LSVSM12

CVPR Workshop 2012:
Large-Scale Video Search and Mining

Automatic Collection of Web Video Shots Corresponding to Specific Actions using Web Images

Do Hang Nga Keiji Yanai

The University of Electro-Communications

Tokyo, Japan

Outline

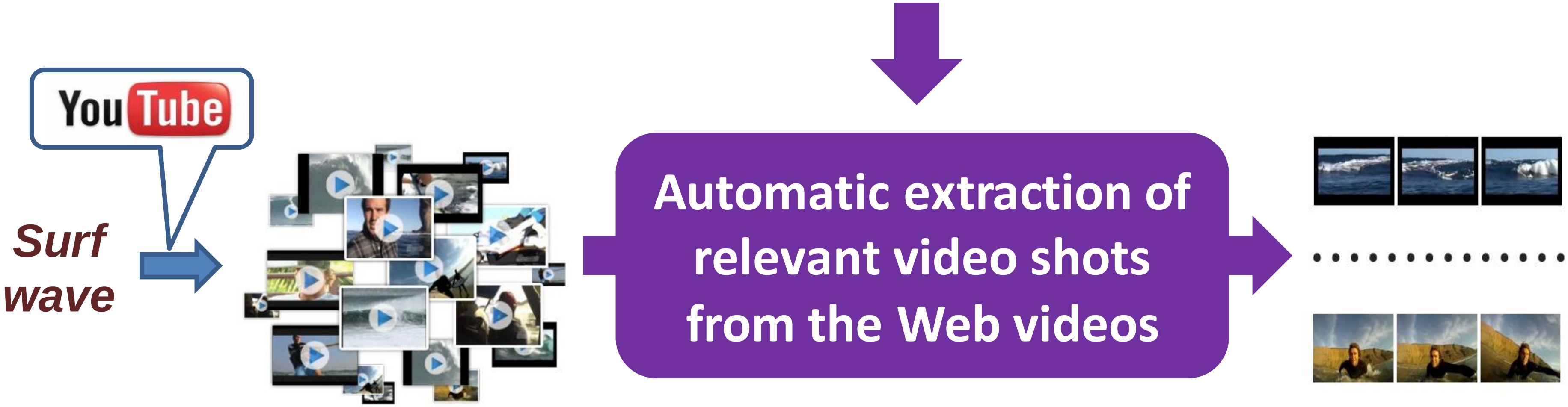
- Motivation & Objective
- Contributions
- Related work
- Previous work
- This work
- Experiments & Results
- Conclusion & Future works

Outline

- **Motivation & Objective**
- **Contributions**
- **Related work**
- Previous work
- This work
- Experiments & Results
- Conclusion & Future works

Motivation & Objective

- **Web data source**: huge + free, *but* **noisy**
- ***Web videos based action database construction***: *extremely* ***time-consuming work***



Automatic extraction of relevant video shots from the Web videos

Unsupervised construction of an action video database

Contributions

- ***Refine our previous approach [ICCV11] by introducing the use of Web action images***

  - propose to select relevant shots based on their similarities with action Web images

- ***Improve results for the failed action categories***

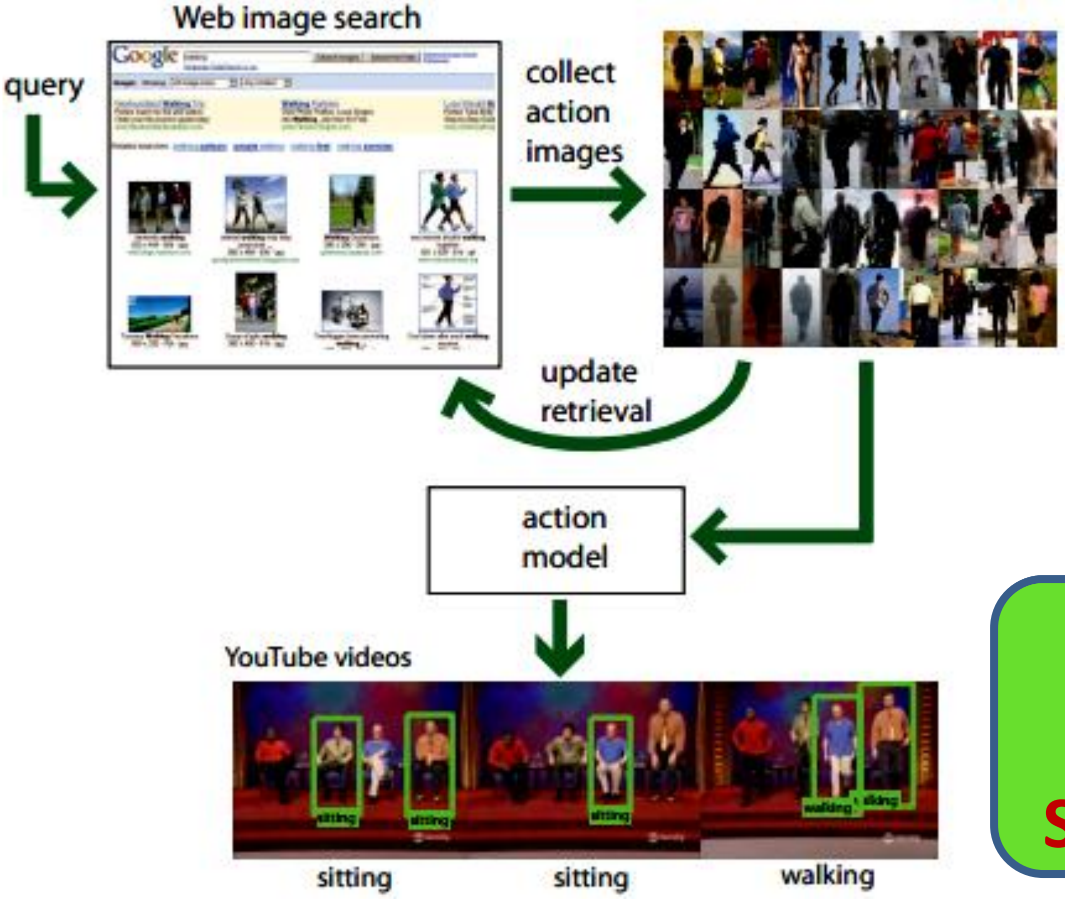
  - 28 human actions: **6%↑**
  - 8 non-human actions: **16%↑**

[ICCV11] Do Hang Nga and Keiji Yanai: Automatic Construction of an Action Video Shot Database using Web Videos. ICCV2011.527-534.

Related work

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

“Learning actions from the web”, ICCV2009



Cinbis et.al.

Web images
Static features

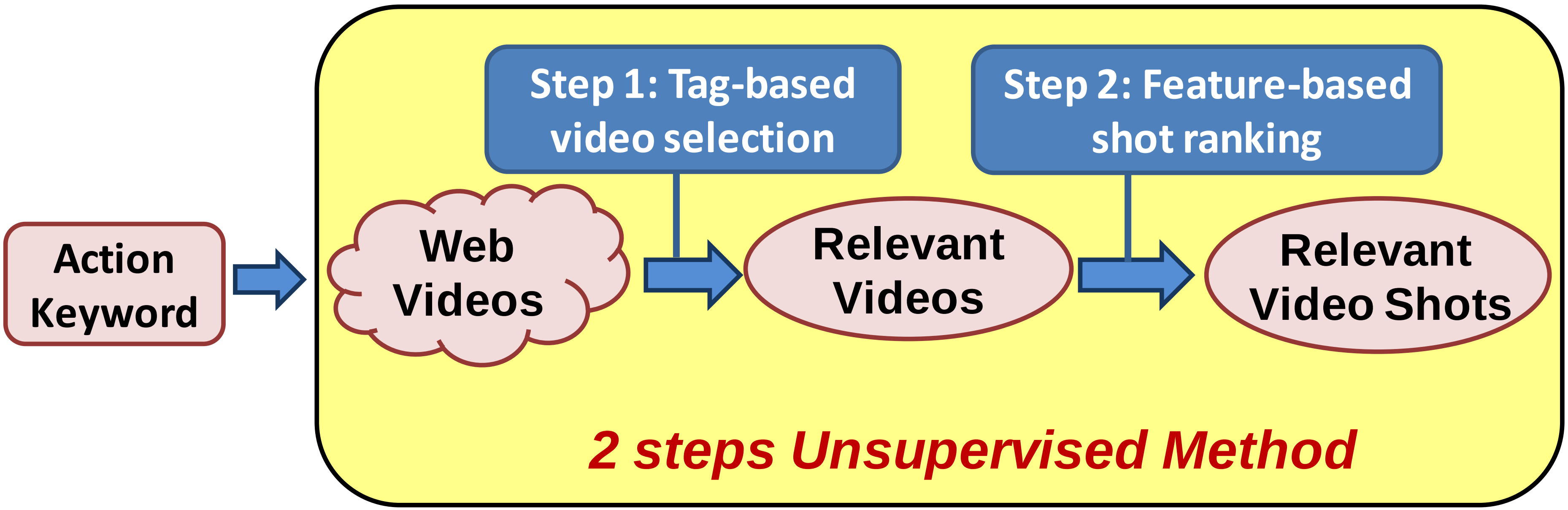
Proposed method

Web videos + images
Spatio-temporal features

Outline

- Objective & Motivation
- Contributions
- Related work
- **Previous Work & its Problems**
- This work
- Experiments & Results
- Conclusion & Future works

Previous work



Action Keyword

Web Videos

Step 1: Tag-based video selection

Relevant Videos

Step 2: Feature-based shot ranking

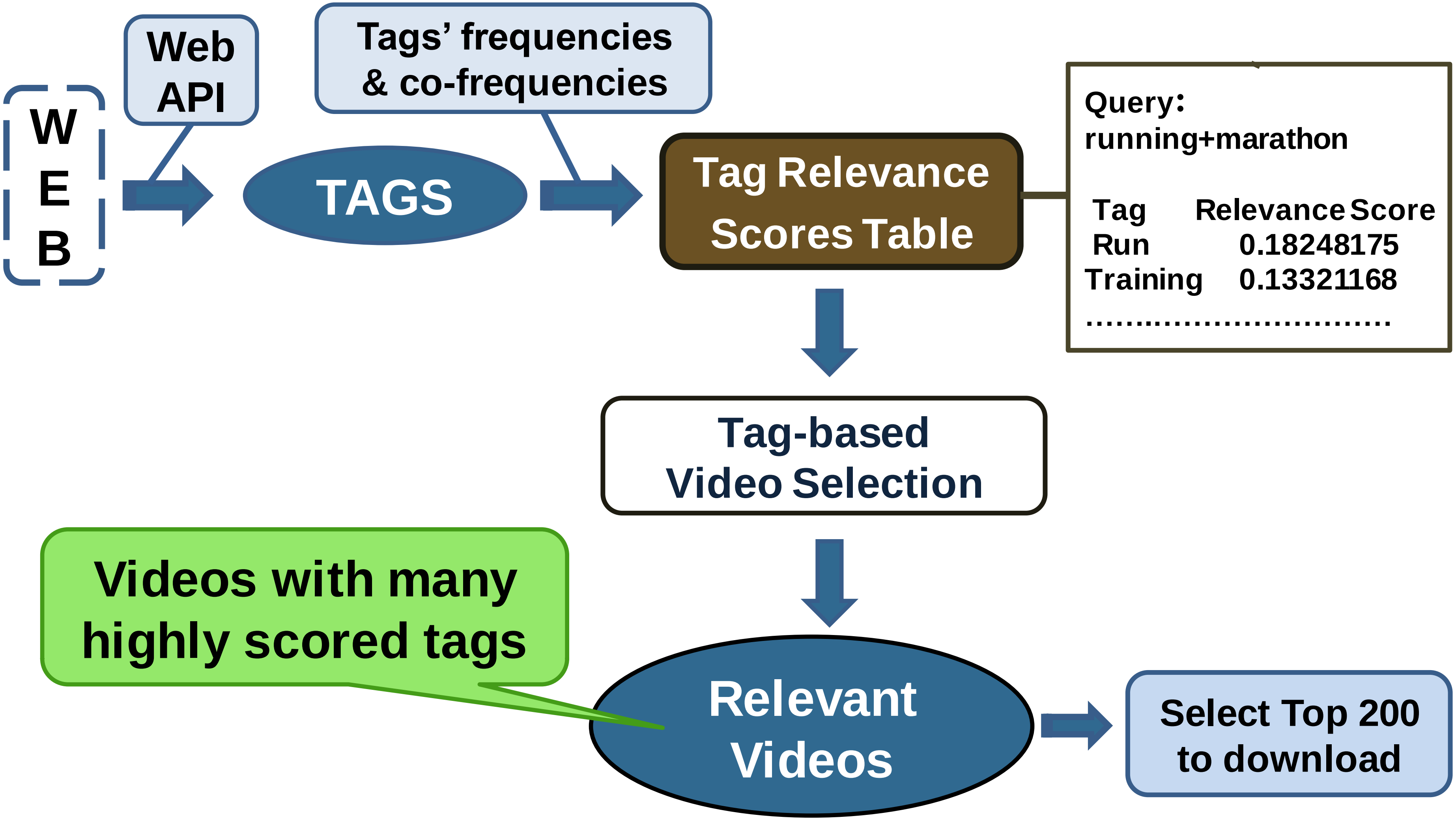
Relevant Video Shots

2 steps Unsupervised Method

✓ **Experimented 100 kinds of human actions** →

e.g. Grill fish
Ride bicycle

Tag-Based Video Selection



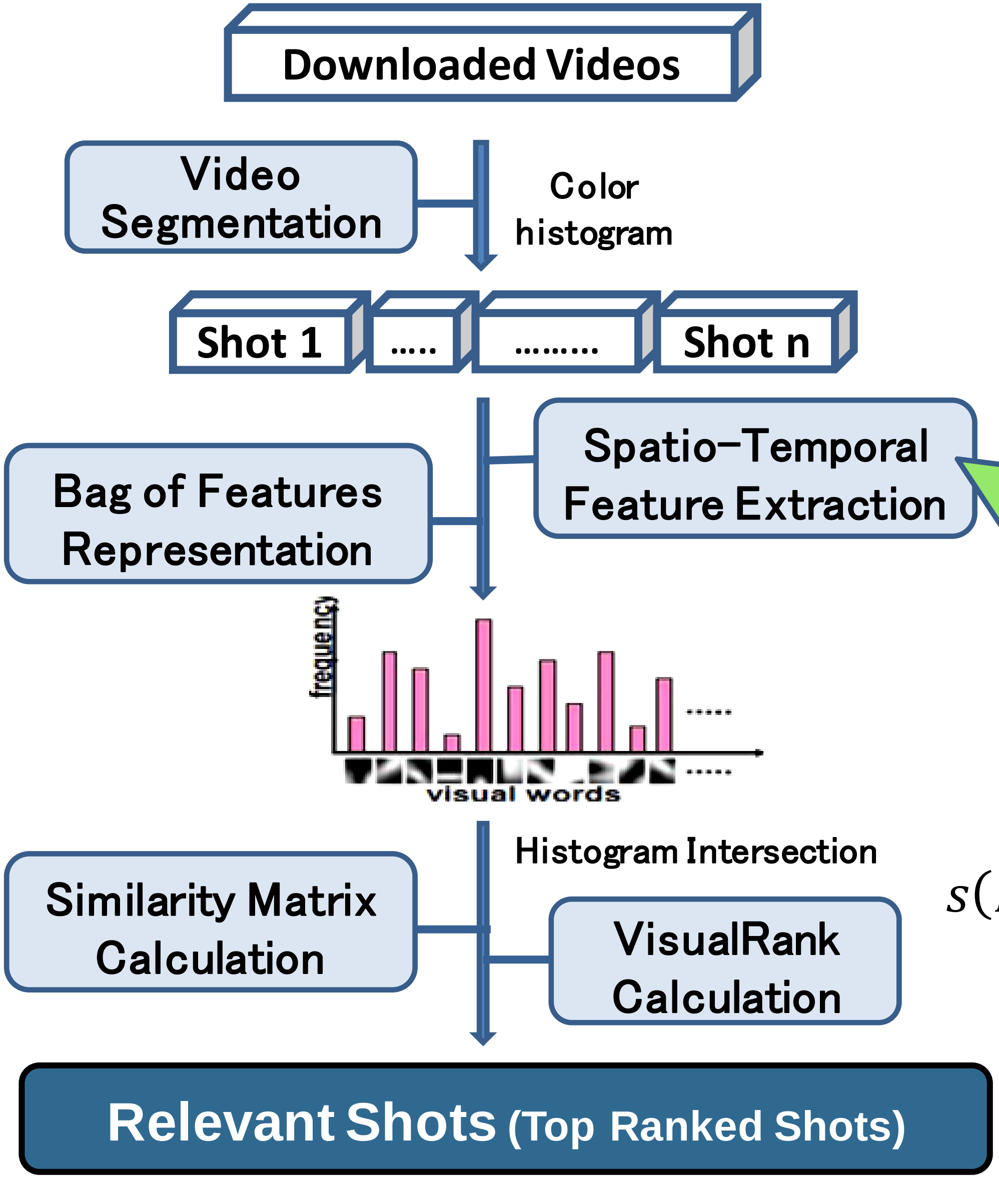
W
E
B
Web API
TAGS
Tags’ frequencies & co-frequencies
Tag Relevance Scores Table
Tag-based Video Selection
Relevant Videos
Select Top 200 to download
Videos with many highly scored tags

Query:
running+marathon

Tag	Relevance Score
Run	0.18248175
Training	0.13321168

……………………………

Feature-based shot ranking



Downloaded Videos

Video Segmentation

Color histogram

Shot 1 .... Shot n

Spatio−Temporal Feature Extraction

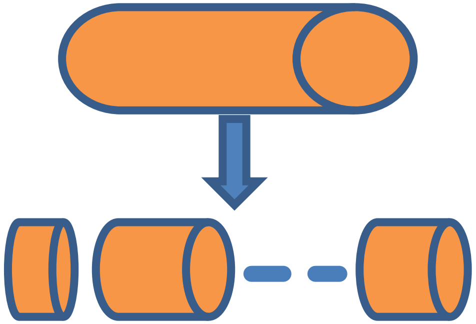
Bag of Features Representation

Histogram Intersection

Similarity Matrix Calculation

VisualRank Calculation

Relevant Shots **(Top Ranked Shots)**



A. Noguchi and K. Yanai. A SURF-based spatio-temporal feature for feature-fusion-based action recognition. ECCV2010 WS on Human Motion.

$$s(H_1,H_2) = \sum_{i=1}^{|H|} min(H_{1i},H_{2i})$$

Feature-based shot ranking

- Shot ranking by VisualRank[*]:

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

- **Our previous work: Bias shots from highly scored videos**

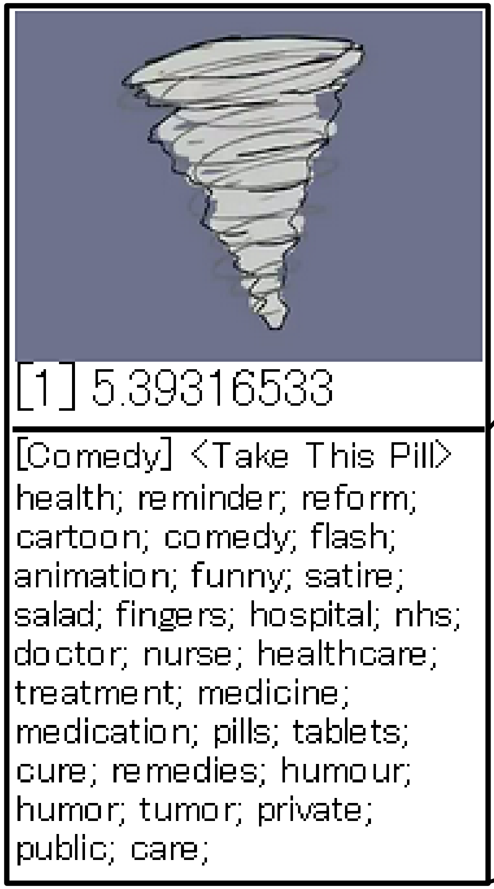
$$p_i = \begin{cases} {}^{1}/_{k} & (i \le k) \\ 0, & (i > k) \end{cases}$$

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

Problems

Exploit only Web Videos and their metadata (tags)

→ Failed in some categories due to noisy tags



[1] 5.39316533

[Comedy] <Take This Pill>
health; reminder; reform; cartoon; comedy; flash; animation; funny; satire; salad; fingers; hospital; nhs; doctor; nurse; healthcare; treatment; medicine; medication; pills; tablets; cure; remedies; humour; humor; tumor; private; public; care;

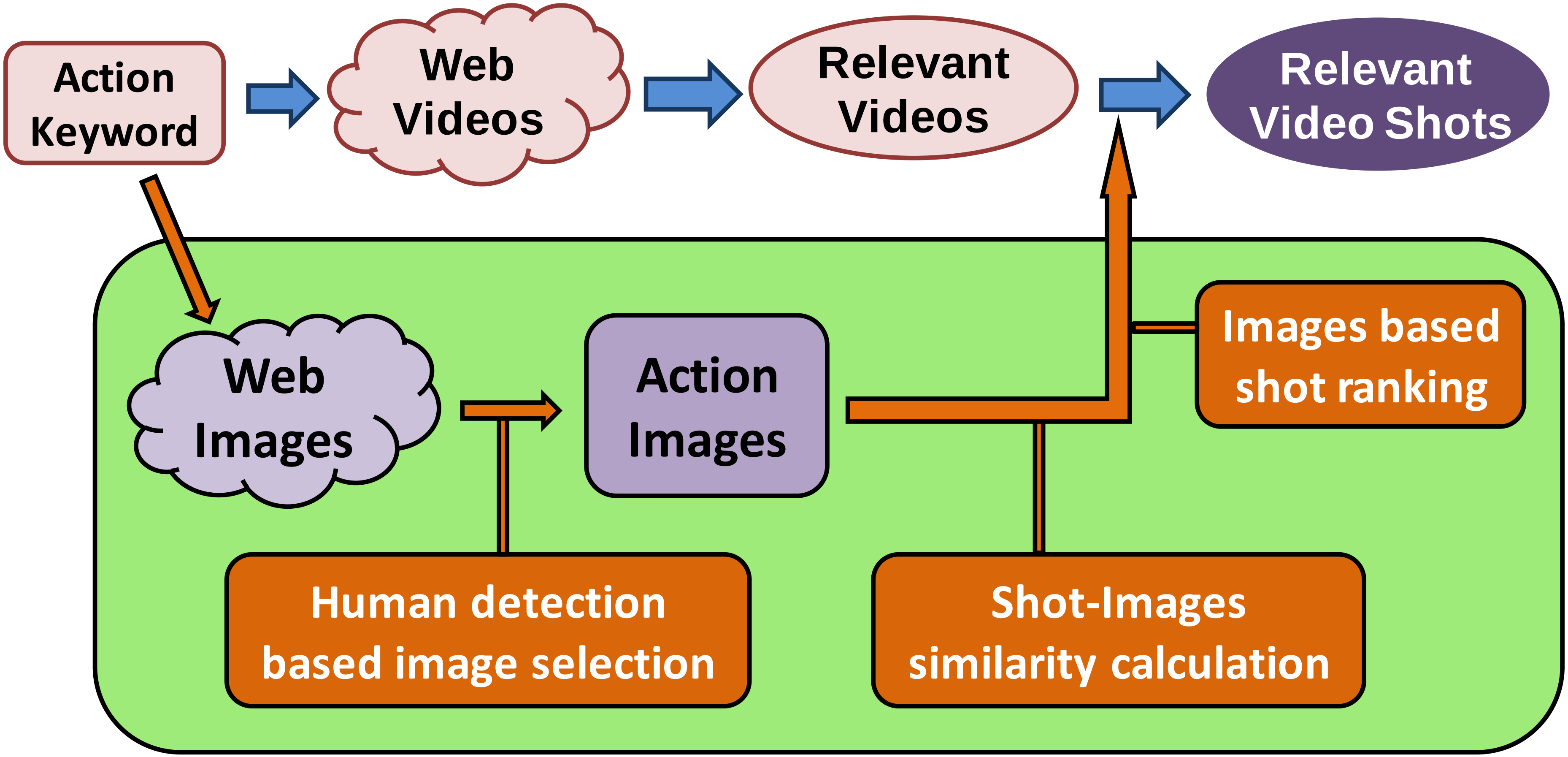
[Comedy] <Take This Pill>
health; **reminder; reform; cartoon; comedy; flash; animation; funny; satire; salad; fingers;** hospital**; nhs;** doctor; nurse; healthcare; treatment; medicine; medication; pills; tablets; cure; remedies; **humour; humor**; tumor; **private; public;** care

The top selected video for ‘take+medicine’ and its tags

Outline

- Motivation & Objective
- Contributions
- Related work
- Previous work
- **This work**
  - Overview of this work
  - The introduction of Web Images
  - Image collection
  - Shot-Images similarity based shot ranking
- Experiments & Results
- Conclusion & Future Works

Overview of this work



Action Keyword

Web Videos

Relevant Videos

Relevant Video Shots

Web Images

Action Images

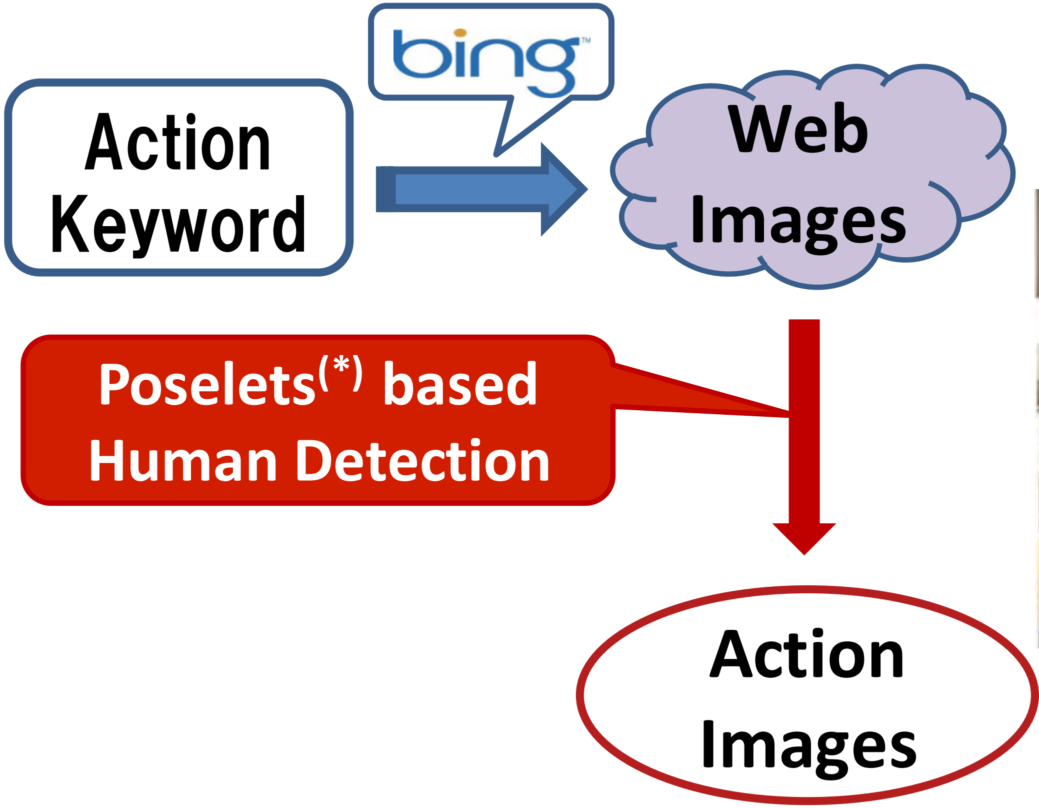
Images based shot ranking

Human detection based image selection

Shot-Images similarity calculation

Extended part

Action Image Collection



Action Keyword

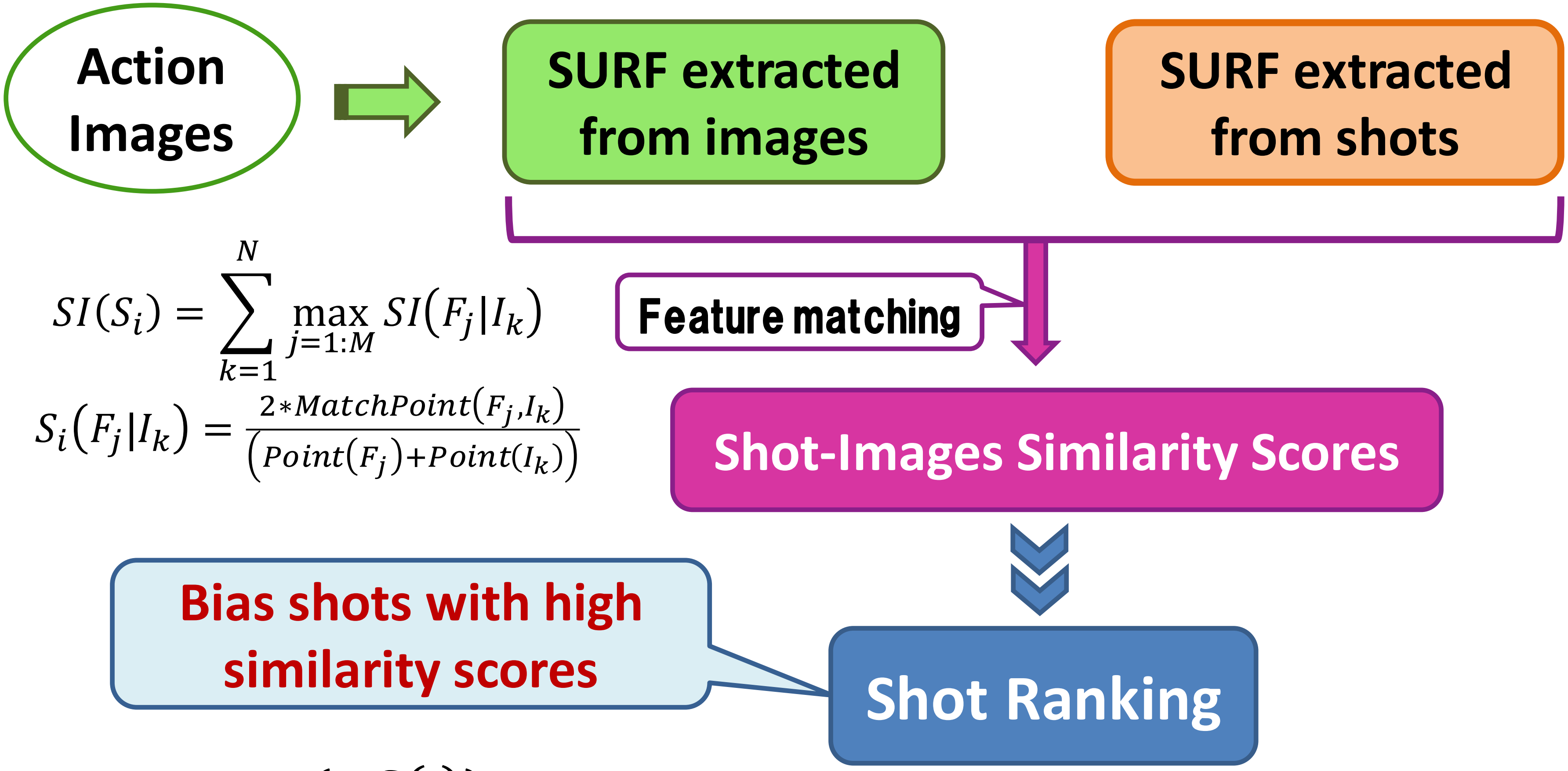
Web Images

Poselets[*] based Human Detection

Action Images

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

Shot-Images similarity based Shot Ranking


Action Images

SURF extracted from images

SURF extracted from shots

Feature matching

Shot-Images Similarity Scores

Bias shots with high similarity scores

Shot Ranking

$$SI(S_i) = \sum_{k=1}^{N} \max_{j=1:M} SI\left(F_j|I_k\right)$$

$$S_i\left(F_j|I_k\right) = \frac{2*MatchPoint\left(F_j,I_k\right)}{\left(Point\left(F_j\right)+Point(I_k)\right)}$$

$$p_i = \frac{\exp(\gamma\ \mathrm{S(i)})}{\sum_{j=1}^{n}\exp(\gamma\ \mathrm{S}(j))} \quad \gamma = log2$$

Outline

- Objective & Motivation
- Contributions
- Related work
- Previous Work & its Problems
- This work
- **Experiments and Results**
- Conclusion & Future works

Experiments & Results

- ***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

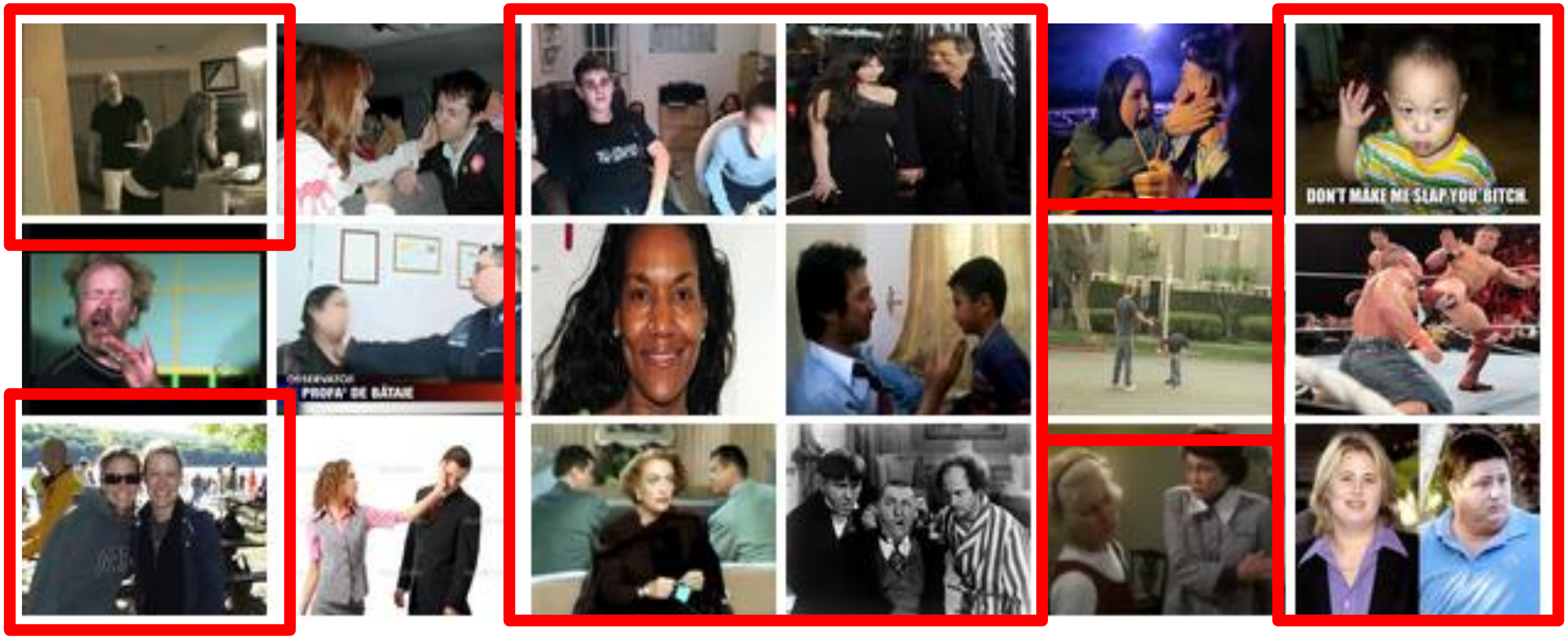
Worst 5 actions in terms of improvement

(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

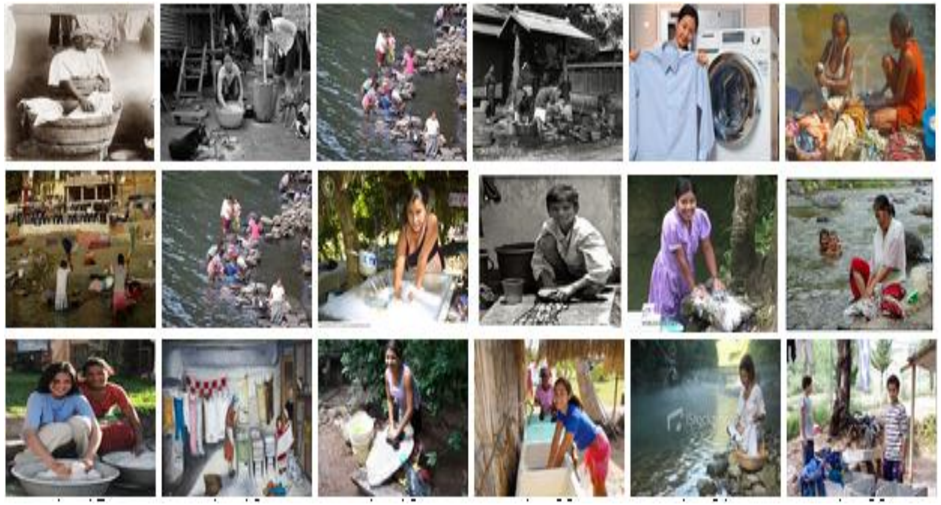


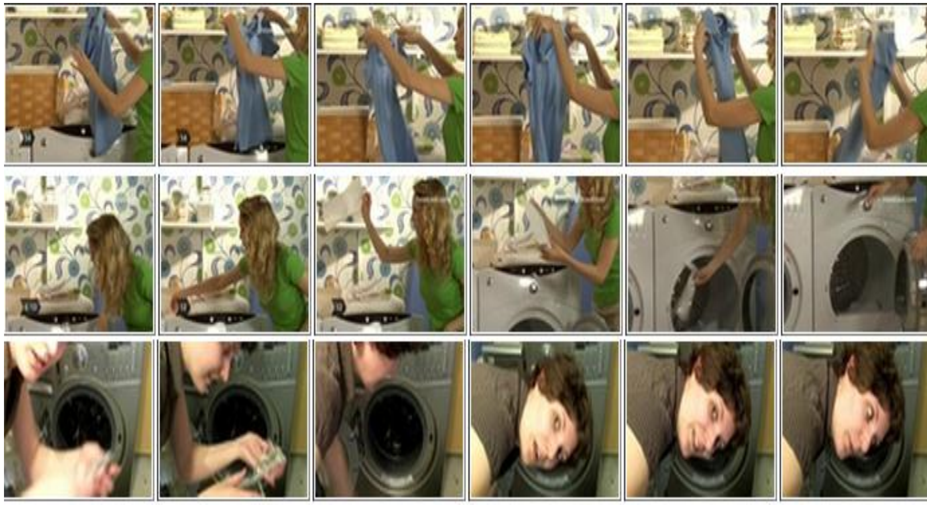
Top selected Web Images for 'slap+face'

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

- *Apply Web action images to the problem of automatically extracting action video shots*
- ***Promising results*** *show effectiveness of our modifications*
- **Future works**:

  - improve video selection step

  - try more features

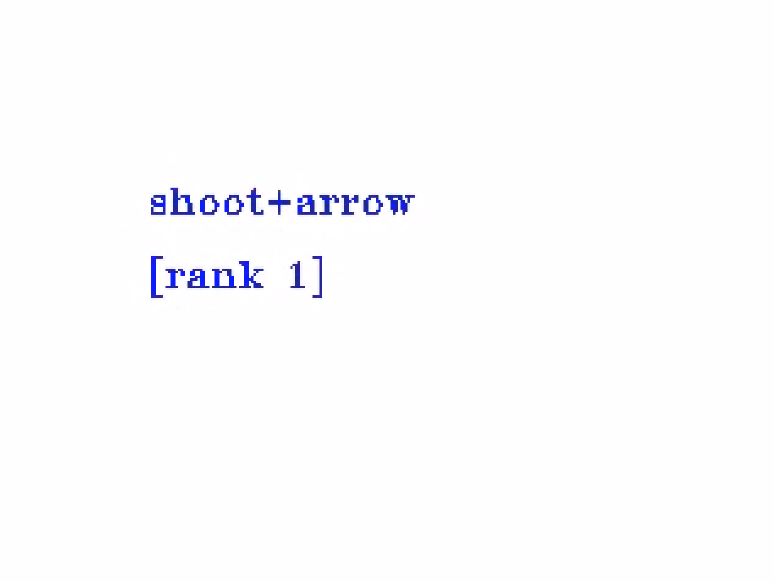
  - apply cross-domain learning

serve+tennis

serve+tennis

[rank 1]

shoot+arrow



shoot+arrow

[rank 1]

snow+falling

snow+falling

[rank 1]

http://www.youtube.com/watch?v=3Rtioh8-nd4

airplane+flying

airplane+flying

[rank 1]

http://www.youtube.com/watch?v=VtMOBrg5-sY