CCF-GAIR 全球人工智能与机器人峰会
30th June 2018, Shenzhen

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Altizure

Computer Vision, Visual Recognition and 3D Reconstruction

计算机视觉，识别与三维重建

Prof. Long QUAN 权龙

The Department of Computer Science and Engineering, HKUST

香港科技大学

Altizure founder 创始人

Overview --- personal reflections

- From AI to Computer vision
  深度学习下人工智能的崛起
- The fundamentals of computer vision
  计算机视觉的基础
- Technological revolution
- Commercial ramifications
- The state-of-the-art Altizure.com
- Conclusions

AI and computer vision
人工智能与计算机视觉

- AI, computer to ‘see’, to ‘hear’ and to ‘read’

  人工智能，让计算机去‘看’， 去‘听’， 去‘读’

  - ‘I see.’ I understand.
  - understanding of visual (image), audial (speech), and languages
    - Computer vision, speech recognition, natural language processing
- Visual is the core

  视觉是核心

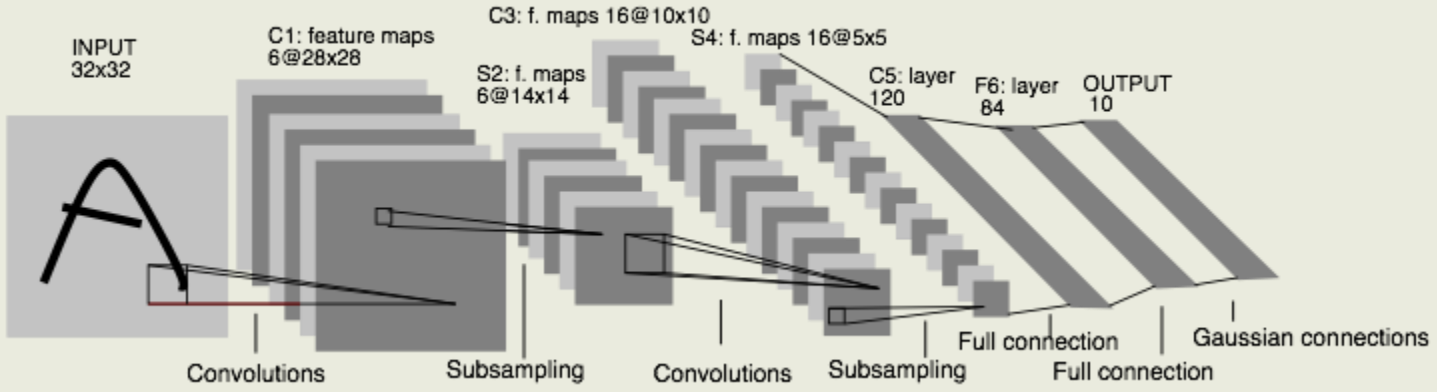
  - 80% of our sensing is visual!
  - The hardest of perception
- The AI (R)evolution comes from computer vision!

  人工智能的发展只能来自视觉

2012, a year of no significance
公元二零一二年

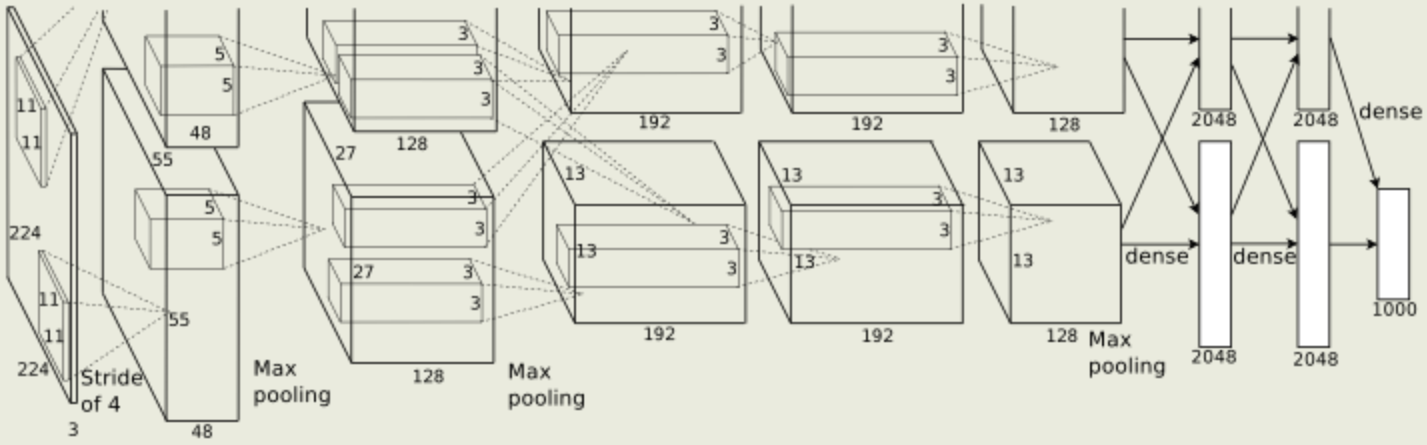
- CVPR 2012
  - 2012, annual gathering of computer vision researchers, like all years
    - An improvement of 10% (from 75 to 85) in ImageNet competition
    - Computer vision researchers use machine learning techniques to recognize labeled objects in large amount of images
- Back to 1998, convolutional neural networks
    - The simple and not trendy visual classification of textual, hand-written digits, from 0 to 9, recognition
- Forward to 2016 AlphaGo beats professionals
    - A narrow AI program

Convolutional Neural Networks: 1989. Input 32*32. CPU
1989 年的卷积神经网络，黑白图像32*32，只有CPU



LeNet: a layered model composed of convolution and subsampling operations followed by a holistic representation and ultimately a classifier for handwritten digits. [LeNet]

2012年的卷积神经网络，三通道的图像 224*224， 有GPU
Convolutional Neural Networks: 2012. Input 224*224*3. GPU.



AlexNet: a layered model composed of convolution, subsampling, and further operations followed by a holistic representation and all-in-all a landmark classifier on
ILSVRC12. [AlexNet]

+ data
+ gpu
+ non-saturating nonlinearity
+ regularization

What happened for 15 years?
漫长的十五年间发生了什么？

- Computer power, GPU
- Millions of labeled image data ImageNet by Fei-fei Li

Six years after, 2018

- 2012, five years ago, alexNet, six days on 2 GPU Nvidia GTX 580 training
- now, 18 minutes in one DGX-2

- CVPR 2018, salt lake city, 10 times growth!
- CVPR 2022, New Orlean, how big?

The fundamentals of computer vision
计算机视觉的基础

- 终极目标是对图像的理解，也是 ‘认知’， 但目前知识 ‘感知’
- Image understanding, ultimately, yet faraway, therefore, visual perception

- 探索最基础的视觉特征
- The search of the fundamental visual features,
  - phoneme in speech, character in textural, word in language …
  - Pixel, picture element, not a feature

- 物体的识别，与三维的重建
- and the two fundamentals of recognition and reconstruction
  - (Features)
  - Recognition
  - Reconstruction

The brief history of computer vision
计算机视觉简史

- Classical period 1980s
  - Primal sketches, edges
  - line segments for reconstruction
- Geometry epoque 1990s and 2000s
  - Hand-engineered SIFT
  - Bag of words, and SVM for recognition
  - 3D point reconstruction
- CNN era 2012
  - Learned features by CNN
  - Classisifcation and regression by NN
  - End-to-end
- It is a revolution, but also an evolution with higher-dimensional features and automatic feature extraction by learning

The history is historical

- “In the late 1990s, neural nets and backpropagation were largely **forsaken by the machine-learning** community and **ignored by the computer-vision and speech-recognition** communities.”

- The 2012 alexnet participated the competition, but not published in CVPR

The human visual system not only recognizes objects, but is able to understand scenes, and processes 3D geometric information to interact with the world
人类的视觉不只是识别，需要三维感知与环境交互

- The second layer of reconstruction in addition to recognition
- CNN mostly successful only in recognition
- Reconstruction drove the computer vision in engineering features before CNN

The convolutional neural networks (CNN) are a computer vision machine!

- The key example of a successful application of insights obtained by studying the brain to machine learning applications.
  - Neural networks greatly inspired by neuros, but technically nothing to do with it, like birds and airplanes
- The first deep model to perform well, long before arbitrary deep models were considered viable
- The first to solve the commercial applications, reading checks
- The CNN by LeCun, “carried the torch for the rest of deep learning and paved the way to the acceptance of neural networks in general”

The state-of-the-art of visual recognition: a statistical classification in supervised CNN

- Any thing you can clearly define and label
- Then show a few thousands examples (labeled data) of these things to the computer
- **A computer recognizes a new image, not seen before, now as good as humans, even better!**
- This is done by deep neural networks.

- It is task-specific
- It memorizes and 'recognizes' better than human, but does not 'understand' better!
- **It is as dump as it was before!**
  - We can label a cat for a dog
  - We can take dogs and wolves for the genus canis
  - We can also distinguish Samoyed dogs from white wolves

The state-of-the-arts of visual features

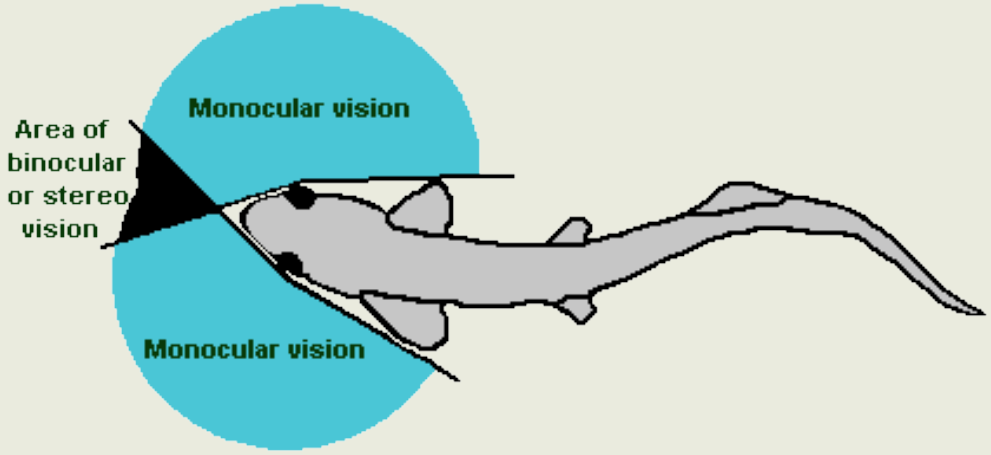
- Hand-craft features (finite, small, well-understood, SIFT) for geometry and reconstruction first, then recognition as bags of words or features
  - A few hundreds
- Learnt CNN features (huge, hierarchical and compositional, less-understood, more powerful) for recognition
  - A few millions

Depth, stereopsis, and reconstruction
深度，视差，重建

- We have two eyes to perceive depth! stereopsis
- Computationally, it is triangulation, while GPS is trilateration!
- This belongs to the classical applied mathematics and optimization, from geometry to photogrammetry and to computer vision …
- Challenge:
  - Locally, for two images
    - how to find the same thing?
    - find the corresponding features → sparse image matching
    - find the corresponding pixel → dense image matching
  - Globally, for millions of images
    - How to find the overlapping images from arbitrary collections of photos? → image retrieval
    - Remove the undesirable things such as skys → **local image recognition**

- Iphone X moved faceID from 2D to 3D, with depth!

- Stereopsis is the computation of depth information from views acquired simultaneously from different points in space.
- It is mostly confined to mammals with front-facing eyes
- The lower animals have less stereo vision



The modern pipeline of 3D reconstruction
现代三维视觉重建的基础

- Feature engineering, detection and matching --- classical features
- Camera pose reconstruction --- reconstruction of ‘feature’ points (Structure from motion)
- Dense reconstruction --- per pixel reconstruction (Multi-view stereo)
- Surface reconstruction --- Geometric representation (Mesh triangulation)

深度学习下的三维重建的革命
The road to ‘deep’ depth, 3D reconstruction

- Feature detection and matching
  - From hand-crafted features to learned features
  - Very large-scale retrieval
  - Pair-wise matching per se
- Structure from motion
  - From relative poses to global motion averaging
  - From small in-core bundle adjustment to a large distributed global BA
- Multi-view stereo
  - From classification to CNN stereo regression
  - End-to-end point reconstruction

- Learned features and recognition
- Recognition

Demos at altizure.com

Our recent directions and works

- Learned local descriptors
  - Large-scale retrieval
  - Pair-wise matching per se and SFM
- Scale-space matching
- Structure from motion
  - motion averaging
  - distributed global BA
- Multi-view stereo
  - End-to-end CNN stereo regression

The-state-of-the-art of reconstruction, **altizure.cn**, by HKUST

- A portal for reality making of the world
  - From small with **smartphones** and big with **drones**!
  - From inside out
  - No limitation …
- A crowd-sourced Earth
  - More than Google earth
  - Understanding of the objects and photos
  - …
- It's a modern AI platform with self-learning ← data

It is a very powerful tool. Yet we are still fundamentally limited by the mathematical tools at our disposition, and our own understanding of what it is meant by knowing …

- The deep convolutional NN
  - It is a still a statistical classification at very large scale
    - Very large and sparse system
  - 'deep' → hierarchical from local and global
  - 'convolutional' → local and sparse
  - 'NN' → linear classifiers
  - 'shallow' understanding if not at all!

- The media calls it AI, and it is the most powerful tool for mathematical and statistical modeling and optimization!
  - Extremely large-scale, from thousands to millions.
  - Easy to 'implement'.

- It does not have a top-down description of our 'knowledge', a high-level cognitive model
  - It is a bottom-up, data-driven, statistical correlation
  - Most of our knowledge is not of probabilistic nature, nor is it statistical

Conclusions

- Personal reflections of computer vision
- Deja-vu, seeing the same in the 80s for the AI and the fifth generation computer
- Yet, hardware evolution is a reality and changing the world!
  - from big (GPU and others) to small (mobile)
  - from flying cameras, drones, to phone cameras, every one and every there!
- Our computer vision community, and broadly AI, is into a golden age, from a few hundreds to a few thousands
- A general AI is defined by visual understanding, and it is yet profoundly limited by it

Thank you, and Q and A.

An open visual deep learning platform of reconstructing and understanding of the world

- ‘Mapping the world’ is also a deep learning close loop!
  - Users capture and collect data
  - Automatic classification and reconstruction
  - Automatic generation of new samples from reconstructed and recognized data
  - Automatic training of the new models
  - Better and more powerful …
- It is a long term endeavor