CSCI446/946 Big Data Analytics

Week 11 Advanced Analytical Theory and Methods: Image Analysis

School of Computing and Information Technology
University of Wollongong Australia

Advanced Analytical Theory and Methods: Image Analysis

- Overview of Image Analysis
- Collecting and Representing Image
- Image Recognition
- Bag-of-Visual-Words model
- Deep Convolutional Neural Networks



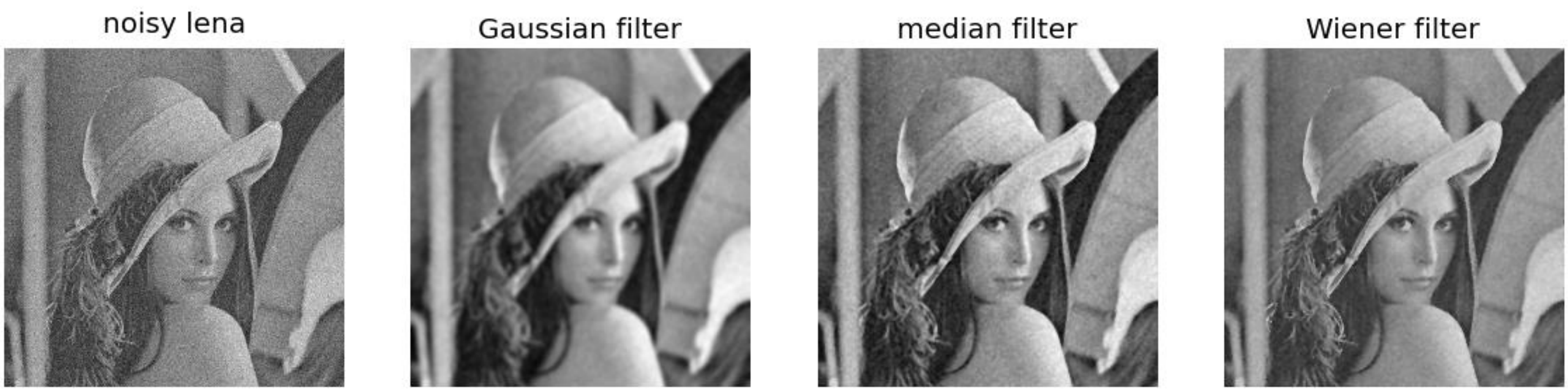


Overview of Image Analysis

- Image analysis
  - Refers to the representation, processing, and modelling of visual data to derive useful insights
  - Suffers from the semantic gap
  - Visual data (image, video, …) is unstructured
- Semantic gap
  - The gap between high-level concepts used by human and the low-level features used by computer

Overview of Image Analysis

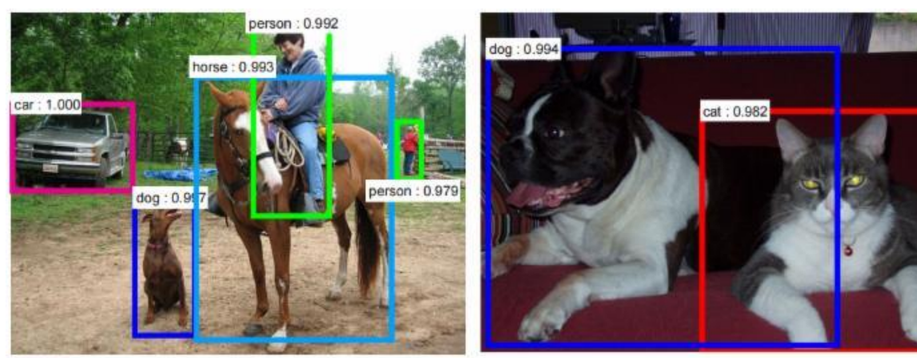
- Image processing
  - Scaling, translation, rotation, ...
  - Filtering, enhancement, histogram equalisation, …



http://www.scipy-lectures.org/scipy.html

Overview of Image Analysis

- Image recognition (in a narrow sense)
  - Image classification
  - Object detection, localisation, tracking
  - Scene segmentation and reconstruction
  - Image search and retrieval



Overview of Image Analysis

- Image classification

Face recognition

OCR recognition

Scene recognition

Object recognition

Overview of Image Analysis

- Object detection, localisation, tracking

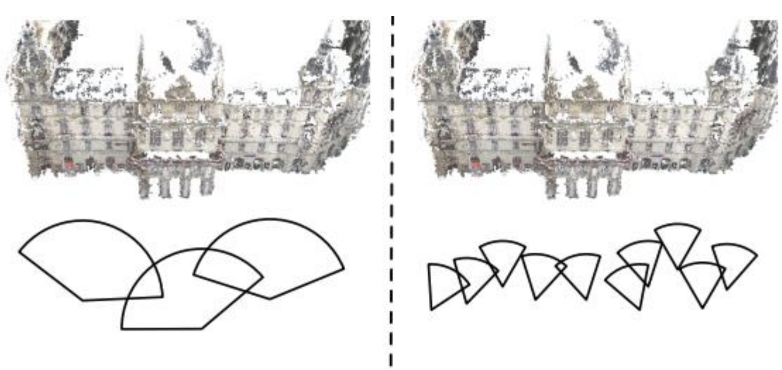
Object detection and localization

Object tracking (https://www.youtube.com/watch?v=dKpRsdYSCLQ)

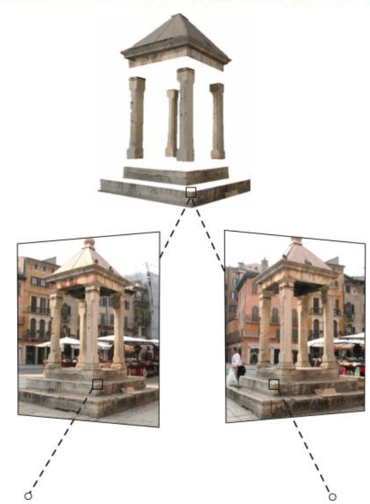
Overview of Image Analysis

- Scene segmentation and reconstruction

[Farabet et al. PAMI 2013]



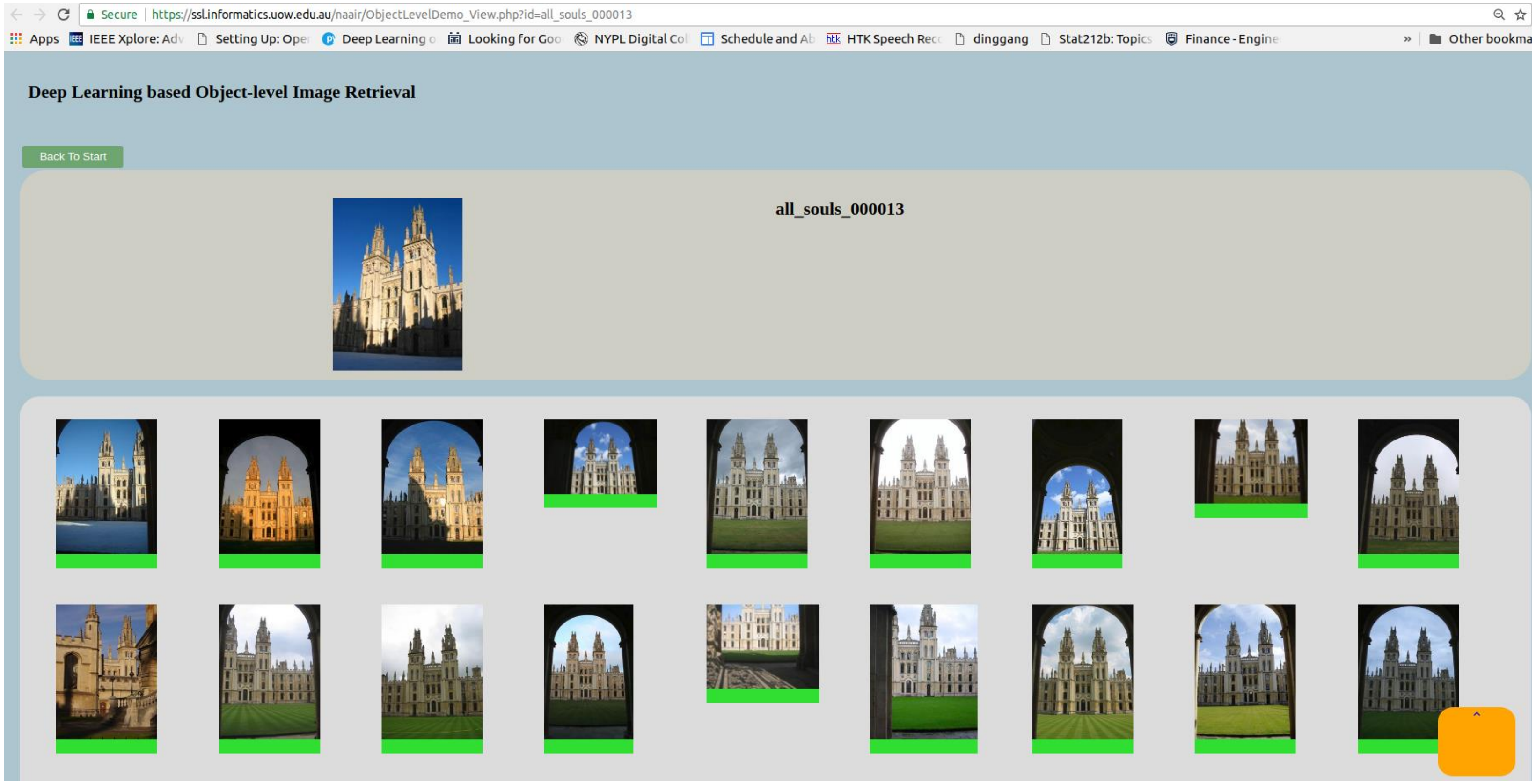
http://twd20g.blogspot.com.au/2011/12/this-work-presents-novel-system-that.html



https://www.3dflow.net/elementsCV/S4.xhtml

Overview of Image Analysis

- Image search and retrieval



https://ssl.informatics.uow.edu.au/naair/ObjectLevelDemo_Home.php

Overview of Image Analysis

- Image recognition benchmark datasets
  - Caltech101 and Caltech256
  - PASCAL VOC project
  - ImageNet IMAGENET
  - Microsoft COCO
  - ORL database of faces
  - Labelled Faces in the Wild
  - A lot more…

Image Analysis Steps

- Collection and labelling
  - Collect representative images from a given task and label the ground truth
- Image representation
  - Select and/or design appropriate image representations (**invariant** and **discriminative**)
- Image analysis techniques
  - Apply and/or design appropriate analysis techniques for the given tasks (classification, detection, tracking, segmentation, etc.)

An Image Analysis Example

- A shop would like to analyse customer shopping behaviour via surveillance video
  - Who are their main customers?
  - What are their shopping patterns?

https://en.wikipedia.org/wiki/Coles_Supermarkets

http://www.dailytelegraph.com.au

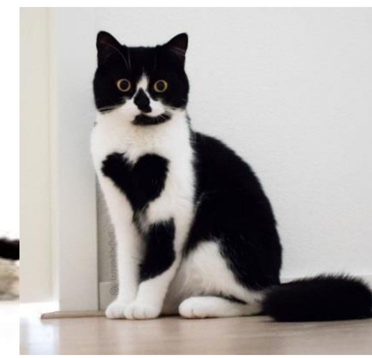
Collecting Raw Image

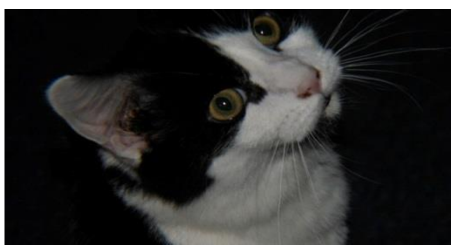
- For image analysis, data must be collected before anything can happen
- The team
  - Starts by actively monitoring various sources for representative images or videos
  - Will deal with unstructured data
- Public APIs and Web scraper
- Be careful about the rights of the owner

http://graphics.cs.cmu.edu/projects/im2gps/flickr_code.html

Representing Image

- Why representing images is difficult?
  - Scale, rotation, illumination, occlusion, background clutter, deformation, …
  - **Invariant** and **Discriminative** representation

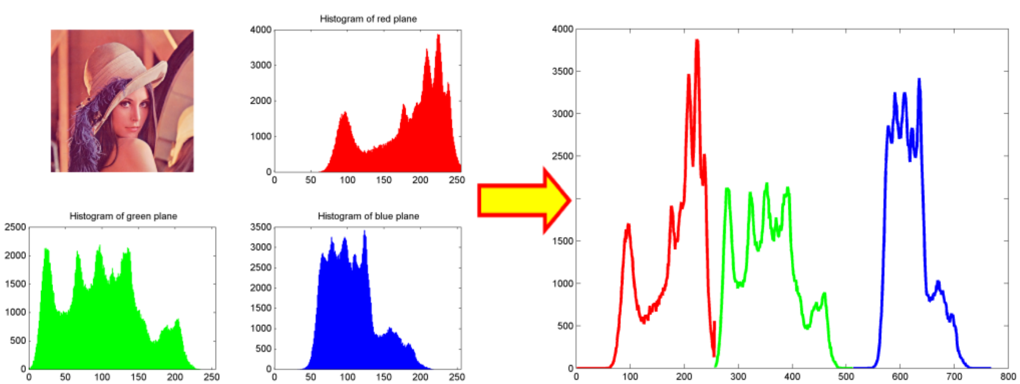




Cat:

Representing Image

- Traditional representation (before year 2000)
  - Hand-crafted, global features
  - Intensity, colour, texture, shape, structure, etc.

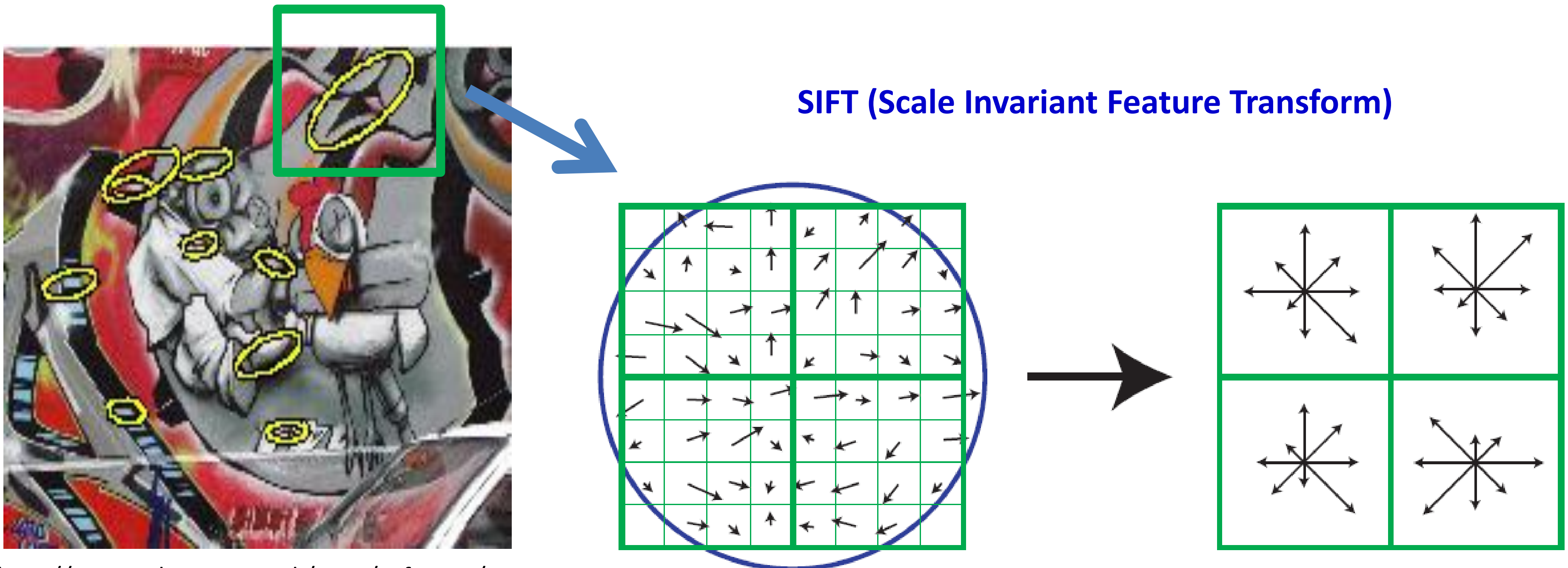


Colour histogram in a RGB space

Face recognition with raw pixel intensities

Representing Image

- Days of the BoVW model (2000 ~ 2012)
  - **SIFT**, HOG, SURF, CENTRIST, filter-based, ...
  - Invariant to view angle, scale, illumination, ...

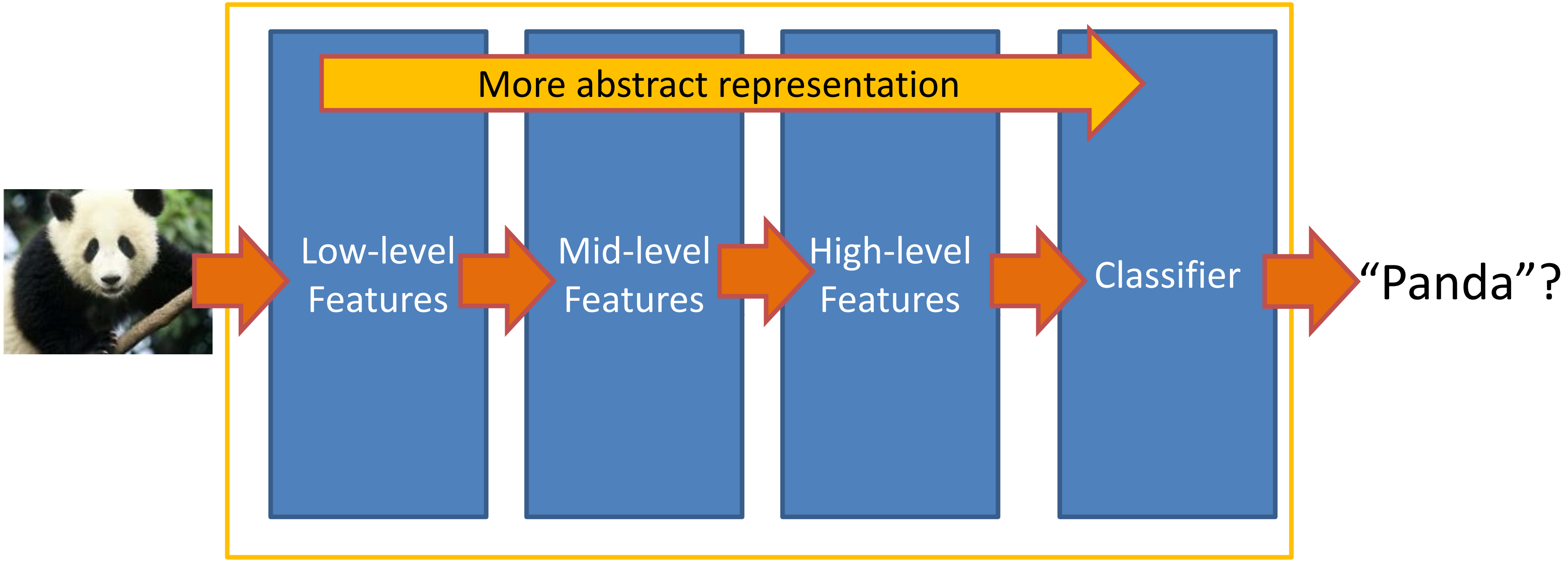


http://www.robots.ox.ac.uk/~vgg/software/

Image courtesy of David Lowe, IJCV04

Representing Image

- Era of Deep Learning (2012 ~ present)
  - Directly learn features representations from data



Deep Learning: train layers of features so that classifier works well.

Image courtesy of M. Ranzato

Bag-of-Visual-Words Model

- Remember the Bag-of-Words representation?
  - A document becomes a high-dimensional vector, indicating the presence/absence/frequency of various words in this document

Grif: An Interactive System for Structured Document Manipulation

Vincent Quint*, Irène Vatton**
**INRIA*
***CNRS*

ABSTRACT
Grif is an interactive system for editing and formatting complex documents. It manipulates structured documents containing objects of various types: tables, mathematical formulæ, programs, pictures, graphics, etc... It is a structure directed editor which guides the user in accordance with the structure of the document and of the objects being edited; the image displayed on the screen also being constructed from that structure. Flexibility is one of the most interesting characteristics of Grif. The user can define new document structures and new types of objects, as well as to specify the way in which the system displays these documents and objects.

1. Presentation
Existing document manipulation systems may be classified into various categories. There are batch formatters [Furuta] and interactive systems [Meyrowitz]. Some formatters such as Scribe [Reid] or Mint [Hibbard] consider the logical structure of the documents they manipulate. Some others, like TEX [Knuth] or Troff [Kernighan], are more concerned with layout, even if macros allow some structure to be introduced in the document.

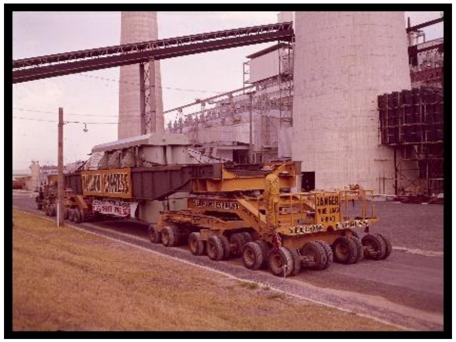
Formatters have also evolved towards more friendly tools, that allow the user to see quickly on the screen the result of his work: TEX, for example, has several 'preview' systems. Janus [Chamberlin] is an original system that has been developed with the same approach. Although they allow the user to see the final form of the document on the screen, these systems cannot be considered as really interactive, as they do not allow the user to interact directly on the final form of the document. Other extensions to formatters have been proposed, by adding a truly interactive editor [André].

Interactive systems may also be classified as layout oriented or structure oriented, but structure oriented interactive systems are mainly used for a particular class of documents: programs. Mentor [Donzeau], the Cornell

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{bmatrix}$$

Bag-of-Visual-Words Model

- BoVW model is borrowed from text analysis



Grif: An Interactive System for Structured Document Manipulation

Vincent Quint*, Irène Vatton**
*INRIA
**CNRS

ABSTRACT
Grif is an interactive system for editing and formatting complex documents. It manipulates structured documents containing objects of various types: tables, mathematical formulae, programs, pictures, graphics, etc... It is a structure directed editor which guides the user in accordance with the structure of the document and of the objects being edited; the image displayed on the screen also being constructed from that structure. Flexibility is one of the most interesting characteristics of Grif. The user can define new document structures and new types of objects, as well as to specify the way in which the system displays these documents and objects.

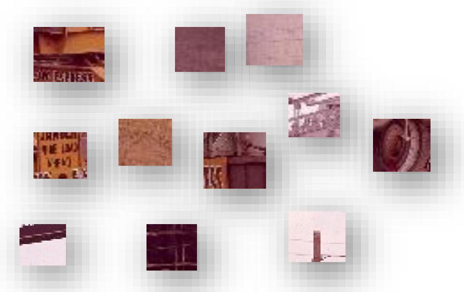
1. Presentation

Existing document manipulation systems may be classified into various categories. There are batch formatters [Furuta] and interactive systems [Meyrowitz]. Some formatters such as Scribe [Reid] or Mint [Hibbard] consider the logical structure of the documents they manipulate. Some others, like TEX [Knuth] or Troff [Kernighan], are more concerned with layout, even if macros allow some structure to be introduced in the document.

Formatters have also evolved towards more friendly tools, that allow the user to see quickly on the screen the result of his work: TEX, for example, has several 'preview' systems. Janus [Chamberlin] is an original system that has been developed with the same approach. Although they allow the user to see the final form of the document on the screen, these systems cannot be considered as really interactive, as they do not allow the user to interact directly on the final form of the document. Other extensions to formatters have been proposed, by adding a truly interactive editor [André].

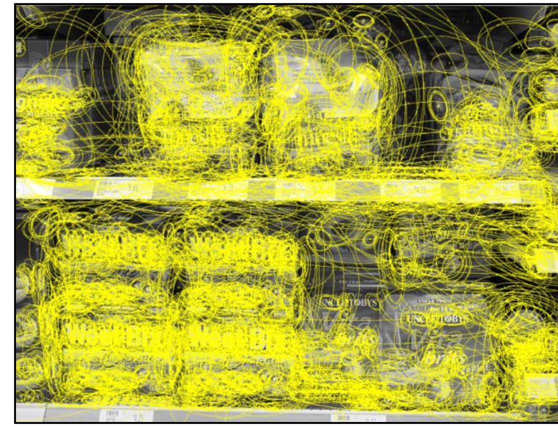
Interactive systems may also be classified as layout oriented or structure oriented, but structure oriented interactive systems are mainly used for a particular class of documents: programs. Mentor [Donzeau], the Cornell

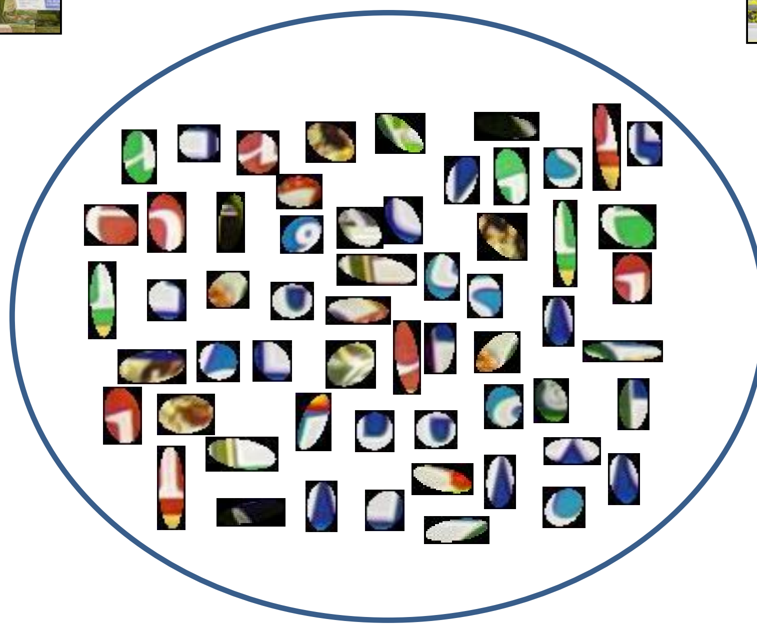
Cambridge Books Online © Cambridge University Press



Bag-of-Visual-Words Model

Interest point detection
or
Dense sampling

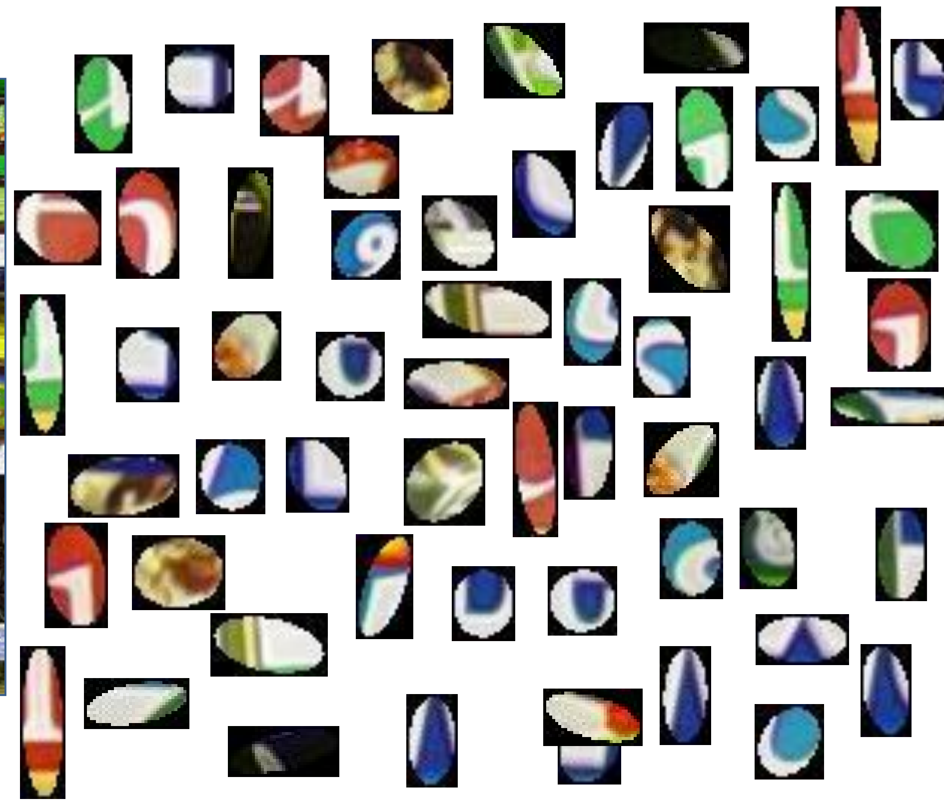




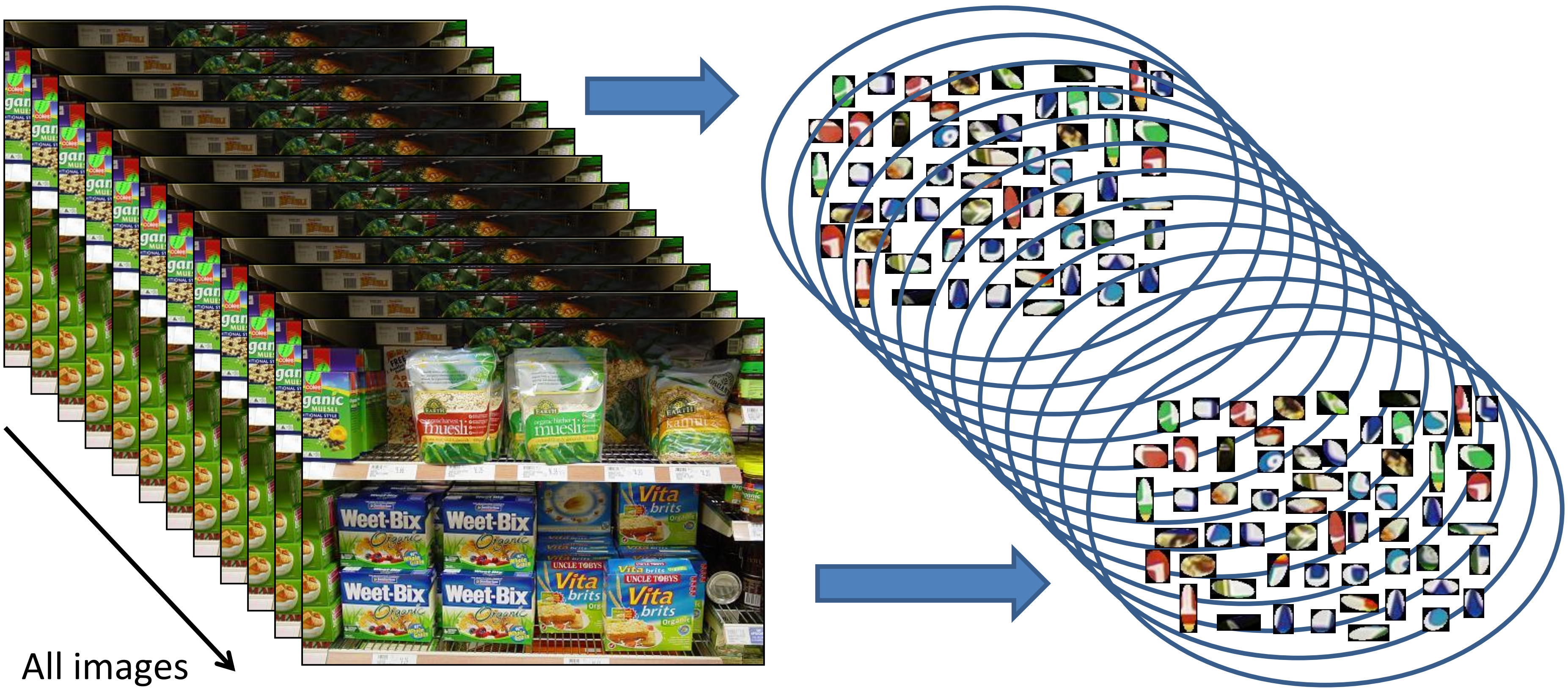
The cropped detected regions

Bag-of-Visual-Words Model

A close-up view



Bag-of-Visual-Words Model

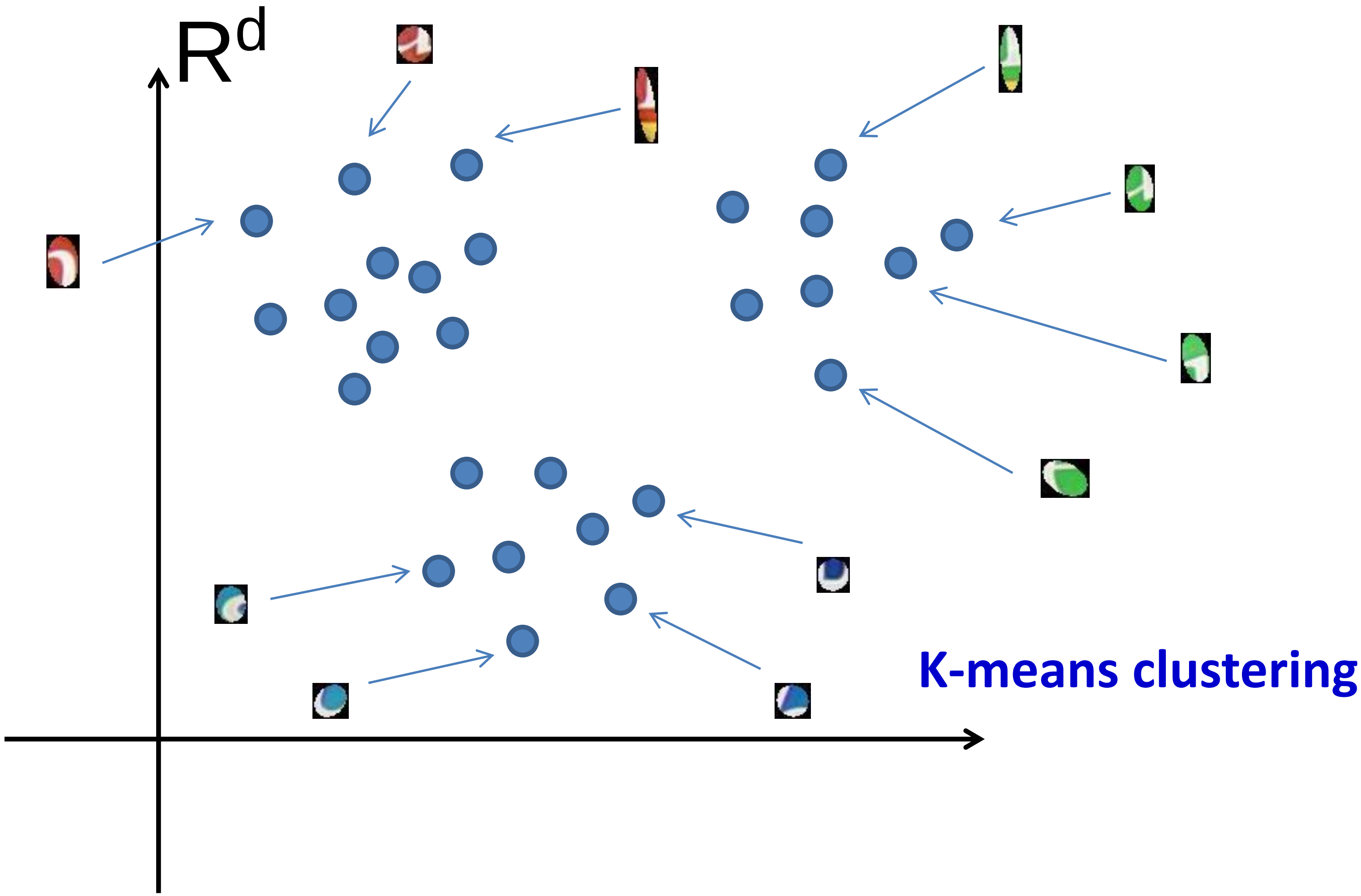


All images
in a dataset

$x \; 2 \; R^d$

Bag-of-Visual-Words Model

Cluster all features to generated "Visual Words"



Bag-of-Visual-Words Model

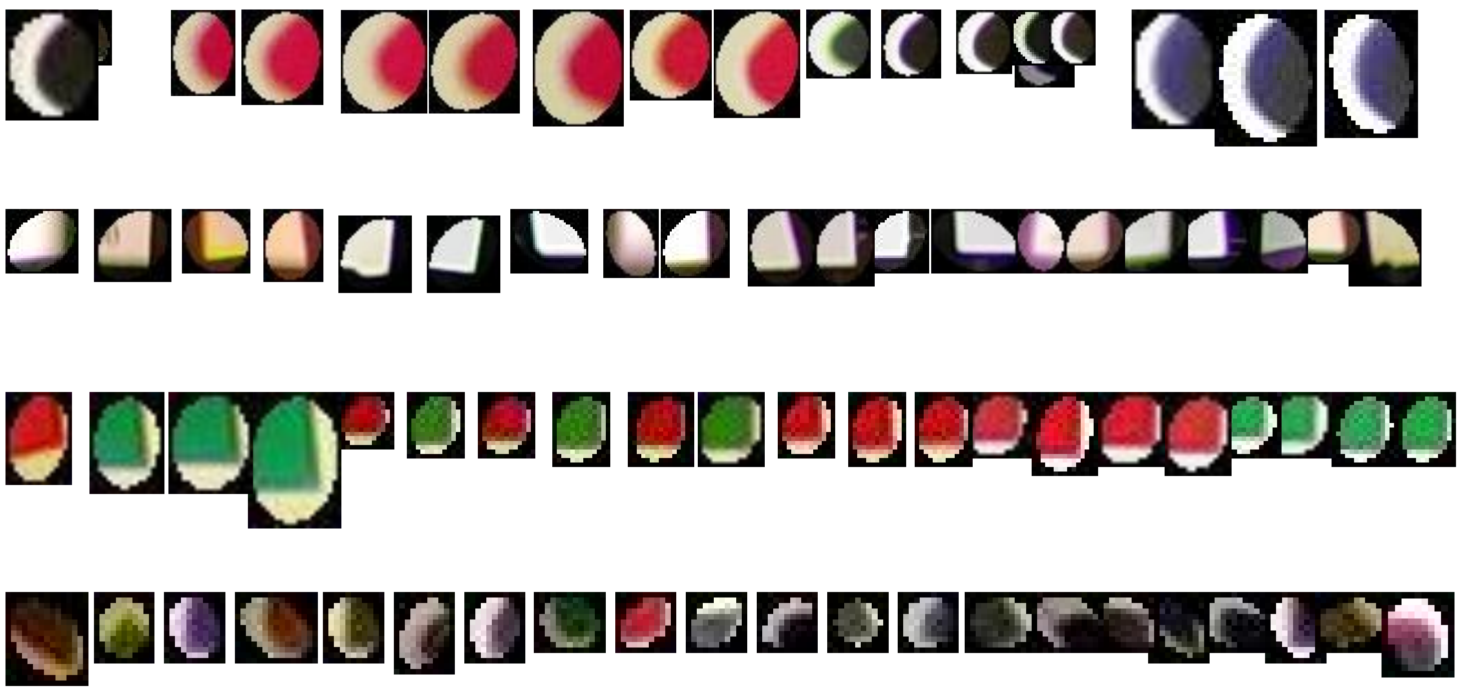
Generated "Visual Words"

Word 1: ...

Word 2: ...

Word 3: ...

Word 4: ...



...

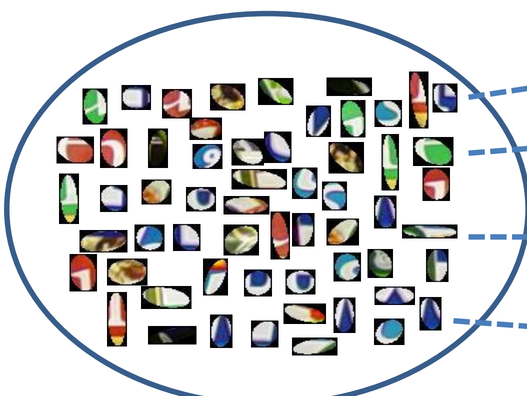
Word *k*: ... ...

Bag-of-Visual-Words Model

From an image to a histogram

The number of occurrence of 1st "word" in this image

$[\, \boxed{n_1,}\ n_2, \ \ldots \ , n_k \,] 2\ R^k$



… … … …

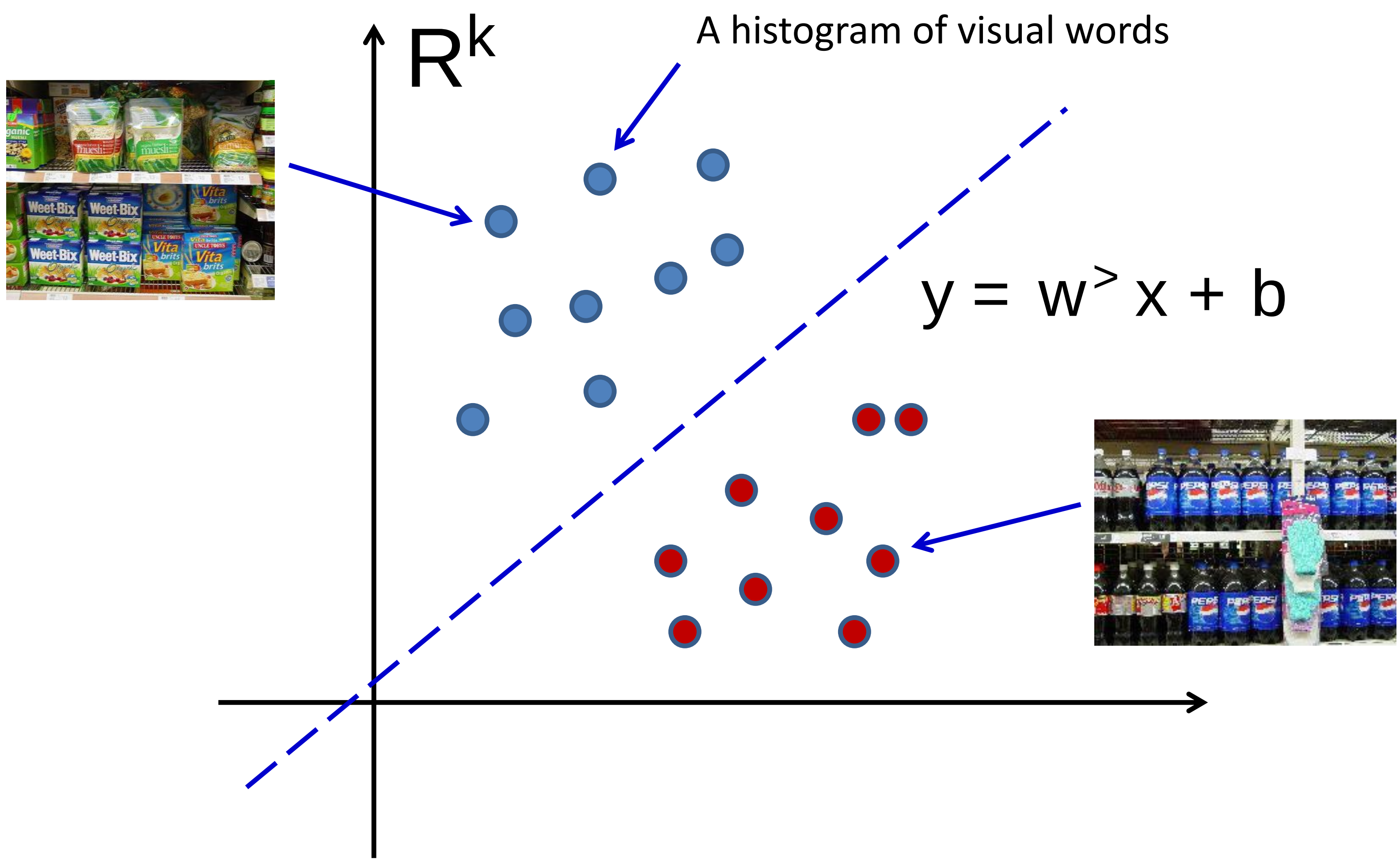
$[\, 0,\ 0,\ 1, \ldots ,0 \,] 2\ R^k$

$[\, 1,\ 0,\ 0, \ldots ,0 \,] 2\ R^k$

$[\, 0,\ 1,\ 0, \ldots ,0 \,] 2\ R^k$

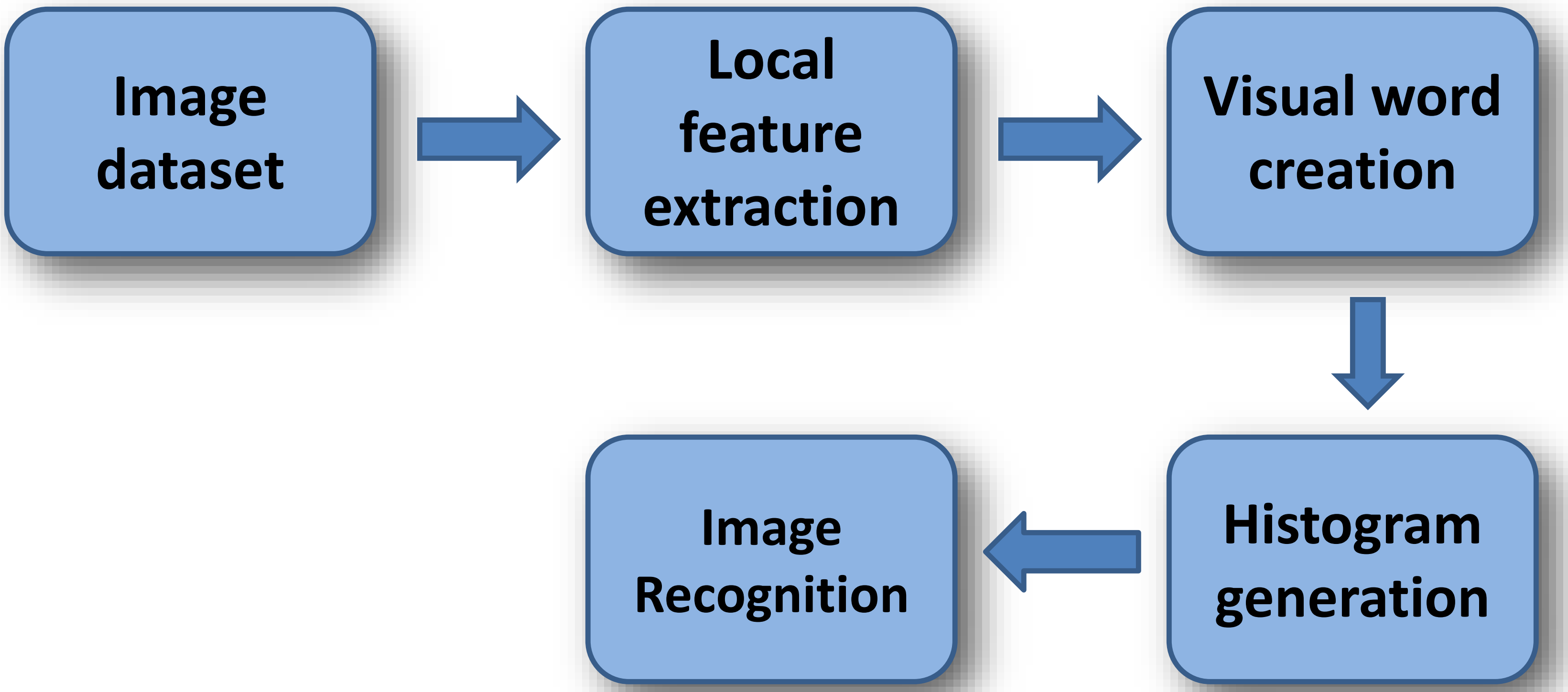
Bag-of-Visual-Words Model

Classifying images

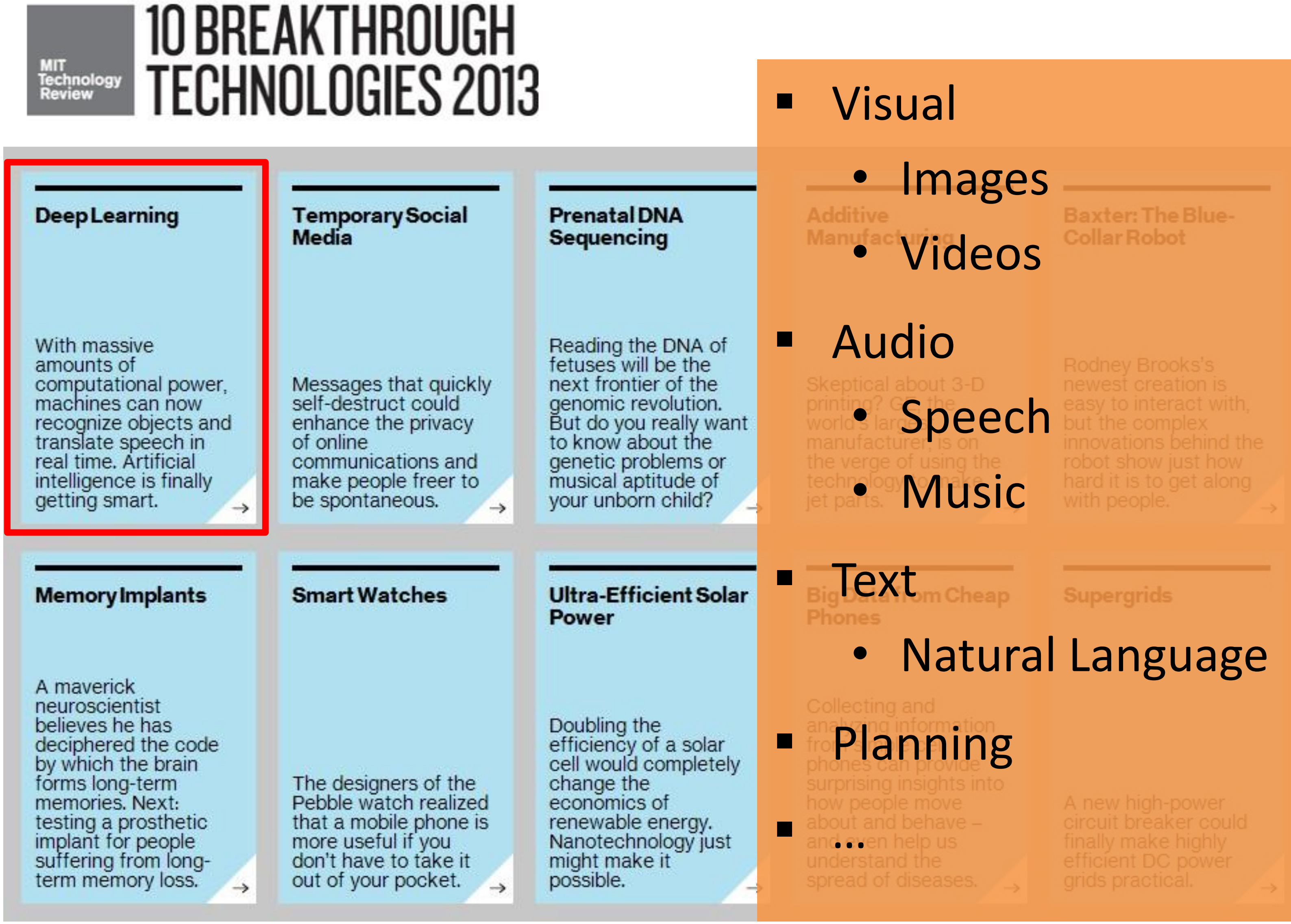


Bag-of-Visual-Words Model

Procedure of the BoVW model based Image Recognition



Deep Learning Model



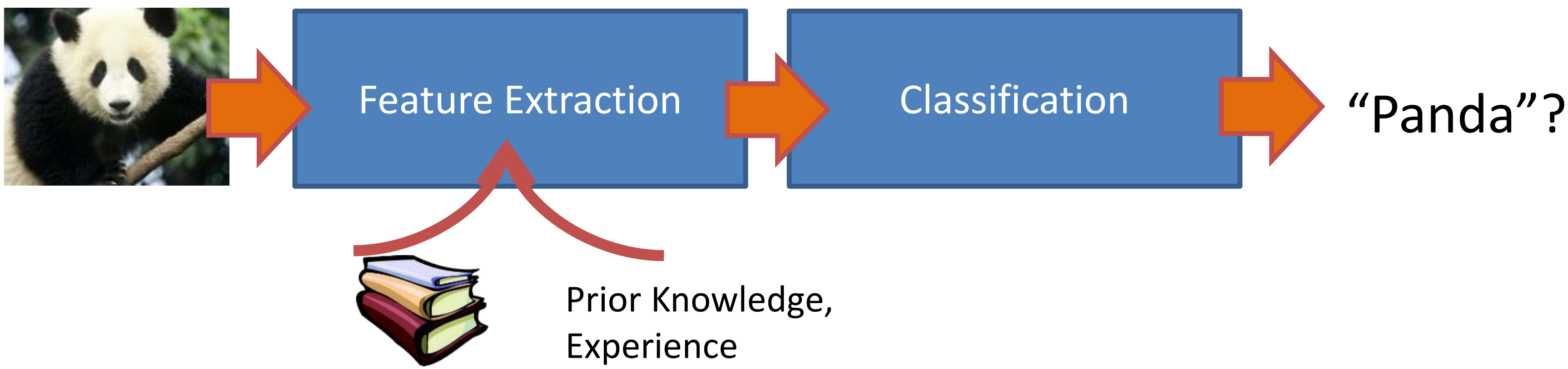
- Visual
  - Images
  - Videos
- Audio
  - Speech
  - Music
- Text
  - Natural Language
- Planning
- …

Deep Learning Model

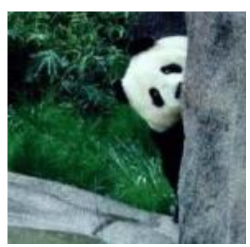
- Image Recognition
  - Faces, objects, poses, scenes, …
- Video content analysis
  - Action, activities, events, summarization, …
- Visual information management
  - Search, retrieval, indexing, browsing, …
- Potential Outcome: AI
  - Computers can see and understand visual information
  - Robotics, self-driving cars, surveillance
  - ….

Deep Learning Model

- We need **Invariant** and **discriminative** features!



Pose



Occlusion

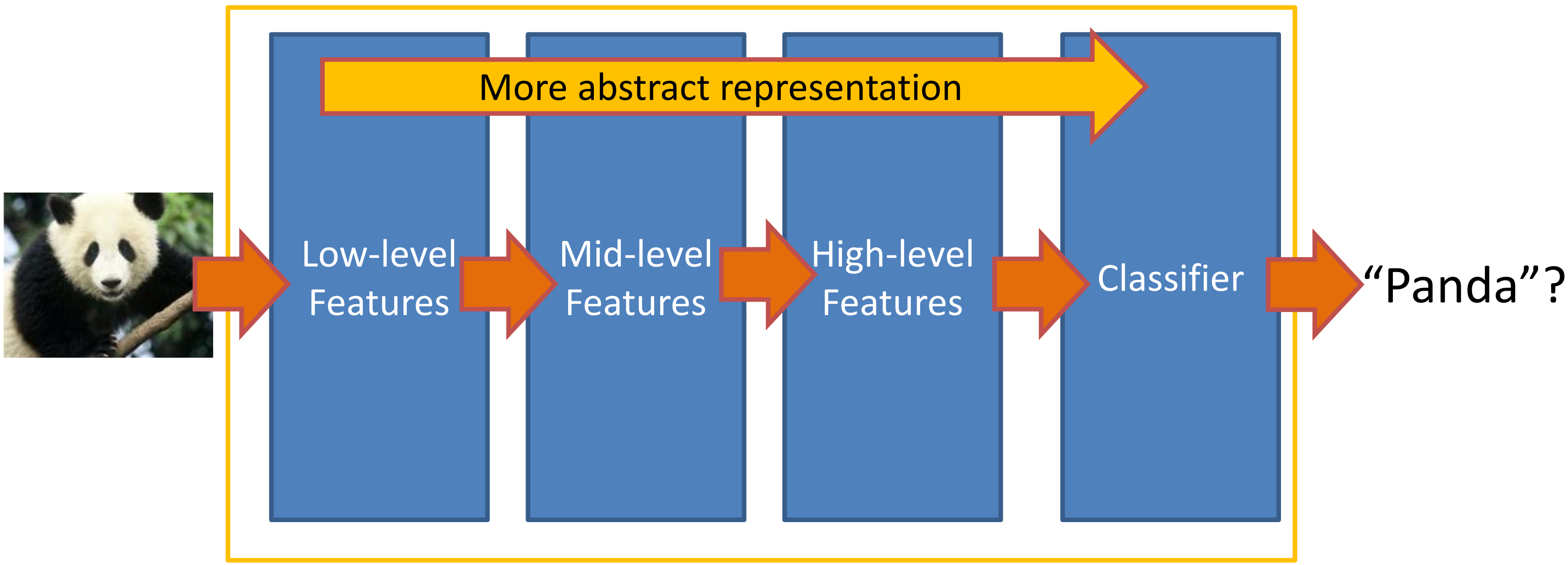

Multiple objects


Inter-class similarity

Image courtesy of M. Ranzato

Deep Learning Model

- **Directly learn** features representations from data.
- **Joint learn** feature representation and classifier.

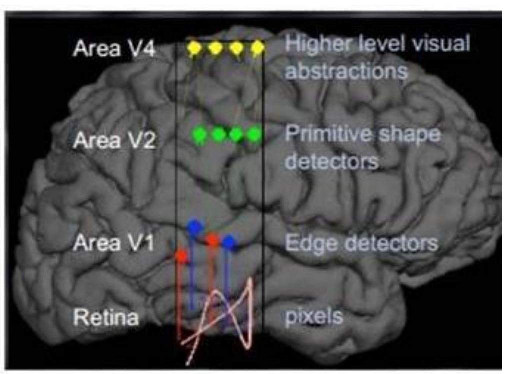


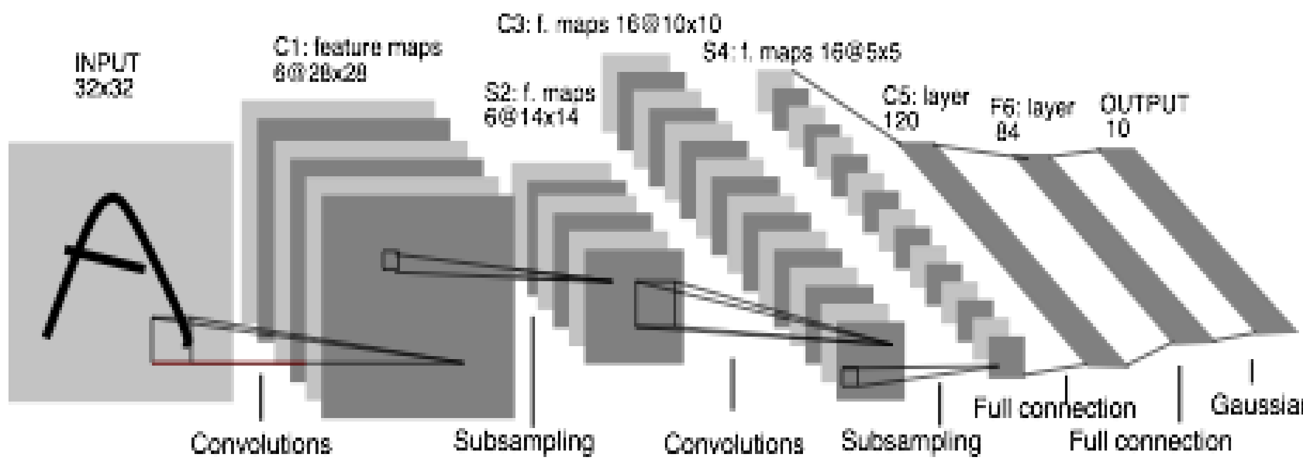
Deep Learning: train layers of features so that classifier works well.

Image courtesy of M. Ranzato

Deep Learning Model

- Inspired by the way that human brain processes information
- Many **layers** of **non-linear** processing stages

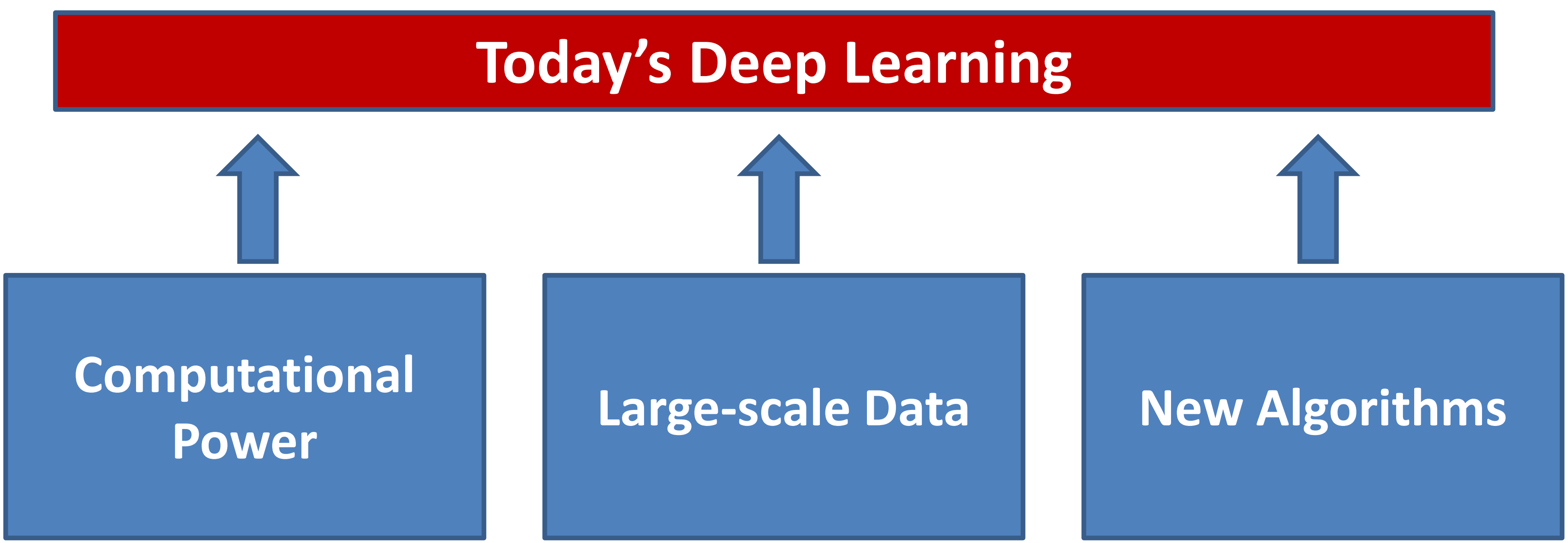




Deep Learning Model

Have we been here before?

- Yes.
  - Basic ideas common to past **neural networks** research
  - Standard machine learning strategies still relevant.
- No.



Deep Learning Model

Convolutional Neural Networks (CNNs)

- A special **multi-stage architecture** inspired by visual system
  - Higher stages compute more global, more invariant features

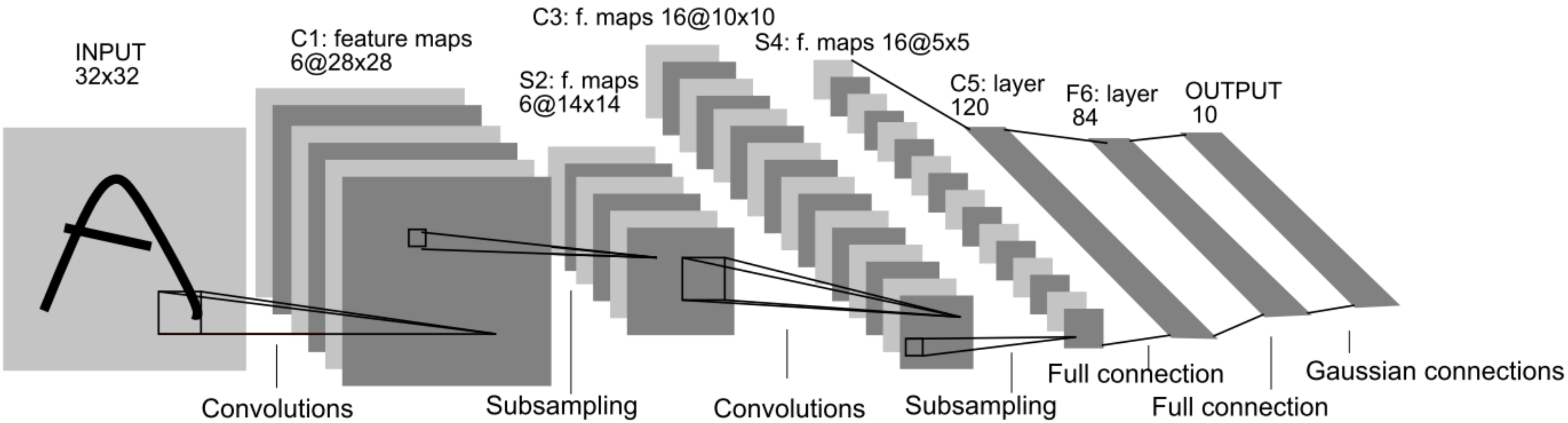
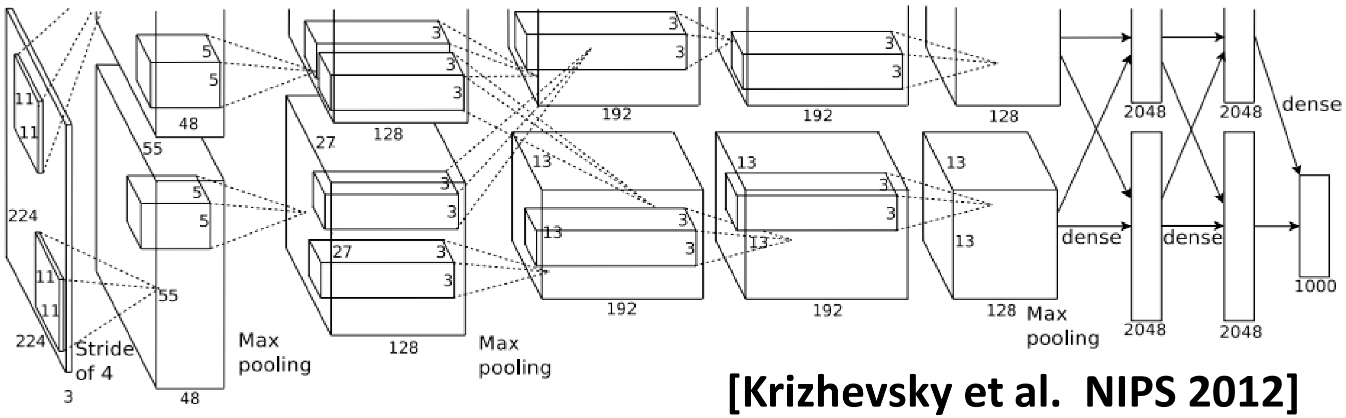


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Deep Learning Model

CNNs: ImageNet Breakthrough



[Krizhevsky et al. NIPS 2012]

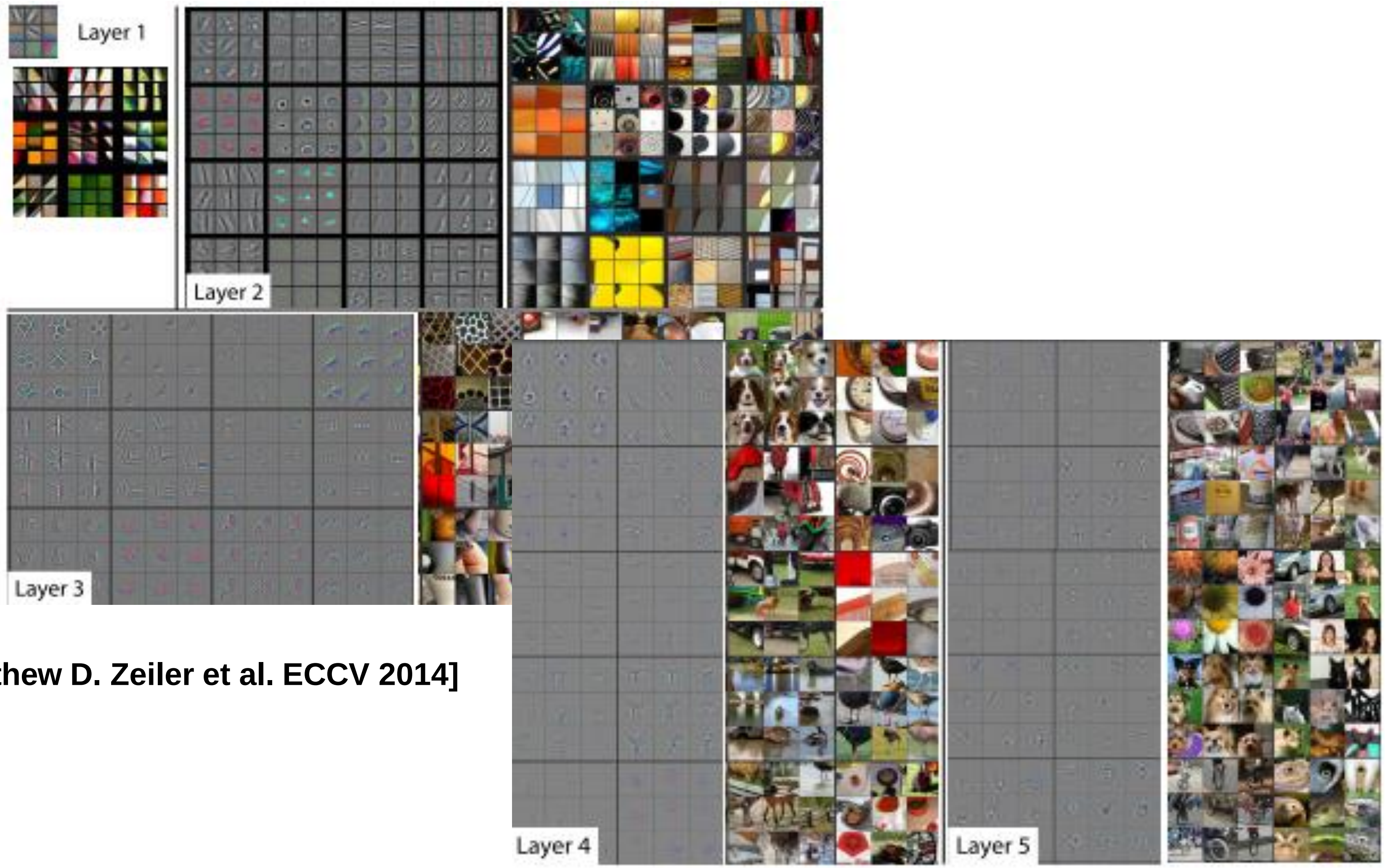
- Krizhevsky et al. win 2012 ImageNet classification with a **much bigger ConvNet**
  - **deeper**: 7 stages vs 3 before
  - **larger**: 60 million parameters vs 1 million before
  - **16.4%** error (top-5) vs Next best 26.2% error

- This was made possible by:
  - **fast hardware**: GPU-optimized code
  - **big dataset**: 1.2 million images vs thousands before
  - **better regularization**: dropout et al.

Image courtesy of Deng et al.

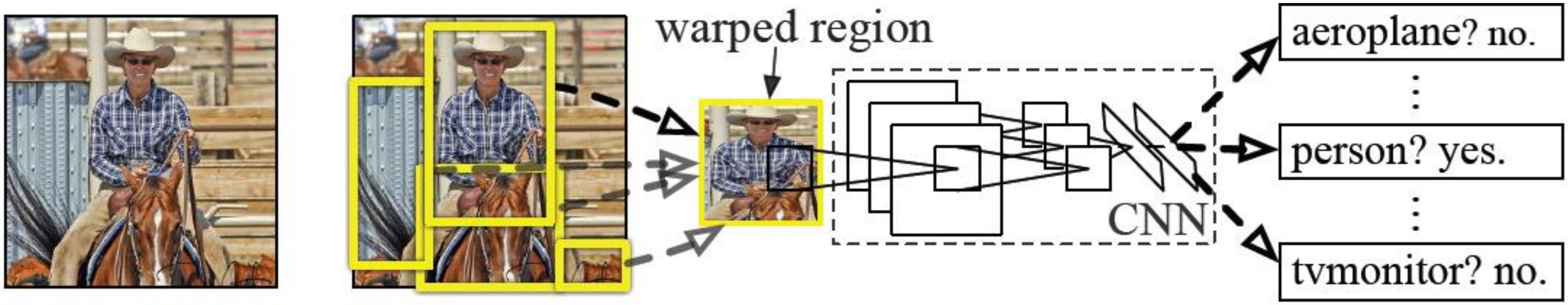
Deep Learning Model

Learned Features of CNNs

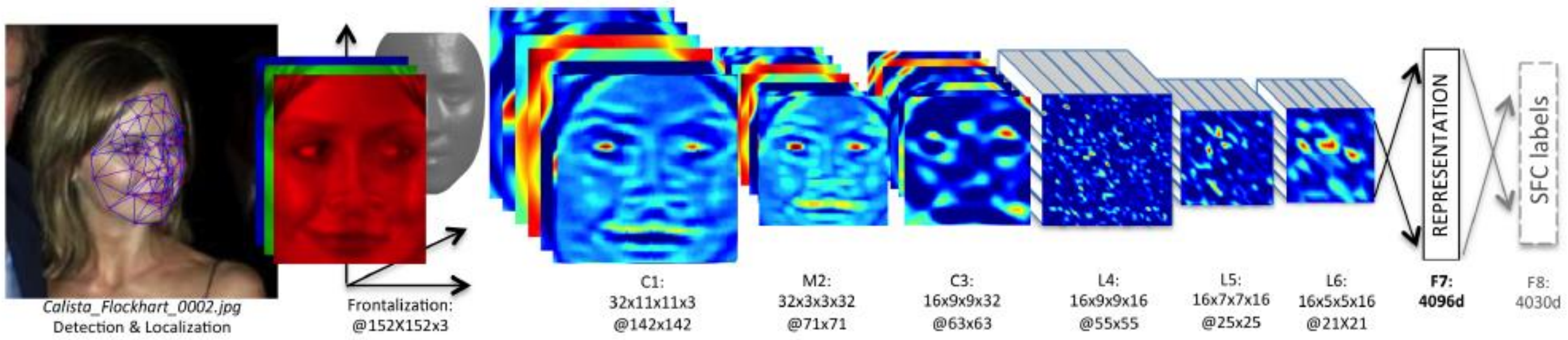


[Matthew D. Zeiler et al. ECCV 2014]

Deep Learning Model

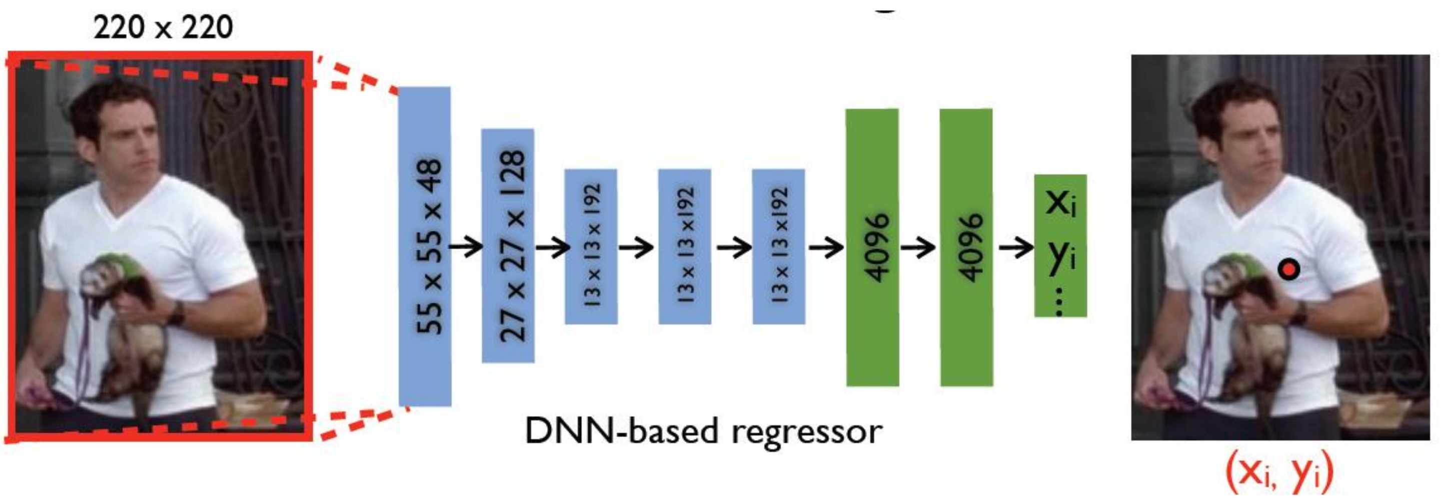


Object detection (Source: Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR 2014)



Face Recognition (Source: DeepFace: Closing the Gap to Human-Level Performance in Face Verification, CVPR 2014)

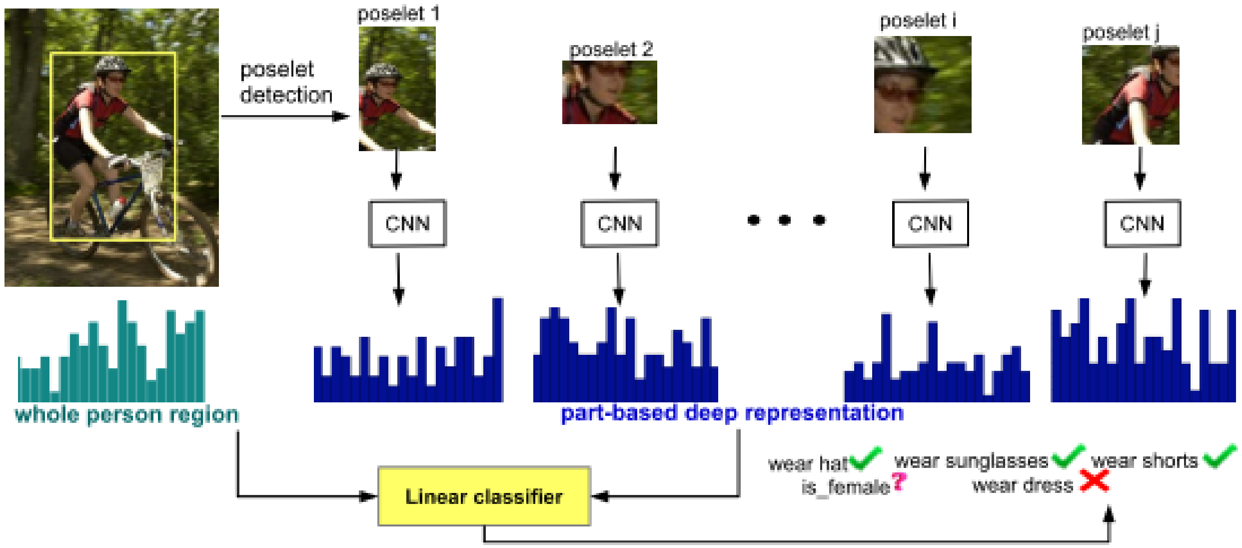
Deep Learning Model



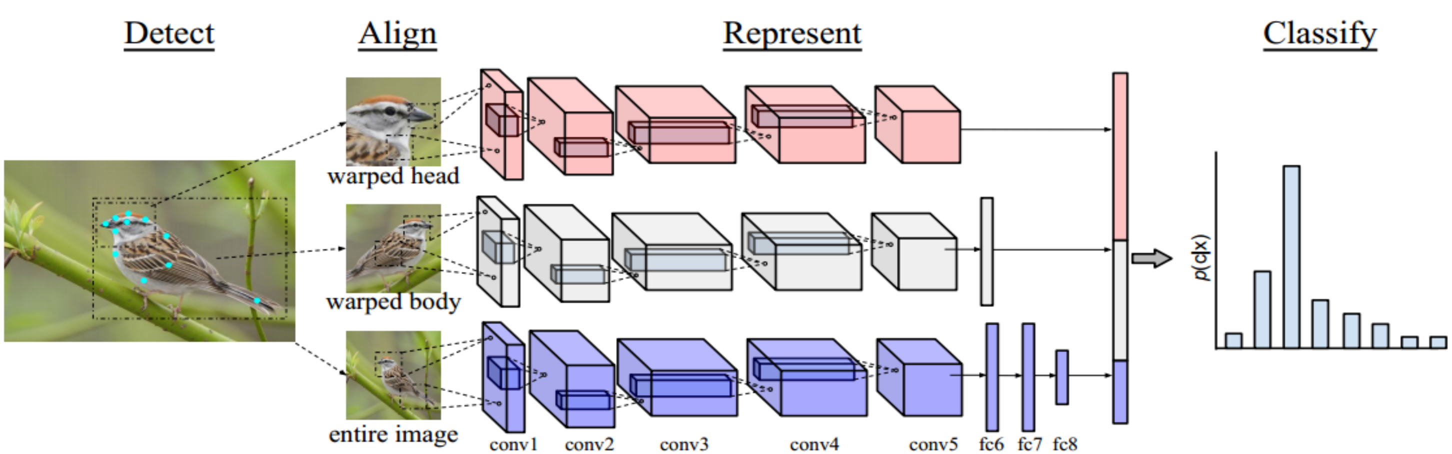
Pose estimation (DeepPose: Human Pose Estimation via Deep Neural Networks, CVPR2014)

Image Segmentation (Source: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE TPAMI 2016)

Deep Learning Model

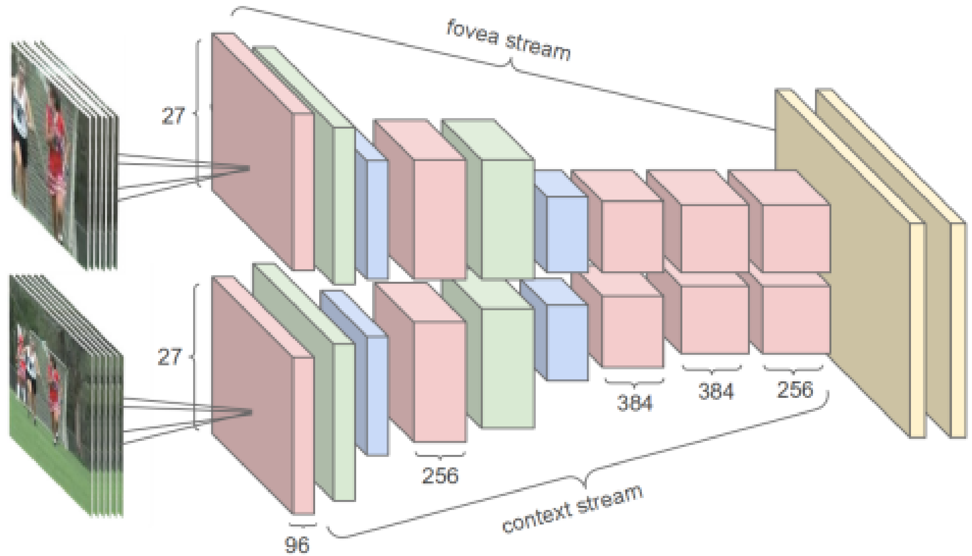


Human attribute classification [Ning Zhang et al. CVPR 2014]

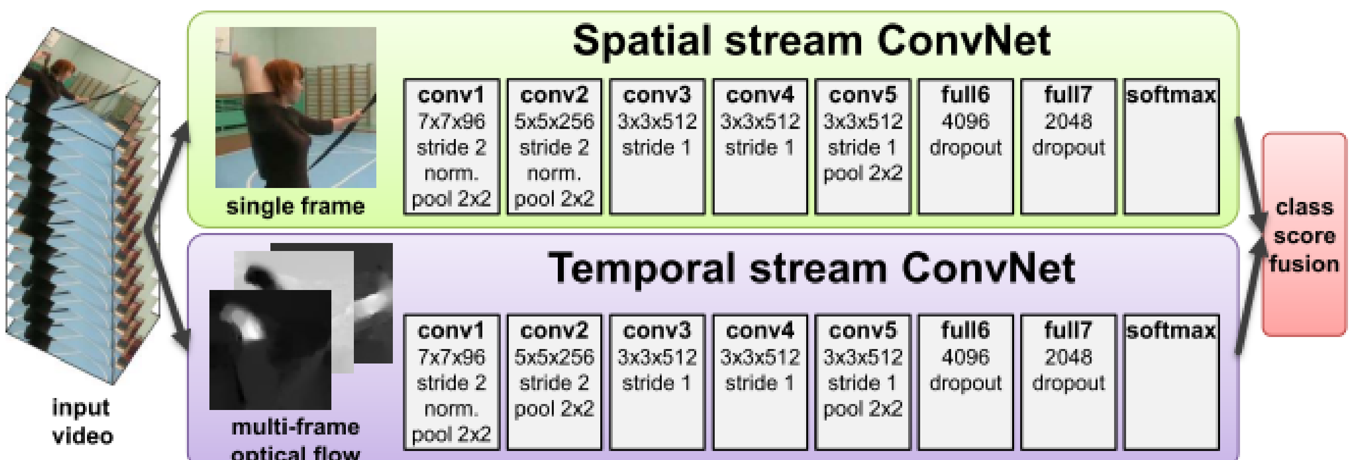


Fine-grained image recognition [Branson et al. arXiv 2014]

Deep Learning Model



Large-scale Video Classification [Karpathy et al. CVPR 2014]



Action Recognition [Simonyan et al. arXiv 2014]

Deep Learning Model

Image Representation: From SIFT to CNNs

- Three main approaches
  - **Directly use pre-trained CNNs** models
    - to extract image feature representations
  - **Fine-tune pre-trained CNNs** models
    - with the images from recognition tasks
  - **BoVW model** based on CNN features
    - "**Deep SIFT**"

Deep Learning Model

- Directly use pre-trained CNNs
  - **Which** layer to use?
  - How to **pool** the features in a convolutional layer?

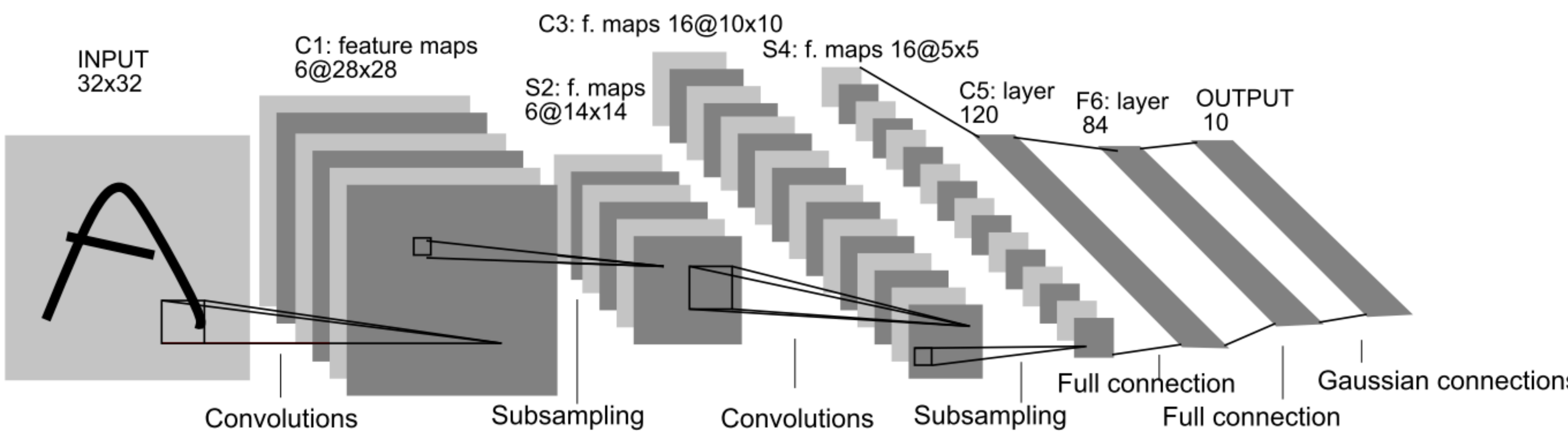


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Deep Learning Model

- Directly use pre-trained CNNs
  - **Which** layer to use?

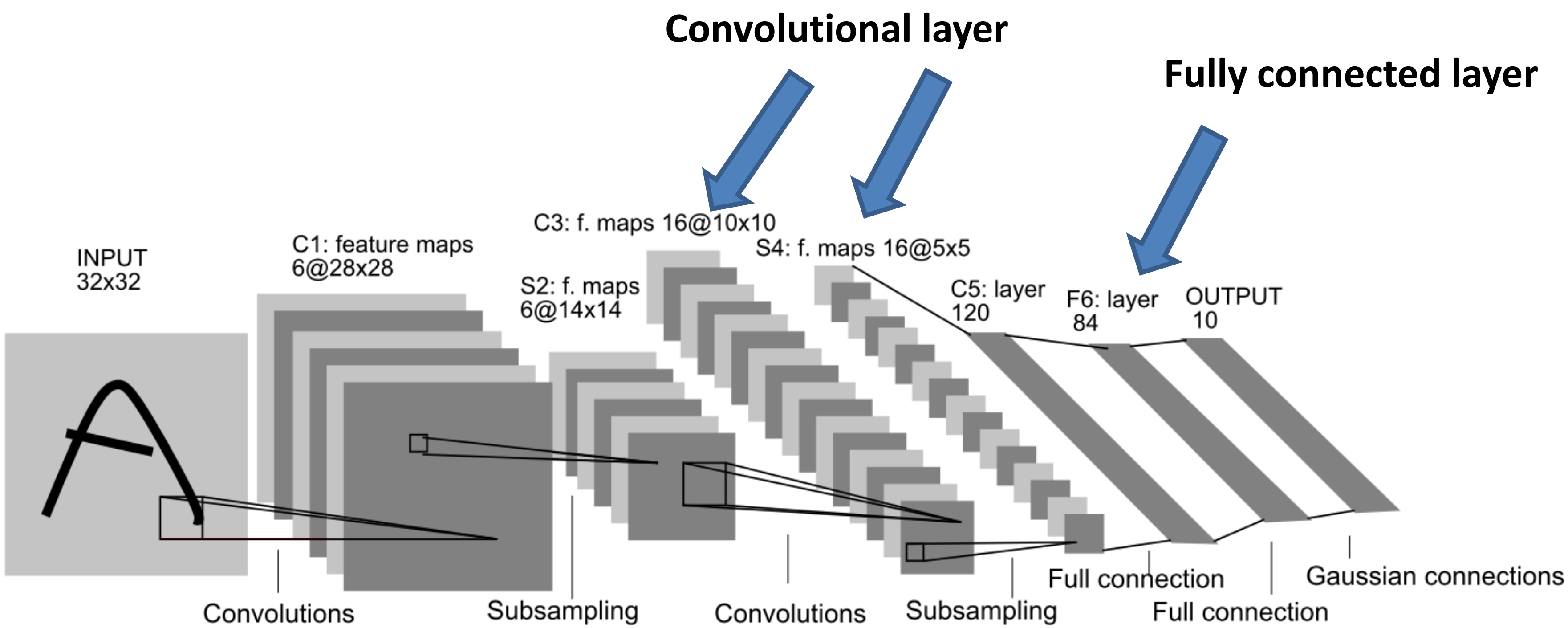


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Deep Learning Model

- Directly use pre-trained CNNs
  - How to **pool** the features in a convolutional layer?

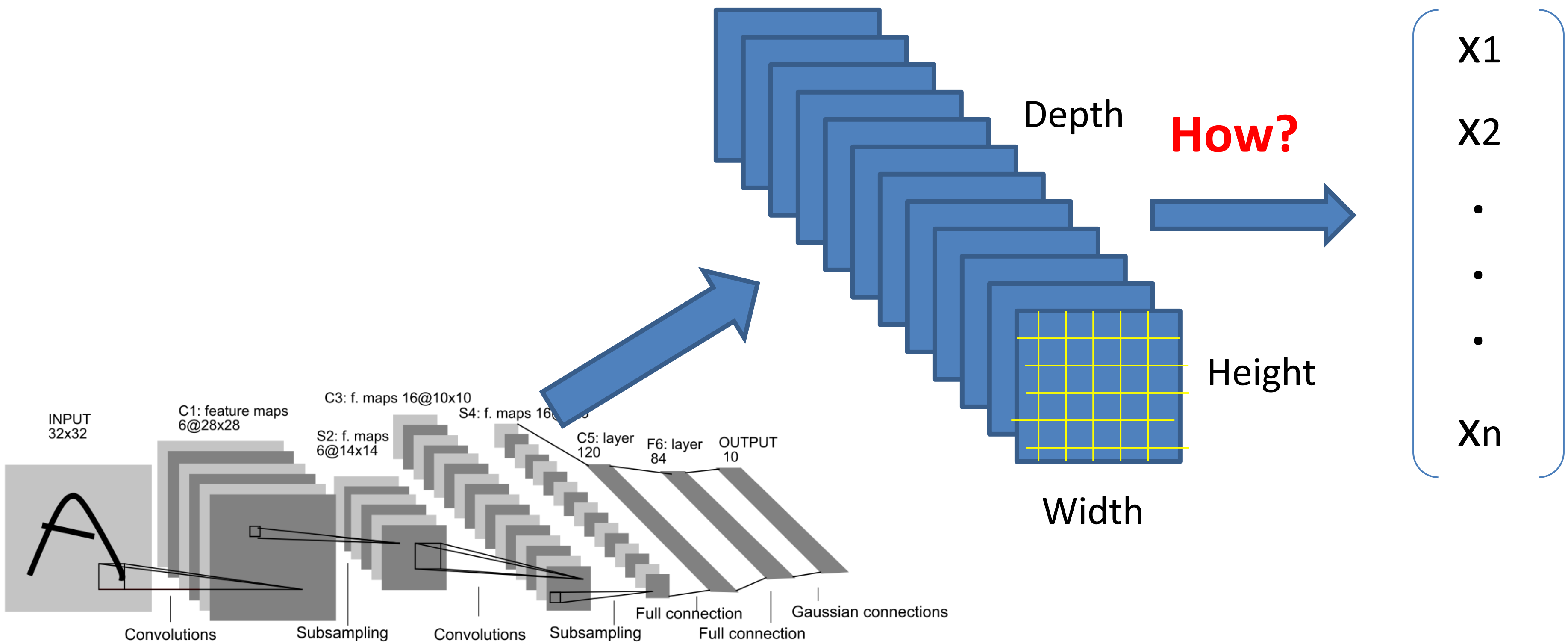
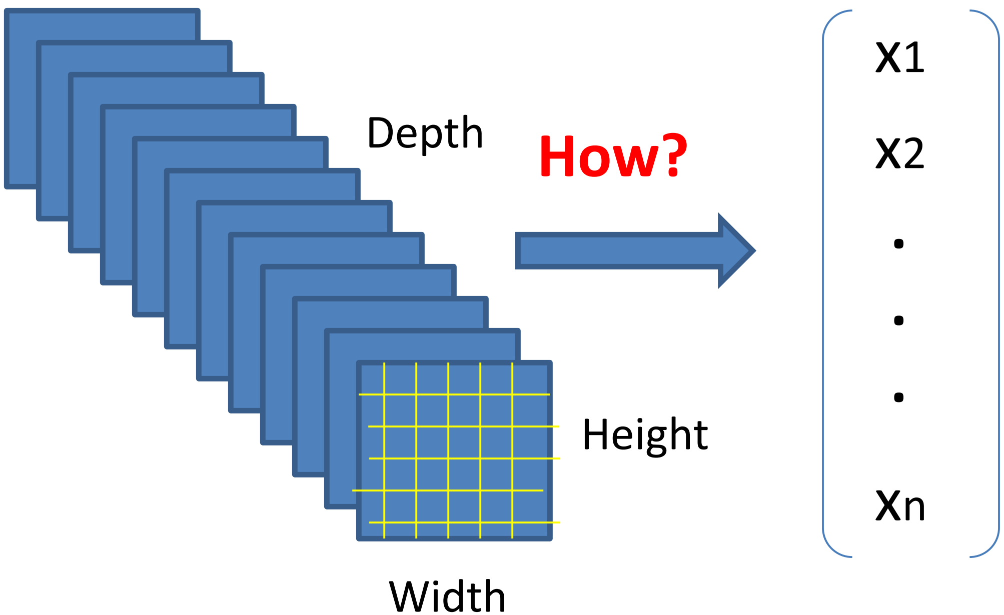


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Deep Learning Model

- Directly use pre-trained CNNs
  - How to **pool** the features in a convolutional layer?



- Sum-pooling
- Max-pooling
- Grid-based max-pooling
- Region-based pooling
- Mixed sum & max pooling

Deep Learning Model

- Fine-tune pre-trained CNNs
  - To incorporate extra information from the images of a new recognition task
  - Make the pre-trained CNNs adapt to this new task

Pre-trained CNNs on

IMAGENET

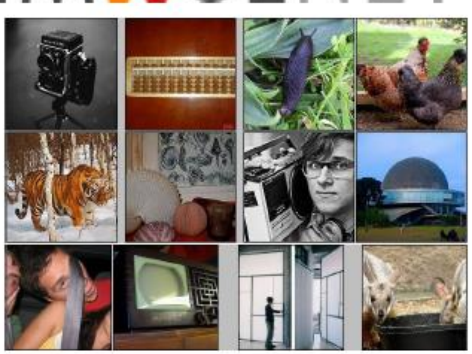


Image courtesy of Deng et al.

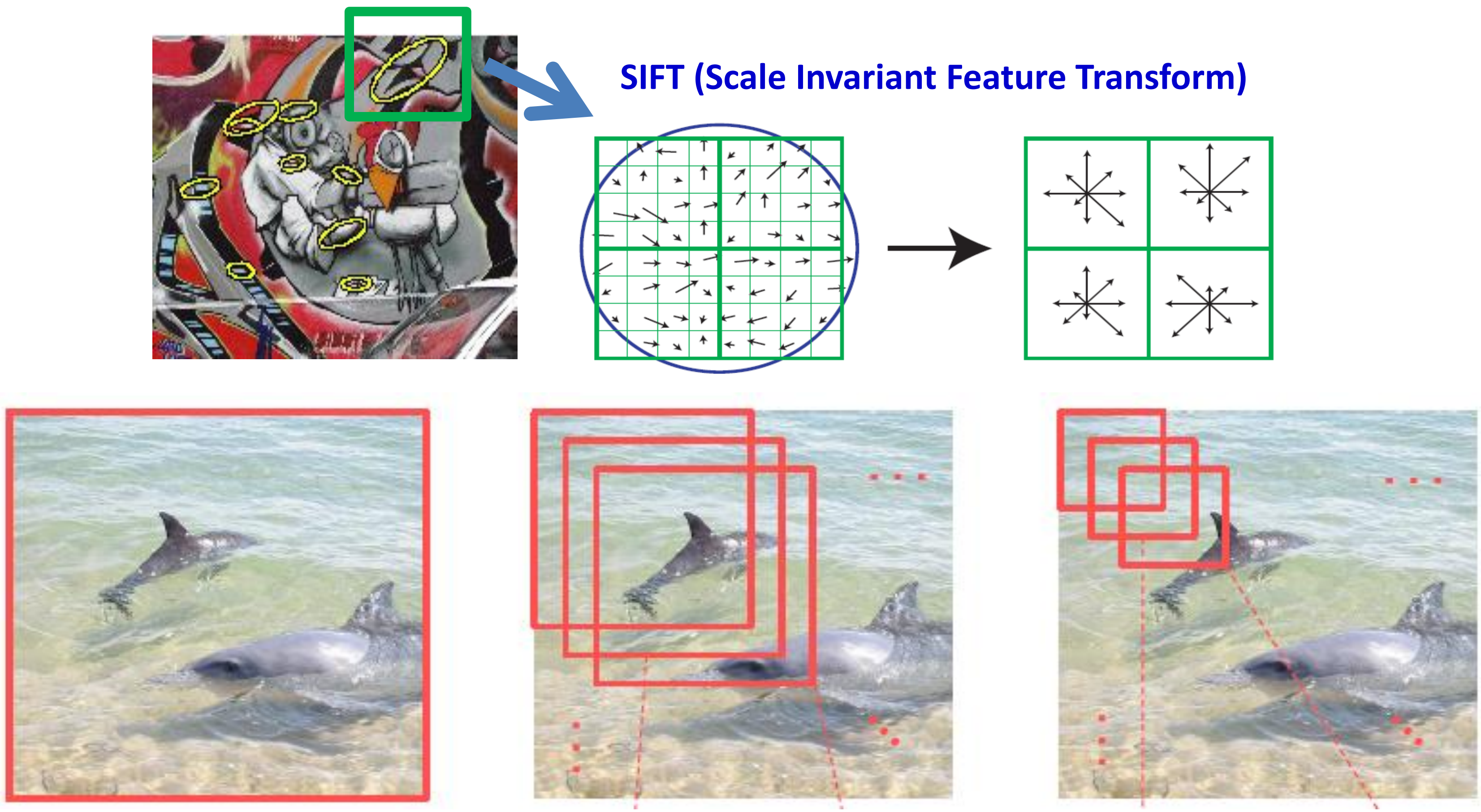
Fine-tune

New recognition task

http://people.csail.mit.edu/bzhou/

Deep Learning Model

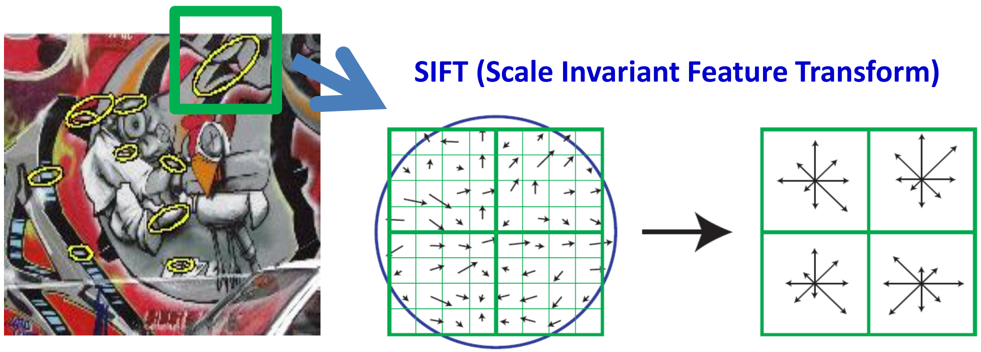
BoVW model based on “Deep SIFT”

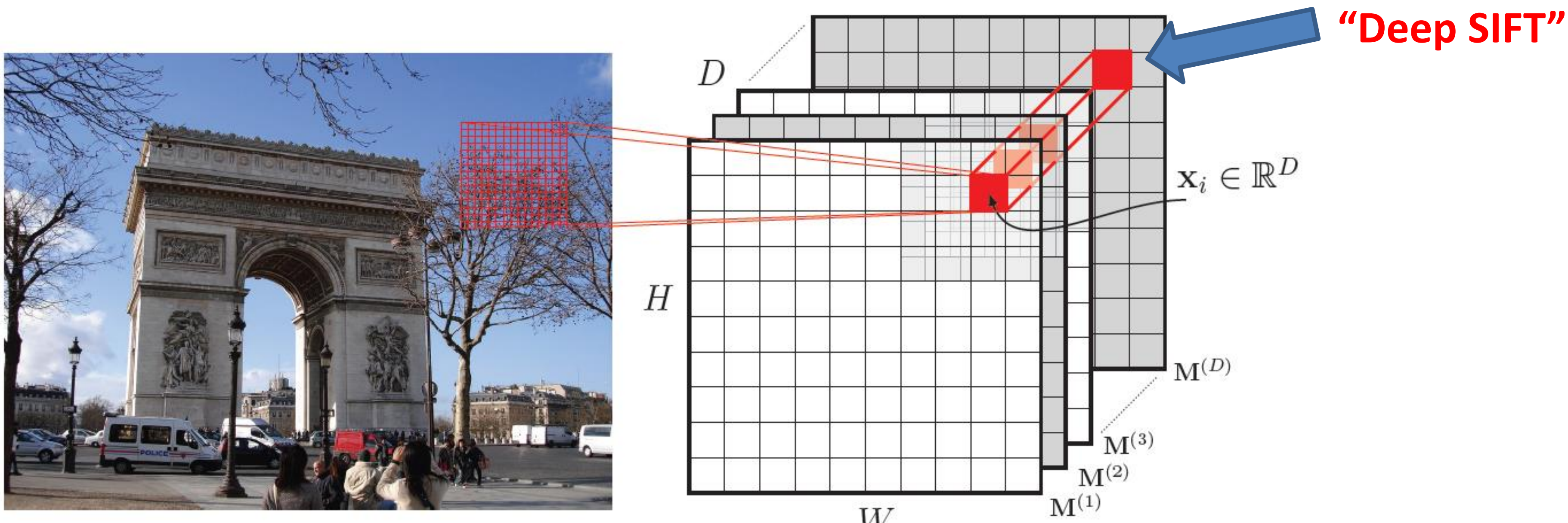


Source: Multi-scale Orderless Pooling of Deep Convolutional Activation Features, ECCV2014

Deep Learning Model

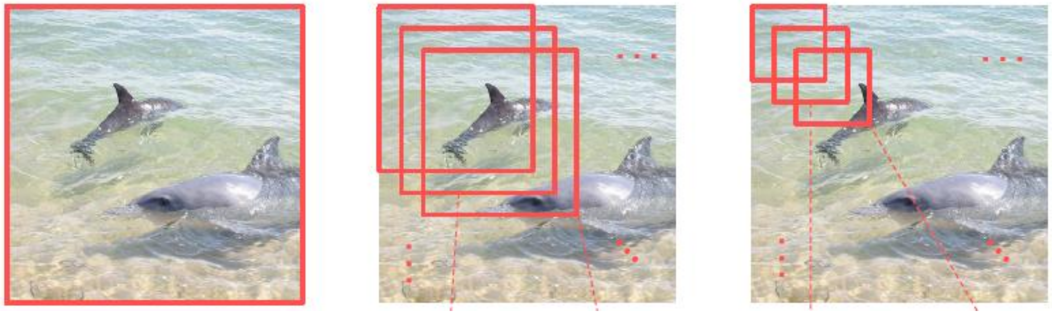
BoVW model based on “Deep SIFT”

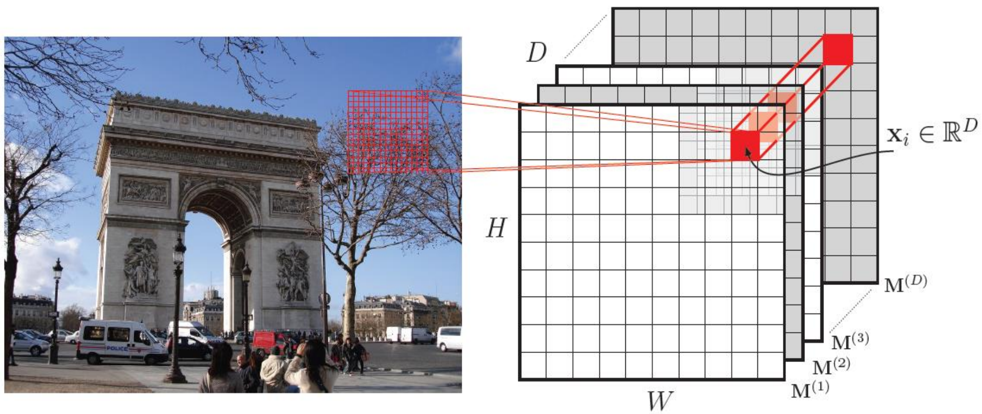




Source: Cao et. al, Where to Focus: Query Adaptive Matching for Instance Retrieval Using Convolutional Feature Maps

Deep Learning Model





Or

Visual word creation

Histogram generation

Image recognition

Deep Learning Package for R

- The support for performing deep learning with R language is increasing

Package	Multiple CPU	[Multiple] GPU	Cluster	Platforms
MXNetR	X	X		Linux\MacOS\ Windows
darch		X		Linux\MaxOS
H2o	X		X	Linux\MacOS\ Windows
deepnet				No information

Source: http://www.rblog.uni-freiburg.de/2017/02/07/deep-learning-in-r/

Deep Learning Package for R

- MXNetR package (link1, link2, Video)
  - R interface of the MXNet library (written in C++)
  - Contains feed-forward neural networks and convolutional neural networks
  - The CPU version can be easily installed in R

```
model <- mx.mlp(train.x, train.y, hidden_node=c(128,64),
out_node=2, activation="relu",
out_activation="softmax",num.round=100,
array.batch.size=15, learning.rate=0.07, momentum=0.9,
device=mx.cpu())

preds = predict(model, testset)
```

Source: http://www.rblog.uni-freiburg.de/2017/02/07/deep-learning-in-r/

Deep Learning Package for R

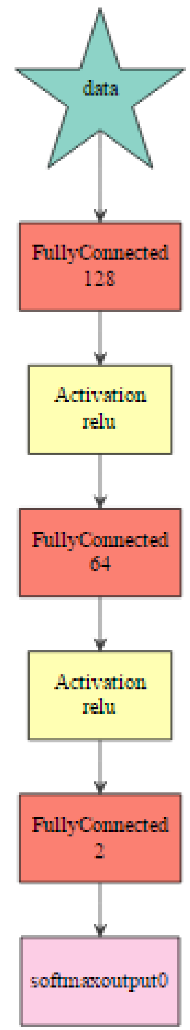
- MXNetR package can visualise the networks

`graph.viz(model$symbol$as.json())`

- A helpful blog on deep learning with R

https://www.r-bloggers.com/image-recognition-tutorial-in-r-using-deep-convolutional-neural-networks-mxnet-package/

Computation graph



Summary

- **Image analysis** is an important and broad area
- **Feature representation** is key for image analysis
- **Deep Learning** techniques are now widely used
- **Comparison** of deep learning software
- Issues to be resolved for image recognition
  - How to **transfer** the benefit of Deep Learning?
  - How to deal with **unsupervised** learning case?
  - How to better handle the **semantic gap**?
  - …

Q&A

Images Courtesy of Google Image