Action Recognition & Categories via Spatial-Temporal Features

华俊豪, 11331007

huajh7@gmail.com

2014/4/9

Talk at “Image & Video Analysis” taught by Huimin Yu.

Outline

- Introduction
- Frameworks
  - Feature extraction and description
  - Action Classification
- Multiple actions categories
- Experiments
- Reference

Introduction

- Motivation: How detection and recognition behavior from video sequences ?





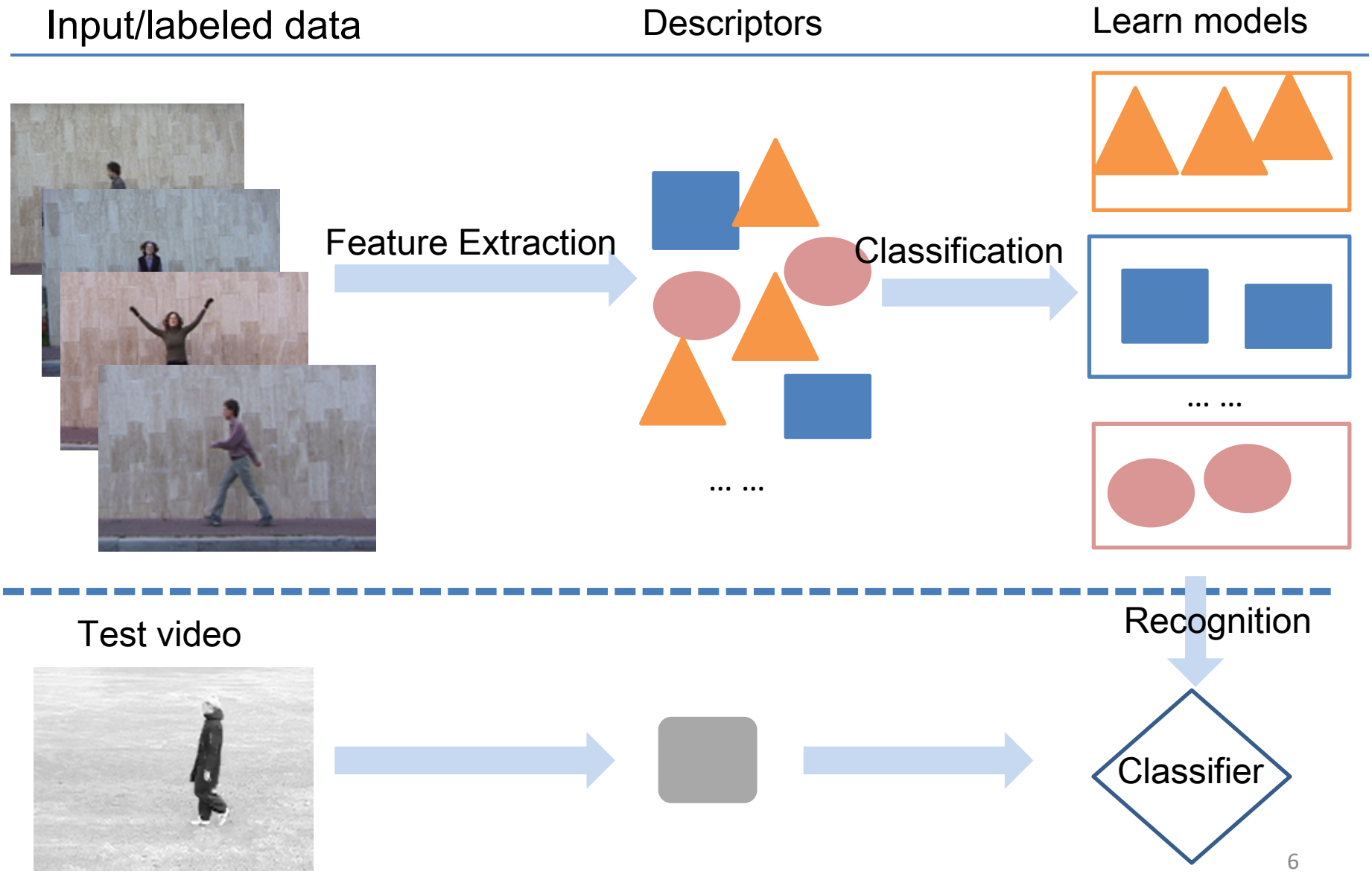

Introduction

- Motivation: How recognize and localize multiple actions in long and complex video sequences containing multiple motions ?

Outline

- Introduction
- Frameworks
  - Feature extraction and description
  - Action Classification
- Multiple actions categories
- Experiments
- Reference

Frameworks


Input/labeled data
Descriptors
Learn models
Feature Extraction
Classification
... ...
... ...
Test video
Recognition
Classifier

Feature Extraction

- Interest point detection
  - Response function:

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

where $g(x, y; \sigma)$ is the 2D Gaussian smoothing kernel. And h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally.

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

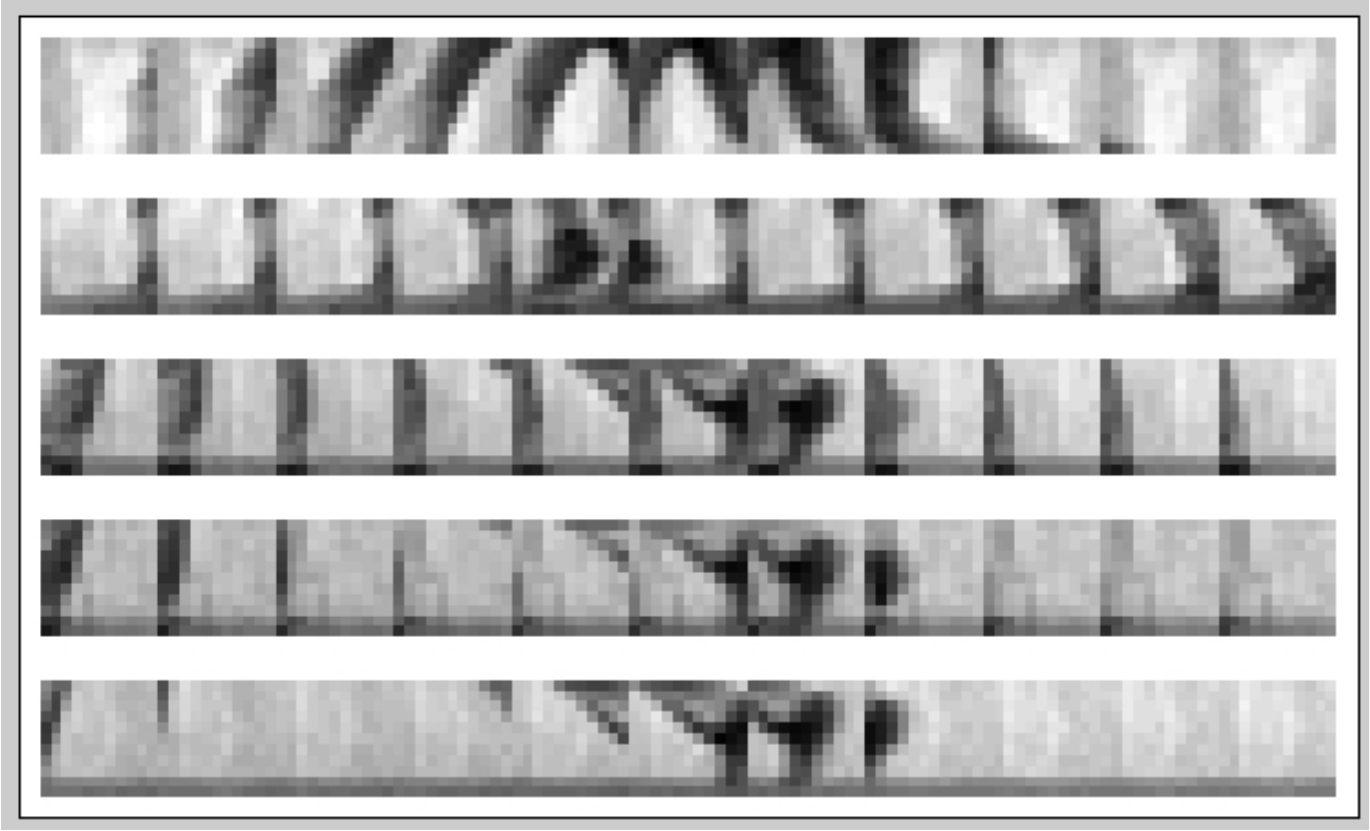
Using $\omega = 4/\tau$. the tuning parameters σ and τ are corresponding roughly to the spatial and temporal scale of the detector.

- A Cuboid
  - A cuboid is extracted which contains the spatio-temporally windowed pixel values.
  - Size(x, y, t) : $(2\lceil 3\sigma \rceil + 1, 2\lceil 3\sigma \rceil + 1, 2\lceil 3\tau \rceil + 1)$

Interest Points visualization

Type: Walk
\sigma = 1.5
\tau = 1.5

The Cuboid visualization

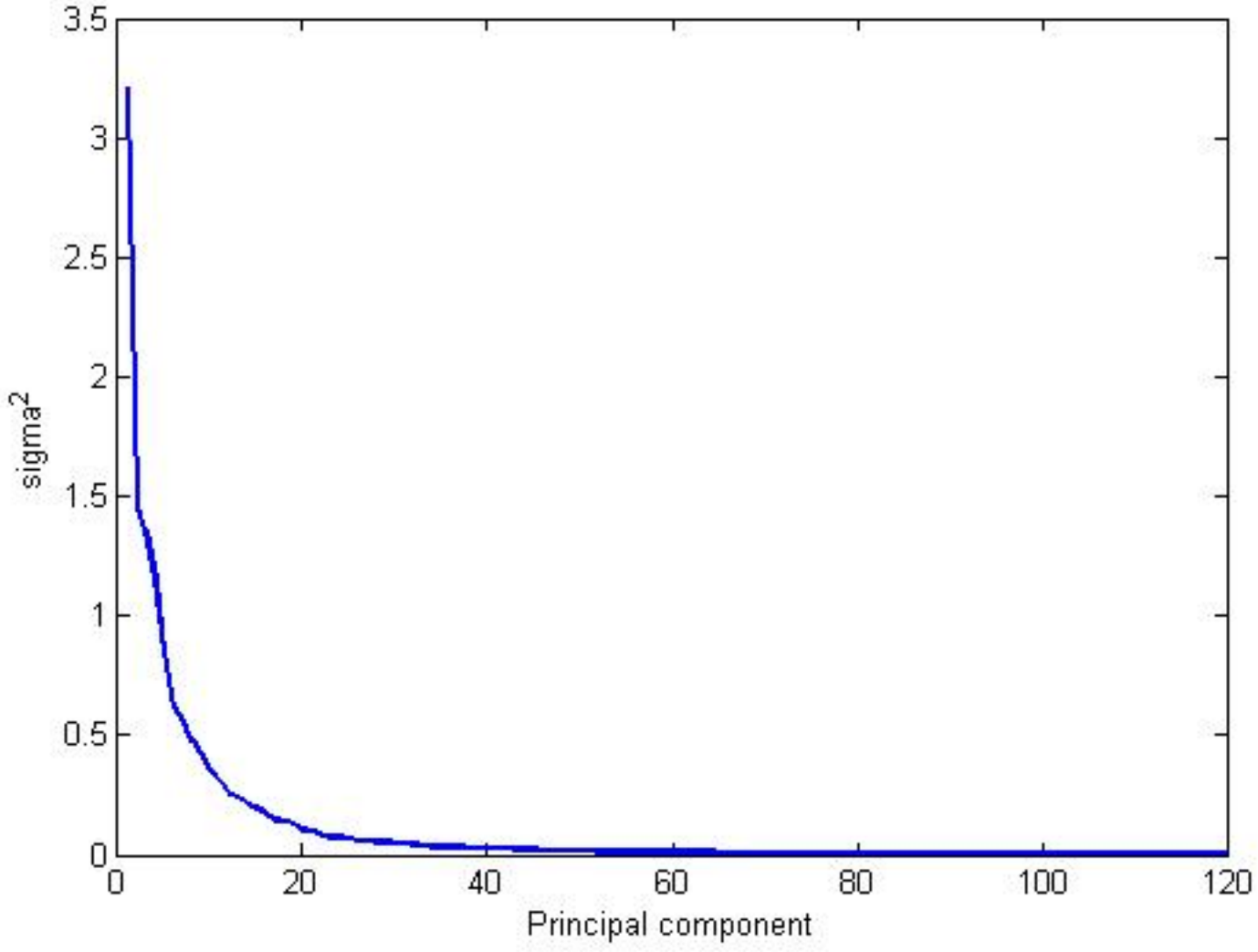


Type: Walk
\sigma = 1.5
\tau = 1.5

Feature Extraction(cont.)

- Cuboid to a feature vector
  - Gaussian smoothing (at different scale, \sigma)
  - Calculate the brightness gradient (G_x,G_y,G_t) at each spatio-temporal location (x, y, t)
  - string out vector
- Dimensionality reduction: PCA
  - Random subsample cuboids from all data
  - Calculate the first k principal components by PCA. (setting k = 100)
  - Project the centralized feature vectors into the new basis. (N x k)
  - (PCA-SIFT descriptor)

Principal components are descending sorted by the corresponding variances.

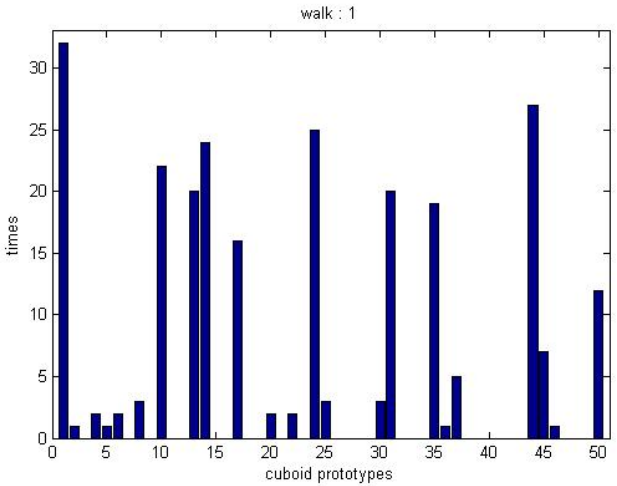


Feature Description

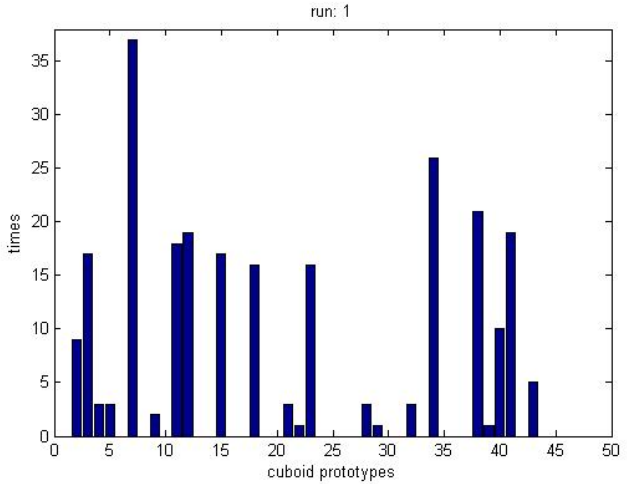
- Cuboid Prototype
  - Create a library of cuboid prototype by clustering the cuboids descriptors.
  - Using K-means and Euclidean distance, setting k=50
- Behavior Descriptor
  - Use a histogram of the cuboid types as the behavior descriptor.
  - Distance between histograms: chi-squared distance

$$d_{\chi^2}(x, y) = \sum_i \frac{(x_i - y_i)^2}{2(x_i + y_i)}$$

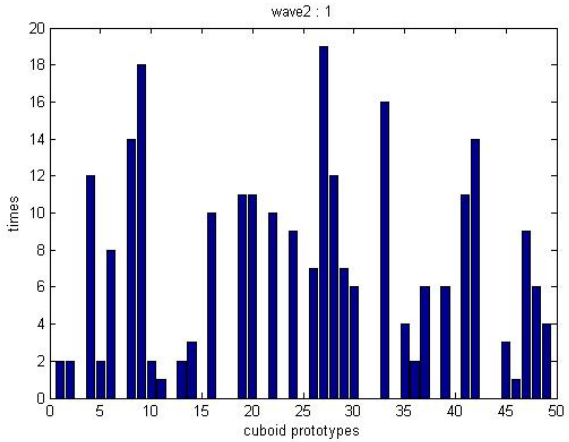
Histogram Descriptor Samples:



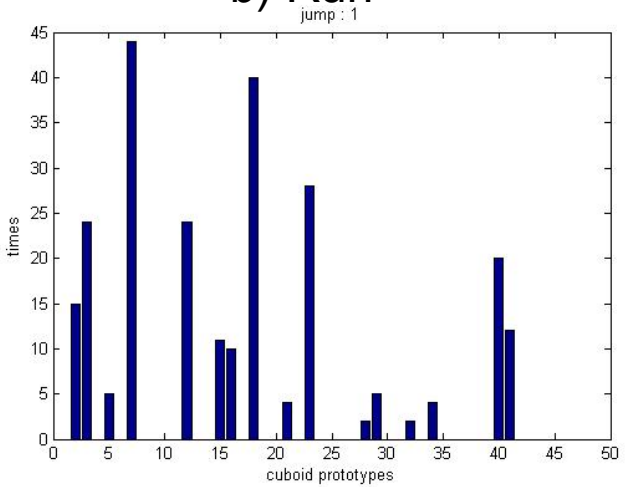
a) Walk



b) Run



c) wave2



d) jump

Outline

- Introduction
- Frameworks
  - Feature extraction and description
  - Action Classification
- Multiple actions categories
- Experiments
- Reference

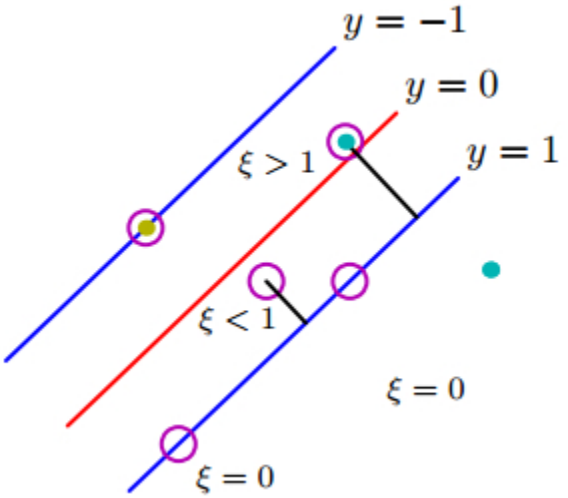
Classification

- Simplest method: 1-NN
  - Find nearest neighbors based on chi-squared distance.
- Support Vector Machine
  - Use Norm2 SVM (least square SVM)
  - Use chi-squared Kernel

$$K_{\chi^2-RBF}(x, x') = \exp(-\rho(1 - d_{\chi^2}(x, x')))$$

  - Multiclass SVM: one-versus-the-rest
    - appropriate scale
    - Imbalance: Positive class: +1; negative class: -1/(K-1)

A Brief Intro. to SVM



The two-class classification problem: minimize

$$C\sum_{n=1}^{N}\xi_n + \frac{1}{2}\|w\|^2$$

subject to

$$t_n(w^T\phi(x)+b) \geq 1-\xi_n, n = 1\ldots,N$$

where $t_n \in \{-1,1\}$, $\xi_n \geq 0$ are slack variables. Using Lagrange multipliers, we obtain the dual Lagrangian,

$$\tilde{L}(a) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M} a_n a_m t_n t_m k(x_n, x_m)$$

subject to $0 \leq a_n \leq C$ (box constraints)and $\sum_{n=1}^{N} a_n t_n = 0$

Optimization methods:

1) Quadratic Programming(QP): Chunking(Vapnik, 1982, Burges,1998)
2) Sequential minimal optimization, or SMO (Platt,1999)

A Brief Intro. to LS-SVM

The classification problem becomes minimizing

$$C\sum_{n=1}^{N} e_n^2 + \frac{1}{2}\|w\|^2$$

- Lost the sparseness

Define the Lagrangian, and Lagrangian multipliers α_n can be either positive or negative due to KKT conditions. The solution now can be written as the set of linear equation,

$$\left[\begin{array}{ccc|c} I & 0 & 0 & -Z^T \\ 0 & 0 & 0 & -Y^T \\ 0 & 0 & \gamma I & -I \\ \hline Z & Y & I & 0 \end{array}\right] \left[\begin{array}{c} w \\ b \\ e \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ \hline \vec{1} \end{array}\right] \qquad (19)$$

where $Z = [\varphi(x_1)^T y_1; \ldots; \varphi(x_N)^T y_N]$, $Y = [y_1; \ldots; y_N]$, $\vec{1} = [1; \ldots; 1]$, $e = [e_1; \ldots; e_N]$, $\alpha = [\alpha_1; \ldots; \alpha_N]$. The solution is also given by

$$\left[\begin{array}{c|c} 0 & -Y^T \\ \hline Y & ZZ^T + \gamma^{-1} I \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \vec{1} \end{array}\right]. \qquad (20)$$

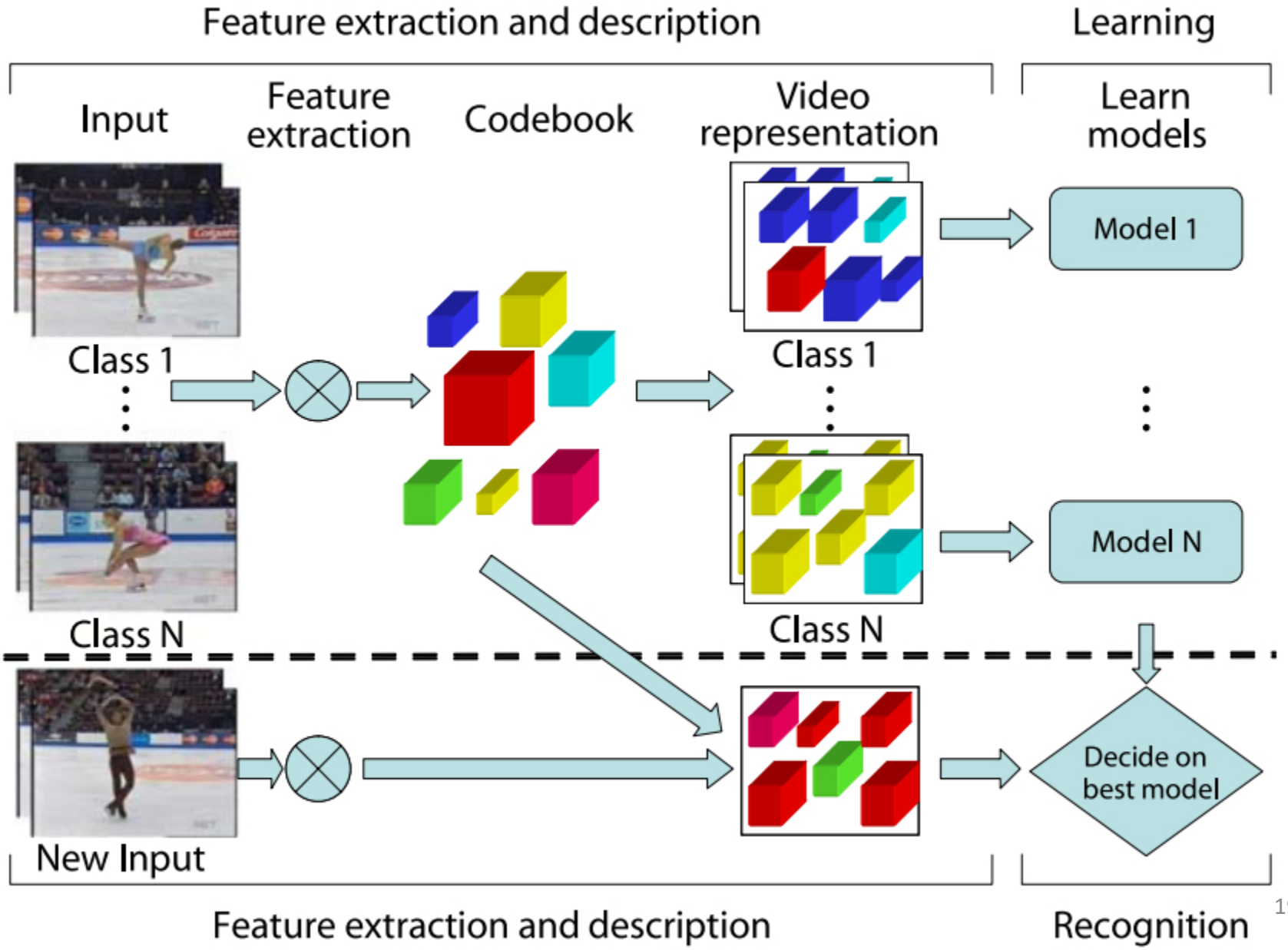
Mercer's condition can be applied again to the matrix $\Omega = ZZ^T$ where

$$\begin{aligned} \Omega_{kl} &= y_k y_l \, \varphi(x_k)^T \varphi(x_l) \\ &= y_k y_l \, \Psi(x_k, x_l). \end{aligned} \qquad (21)$$

Outline

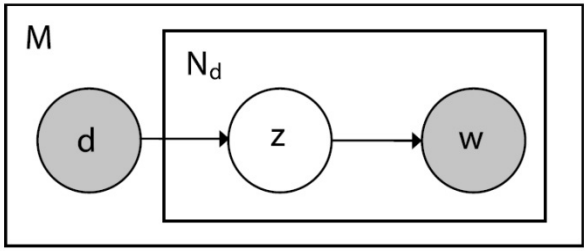
- Introduction
- Frameworks
  - Feature extraction and description
  - Action Classification
- Multiple actions categories
- Experiments
- Reference

Multi-actions categories Framework



Multi-actions categories scheme

- Feature Extraction: Cuboid descriptor(PCA-SIFT)
- Build vocabulary(codebook):
  - by clustering using k-means and Euclidean distance, k = 1000
- Label each codeword (w_i) with a topic/category (z_k)
  - Assign each interest point (training) to a codeword
  - By maximizing the posterior P(z_k | w_i)
- Assign detected interest points a topic
  - By assign each point to a codeword
- Topic Localization
  - how many action categories are significantly in a single sequence?
  - k-means to the spatial position of space-times patches.
  - Each word votes for its assigned action within its cluster.



- Topic Localization(cont.)
  - how many action categories are significantly in a single sequence?
  - Tuning parameters.

```
% how many action categories ?
categ_idx = [];
k = 0;
for j=1:ntype
    if type_prob(j)>0.5 ||...
        (type_prob(j)>0.4 && type_cnts(j) > 5) ||...
        (type_prob(j)>0.3 && type_cnts(j) > 14) ||...
        (type_prob(j)>0.2 && type_cnts(j) > 13) ||...
        (type_prob(j)>0.1 && type_cnts(j) > 15)
        categ_idx = [categ_idx,j];
        k = k + 1;
    end
end
```

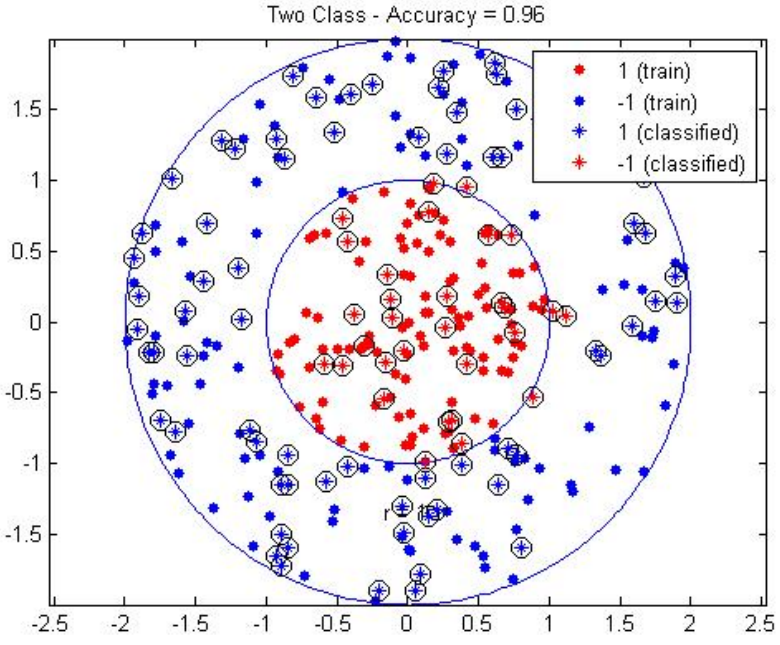
- We call this method Voting.

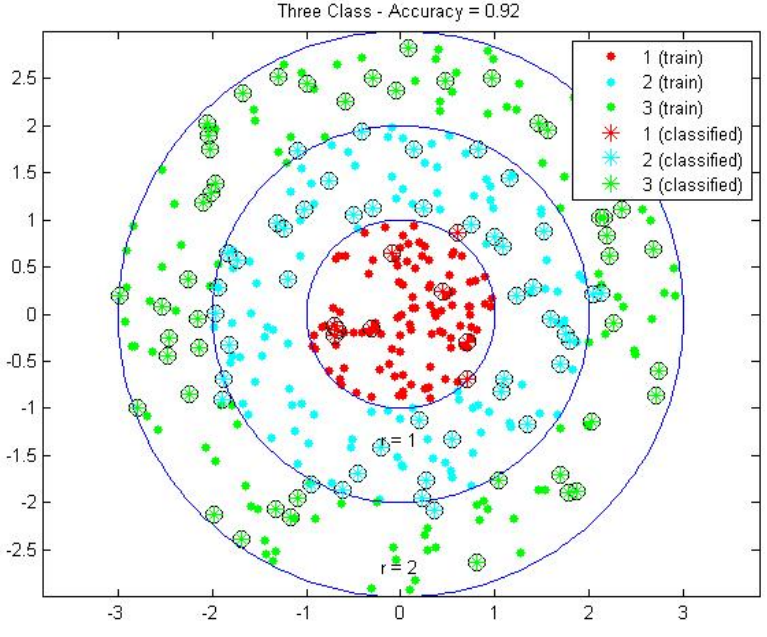
Outline

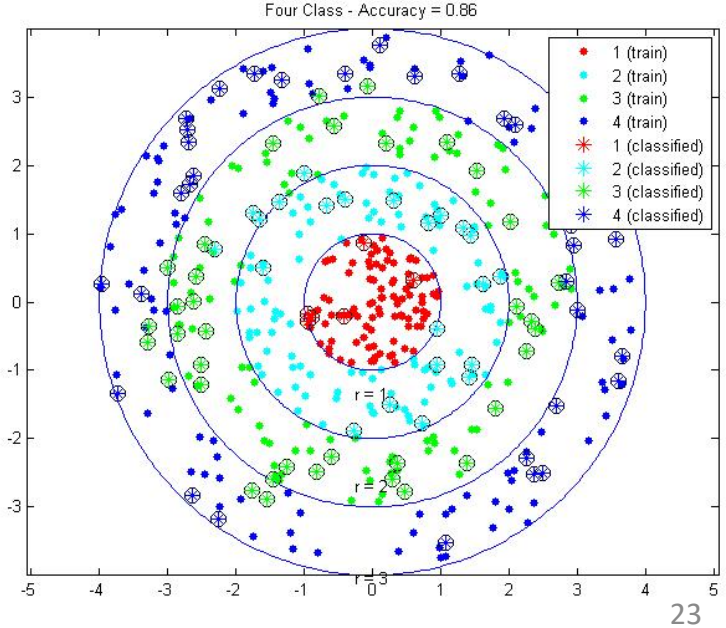
- Introduction
- Frameworks
  - Feature extraction and description
  - Action Classification
- Multiple actions categories
- Experiments
- Reference

SVM Classifier on Toy dataset

Accuracy:
2 class: 0.9600
3 class: 0.9200
4 class: 0.8600







Experiment 1: Weizmann dataset

- Dataset Setup
  - Weizmann human action dataset [1]
    - Contains 90 low-resolution (180x144,50fps) video sequences showing 9 different people, each performing 10 actions (bend, jack, jump, pjump, run, side, skip, walk, wave1,wave2)


wave1


walk


wave2


side


jack


jump


run


bend

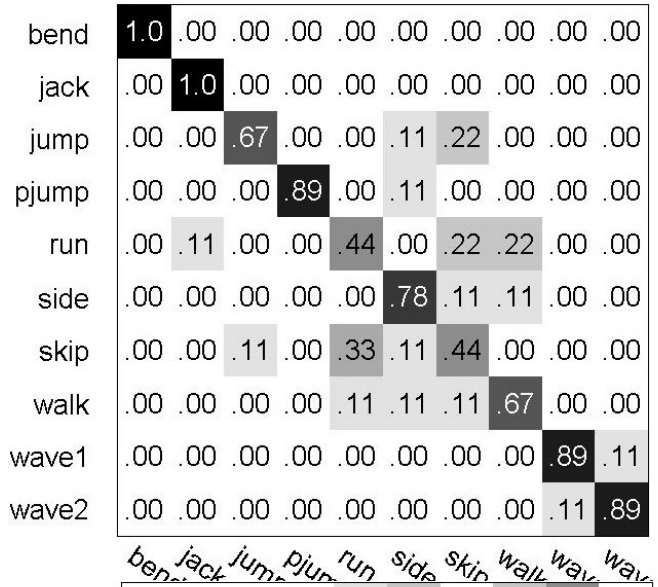
[1] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

Result: Weizmann by 1NN

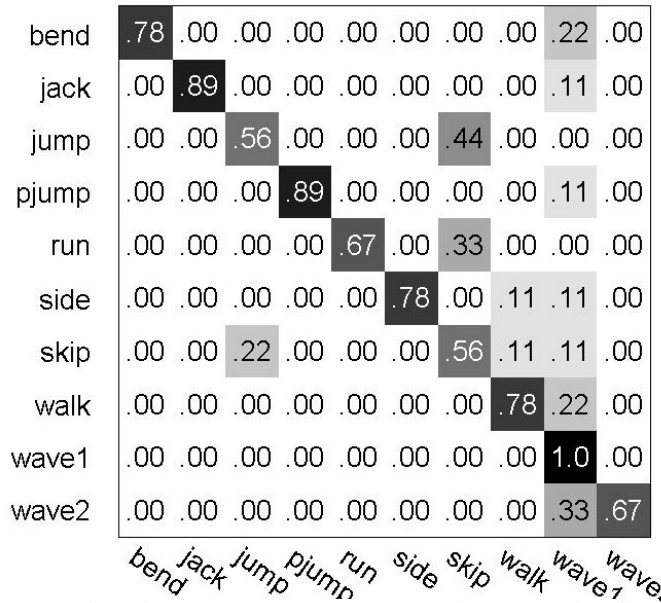
- \sigma = 1.2, \tau = 1.2,
- maxn(interest points) = 200
- Kpca = 100
- Divide into 9 subset
- Leave-one-out test
- Repeat: 5 times
- Overall accuracy: 0.7933

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00
jack	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00
jump	.00	.00	.71	.00	.04	.00	.24	.00	.00	.00
pjump	.00	.00	.00	1.0	.00	.00	.00	.00	.00	.00
run	.00	.00	.00	.00	.67	.02	.31	.00	.00	.00
side	.00	.00	.02	.00	.04	.87	.02	.04	.00	.00
skip	.00	.00	.33	.00	.29	.00	.29	.09	.00	.00
walk	.00	.00	.04	.00	.00	.04	.11	.80	.00	.00
wave1	.00	.00	.00	.00	.00	.00	.00	.00	.87	.13
wave2	.02	.02	.00	.00	.00	.00	.00	.00	.22	.73

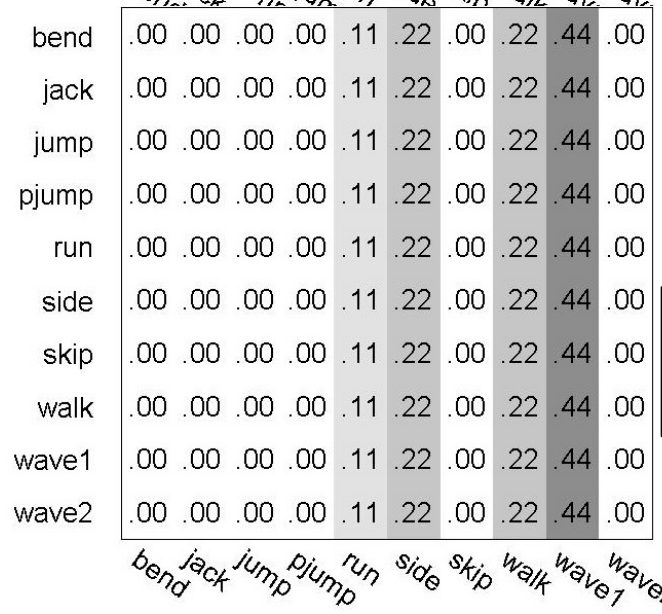
Result: Weizmann by LS-SVM



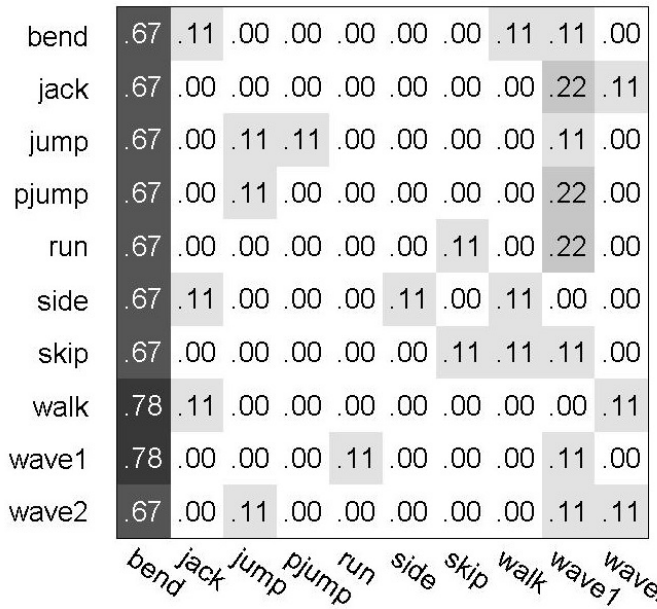
Linear:
0.7467



Polynomial:
0.7556



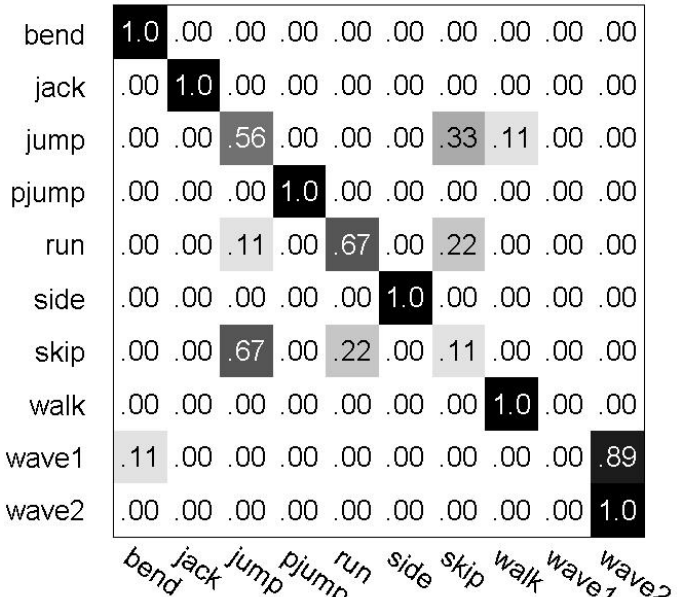
RBF:
0.1000



Chi-squared:
0.1022

Result: Weizmann by Voting

- \sigma = 1.2, \tau = 1.2, maxn(interest points) = 200
- Kpca = 100
- ncodeword = 1000
- Divide into 9 subset
- Leave-one-out test
- Repeat: 5 times
- Overall accuracy: 0.7333



Experiments 2: KTH dataset

- Dataset setup:
  - KTH human action dataset [2]
    - Contains 6 types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) preformed by 25 difference people in 4 different scenarios of outdoor and indoor environment with scale change.
    - Total 598 (160x140) video sequences

boxing

handclapping

jogging

handwaving

running

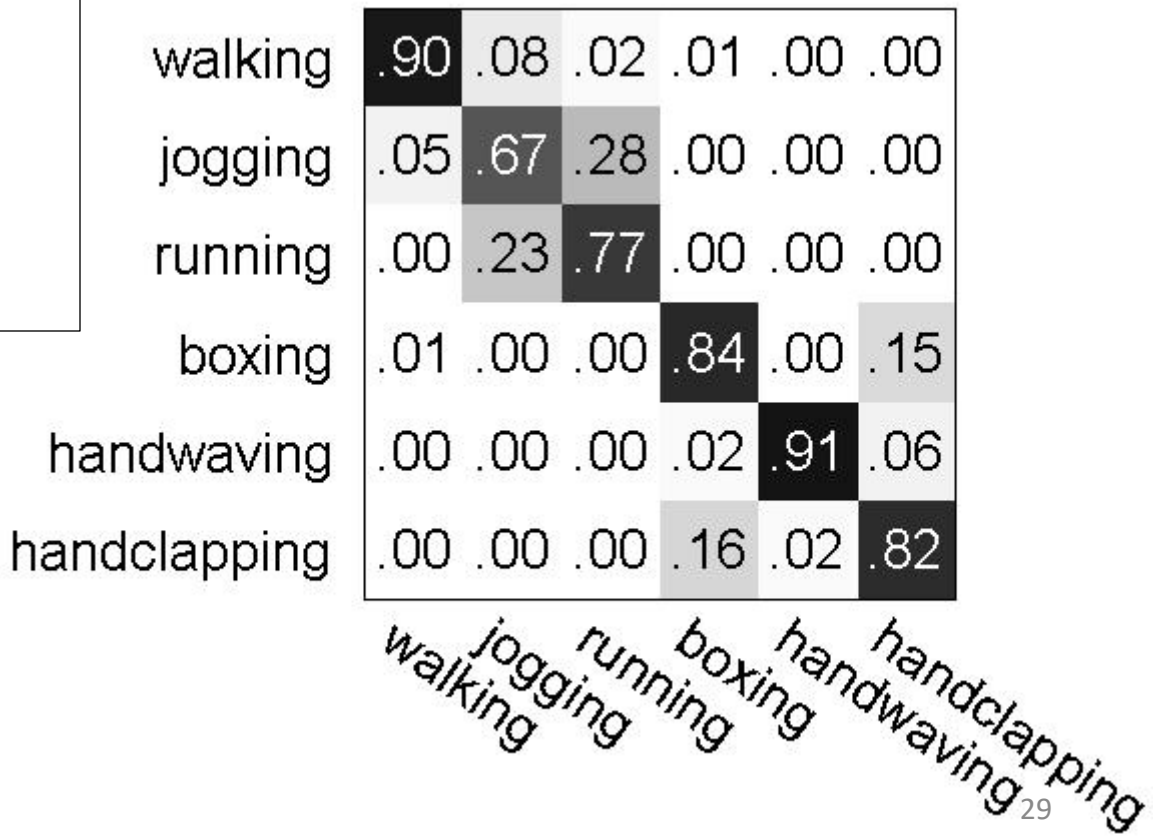
walking

Great thanks to xcy for downloading the dataset.

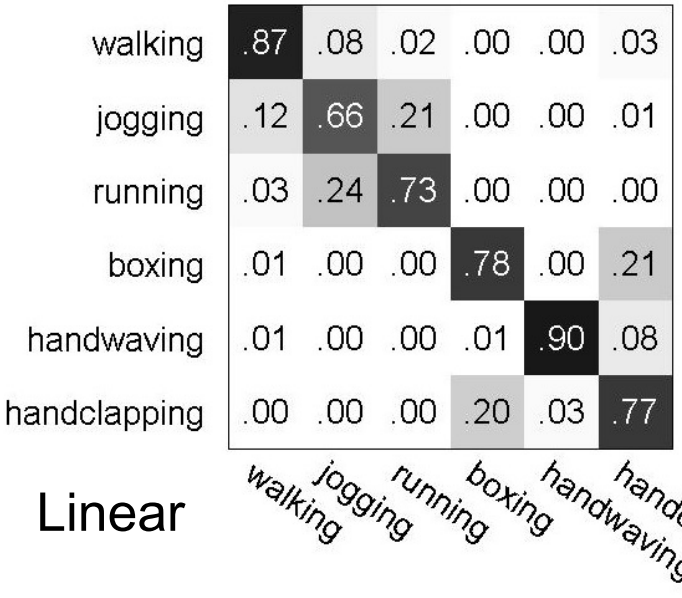
[2] http://www.nada.kth.se/cvap/actions/

Result: KTH dataset by 1NN

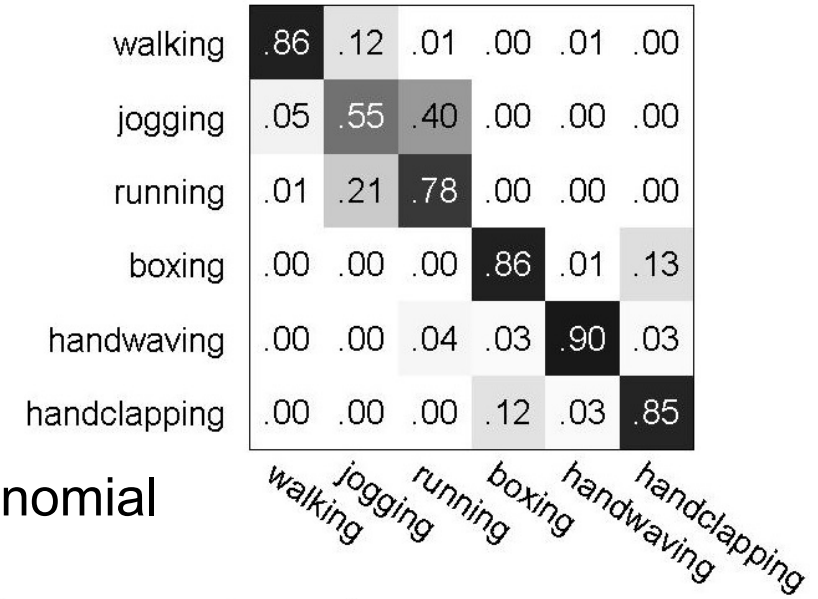
- \sigma = 2, \tau = 2.5,
- maxn(interest points) = 300
- Kpca = 200
- Divide into 25 subsets
- Leave-one-out test
- Repeat: 5 times
- Overall accuracy: 0.8179



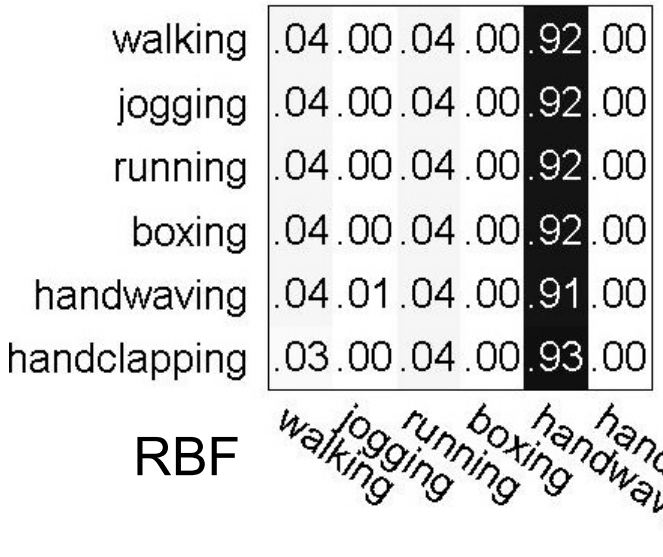
Result: KTH dataset by LS-SVM



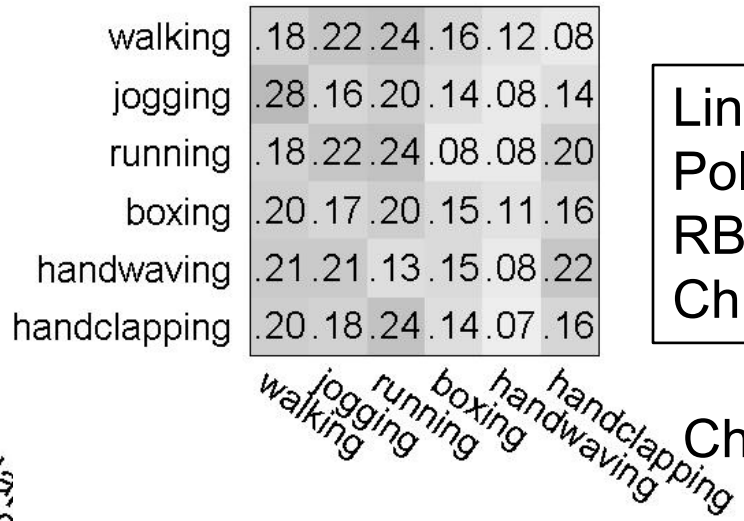
Linear



Polynomial



RBF

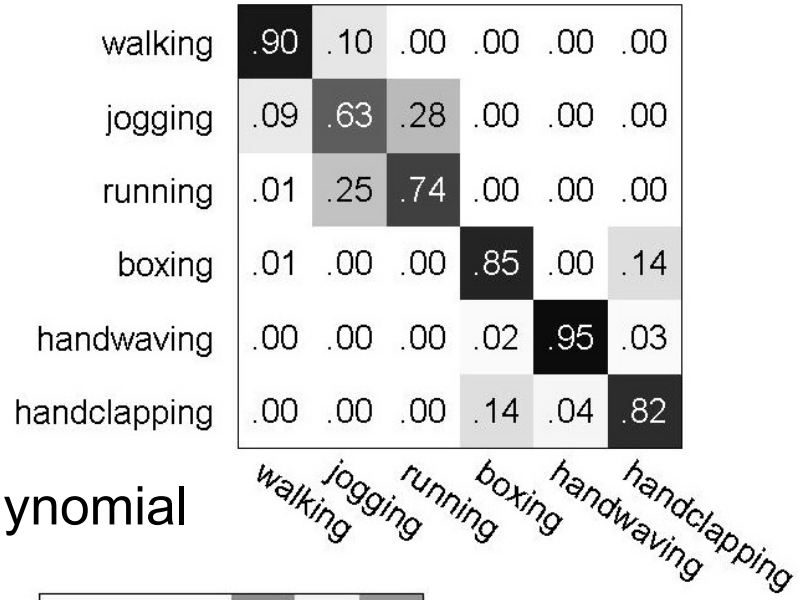


Chi-squared

Linear: 0.7844
Polynomial: **0.7996**
RBF: 0.1656
Chi-squared: 0.1624

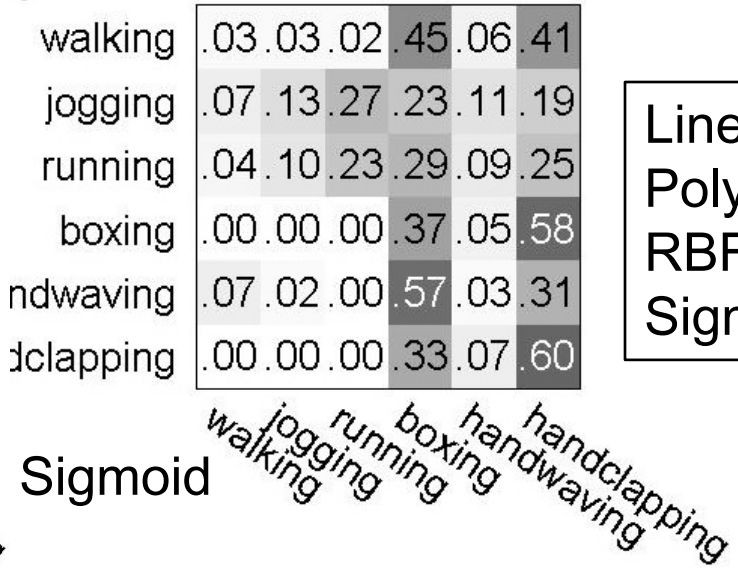
Result: KTH dataset by Norm1 SVM

Linear



Polynomial

RBF

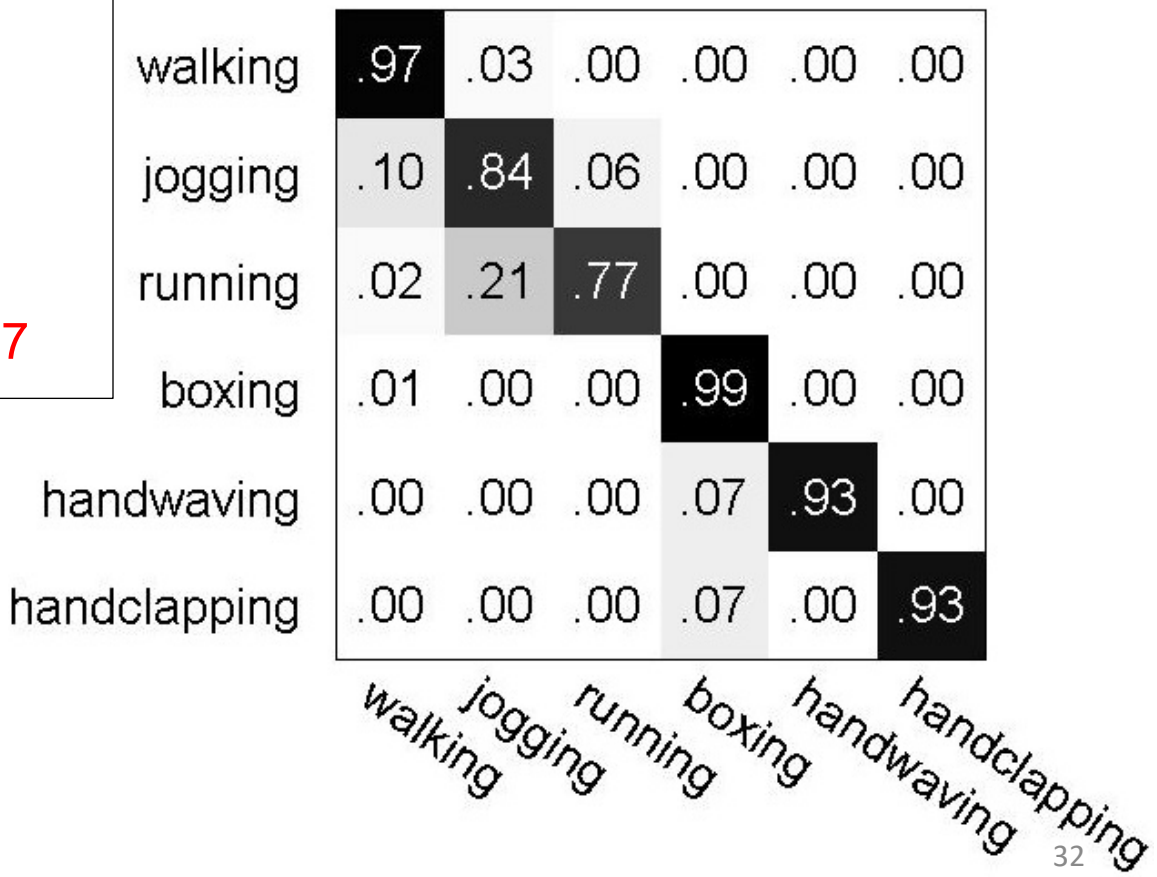


Sigmoid

Linear: 0.7429
Polynomial: **0.8146**
RBF: 0.1873
Sigmoid: 0.2309

Result: KTH dataset by Voting

- \sigma = 2, \tau = 2.5, maxn(interest points) = 300
- Kpca = 200
- Ncodeword = 1000
- Divide into 25 subset
- Leave-one-out test
- Repeat: 5 times
- Overall accuracy: 0.9047



Experiments 3: Multi-Actions

- Thanks to Lin**, liu*, wang**




Result by Voting Scheme

- Using the Weizmann dataset for training(\sigma = 1.2, \tau = 1.2)




Result by Voting Scheme(cont.)

- Using the KTH dataset for training(\sigma = 2, \tau = 2.5)




Reference

- Dollár, Piotr, et al. "Behavior recognition via sparse spatio-temporal features." *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005.
- Niebles, Juan Carlos, Hongcheng Wang, and Li Fei-Fei. "Unsupervised learning of human action categories using spatial-temporal words."*International journal of computer vision* 79.3 (2008): 299-318.
- Chapelle, Olivier, Patrick Haffner, and Vladimir N. Vapnik. "Support vector machines for histogram-based image classification." *Neural Networks, IEEE Transactions on* 10.5 (1999): 1055-1064.
- Schuldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach." *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Vol. 3. IEEE, 2004.
- Van Gestel, Tony, et al. *Least squares support vector machines*. Vol. 4. Singapore: World Scientific, 2002.

Any Question ?