ISTITUTO ITALIANO DI TECNOLOGIA

POLYTECHNIQUE MONTRÉAL

Action Recognition with Image Based CNN Features

Hossein Mousavi

December 2017

Outline

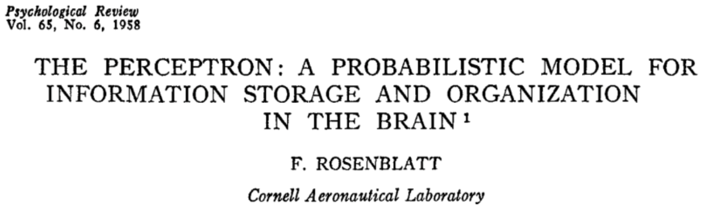
- **Neural Networks A Brief History**

- **Action Recognition with Image Based CNN Features**
  - **Introduction**
  - **Method**
  - **Experiment Results**

- **CNN-aware Binary Map For General Image Segmentation**

Neural Networks A Brief History

- The 1950s and 1960s: The First Golden Age of Neural Networks
  - Frank Rosenblatt (1958) created the perceptron

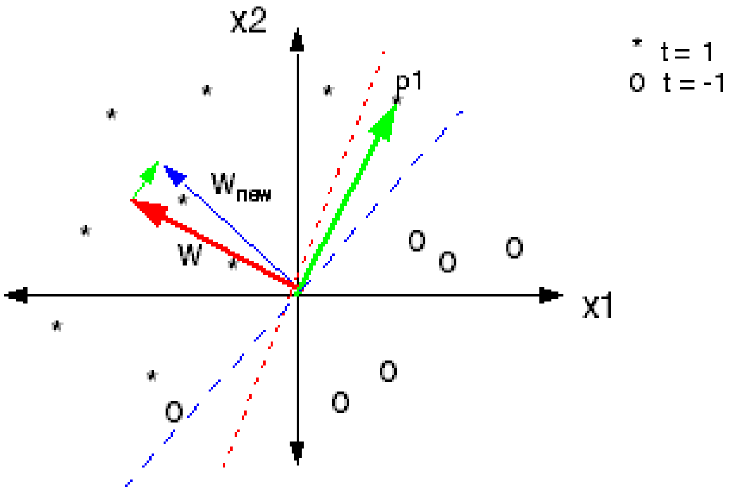


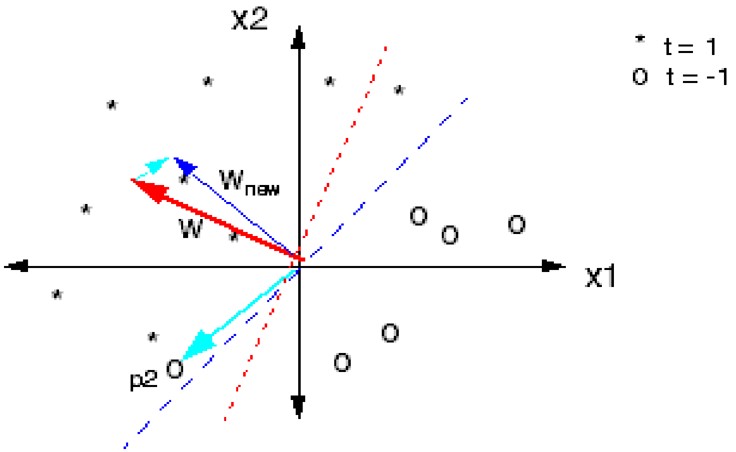
Psychological Review
Vol. 65, No. 6, 1958

THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN [1]

F. ROSENBLATT

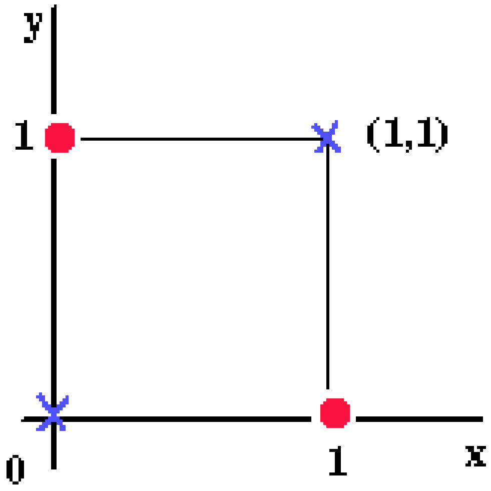
Cornell Aeronautical Laboratory



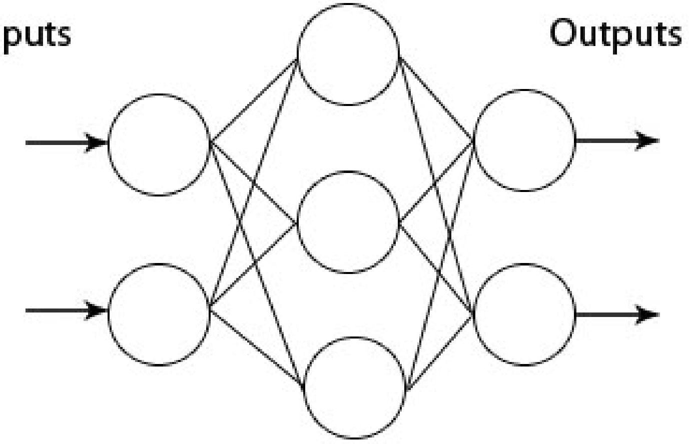


Neural Networks A Brief History

- The 1970s: The Quiet Years
  - Perceptron could not solve simple XOR problem
  - Overestimating the success of AI in research papers

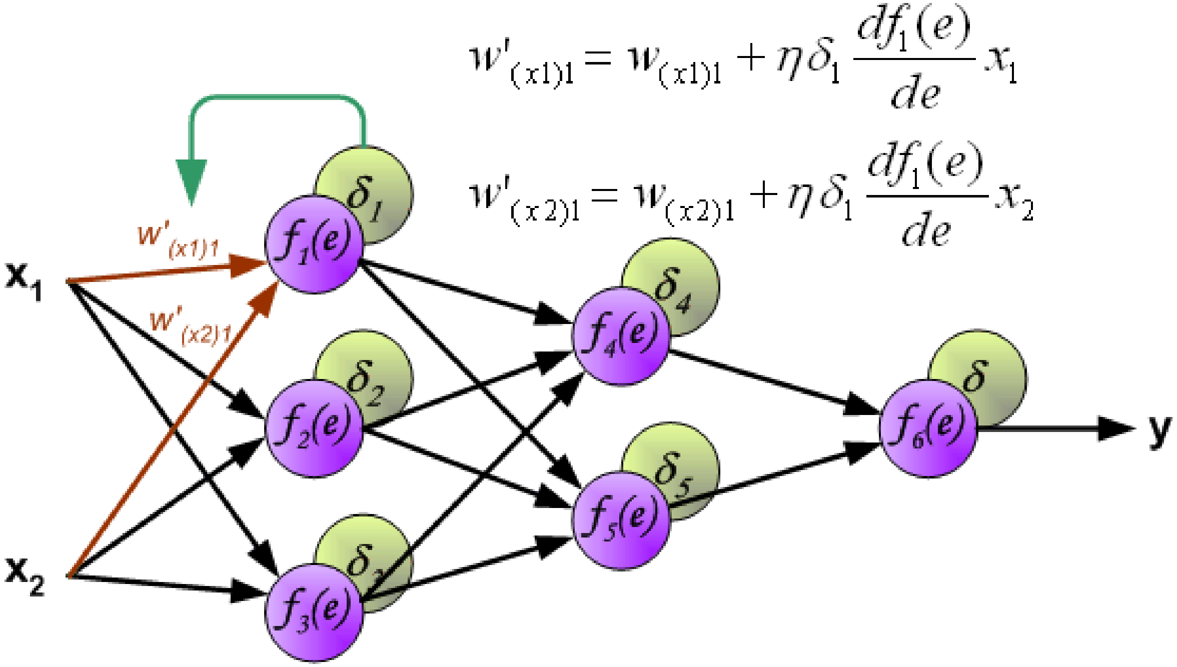


Multi-Layer Perceptron : How to train?!!!



Neural Networks A Brief History

- After 1975 up to 1990: Renewed Enthusiasm
  - The Backpropagation algorithm was created by Paul Werbos (1975)

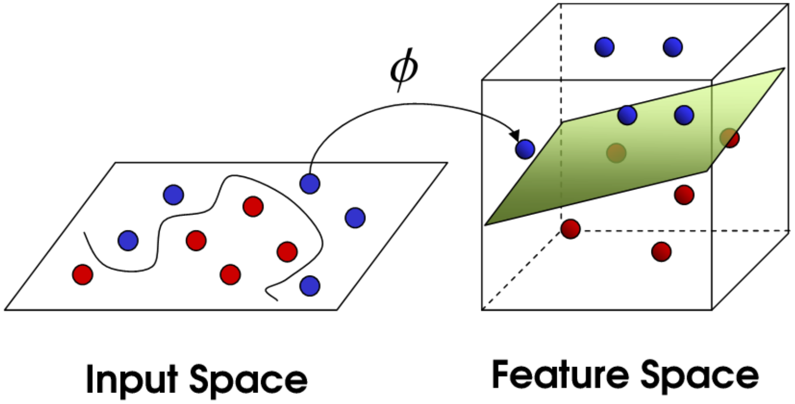


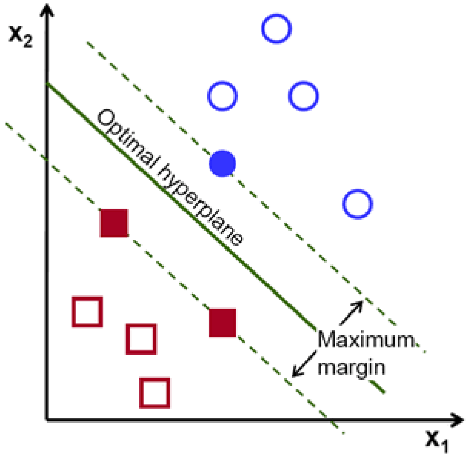
$$w'_{(x1)1} = w_{(x1)1} + \eta \delta_1 \frac{df_1(e)}{de} x_1$$

$$w'_{(x2)1} = w_{(x2)1} + \eta \delta_1 \frac{df_1(e)}{de} x_2$$

Neural Networks A Brief History

- 1990 -2012 : Long Quiet Years !!!
  - Learning large network was computationally expensive
  - Support Vector Machine took over
    - Convex Optimization
    - Nonlinear Models by Kernel Tricks

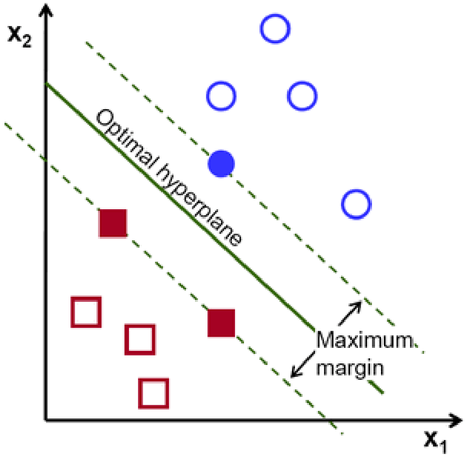




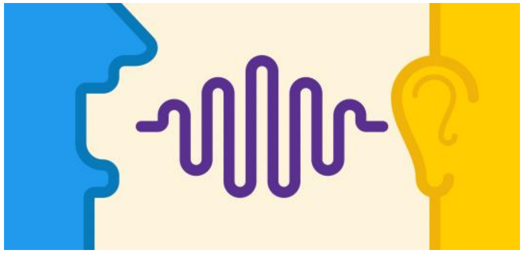
Feature Engineering

- Converting everything to a vector representation

X = [x1,x2,…,xD]



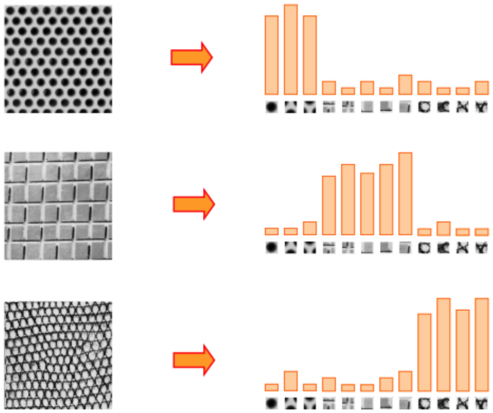
Feature Engineering

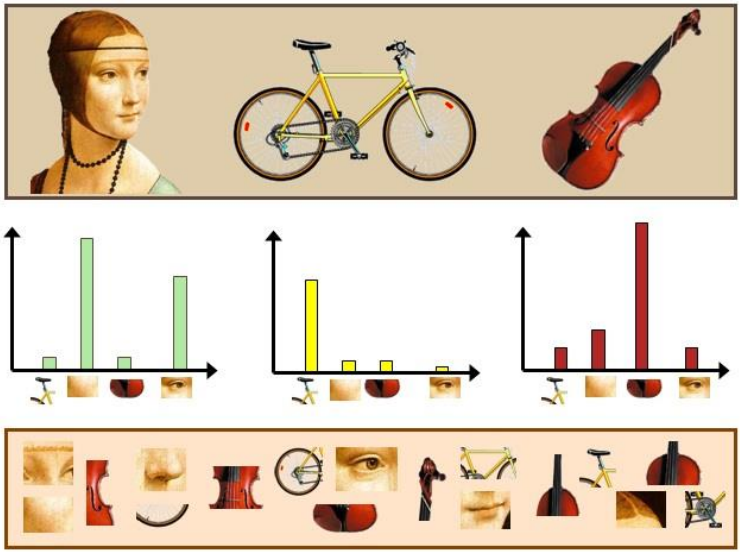


Machine Learning

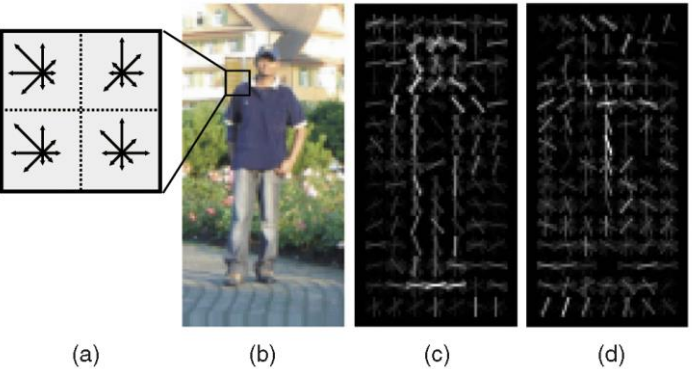
Feature Engineering

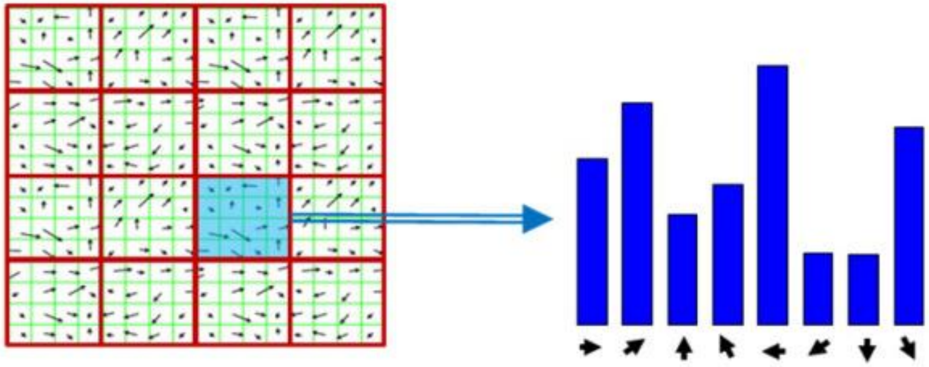
- Bag of Words





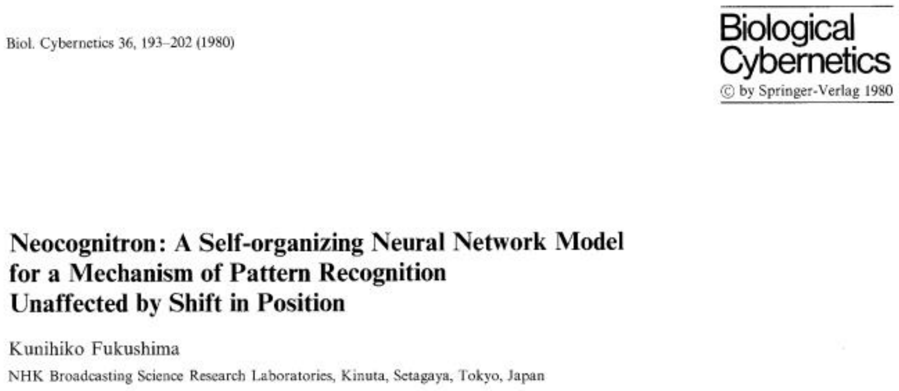
- Histogram of Oriented Gradients





Feature Learning

- Convolutional Neural Networks



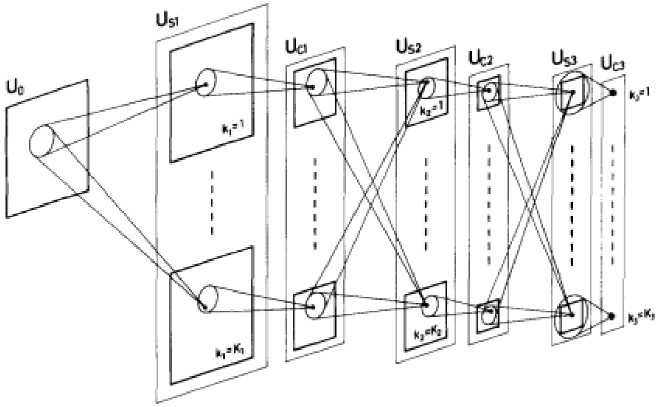
Biol. Cybernetics 36, 193–202 (1980)

Biological Cybernetics
© by Springer-Verlag 1980

Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position

Kunihiko Fukushima

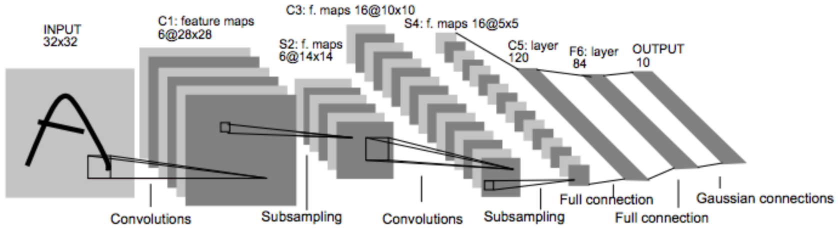
NHK Broadcasting Science Research Laboratories, Kinuta, Setagaya, Tokyo, Japan



PROC. OF THE IEEE, NOVEMBER 1998 1

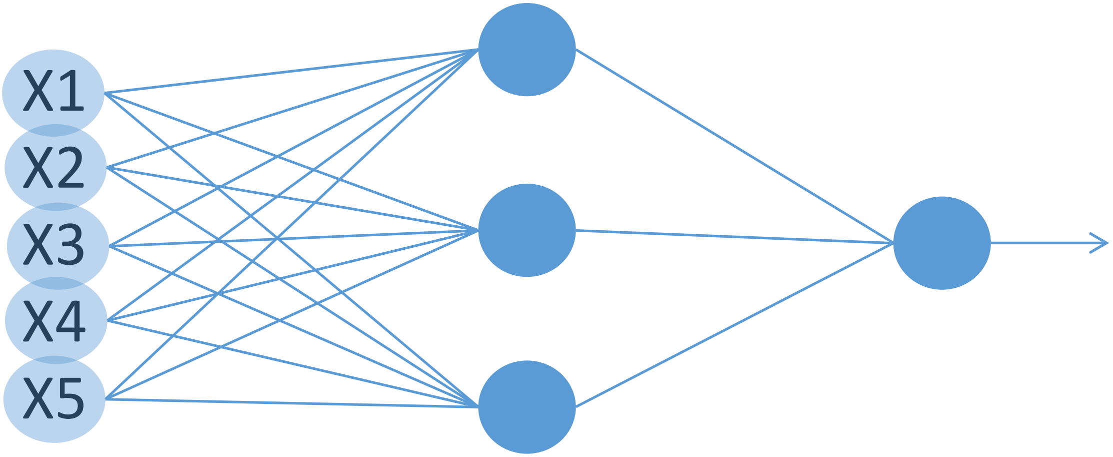
Gradient-Based Learning Applied to Document Recognition

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

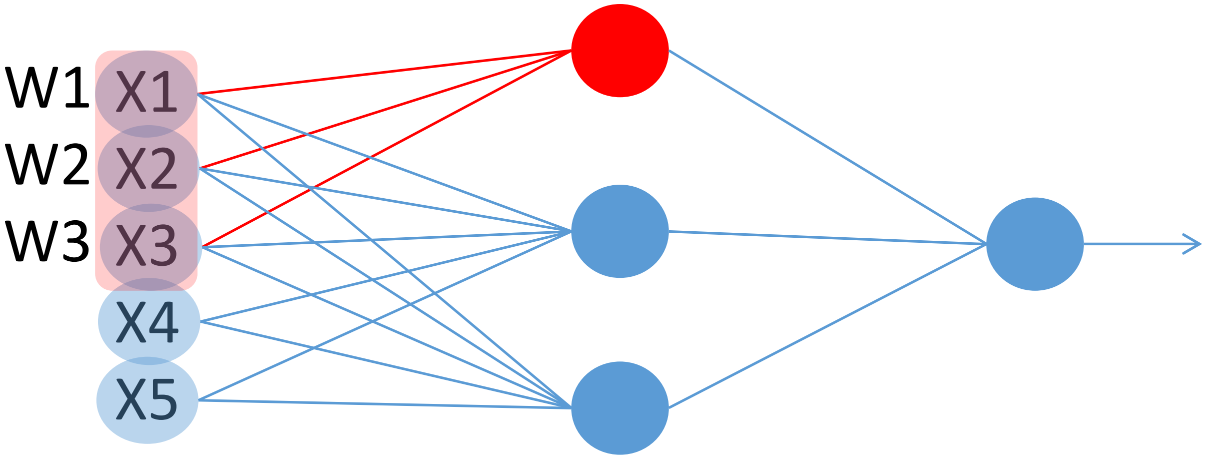


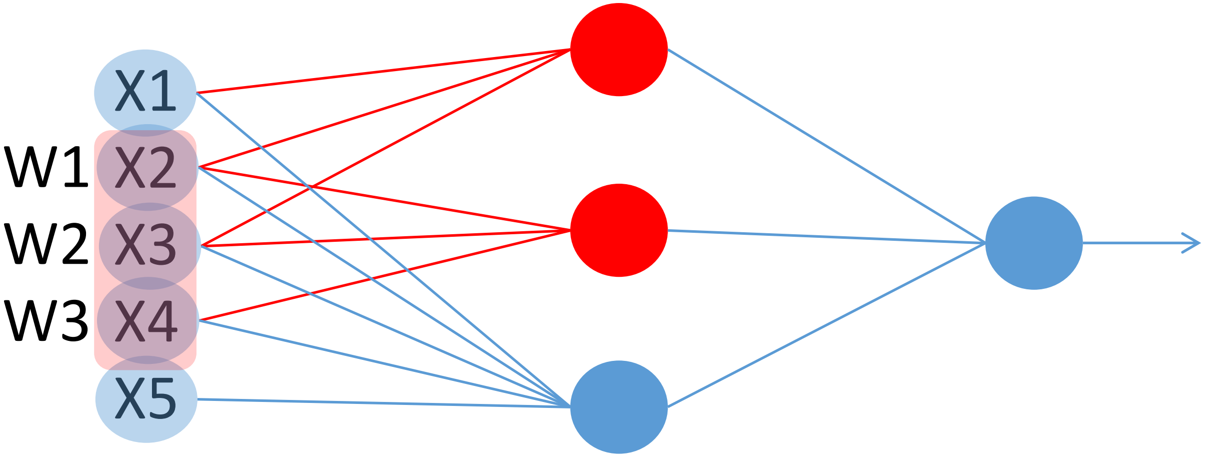
https://www.youtube.com/watch?v=Qil4kmvm2Sw

Convolutional Neural Networks (CNN)

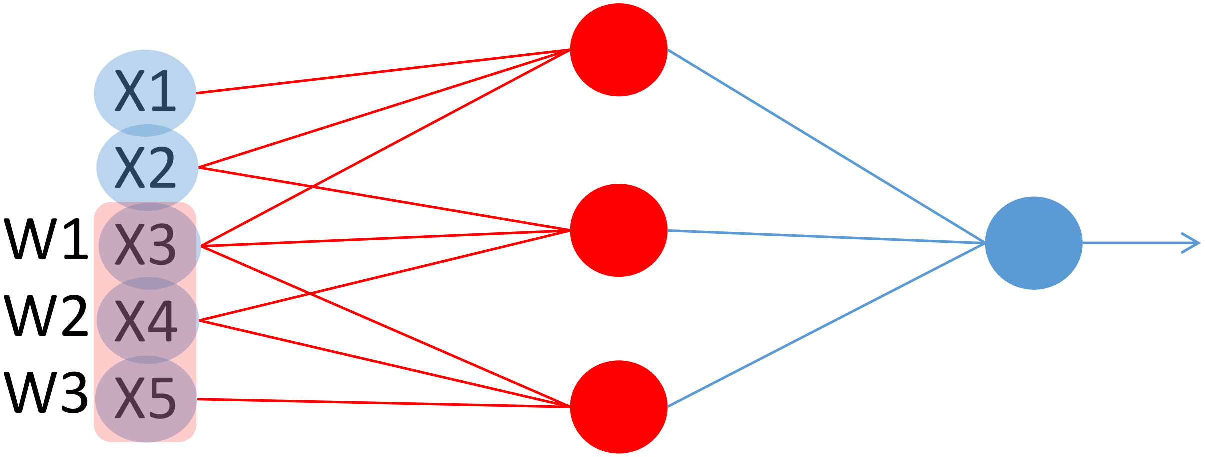


Convolutional Neural Networks (CNN)

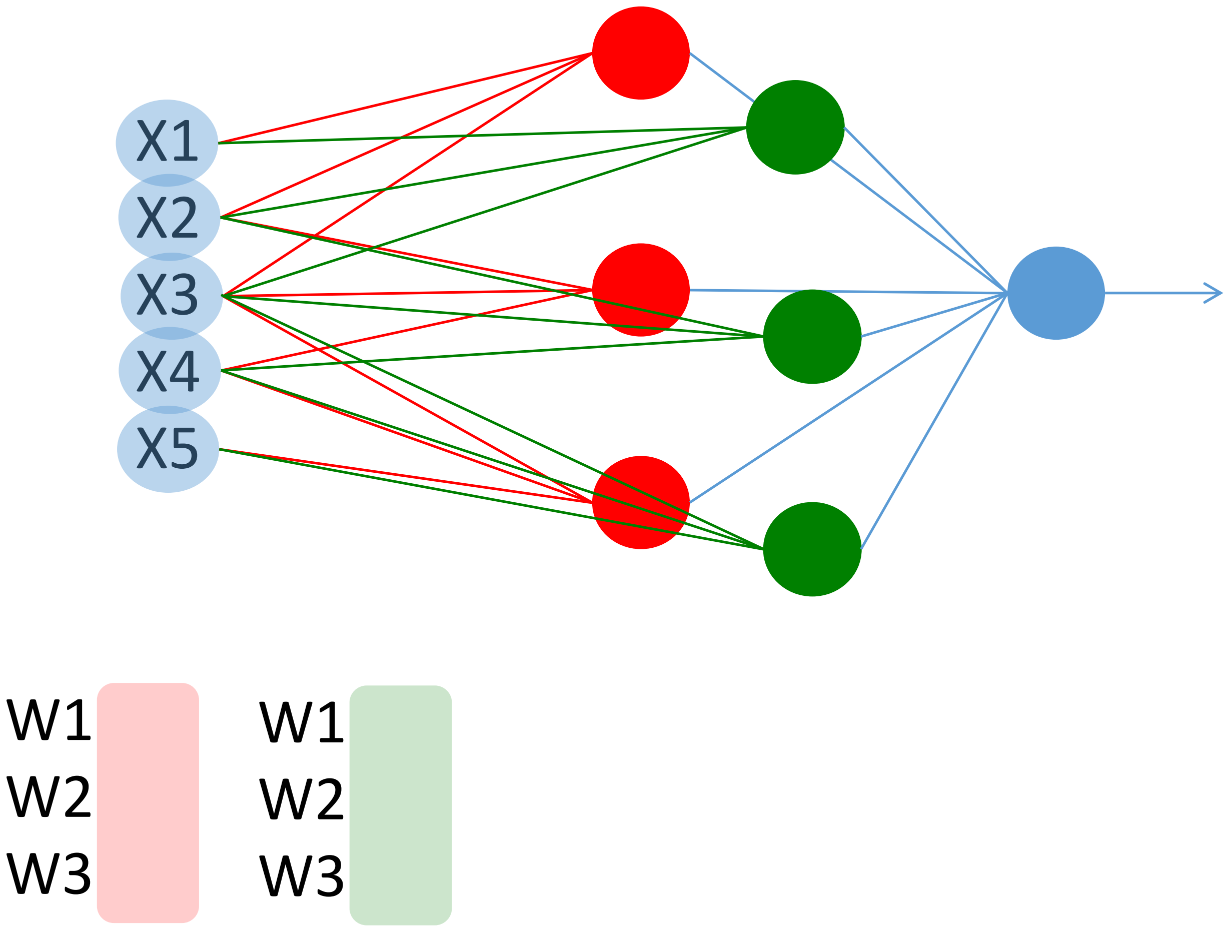


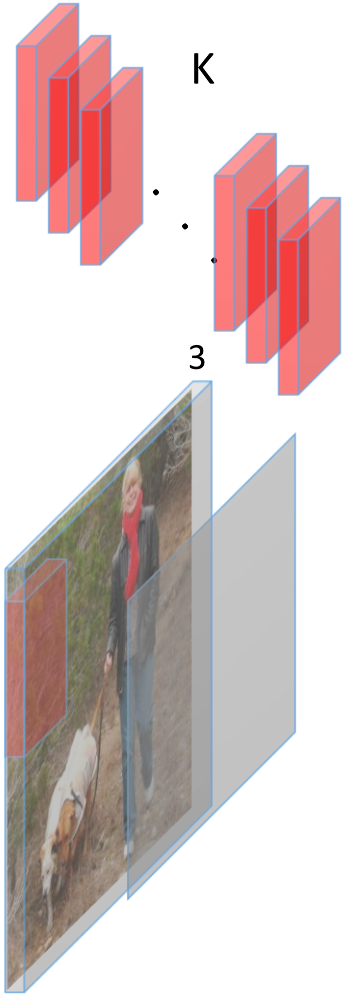

X1
W1
X2
W2
X3
W3
X4
X5

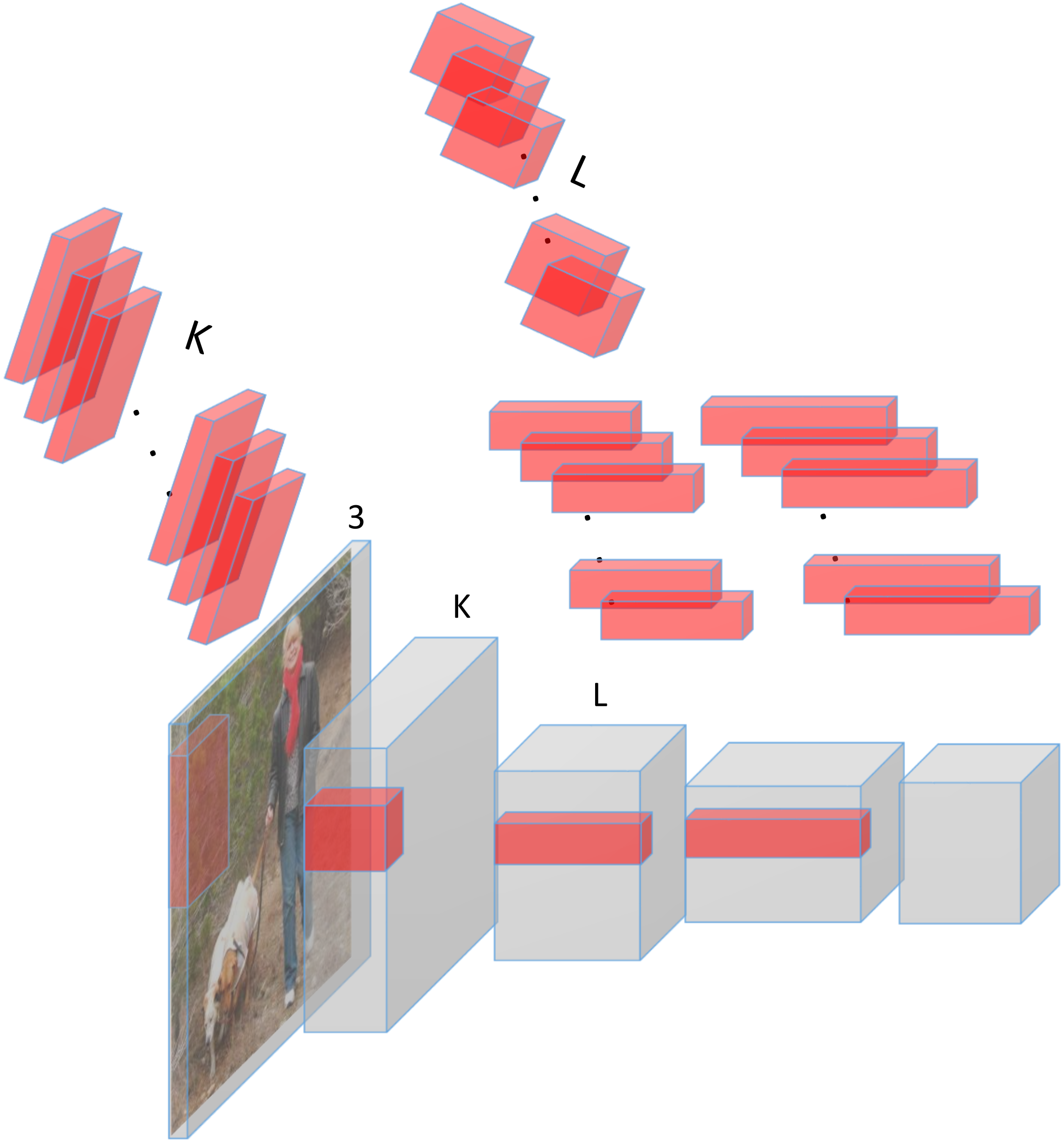
Convolutional Neural Networks (CNN)

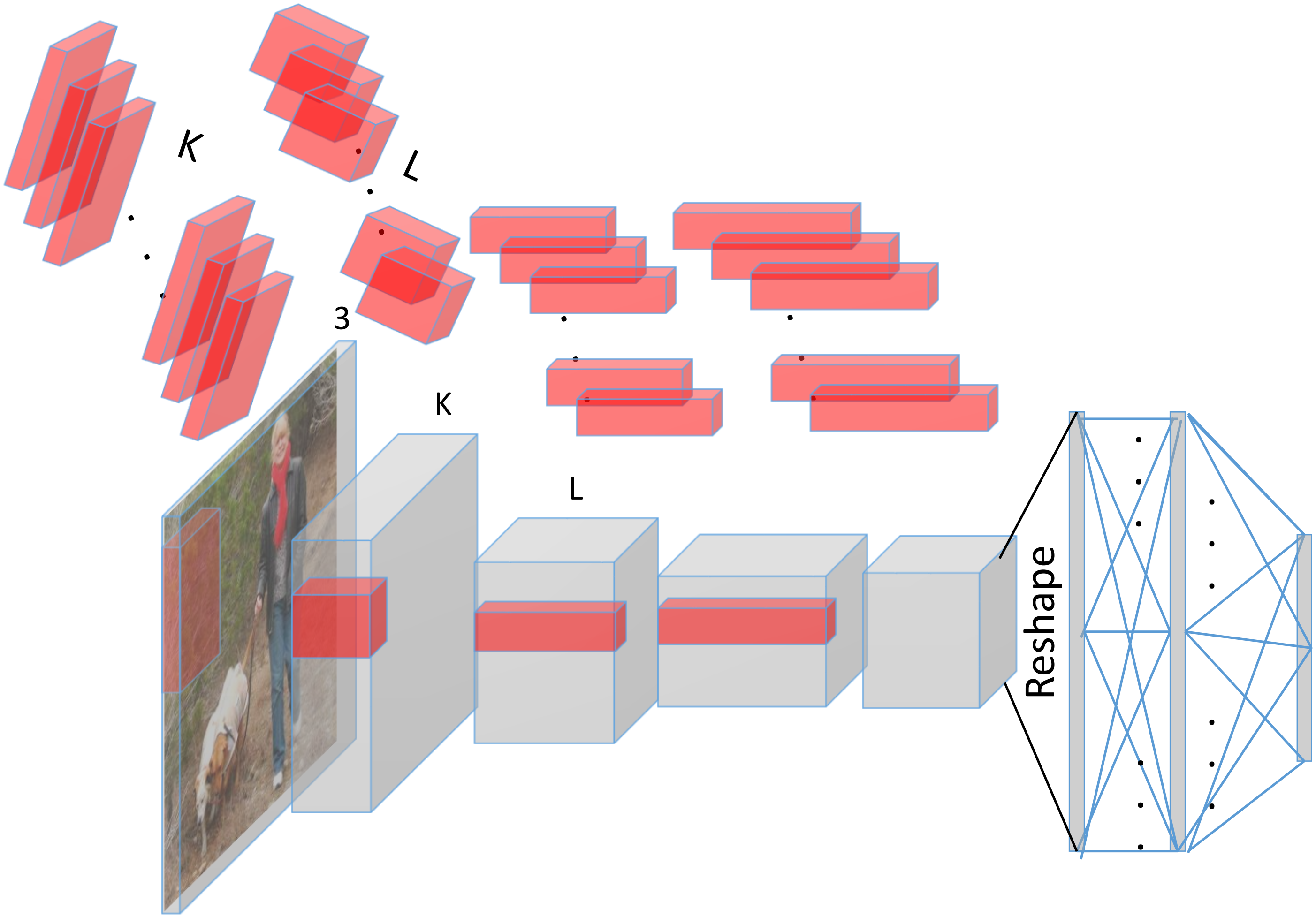


Convolutional Neural Networks (CNN)



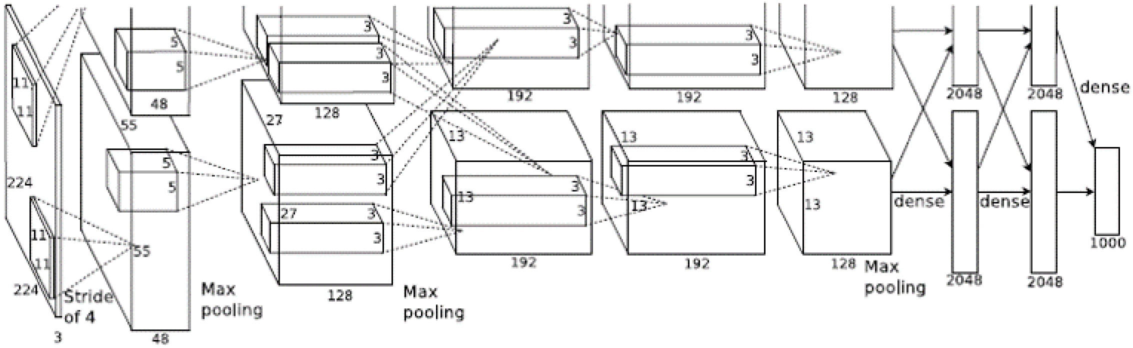

K
3

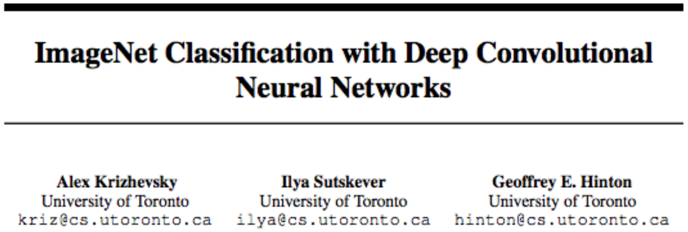

L
K
3
K
L


K
L
3
K
L
Reshape

GPU and BigData

- AlexNet (2012)



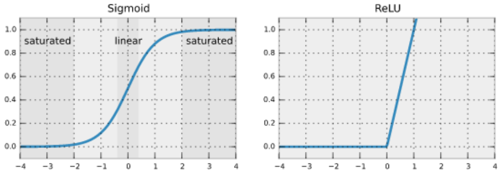


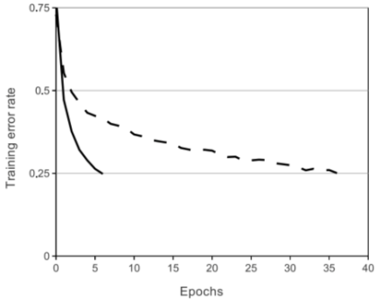
ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

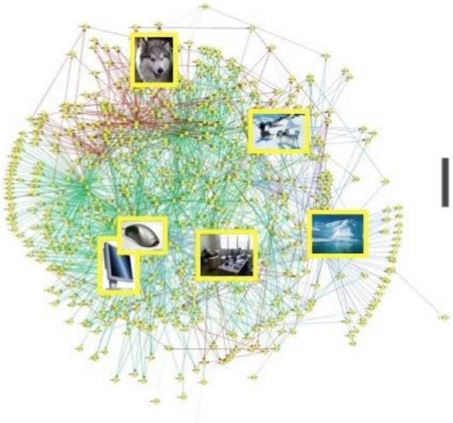
Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca





- ImageNet



IMAGENET

Outline

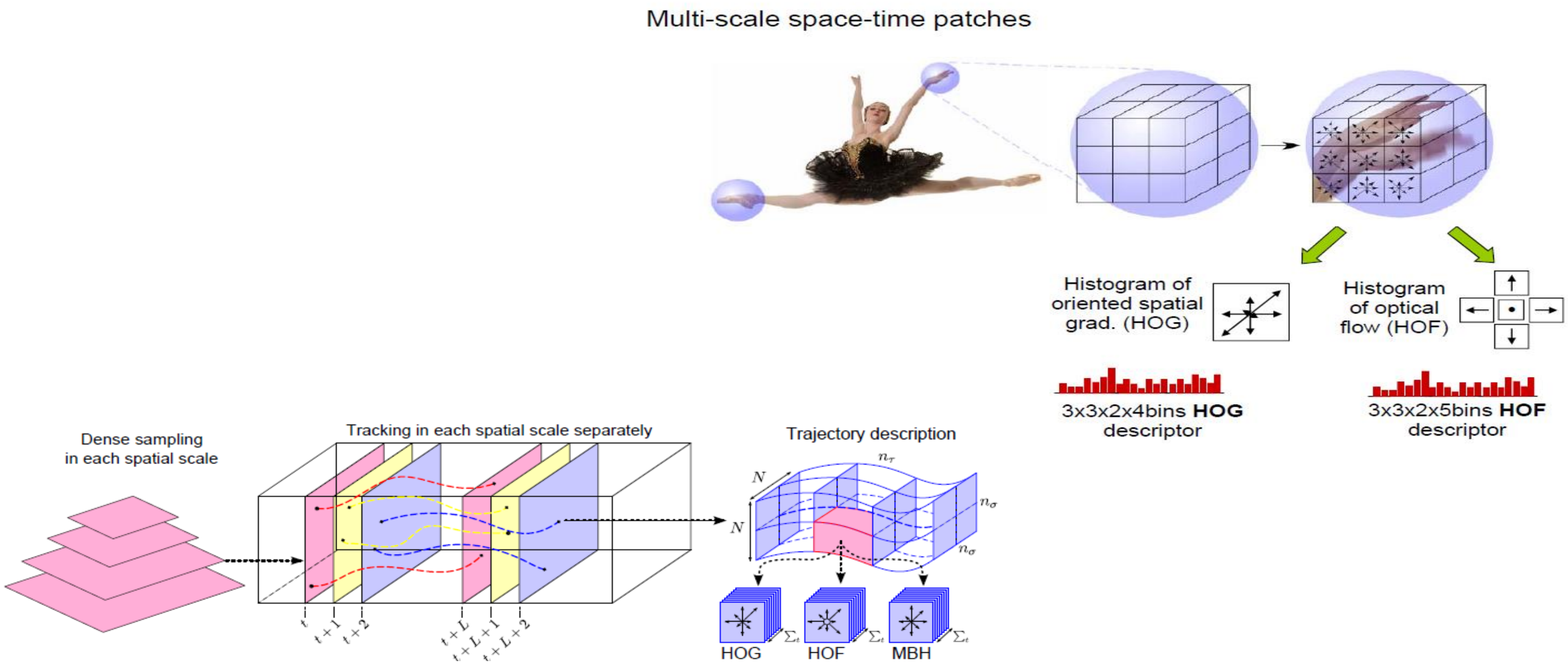
- **Neural Networks A Brief History**

- **Action Recognition with Image Based CNN Features**
  - **Introduction**
  - **Method**
  - **Experiment Results**

- **CNN-aware Binary Map For General Image Segmentation**

Introduction

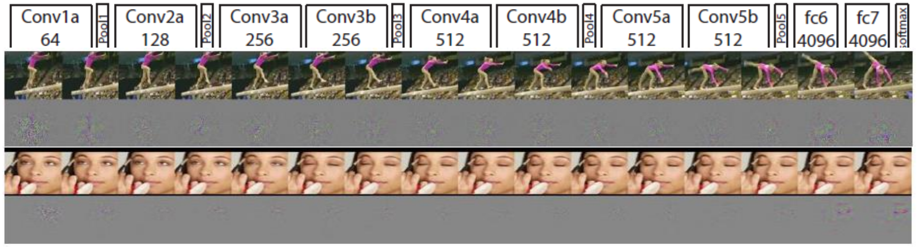
- Classical Models
    - Rely on complex handcrafted structures


Multi-scale space-time patches
Histogram of oriented spatial grad. (HOG)
Histogram of optical flow (HOF)
3x3x2x4bins **HOG** descriptor
3x3x2x5bins **HOF** descriptor
Dense sampling in each spatial scale
Tracking in each spatial scale separately
Trajectory description
t
$t+1$
$t+2$
$t+L$
$t+L+1$
$t+L+2$
N
N
n_τ
n_σ
n_σ
HOG
HOF
MBH
Σ_t
Σ_t
Σ_t

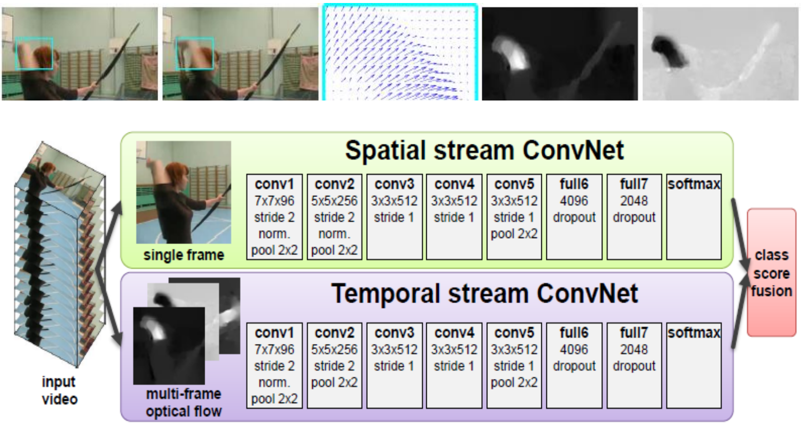
Wang et al. 'Action recognition by dense trajectories', ICCV 2011&2013

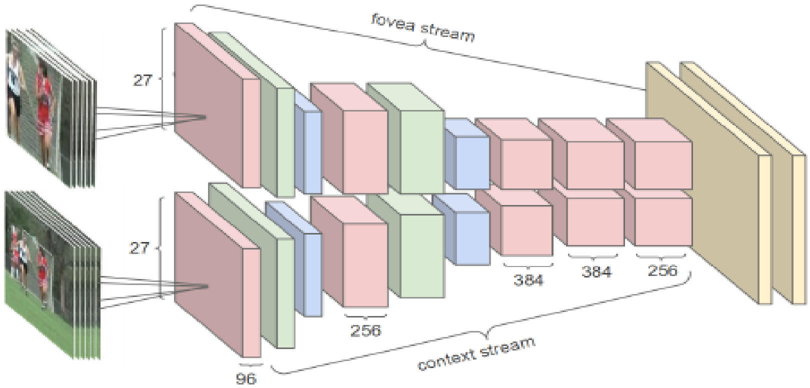
Introduction

- **Deep Models**
  - CNN network based on spatial-temporal domain



Tran, Du. et al. "Learning spatiotemporal features with 3d convolutional network





Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." *CVPR* (2014)

Simonyan and Zisserman. "Two-stream convolutional networks for action recognition in videos." *NIPS* (2014)

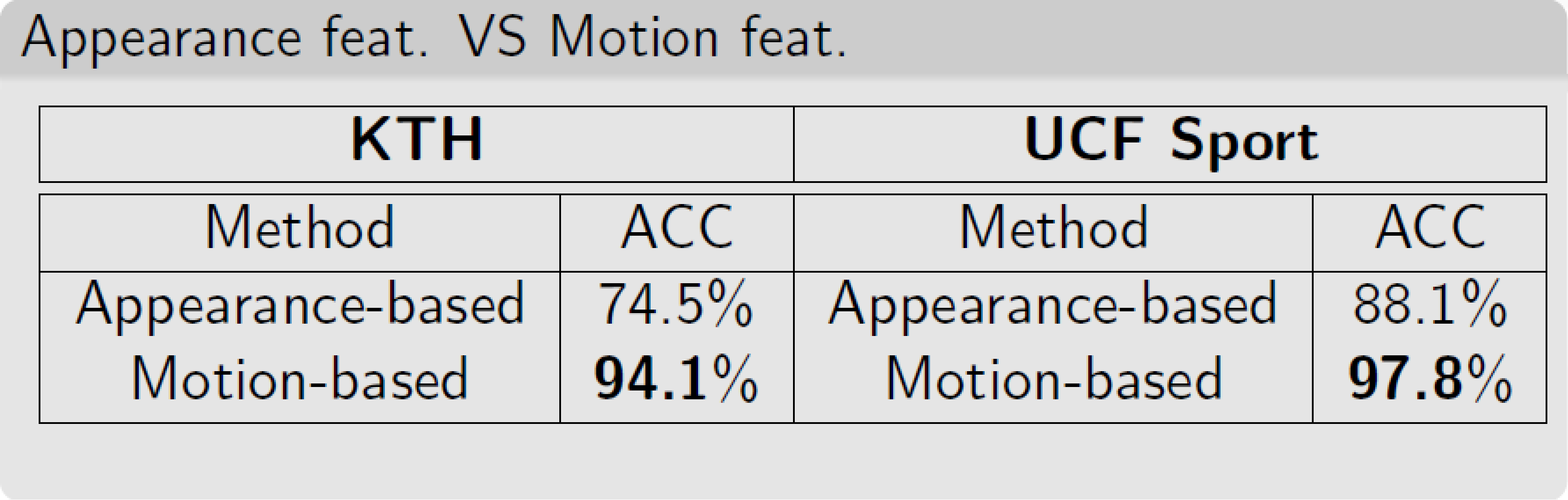
Outline

- Neural Networks A Brief History
- Action Recognition with Image Based CNN Features
  - Introduction
  - Method
  - Experiment Results
- CNN-aware Binary Map For General Image Segmentation

Method

- DeepNets have shown promising results on Image classification.

- Image-based CNNs can not overcome performance on Action Recognition task.



Appearance feat. VS Motion feat.

KTH		UCF Sport	
Method	ACC	Method	ACC
Appearance-based	74.5%	Appearance-based	88.1%
Motion-based	**94.1%**	Motion-based	**97.8%**

Contribution

- Spatial Features

  Use the ***AlexNet*** architecture pre-trained on ImageNet dataset.

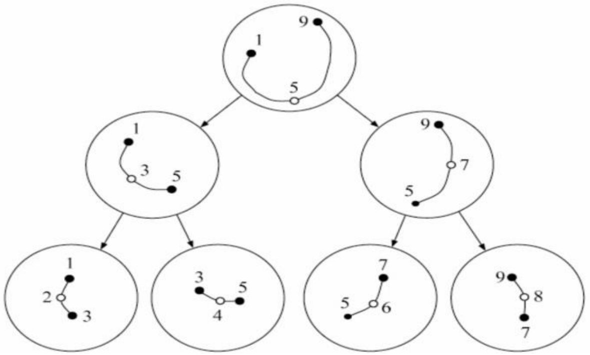
  We utilize the output of fc7 layer of this CNN for representing spatial information.

- Temporal Component

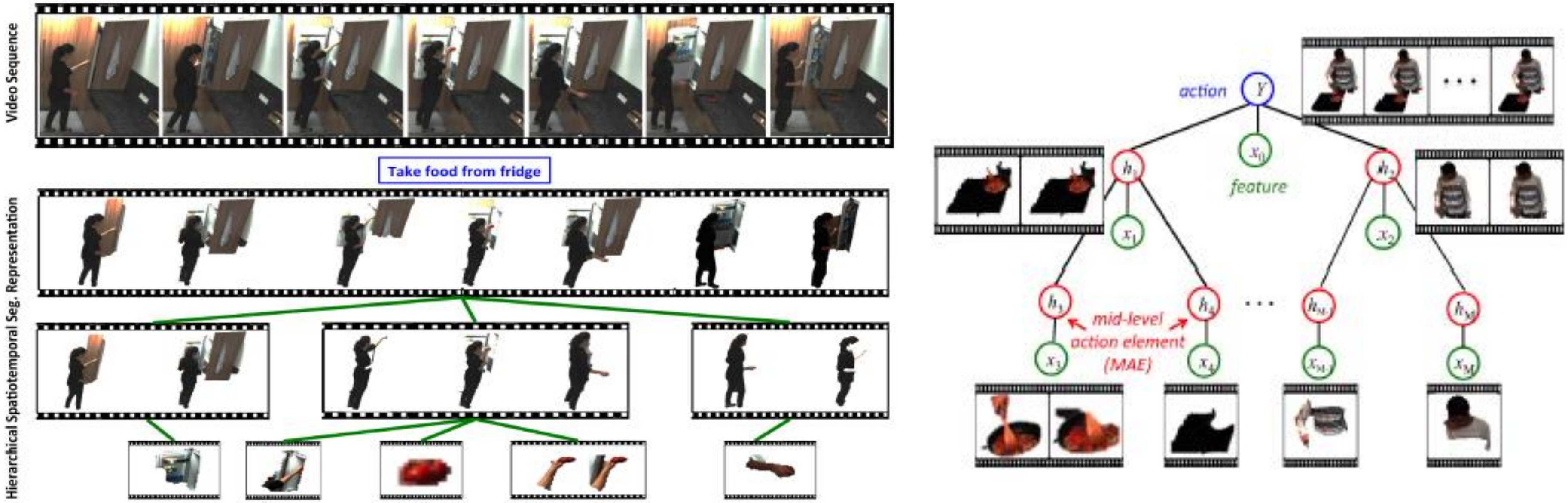
  ”***CNN flow***” captures informative features about image movement inspired from optical flow.

Introduction

- **Hierarchal structure**
  - ability to represent the information of a video in a multi-level.



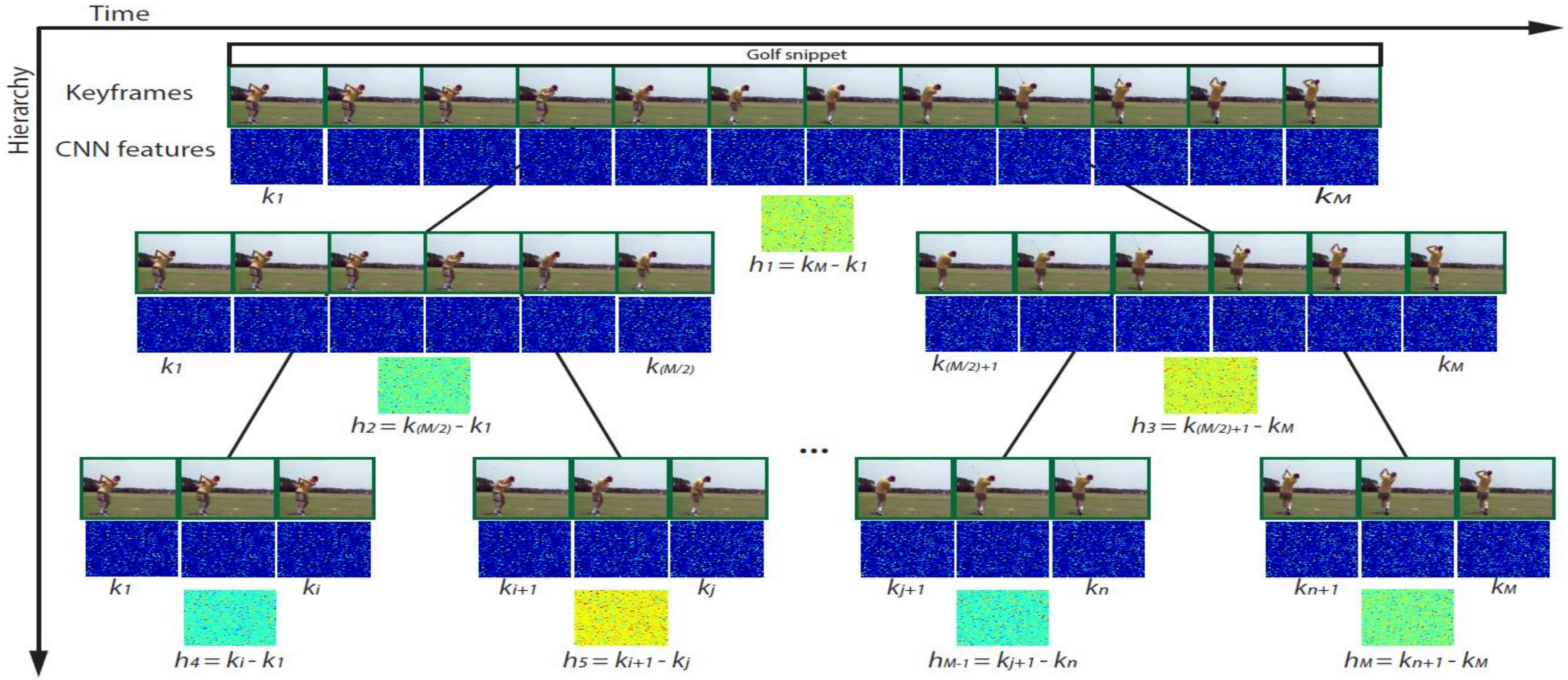
Felzenszwalb et al “Hierarchical matching of deformable shapes.”CVPR(2007)



Lan, Tian, et al. "Action Recognition by Hierarchical Mid-level Action Elements." *ICCV (*2015)

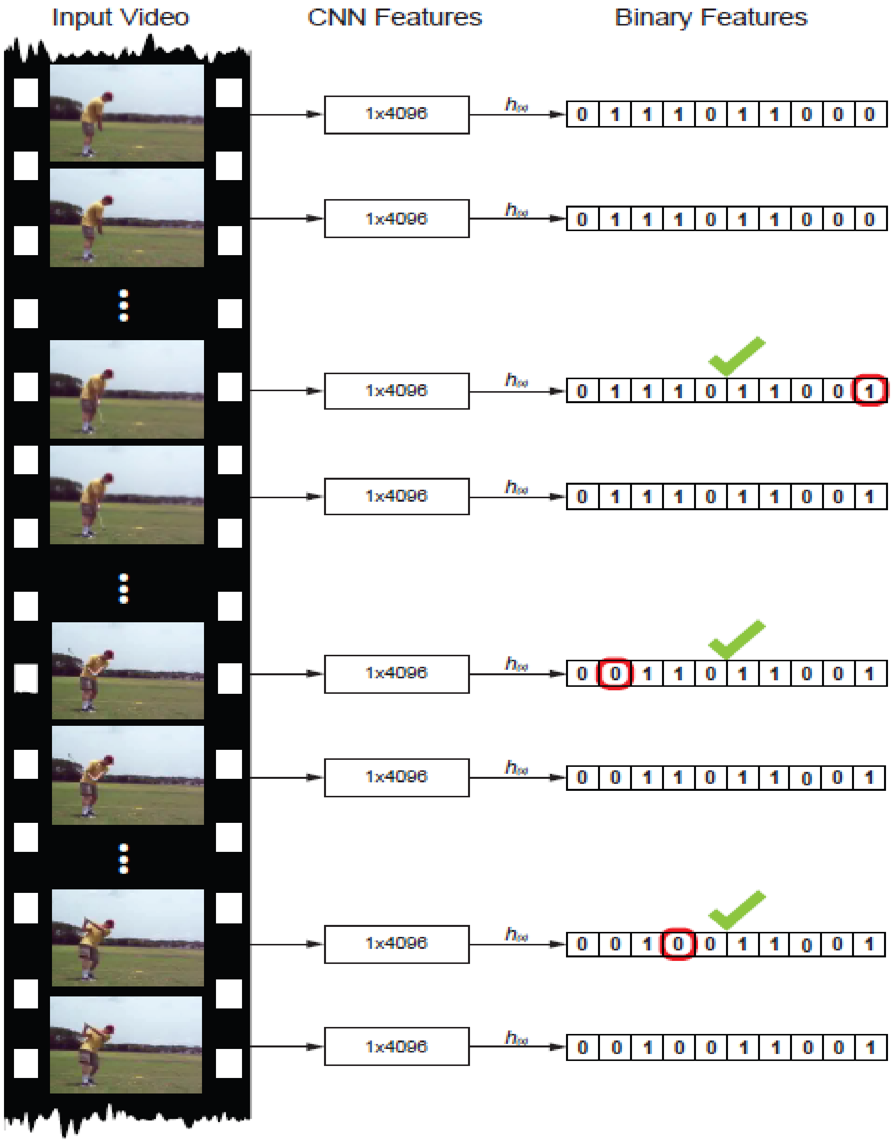
Hierarchical Model

- Enables to capture sub-actions from a complex action.
- Hierarchy can represent the information of a video in a multi-level of resolution.
- Coarse to fine representation (higher levels coarse action sequence, lower levels represent ne action elements)



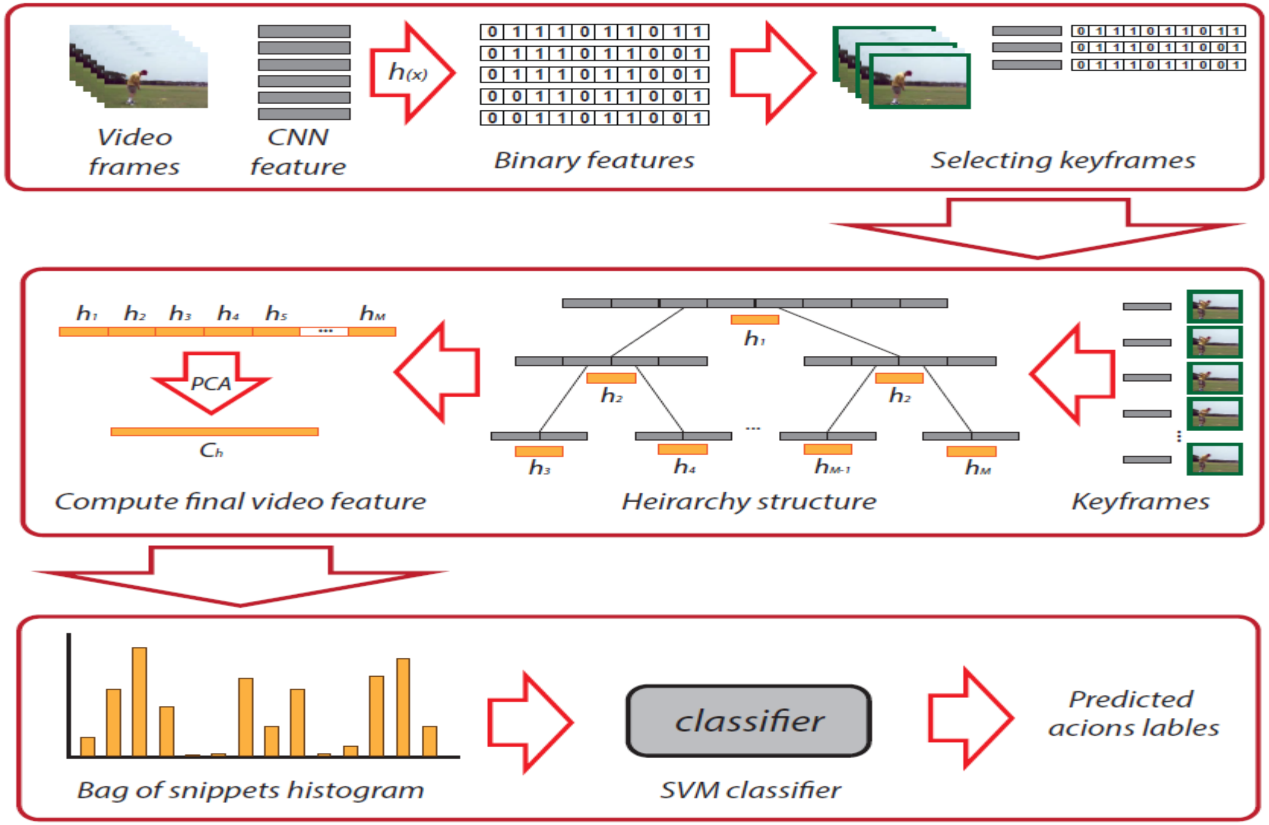
Key-frame Selection

- Extract CNN feature.
- Generate binary codes by Iterative Quantization (ITQ).
- Select key-frames regarding to binary code changes.



Pipeline Overview

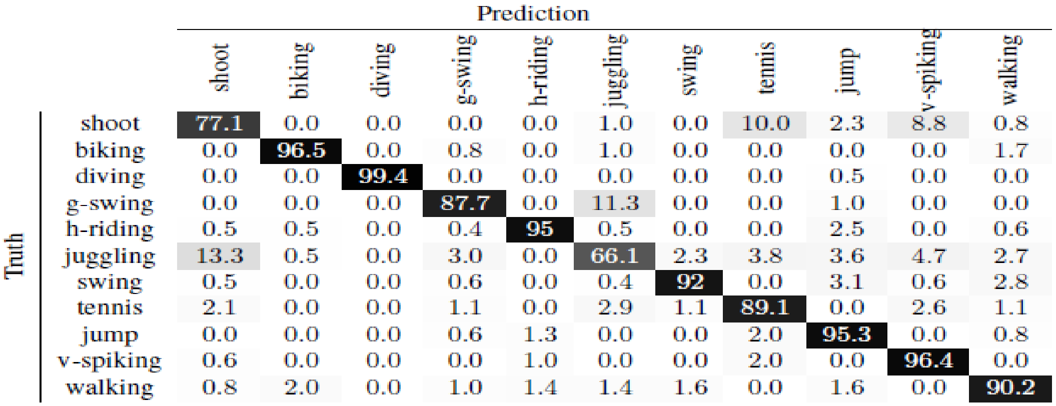
- Extract CNN feature, key-frames.
- Build pyramid, compute video feature.
- Bag of snippets, and classifier.



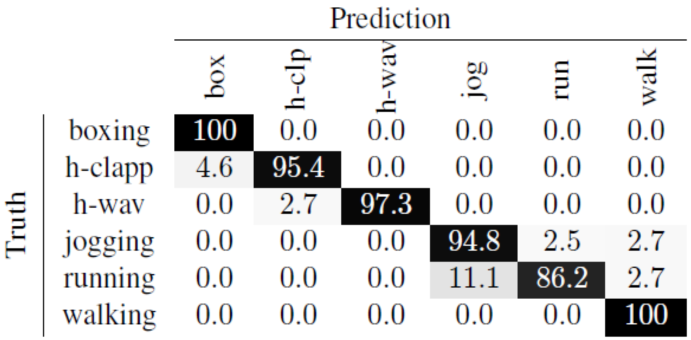
Outline

- **Neural Networks A Brief History**
- **Action Recognition with Image Based CNN Features**
  - **Introduction**
  - **Method**
  - **Experiment Results**
- **CNN-aware Binary Map For General Image Segmentation**

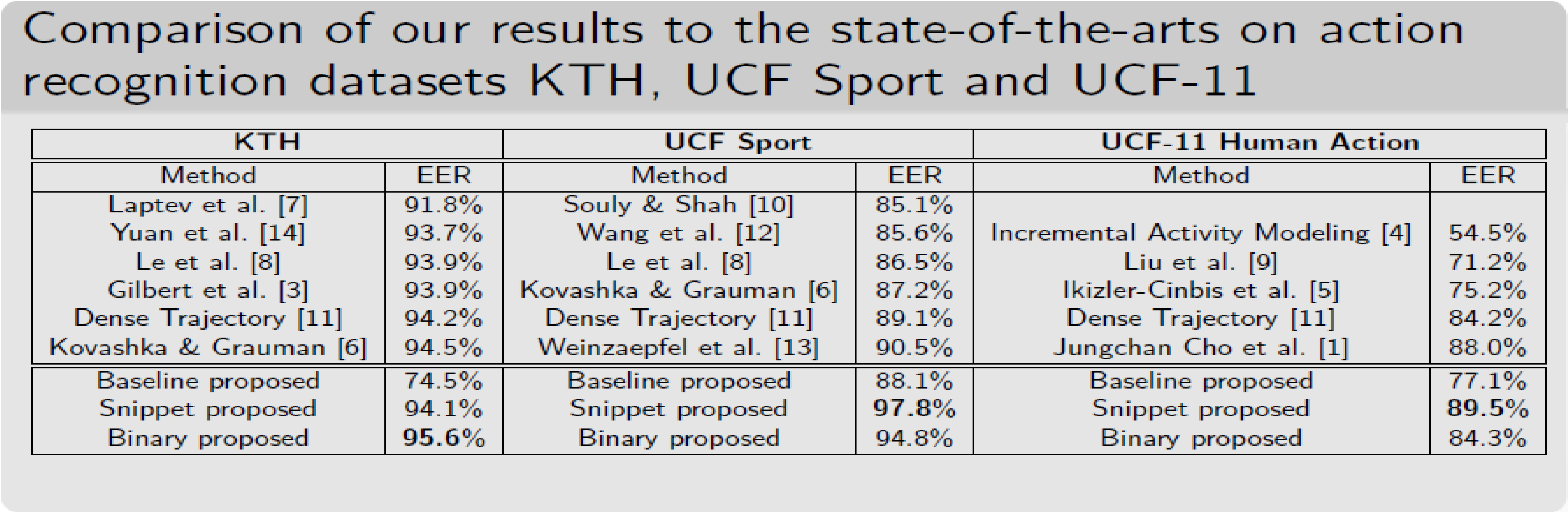
Experimental Results



		Prediction										
		shoot	biking	diving	g-swing	h-riding	juggling	swing	tennis	jump	v-spiking	walking
Truth	shoot	77.1	0.0	0.0	0.0	0.0	1.0	0.0	10.0	2.3	8.8	0.8
	biking	0.0	96.5	0.0	0.8	0.0	1.0	0.0	0.0	0.0	0.0	1.7
	diving	0.0	0.0	99.4	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0
	g-swing	0.0	0.0	0.0	87.7	0.0	11.3	0.0	0.0	1.0	0.0	0.0
	h-riding	0.5	0.5	0.0	0.4	95	0.5	0.0	0.0	2.5	0.0	0.6
	juggling	13.3	0.5	0.0	3.0	0.0	66.1	2.3	3.8	3.6	4.7	2.7
	swing	0.5	0.0	0.0	0.6	0.0	0.4	92	0.0	3.1	0.6	2.8
	tennis	2.1	0.0	0.0	1.1	0.0	2.9	1.1	89.1	0.0	2.6	1.1
	jump	0.0	0.0	0.0	0.6	1.3	0.0	0.0	2.0	95.3	0.0	0.8
	v-spiking	0.6	0.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0	96.4	0.0
	walking	0.8	2.0	0.0	1.0	1.4	1.4	1.6	0.0	1.6	0.0	90.2



		Prediction					
		box	h-clp	h-wav	jog	run	walk
Truth	boxing	100	0.0	0.0	0.0	0.0	0.0
	h-clapp	4.6	95.4	0.0	0.0	0.0	0.0
	h-wav	0.0	2.7	97.3	0.0	0.0	0.0
	jogging	0.0	0.0	0.0	94.8	2.5	2.7
	running	0.0	0.0	0.0	11.1	86.2	2.7
	walking	0.0	0.0	0.0	0.0	0.0	100



Comparison of our results to the state-of-the-arts on action recognition datasets KTH, UCF Sport and UCF-11

KTH		UCF Sport		UCF-11 Human Action	
Method	EER	Method	EER	Method	EER
Laptev et al. [7]	91.8%	Souly & Shah [10]	85.1%		
Yuan et al. [14]	93.7%	Wang et al. [12]	85.6%	Incremental Activity Modeling [4]	54.5%
Le et al. [8]	93.9%	Le et al. [8]	86.5%	Liu et al. [9]	71.2%
Gilbert et al. [3]	93.9%	Kovashka & Grauman [6]	87.2%	Ikizler-Cinbis et al. [5]	75.2%
Dense Trajectory [11]	94.2%	Dense Trajectory [11]	89.1%	Dense Trajectory [11]	84.2%
Kovashka & Grauman [6]	94.5%	Weinzaepfel et al. [13]	90.5%	Jungchan Cho et al. [1]	88.0%
Baseline proposed	74.5%	Baseline proposed	88.1%	Baseline proposed	77.1%
Snippet proposed	94.1%	Snippet proposed	**97.8%**	Snippet proposed	**89.5%**
Binary proposed	**95.6%**	Binary proposed	94.8%	Binary proposed	84.3%

Outline

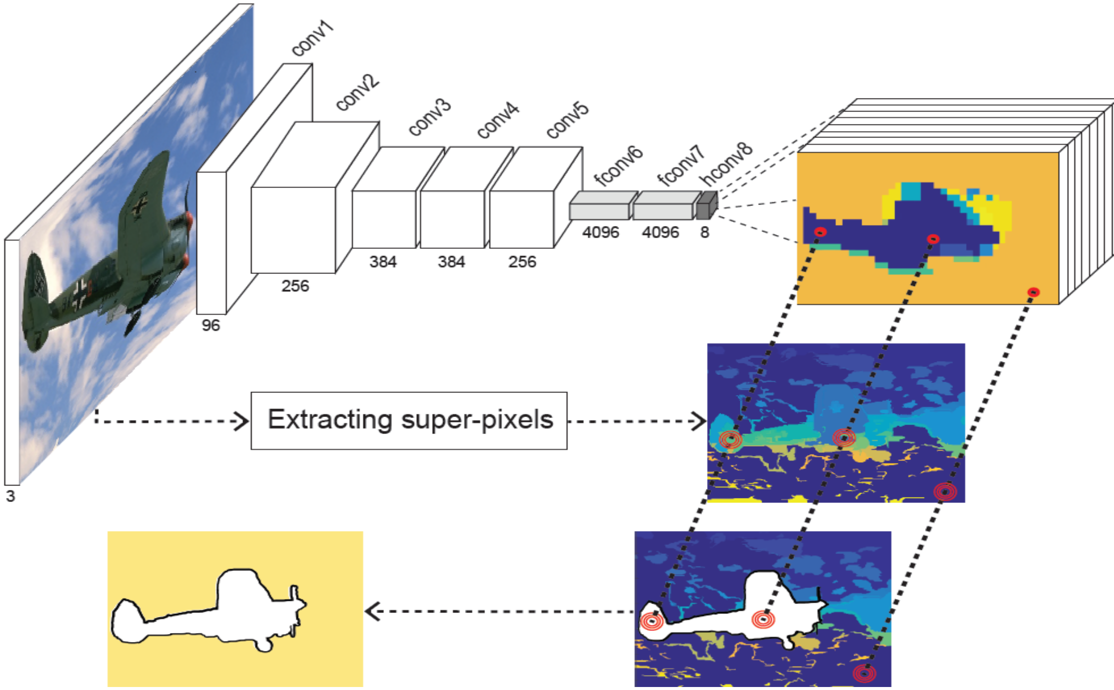
- **Neural Networks A Brief History**

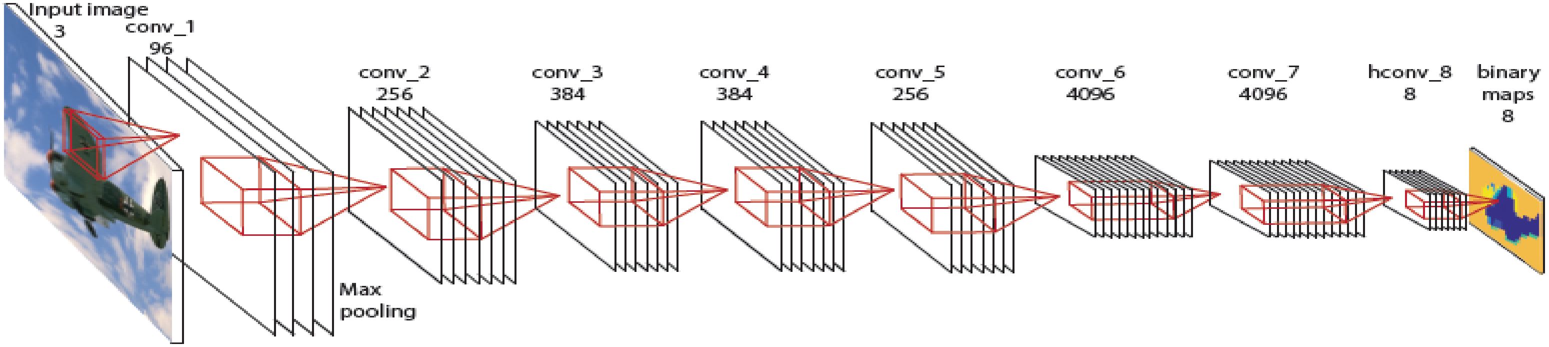
- **Action Recognition with Image Based CNN Features**
  - **Introduction**
  - **Method**
  - **Experiment Results**

- **CNN-aware Binary Map For General Image Segmentation**

CNN-aware Binary Map For General Image Segmentation

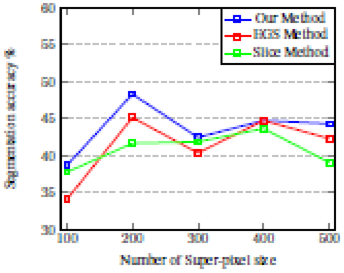
- Visually and semantically coherent image segments



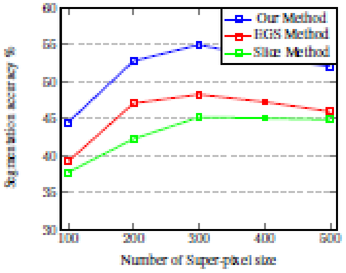


Experimental Result

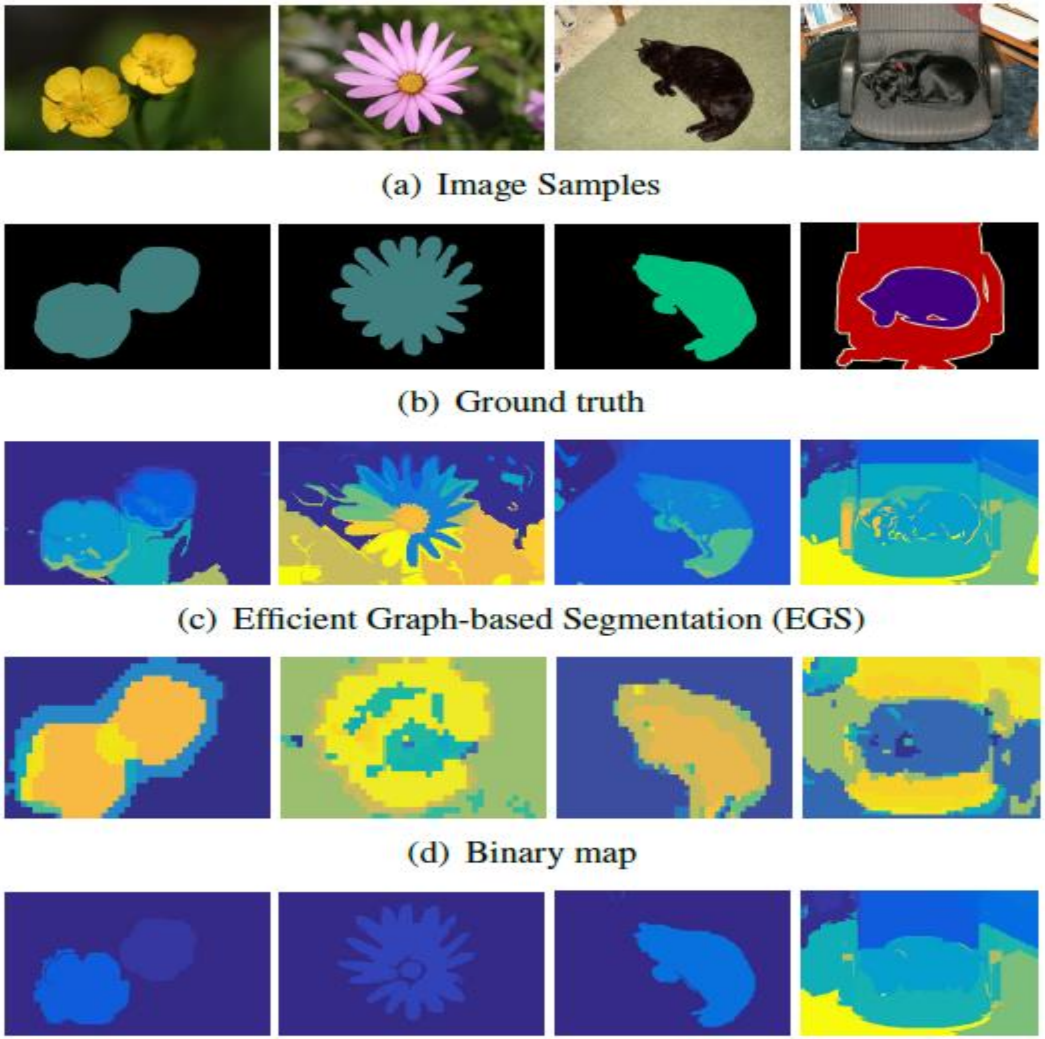
MSRC		Berkeley	
Method	IoU	Method	IoU
EGS [1]	50.3%	EGS [1]	45.19%
SLIC [2]	48.7%	SLIC [2]	43.70%
Our method	**55.03 %**	Our method	**48.35%**



(a) Accuracy on Berkeley



(b) Accuracy on MSRC



(a) Image Samples

(b) Ground truth

(c) Efficient Graph-based Segmentation (EGS)

(d) Binary map

(e) Our method

Conclusion&Future work

- Proposed hierarchical structure of CNN features
- we introduced CNN-flow Inspired by optical flow
- Find key-frames to build snippets
- Train network based on our proposed approaches toward action recognition. (Model CNN flow)
- Apply proposed segmentation for action detection

Thank you!