同济大学计算机科学与技术系

多媒体与智能计算实验室

Encoding Scale into Fisher Vector for Human Action Recognition

Bowen Zhang and Hanli Wang

Department of Computer Science and Technology,
Tongji University, Shanghai, China

Key Laboratory of Embedded System and Service Computing,
Ministry of Education,
Tongji University, Shanghai, China

Content

Outline

Action Recognition Task

Testing Video

Classification

Results

- Brush hair
- Cartwheel
- Catch
- Chew
- Clap
- Climb
- Climb stairs
- Dive
- Draw sword
- Dribble

- Camera motions
- Different viewpoints/backgrounds/…
- Variety of movements

Outline

Related Work

- **Trajectory based Feature Extraction and Description**
  - KLT trajectory [Lucas et al 1981]
  - SIFT trajectory [Sun et al 2009]
  - Dense Trajectory (DT) [Wang et al 2013]
  - Improved Dense Trajectory (iDT) [Wang et al 2013]
  - ……
- **Feature Encoding Method**
  - Bag-of-Visual-Words (BoVW) [Sivic et al 2003]
  - Fisher Vector (FV) [Perronin et al 2010]
  - Spatial Fisher Vector (SFV) [Krapac et al 2011]
  - Spatial Temporal Fisher Vector (STFV) [Oneata et al 2013]
  - ……

Outline

Temporal Scale in Video

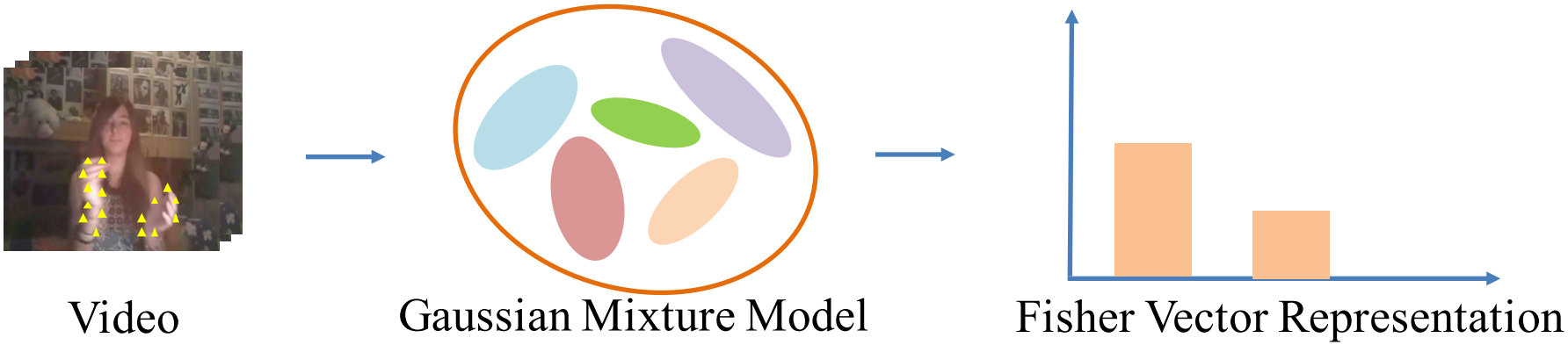
Motivation

- Scale information plays a great role in image area.

  Low level features like SIFT and SURF use spatial scale information to determine salient points. Recent work like Domain-size Pooling [Dong et al 2015] utilizes scale information to improve matching performance.

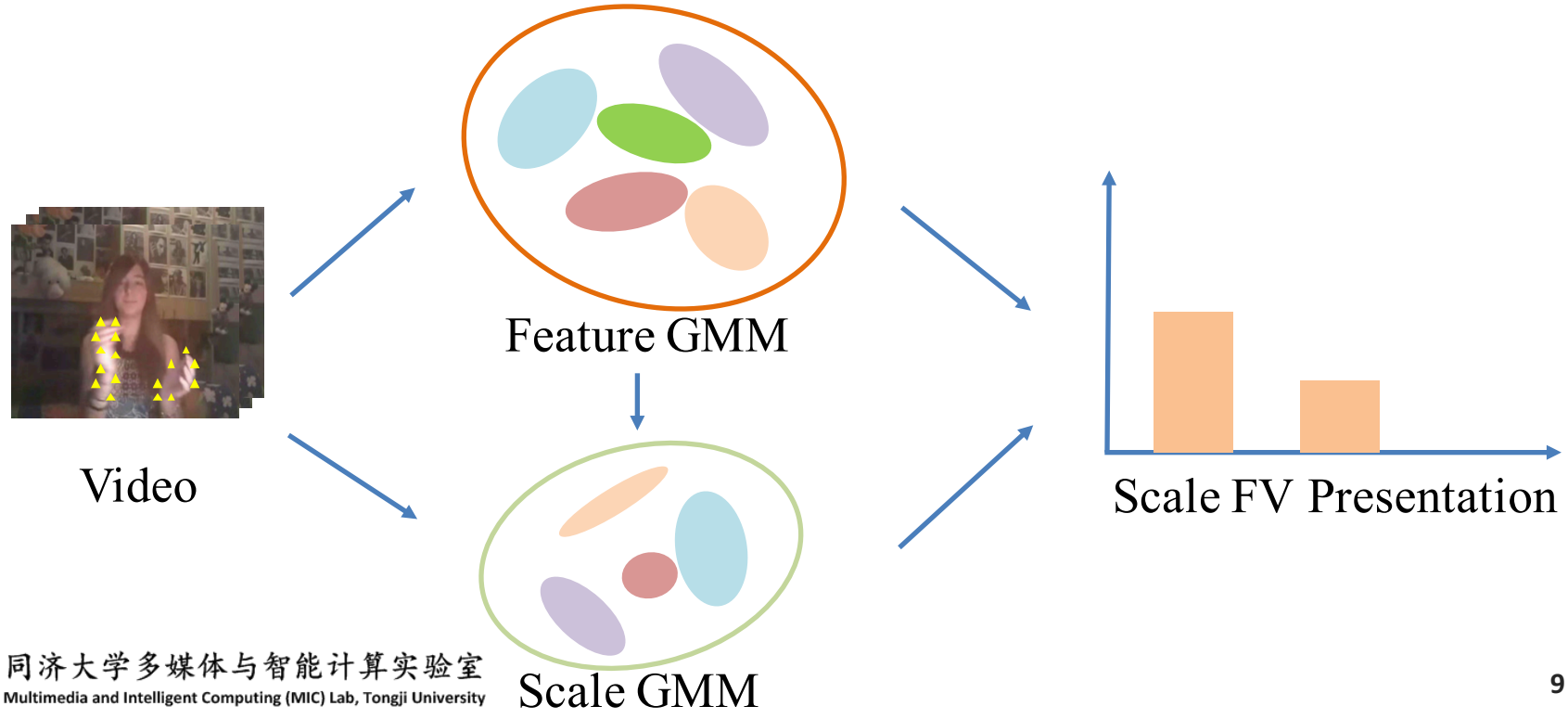
- Different actions should have divergent movement patterns.

  Movement pattern contains discriminative information.



Temporal Scale in Video

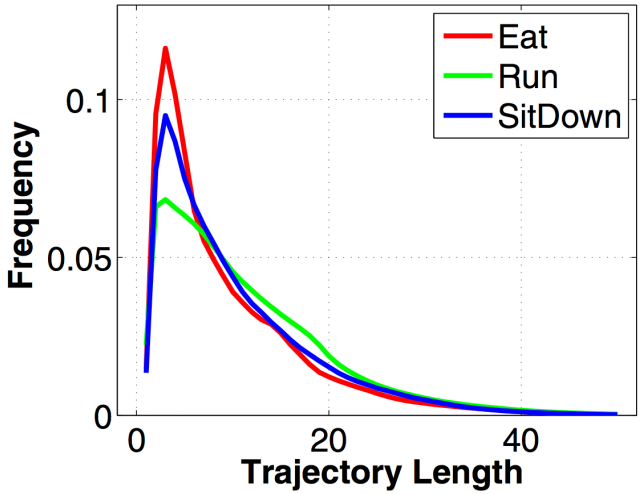
- **We want to …**
  - Give a definition to temporal scale information in video.
  - Encode scale information into Fisher Vector.
  - Combine scale and position information with low level features to strengthen the performance.



Temporal Scale in Video

Definition

- We define the temporal scale in videos as the length of an event trajectory within a specific time period.



Frequency of trajectory length about three action classes ('Eat', 'Run' and 'SitDown') in Hollywood-2

Outline

Fisher Vector Revisited

- **Gaussian Mixture Model**

GMM: $p(y_t) = \sum_{i=1}^{K} \pi_i N(y_t; \mu_i, \Sigma_i)$

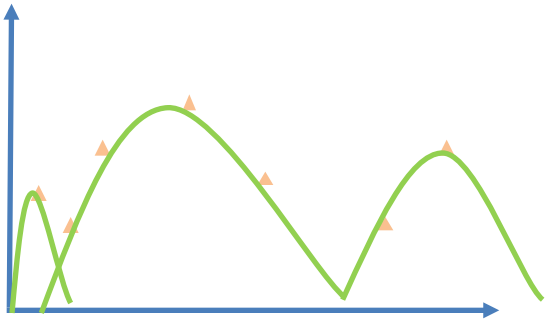
Features: y_t

Prior possibility: $\pi_i = \frac{e^{\alpha_i}}{\sum_j e^{\alpha_j}}$

Mean: μ_i

Covariance Matrix: Σ_i

Total number of GMM: K



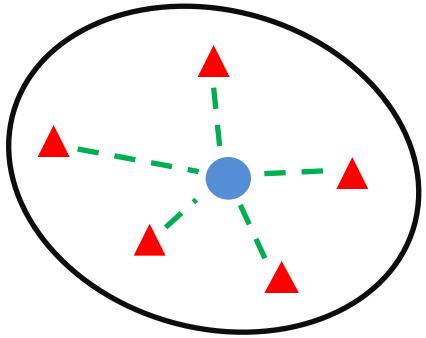
Fisher Vector Revisited

- **Fisher Vector**

$$\frac{\partial \ln p(y_t)}{\partial \alpha_i} = q_i(t) - \pi_i,$$
$$\frac{\partial \ln p(y_t)}{\partial \mu_i} = q_i(t)\Sigma_i^{-1}(y_t - \mu_i),$$
$$\frac{\partial \ln p(y_t)}{\partial \Sigma_i^{-1}} = q_i(t)\frac{\Sigma_i - diag\left((y_t - \mu_i)^2\right)}{2}.$$

Posterior possibility: $q_i(t) = \frac{\pi_i N(y_t; \mu_i, \Sigma_i)}{\sum_{j=1}^{K} \pi_j N(y_t; \mu_j, \Sigma_j)}$

Dimensions: $(2D + 1)K$



Outline

Scale Fisher Vector

- **Scale Gaussian Mixture Model**

Scale GMM: $p(y_t, s_t) = \sum_{i=1}^{K} \pi_i N(y_t; \mu_i, \Sigma_i) \sum_{j=1}^{J} \lambda_{ji} N(s_t; \delta_{ji}, \mathrm{Z}_{ji})$

Scale information: $s_t = \{\delta_t, \tau_t\}$

Prior possibility for scale: $\lambda_{ji} = \frac{e^{\alpha_{ji}}}{\sum_m e^{\alpha_{mi}}}$

Mean for scale information: δ_{ji}

Covariance Matrix for scale information: Z_{ji}

Total number of GMM for scale information: J

Scale Fisher Vector

- **Scale Fisher Vector**

$$\frac{\partial \ln p(y_t)}{\partial \alpha_i} = q_i(t) - \pi_i,$$
$$\frac{\partial \ln p(y_t)}{\partial \mu_i} = q_i(t)\Sigma_i^{-1}(y_t - \mu_i),$$
$$\frac{\partial \ln p(y_t)}{\partial \Sigma_i^{-1}} = q_i(t)\frac{\Sigma_i - diag\left((y_t - \mu_i)^2\right)}{2},$$
$$\frac{\partial \ln p(y_t, s_t)}{\partial \alpha_{ji}} = q_i(t)\left(r_{ji}(t) - \lambda_{ji}\right),$$
$$\frac{\partial \ln p(y_t, s_t)}{\partial \delta_{ji}} = q_i(t) r_{ji}(t) \mathrm{Z}_{ji}^{-1}(s_t - \delta_{ji}),$$
$$\frac{\partial \ln p(y_t, s_t)}{\partial \mathrm{Z}_{ji}^{-1}} = q_i(t) r_{ji}(t)\frac{\mathrm{Z}_{ji} - diag\left((s_t - \delta_{ji})^2\right)}{2}.$$

Posterior possibility: $$q_i(t) = \frac{\pi_i N(y_t; \mu_i, \Sigma_i)\sum_{j=1}^{J}\lambda_{ji}N(s_t; \delta_{ji}, \mathrm{Z}_{ji})}{\sum_{l=1}^{K}\pi_j N(y_t; \mu_l, \Sigma_l)\sum_{j=1}^{J}\lambda_{jl}N(s_t; \delta_{jl}, \mathrm{Z}_{jl})}$$

Dimensions: $$((2D + 1) + 1 + J(2d + 1))K$$

Scale Fisher Vector

- **Combined Fisher Vector**

Combined GMM: $p(y_t, s_t) = \sum_{i=1}^{K} \pi_i N(y_t; \mu_i, \Sigma_i) \sum_{j=1}^{J} \lambda_{ji} N(u_t; \delta_{ji}, \mathrm{Z}_{ji})$

CombFV:

$$\frac{\partial \ln p(y_t)}{\partial \alpha_i} = q_i(t) - \pi_i,$$

$$\frac{\partial \ln p(y_t)}{\partial \mu_i} = q_i(t)\Sigma_i^{-1}(y_t - \mu_i),$$

$$\frac{\partial \ln p(y_t)}{\partial \Sigma_i^{-1}} = q_i(t)\frac{\Sigma_i - diag\left((y_t - \mu_i)^2\right)}{2},$$

$$\frac{\partial \ln p(y_t, s_t)}{\partial \alpha_{ji}} = q_i(t)(r_{ji}(t) - \lambda_{ji}),$$

$$\frac{\partial \ln p(y_t, s_t)}{\partial \delta_{ji}} = q_i(t) r_{ji}(t) \mathrm{Z}_{ji}^{-1}(u_t - \delta_{ji}),$$

$$\frac{\partial \ln p(y_t, s_t)}{\partial \mathrm{Z}_{ji}^{-1}} = q_i(t) r_{ji}(t)\frac{\mathrm{Z}_{ji} - diag\left((u_t - \delta_{ji})^2\right)}{2}.$$

$u_t = \{x_t, y_t, \gamma_t, \delta_t, \tau_t\}$

Position

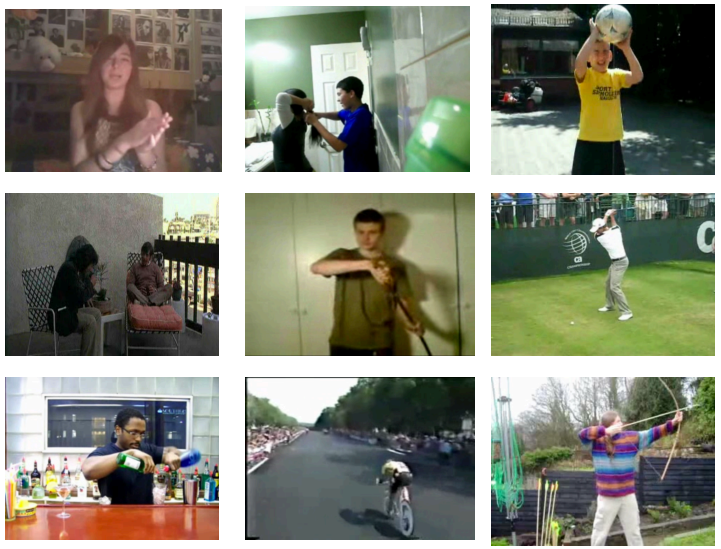
Scale information

Outline

Results on Standard Datasets

Datasets

- Hollywood-2 score: mAP
- HMDB51 score: accuracy



Experiment Setup

- Use improve Dense Trajectory (iDT) as feature extraction and description method.
- K of GMM is set as 256 for fair comparison with state-of-the-art results.
- Linear Support Vector Machine is employed

Results on Standard Datasets

- **Parameter *J* for Scale FV, STFV and CombFV**

Methods	*J*	Hollywood-2	HMDB51
STFV [14]	1	65.92%	60.39%
	2	65.92%	60.35%
	3	66.52%	61.07%
	4	65.69%	60.61%
	5	65.80%	60.96%
ScaleFV	1	64.69%	58.36%
	2	64.84%	58.65%
	3	64.79%	58.61%
	4	64.61%	58.41%
	5	64.65%	57.87%
CombFV	1	66.28%	60.41%
	2	66.29%	60.50%
	3	66.48%	60.41%
	4	66.53%	60.28%
	5	66.50%	60.56%
STFV+CombFV	-	66.72%	**61.50%**
STFV+CombFV+ScaleFV	-	**66.96%**	61.07%

Impact of parameter J on Hollywood-2 and HMDB51 for different methods

Results on Standard Datasets

- **Comparison with state-of-the-art results**

	Hollywood-2	HMDB51
Jain *et al.* [18]	62.5%	52.1%
Oneata *et al.* [14]	63.3%	54.8%
Wang *et al.* [12]	64.3%	57.2%
Wu *et al.* [19]	64.5%	-
Simonyan *et al.* [20]	-	59.4%
STFV+CombFV	66.72%	**61.50%**
STFV+CombFV+ScaleFV	**66.96%**	61.07%

Comparison with state-of-the-art performance on Hollywood-2 and HMDB51

Thank you!

1023zhangbowen@tongji.edu.cn
http://zbwglory.github.io