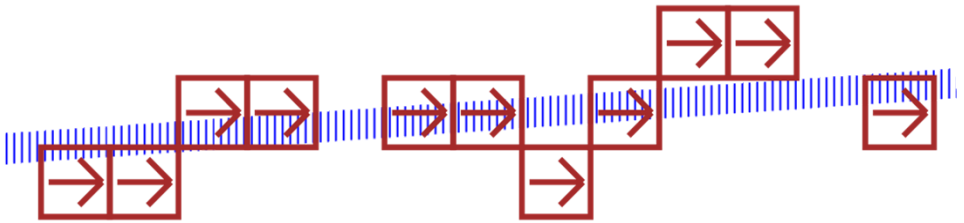
۱) تصویر را با فیلترهای مایل در حوزه‌ی مکان (احتمالاً multiscale) کانوالو کنید.





۲) پیکسل‌های بالای مقدار آستانه (above-threshold) را با جهت لبه برچسب‌گذاری کنید.

۳) پاره‌خطهای «تمیز» را با ترکیب پیکسل‌های لبه‌ی متوالی با جهت یکسان استنتاج کنید.



هوش مصنوعی

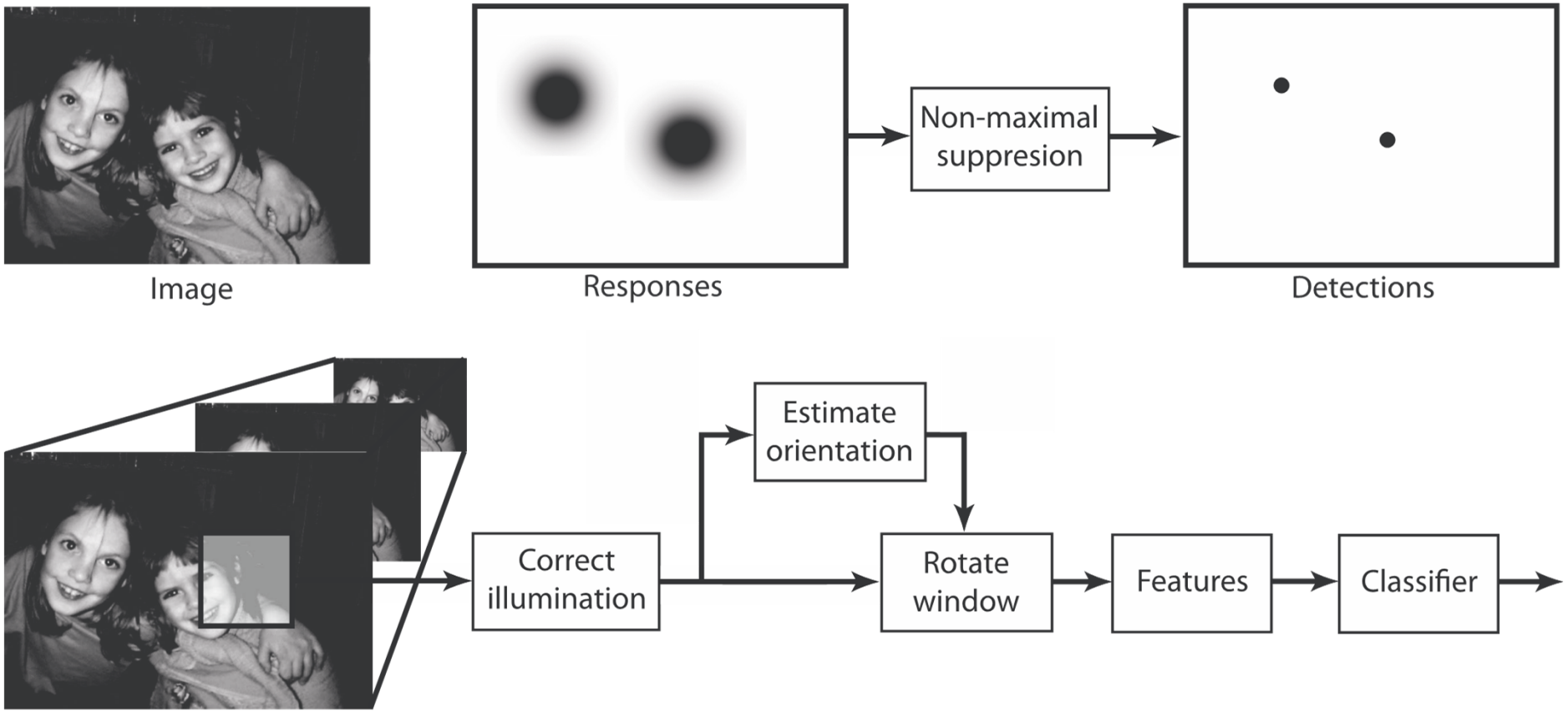
ادراک: بینایی ماشینی

۳

بازشناسی اشیا از روی ظاهر

بازشناسی اشیا از روی ظاهر

مثال



هوش مصنوعی

ادراک: بینایی ماشینی

۴

بازسازی دنیای سه‌بعدی

بازسازی دنیای سه‌بعدی

اشاره‌هایی از دانایی پیشین

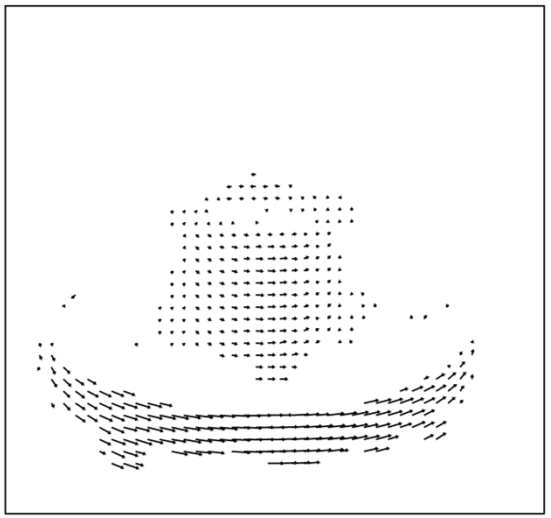
RECONSTRUCTING THE 3D WORLD: CUES FROM PRIOR KNOWLEDGE

شکل به‌دست‌آمده از ... *Shape From …*	فرض می‌کند ... *Assumes …*
حرکت *Motion*	بدنه‌های صلب، حرکت پیوسته *rigid bodies, continuous motion*
استریو *Stereo*	جامد، همجوار، بدنه‌های غیرتکراری *solid, contiguous, non-repeating bodies*
بافت *Texture*	بافت یکنواخت *uniform texture*
سایه‌زنی *Shading*	بازتاب یکنواخت *uniform reflectance*
کانتور *Contour*	انحنای می‌نیمم *minimum curvature*

بازسازی دنیای سه‌بعدی

حرکت

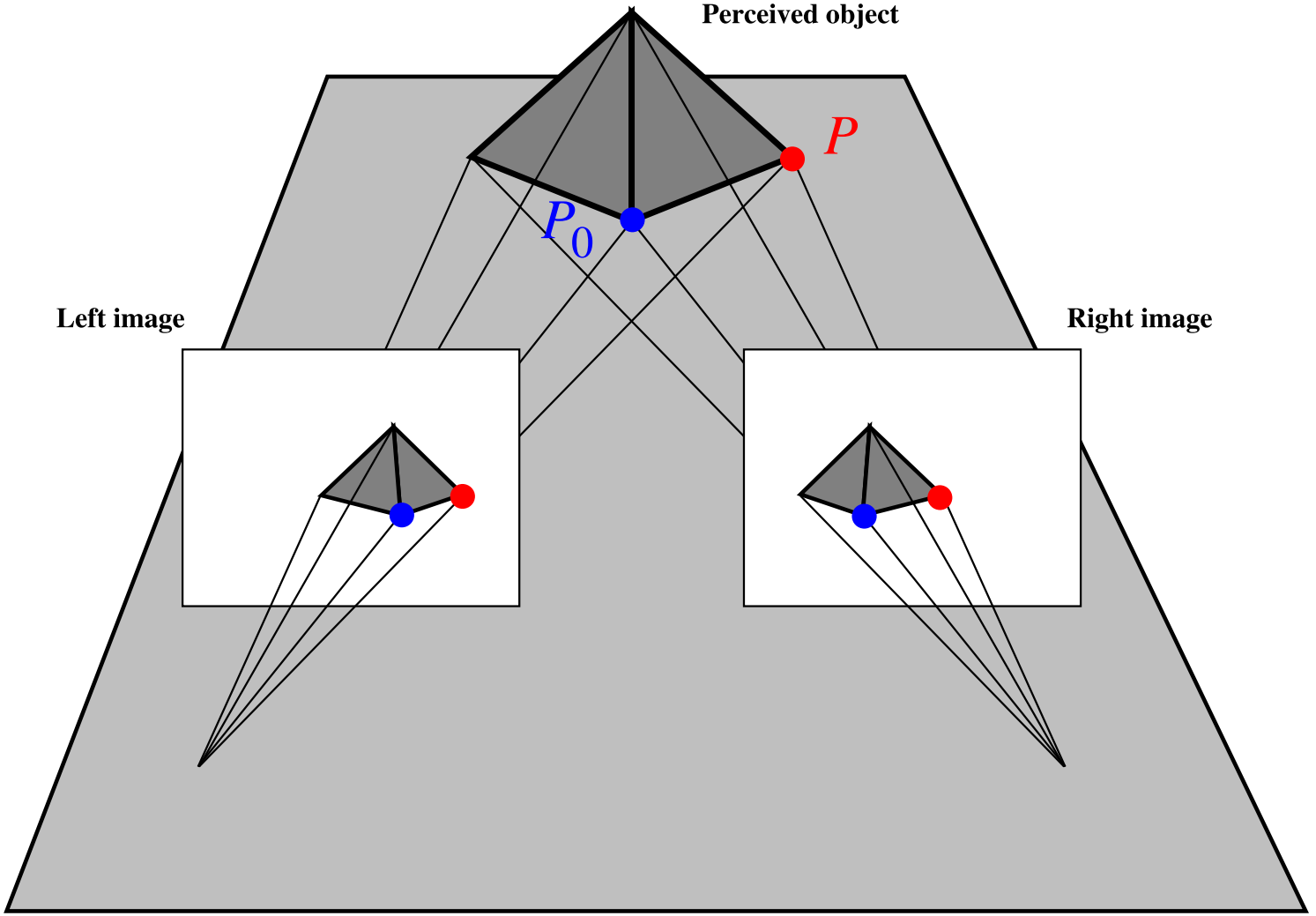
MOTION



بازسازی دنیای سه‌بعدی

استریو

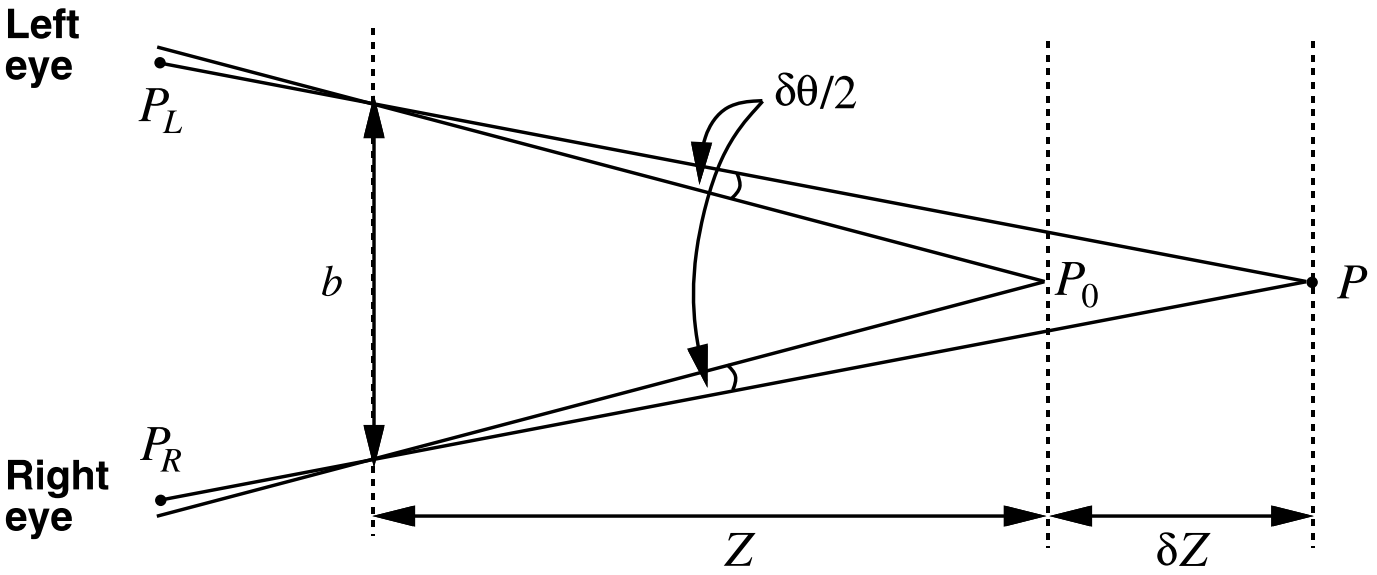
STEREO



بازسازی دنیای سه‌بعدی

استریو : تعیین عمق

STEREO



Simple geometry: $\delta Z = Z^2 \delta\theta/(-b)$

Physiology: $\delta\theta \geq 2.42 \times 10^{-5}$ radians, $b = 6$cm

$Z = 30\text{cm} \Rightarrow \delta Z \approx 0.04\text{mm}$

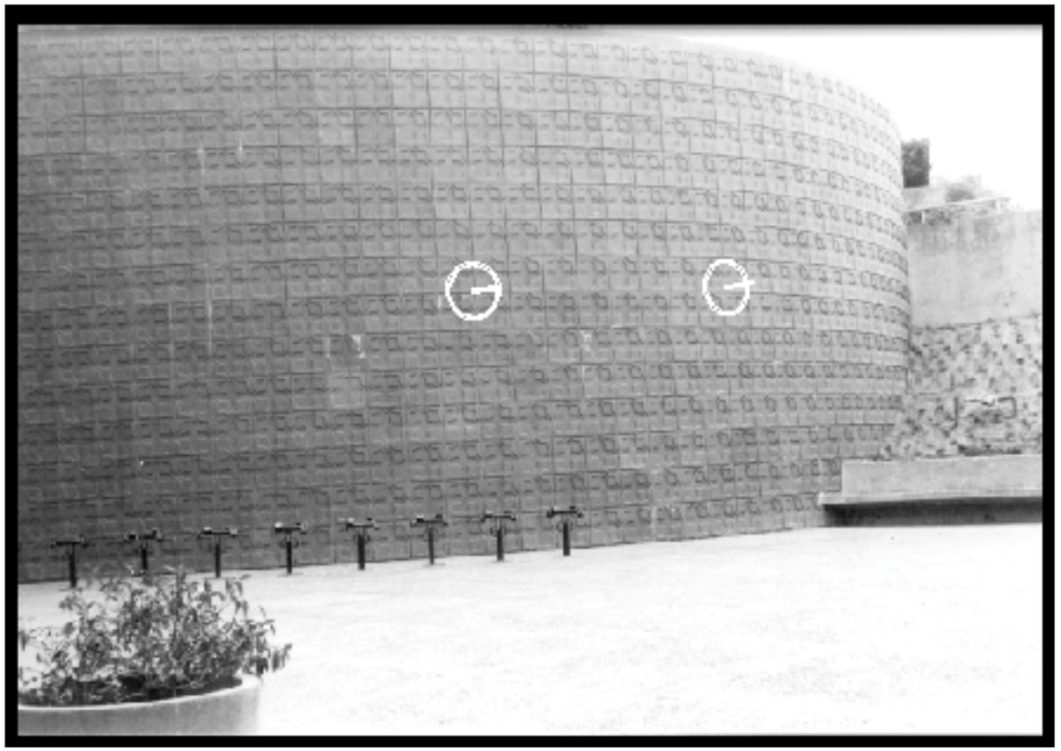
$Z = 30\text{m} \Rightarrow \delta Z \approx 40\text{cm}$

Large baseline $\Rightarrow$ better resolution!

بازسازی دنیای سه‌بعدی

بافت

TEXTURE



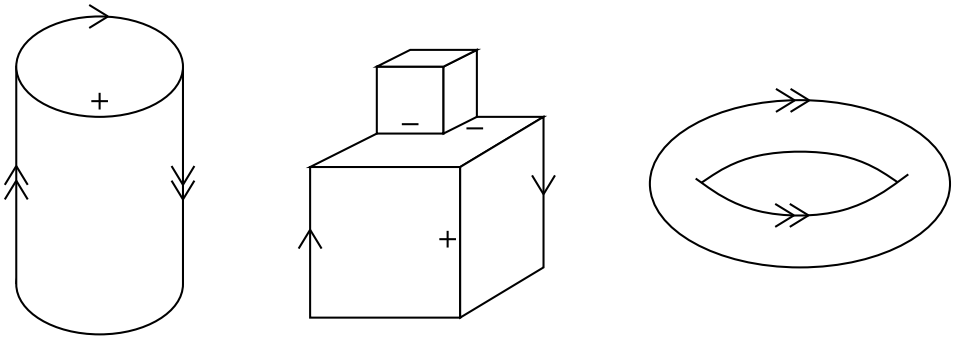
ایده: فرض می‌شود بافت واقعی یکنواخت است،
شکل رویه‌ای که این اعوجاج را ایجاد کرده است، محاسبه کنید.

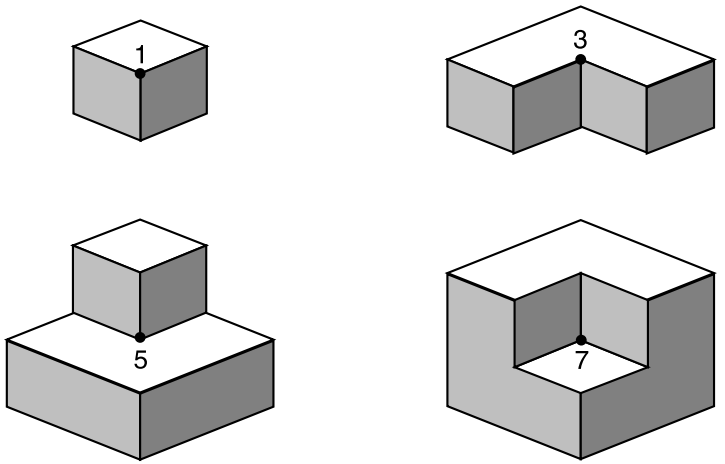
ایده‌ی مشابه برای سایه هم کار می‌کند:
فرض می‌کنیم بازتاب و ... یکنواخت باشد، **اما** عدم بازتاب‌ها محاسبات غیرمحلی شدت نور دریافت شده را موجب می‌شوند.
⇐ گودی‌ها کم‌عمق‌تر از آنچه هستند به نظر می‌رسد.

بازسازی دنیای سه‌بعدی

انواع لبه‌ها و رأس‌ها

EDGE AND VERTEX TYPES



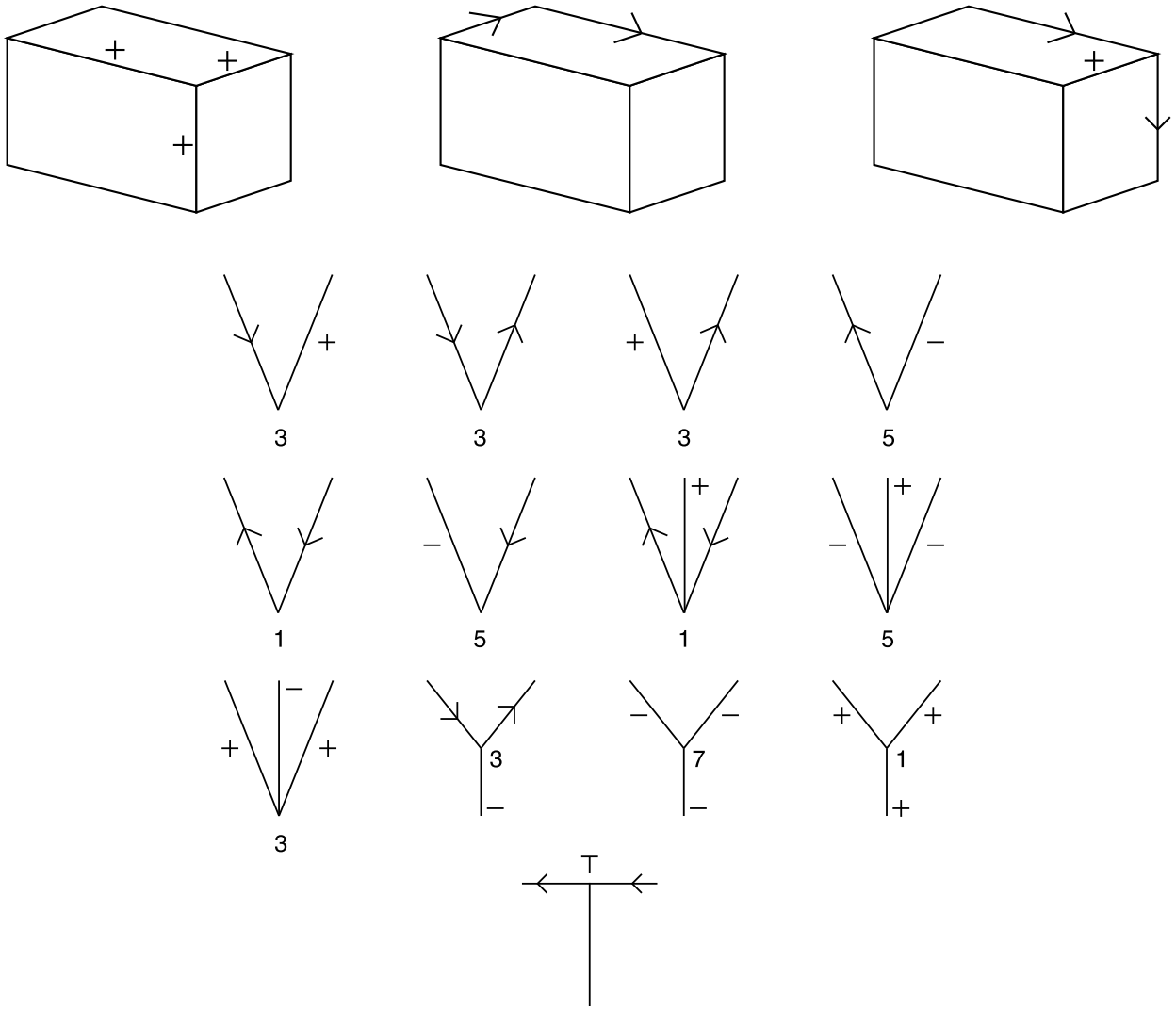


دنیای **اشیای چندوجهی** با **رئوس سه‌وجهی** را در نظر می‌گیریم.

بازسازی دنیای سه‌بعدی

برچسب‌های رأس/لبه

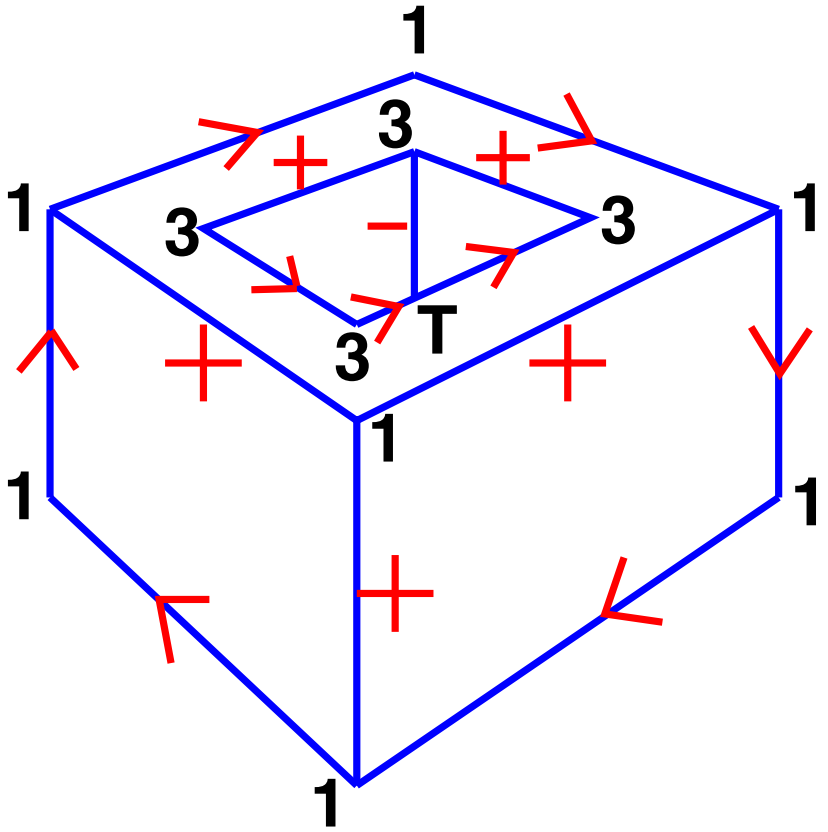
VERTEX/EDGE LABELS

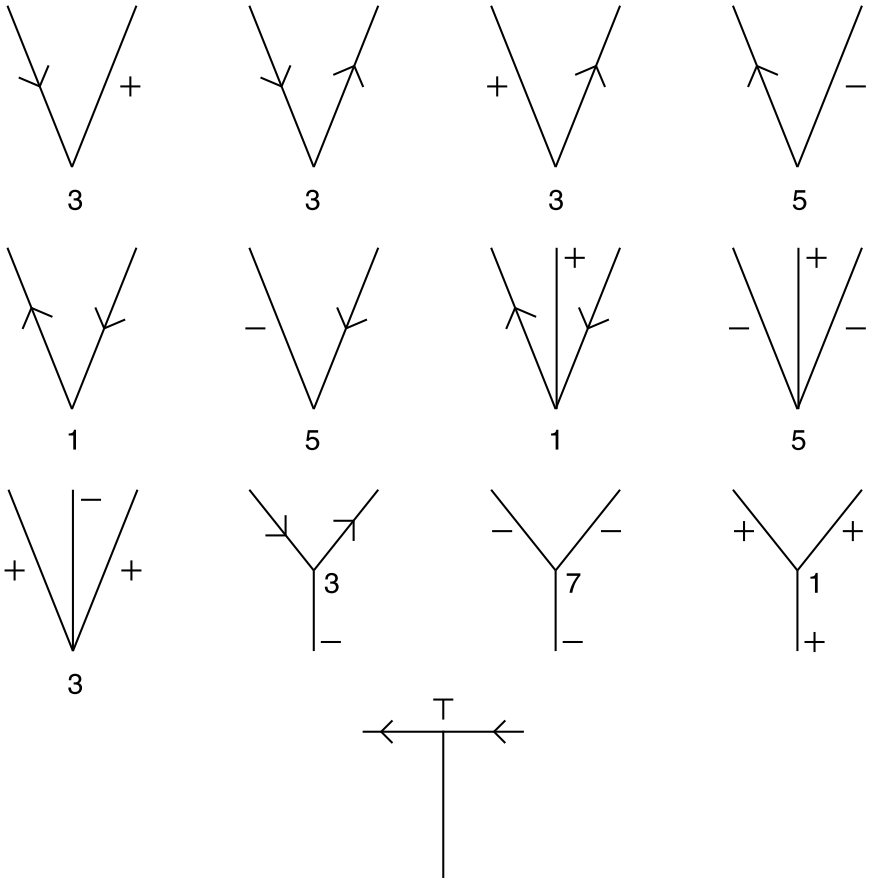


بازسازی دنیای سه‌بعدی

برچسب‌زنی رأس/لبه: مثال

VERTEX/EDGE LABELLING: EXAMPLE





مسئله‌ی برچسب‌زنی رأس/لبه به‌عنوان یک مسئله‌ی ارضای قید **(CSP)**:

متغیرها = لبه‌ها

قیدها = پیکربندی‌های ممکن گره‌ها

هوش مصنوعی

ادراک: بینایی ماشینی

۵

بازشناسی اشیا از روی اطلاعات ساختاری

بازشناسی اشیا

OBJECT RECOGNITION

ایده‌ی ساده برای بازشناسی اشیا:
– استخراج شکل‌های سه‌بعدی از تصویر
– تطابق در برابر «کتابخانه‌ی شکل‌ها»

مسائل بازشناسی شیئ

- استخراج رویه‌های انحنادار از تصویر
- بازنمایی شکل شیئ استخراج‌شده
- بازنمایی شکل و تغییرپذیری طبقه‌های اشیا در کتابخانه
- قطعه‌بندی نامناسب، پنهان شدن (قرار گرفتن شیئی روی دیگری)
- نامعلوم بودن نورپردازی، سایه‌ها، علامت‌گذاری‌ها، نویز، پیچیدگی و ...

روی‌کردها

- نمایه‌سازی کتابخانه با اندازه‌گیری خصوصیات نامتغیر اشیا
- ترازبندی ویژگی‌های تصویر با ویژگی‌های شیئ پروجکت شده در کتابخانه
- تطابق تصویر در برابر چندین نمای ذخیره شده (جنبه‌ها) از کتابخانه‌ی اشیا
- روش‌های یادگیری ماشینی بر اساس آماره‌های تصویر

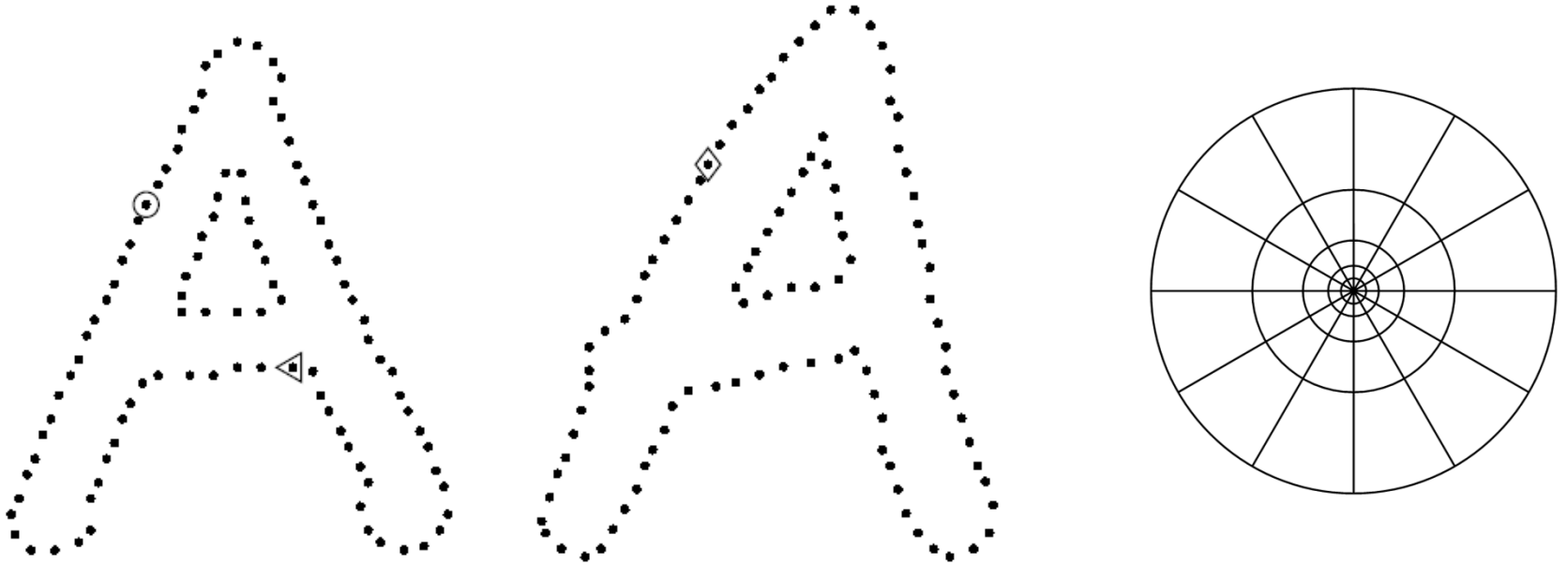
بازشناسی اشیا

روش تطابق زمینه‌ی شکل (۱ از ۳)

SHAPE-CONTEXT MATCHING

ایده‌ی پایه:

شکل (یک مفهوم رابطه‌ای) را به مجموعه‌ای ثابت از خصیصه‌ها تبدیل کنید؛
با استفاده از بستر مکانی هر یک از مجموعه نقاط ثابت روی رویه‌ی شکل.

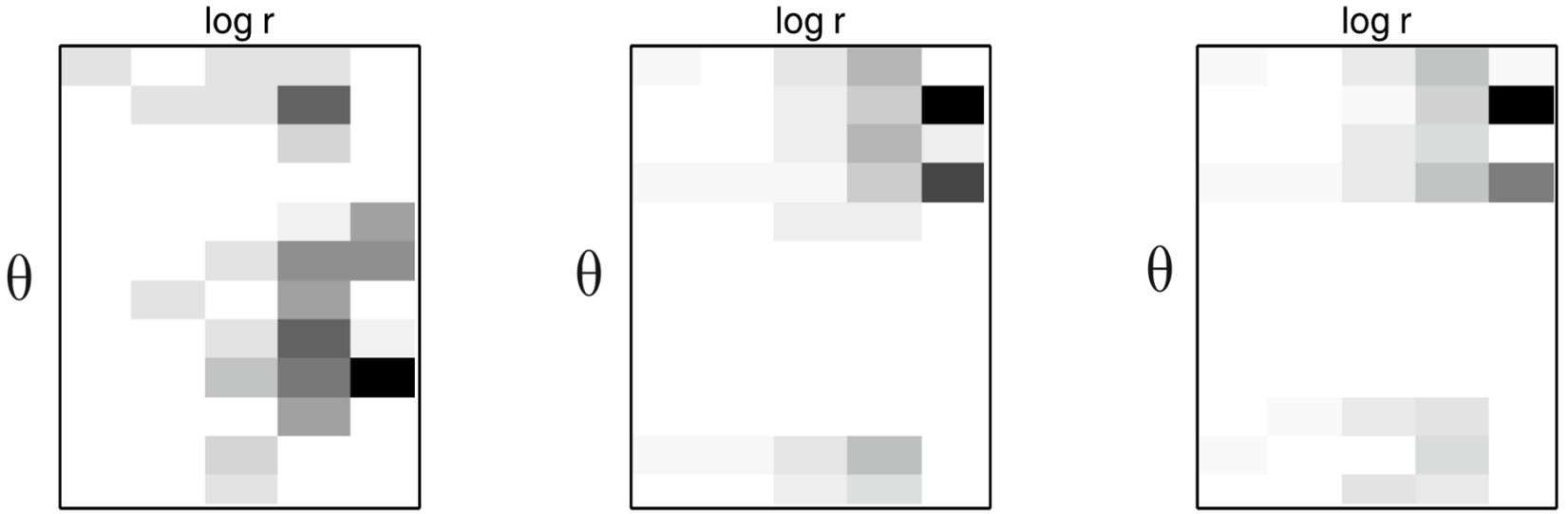


بازشناسی اشیا

روش تطابق زمینه‌ی شکل (۲ از ۳)

SHAPE-CONTEXT MATCHING

هر نقطه با هیستوگرام بستر محلی خودش توصیف می‌شود.
(تعداد نقاطی که در هر بسته‌ی مشبکه‌ی نمودار لگاریتمی–قطبی می‌افتد)

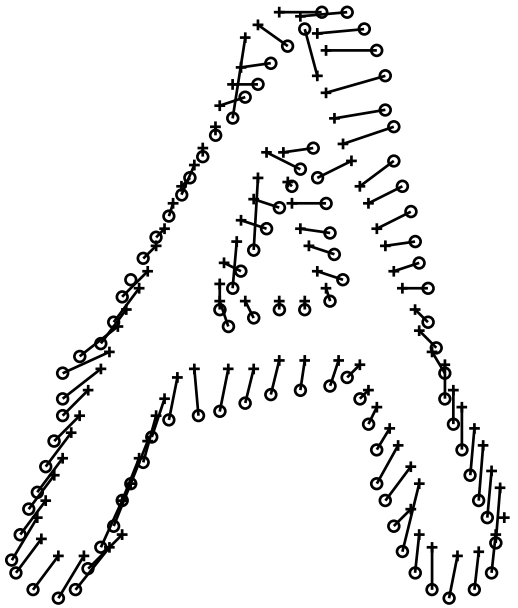


بازشناسی اشیا

روش تطابق زمینه‌ی شکل (۳ از ۳)

SHAPE-CONTEXT MATCHING

فاصله‌ی کل بین شکل‌ها
با مجموع فاصله‌های میان نقاط متناظر تحت بهترین تطابق تعیین می‌شود.



الگوریتم یادگیری ساده‌ی نزدیک‌ترین همسایه، 0.63% خطا بر روی مجموعه‌داده اعداد NIST نشان می‌دهد.

هوش مصنوعی

ادراک: بینایی ماشینی

۶

استفاده از بینایی

استفاده از بینایی کامپیوتری

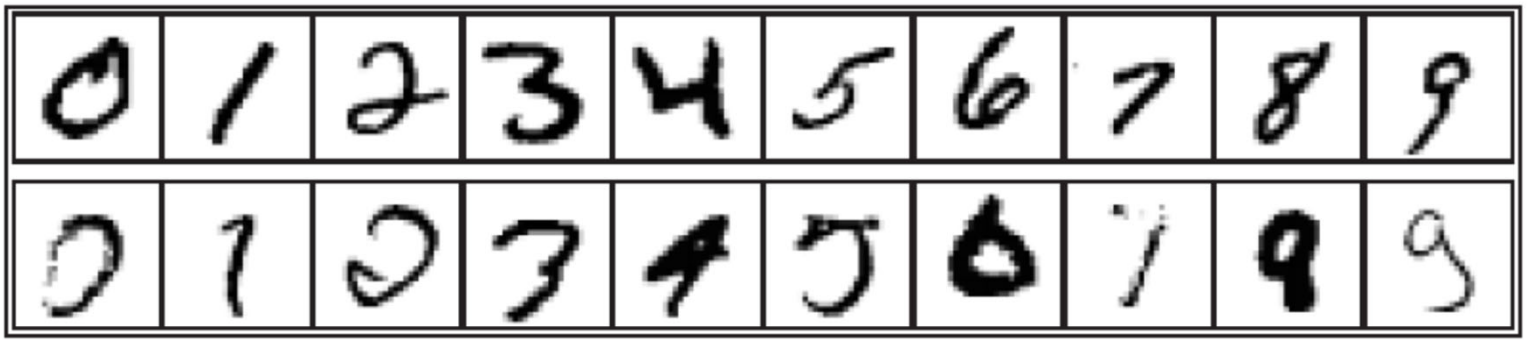
کاربردها

شناسایی زیست‌سنجی *Biometric Identification*	بازیابی تصویر مبتنی بر محتوا *Content-Based Image Retrieval*	آشکارسازی چهره *Face Detection*	بازشناسی چهره *Face Recognition*	بازشناسی پلاک خودرو *Car Plate Recognition*
حفاظت مبتنی بر تصویر *Image-Based Protection*	واسط کاربر مبتنی بر بینایی *Vision-Based User Interface*	بازشناسی امضا *Signature Recognition*	بازشناسی دست‌خط *Handwriting Recognition*	بازشناسی نویسه‌ها *Character Recognition (OCR)*

بازشناسی اعداد دست‌نویس

مثالی از بازشناسی اشیا

HANDWRITTEN DIGIT RECOGNITION



3-nearest-neighbor = 2.4% error
400–300–10 unit MLP = 1.6% error
LeNet: 768–192–30–10 unit MLP = 0.9% error

Current best (kernel machines, vision algorithms) ≈ 0.6% error

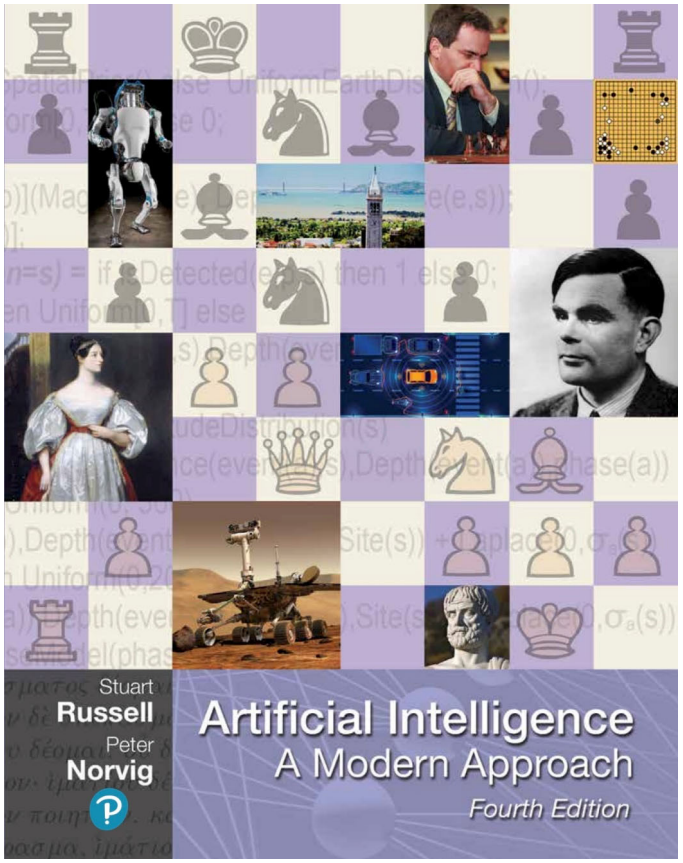
هوش مصنوعی

ادراک: بینایی ماشینی

۷

منابع، مطالعه، تکلیف

منبع اصلی



Stuart Russell and Peter Norvig,
Artificial Intelligence: A Modern Approach,
4th Edition, Prentice Hall, 2020.

Chapter 25

CHAPTER 25

COMPUTER VISION

In which we connect the computer to the raw, unwashed world through the eyes of a camera.

Most animals have eyes, often at significant cost: eyes take up a lot of space; use energy; and are quite fragile. This cost is justified by the immense value that eyes provide. An agent that can see can predict the future—it can tell what it might bump into; it can tell whether to attack or to flee or to court; it can guess whether the ground ahead is swampy or firm; and it can tell how far away the fruit is. In this chapter, we describe how to recover information from the flood of data that comes from eyes or cameras.

25.1 Introduction

Vision is a perceptual channel that accepts a **stimulus** and reports some representation of the world. Most agents that use vision use **passive sensing**—they do not need to send out light to see. In contrast, **active sensing** involves sending out a signal such as radar or ultrasound, and sensing a reflection. Examples of agents that use active sensing include bats (ultrasound), dolphins (sound), abyssal fishes (light), and some robots (light, sound, radar). To understand a perceptual channel, one must study both the physical and statistical phenomena that occur in sensing and what the perceptual process should produce. We concentrate on vision in this chapter, but robots in the real world use a variety of sensors to perceive sound, touch, distance, temperature, global position, and acceleration.

A **feature** is a number obtained by applying simple computations to an image. Very useful information can be obtained directly from features. The wumpus agent had five sensors, each of which extracted a single bit of information. These bits, which are features, could be interpreted directly by the program. As another example, many flying animals compute a simple feature that gives a good estimate of time to contact with a nearby object; this feature can be passed directly to muscles that control steering or wings, allowing very fast changes of direction. This **feature extraction** approach emphasizes simple, direct computations applied to sensor responses. Feature

The **model-based** approach to vision uses two kinds of models. An **object model** could be the kind of precise geometric model produced by computer aided design systems. It could also be a vague statement about general properties of objects, for example, the claim that all faces viewed in low resolution look approximately the same. A **rendering model** describes the physical, geometric, and statistical processes that produce the stimulus from the world. While rendering models are now sophisticated and exact, the stimulus is usually ambiguous. A white object under low light may look like a black object under intense light. A small, nearby object may look the same as a large, distant object. Without additional evidence,