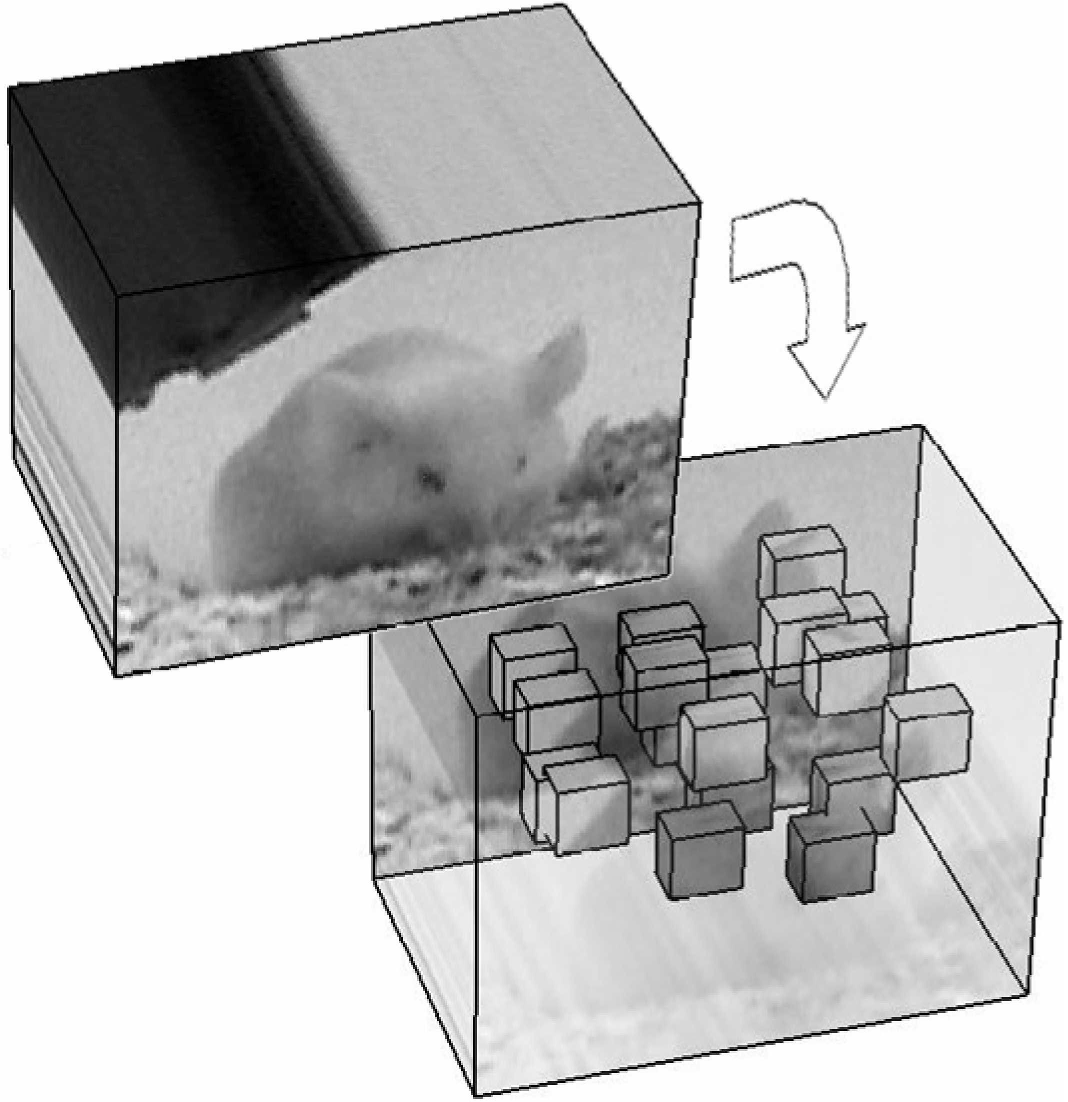
# Behavior Recognition via Sparse Spatio-Temporal Features

Piotr Dollár
Vincent Rabaud
Garrison Cottrell
Serge Belongie

Antón R. Escobedo
cse 252c

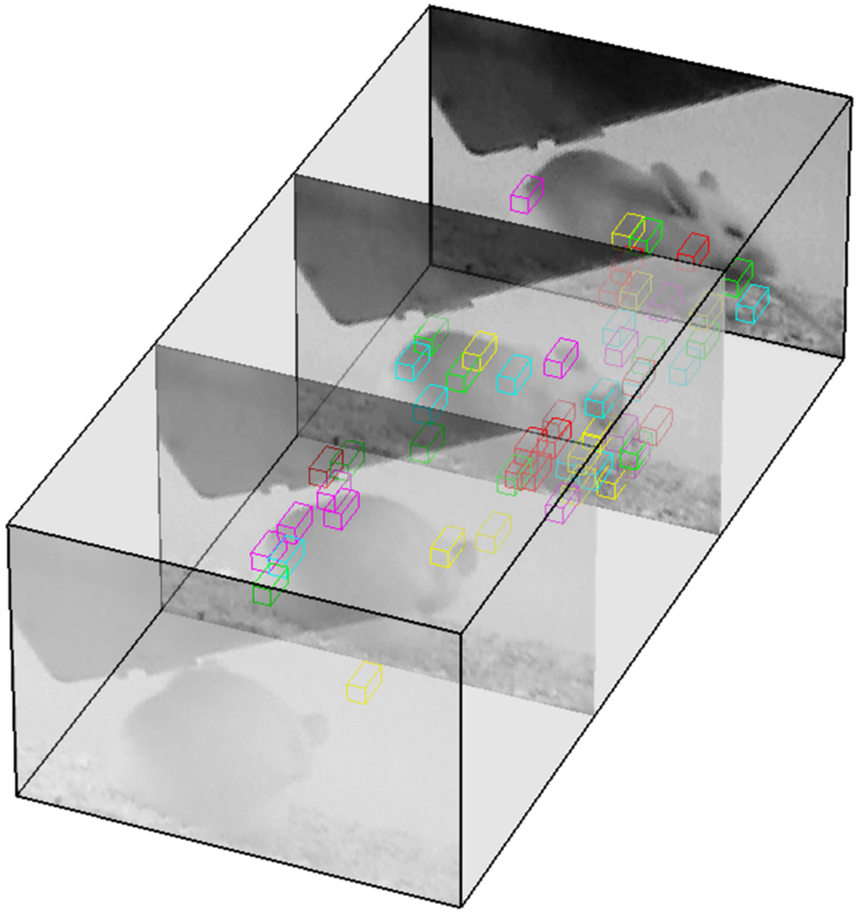
# Outline

- I. Introduction
- II. Related Work
- III. Algorithm
- IV. Experiments
- V. Current Work



# Part I: Introduction

- Motivation:
  - Sparse feature points extended to the spatio-temporal case



# Part I: Introduction

- Motivation:
  - Behavior detection from video sequences
    - Behavior recognition faces similar issues to those seen in object recognition.
    - Posture, appearance, size, image clutter, variations in the environment such as illumination.
    - Imprecise nature of feature detectors.

# Part I: Introduction

- Inspiration: Sparsely detected features in object recognition.
  - Fergus et al. “Object Class Recognition by Unsupervised Scale-Invariant Learning”
  - Agarwal et al. “Learning to Detect Objects in Images via a Sparse, Part-Based Representation”
  - Leibe, Schiele “Scale invariant Object Categorization Using a Scale-Adaptive Mean-Shift Search”

# Part I: Introduction

## Advantages of Sparse Features

- Robustness
- Very good results

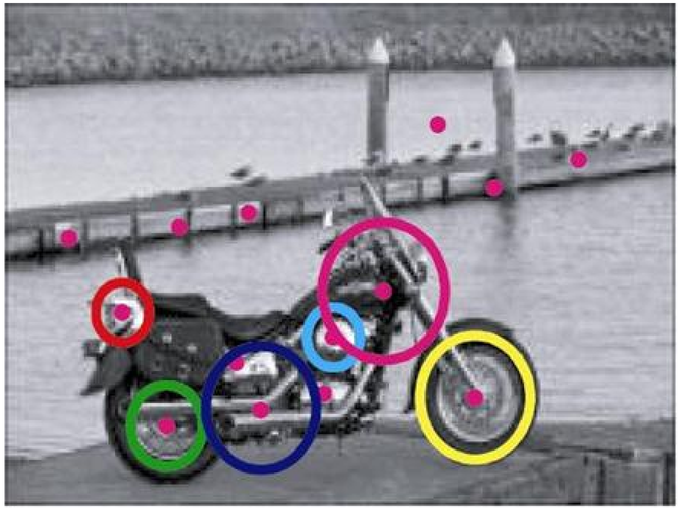
example from: http://www.robots.ox.ac.uk/~fergus/research/index.html

# object recognition

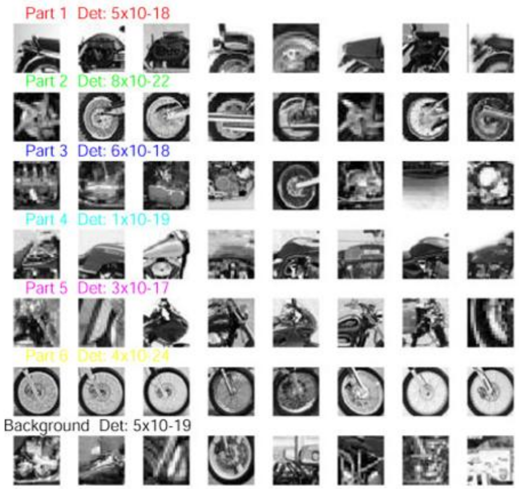
example from: http://www.robots.ox.ac.uk/~fergus/research/index.html

[training data & features]

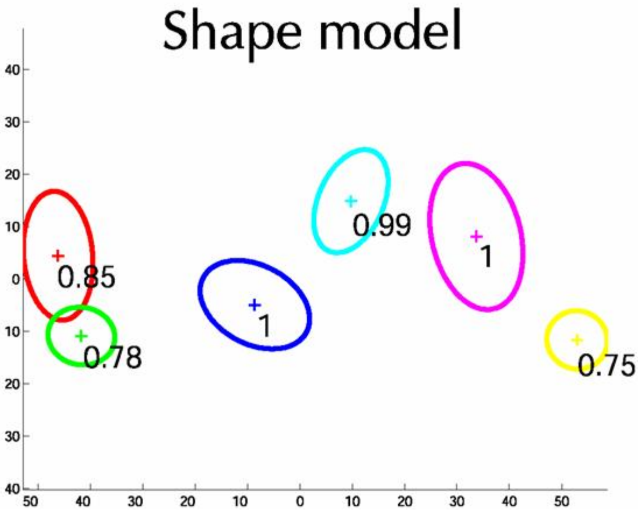
[motorcycle detected]



train




[parts]




[object model]

classify

# Spatial-Temporal Features

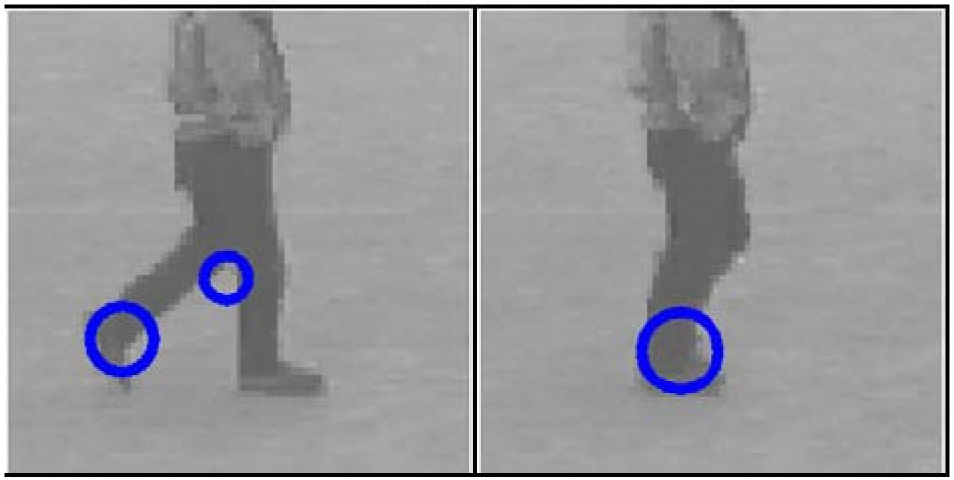
- Short, local video sequence that can be used to describe a behavior.
- Behavior recognition based on features detected and compared with a rich set of features.
- 3rd dimension
  - Temporal, not spatial

# Part I: Introduction

- Will Show:
  - Direct 3D counterparts to feature detectors are inadequate.
  - Development and testing of descriptors used in this paper.
  - A dictionary of descriptors is all that is needed to recognize behavior.
    - Proven on human behavior, facial expressions and mouse behavior dataset

# Part II: Related Work

- Articulated models
- Efros et al.
  - 30 pixel man
- Schuldt et al.
  - Spatio-Temporal features



Images from:
http://www.eecs.berkeley.edu/Research/Projects/CS/vision/human/efros-iccv03_slides/efrosbmm_iccv03.ppt
ftp://ftp.nada.kth.se/CVAP/users/laptev/icpr04actions.pdf

# Part III: Proposed Algorithm

- Feature Detection
- Cuboids
- Cuboid Prototypes
- Behavior Descriptors

# Feature Detection (spatial domain)

- Corner Detectors
- Laplacian of Gaussian (SIFT)
- Extensions to Spatio-Temporal Case
  - Stacks of images denoted by: I(x,y,t)
  - Detected features also have temporal extent.

# Feature Detection

- Harris in 3D
  - Spatio-Temporal corners:
    - Regions where the local gradient vectors point in orthogonal directions for x,y and t.
  - Why this doesn't work
- Develop an Alternative detector
  - Err on the side of too many features
  - Why this works

# Feature Detection

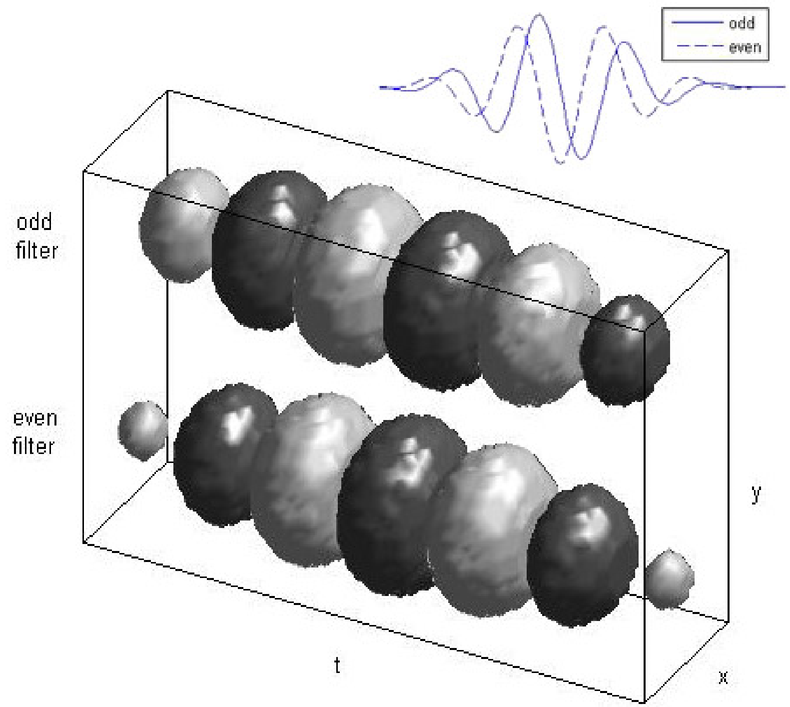
- Response Function

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$$

  - Spatial Filter:
    Gaussian
  - Temporal Filter:
    Gabor

$$h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$$

$$h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$$

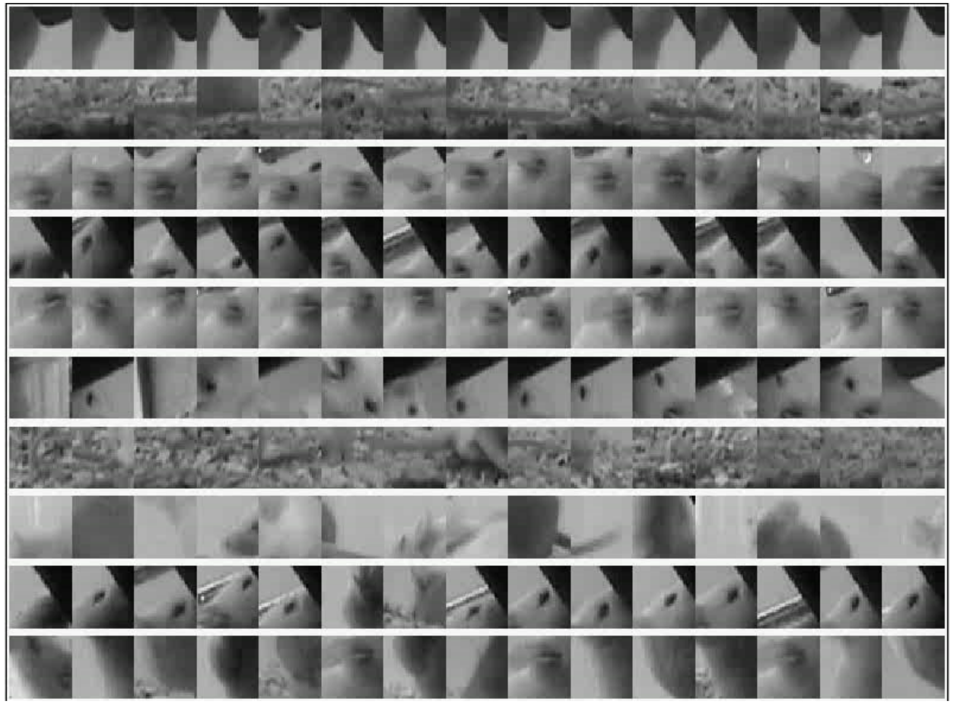


# Feature Detection

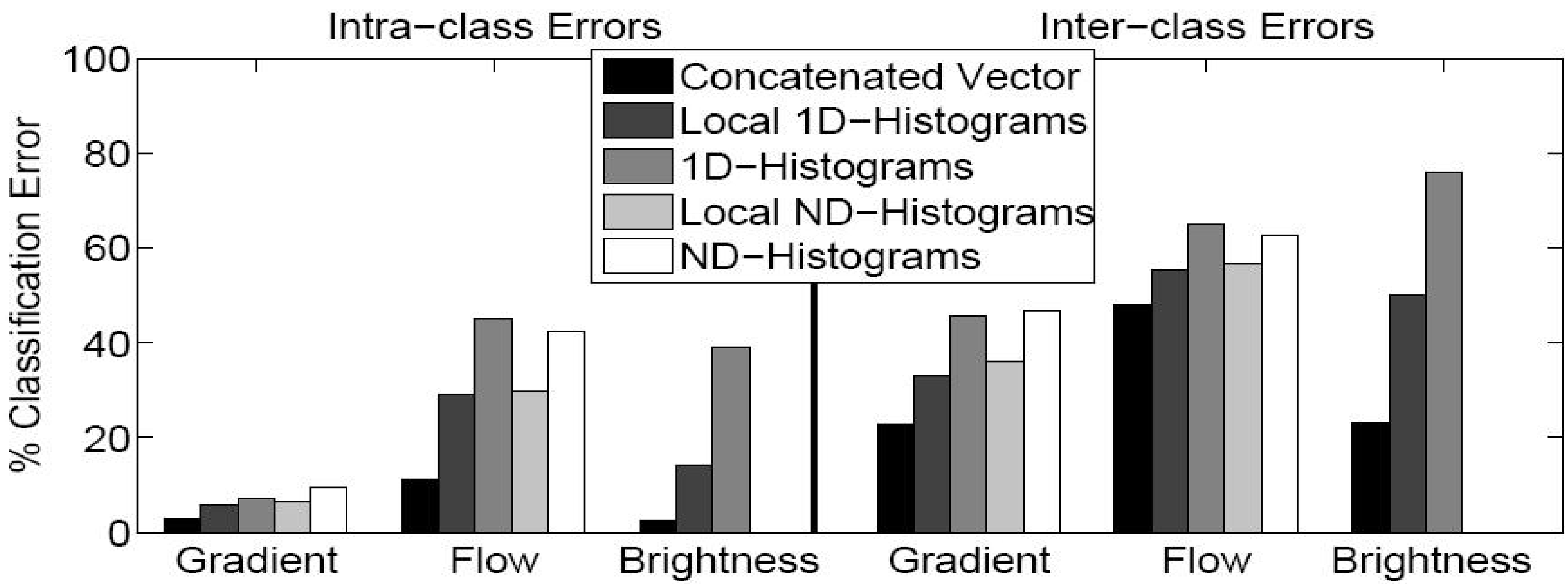
- What this implies:
  - Any region with spatially distinguishing characteristics undergoing a complex motion will induce a strong response.
  - Pure translation will not induce a response.

# Cuboids

- Extracted at each interest point
  - ~6x scale at which detected
- Descriptor: Feature Vector
  - Transformations Applied
    - Normalize Pixel Values
    - Brightness Gradient
    - Optical Flow
  - Feature Vector from local histograms



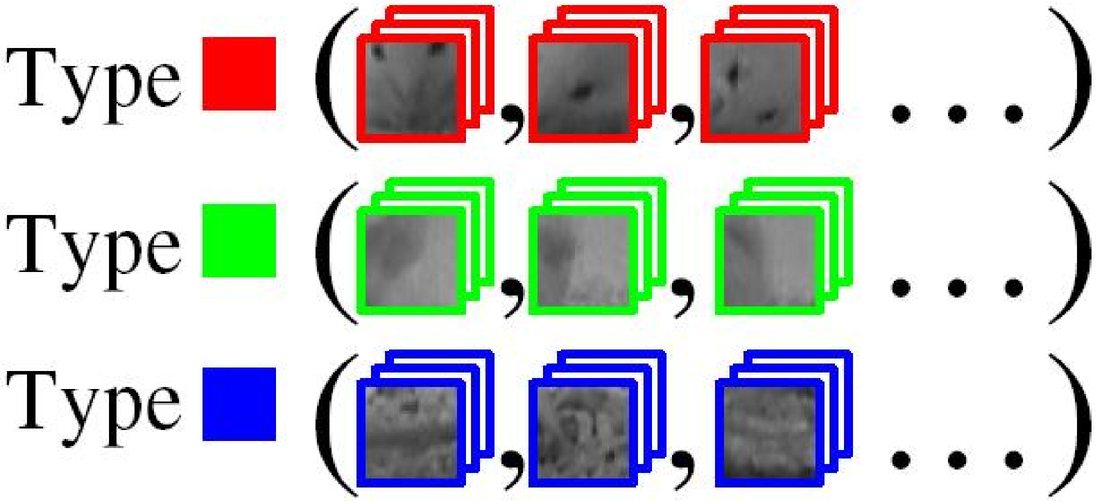
# Cuboid Descriptor



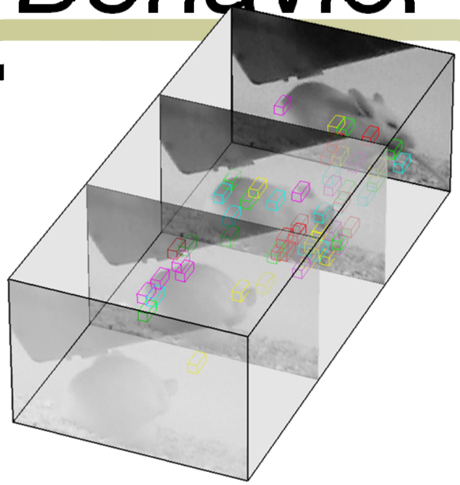

- Flattened Gradient vector gave best results
- Generalization of PCA-SIFT descriptor

# Cuboid Prototypes

- Unlimited cuboids are possible, but only a limited number of types exist.
- Use k-means algorithm to cluster extracted cuboids together by type.



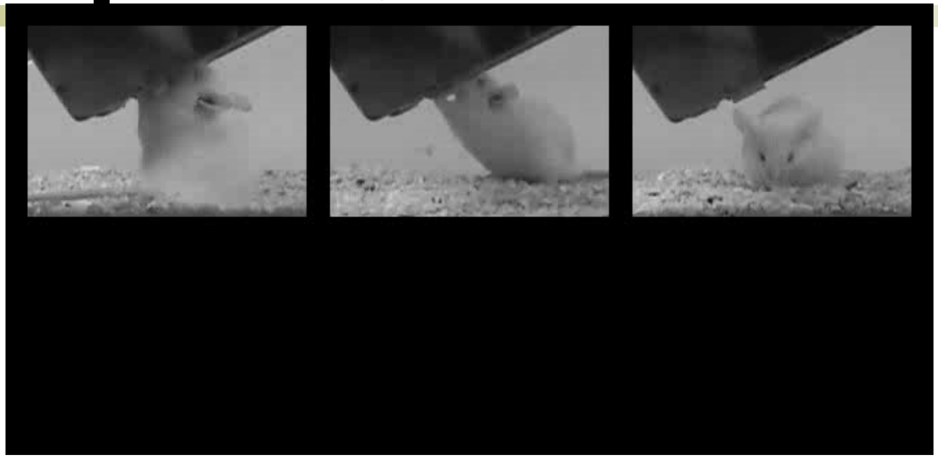
# Behavior Descriptor

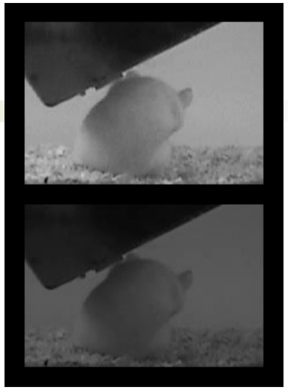


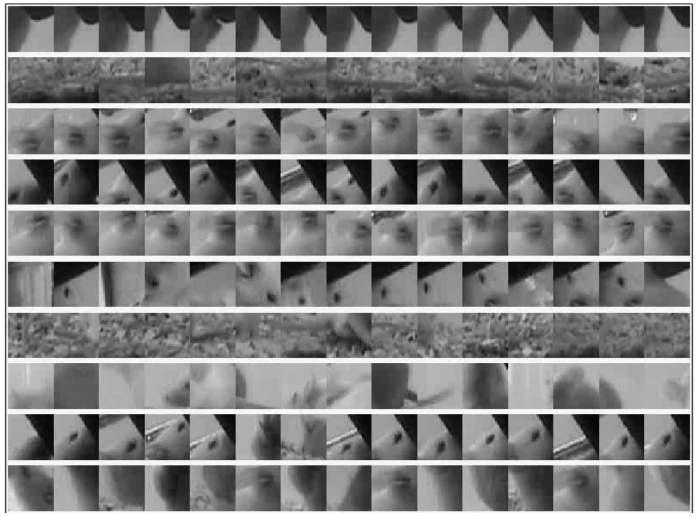
- Assumption is that cuboid types present capture all information about behavior.

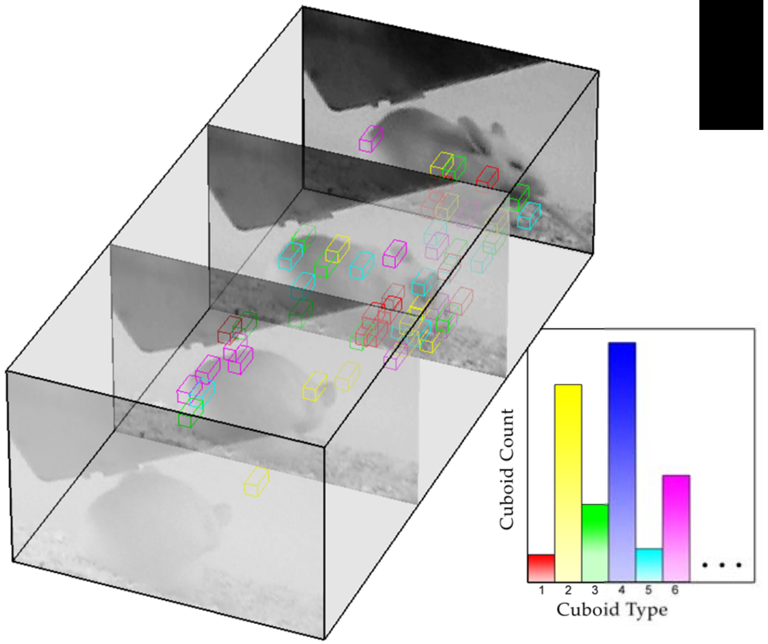
- Behavior descriptor: histogram of cuboid types
  - Simple.
  - Distance measured using chi-squared distance.
  - Can easily be used in classification framework.
  - Discards spatial layout and temporal order of cuboids.

# spatio-temporal features


[training data & features]


[grooming detected]


[cuboid prototypes]



[behavior model]

# domain 1:  human activity

training examples: test example:

boxing clapping ? boxing

clips from Schüldt et al.

# domain 2: facial expressions

training examples:

test example:

disgust

happiness

?

disgust

# domain 3: mouse behavior

training examples: test example:

eating exploring ? eating

# near vs. medium field

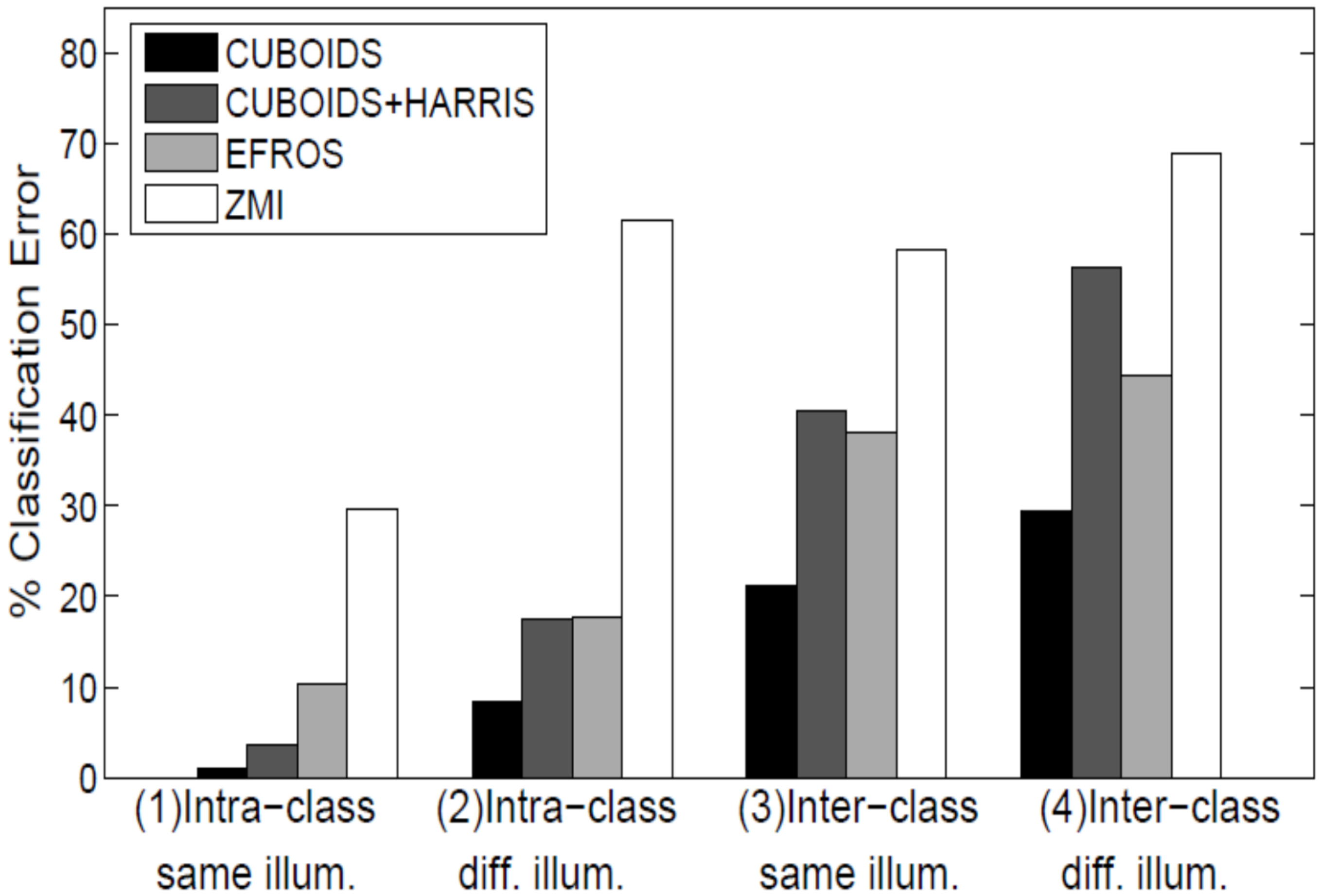
Efros et al. 2003

near field　medium field　far field

# performance evaluation

- compared 4 methods:
  - **CUBOIDS** – our approach
  - **CUBOIDS+HARRIS** – our approach using Laptev's 3D corner detector
  - **ZMI** – Zelnik-Manor & Irani 2001
    - Statistical measure of gross activity using histograms of spatio-temporal gradients gives activity descriptor
  - **EFROS** – Efros et al. 2003
    - Normalized cross correlation of optical flow gives distance measure between activities
- analysis in terms of **relative** performance
- not all algorithms are always applied
  - format of data, computational complexity

# facial expressions I



# facial expressions II

**Train on Sub 1 / Test on Sub 2**

| | anger | disgust | fear | joy | sadness | surprise |
|---|---|---|---|---|---|---|
| anger | .29 | .13 | .47 | .11 | .00 | .00 |
| disgust | .02 | .94 | .04 | .00 | .00 | .00 |
| fear | .00 | .00 | .90 | .00 | .06 | .04 |
| joy | .00 | .00 | .00 | 1.0 | .00 | .00 |
| sadness | .00 | .01 | .11 | .00 | .88 | .00 |
| surprise | .00 | .00 | .00 | .00 | .00 | 1.0 |

**Train on Sub 2 / Test on Sub 1**

| | anger | disgust | fear | joy | sadness | surprise |
|---|---|---|---|---|---|---|
| anger | .61 | .32 | .00 | .07 | .00 | .00 |
| disgust | .06 | .94 | .00 | .00 | .00 | .00 |
| fear | .17 | .13 | .64 | .04 | .00 | .03 |
| joy | .07 | .00 | .00 | .93 | .00 | .00 |
| sadness | .00 | .00 | .00 | .00 | 1.0 | .00 |
| surprise | .00 | .00 | .00 | .00 | .00 | 1.0 |

**confusion matrices, row normalized**

# mouse behavior
# full database results

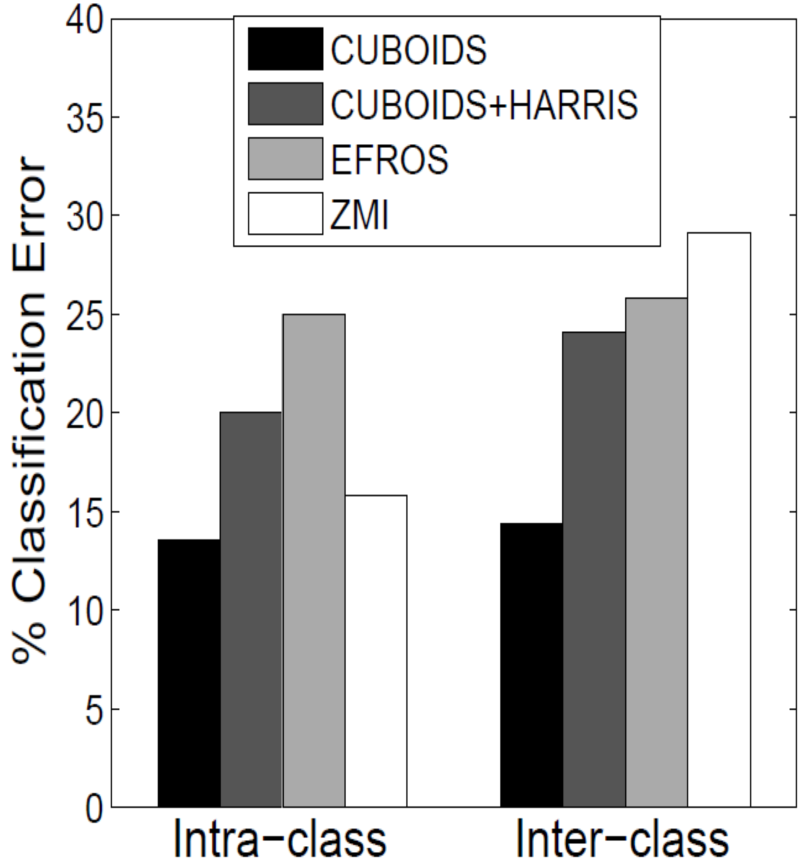
1−NN

| | drink | eat | explore | groom | sleep |
|---|---|---|---|---|---|
| drink | .65 | .00 | .24 | .06 | .06 |
| eat | .00 | .89 | .09 | .02 | .00 |
| explore | .02 | .04 | .86 | .07 | .02 |
| groom | .05 | .00 | .64 | .32 | .00 |
| sleep | .02 | .00 | .09 | .02 | .87 |

LDA

| | drink | eat | explore | groom | sleep |
|---|---|---|---|---|---|
| drink | .76 | .06 | .00 | .00 | .18 |
| eat | .01 | .88 | .07 | .02 | .01 |
| explore | .04 | .02 | .74 | .14 | .05 |
| groom | .09 | .00 | .34 | .55 | .02 |
| sleep | .02 | .00 | .09 | .00 | .89 |

# mouse behavior
## pilot study



# human activity

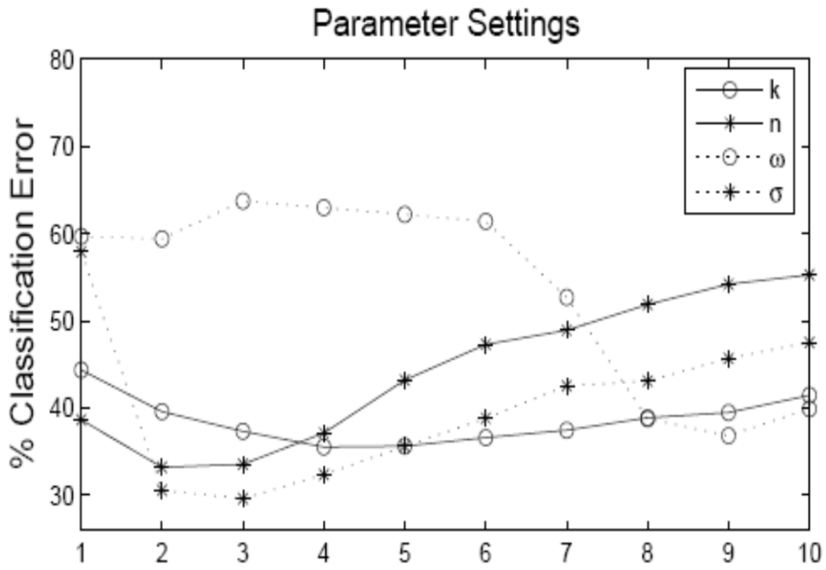
1−NN

| | walking | jogging | running | boxing | handclapping | handwaving |
|---|---|---|---|---|---|---|
| walking | .89 | .10 | .00 | .00 | .01 | .02 |
| jogging | .25 | .63 | .12 | .00 | .00 | .00 |
| running | .05 | .23 | .73 | .00 | .00 | .00 |
| boxing | .00 | .00 | .00 | .80 | .15 | .05 |
| handclapping | .00 | .00 | .00 | .09 | .82 | .09 |
| handwaving | .00 | .00 | .00 | .06 | .10 | .84 |

SVM

| | walking | jogging | running | boxing | handclapping | handwaving |
|---|---|---|---|---|---|---|
| walking | .90 | .04 | .05 | .01 | .01 | .00 |
| jogging | .20 | .57 | .24 | .00 | .00 | .00 |
| running | .03 | .13 | .85 | .00 | .00 | .00 |
| boxing | .00 | .00 | .00 | .93 | .06 | .01 |
| handclapping | .00 | .00 | .01 | .22 | .77 | .01 |
| handwaving | .03 | .00 | .01 | .07 | .04 | .85 |

# parameter settings




- k, 50 < k < 500, number of clusters
- n, 10 ≤ n ≤ 200, number of cuboids per clip
- ω, 0 < ω < 1 overlap allowed between cuboids
- σ, 2 < σ < 9, spatial scale of the detector
- Base settings used were approximately:
  - k = 250, n = 30, ω = .9, and σ = 2

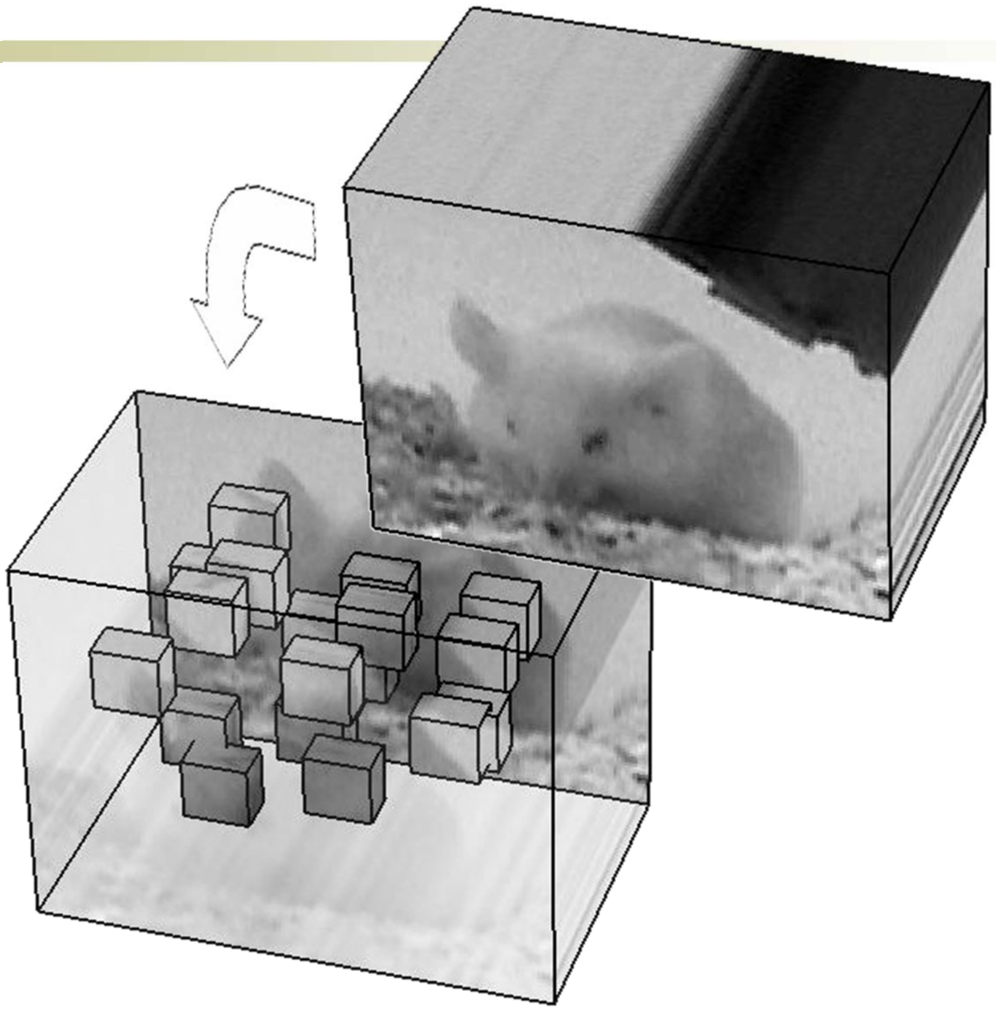
# summary of results

- achieved good performance in all domains [typically 10-20% error]
- achieved best performance of algorithms tested in all domains
- comparison to domain specific algorithms necessary

# Current Work

- Niebles et al.
  - "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words" BMVC, 2006
    - Recognizes multiple activities in a single video sequence
    - Using the same interest point detector, cluster cuboids into a set of video codewords, then use pLSA graphical model to determine probability distributions.
    - 81.50% accuracy vs 81.17% for Dollár et al.
      - However, learning is unsupervised for Niebles et al.

# Questions?



Acknowledgements: Piotr Dollar