Automatic Collection of Web Video Shots Corresponding to Specific Actions

Do Hang Nga **Keiji Yanai**

The University of Electro-Communications

Tokyo, Japan

References

- **Do Hang Nga and Keiji Yanai: Automatic Construction of an Action Video Shot Database using Web Videos. ICCV. Nov. 2011. pp 527-534**
- Do Hang Nga and Keiji Yanai: Automatic Collection of Web Video Shots Corresponding to Specific Actions using Web Images. CVPR Workshop on Large-Scale Video Search and Mining (LSVSM). Jun. 2012.
- **Demo page:**

  http://mm.cs.uec.ac.jp/webvideo/

Outline

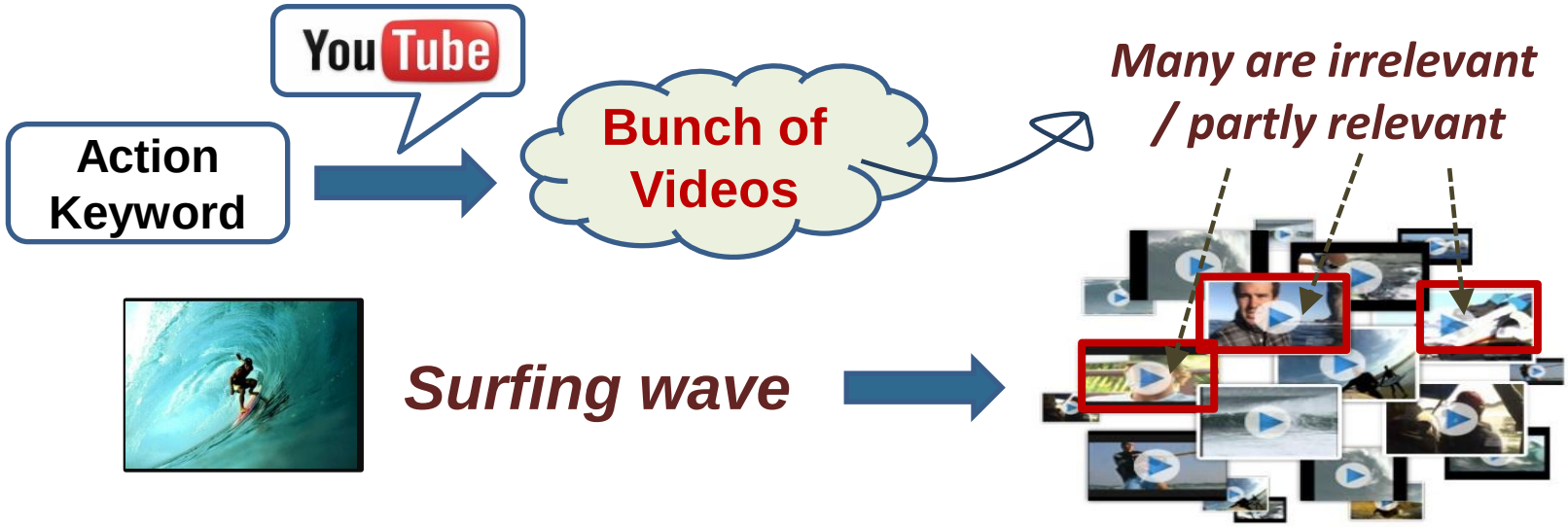
- Motivation
- Objective
- Related work
- Proposed Method
- Experiments & Results
- Conclusion & Future works

Outline

- **Motivation**
- **Objective**
- **Related Work**
- Proposed Method
- Experiments & Results
- Conclusion & Future works

Motivation

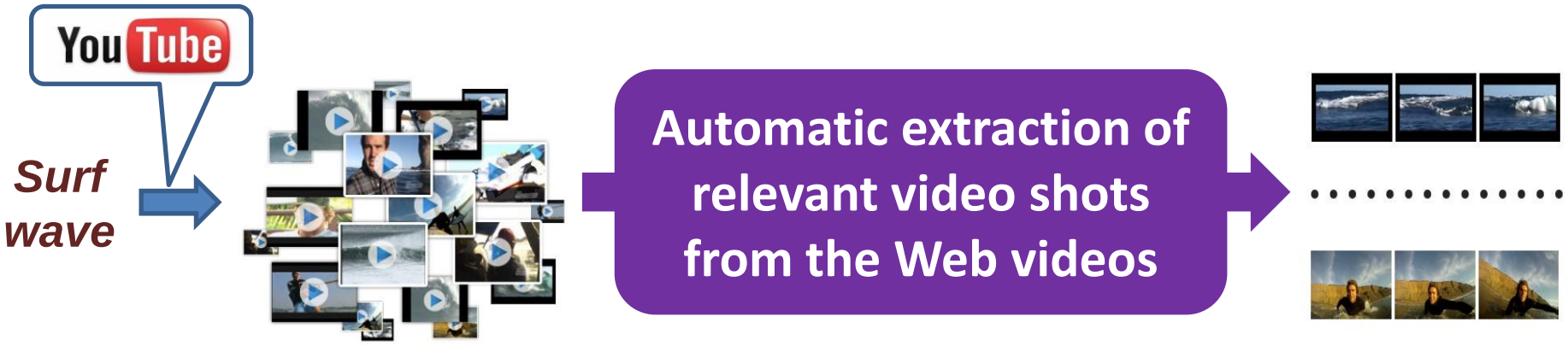
- **Web data source**: huge + free, *but* **noisy**



- ***Web videos based action database construction***: *extremely* ***time-consuming work***

Objective

Unsupervised construction of an action video database



Examples of Results

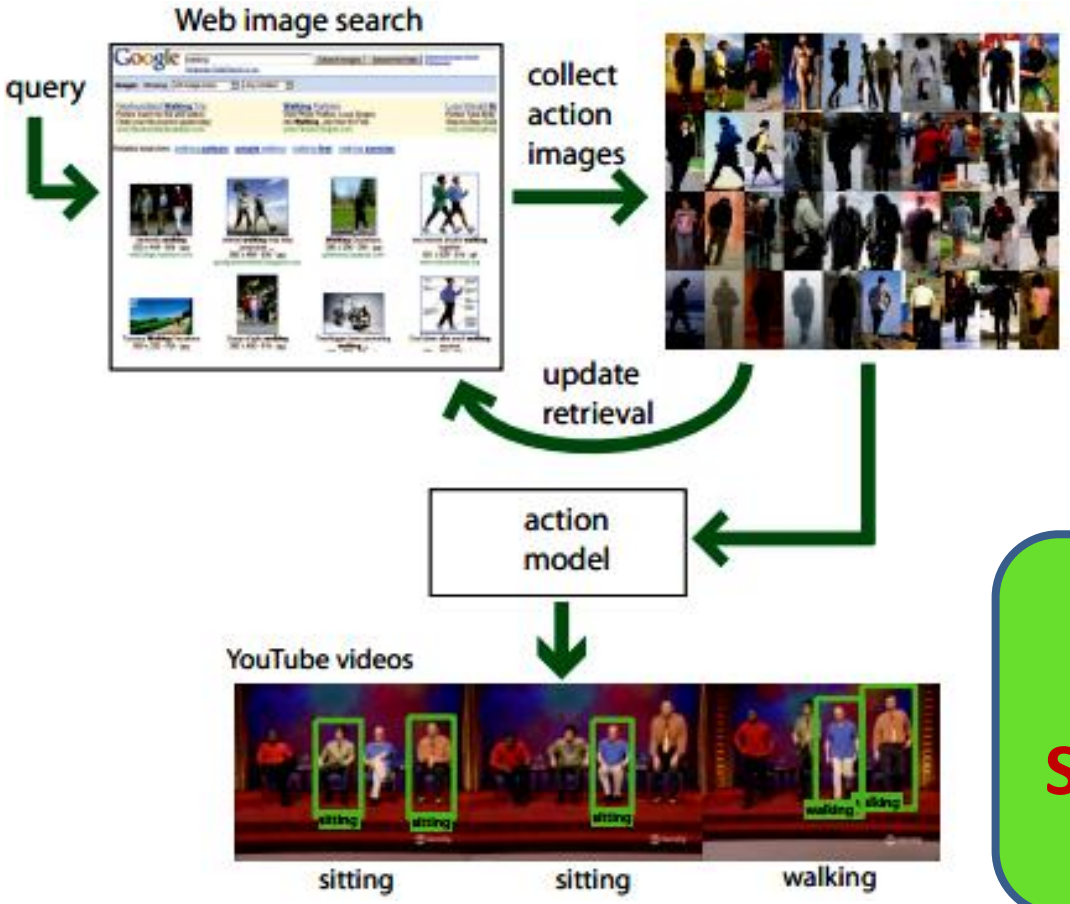
Dancing+flamenco

Surfing+wave

Related work

N. I. Cinbis, R. G. Cinbis and S. Sclaroff:

"Learning actions from the web", ICCV2009



Cinbis et.al.

Web images
Static features

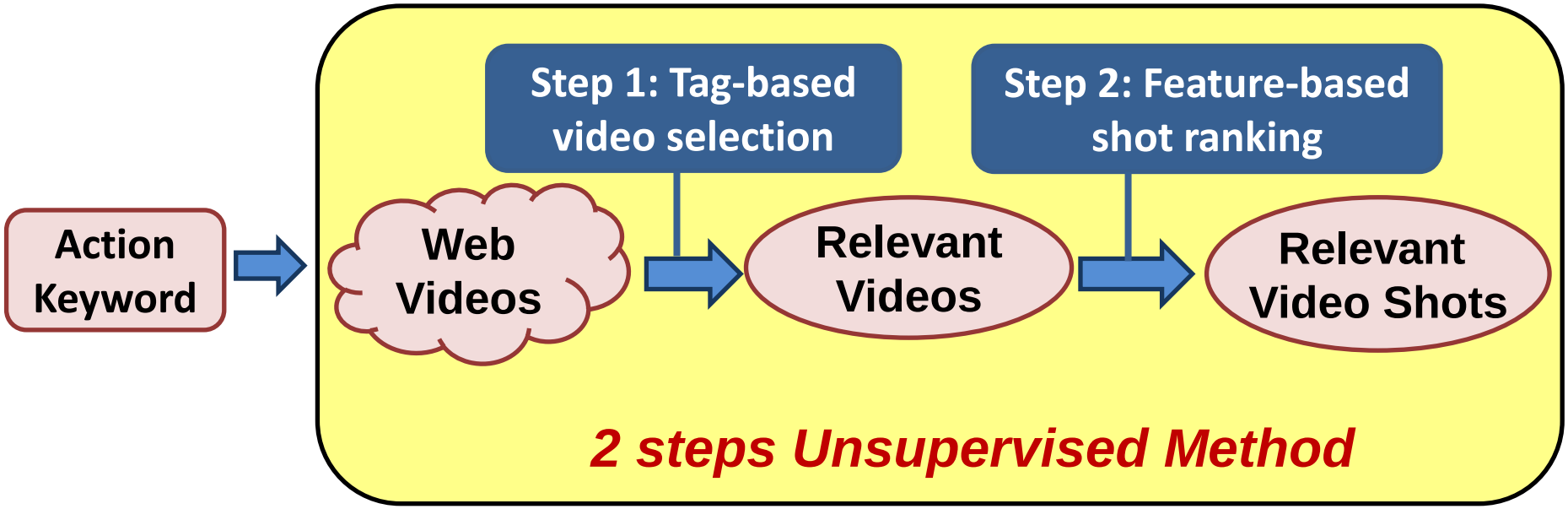
Proposed method

Web videos＋images
Spatio-temporal features
Metadata (Tags)

Outline

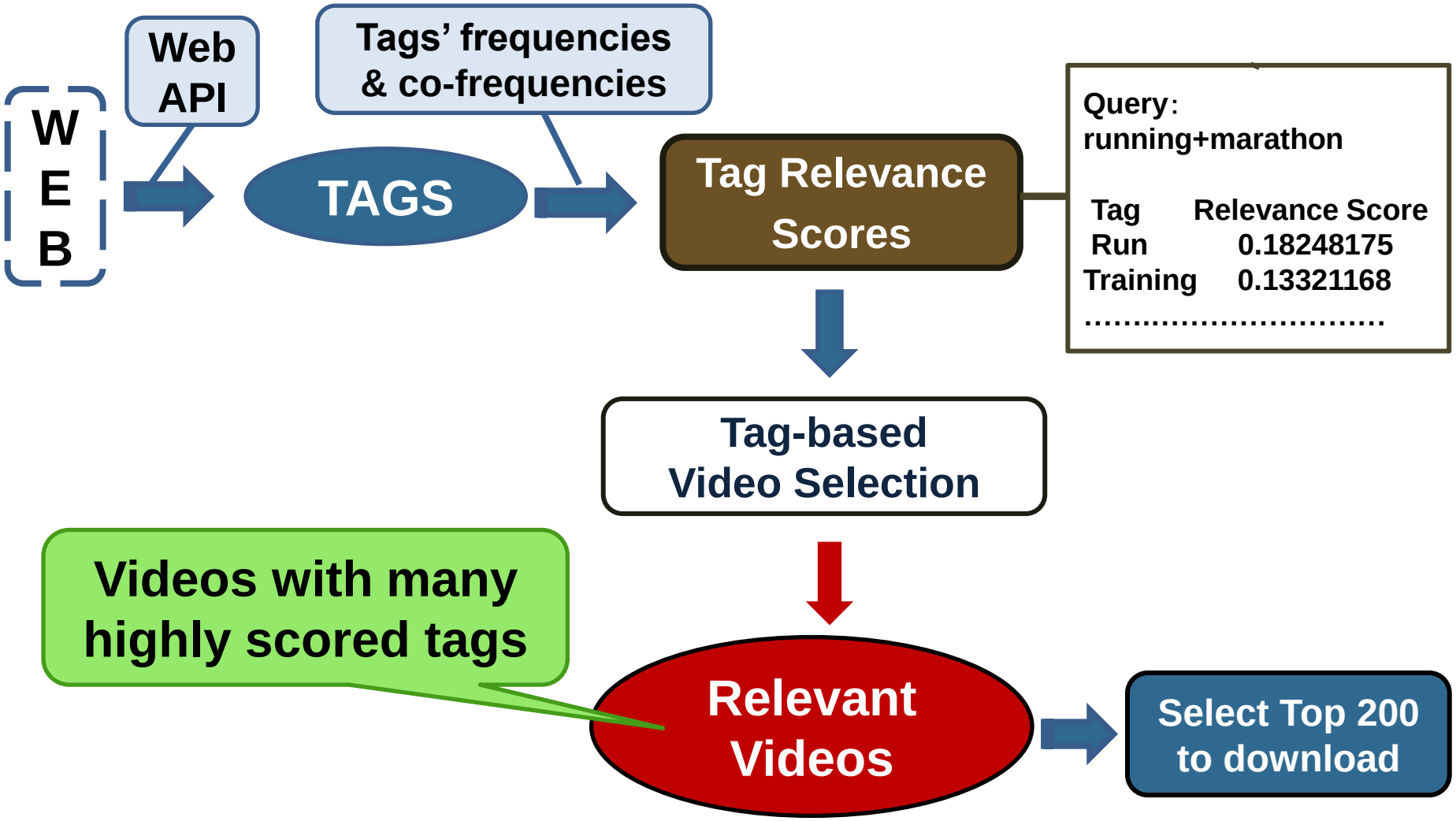
- Motivation
- Objective
- Related work
- **Proposed Method**
- **Experiments & Results**
- Conclusion
- Future works

Proposed Method

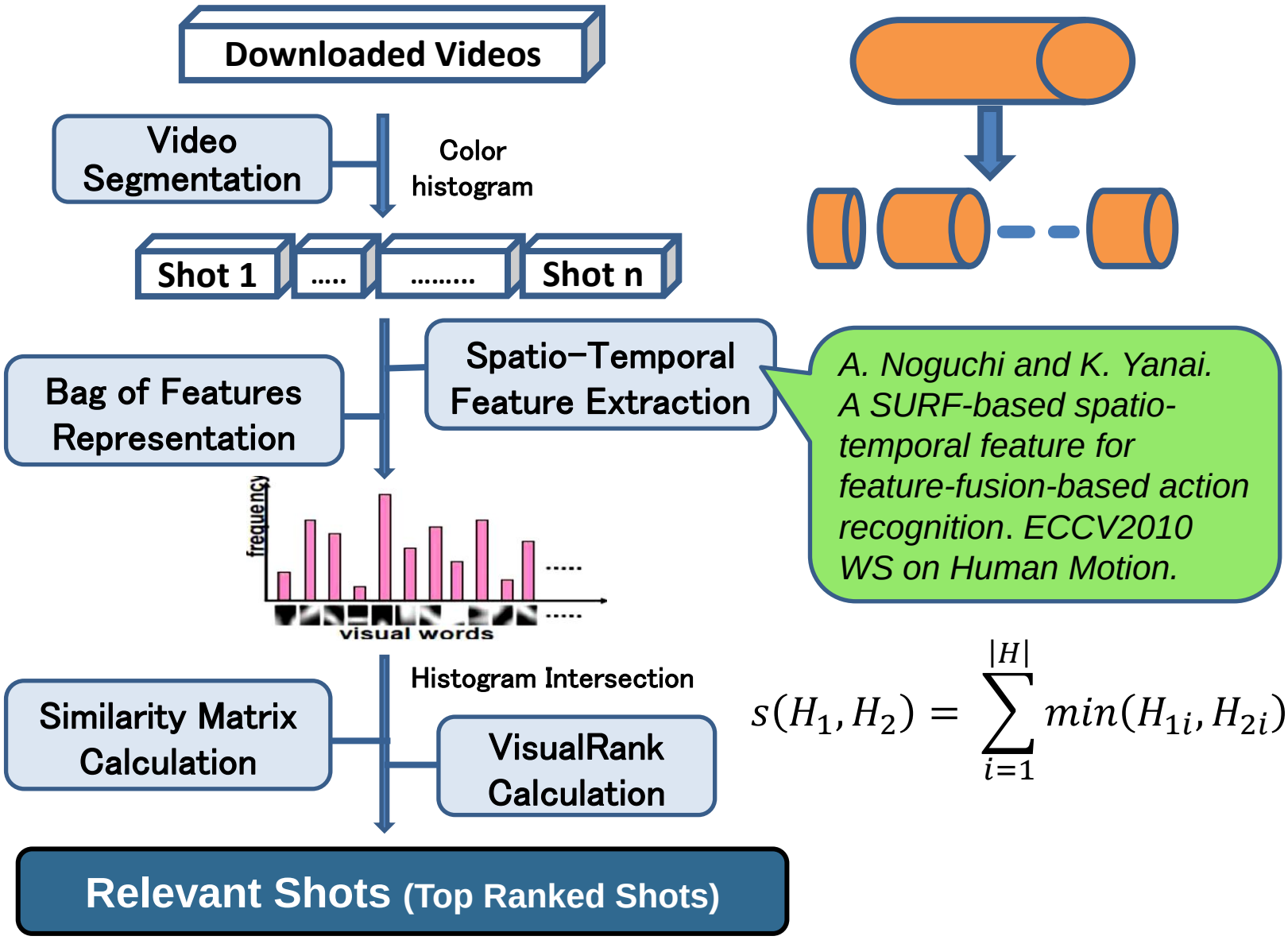


✓ Experimented 100 kinds of human actions → *e.g. Grill fish, Ride bicycle*

Tag-Based Video Selection

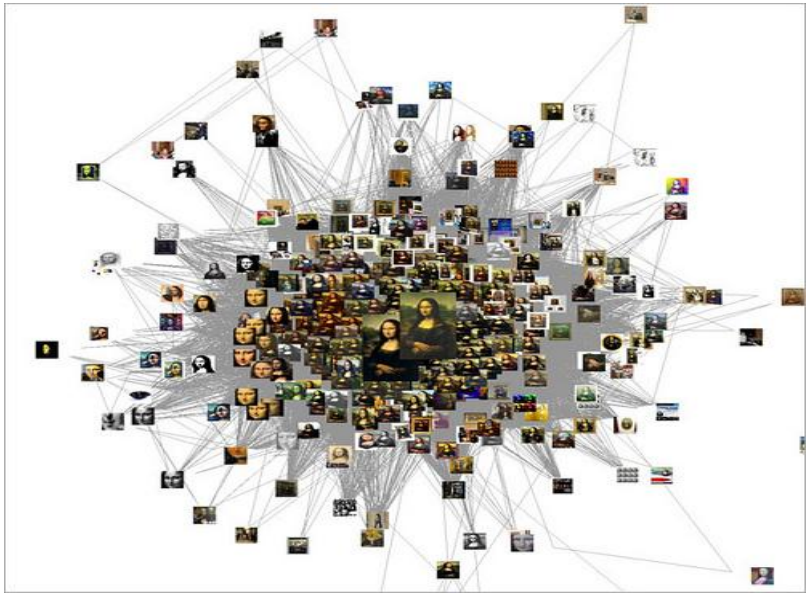


Feature-based shot ranking



Shot ranking by VisualRank

- VisualRank(*):



- **How to calculate**

$$r = dS^{*}r + (1-d)p$$

Damping vector

(*) *Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. PAMI, 30(11):1870–1890, 2008.*

Shot ranking by VisualRank

- Image ranking by VisualRank:

$$r = dS^{*}r + (1-d)p, where\ p = \left[\frac{1}{n}\right]_{n\times 1}$$

- **Our method: damping value is high if shot is from video with high tag scores**

$$p_i = \begin{cases} {}^{1}/_{k} & (i \le k) \\ 0, & (i > k) \end{cases}$$

k top score shots
k~1000

Experiments & Results

✓**Experimented 100 kinds of human actions**

✓**36.6% prec@100 for 100 actions**

79.8% for the best 10

e.g. Painting picture
Playing piano
Counting money
Riding bicycle

- How to evaluate results

prec@100 = number of relevant shots in top 100 shots

Results

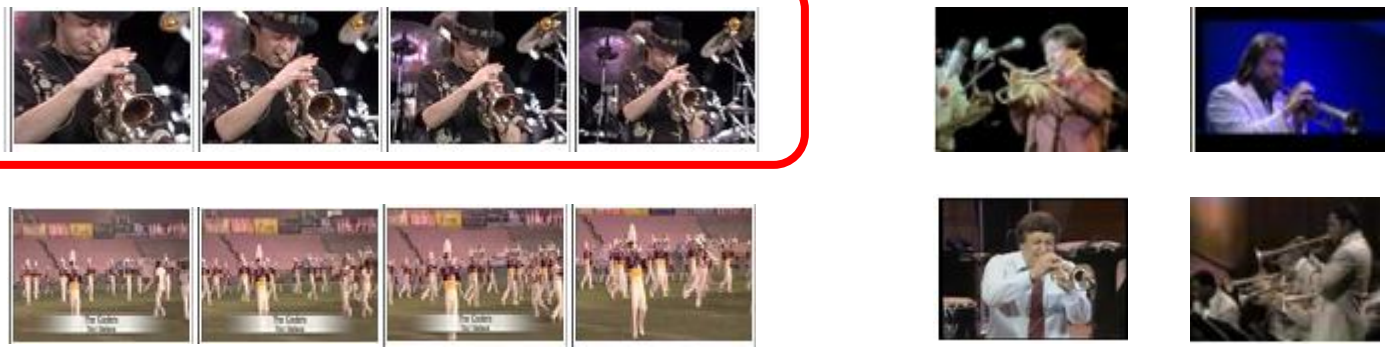
See relevant shot extraction **results** of **100 actions** at **http://mm.cs.uec.ac.jp/webvideo/**

Improve Shot Ranking Step by Exploiting Web Images

- **Idea : Bias shots with many scenes similar to related images**

$$r = dS^*r + (1-d)p$$

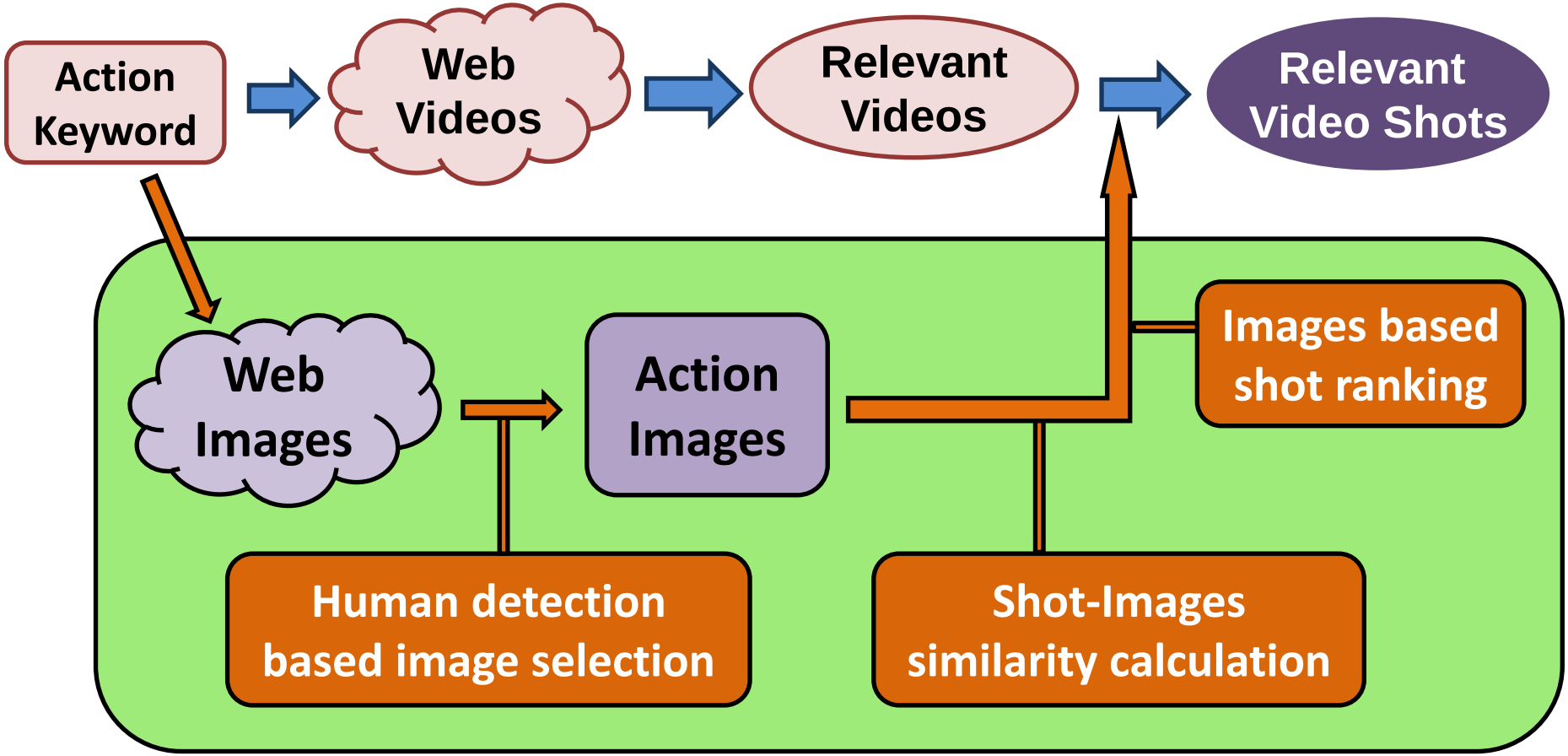
Play trumpet



Video Shots

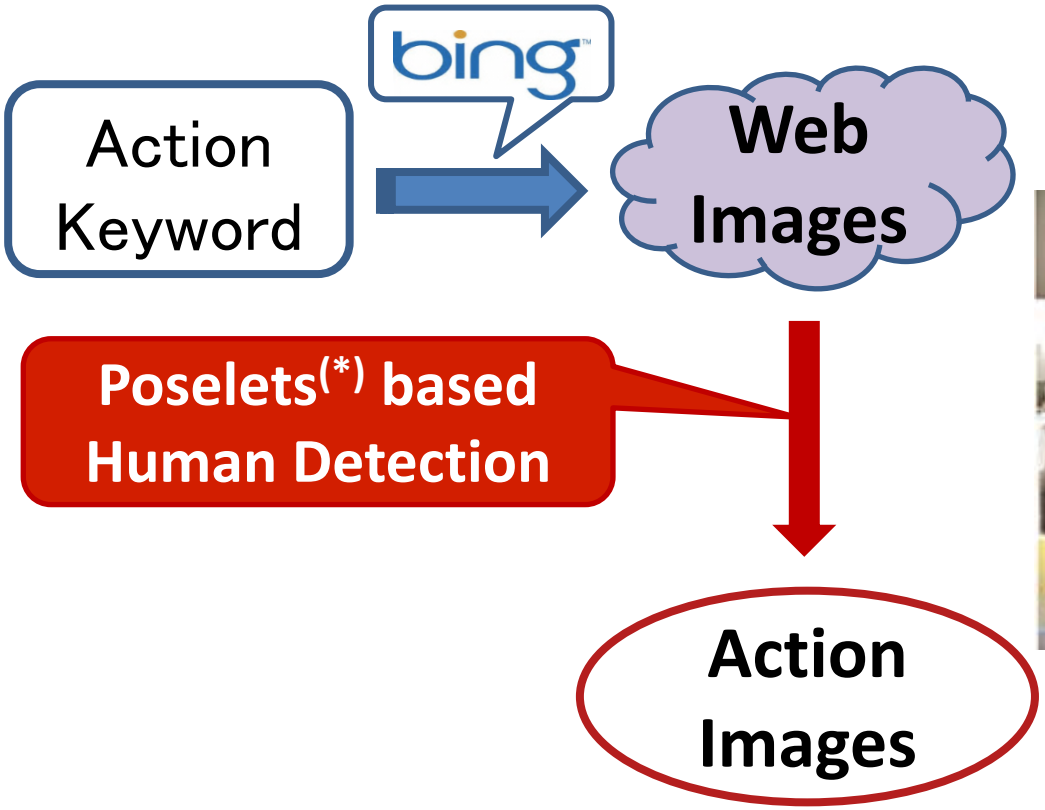
Action Images

Introducing Web Images



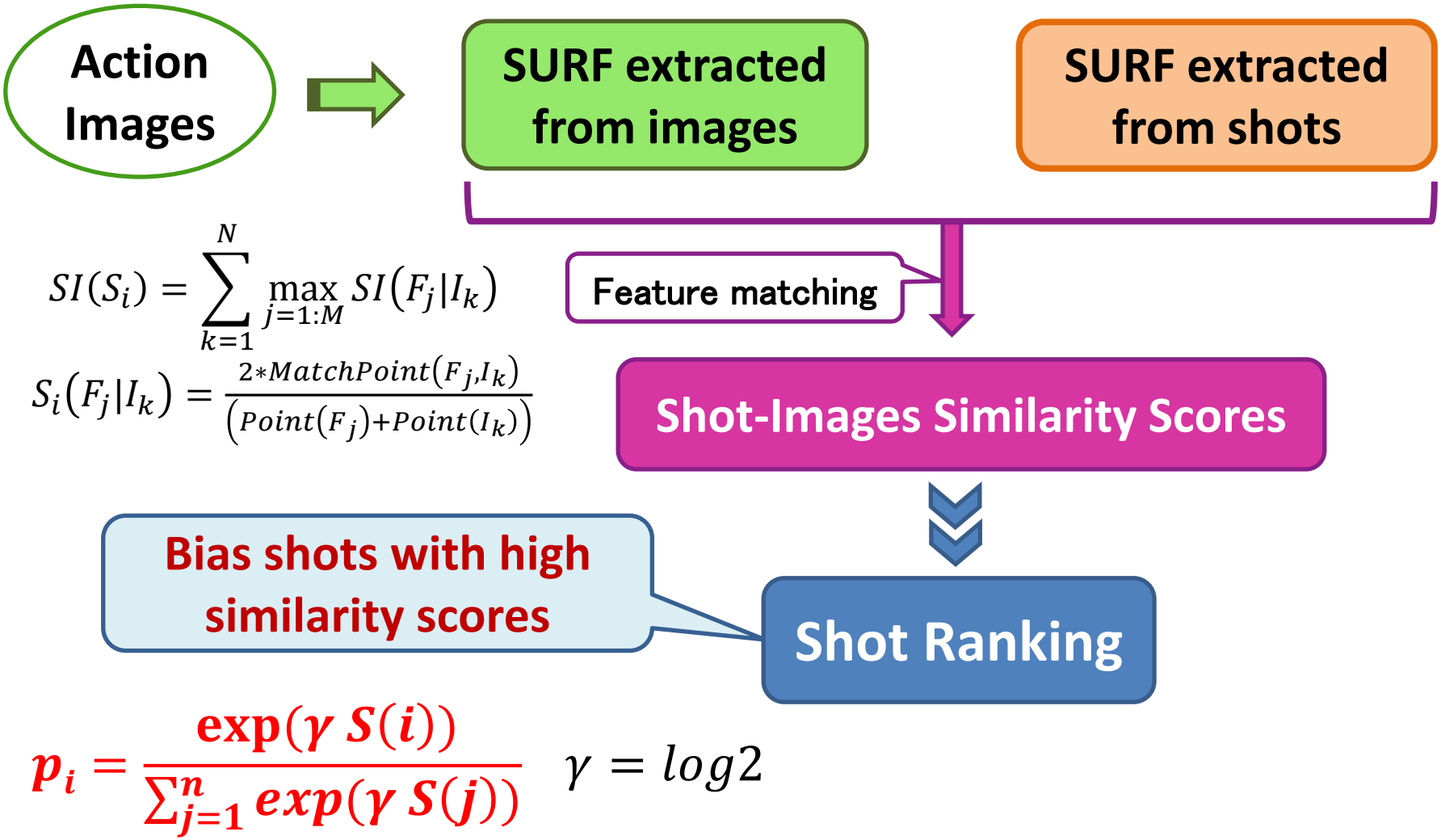
Extended part

Action Image Collection



(*) *Lubomir Bourdev, Jitendra Malik, Poselets: Body Parts Detectors Trained using 3D Human Pose Annotations, ICCV 2009*

Shot-Images similarity based Shot Ranking



Efficiency of introducing Web Images

- ***Dataset**: failed categories in the previous work*
  - 28 human action categories (Prec@100 < 20%)
  - 8 non-human action categories (Prec@100 < 15%)
- ***Evaluation**: percentage of relevant shots over 100 top ranked shots (Precision@100)*
- ***Results**:*
  - human actions: 10.1% → 16.3% **(6.2%↑)**
  - non-human actions: 2% → 18.6% **(16.6%↑)**

Improved categories

Top 5 actions in terms of improvement

(1): previous work (2): this work

Actions	(1)	(2)	*gain*
swim+butterfly	**7**	**31**	**+24**
serve+volleyball	**7**	**31**	**+24**
grill+fish	**5**	**26**	**+21**
squat	**19**	**32**	**+13**
bake+bread	**6**	**18**	**+12**

Degraded categories

Worst 5 actions in terms of improvement

(1): previous work (2): this work

Actions	(1)	(2)	***gain***
slap+face	**20**	**13**	**-7**
wash+clothes	**15**	**10**	**-5**
drink+coffee	**14**	**9**	**-5**
boil+egg	**9**	**6**	**-3**
slice+apple	**5**	**2**	**-3**

Why worsen some categories?

(1) human-detection-based image selection selects very few relevant images

Top selected Web Images for ‘slap+face’

Why worsen some categories?

(2) shots-images similarity calculation method is not effective

- gaps between selected images and downloaded videos

Selected Web Images (washing+clothes) Downloaded Videos

Conclusion & Future works

- ***Proposed a framework of automatically extracting action video shots from the Web***
- ***Applied Web action images to improve the framework***
- **Future works:**

  - improve video selection step

  - try more features

serve+tennis

serve+tennis

[rank 1]

shoot+arrow

shoot+arrow

[rank 1]

snow+falling

snow+falling

[rank 1]

http://www.youtube.com/watch?v=3RtiohB-nd4

airplane+flying

airplane+flying

[rank 1]

http://www.youtube.com/watch?v=VtM0Brg5-sY