Learning Human Interaction by Interactive Phrases

Yu Kong[1,3], Yunde Jia[1] and Yun Fu[2]

[1]Beijing Institute of Technology

[2]Northeastern University

[3]University at Buffalo

Activity landscape

Individual action | Interaction | Group action | Crowd action




One person | Few people | Several people | Crowd of people

Number of people

Identification of each person

Easy	Easy	Not accurate but we can	Very challenging, open problem

Our work

Interaction: Boxing

Objective: recognizing human interactions from videos.

Applications

Motion analysis

Judge sports automatically
Video game interfaces

Group activity understanding

Scene analysis
Smart surveillance

Detect unusual behavior

Smart surveillance

Motivation

An interaction is determined by individual actions.

Recognize interaction by **action co-occurrence**



Action co-occurrence

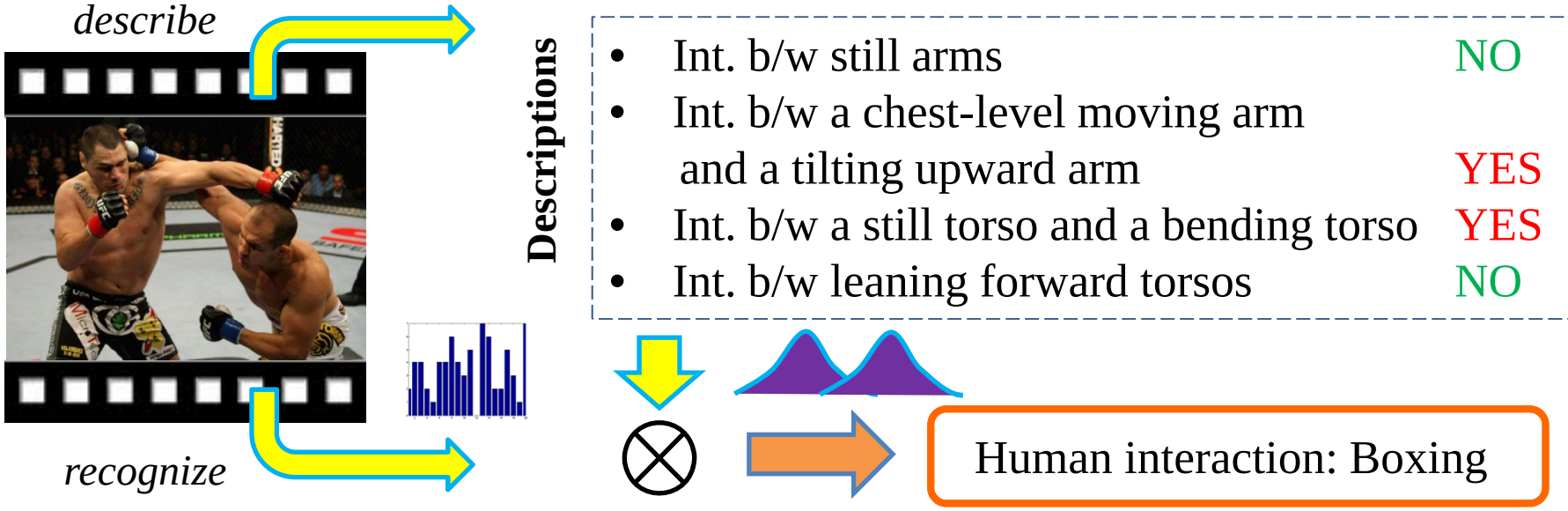
Attack-Protect head Attack-Dodge Attack-Hit back

Interaction: Boxing

Problem: co-occurrence relationships are not expressive enough to deal with interactions with large variations.

Motivation

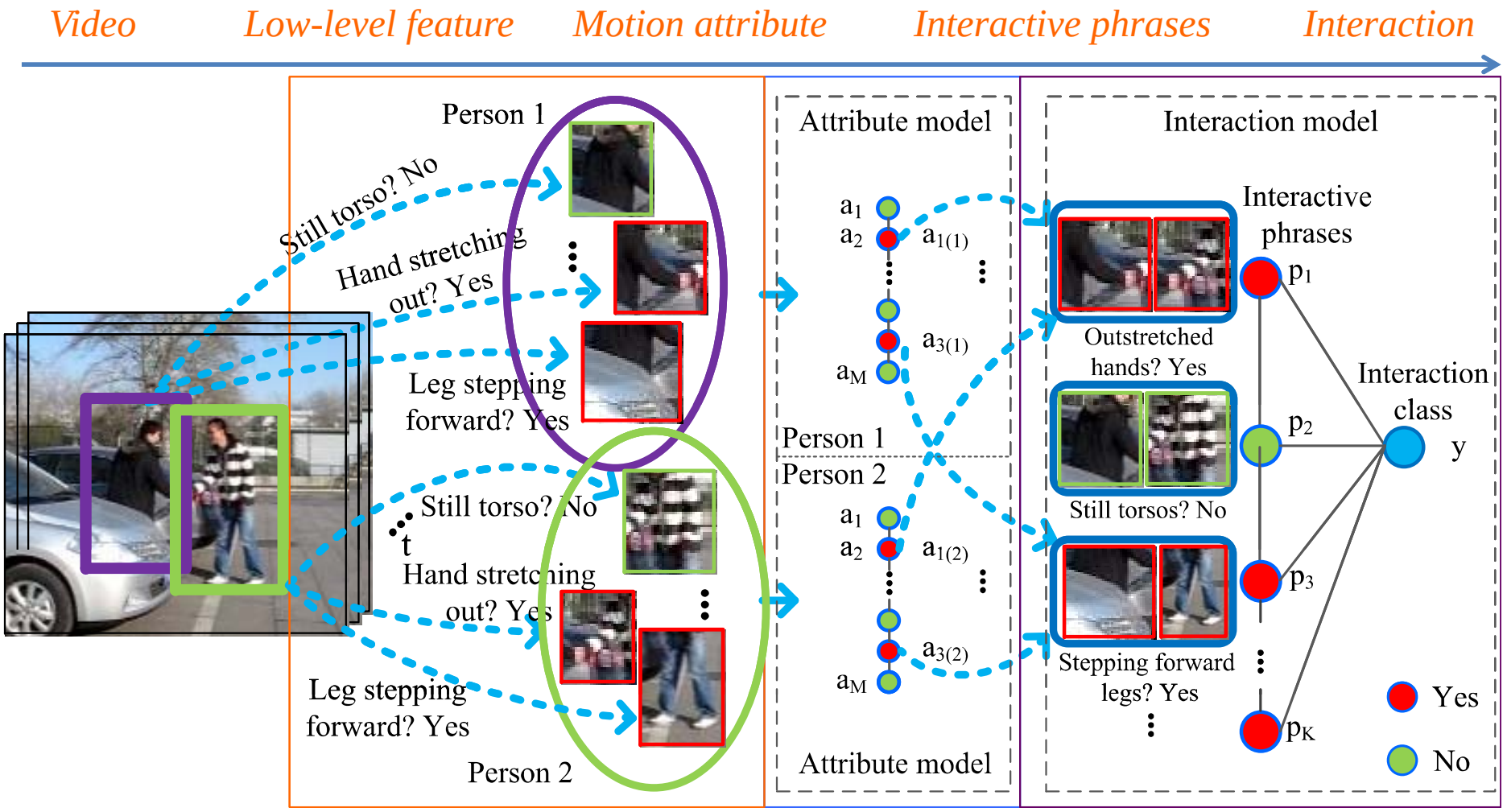
We introduce *interactive phrases* to describe human interactions.



Interactive phrases:

- More expressive to describe complicated human interactions.
- Binary motion relationships between people. E.g., relationships between arms, legs, and torsos, etc.
- Mid-level feature learned from data

Flowchart of our method

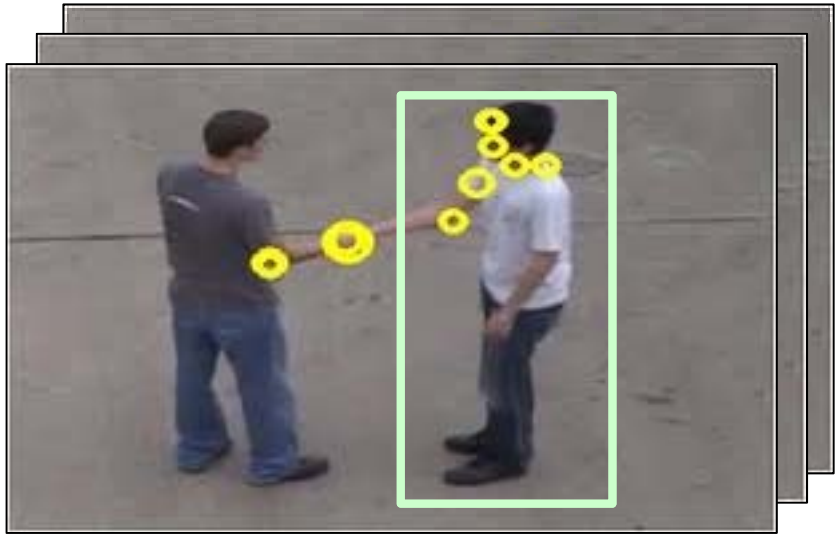


Feature extraction
Build individual action representation

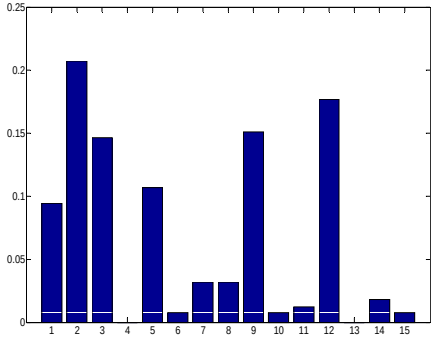
Attribute model
Detect individual motion attribute

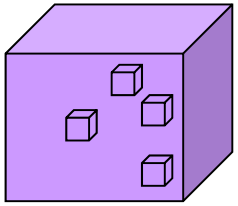
Interaction model
Learn interactive phrases and recognize interaction

Individual action representation

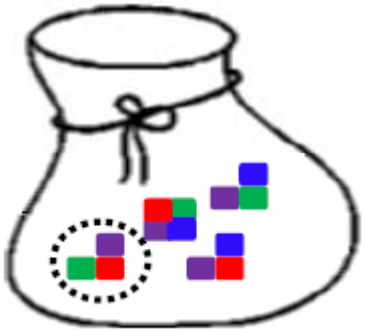


[d1, d2, …,dn]





Low-level local feature

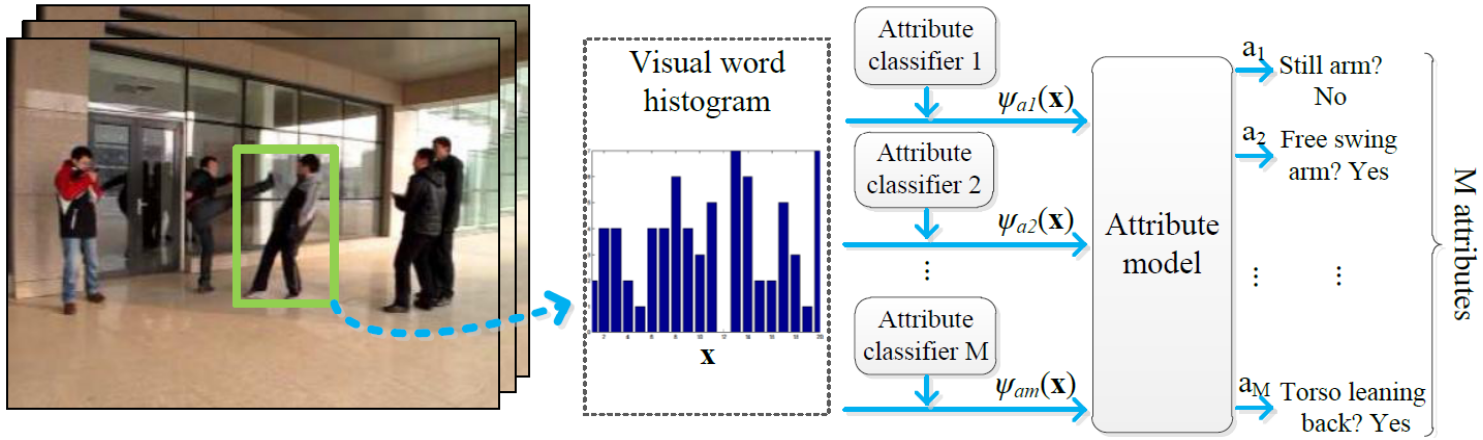


Learned dictionary

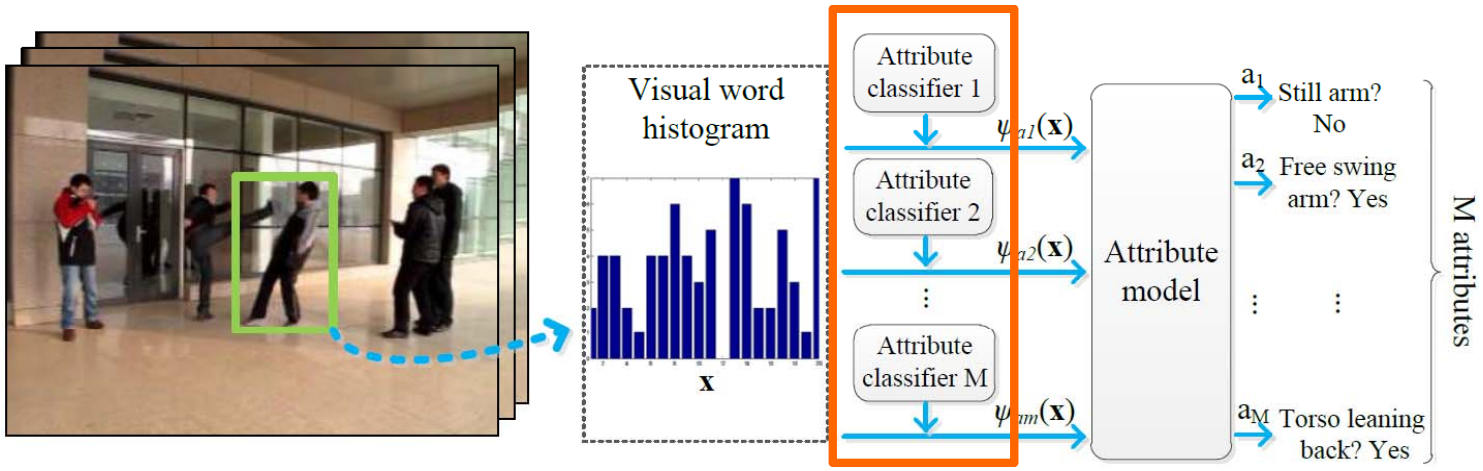
Attribute model

Objective: **Jointly** detect individual motion attributes.

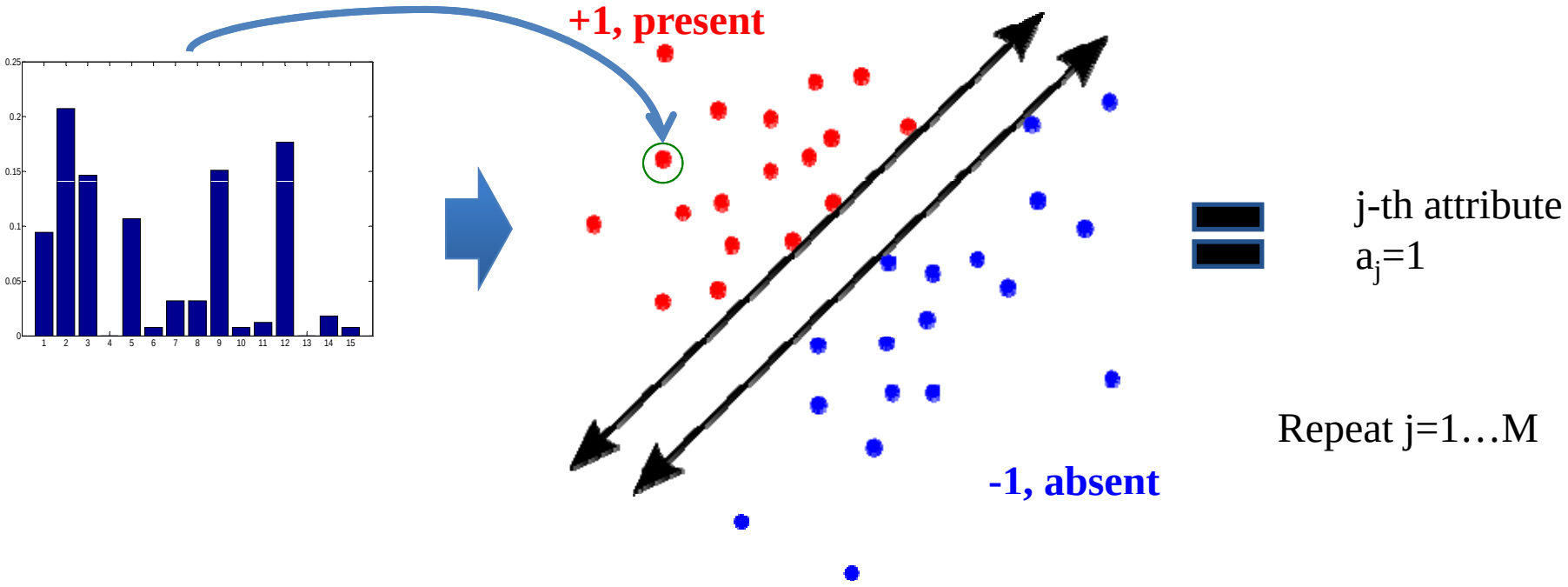
Motion attributes: describe individual motion, e.g. arm stretching out, leg stepping forward, etc.



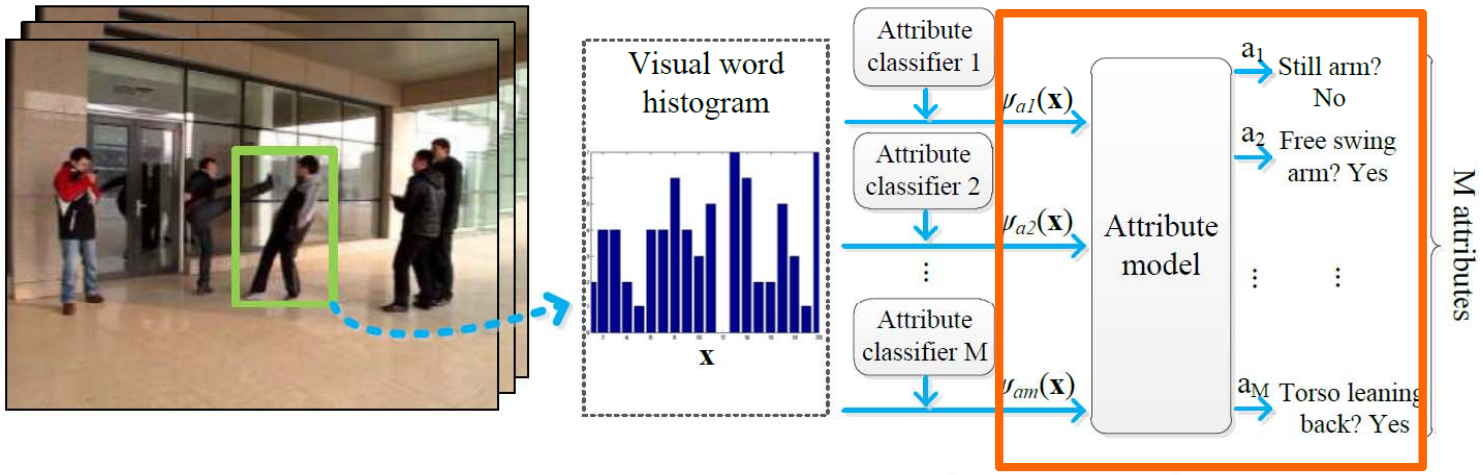
id	attributes a_m
1	still arm
2	hand stretching out motion
3	arm chest-level motion
4	two arms chest-level motion
5	arm raising up motion
6	arm embracing motion
7	arm free swinging motion
8	arm intense motion
9	still leg
10	leg stepping forward motion
11	leg kicking motion
12	leg stepping back motion
13	still torso
14	torso leaning back motion
15	torso leaning forward motion
16	torso bending motion
17	friendly motion



Individual attribute detection



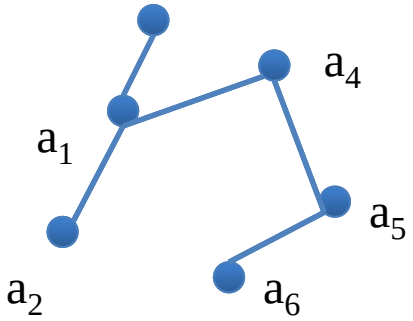
Attribute detector



Jointly detect individual motion attributes.

Infer the optimal configuration of attributes ($a_1 \ldots a_M$)

$$\lambda^{\mathrm{T}}\Phi(\mathbf{x},\mathbf{a}) = \sum_{a_j \in \mathcal{V}_a} \lambda_{a_j}^{\mathrm{T}} \phi_1(\mathbf{x}, a_j) + \sum_{(a_j,a_k)\in\mathcal{E}_a} \lambda_{a_j a_k}^{\mathrm{T}} \phi_2(a_j, a_k)$$



Attribute graph

Unary attribute potential

Score attribute label from feature

Pairwise attribute potential

Score pairwise attribute relationship

Attribute model

Motion attribute $a = \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \\ 1 \\ 0 \end{pmatrix}$

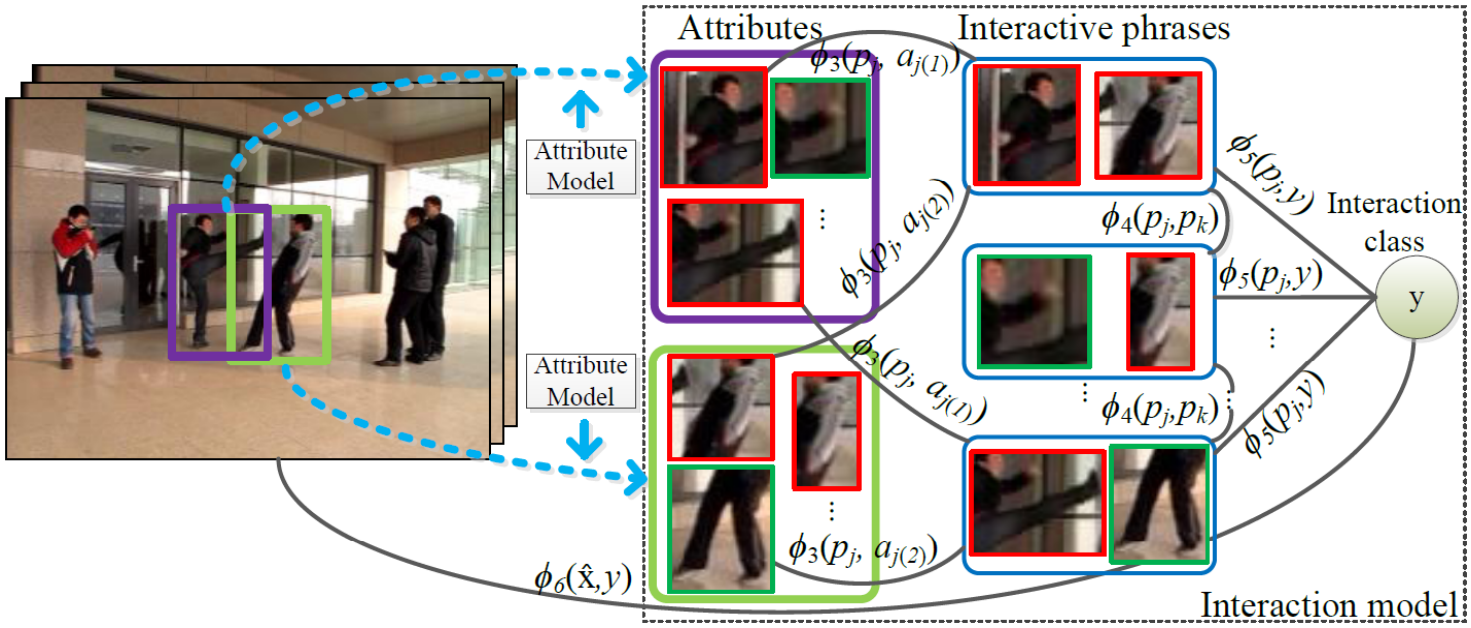
$$a_i = \begin{cases} 0 & \text{if attribute } i \text{ is absent;} \\ 1 & \text{if attribute } i \text{ is present.} \end{cases}$$

id	attributes a_m	
1	still arm	
2	hand stretching out motion	
3	arm chest-level motion	1
4	two arms chest-level motion	
5	arm raising up motion	
6	arm embracing motion	
7	arm free swinging motion	
8	arm intense motion	
9	still leg	
10	leg stepping forward motion	1
11	leg kicking motion	
12	leg stepping back motion	
13	still torso	
14	torso leaning back motion	
15	torso leaning forward motion	1
16	torso bending motion	
17	friendly motion	

=

Interaction model

Objective: learn interactive phrases and infer interaction class
Interactive phrases: motion relationships between people, e.g. relationships between arms, legs, torsos, etc.



Interactive phrases

id	interactive phrases p_j	id of associated attributes $a_{j(1)}, a_{j(2)}$
1	b/w still arms	1,1
2	b/w a chest-level moving arm and a free swinging arm	3,7
3	b/w outstretched hands	2,2
4	b/w raising up arms	5,5
5	b/w embracing arms	6,6
6	b/w a chest-level moving arm and a still arm	3,1
7	b/w two chest-level moving arms and a free swinging arm	4,7
8	b/w free swinging arms	7,7
9	b/w intense moving arms	8,8
10	b/w a chest-level moving arm and a leaning backward torso	3,14
11	b/w two chest-level moving arms and a leaning backward torso	4,14
12	b/w still legs	9,9
13	b/w a stepping forward leg and a stepping backward leg	10,12
14	b/w stepping forward legs	10,10
15	b/w a stepping forward leg and a still leg	10,9
16	b/w a kicking leg and a stepping backward leg	11,12
17	b/w a bending torso and a still torso	16,13
18	b/w a leaning forward torso and a leaning backward torso	15,14
19	b/w leaning forward torsos	15,15
20	b/w leaning backward torsos	14,14
21	b/w a leaning forward torso and a still torso	15,13
22	b/w still torsos	13,13
23	cooperative interaction	17,17

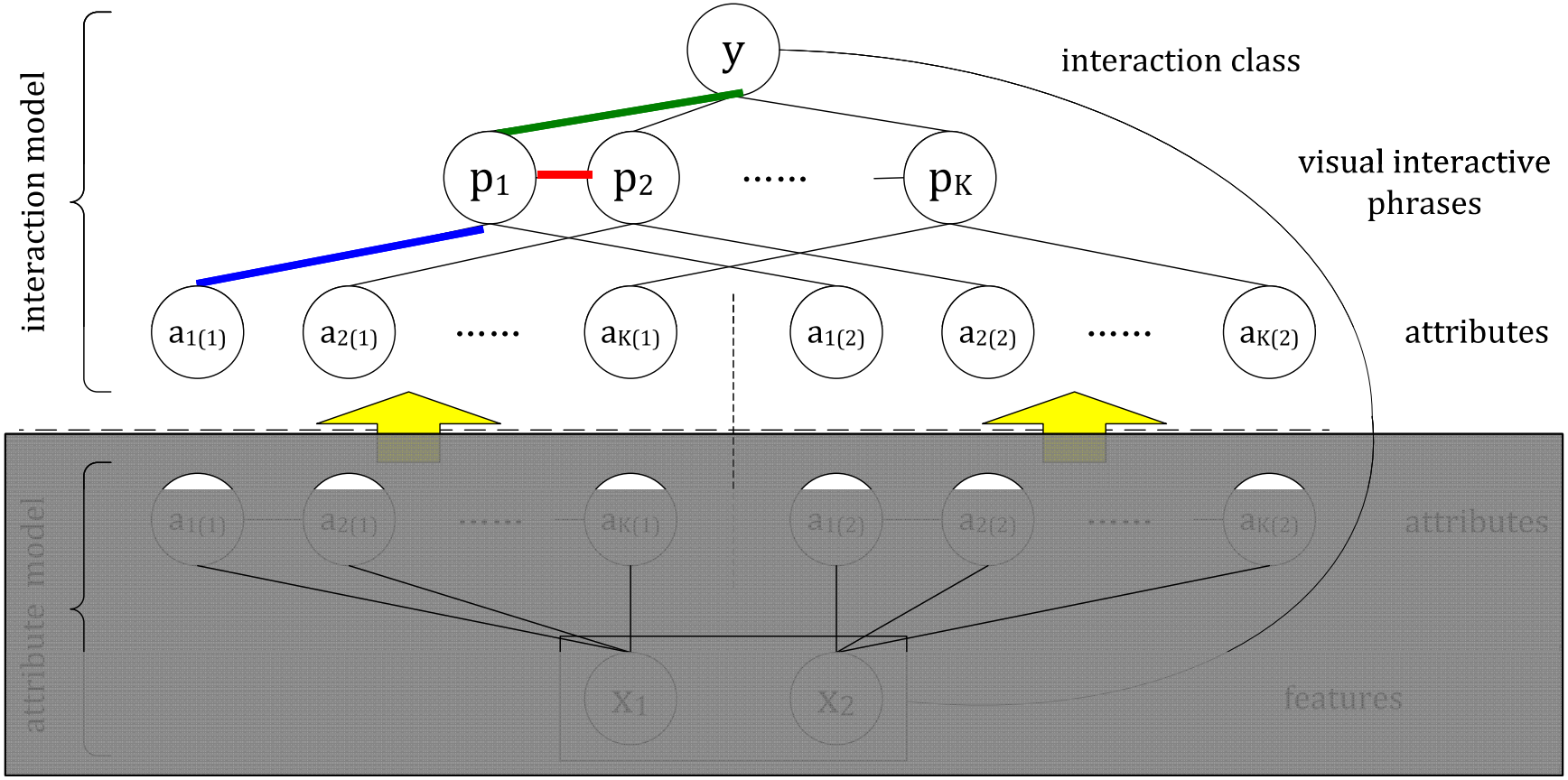
Attributes

Person 1	Person 2
Still arm	Still arm
Stepping forward leg/still leg	Still leg/stepping forward leg

Interactive phrases

$$p = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \qquad p_j = \begin{cases} 1 & \text{if j-th phrase is present} \\ 0 & \text{otherwise} \end{cases}$$

Latent variable, learned from data
mid-level feature, used for inferring interaction class

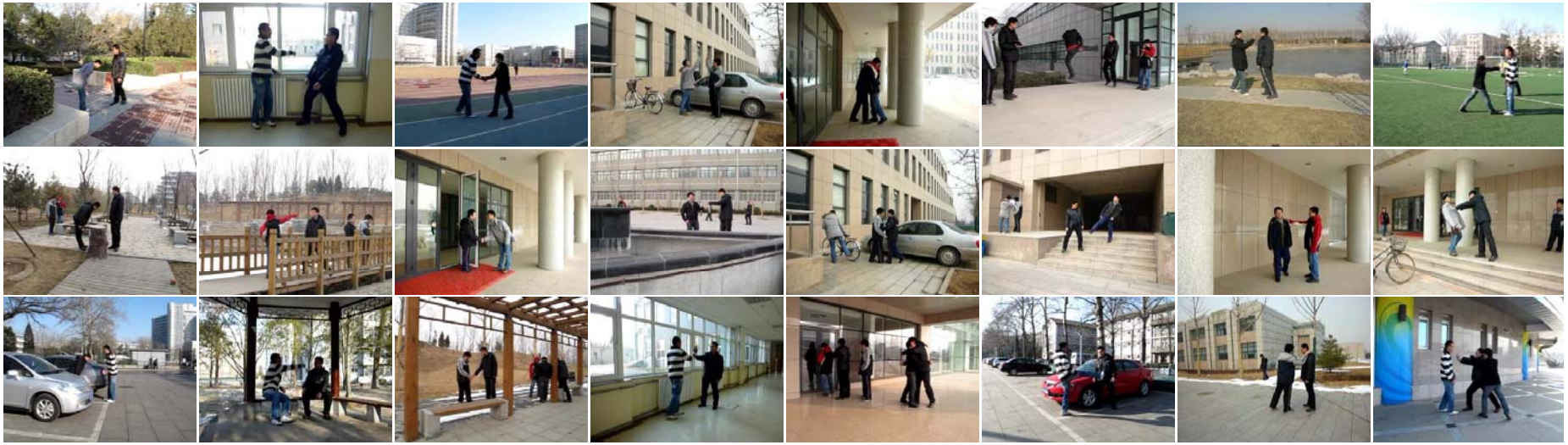


$$f_{\mathbf{w}}(\hat{\mathbf{x}}, \hat{\mathbf{a}}, y) = \max_{\mathbf{p}} \mathbf{w}^{\mathrm{T}} \Phi(\hat{\mathbf{x}}, \hat{\mathbf{a}}, \mathbf{p}, y)$$

$$\mathbf{w}^{\mathrm{T}} \Phi(\hat{\mathbf{x}}, \hat{\mathbf{a}}, \mathbf{p}, y) = \sum_{p_j \in \mathcal{V}_p} \sum_{i=1}^{2} \mathbf{w}_{p_j a_{j(i)}}^{\mathrm{T}} \phi_3(p_j, a_{j(i)}) + \sum_{p_j \in \mathcal{V}_p} \mathbf{w}_{p_j y}^{\mathrm{T}} \phi_4(p_j, y)$$
$$+ \sum_{(p_j, p_k) \in \mathcal{E}_p} \mathbf{w}_{p_j p_k}^{\mathrm{T}} \phi_5(p_j, p_k) + \mathbf{w}_{\hat{\mathbf{x}} y}^{\mathrm{T}} \phi_6(\hat{\mathbf{x}}, y),$$

Experiments

- BIT-Interaction dataset
  - 8 classes, 400 videos



bow **boxing** **handshake** **high-five** **hug** **kick** **pat** **push**

- UT-Interaction dataset
  - 6 classes, 60 videos

handshake hug kick point punch push

Results on BIT-Interaction dataset

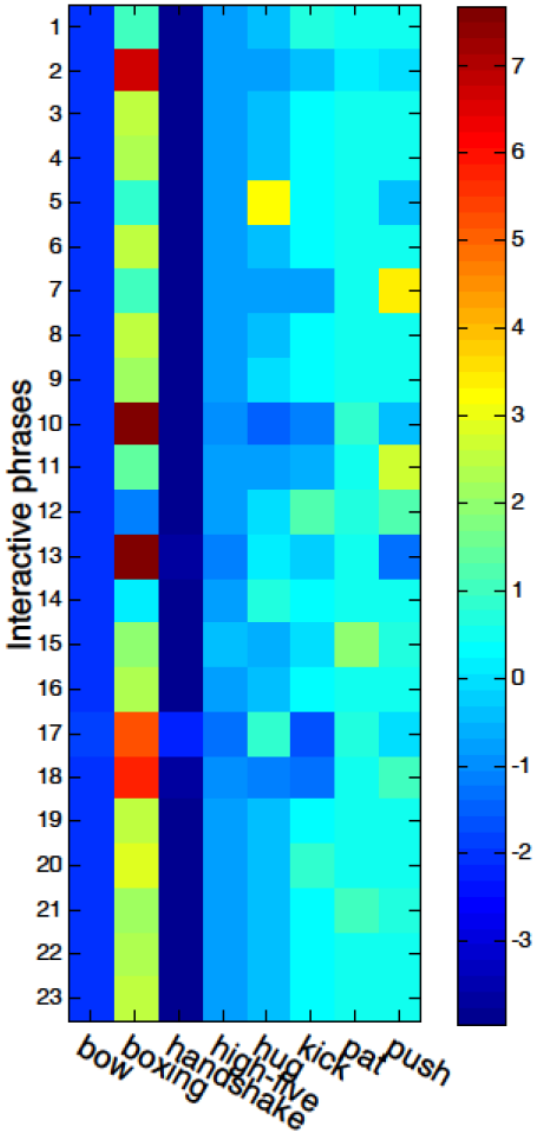
- 8 interaction classes, 400 videos, 23 interactive phrases, 17 motion attributes

	bow	boxing	handshake	high-five	hug	kick	pat	push
bow	0.81	0.06	0.00	0.00	0.00	0.00	0.13	0.00
boxing	0.00	0.81	0.00	0.00	0.06	0.00	0.00	0.13
handshake	0.06	0.00	0.81	0.00	0.13	0.00	0.00	0.00
high-five	0.00	0.06	0.00	0.94	0.00	0.00	0.00	0.00
hug	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.06
kick	0.00	0.06	0.00	0.06	0.00	0.81	0.00	0.06
pat	0.00	0.13	0.00	0.06	0.00	0.00	0.81	0.00
push	0.00	0.06	0.00	0.06	0.00	0.00	0.00	0.88

Confusion matrix of our method
Accuracy = 85.16%

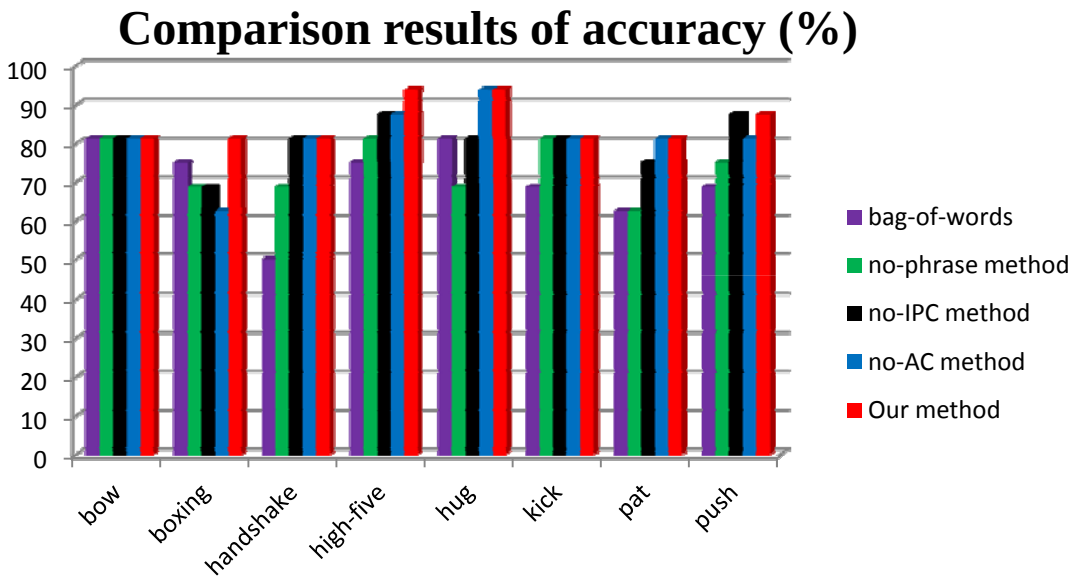
Classification examples of our method

Results on BIT-Interaction dataset



Interactions	Interactive phrases p_j	id
bow	b/w a bending torso and a still torso	17
	b/w still arms	1
	b/w a chest-level moving arm and a free swinging arm	2
boxing	b/w a chest-level moving arm and a leaning backward torso	10
	b/w a stepping forward leg and a stepping backward leg	13
	b/w a chest-level moving arm and a free swinging arm	2
handshake	b/w a bending torso and a still torso	17
	b/w outstretched hands	3
	b/w a leaning forward torso and a leaning backward torso	18
high-five	b/w a stepping forward leg and a still leg	15
	b/w raising up arms	4
	b/w outstretched hands	3
hug	b/w embracing arms	5
	b/w a bending torso and a still torso	17
	b/w stepping forward legs	14
kick	b/w still legs	12
	b/w leaning backward torsos	20
	b/w a kicking leg and a stepping backward leg	16
pat	b/w a stepping forward leg and a still leg	15
	b/w a leaning forward torso and a still torso	21
	b/w still legs	12
push	b/w two chest-level moving arms and a free swinging arm	7
	b/w two chest-level moving arms and a leaning backward torso	11
	b/w a leaning forward torso and a leaning backward torso	18

Results on BIT-Interaction dataset



Recognition accuracy (%) of methods

Methods	Overall
bag-of-words	70.31
no-phrase method	73.43
no-IPC method	80.47
no-AC method	81.25
Our method	**85.16**

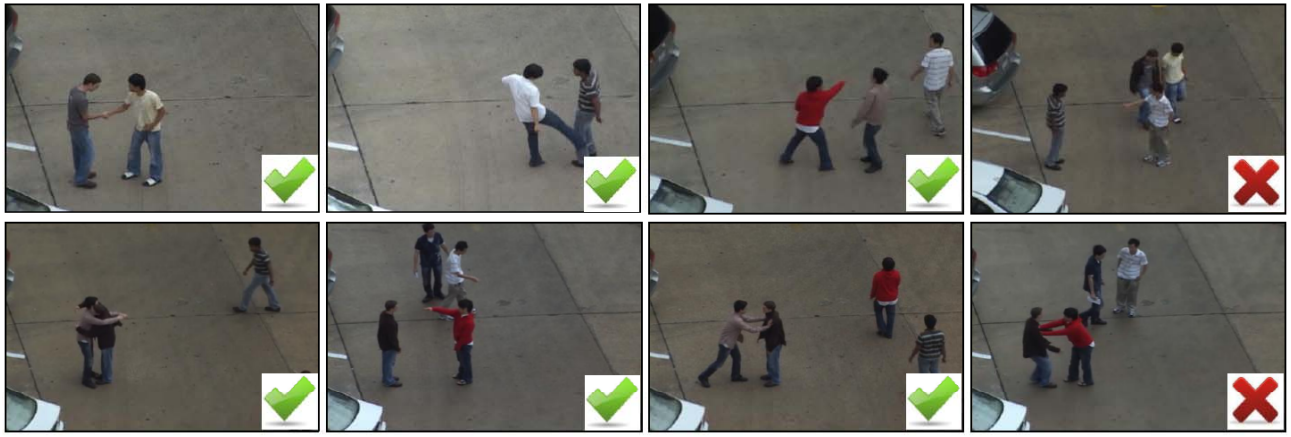
No-phrase method: remove phrase layer from the full model
No-IPC method: remove phrase connection component from the full model
No-AC method: remove attribute connection component from the full model

Results on UT-Interaction dataset

- 6 interaction classes, 60 videos, 23 interactive phrases, 16 motion attributes

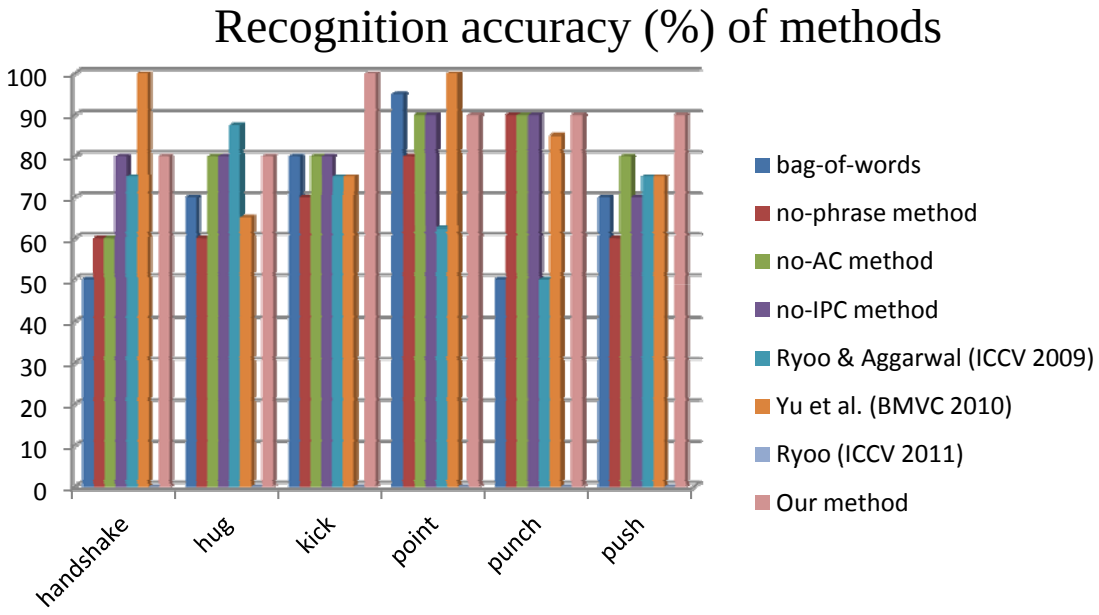
	handshake	hug	kick	point	punch	push
handshake	0.80	0.00	0.10	0.00	0.00	0.10
hug	0.00	0.80	0.20	0.00	0.00	0.00
kick	0.00	0.00	1.00	0.00	0.00	0.00
point	0.00	0.00	0.00	0.90	0.10	0.00
punch	0.00	0.00	0.00	0.00	0.90	0.10
push	0.00	0.00	0.10	0.00	0.00	0.90

Confusion matrix of our method
Accuracy = 88.33%



Classification examples of our method

Results on UT-Interaction dataset



Recognition accuracy (%) of methods

Methods	Overall
bag-of-words	68.33
no-phrase method	70
no-AC method	80
no-IPC method	81.67
Ryoo & Aggarwal (ICCV 2009)	70.8
Yu et al.(BMVC 2010)	83.33
Ryoo (ICCV 2011)	85
Our method	**88.33**

[1] Ryoo, M., Aggarwal, J.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV. (2009) 1593–1600

[2] Yu, T.H., Kim, T.K., Cipolla, R.: Real-time action recognition by spatiotemporal semantic and structural forests. In: BMVC. (2010)

[3] Ryoo, M.S.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: ICCV. (2011)

Thank you!

Please email yukong@ece.neu.edu

if you have any questions.