✅

Lecture 1

Title	Introduction to Convolutional Neural Networks for Visual Recognition
slide	http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture1.pdf

소개

컴퓨터 비전(Computer Vision)이란? 시각 데이터를 활용한 알고리즘이다.

시각 데이터를 암흑 물질(dark matter)로 비유한다. 암흑 물질이란 우주 대부분의 질량을 차지하고 있는 물질로 “존재”하지만 직접 “관측”할 수는 없는 물질이다. 시각 데이터도 이와 마찬가지로 데이터의 대부분을 차지하지만 이를 이해하고 해석하는 일이 상당히 어렵다.

> ✅ ***즉, 문제는 시각 데이터를 해석하기가 까다롭다는 점이다.***

Computer Vision의 역사

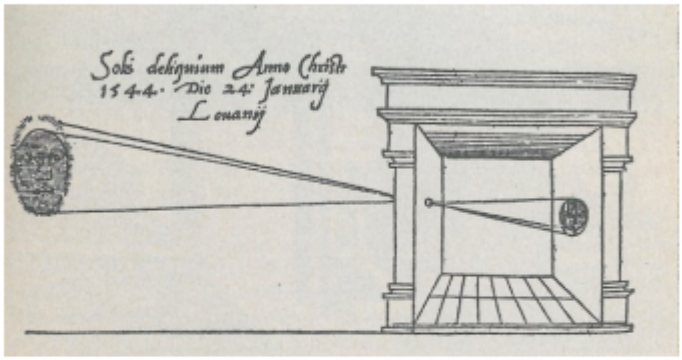
1. **생물학적 비전**

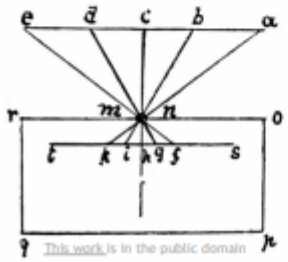
5억 4천만 년 전, 천만 년이라는 아주 짧은 시기 동안 생물의 종이 폭발적으로 늘어났다. 생물학자들은 이것을 캄브리아 폭발이라고 불렀다. 이 현상을 이유로 오스트레일리아의 동물학자가 가장 설득력 있는 이론 중 하나를 제안했는데, 바로 시각(비전)의 탄생이 폭발적인 종 분화의 시기를 촉발 시켰다는 것이다. 이 시기에 동물에게 처음 시각이 생겼고, 시각의 도래는 진화적 군비경쟁을 촉발 시켜 생물들이 하나의 종으로 살아남기 위해서 빠르게 진화해야 했다는 것이다. 현재 시각은 거의 모든 지능을 가진 동물 특히 인간에게 가장 큰 감각 체계로 발전했다.

2. **공학적 비전**

Camera Obscura

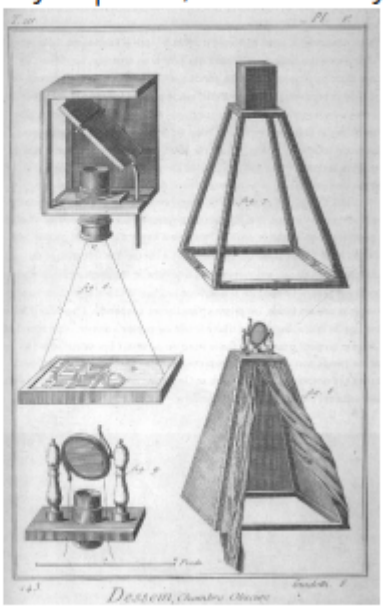
Gemma Frisius, 1545





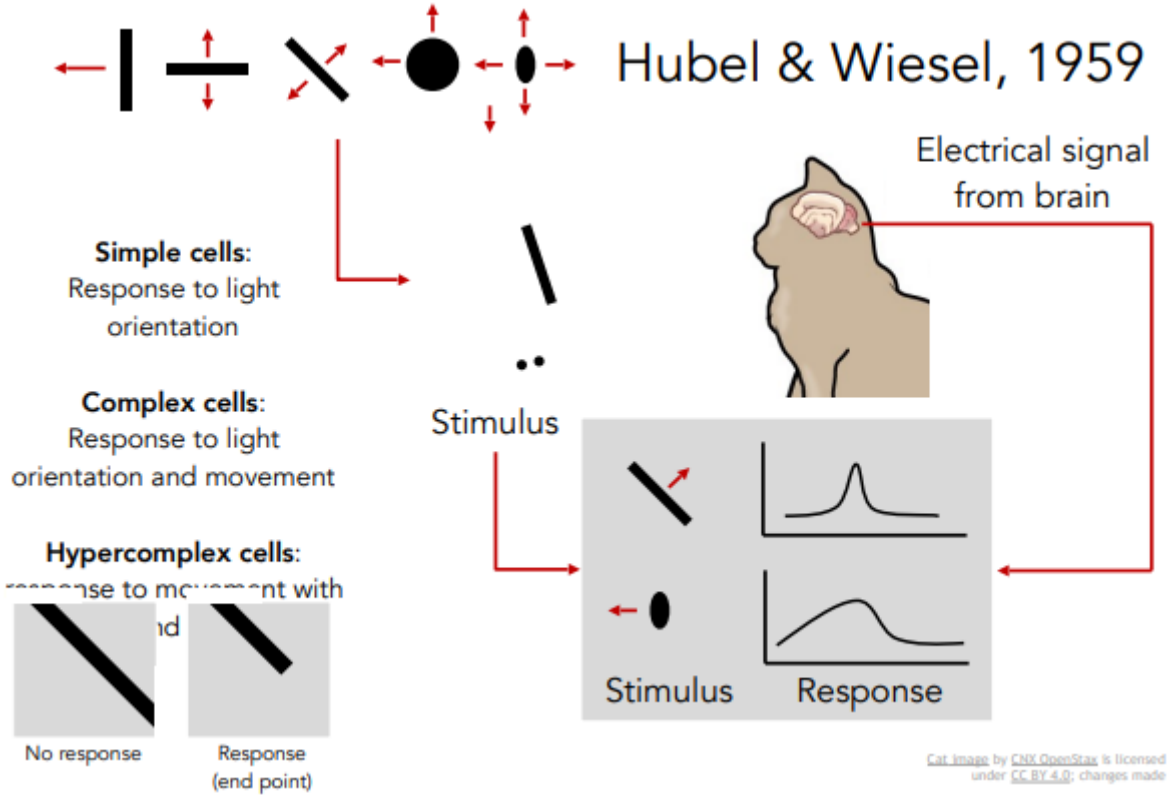
Leonardo da Vinci,
16th Century AD

Encyclopedie, 18th Century



1600년대 르네상스 시대의 카메라인 Obscura는 핀홀 카메라 이론은 기반으로 발명되었다. 이는 생물학적으로 발전한 초기의 눈과 상당히 유사한데, 빛을 모아주는 구멍이 하나 있고 카메라 뒤편의 평평한 면은 정보를 모으고 이미지를 투영한다. 현재는 카메라 기술이 발전하여 스마트폰 카메라나 다른 여러 기기에 이르기까지 사람들이 사용하는 가장 인기 있는 센서 중 하나가 되었다.

3. **비전 매커니즘**



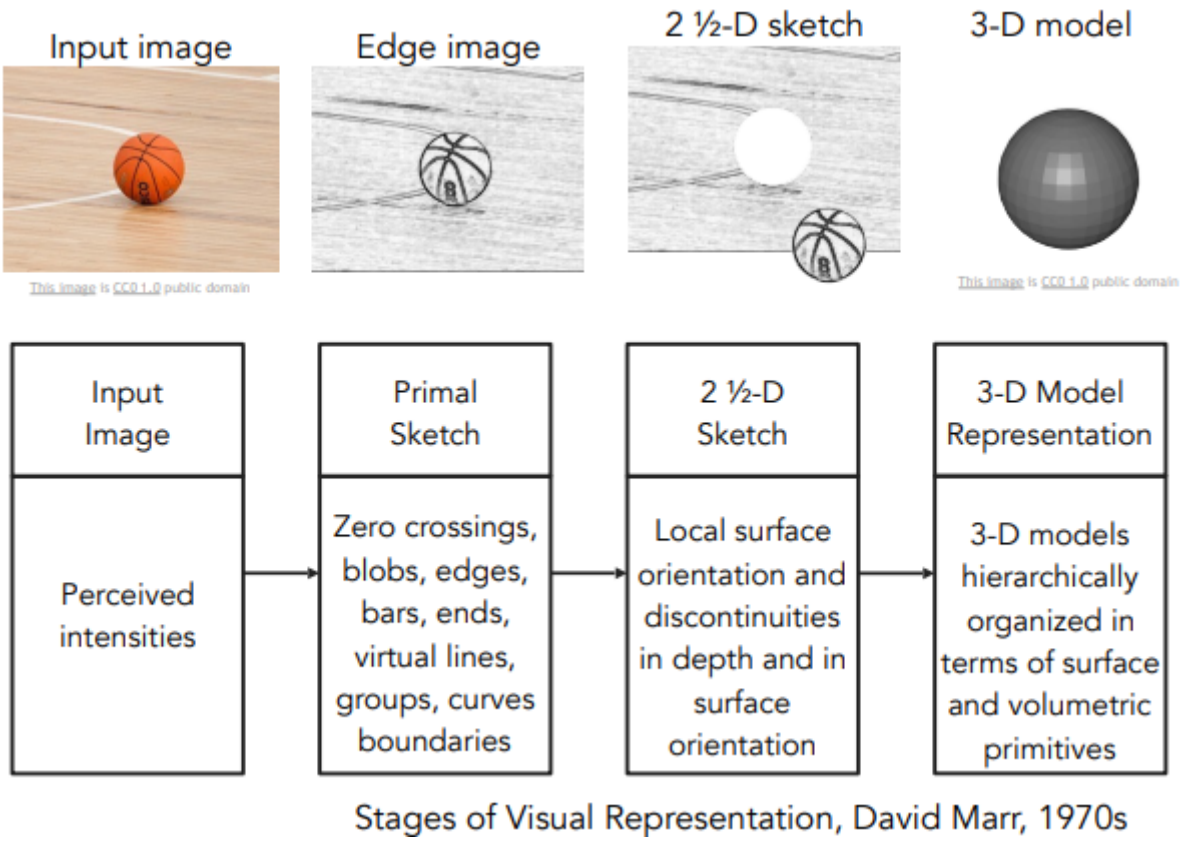
1950~1960년대 “포유류의 시각적 처리 메커니즘은 무엇인가?”의 물음으로 Hubel과 Wiesel가 연구를 진행했다. 인간과 고양이의 시각 처리 매커니즘이 비슷했기 때문에 고양이의 뇌를 통해 연구했다. 고양이 두뇌 뒷면에 전극 몇 개를 꽂고, 어떤 자극을 줘야 일차 시각 피질의 뉴런들이 반응하는지 관찰했다. 이를 통해 일차 시각 피질에 다양한 종류의 세포가 있다는 것을 알았는데, 그 중 가장 중요한 세포들은 경계(edges)가 움직이면 이에 반응하는 세포들이었다. 이 연구의 주된 발견은 **시각 처리가 처음에는 단순한 구조로 시작되며 그 정보가 통로를 거치면서 점점 복잡해진다는 것이다.**

4. 컴퓨터 비전

1960년대

1960년대 초 Larry Roberts는 Vlock World를 통해 우리 눈에 보이는 사물을 기하학적 모양으로 단순화 시켰다. 이 연구의 목표는 우리 눈에 보이는 세상을 인식하고 그 모양을 재구성하는 일이었다. 1966년 MIT The Summer Vision Project를 통해 시각 시스템의 전반을 구현하기 위해서 프로젝트 참가자들을 효율적으로 이용하는 것을 목표로 하였다.

1970년대

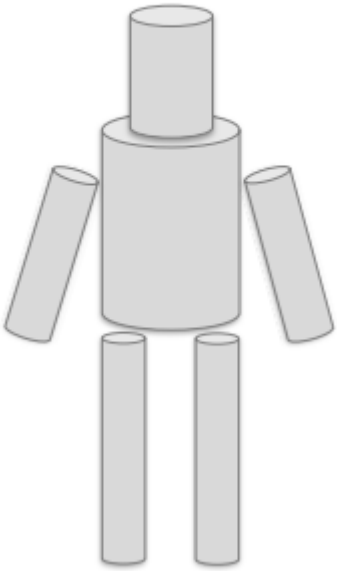


Stages of Visual Representation, David Marr, 1970s

1970년 후반 David Marr은 책을 통해 비전이란 무엇인지, 그리고 어떤 방향으로 컴퓨터 비전이 나아가야 하는지, 그리고 컴퓨터가 비전을 인식하게 하기 위해 어떤 방향으로 알고리즘을 개발해야 하는지를 다뤘다. 그의 저서에서, 우리가 눈으로 받아들인 "이미지"를 "최종적인 full 3D 표현"으로 만들려면 몇 단계의 과정을 거쳐야만 한다고 주장했다. 첫 단계는, "Primal Sketch"라고 하는 단계이다. 이 과정은 주로 경계(edges), 막대(bars), 끝(ends), 가상의 선(virtual lines), 커브(curves), 경계(boundaries)가 표현되는 과정이다. Hubel과 Wiesel은 시각 처리의 초기 단계는 경계와 같은 단순한 구조와 아주 밀접한 관계가 있다고 했었는데, 경계와 커브 이후의 다음 단계는 "2.5-D sketch"라는 단계이며 이 단계에서는 시각 장면을 구성하는 표면(surfaces) 정보, 깊이 정보, 레이어, 불연속 점과 같은 것들을 종합한다. 그리고 결국에 그 모든 것을 한데 모아서 surface and volumetric primives의 형태의 계층적으로 조직화된 최종적인 3D 모델을 만들어 낸다. 이런 방식의 사고방식은 실제로 수십 년 간 컴퓨터 비전 분야를 지배했다.

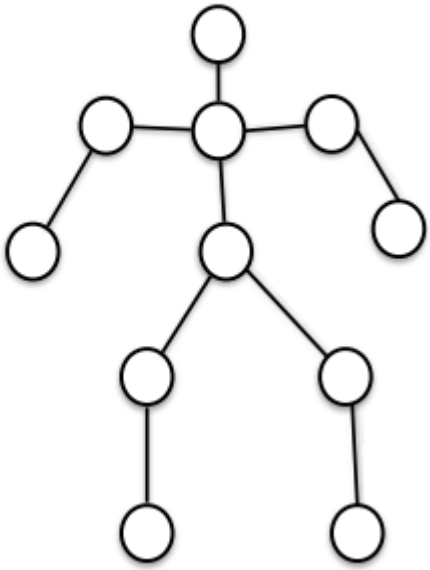
- Generalized Cylinder

Brooks & Binford, 1979



- Pictorial Structure

Fischler and Elschlager, 1973



1970년대 사람들은 "우리는 어떻게 해야 장난감 같은 단순한 블록 세계를 뛰어넘어서 실제 세계를 인식하고 표현할 수 있을까?"라는 질문을 하기 시작했다. 이에 Stanford와 SRI에서 과학자들이 서로 비슷한 아이디어를 제안했다. 하나는 "generalized cylinder"이고 하나는 "pictorial structure"이다. 기본 개념은 "모든 객체는 단순한 기하학적 형태로 표현할 수 있다"라는 것이다. 가령 사람은 원통 모양을 조합해서 만들 수 있고, 또는 "주요 부위"와 "관절"로 표현할 수도 있다. 두 방법 모두 단순한 모양과 기하학적인 구성을 이용해서 복잡한 객체를 단순화시키는 방법이다. 이러한 연구들은 수년간 다른 연구에 상당히 많은 영향을 미쳤다.

1960~1980년까지 계속해서 컴퓨터 비전에 대한 연구와 고민이 이루어졌지만, 객체 인식 문제를 해결하기는 어려웠다. 객체인식이 너무 어렵다면 우선 객체 분할(segmentation)이 우선이 아니었을까 라는 생각으로 객체 분할이 발달하게 되었다. 객체분할은 이미지의 각 픽셀을 의미 있는 방향으로 군집화하는 방법이다. 픽셀을 모아 놔도 사람을 정확히 인식할 수 없지만 배경인 픽셀과 사람이 속해 있을지도 모르는 픽셀을 가려낼 수는 있었다.

1990~2000년대

컴퓨터 비전에서 유난히 발전 속도가 빨랐던 분야는 **“얼굴인식”분야**이다.대략 1999/2000년대에는 "기계학습", 특히나 "통계적 기계학습” 이라는 방법이 점차 탄력을 얻기 시작했다. 가령 "Support Vector Machine", "Boosting", "Graphical models” 그리고 초기 "Neural Network" 등이 있었다. 2001년 Paul Viola와 Michael Jones이 AdaBoost를 이용해 실시간 얼굴인식에 성공 하였다.

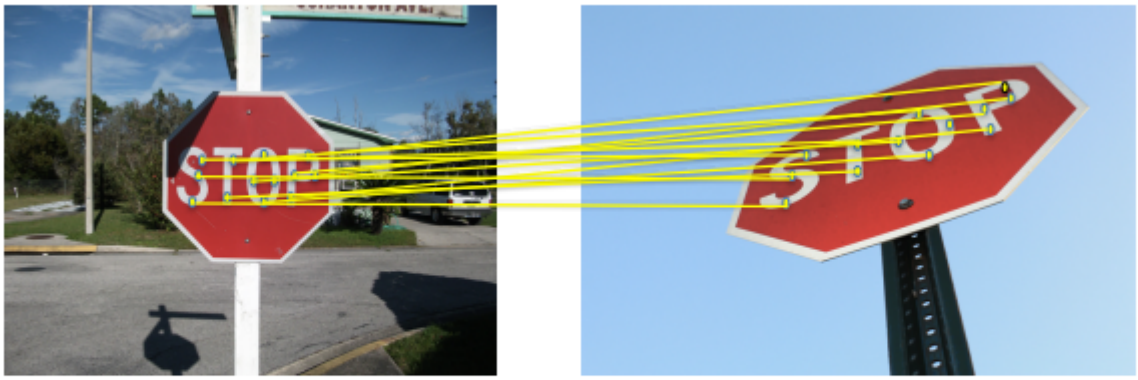


Image is public domain

Image is CC BY-SA 2.0

"SIFT" & Object Recognition, David Lowe, 1999

1990년대 후반부터 2010년도까지의 시대를 풍미했던 알고리즘은 **"특징기반 객체인식 알고리즘"** 이었다. 이 시절 나온 아주 유명한 알고리즘이 바로 David Lowe의 SIFT feature이다. 카메라 앵글이 변할 수 있고, 겹치거나 화각이 변하고 빛도 변하고 객체 자체도 얼마든지 변할 수 있지만 객체의 특징 중 일부는 다양한 변화에 조금 더 강인하고 불변하다는 점을 발견했다. 그리하여 객체인식은 객체에서 중요한 특징들을 찾아내고 그 특징들을 다른 객체에 매칭시키는 과제가 되었다. 미지에 존재하는 "특징"을 사용하게 되면서 컴퓨터 비전은 또 한 번의 도약을 할 수 있었다.

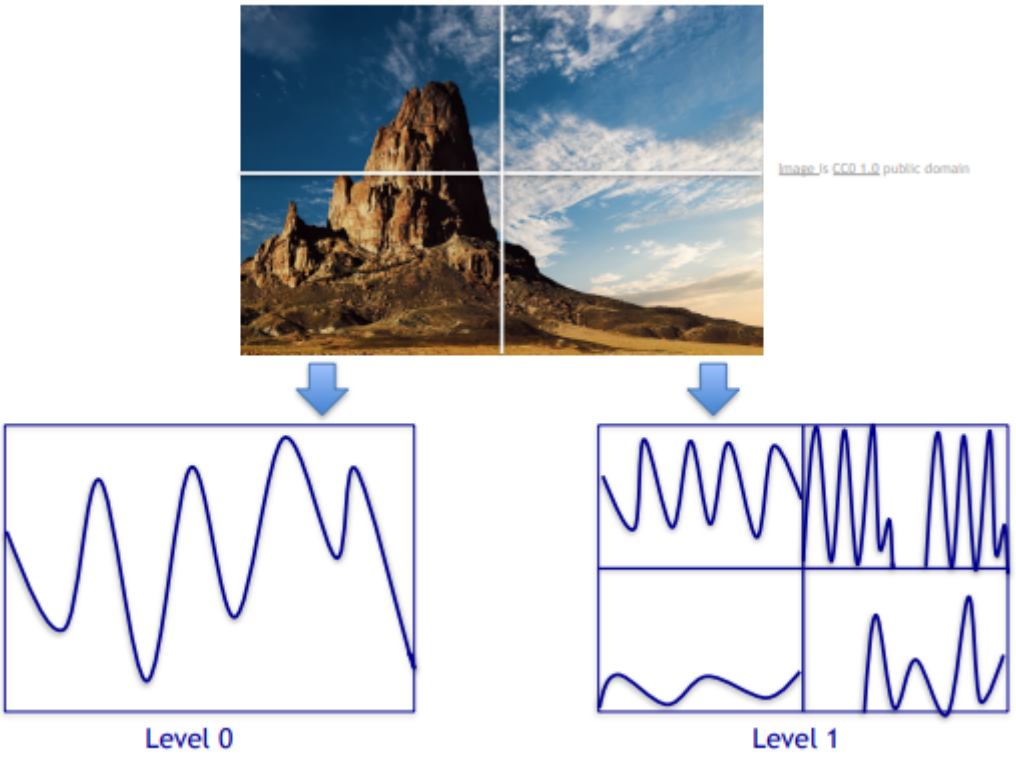
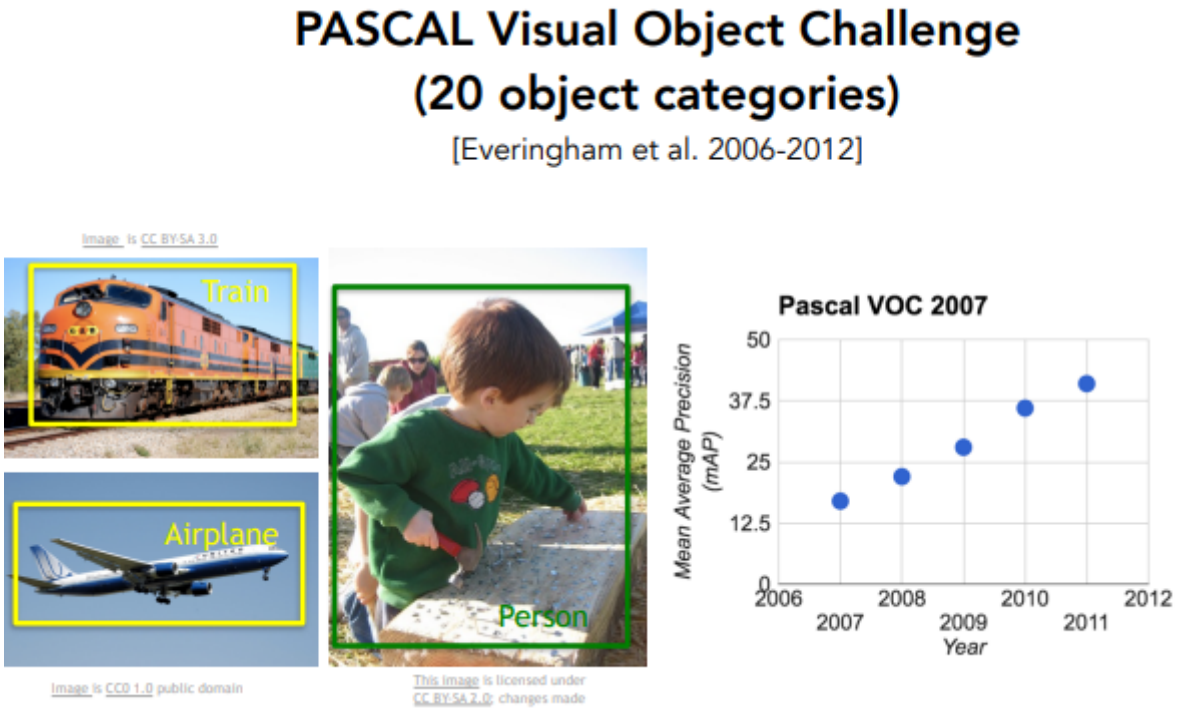


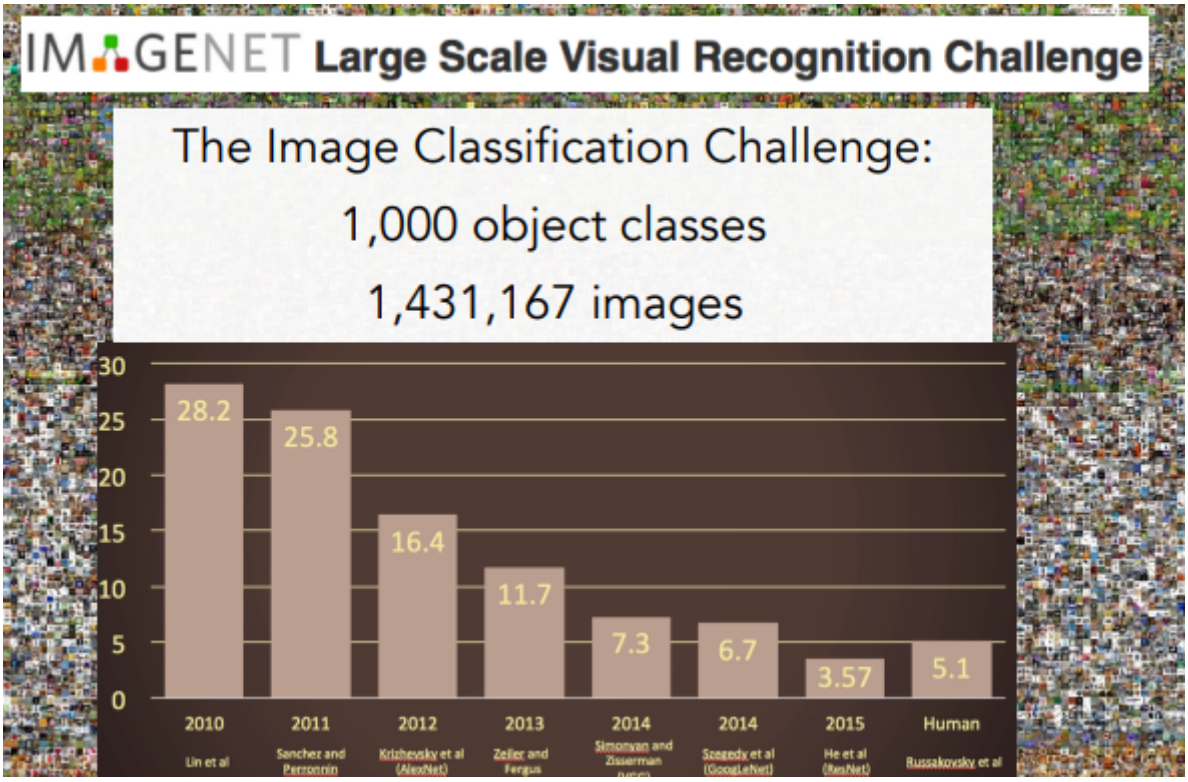
Image is CC0 1.0 public domain

Spatial Pyramid Matching, Lazebnik, Schmid & Ponce, 2006

그리고 **장면 전체를 인식**하기에 이르렀다. 한 예로, Spatial Pyramid Matching이다. 기본 아이디어는 우리가 특징들을 잘 뽑아낼 수만 있다면 그 특징들이 일종의 "단서"를 제공해 줄 수 있다는 것이다. 이미지가 풍경인지, 부엌인지, 또는 고속도로인지 등등이다. 이 연구는 이미지 내의 여러 부분과 여러 해상도에서 추출한 특징을 하나의 특징 기술자로 표현하고 Support Vector Algorithm을 적용한다.



21세기를 맞이하고 사진의 품질이 점점 좋아졌다. 객체인식 성능 측정을 위해 Benchmark Dataset를 모으기 시작했다. 그 중 하나는 PASCAL Visual Object Challenge(VOC)이다. 이 데이터셋에는 20개의 클래스가 있고 기차, 비행기, 사람이 있고 소, 병, 고양이 등으로 구성되어 있다. 데이터셋은 클래스당 수천 수만 개의 이미지들이 있었으며, 다양한 연구 집단에서 이를 통해 알고리즘의 자신들의 알고리즘을 테스트하여 객체인식 성능이 얼마나 좋아졌는지를 판단할 수 있었다.



그러나 아직 두 가지의 motivation이 남아있었다. 하나는 이 세상의 모든 것들을 인식하고 싶다는 것이고 또 하나는 기계학습의 Overfiting 문제를 극복해보자는 것이다. 이 동기를 바탕으로 **ImageNet** 프로젝트를 시작했다. 인터넷에서 수십억 장의 이미지를 다운받았고 WordNet이라는 Dictionary로 정리하여 대략 15만 장에 달하는 이미지와 22만 가지의 클래스 카테고리를 보유하게 되었다. ImageNet 팀은 2009년부터 국제 규모의 대회를 주최하였는데, 2010년도부터 2015년도 오류율이 점차 감소하고 있습니다. 가장 주목해야 할 것은 2012년인데, 처음 2년 동안은 오류율이 약 25%를 맴돌았던 것에 반해, 2012년에 오류율이 16%로 거의 10%가량 떨어졌다. 이때 우승한 알고리즘은 convolutional neural network 모델입니다. CNN은 그 당시 다른 알고리즘들을 능가하고 ImageNet Challenge에서 우승하였습니다.

CS231n overview

이 수업에서 중점적으로 다룰 문제는 **Image Classification**이다. Image Classification의 문제 정의는 알고리즘이 이미지 한 장을 보고, 몇 개의 고정된 카테고리 안에서 정답 하나를 고르는 것이다. 그리고 **object detection**과 **image captioning**도 배울 것이다. 이 이미지가 고양이다, 개다, 말이다 하는 것을 실제로 어디에 있는지 네모박스를 그릴 수 있다. image captioning도 배울 것입니다. 이미지가 입력으로 주어지면 이미지를 묘사하는 적절한 문장을 생성해야 합니다.

최근 컴퓨터 비전 분야의 진보를 이끌어낸 주역은 바로 Convolutional neural networks, 즉 CNN(convnet)이다. 2012년의 CNN의 시대가 도래했고, 이후 CNN을 개선하고 튜닝하려는 많은 시도들이 있었다. 이 강의 전반에 걸쳐 CNN 모델들이 어떻게 동작하는지는 심도 깊게 살펴볼 것이다.