Digitální zpracování obrazu × počítačové vidění zakotvení

Václav Hlaváč

České vysoké učení technické v Praze

Český institut informatiky, robotiky a kybernetiky

160 00 Praha 6, Jugoslávských partyzánů 1580/3

http://people.ciirc.cvut.cz/hlavac, vaclav.hlavac@cvut.cz

také z Centra strojového vnímání, http://cmp.felk.cvut.cz

Osnova přednášky:

- Digitální zpracování obrazů × analýza obrazů × počítačové vidění.
- Vidění × vnímání.
- Proč je vidění těžké?
- Interpretace, význam pro obrazy.

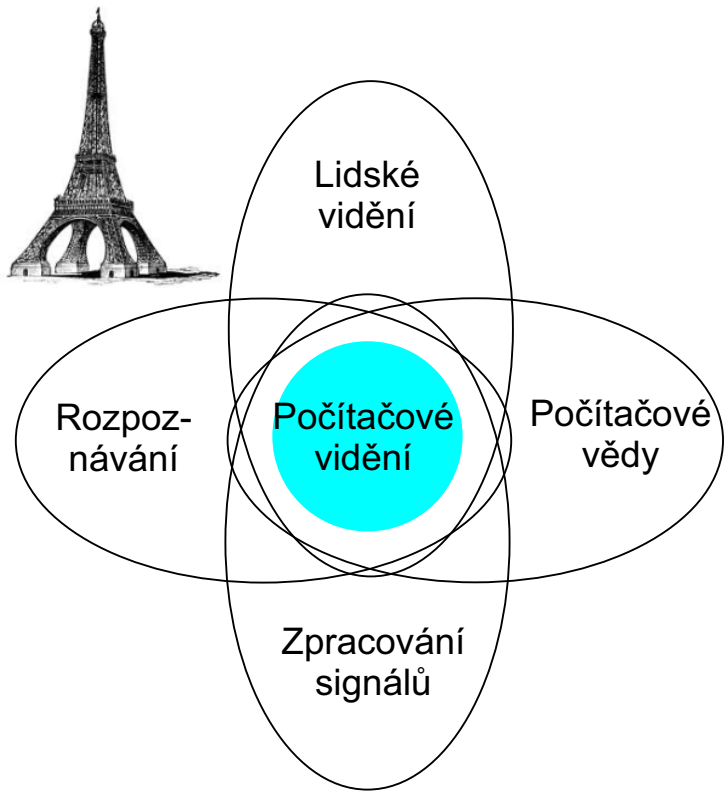
Co je počítačové vidění?

Počítačové vidění, část umělé inteligence, je jak vědou tak i technologií usilující o vytváření strojů schopných vidět a vnímat.

- Vědecký obor:
  hledá teorie pro vytváření umělých systémů získávajících informace z obrazů.
- Technický obor:
  Počítačové vidění = kamera + počítač + ?
- Obrazy (například):
  - pohledy z více kamer,
  - video sekvence,
  - vícerozměrná data z lékařského skeneru, např. tomografu.

Proč studujeme zpracování obrazu, analýzu obrazu a počítačové vidění?

- Počítačové vidění vyrostlo přinejmenším na čtyřech pilířích: (1) vědě o počítačích; (2) teorii signálů; (3) rozpoznávání (též strojové učení); (4) porozumění lidskému vidění.
- Zkoumání započala v 60. letech 20. století.
- Má bohatou metodologii.
- Poskytuje a využívá zajímavé mezioborové vazby.
- Poskytuje vhled do lidského vidění.
- Je důležitým zdrojem informace v moderní informační době.



Na co se používá počítačové vidění?

- Pro detekci, segmentaci, popisování, rozpoznávání (vzletněji: porozumění, vnímání) objektů, které nás zajímají ve 2D či 3D obrazech nebo sekvencích.
- Pro detekci událostí (např. při sledování bezpečnostními kamerami, počítání lidí, při detekci startující balistické rakety ze senzorů na družici).
- Pro uspořádání informace na základě obrazů (např. pro indexování obrazových databází nebo video sekvencí).
- Jako součást řídicích systému (např. u průmyslových robotů nebo autonomně jedoucích aut).
- Pro modelování objektů nebo okolního světa (např. při obrazové kontrole kvality výrobku v průmyslu, při analýze lékařských obrazů, při získávání 3D modelu ze série 2D obrazů).
- Pro interakci mezi člověkem a strojem (např. ovládání počítačové hry pomocí hloubkoměru Kinect).
- atd.

Vnímání

- Postupy k upoutání pozornosti a/nebo porozumění vjemům ze senzorů.
- Úloha je mnohem složitější, než si vědci byli schopni představit okolo let 1950 a 1960:
  - Tehdy: "Vytvoření strojů vnímat potrvá zhruba jedno desetiletí."
  - Přitom jsme od tohoto cíle stále velmi vzdáleni.
- Pět Aristotelových smyslů: zrak, slyšení, hmat, čich a chuť.
- Vnímání předpokládá dynamický vztah mezi: "reprezentaci světa v mozku" na základě $\leftrightarrow$ smyslů, $\leftrightarrow$ bezprostředního okolního světa, $\leftrightarrow$ paměti.

Poděkování za obrázek:
http://www.richardsonthebrain.com/

Poznámky k lidskému (vizuálnímu) vnímání

Co vidíte na obrázku?

Poznámky k lidskému (vizuálnímu) vnímání

Co vidíte na obrázku?

Vidění a vnímání je pro člověka
a mnoho zvířat snadné a přirozené.

-
-
-

Poznámky k lidskému (vizuálnímu) vnímání

Co vidíte na obrázku?

Vidění a vnímání je pro člověka
a mnoho zvířat snadné a přirozené.

- Není to zadarmo:
-
-

Poznámky k lidskému (vizuálnímu) vnímání

Co vidíte na obrázku?

Vidění a vnímání je pro člověka
a mnoho zvířat snadné a přirozené.

- Není to zadarmo:
  - Asi 50 % šedé mozkové kůry u primátů se věnuje zpracování vizuální informace (Felleman-van Essen 1991).
-
-

Poznámky k lidskému (vizuálnímu) vnímání

Co vidíte na obrázku?

Vidění a vnímání je pro člověka
a mnoho zvířat snadné a přirozené.

- Není to zadarmo:
  - Asi 50 % šedé mozkové kůry u primátů se věnuje zpracování vizuální informace (Felleman-van Essen 1991).
  - Lidský mozek spotřebovává asi 20 % veškeré energie v lidském těle.
-
-

Poznámky k lidskému (vizuálnímu) vnímání

Co vidíte na obrázku?

Vidění a vnímání je pro člověka
a mnoho zvířat snadné a přirozené.

- Není to zadarmo:
  - Asi 50 % šedé mozkové kůry u primátů se věnuje zpracování vizuální informace (Felleman-van Essen 1991).
  - Lidský mozek spotřebovává asi 20 % veškeré energie v lidském těle.
- Aby počítač viděl a vnímal jako člověk, museli bychom vyřešit většinu úkolů umělé inteligence (což je velmi těžké, blízké k nemožnému).
-

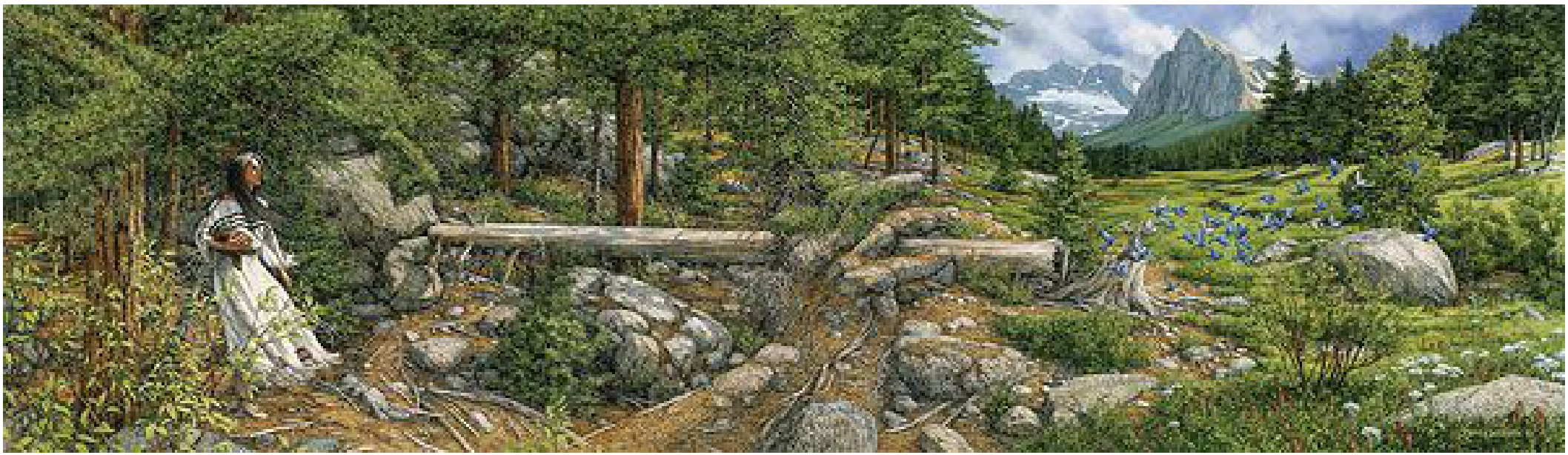
Poznámky k lidskému (vizuálnímu) vnímání

Co vidíte na obrázku?

Vidění a vnímání je pro člověka
a mnoho zvířat snadné a přirozené.

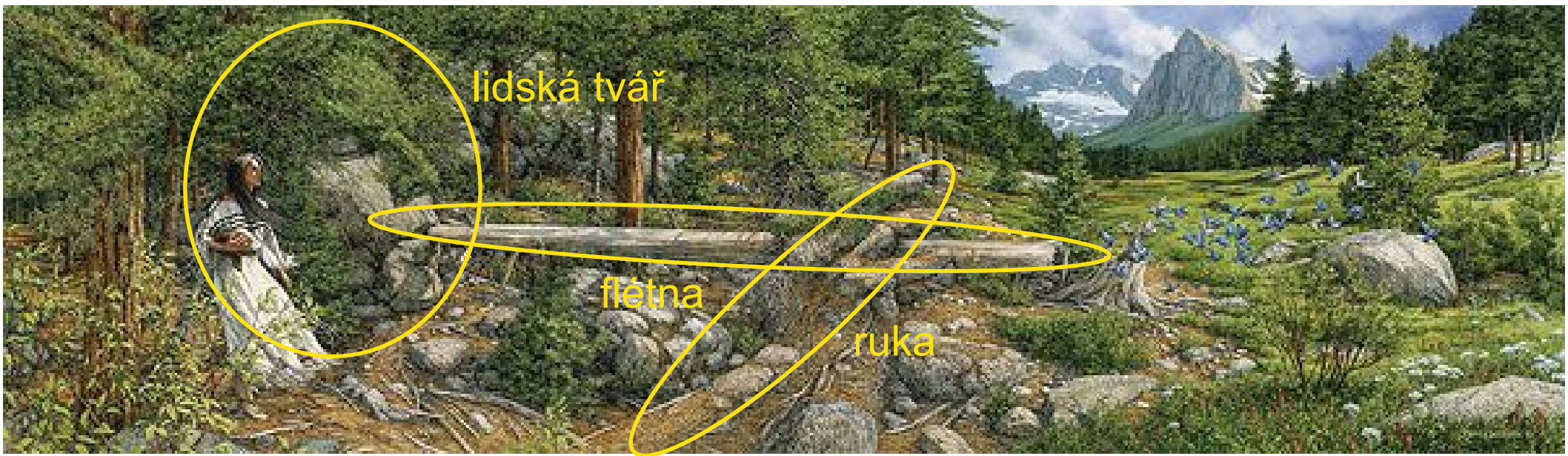
- Není to zadarmo:
  - Asi 50 % šedé mozkové kůry u primátů se věnuje zpracování vizuální informace (Felleman-van Essen 1991).
  - Lidský mozek spotřebovává asi 20 % veškeré energie v lidském těle.
- Aby počítač viděl a vnímal jako člověk, museli bychom vyřešit většinu úkolů umělé inteligence (což je velmi těžké, blízké k nemožnému).
- Využívá se znalost vyšší úrovně, sémantická informace a kontext.

Příklad, obraz s objekty skrytými kamufláží A



Americká malířka Bev Dolittle *1949, protagonistka techniky kamufláže

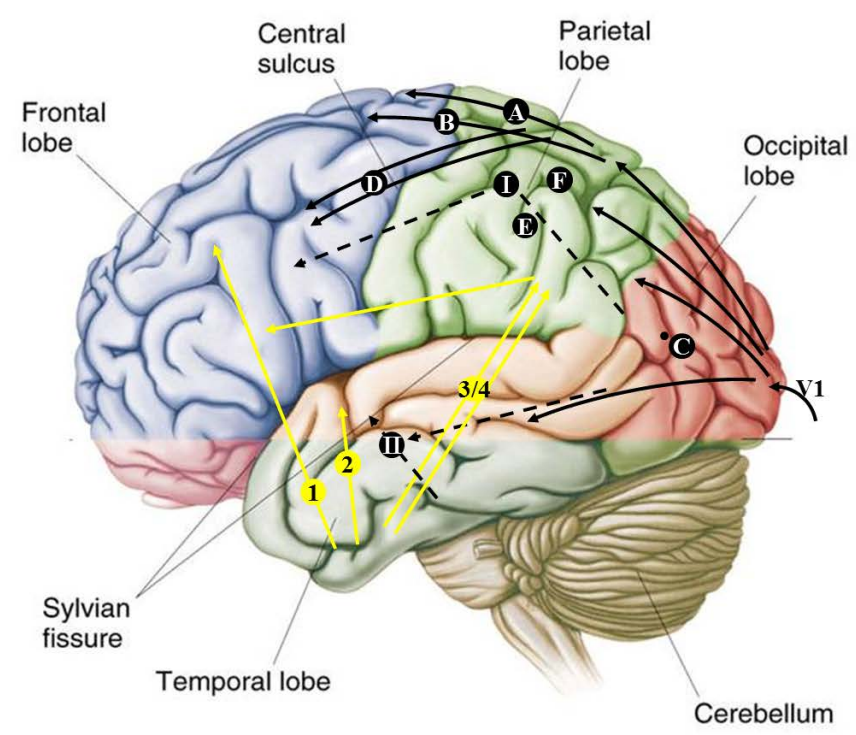
Příklad, obraz s objekty skrytými kamufláží B



Americká malířka Bev Dolittle *1949, protagonistka techniky kamufláže

Lidské vidění

- Část šedé kůry mozkové věnující se vidění zaobírá 50 % mozku makaka.
- U člověka se větší část mozku věnuje vidění než jiným úkolům.



- A Navigation, direction, obstacles
- B Saccade control
- C Pursuit control
- D Understanding, focus, manipulation control
- E Drawing, writing
- F Calculation
- 1 Categorization
- 2 Recognition, objects in context
- 3 Calculation
- 4 Reading
- I Recognition of actions
- II Emotions

Lidské vidění na rozdíl od počítačového vidění

Vidění dovoluje člověku i zvířeti vnímat a porozumět světu, který je obklopuje.

Kognitivní vědy zkoumají také vidění v biologických systémech:

- Hledají empirické modely popisující biologické vidění.
- Někdy popisují vidění, jako by šlo o výpočetní systém.

Počítačové vidění usiluje o technická řešení, i když se někdy inspiruje v biologickém vidění:

- Biologické vidění zvládá úlohy, na něž je počítačové vidění stále krátké. Přesto poskytuje biologické vidění inspiraci i pro technická řešení.
- Technické požadavky na systémy počítačového vidění se často shodují s požadavky na biologické vidění.

Varování: Napodobování biologického vidění nemusí být nejlepším příkladem řešení technické úlohy.

Proč je počítačové vidění těžké ?
Najděme alespoň 6 příčin.

Proč je počítačové vidění těžké ?
Najděme alespoň 6 příčin.

3D $\rightarrow$ 2D přináší ztrátu informace díky vlastnostem perspektivní transformace (matematická abstrakce, dírková komora).

Měřený jas je dán složitým fyzikálním postupem vytváření obrazu. Zář (angl. radiance) ($\approx$ jas) závisí na typu světelných zdrojů, jejich poloze, intenzitě, poloze pozorovatele, lokální geometrii povrchu a odrazivosti povrchu. Obrácená úloha je špatně podmíněna.

Nevyhnutelná přítomnost šumu v každém měření ve skutečném světě.

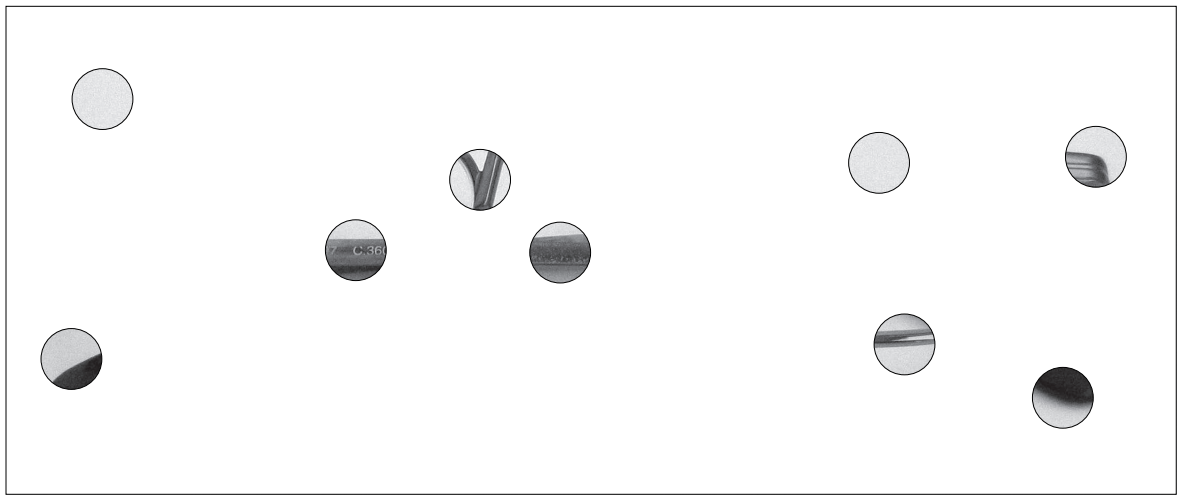
Příliš mnoho dat Stránka A4, 300 dpi, 8 bitů na pixel = 8,5 Mbajtů.

Neprokládané 4k video 3840 $\times$ 2160, RGB (24 bitů) $\approx$ 9.96 Gbitů/sekundu.

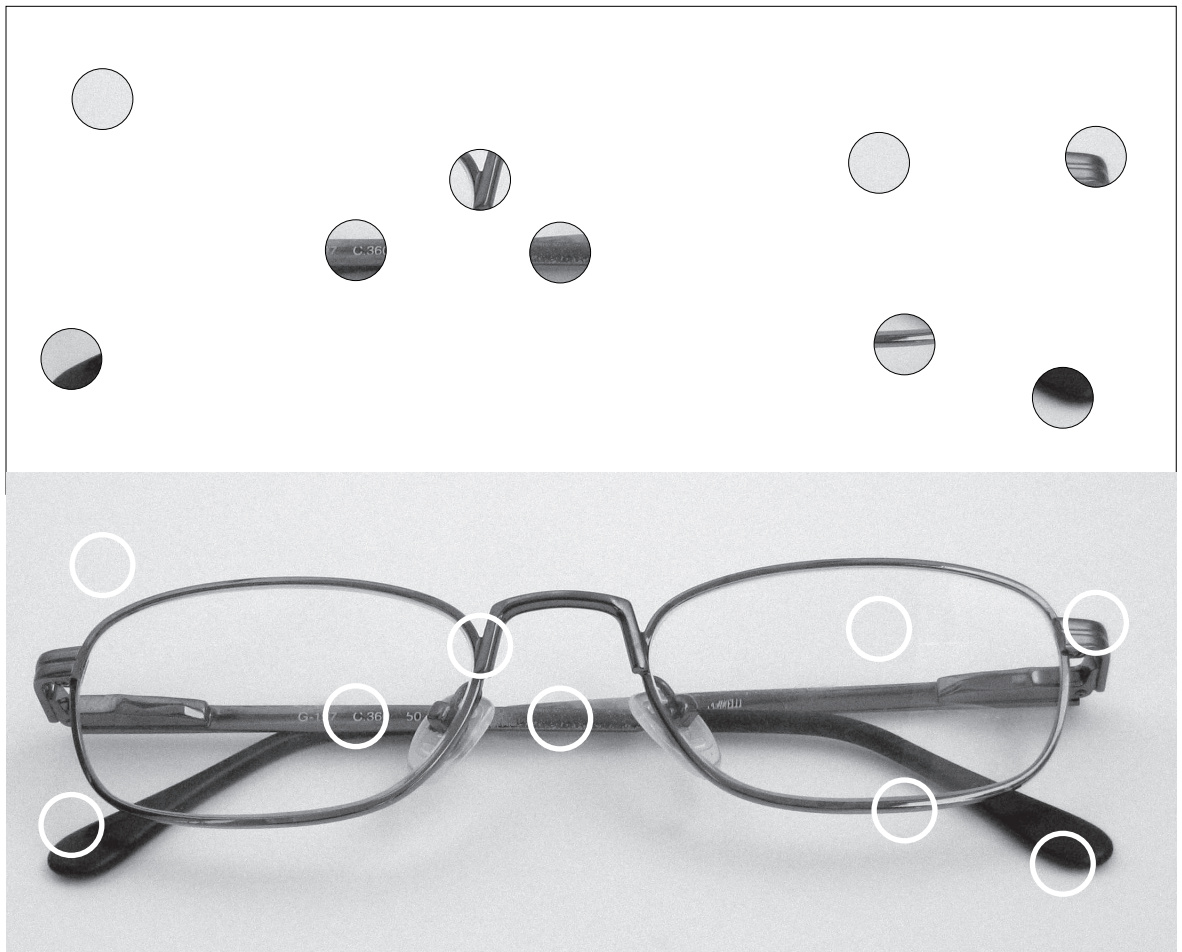
Nutnost zahrnout interpretaci (bude brzy diskutováno).

Lokální okno v kontrastu s potřebou globálního pohledu.

Ilustrace nedostatečnosti lokálního pohledu



Ilustrace nedostatečnosti lokálního pohledu



Od pixelů k symbolům: úloha ukotvení

Člověk dokáže zasadit vjemy do kontextu, který zná. Tato schopnost je formalizována pojmem "interpretace". Pokusím se tento pojem ilustrovat matematicky, a to prostřednictvím konceptu zobrazení, $\longrightarrow$

Přiblížení:

obrázek $\longrightarrow$ pozorování (ze senzorů) $\longrightarrow$ příznaky $\longrightarrow$ kategorie $\longrightarrow$ symbol, jeho jméno

Příklad:

obrázek jablka $\longrightarrow$ červený kulatý objekt na běžícím pásu $\longrightarrow$ "jablko, třída kvality 2"

Hlavní zpráva:

Symbol, jako je například "jablko", sám o sobě nemá žádný význam. Užitečným se stává až ve spojení se smyslovými vjemy, činností a kontextem daného úkolu. V opačném případě představuje interpretace pouze "symboly odkazující na jiné symboly."

Interpretace: vybíráme model vysvětlující naše pozorování

Přiblížení:

Observation $\longrightarrow$ grounded symbols $\longrightarrow$ logical/task model $\longrightarrow$ decision/action

Příklady:

Window image $\longrightarrow$ rain, no-rain $\longrightarrow$ take umbrella?

Apple on belt $\longrightarrow$ class 1, class 2, class 3 $\longrightarrow$ sort apple

Traffic scene $\longrightarrow$ car, plate region, characters $\longrightarrow$ read number plate

Hlavní zpráva:

Computer vision is not only measuring images. It maps visual evidence into a representation meaningful for the task. The "right" interpretation depends on prior knowledge and purpose.

Theoretical background: mathematical logic, theory of formal languages.

Čistě formální usuzování čelí teoretickému prokletí

Varování z matematické logiky – neformálně

- Formální systémy manipulují se symboly podle pravidel, avšak "samotná syntaxe neposkytuje sémantiku". Gödelovy věty o neúplnosti (1929, 1931) přinášejí závažný příklad: každý dostatečně vyjadřovací schopný a bezesporný formální systém aritmetiky obsahuje pravdivá tvrzení, která v rámci daného systému nelze dokázat.
- Pro oblast počítačového vidění je praktické ponaučení skromnější: "interpretaci nelze redukovat pouze na manipulaci se symboly, které nemají oporu v realitě". Symboly musí být propojeny s pozorováními, s předchozími znalostmi a v konečném důsledku i s vnějším světem.

Závažný filozofický problém: Gödelovy věty o neúplnosti; program Davida Hilberta, jehož cílem bylo nalézt úplnou a bezespornou soustavu axiomů pro veškerou matematiku, je neuskutečnitelný. Přesněji řečeno, žádný bezesporný systém axiomů, jehož teorémy lze vypsat pomocí algoritmu, není schopen dokázat všechny pravdy o aritmetice přirozených čísel. B. Russell a A.N. Whitehead byli spoluautory monumentálního díla Principia Mathematica (1910–1913). Ludwig Wittgenstein a jeho Tractatus Logico-Philosophicus (sepsáno 1918, vydáno 1921).

Od nízké k vyšší úrovni zpracování
z hlediska využívané apriorní znalosti

Nízká (až žádná) znalost ≈ digitální zpracování obrazu

- Obrazy se neinterpretují, a proto jsou postupy nezávislé na konkrétní aplikační oblasti.
- Používají se metody zpracování signálů, např. 2D Fourierova transformace.

Střední znalost ≈ analýza obrazu

- Často jen 2D obrazy, např. obrazy buněk v optickém mikroskopu.
- Interpretace přináší důležitou dodatečnou znalost umožňující řešit i úlohy, které by jinak řešit nešly.

Vyšší znalost ≈ počítačové vidění, porozumění obsahu 3D scény z obrazů a videí

- Nejobecnější formulace úloh, 3D svět, měnící se scéna (videosekvence).
- Složité, využívá se interpretace, zpětné vazby a techniky umělé inteligence.
- Příliš ambiciózní cíle. Často špatně podmíněné a příliš těžké úlohy. Obvykle se musí radikálně zjednodušit.

Role apriorní znalosti, protipříklad

- Apriorní znalost "našeho světa" umožňuje člověku porozumět i mnohoznačným obrázkům.
- Ovšem, apriorní očekávání mohou také přivést k chybné interpretaci ...
- Protipříklad: Amesova židle.

Vidíme židle.

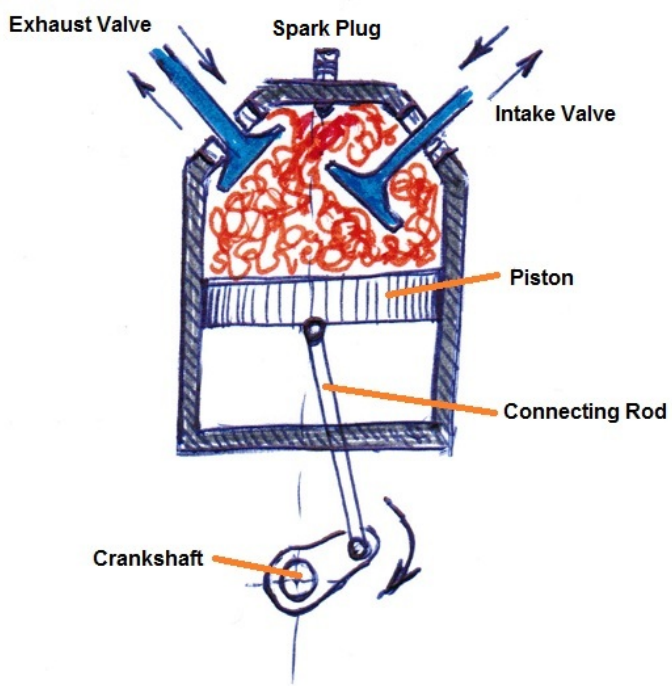
Role apriorní znalosti, protipříklad

- Apriorní znalost "našeho světa" umožňuje člověku porozumět i mnohoznačným obrázkům.
- Ovšem, apriorní očekávání mohou také přivést k chybné interpretaci ...
- Protipříklad: Amesova židle.

Vidíme židle.

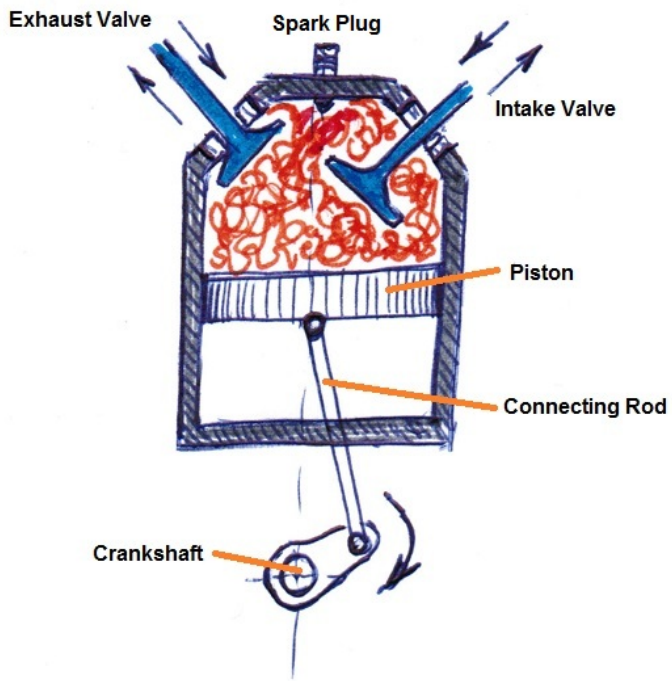
Ve scéně židle nejsou.

Jsou úlohy počítačového vidění typickou inženýrskou úlohou?



Jednoduchá myšlenka . . .

Jsou úlohy počítačového vidění typickou inženýrskou úlohou?



Jednoduchá myšlenka . . .

Myšlenka potřebuje jen trochu doladit . . .

Ultrakrátká historie počítačového vidění

1966 Seymour Papert zadává "summer vision project" jako letní práci pro tým studentů.

∼**1960** Interpretace v omezeném umělém světě, např. svět kostek robotu.

∼**1970** Jistý pokrok v interpretaci obrazů v omezeném světě.

∼**1980** Umělé neuronové sítě přišly a odešly; posun zájmu ke geometrii a rigoróznějšímu použití matematiky; inspirace biologickým viděním (D. Marr a spolupracovníci).

∼**1990** Detekce a rozpoznávání lidských obličejů; růst popularity statistické analýzy; zájem o geometrické úlohy vidění.

∼**2000** Rozpoznávání ve větším; k dispozici začínají být rozsáhlé anotované databáze; počátek prakticky použitelných metod analýzy videa.

∼**2000** Broader recognition; large annotated datasets available; video processing starts; ability to use Markov random fields for images.

∼**2012** AlexNet (deep learning breakthrough)

∼**2015** ResNet, ImageNet dominance

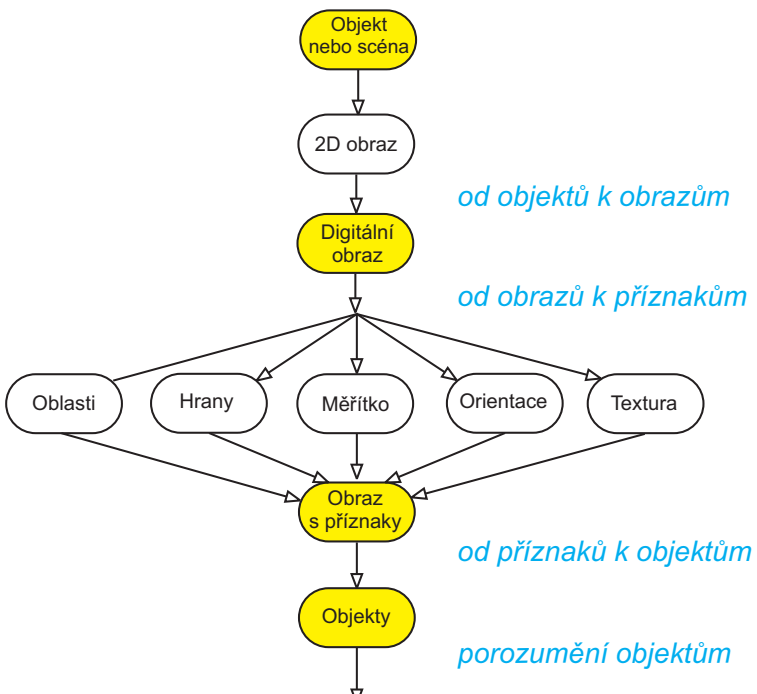
∼**2020** Vision Transformers

∼**2022** Foundation models (CLIP, diffusion)

∼**2022** Multimodal models (GPT-4V)

Rozpoznávání na základě obrazů: Hierarchie reprezentací

Computer vision is seen as an engineering pipeline



After the year 2012

Computer vision is seen as

- data driven and
- uses learned representations