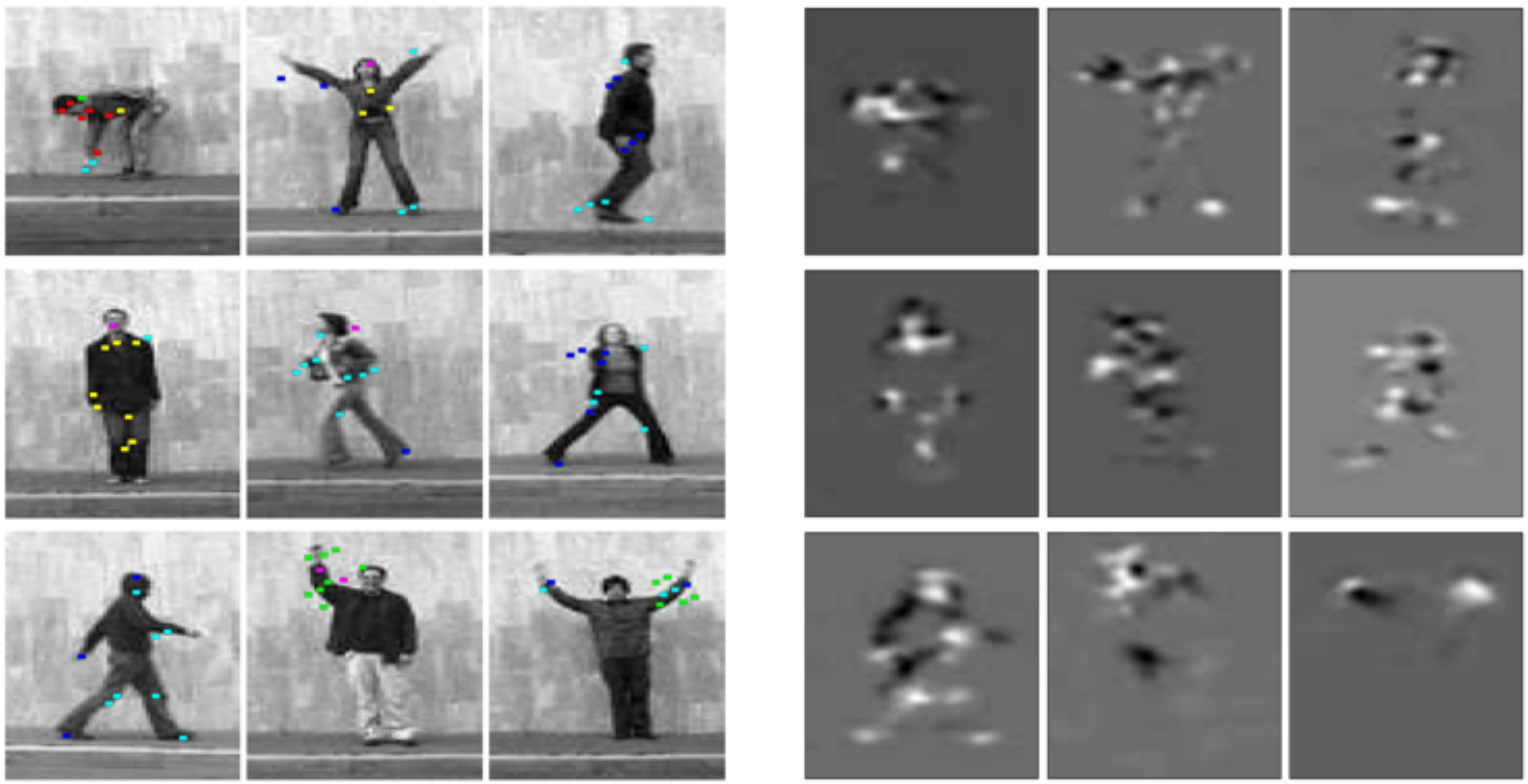
Hidden Part Models for Human Action Recognition: Probabilistic vs. Max-Margin



Yang Wang & Greg Mori (PAMI2011)

Colin Lea - Summer 2014

Data

Goal:
<u>Recognize</u> an action in a short video with a single actor

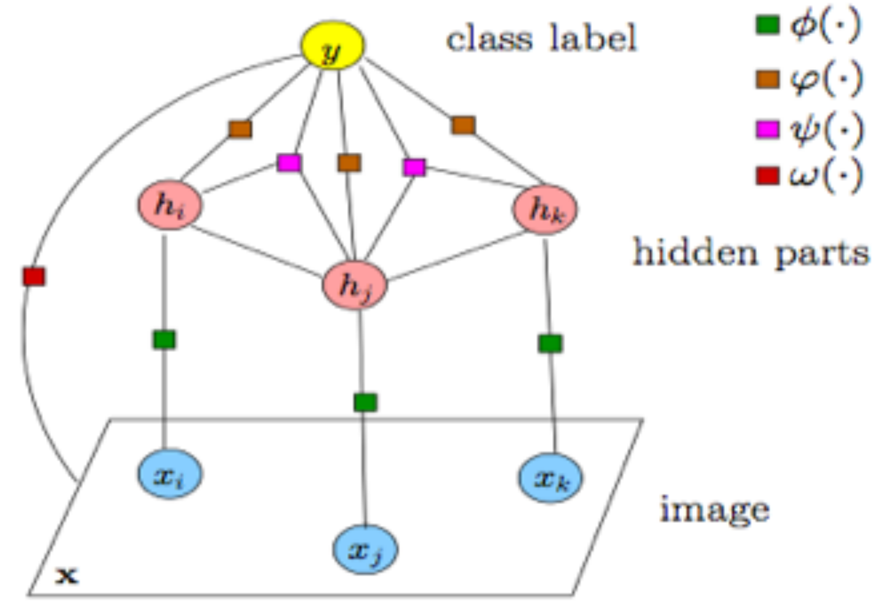
Weizmann dataset
10 simple actions
Run, walk, skip, jump, skip gallup, bend, wave1, wave2, jumping jacks

KTH Dataset
6 simple actions:
Walk, Jog, Run, Box, Wave, Clap

Overview

Goal:
Recognize an action in a short video with a single actor

Approach:
- Inspired by part-based models for humans
  - Use constellation of parts conditioned on image seq.
- Hidden CRF for parts model
- Global + local features



Contributions:
- Novel part-based approach
- Compare probabilistic vs max-margin approach

Model

phi: (unary) concatenation of appearance+spatial features

$$\alpha^\top \cdot \phi(x_j, h_j) = \sum_{c\in\mathcal{H}} \alpha_c^\top \cdot \mathbb{1}_{\{h_j=c\}} \cdot [f^a(x_j)\ f^s(x_j)]$$

phi2: (unary) likelihood of 1 part label & class

$$\beta^\top \cdot \varphi(y, h_j) = \sum_{a\in\mathcal{Y}} \sum_{b\in\mathcal{H}} \beta_{a,b} \cdot \mathbb{1}_{\{y=a\}} \cdot \mathbb{1}_{\{h_j=b\}}$$

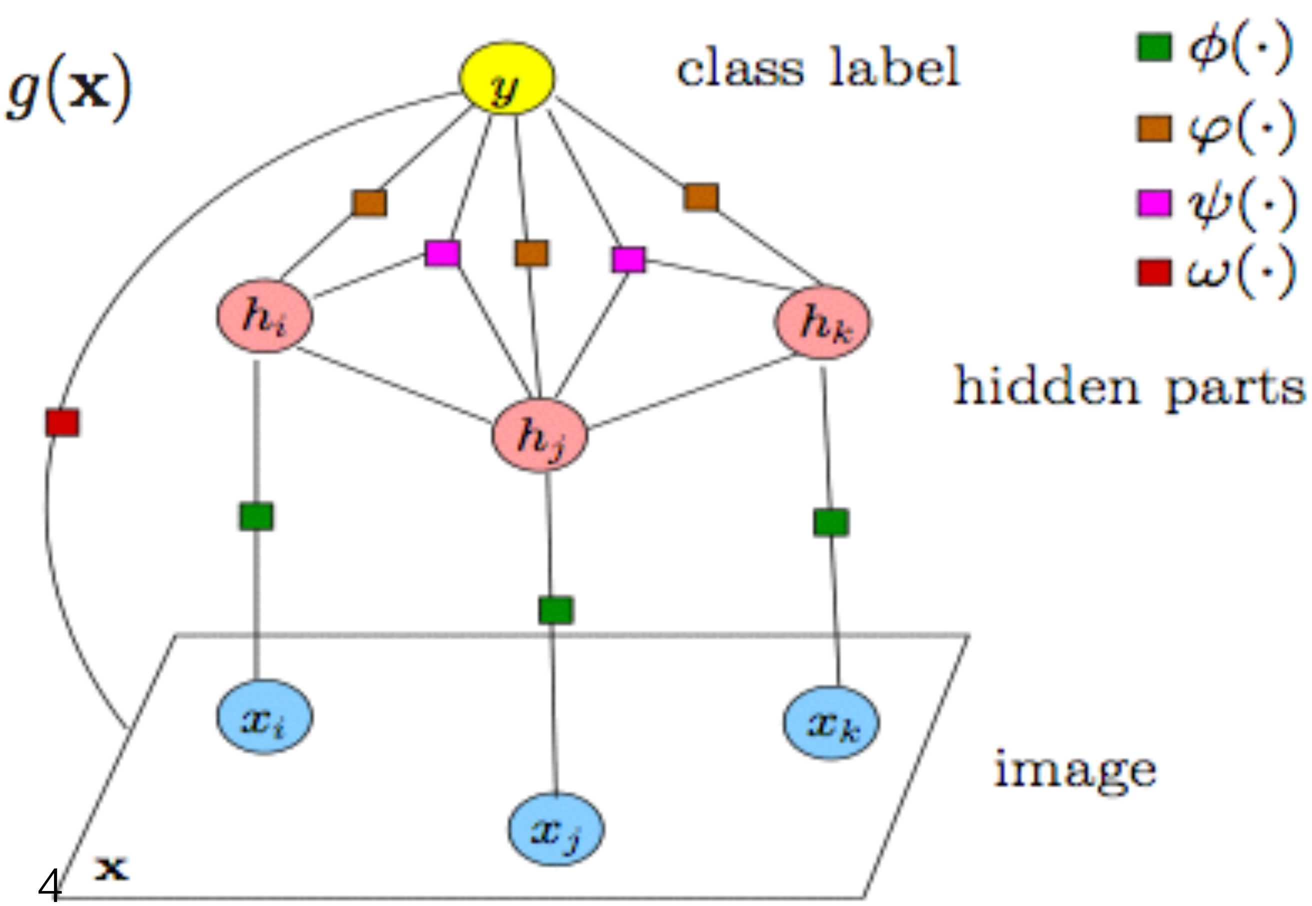
psi: (pairwise) likelihood of 2+ part labels & class

$$\gamma^\top \cdot \psi(y, h_j, h_k) = \sum_{a\in\mathcal{Y}} \sum_{b\in\mathcal{H}} \sum_{c\in\mathcal{H}} \gamma_{a,b,c} \cdot \mathbb{1}_{\{y=a\}} \cdot \mathbb{1}_{\{h_j=b\}} \cdot \mathbb{1}_{\{h_k=c\}}$$

...

omega: (unary) root filter

$$\eta^\top \cdot \omega(y, \mathbf{x}) = \sum_{a\in\mathcal{Y}} \eta_a^\top \cdot \mathbb{1}_{\{y=a\}} \cdot g(\mathbf{x})$$



y: class label
h: part label
x: image feature

Model

p(label | data, params)

$$p(y|\mathbf{x};\theta) = \sum_{\mathbf{h}\in\mathcal{H}^{\mathbf{m}}} p(y,\mathbf{h}|\mathbf{x};\theta)$$

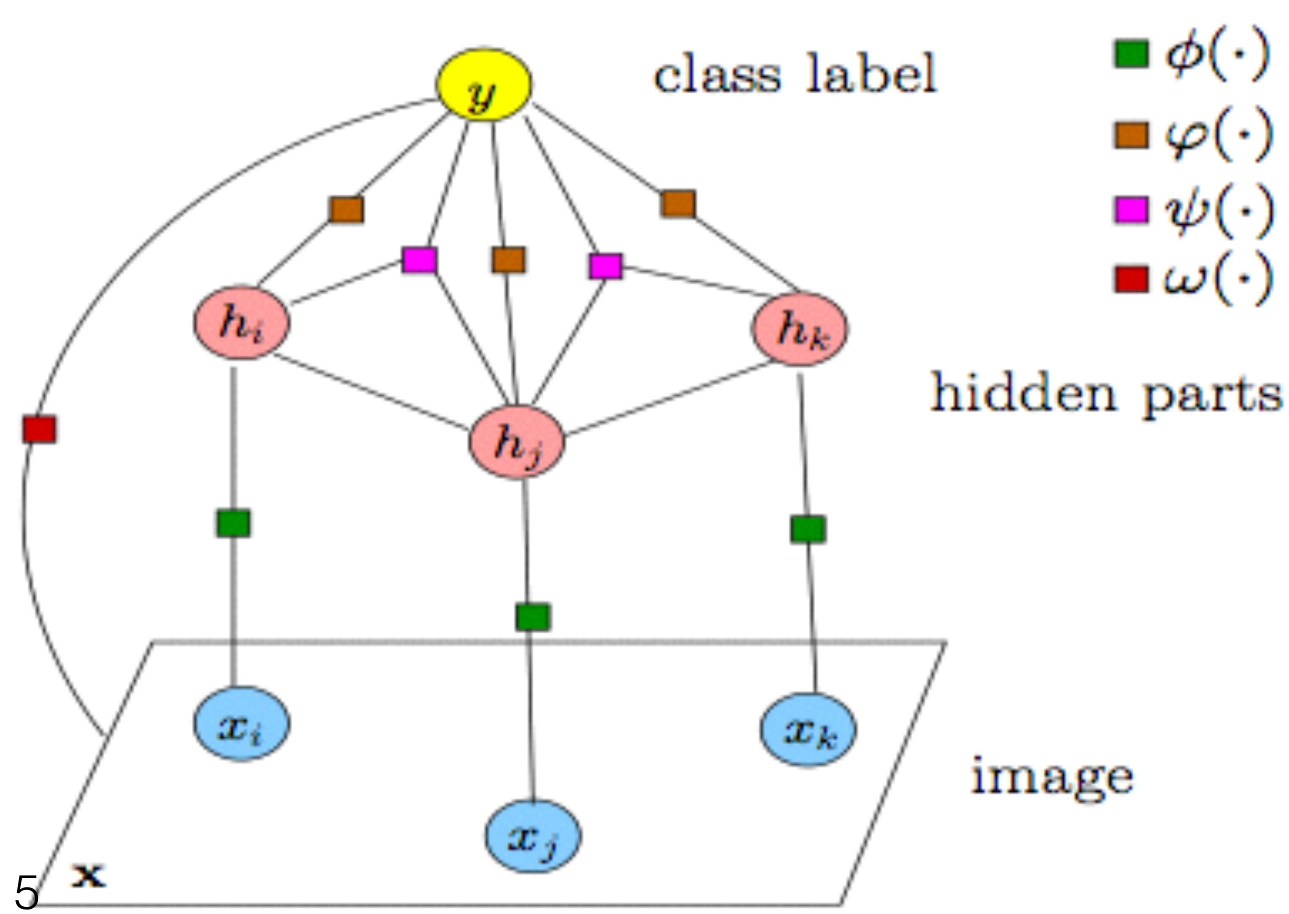
Marginalize over **hidden variables**

$$= \frac{\sum_{\mathbf{h}\in\mathcal{H}^m} \exp(\theta^\top \cdot \Phi(\mathbf{x},\mathbf{h},y))}{\sum_{\hat{y}\in\mathcal{Y}} \sum_{\mathbf{h}\in\mathcal{H}^m} \exp(\theta^\top \cdot \Phi(\mathbf{x},\mathbf{h},\hat{y}))}$$

- Cluster features into **hidden parts h**.
- Action ∝ **action-cluster** co-occurrences & **features**
- Ignore **psi** (pairwise). Small affect on model.

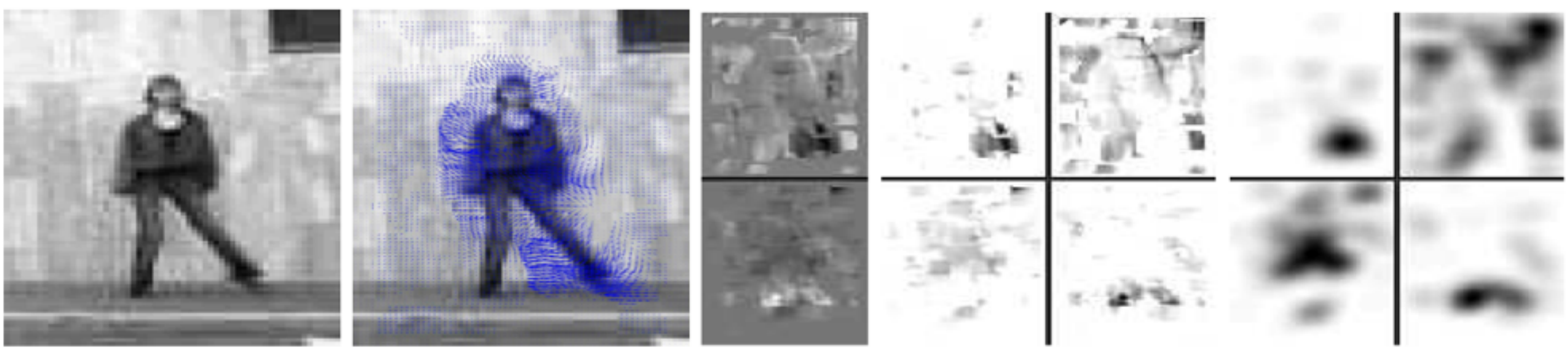
Features:

$$\theta^\top \cdot \Phi(\mathbf{h},\mathbf{x},y) = \sum_{j\in\mathcal{V}} \alpha^\top \cdot \phi(x_j,h_j) + \sum_{j\in\mathcal{V}} \beta^\top \cdot \varphi(y,h_j) + \sum_{(j,k)\in\mathcal{E}} \gamma^\top \cdot \psi(y,h_j,h_k) + \eta^\top \cdot \omega(y,\mathbf{x}) \quad (1)$$



Features

- Motion features: optical flow
  - Lucas Kanade used to track person
  - Assume person front/center of image
  - $F b^{+}x$,$F b^{-}x$,$F b^{+}y$,$F b^{-}y$ (half wave rectified+blur)
- Spatial features:
  - Bit-vector defining relative location of image patches
  - (Vector of length L (=#bins) with 0/1 in each)



Original — Optical Flow — Fx Fy — +half-wave rectified — +blur

Hidden parts *use image patches.* ***Root*** *uses whole image*

Other details

- Initialize filters/parameters
  - Root filter: (compute over whole image)
    - (omega = feature vector)

$$
\begin{aligned}
\eta^* &= \arg\max_{\eta} \sum_{t=1}^{N} \log \mathcal{L}^{root}(y^{(t)}|\mathbf{x}^{(t)};\eta) \\
&= \arg\max_{\eta} \sum_{t=1}^{N} \log \frac{\exp\left(\eta^\top \cdot \omega(y^{(t)},\mathbf{x}^{(t)})\right)}{\sum_y \exp\left(\eta\top \cdot \omega(y,\mathbf{x}^{(t)})\right)}
\end{aligned}
$$

  - Hidden parts {6,10,20}: Find top patches from previous equation in training. For test, compute score for all hidden parts.

- Background subtraction is performed (from dataset)

Probabilistic Formulation

Maximize the conditional likelihood:

$$\theta^* = \arg\max_\theta \mathcal{L}(\theta) = \arg\max_\theta \sum_{t=1}^{N} \mathcal{L}^t(\theta)$$

$$= \arg\max_\theta \sum_{t=1}^{N} \log p(y^{(t)}|\mathbf{x}^{(t)};\theta)$$

$$= \arg\max_\theta \sum_{t=1}^{N} \log \left(\sum_{\mathbf{h}} p(y^{(t)},\mathbf{h}|\mathbf{x}^{(t)};\theta) \right)$$

N=training samples
theta=params

Marginalize over **hidden parts**

Gradient descent:

$$\frac{\partial \mathcal{L}^t(\theta)}{\partial \alpha} = \sum_{j\in\mathcal{V}} \Big[\mathbb{E}_{p(h_j|y^{(t)},\mathbf{x}^{(t)};\theta)}\phi(x_j^{(t)},h_j) - \mathbb{E}_{p(h_j,y|\mathbf{x}^{(t)};\theta)}\phi(x_j^{(t)},h_j)\Big]$$

Appearance

$$\frac{\partial \mathcal{L}^t(\theta)}{\partial \gamma} = \sum_{(j,k)\in\mathcal{E}} \Big[\mathbb{E}_{p(h_j,h_k|y^{(t)},\mathbf{x}^{(t)};\theta)}\psi(y^{(t)},h_j,h_k) - \mathbb{E}_{p(h_j,h_k,y|\mathbf{x}^{(t)};\theta)}\psi(y,h_j,h_k)\Big]$$

Hidden Pairwise

$$\frac{\partial \mathcal{L}^t(\theta)}{\partial \beta} = \sum_{j\in\mathcal{V}} \Big[\mathbb{E}_{p(h_j|y^{(t)},\mathbf{x}^{(t)};\theta)}\varphi(h_j,y^{(t)}) - \mathbb{E}_{p(h_j,y|\mathbf{x}^{(t)};\theta)}\varphi(h_j,y)\Big]$$

Hidden Unary

$$\frac{\partial \mathcal{L}^t(\theta)}{\partial \eta} = \omega(y^{(t)},\mathbf{x}^{(t)}) - \mathbb{E}_{p(y|\mathbf{x}^{(t)};\theta)}\omega(y,\mathbf{x}^{(t)})$$

Root

Max Margin Formulation (1/3)

Same as LSSVM

$$f_\theta(\mathbf{x}, y) = \max_{\mathbf{h}} \theta^\top \Phi(\mathbf{x}, \mathbf{h}, y)$$

$$\begin{aligned}
\min_{\theta,\xi} \quad & \frac{1}{2}||\theta||^2 + C\sum_{t=1}^{N} \xi^{(t)} \\
\text{s.t.} \quad & \max_{\mathbf{h}} \theta^\top \Phi(\mathbf{x}^{(t)}, \mathbf{h}, y) - \max_{\mathbf{h}'} \theta^\top \Phi(\mathbf{x}^{(t)}, \mathbf{h}', y^{(t)}) \\
& \leq \xi^{(t)} - \delta(y, y^{(t)}), \quad \forall t, \quad \forall y \\
\text{where} \quad & \delta(y, y^{(t)}) = \begin{cases} 1 & \text{if } y \neq y^{(t)} \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{11}$$

Loss function

Not convex because of hidden nodes

Alternative: Use CCCP from the LSSVM paper

Max Margin Formulation (2/3)

Coordinate Descent

1) Fixing θ, ξ, optimize the latent variable $\mathbf{h}$ for each pair $\langle \mathbf{x}^{(t)}, y \rangle$ of an example $\mathbf{x}^{(t)}$ and a possible labeling y:

$$\mathbf{h}_y^{(t)} = \arg\max_{\mathbf{h}} \theta^\top \Phi(\mathbf{x}^{(t)}, \mathbf{h}, y)$$

Infer with Viterbi
(Describe LP but don't use)

2) Fixing $\mathbf{h}_y^{(t)} \quad \forall t, \quad \forall y$, optimize θ, ξ by solving the following optimization problem:

$$\begin{aligned} \min_{\theta,\xi} \quad & \frac{1}{2}||\theta||^2 + C\sum_{t=1}^{N} \xi^{(t)} \\ \text{s.t.} \quad & \theta^\top \Phi(\mathbf{x}^{(t)}, \mathbf{h}_y^{(t)}, y) - \theta^\top \Phi(\mathbf{x}^{(t)}, \mathbf{h}_{y^{(t)}}^{(t)}, y^{(t)}) \\ & \leq \xi^{(t)} - \delta(y, y^{(t)}), \quad \forall t, \quad \forall y \qquad (16) \end{aligned}$$

SMO-like algorithm

Max Margin Formulation (3/3)

Similar to SMO
(Except h varies with x)

2) Fixing $\mathbf{h}_y^{(t)}$ $\forall t$, $\forall y$, optimize θ, ξ by solving the following optimization problem:

$$\begin{aligned} \min_{\theta,\xi} \quad & \frac{1}{2}||\theta||^2 + C\sum_{t=1}^{N}\xi^{(t)} \\ \text{s.t.} \quad & \theta^\top \Phi(\mathbf{x}^{(t)}, \mathbf{h}_y^{(t)}, y) - \theta^\top \Phi(\mathbf{x}^{(t)}, \mathbf{h}_{y^{(t)}}^{(t)}, y^{(t)}) \\ & \leq \xi^{(t)} - \delta(y, y^{(t)}), \quad \forall t, \quad \forall y \end{aligned} \tag{16}$$

Optimize parameters w/ dual:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{t=1}^{N}\sum_{y} \alpha_{t,y}\delta(y, y^{(t)}) - \frac{1}{2}||\sum_{t=1}^{N}\sum_{y}\alpha_{t,y}\Psi(\mathbf{x}^{(t)}, y)||^2 \\ \text{s.t.} \quad & \sum_{y}\alpha_{t,y} = C, \quad \forall t \\ & \alpha_{t,y} \geq 0, \quad \forall t, \quad \forall y \end{aligned} \tag{17}$$

Quadratic programing problem:

$$\begin{aligned} & \mathcal{L}(\{\alpha_{t,y} : \forall y\}) \\ = & \sum_{y}\alpha_{t,y}\delta(y, y^{(t)}) - \frac{1}{2}\Bigg[||\sum_{y}\alpha_{t,y}\Psi(\mathbf{x}^{(t)}, y)||^2 \\ & +2\Big(\sum_{y}\alpha_{t,y}\Psi(\mathbf{x}^{(t)}, y)\Big)^\top\Big(\sum_{s:s\neq t}\sum_{y'}\alpha_{s,y'}\Psi(\mathbf{x}^{(s)}, y')\Big)\Bigg] \\ & + \text{other terms not involving } \{\alpha_{t,y} : \forall y\} \end{aligned}$$

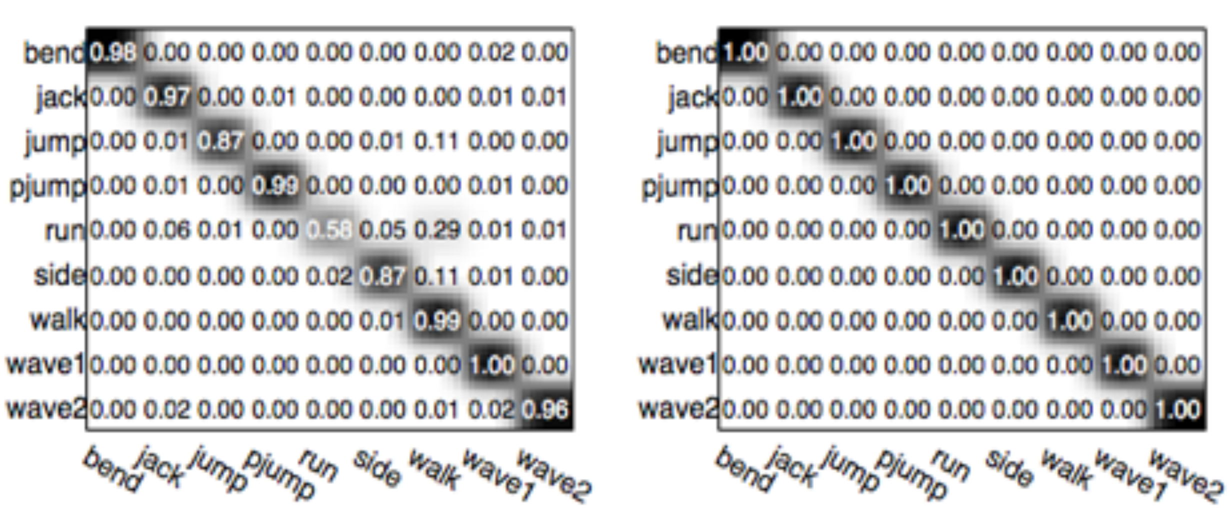
$$\begin{aligned} \max_{\alpha_{t,y}:\forall y} \quad & \mathcal{L}(\{\alpha_{t,y} : \forall y\}) \\ \text{s.t.} \quad & \sum_{y}\alpha_{t,y} = C \\ & \alpha_{t,y} \geq 0, \quad \forall y \end{aligned}$$

Weizmann Results

83 videos. 9 people. 9 actions
Train on 5, test on 4

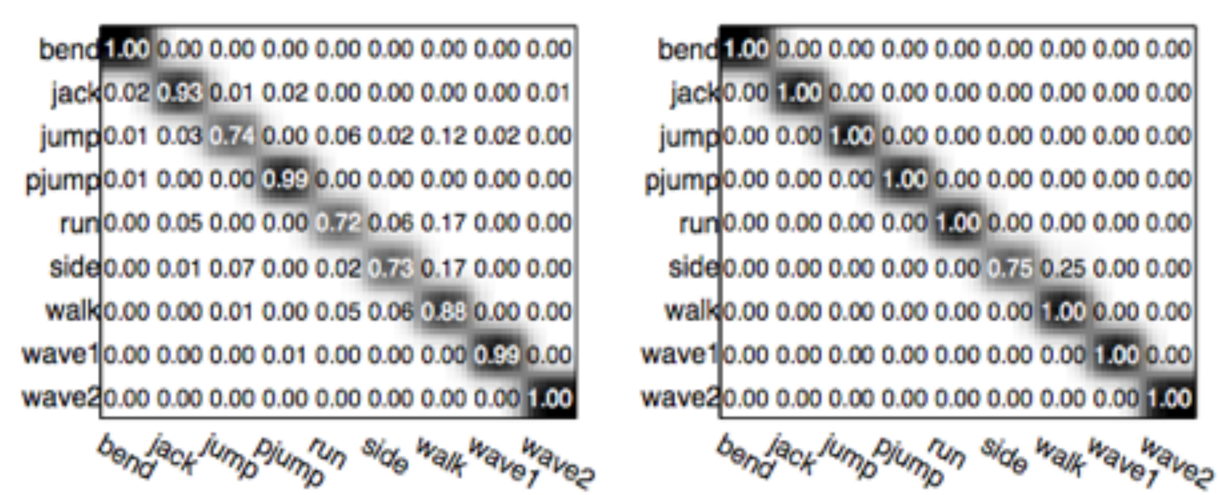
	per-frame	per-video	per-cube
MMHCRF	**0.9311**	**1**	N/A
HCRF	**0.9029**	**0.9722**	N/A
Jhuang et al. [22]	N/A	0.988	N/A
Niebles & Fei-Fei [24]	0.55	0.728	N/A
Blank et al. [9]	N/A	N/A	0.9964

* No tracking



(a) per-frame (b) per-video

Max Margin



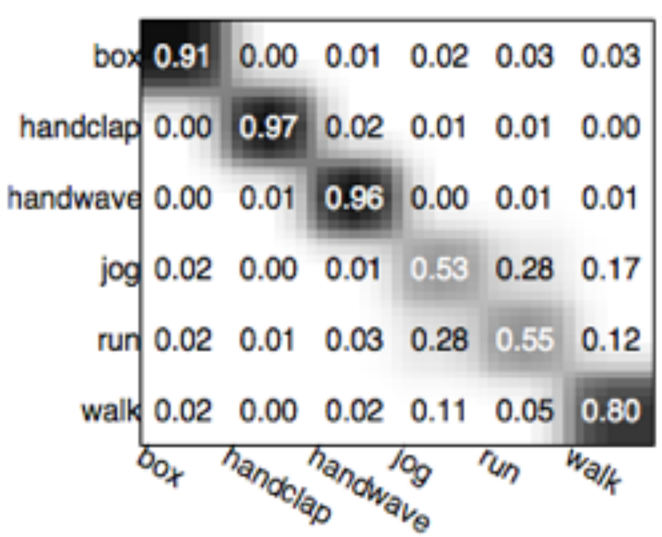
(a) per-frame (b) per-video

Probabilistic

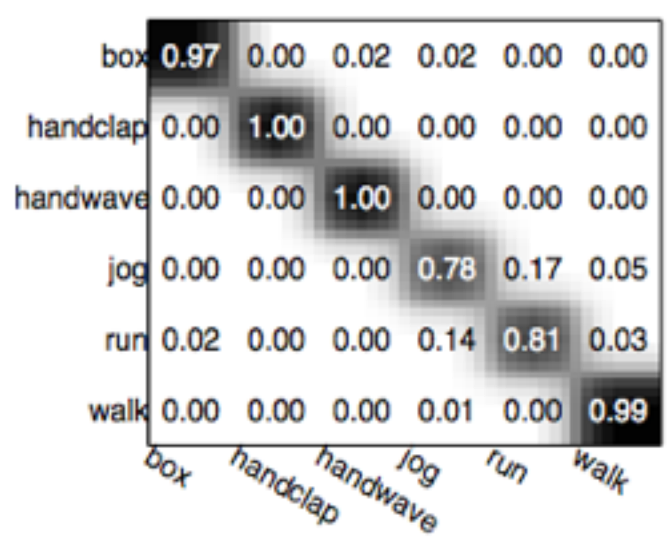
KTH Results

25 users, 4 scenes, 6 actions
“roughly half” train/test

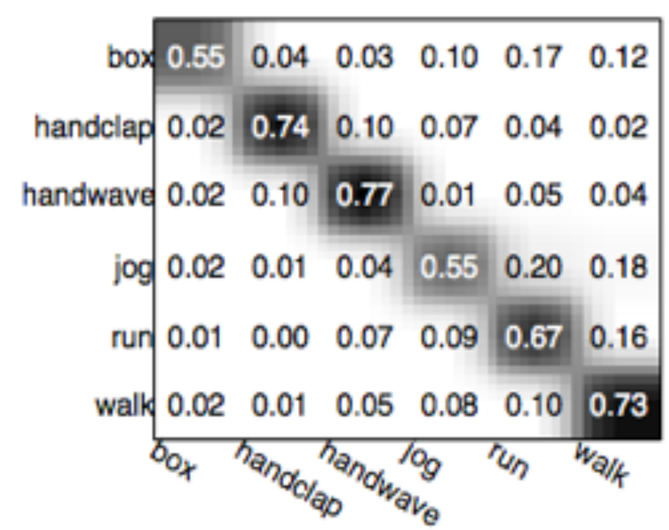
methods	accuracy
MMHCRF	**0.9251**
HCRF	**0.8760**
Liu & Shah [50]	0.9416
Jhuang et al. [22]	0.9170
Nowozin et al. [13]	0.8704
Niebles et al. [12]	0.8150
Dollár et al. [11]	0.8117
Schuldt et al. [14]	0.7172
Ke et al. [51]	0.6296



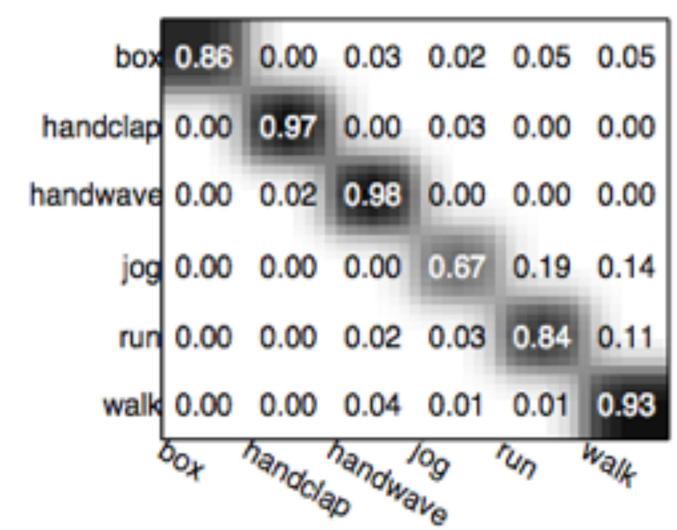
(a) per-frame



(b) per-video



(a) per-frame



(b) per-video

Max Margin

Probabilistic

Alternative models (1/2)

		method	Weizmann		KTH	
			per-frame	per-video	per-frame	per-video
	1	root model	0.7470	0.8889	0.5377	0.7339
		local HCRF				
		$\|\mathcal{H}\|$=6	0.5722	0.5556	0.4749	0.5607
No root	2	$\|\mathcal{H}\|$=10	0.6656	0.6944	0.4452	0.5814
		$\|\mathcal{H}\|$=20	0.6383	0.6111	0.4282	0.5504
		HCRF				
Prob	3	$\|\mathcal{H}\|$=6	**0.8682**	**0.9167**	**0.6633**	**0.7855**
		$\|\mathcal{H}\|$=10	**0.9029**	**0.9722**	**0.6698**	**0.8760**
		$\|\mathcal{H}\|$=20	**0.8557**	**0.9444**	**0.6444**	**0.7512**
		MMHCRF				
		$\|\mathcal{H}\|$=6	**0.8996**	**0.9722**	**0.7064**	**0.8475**
Max Margin	4	$\|\mathcal{H}\|$=10	**0.9311**	**1.0000**	**0.7853**	**0.9251**
		$\|\mathcal{H}\|$=20	**0.8891**	**0.9722**	**0.7486**	**0.8966**

Alternative models (2/2)

Experiment 1: Remove pairwise terms
Experiment 2: Exp1 + Convert to N 1-vs-all LSVMs (N=class)
Experiment 3: Train SVM on Exp 2 outputs.

All experiments using Max Margin version
Pairwise on avg. ~1% better

Full model

method	Weizmann		KTH	
	per-frame	per-video	per-frame	per-video
MMHCRF				
$\|\mathcal{H}\|=6$	**0.8996**	**0.9722**	**0.7064**	**0.8475**
$\|\mathcal{H}\|=10$	**0.9311**	**1.0000**	**0.7853**	**0.9251**
$\|\mathcal{H}\|=20$	**0.8891**	**0.9722**	**0.7486**	**0.8966**

Reduced model

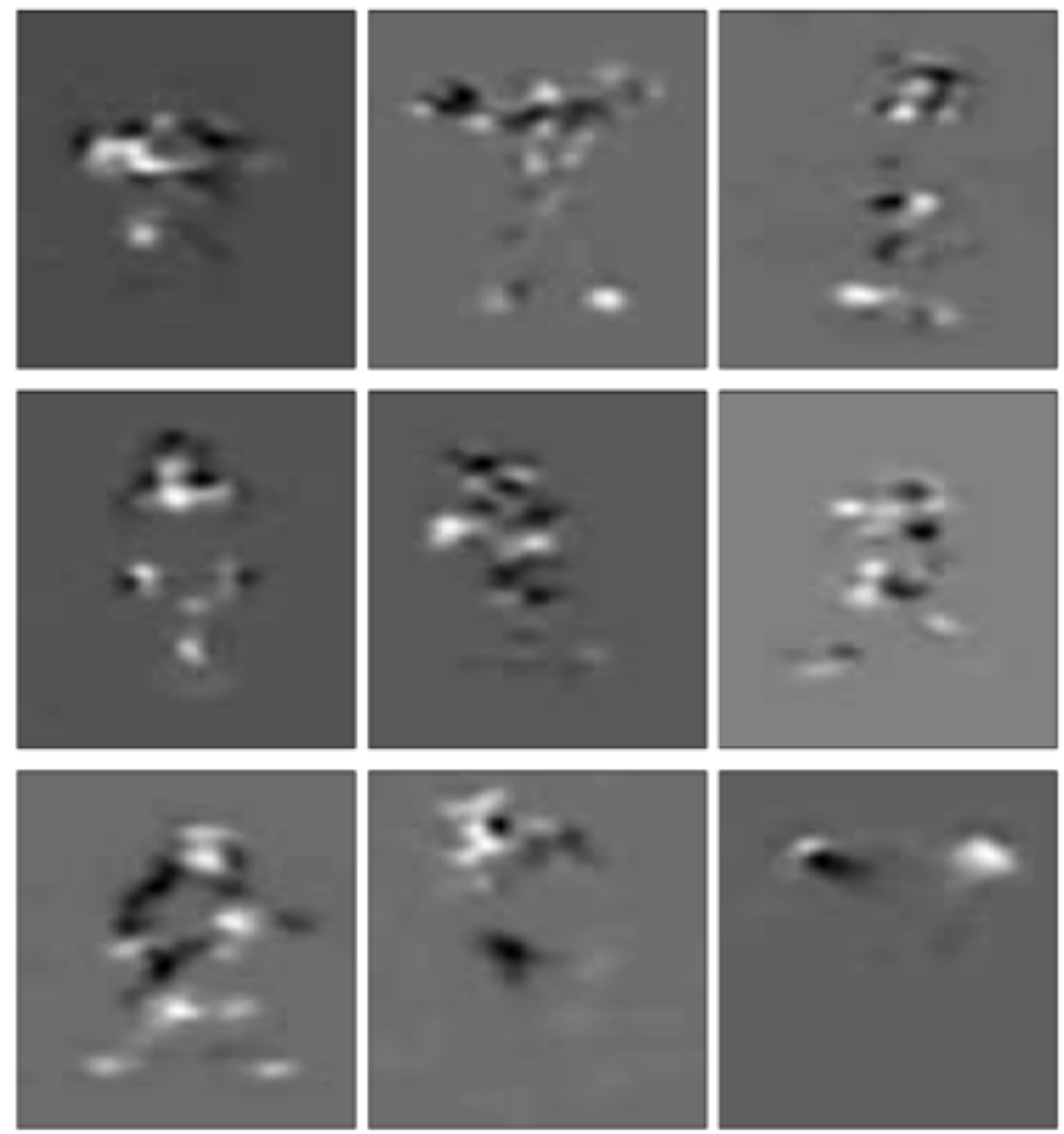
method	Weizmann		KTH	
	per-frame	per-video	per-frame	per-video
no pairwise				
$\|\mathcal{H}\| = 6$	0.8344	0.9062	0.6767	0.8527
$\|\mathcal{H}\| = 10$	0.8414	0.9688	0.7005	0.8941
$\|\mathcal{H}\| = 20$	0.8358	0.9688	0.6891	0.8734
one-against-all				
$\|\mathcal{H}\| = 6$	0.7525	0.8889	0.5171	0.6589
$\|\mathcal{H}\| = 10$	0.7507	0.8611	0.5052	0.6589
$\|\mathcal{H}\| = 20$	0.7447	0.8889	0.5052	0.6899
one-against-all + SVM				
$\|\mathcal{H}\| = 6$	0.8173	0.9444	0.5705	0.7209
$\|\mathcal{H}\| = 10$	0.8460	0.9444	0.5610	0.7287
$\|\mathcal{H}\| = 20$	0.8145	0.9444	0.5623	0.7442

Part/filter visualization (Weizmann)

Colors = class of hidden part

Red: moving down

Green: hand waving



Filters (per class)

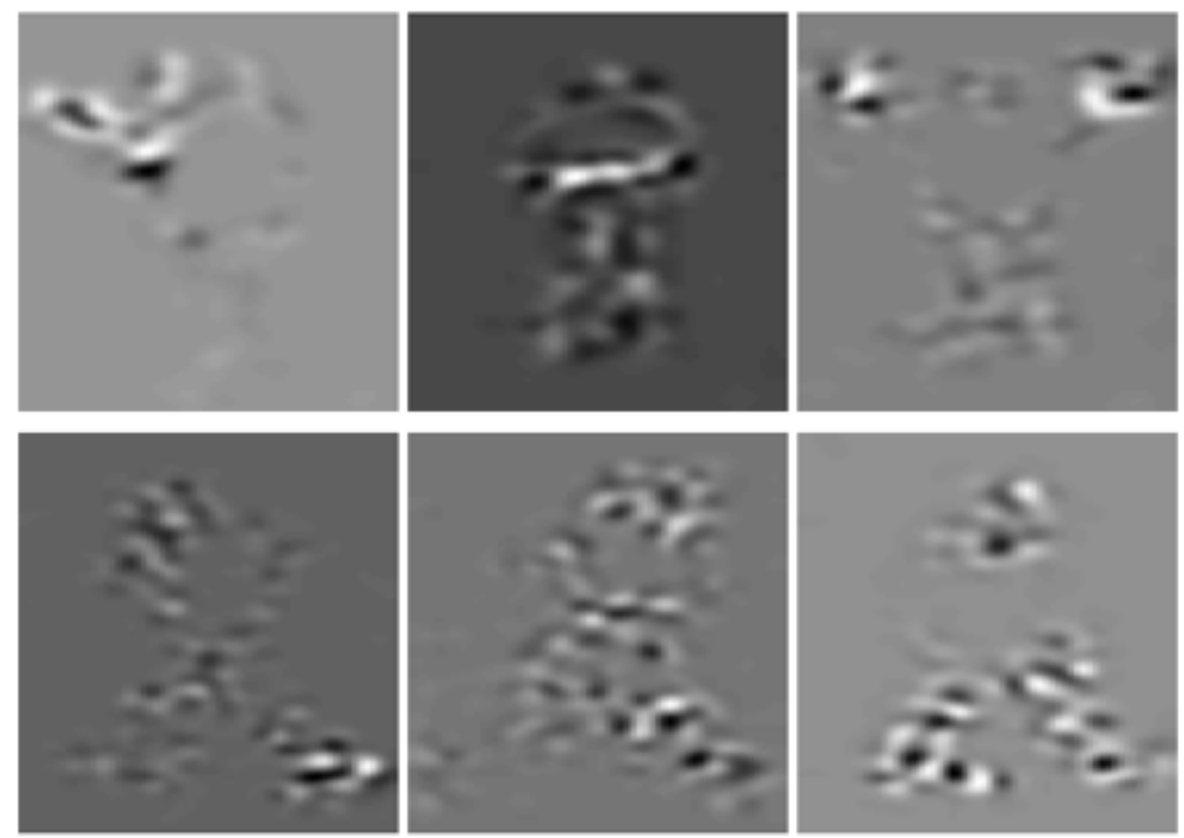
Part/filter visualization (KTH)

Colors = class of hidden part

Pink: boxing

Red: clapping

Green: waving



Filters (per class)

Takeaways

Good

- Max margin > Probabilistic (~5% acc. here)
- Nice analysis of each component
  - e.g. Root filter + pairwise analysis
- Local+global features >> Global features

Bad

- Does this generalize to other datasets??
  - Weizmann and KTH are very similar and too simple
- No temporal component
- [Pet peeve: introduce a lot of unused math due to model assumptions]