A Hierarchical Model of Shape and Appearance for Human Action Classification

Juan Carlos Niebles [1,2]
jnieble2@uiuc.edu

Li Fei-Fei [1,3]
feifeili@cs.princeton.edu

http://vision.cs.princeton.edu

[1] University of Illinois at Urbana – Champaign

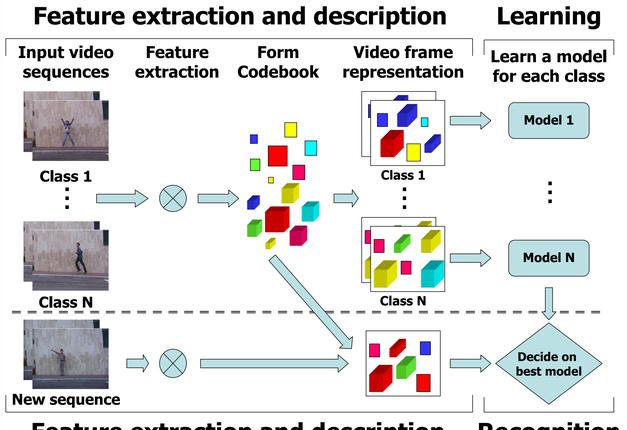
[2] Universidad del Norte, Colombia

[3] Princeton University

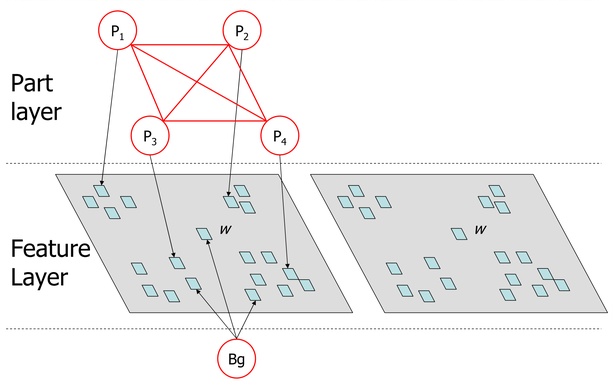
Highlights and Summary

- A novel model for human action categorization from video sequences.
- Our model can be characterized as a constellation of bags-of-features.
- Use of hybrid features: combines both static shape and spatio-temporal features.

Algorithm



Previous Works



constellation

☹ Small number of features

☺ Strong shape representation

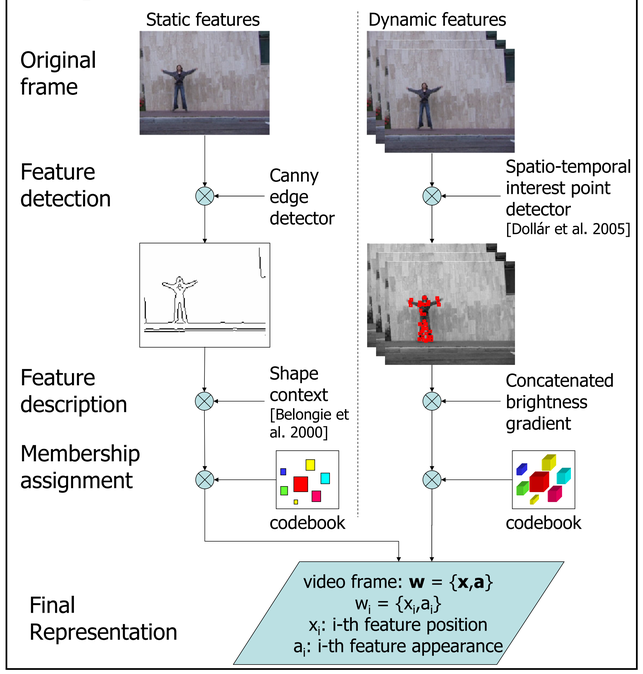
bag-of-features

☺ Large number of features

☹ No geometrical or shape information

[Weber et al. 2000, Csurka et al. 2004, Sudderth et al. 2006]

Hybrid features



Learning

Estimate model parameters using EM

$$\theta_\omega = \left\{\boldsymbol{\mu}_{L,\omega}, \boldsymbol{\Sigma}_{L,\omega}, \boldsymbol{\Sigma}^X_{p,\omega}, \theta^A_{p,\omega}, \theta^X_0, \theta^A_0\right\} \quad p = 1 \ldots P \quad \omega = 1 \ldots \Omega$$

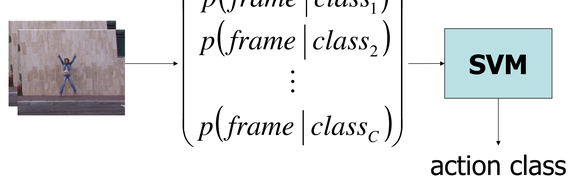
- **E-step:**

$$p\left(\mathbf{h}, \omega \middle| \mathbf{w}, \mathbf{Y}, \theta^{old}\right) \approx \frac{\pi_\omega p\left(\mathbf{Y}\middle|\mathbf{h}, \theta^{old}_\omega\right) p\left(\mathbf{h}\middle|\theta^{old}_\omega\right) p\left(\mathbf{w}\middle|\mathbf{Y}, \mathbf{h}, \mathbf{m}^*, \theta^{old}_\omega\right)}{p\left(\mathbf{w}, \mathbf{Y}\middle|\theta^{old}_\omega\right)}$$

- **M-Step:**

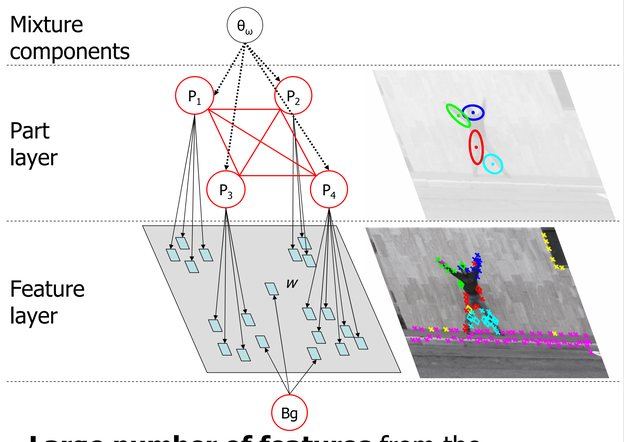
$$\theta^{new} = \arg\max_\theta \sum_{\mathbf{h}} p\left(\mathbf{h}, \omega \middle| \mathbf{w}, \mathbf{Y}, \theta^{old}\right) \ln p\left(\mathbf{w}, \mathbf{Y}, \mathbf{h}, \omega \middle| \theta\right)$$

Recognition



- Classify actions in both frame based and video based manner
- Video classification based on majority votes of frames

Hierarchical model



- **Large number of features** from the **bag-of-features** model
- **Strong shape representation** from the **constellation** model.

Approximated data likelihood:

$$p\left(\mathbf{w}, \mathbf{Y}\middle|\boldsymbol{\theta}\right) \approx \sum_{\omega=1}^{\Omega}\left[\pi_\omega \sum_{\mathbf{h}\in H} p\left(\mathbf{h}\middle|\theta_\omega\right) \underbrace{p\left(\mathbf{Y}\middle|\mathbf{h}, \theta_\omega\right)}_{\text{Part layer}} \underbrace{p\left(\mathbf{w}\middle|\mathbf{Y}, \mathbf{m}^*, \mathbf{h}, \theta_\omega\right)}_{\text{Local feature layer}}\right]$$

Part layer term:

$$p\left(\mathbf{Y} \mid \mathbf{h}, \theta_\omega\right) = N\left(\mathbf{Y}_\mathbf{T}(\mathbf{h}) \mid \boldsymbol{\mu}_{L,\omega}, \boldsymbol{\Sigma}_{L,\omega}\right)$$

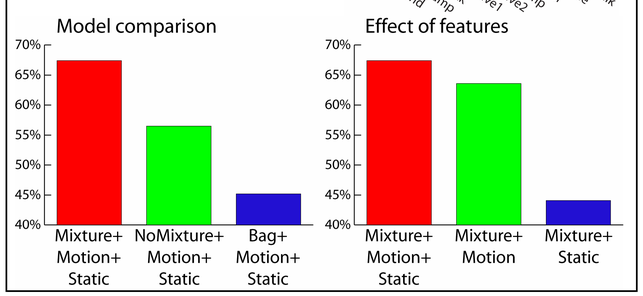
Local feature layer term:

$$p\left(\mathbf{w}\middle|\mathbf{Y}, \mathbf{m}^*, \mathbf{h}, \boldsymbol{\theta}_\omega\right) = \prod_{\mathbf{w}_j \in Bg} \underbrace{p\left(x^r_j\middle|\theta^X_0\right)}_{\text{Bg Shape}} \underbrace{p\left(a_j\middle|\theta^A_0\right)}_{\text{Bg Appearance}} \prod_{p=1}^{P} \prod_{\mathbf{w}_i \in P_p} \underbrace{p\left(x^r_i\middle|\mathbf{Y}, h_p, \theta^X_p\right)}_{\text{Part Shape}} \underbrace{p\left(a_i\middle|\theta^A_p\right)}_{\text{Part Appearance}}$$

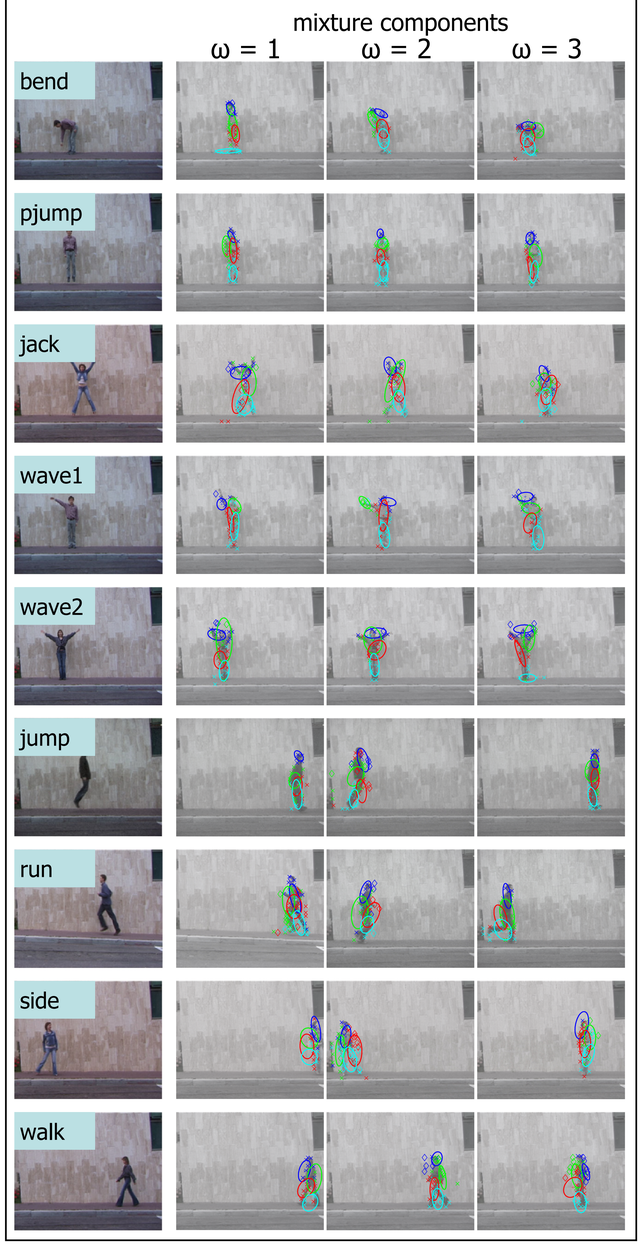
Experimental Results

- 9 action classes, performed by 9 subjects [Blank et al 2005]
- Leave one out cross-validation
- Video Classification performance: 72.8%

	bend	pjump	jack	wave1	wave2	jump	run	side	walk
bend	1.0	.00	.00	.00	.00	.00	.00	.00	.00
pjump	.00	1.0	.00	.00	.00	.00	.00	.00	.00
jack	.00	.00	1.0	.00	.00	.00	.00	.00	.00
wave1	.22	.11	.11	.44	.11	.00	.00	.00	.00
wave2	.00	.00	.11	.22	.67	.00	.00	.00	.00
jump	.00	.00	.00	.00	.00	.78	.00	.11	.11
run	.00	.00	.11	.00	.00	.11	.56	.11	.11
side	.00	.00	.00	.00	.00	.33	.11	.56	.00
walk	.00	.00	.00	.00	.00	.11	.00	.33	.56



Action Models



Conclusions

- The constellation of bags-of-features is able to capture semantic information of human action classes.
- Combines hybrid features: static shape features and dynamic motion features.
- Capable of classifying in both frame based and video based manner.

Ref: J.C. Niebles & L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. *CVPR* 2007. Minneapolis, USA.