Concept-Based Framework for Detecting High-Level Events in Video

Shih-Fu Chang
Columbia University

September 18th, 2014
University of Maryland

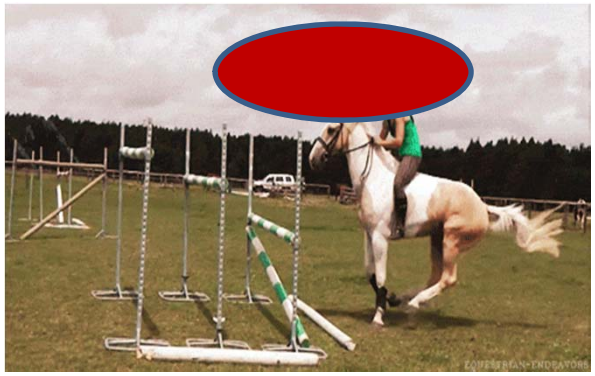
Acknowledgments: This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of IARPA, DoI/NBC, or the U.S. Government.

Complex Video Event Detection

What “event” is described in these videos?

Felling a Tree

Attempting Board Trick



Horseriding Competition

Automatic Detection will help:

High-Level Summarization

Targeted Advertisement/Learning

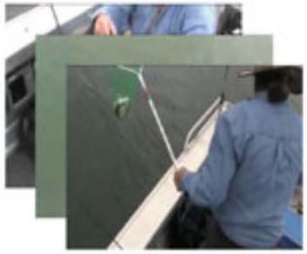
Recognize Complex Events

- NIST TRECVID Multimedia Event Detection (MED) contest
- Detecting complex events in ~200,000 videos

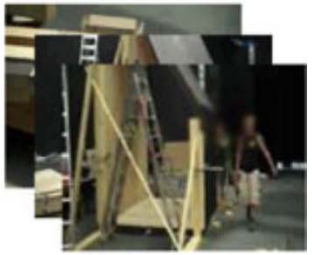
MED 2011 devel. events


Attempting a board trick


Feeding an animal


Landing a fish


Wedding ceremony


Working on a woodworking project

MED 2011 testing events


Birthday party


Changing a vehicle tire

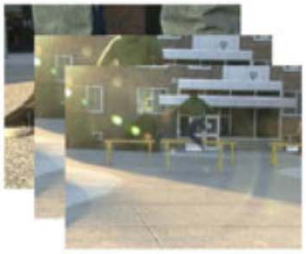

Flash mob gathering


Getting a vehicle unstuck


Grooming an animal


Making a sandwich


Parade

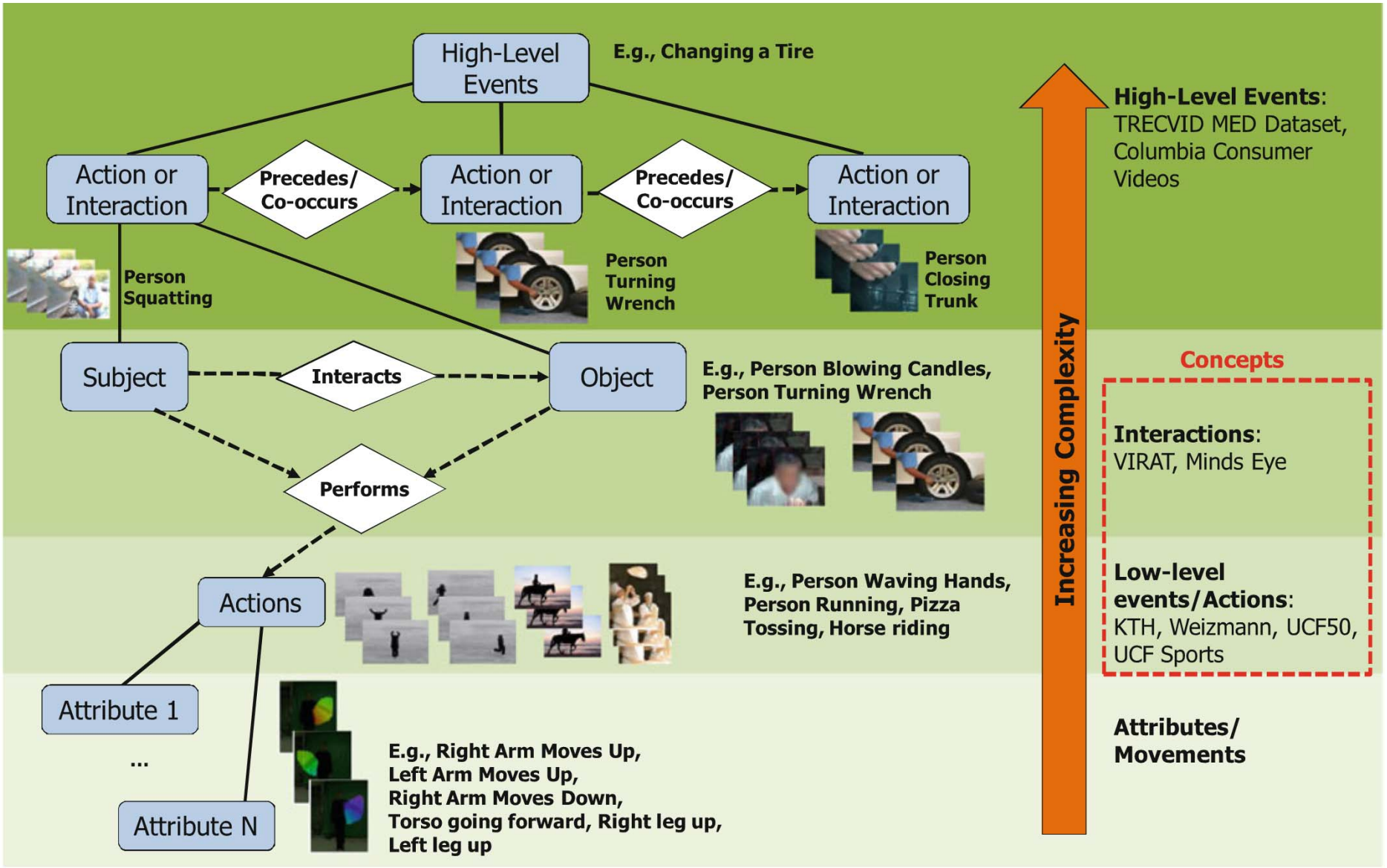

Parkour


Repairing an appliance

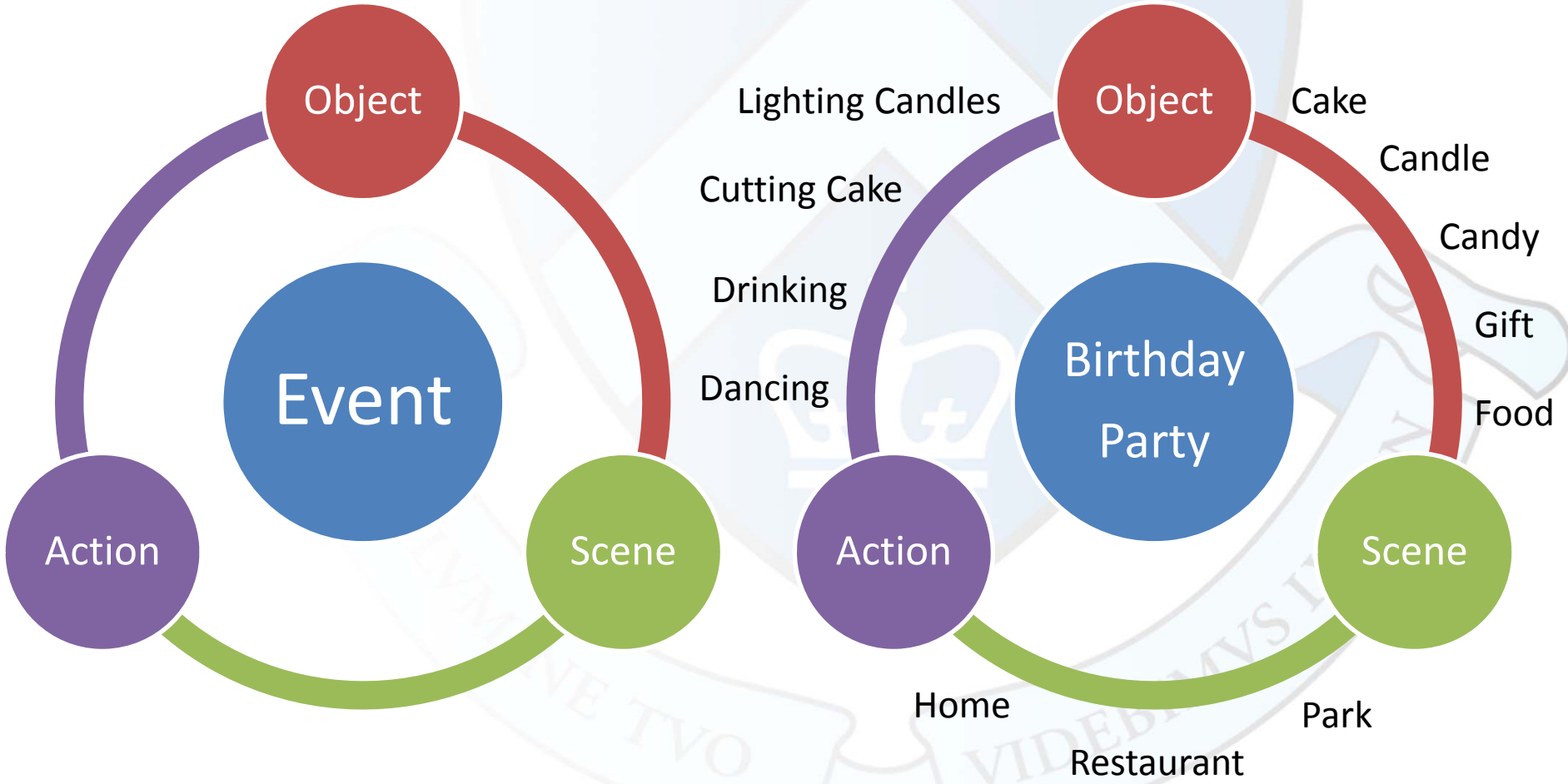

Working on a sewing project

High-level Events are Complex

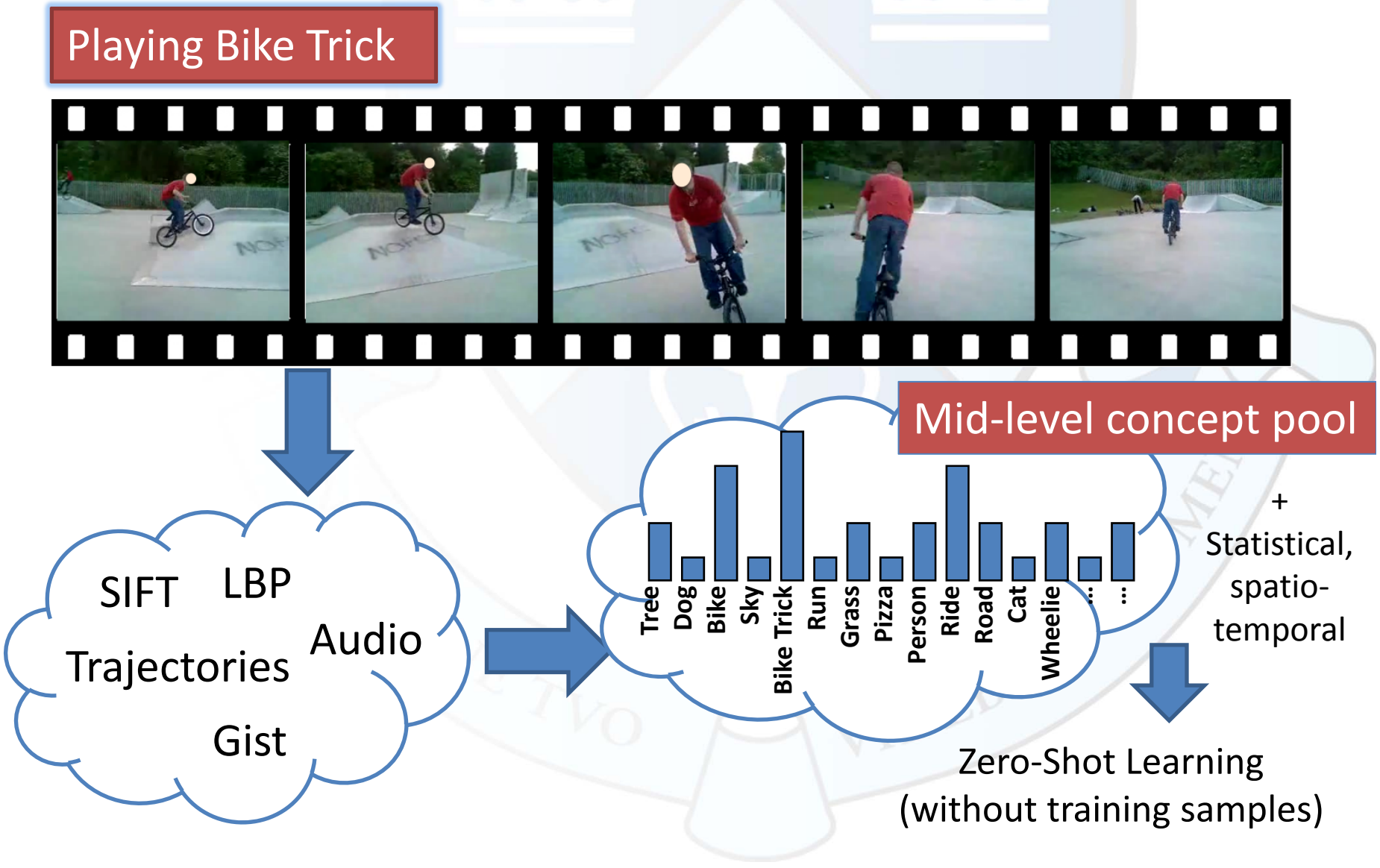
- Y. Jiang et al., high level events recognition in unconstrained videos, IJMIR, 2012 (survey)



Decompose an event into concepts



Concepts as Mid-Level Representation for Video



How to Discover Relevant Concepts?

A question often skipped by computer vision community.

Some work in attribute discovery, but not for video evens.

Ways of discovering new complex event concepts:

- Use dictionary definitions
- Discover concepts from the Web
- Explore knowledge source, e.g., *WikiHow* events
- Perform interactive cognitive studies

Dictionary: TRECVID MED Event Kits

Groom An Animal

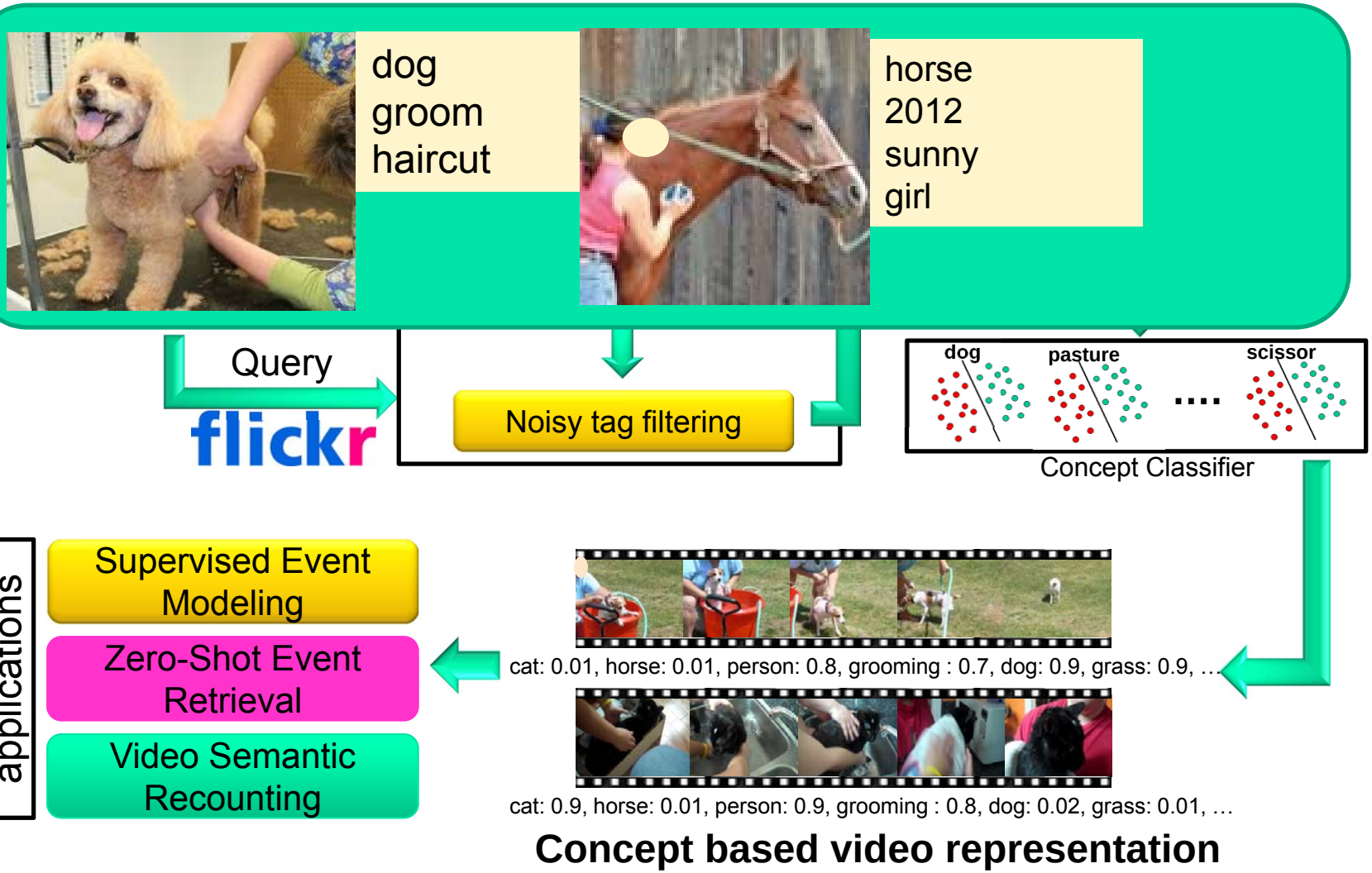
- **Definition:** One or more people groom an animal.
- **Explication**: Grooming refers to caring for the hygiene/cleanliness and appearance of the animal. A very common form of grooming is bathing the animal, usually accomplished by either immersing the animal in water or spraying the animal with water, often followed by application of soap/shampoo and then additional rinsing with water. Other grooming activities include trimming of hair and nails, cleaning of teeth, eyes, and ears, and brushing, combing, and styling the fur of the animal.
- **Evidence Description:**
  - **Scene:** yard, corral, bathroom, grooming salon, exhibition center.
  - **Object/ people:** sink, bathtub, hose, shower, soap, shampoo, scissors, clippers.
  - **Activity :** spraying hose, putting animal on table, rinsing, blow drying fur, cutting fur, clipping nails.

Obviously hard to scale up!

How to Discover Relevant Concepts?

- Use dictionary definitions
- Discover Concepts from the Web
- Use knowledge source like *WikiHow*
- Perform interactive cognitive studies

Discover Concepts via Web Data Expansion



Jiawei Chen, Yin Cui, Guangnan Ye, Dong Liu, Shih-Fu Chang. **Event-Driven Semantic Concept Discovery by Exploiting Weakly Tagged Internet Images**. In *ACM International Conference on Multimedia Retrieval (ICMR), full paper (oral)*, 2014.

Concepts discovered from Web

(Attempting a bike trick)

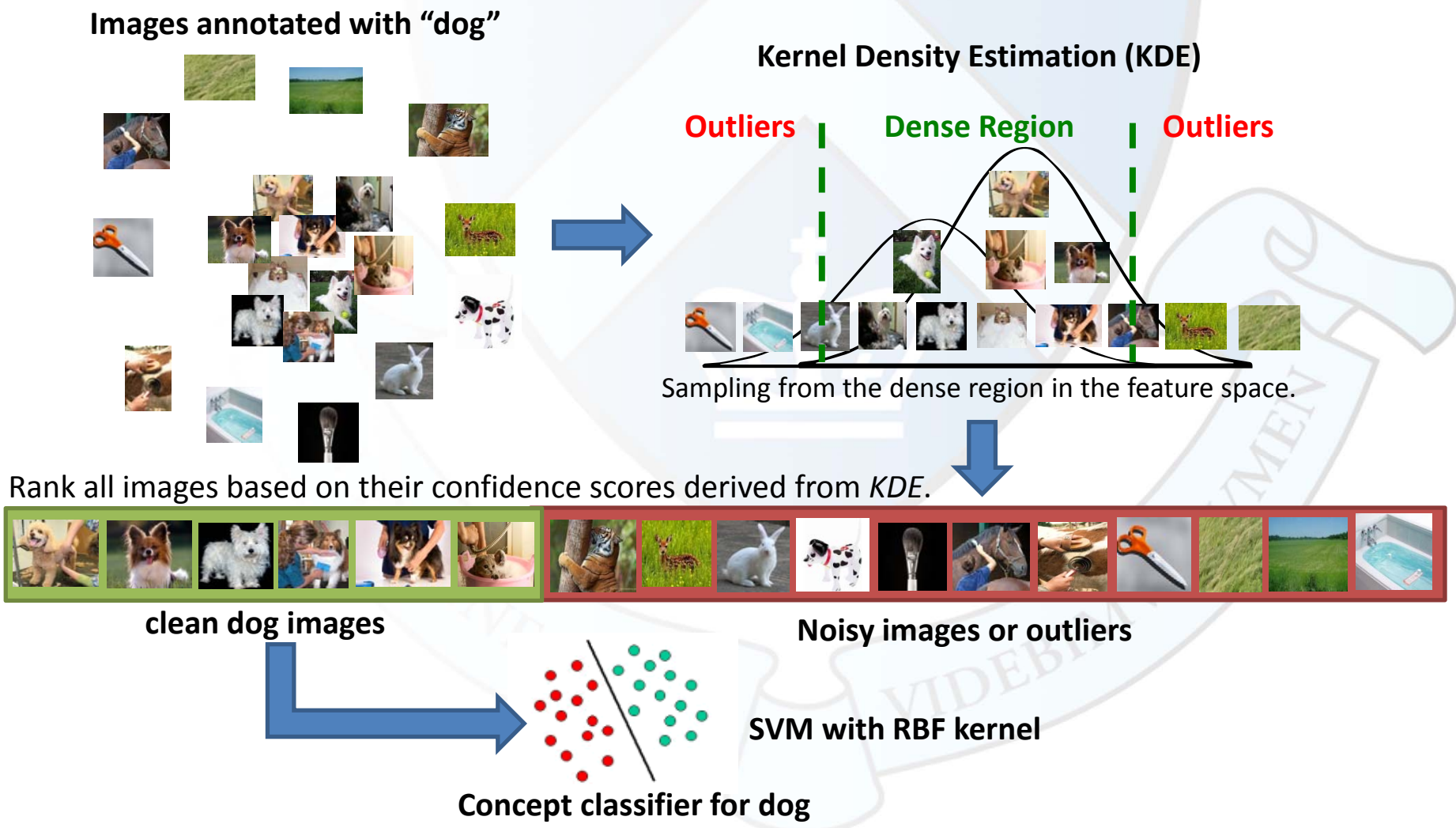
Web-Concept Expansion Finds Novel Concepts

Event Name	Concepts Discovered from Different Resources	
getting a vehicle unstuck	Classemes	air transportation vehicle, all terrain vehicle, amphibious vehicle, armed person, armored fighting vehicle, armored recovery vehicle, armored vehicle, armored vehicle heavy, armored vehicle light, command vehicle
	ImageNet	vehicle, bumper car, craft, military vehicle, rocket, skibob, sled, steamroller, wheeled vehicle, conveyance
	Web	**tire, car, snow, stick, stuck, winter, vehicle, truck, night, blizzard**
grooming an animal	Classemes	adult animal, animal, animal activity, animal blo, animal body part, animal body region, animal cage, animal container, animal pen, animal shelter
	ImageNet	groom, animal, invertebrate, homeotherm, work animal, darter, range animal, creepy-crawly, domestic animal, molter
	Web	**dog, pet, grooming, cat, animal, bath, cute, canine, puppy, water**
making a sandwich	Classemes	baking dish, cafe place, classroom setting, collection display setting, cutting device, dish drying rack, food utensil, hair cutting razor, hdtv set, hole making tool
	ImageNet	sandwich, open-face sandwich, butty, reuben, ham sandwich, gyro, chicken sandwich, hotdog, club sandwich, wrap
	Web	**sandwich, food, bread, cooking, cheese, spice, baking, pan, kitchen, breakfast**

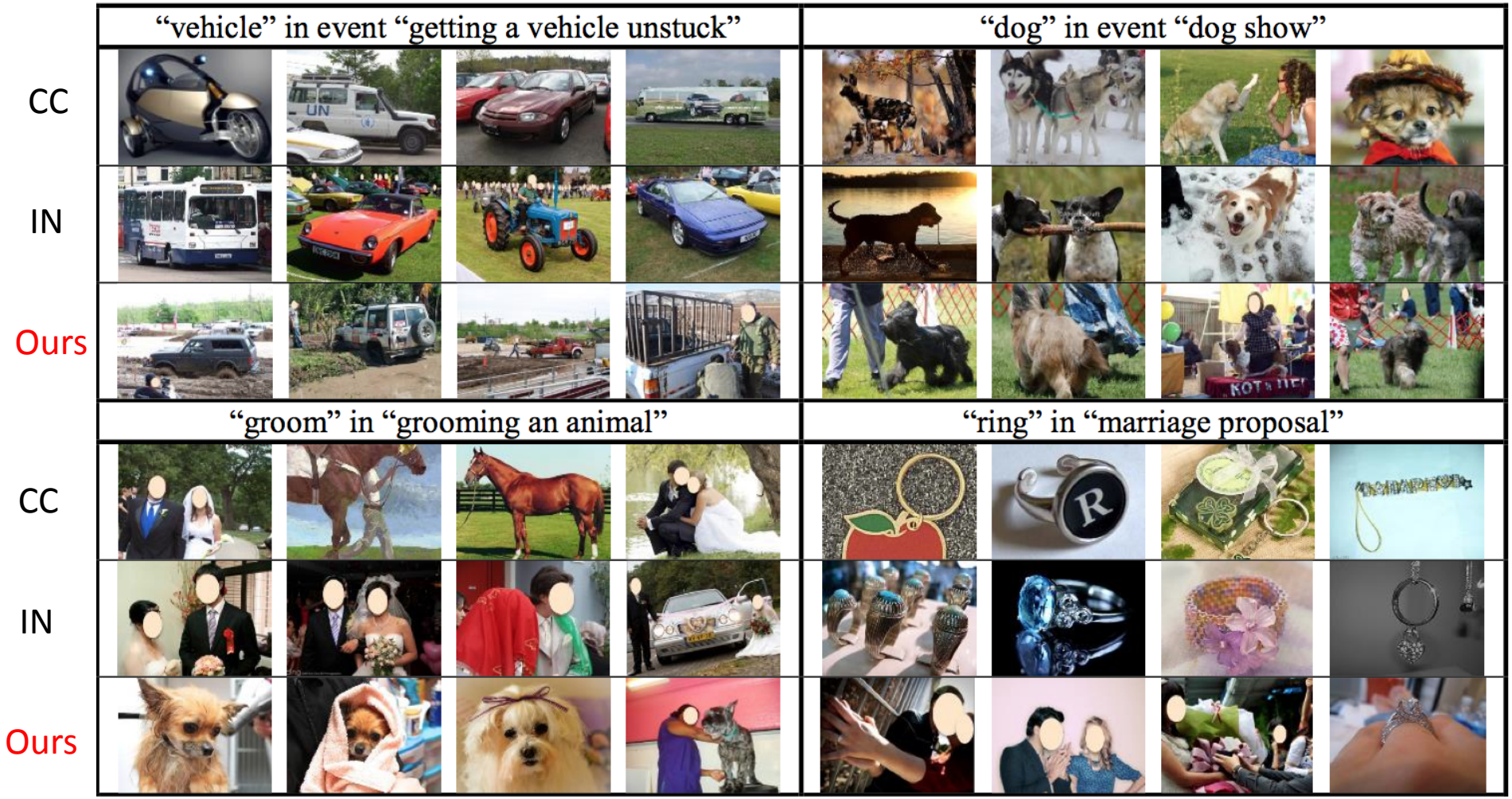
Top 10 concepts discovered from different resources

Concept images are noisy

- Training Image Selection



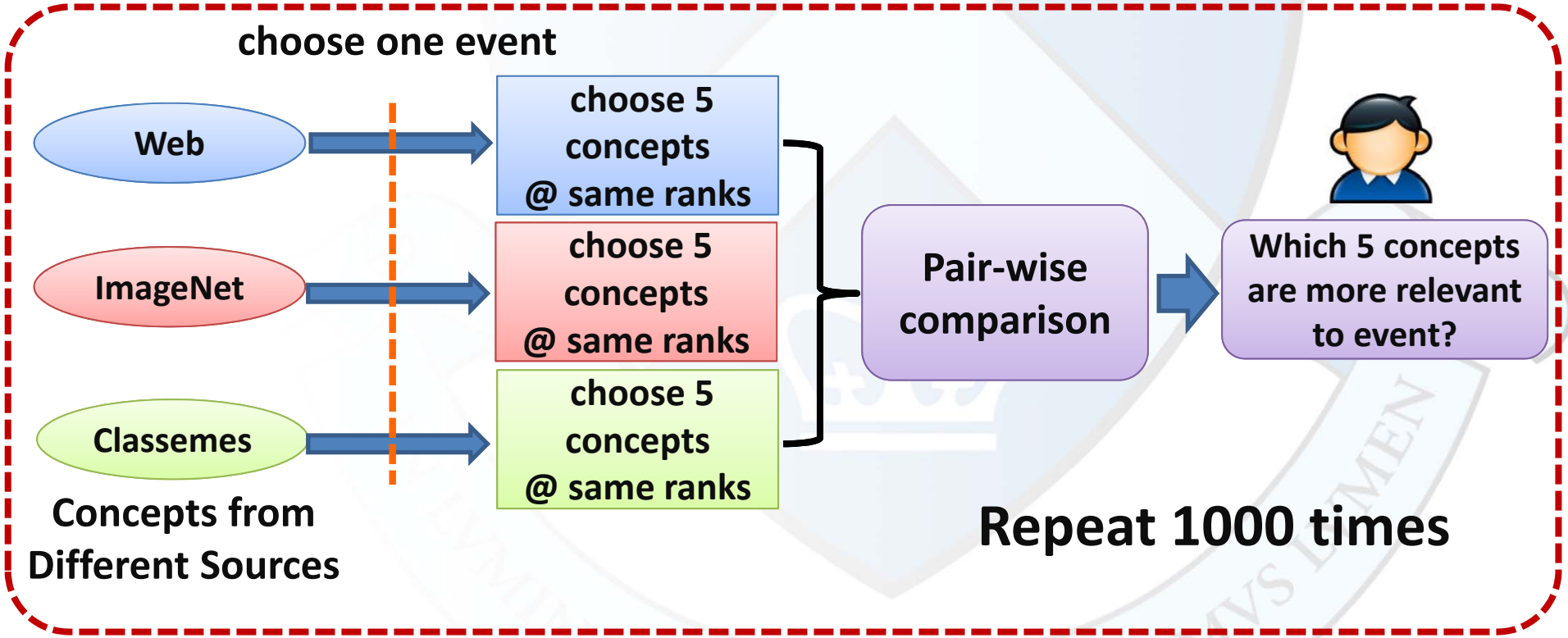
Event-specific expansion finds relevant images



CC: Classemes IN: ImageNet

Human Evaluation

- For each event, choose top 100 concepts based on similarity between concept name and event name.



- Our concept is better than others with ***81.29%*** chance.
- ImageNet concept is better than others with ***34.19%*** chance.
- Classemes concept is better than others with ***32.46%*** chance.

Zero-Shot Event Retrieval

- Given an event name as textual query without any training videos, rank all videos.

- A simple search method using concepts

Semantic Similarity from WordNet

Concept Score

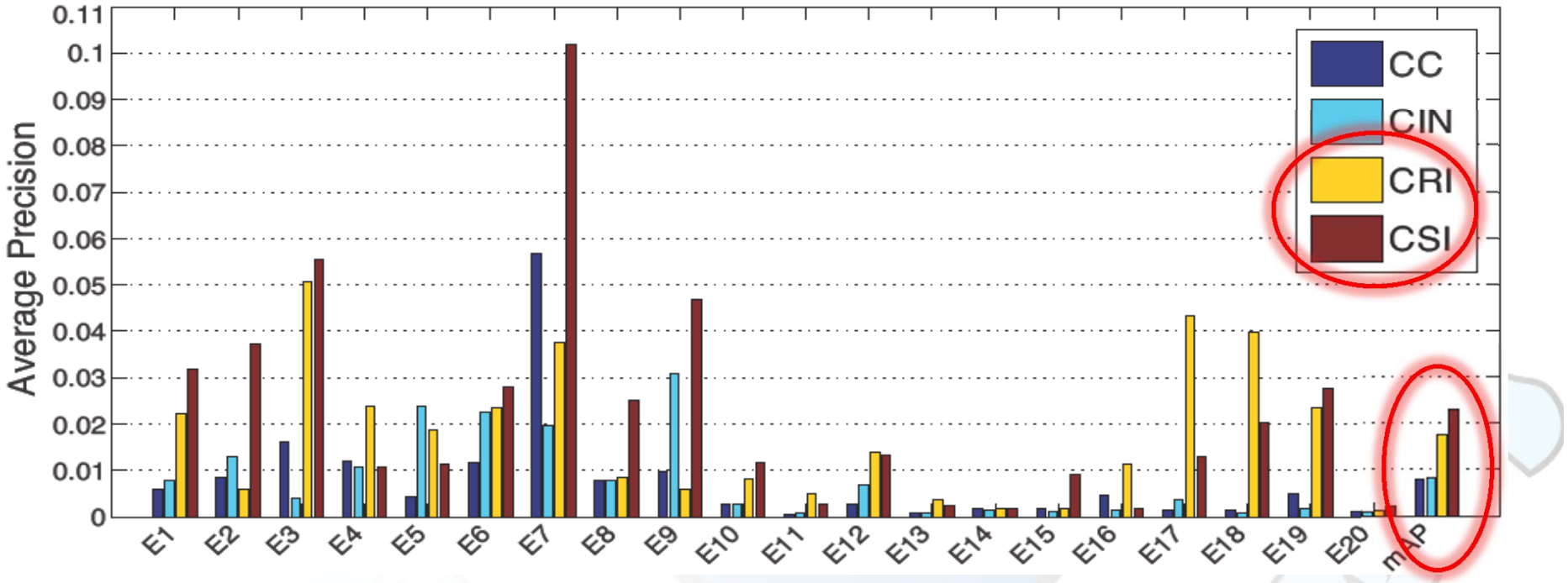
$$RankingScore = \sum_{i=1}^{T} sim(concept_i, Event) \times score_i$$

T : # of concepts

Experiment Setup

- 20 Pre-specified MED events
  - No training video for each event.
  - Use 100 dimension concept score vector for each event as feature representation.
  - Evaluation metric : full-length Mean Average Precision (MAP).

Performance: Zero-Shot Concept-based Event Retrieval



CC: Classeme concepts, CIN: ImageNet, CRI/CSI: Ours

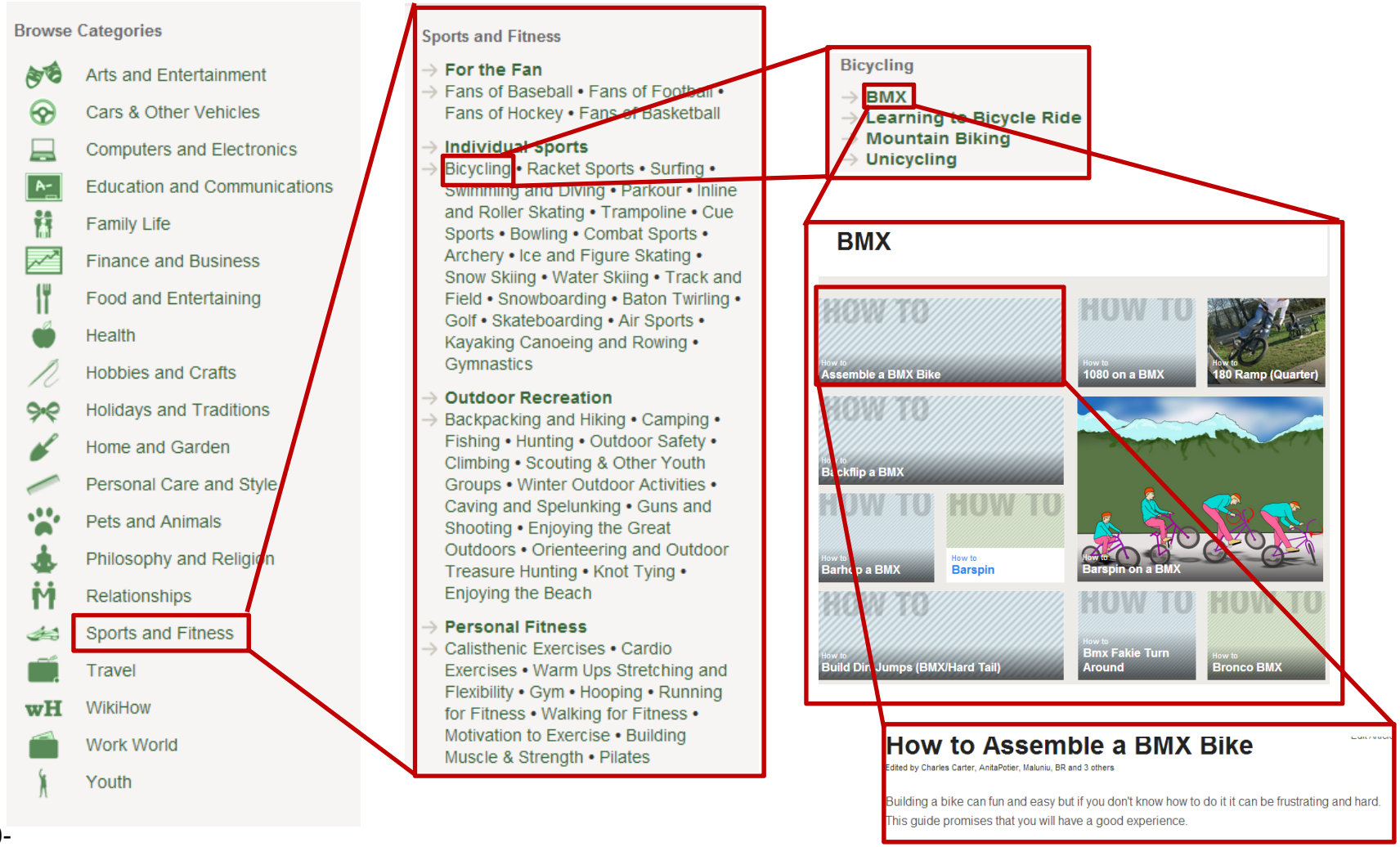
E1: birthday party, **E2**: changing a vehicle tire,**E3**: flash mob gathering,**E4**: getting a vehicle unstuck, **E5**: grooming an animal,**E6**: making a sandwich,**E7**: parade, **E8**: parkour, **E9**: repair an appliance,**E10**: working on a sewing project, **E11**: attempting a bike trick, **E12**: cleaning an appliance, **E13**: dog show, **E14**: giving directions to a location, **E15**: marriage proposal, **E16**: renovating a home, **E17**: rock climbing, **E18**: town hall meeting,**E19**: winning a race without a vehicle, **E20**: working on a metal crafts project

How to Discover Relevant Concepts?

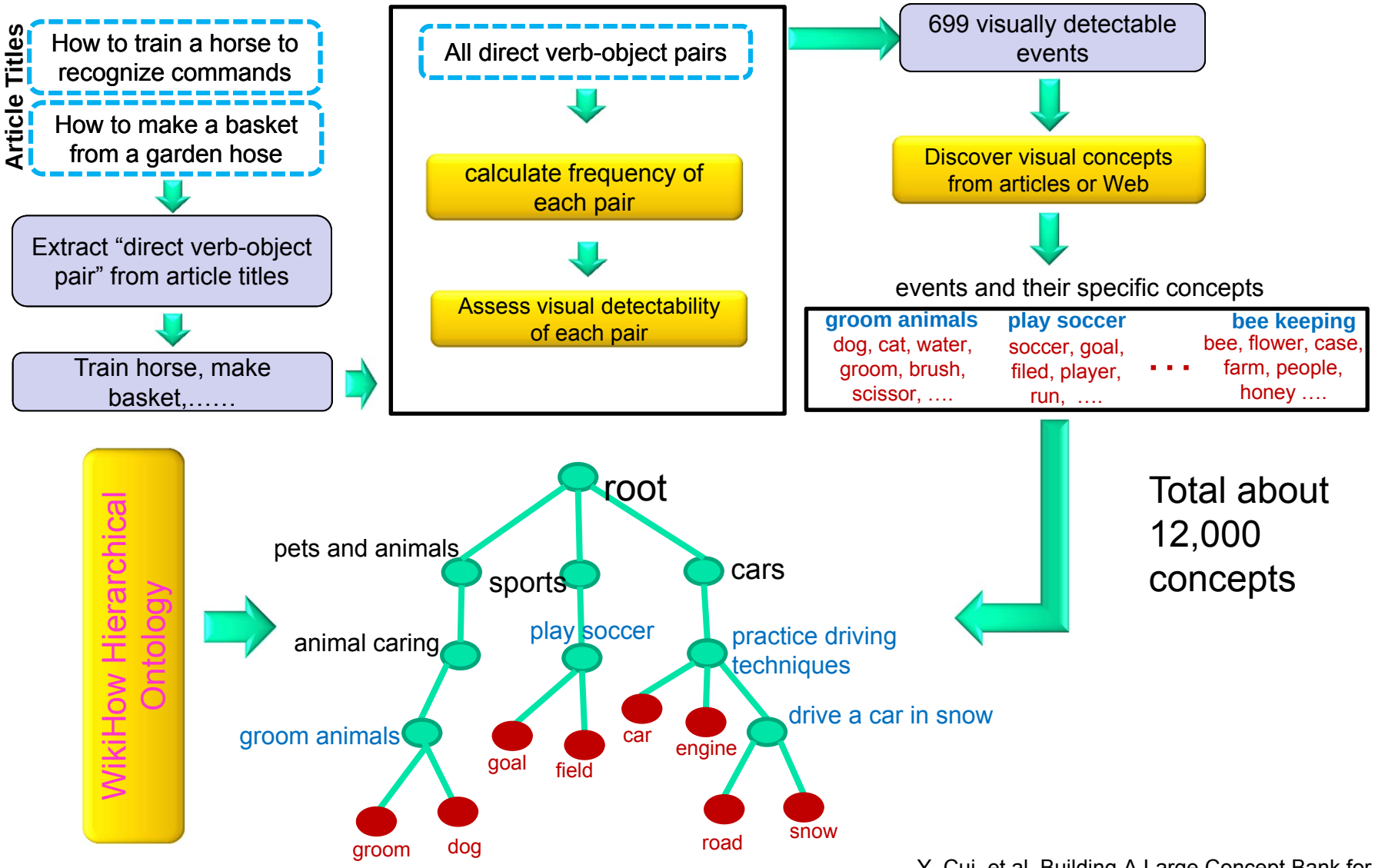
- Use dictionary definitions
- Discover Concepts from the Web
- Use knowledge source like *WikiHow*
- Perform interactive cognitive studies

Knowledge Source: WikiHow

- A wiki contains ~180,000 articles on 2,803 “how to” categories.
- All articles are organized into a hierarchical structure.

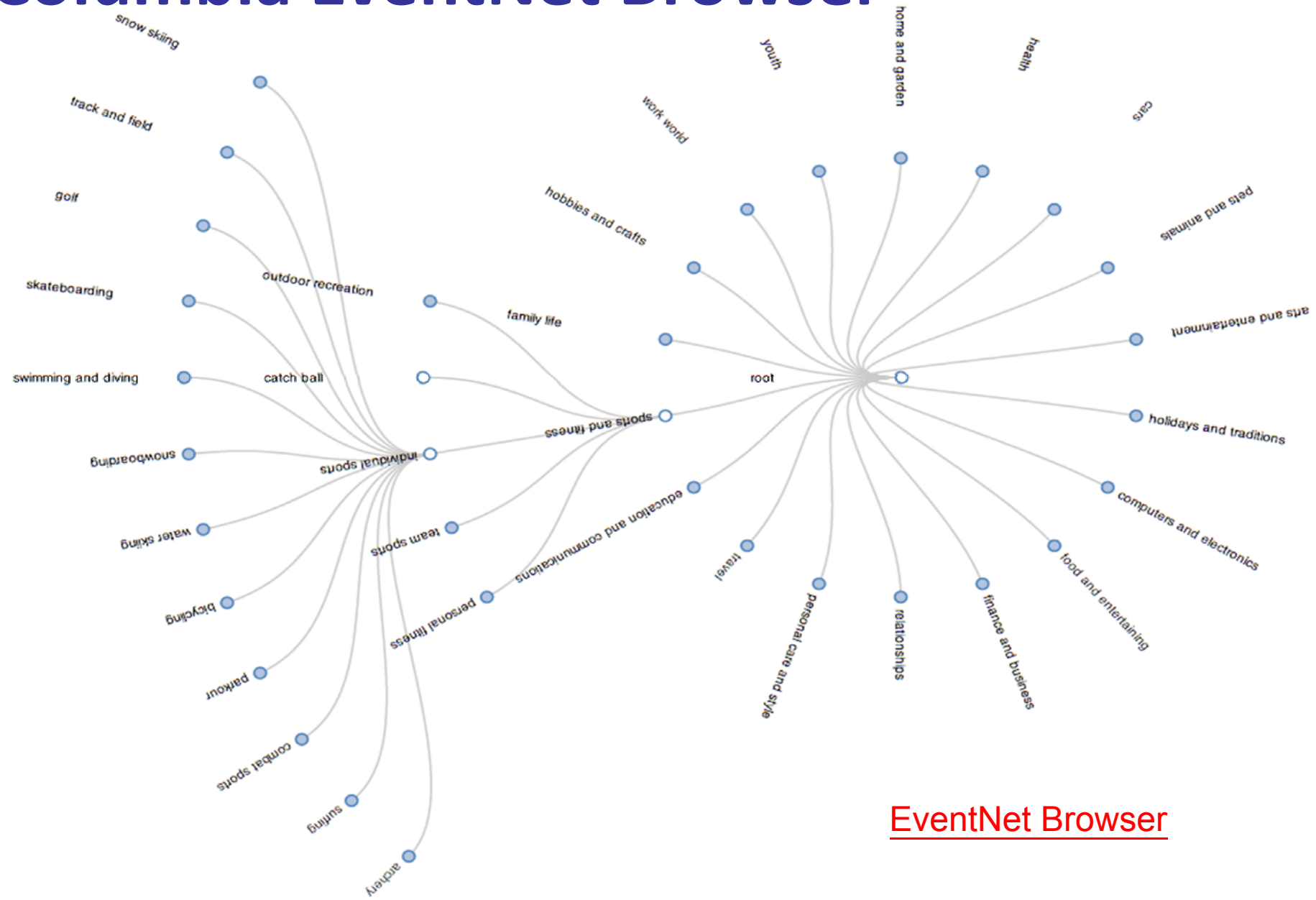


Construct EventNet from WikiHow



Y. Cui, et al, Building A Large Concept Bank for Representing Events in Video, arXiv, March 2014

Columbia EventNet Browser



EventNet Browser

How to Discover Relevant Concepts?

- Use dictionary definitions
- Discover Concepts from the Web
- Use knowledge source like *WikiHow*
- Perform interactive cognitive studies

How Do Humans Judge Events?

Conventional: Passive **Linear Video Playback** and Judge

- Used extensively in video summarization [1,2], persistent surveillance [3].

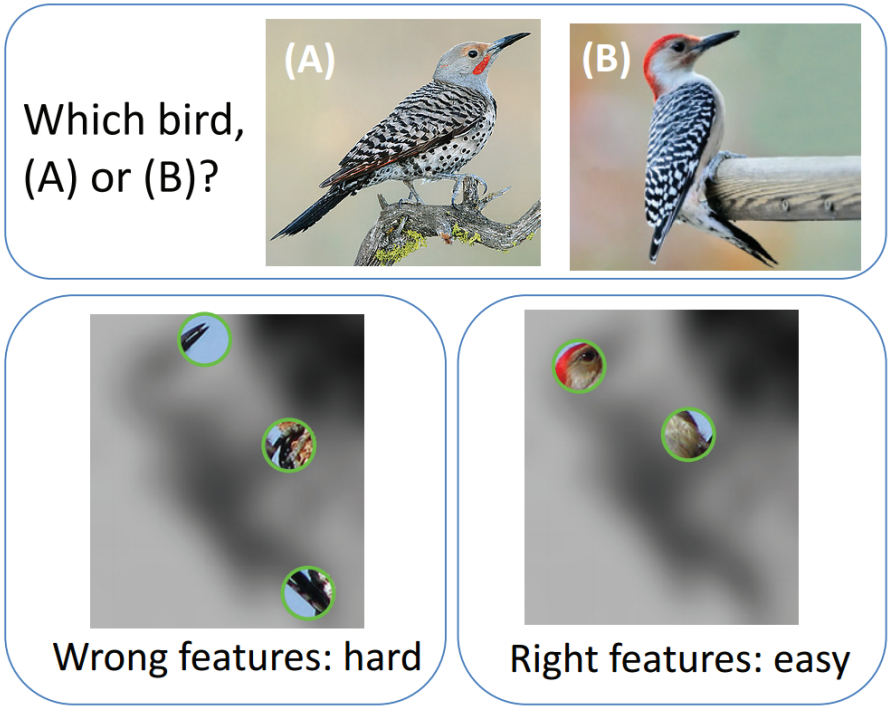
[1] Wu et. al, **Video summarization via crowdsourcing**, *CHI '11*, pp. 1531-1536.

[2] Ma et. al, **A user attention model for video summarization**. MM '02, pp. 533-542.

[3] Kim et. al, **Intelligent visual surveillance - A survey**, IJCAS' 10, pp 926-939

Human Decision Strategy

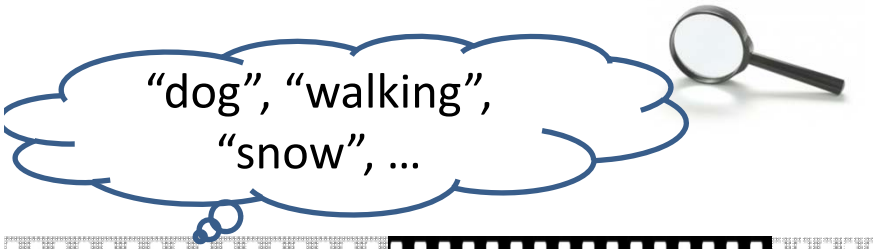
- Do humans actually employ linear playback to judge a complex event?
- Probably not!
- Bubble-game [5] to identify "Human" discovered discriminative patches for fine-grained recognition



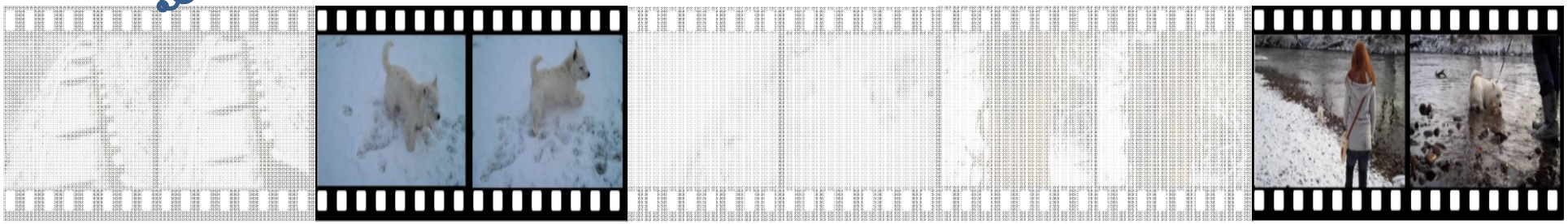
[5] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In IEEE CVPR, 2013.

Evidences needed by humans for event judgment?

Look for **Needed Evidence** in Events **(Proposed)**



- Not necessarily linear
- Not necessarily sequential



S. Bhattacharya, F.X. Yu, S.-F. Chang. **Minimally Needed Evidence for Complex Event Recognition in Unconstrained Videos**. In *ACM Conf. on Multimedia Retrieval (ICMR),* April 2014.

Let's take a test …

Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

Reveal?	Reveal?	Reveal?	Reveal?	Reveal?	12

Lets take a test …

Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

Reveal?		Reveal?	Reveal?	Reveal?	10

Lets take a test …

Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

Reveal?

Reveal?

Reveal?

8

Lets take a test …

Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

	Reveal?		Reveal?	Reveal?		8
	Reveal?	Reveal?	Reveal?	Reveal?	Reveal?	12

Lets take a test …

Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

	Reveal?		Reveal?	Reveal?		8
	Reveal?	Reveal?		Reveal?	Reveal?	10

Lets take a test …

Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

✓	Reveal?		Reveal?	Reveal?		8
✗	Reveal?	Reveal?			Reveal?	8

Lets take a test …

Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

	Reveal?		Reveal?	Reveal?		8
	Reveal?	Reveal?			Reveal?	8
	Reveal?	Reveal?	Reveal?	Reveal?	Reveal?	12

Lets take a test …

Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

	Reveal?		Reveal?	Reveal?		8
	Reveal?	Reveal?			Reveal?	8
		Reveal?	Reveal?	Reveal?	Reveal?	10

Lets take a test …

Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

✓	Reveal?		Reveal?	Reveal?		8
✗	Reveal?	Reveal?			Reveal?	8
✗		Reveal?	Reveal?	Reveal?	Reveal?	10

Lets take a test …

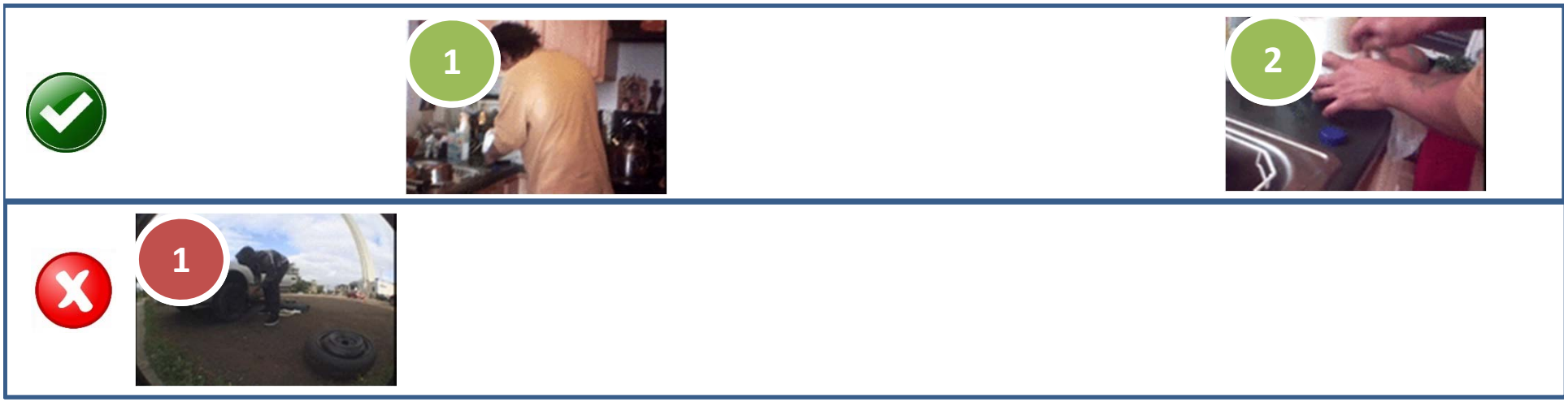
Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

✓	Reveal?		Reveal?	Reveal?		8
✗	Reveal?	Reveal?			Reveal?	8
✗		Reveal?	Reveal?	Reveal?	Reveal?	10

Congratulations you got all correct! You scored **26** out of **30**.

Minimally Needed Evidence

For the event Cleaning an appliance:



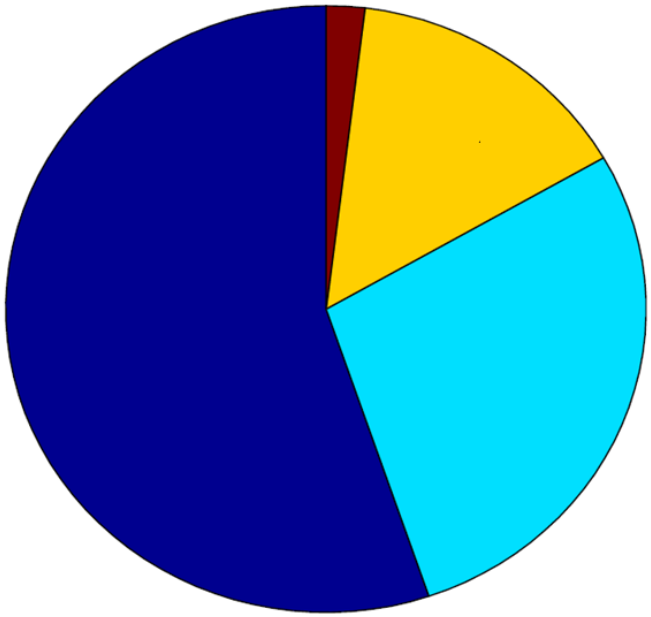
- **Practical way** of finding Minimally Needed Evidence
  → **Event Quiz Interface**
- **Clever annotation tool** → Enables **judicious use** of Human feedback
- Can **reduce computational overhead** for feature extraction

Microshot Selection

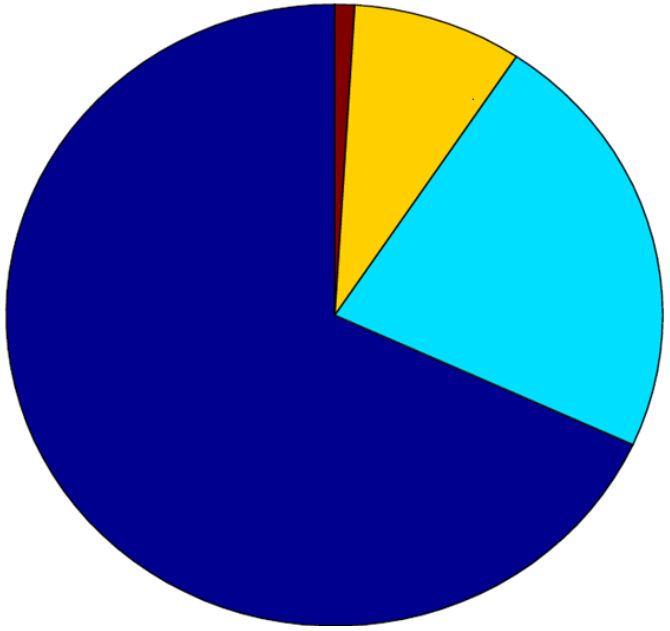
- Most action concepts e.g. jogging, boxing, can be captured in 1.5s of continuous footage (30hz)

- Divide video into non-overlapping 1.5s blocks; Filter out non-interesting microshots (low appearance + motion entropy)

Surprisingly, Humans can



Correctly Identify (positives)

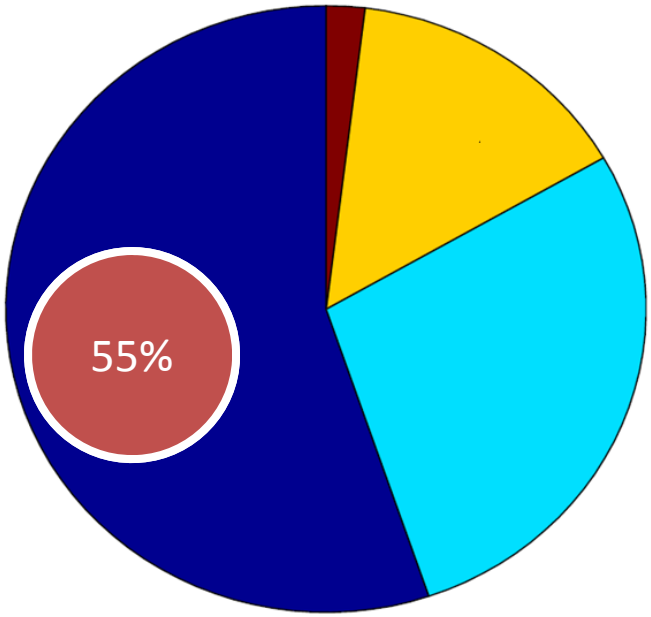


Correctly Reject (negatives)

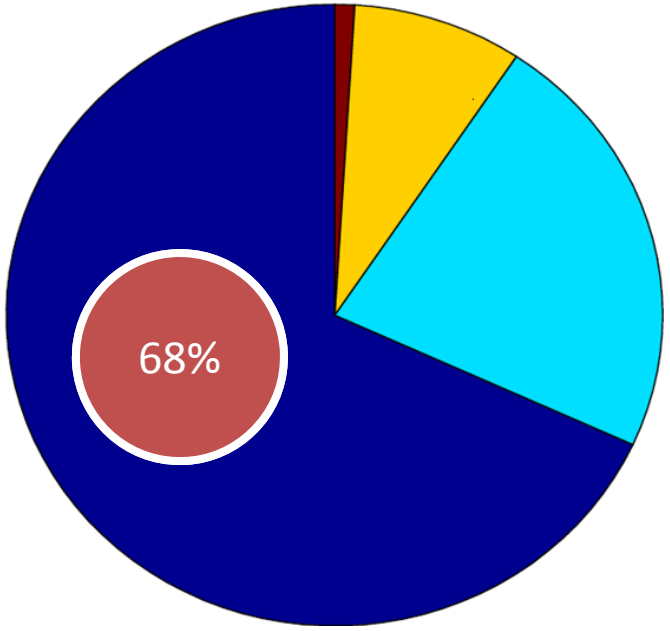
Number of microshots Revealed 1 2 3 >3

- Correctly judge an event **in ~87% cases** from just **1 or 2 microshots** (1.5s footage)

Surprisingly, Humans can



Correctly Identify (positives)

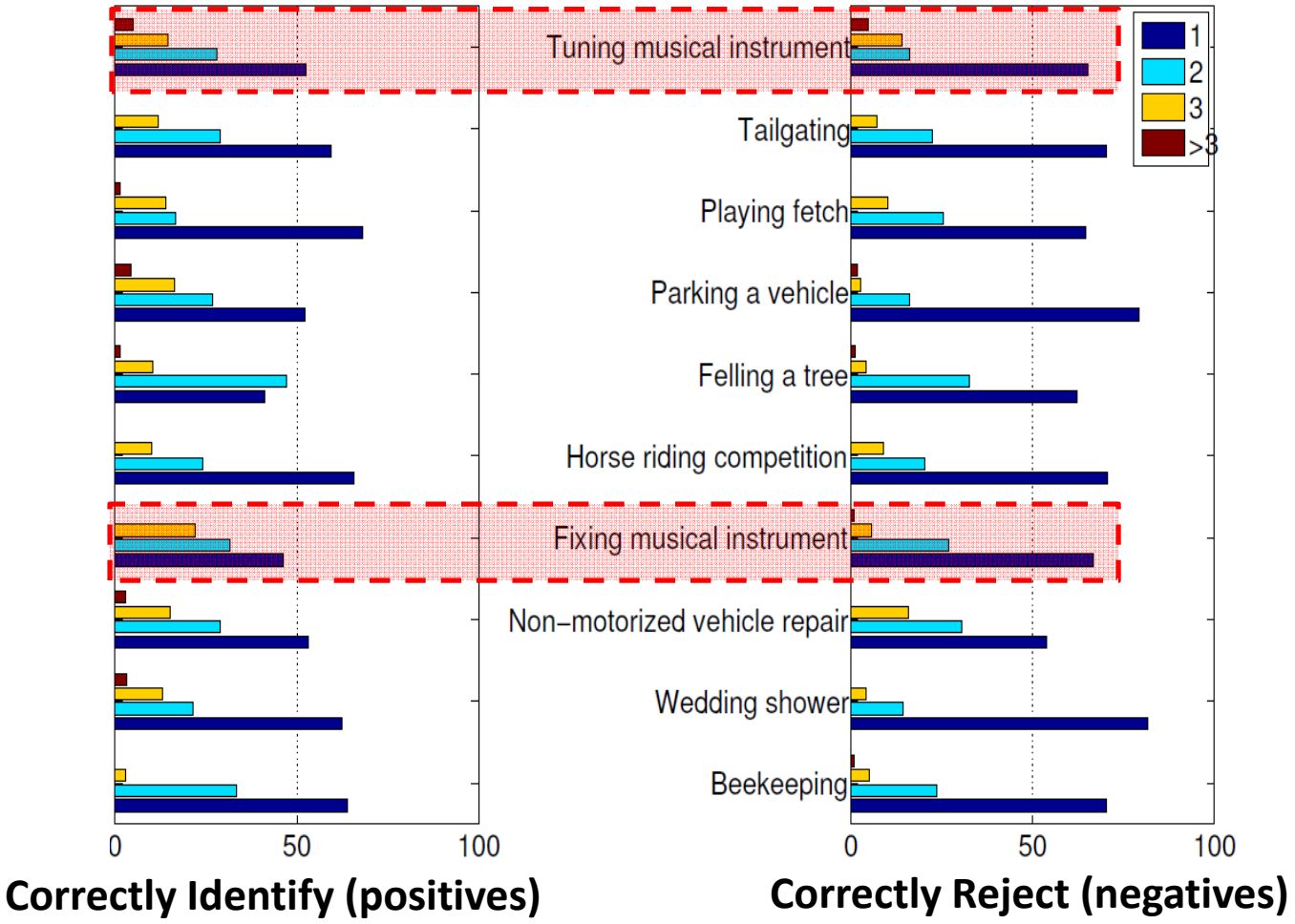


Correctly Reject (negatives)

Number of microshots Revealed 1 2 3 >3

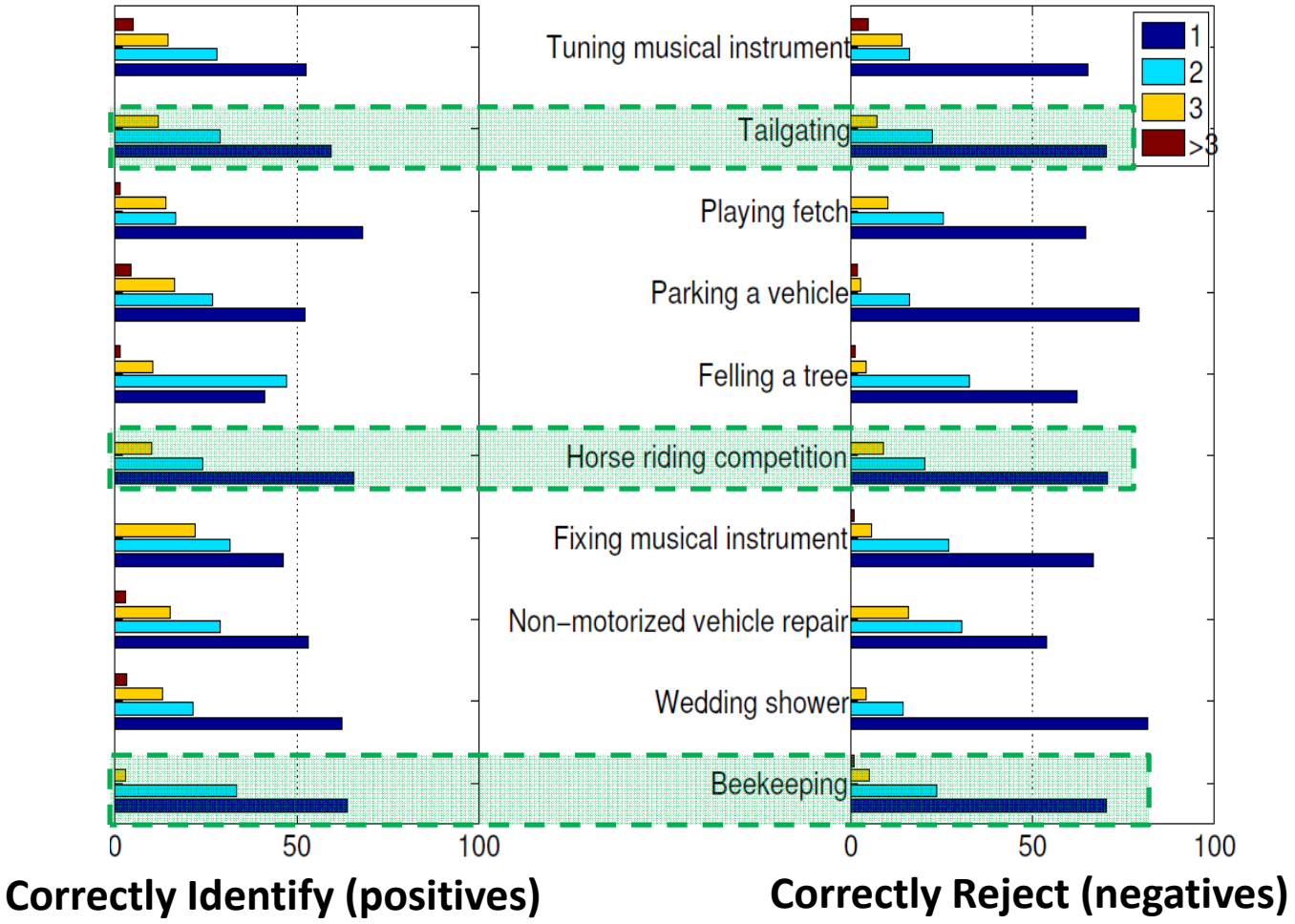
- **Correctly identify** a video **containing an event** in videos in 55% cases with 1 shot
- **Correctly reject** a video for **not containing an event** in 68% cases

Additionally- Event Complexity varies



- "Tuning a musical instrument" **more visually challenging** than "Fixing musical instrument" (needs more microshot revealation than other events)

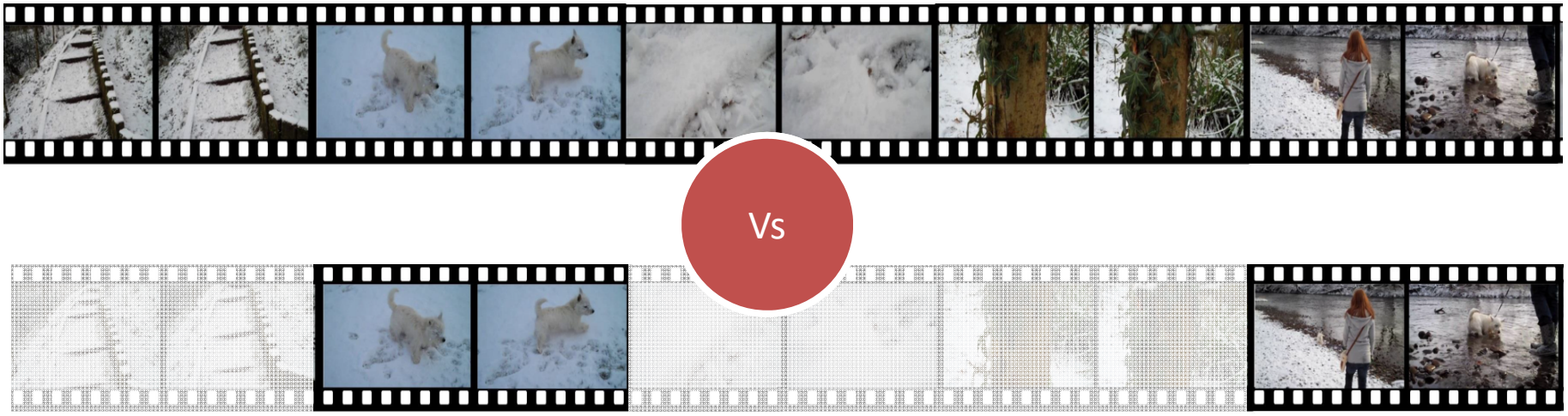
Additionally- Event Complexity varies



- "Tailgating", "Horseriding Competition" and "Beekeeping" **require less evidence**

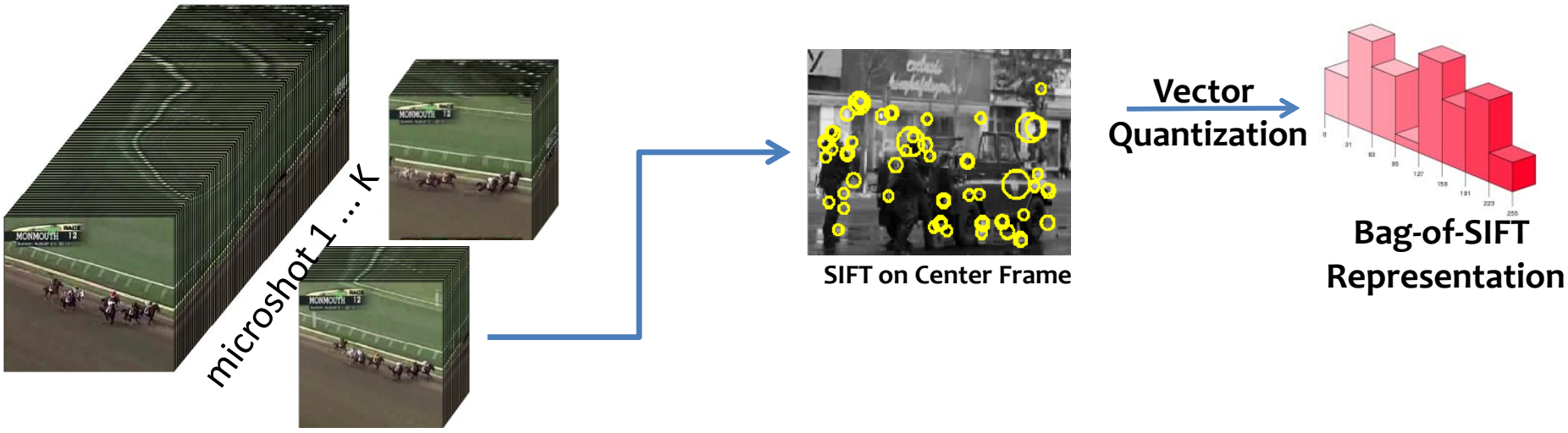
Hypothesis

- Human Discovered Evidences provide better event representation for recognition



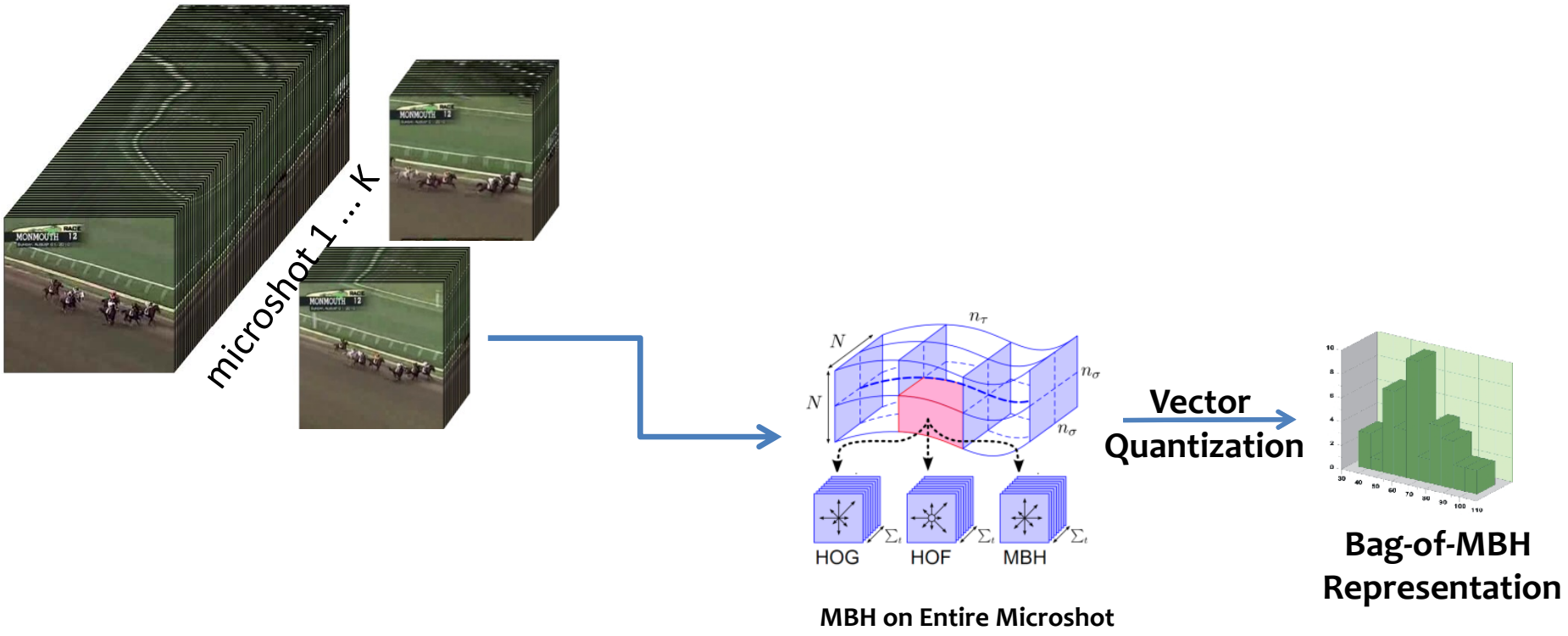
- To validate: Basic Retrieval experiment

Representation for Retrieval



- Standard Bag-of-visual Words approach
- Empirically determined vocabulary size for
  - **Appearance Features : 2,000**

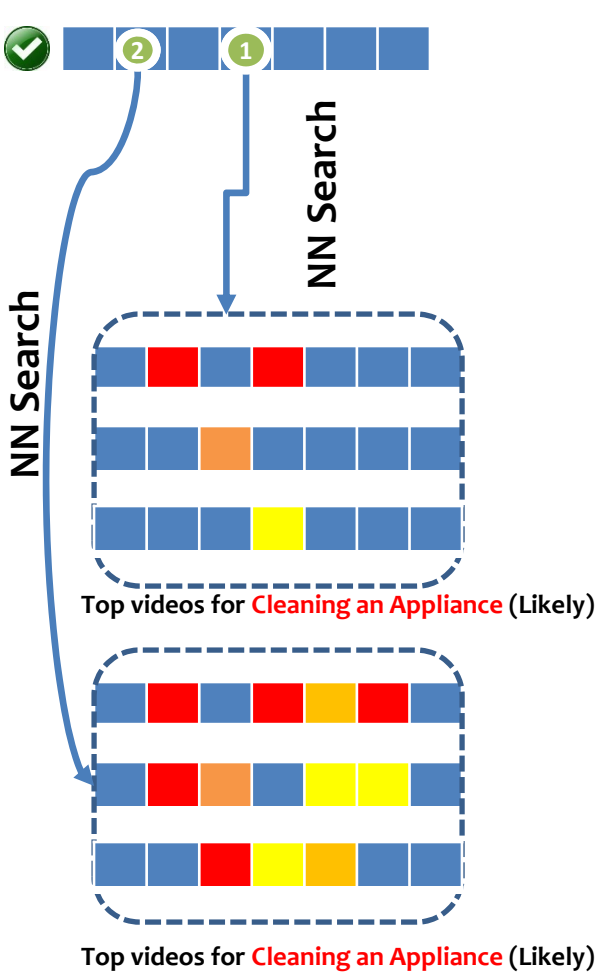
Representation for Retrieval

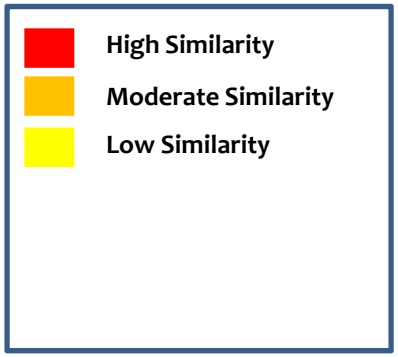


- Standard Bag-of-visual Words approach
- Empirically determined vocabulary size for
  - **Motion Features : 5,000**

Retrieval Methodology

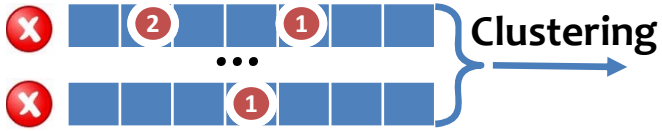
Target Event: Cleaning an Appliance





Use Negative Cues for Quick Reject

Target Event: Cleaning an Appliance



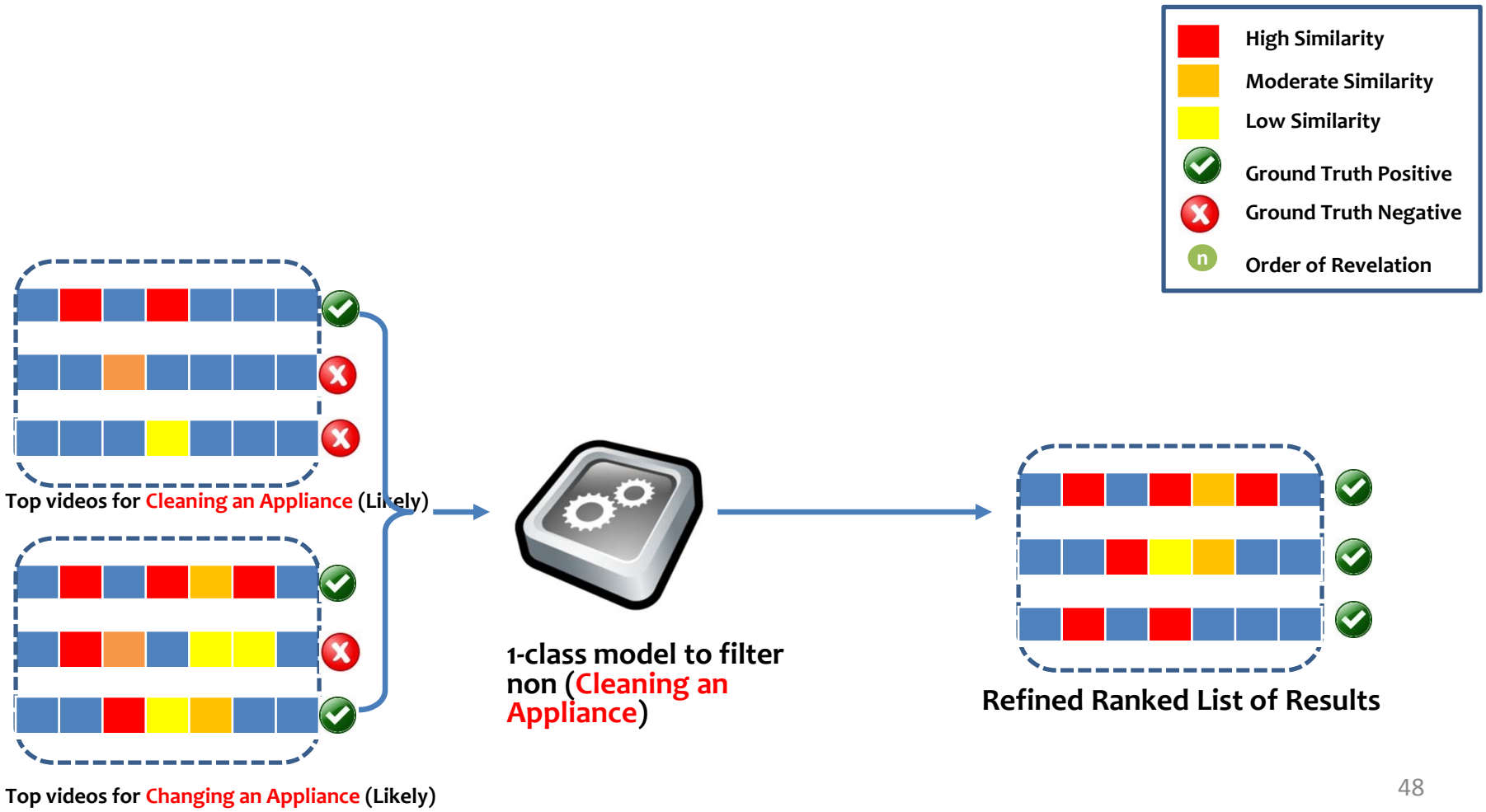
Top-3 Negative Clusters for "Cleaning an Appliance"

Train 1-class SVM

1-class model to filter non (Cleaning an Appliance)

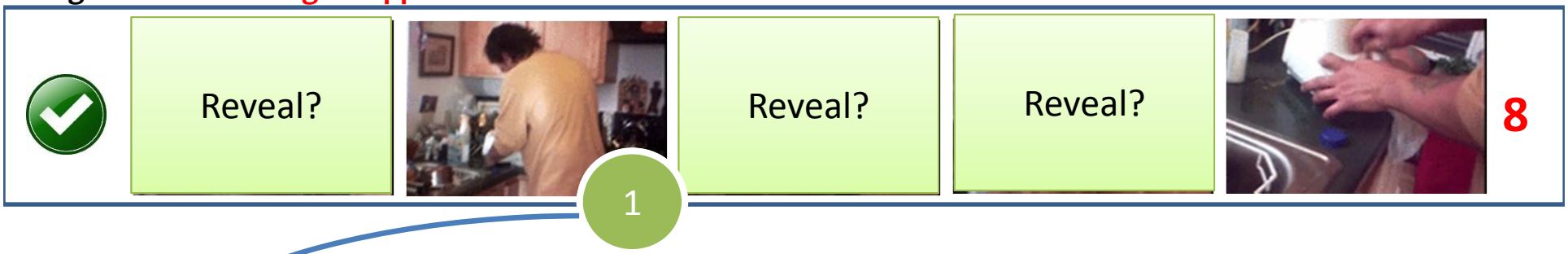
Retrieval Methodology

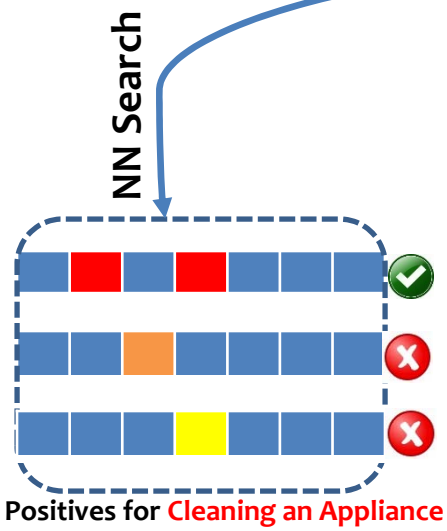
Target Event: Cleaning an Appliance



Rewarding Decisive Microshots

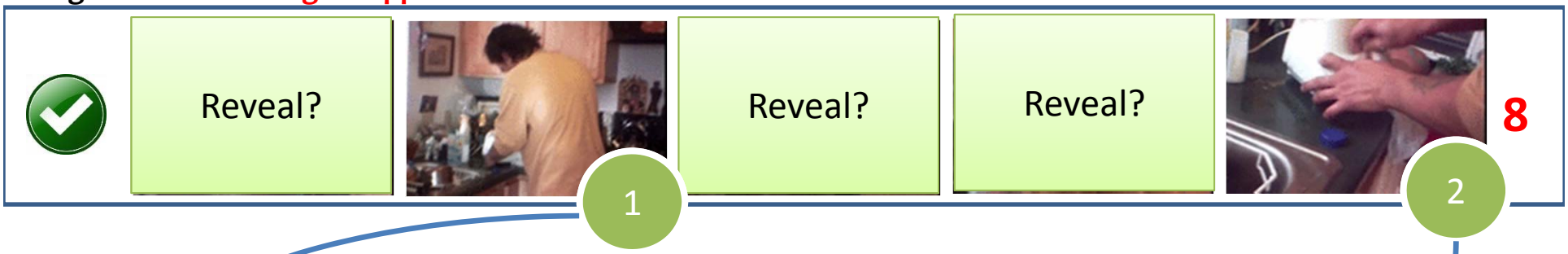
Target Event: Cleaning an Appliance



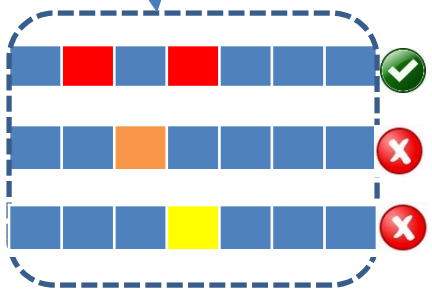


Rewarding Decisive Microshots

Target Event: Cleaning an Appliance



NN Search



Positives for Cleaning an Appliance (Likely)

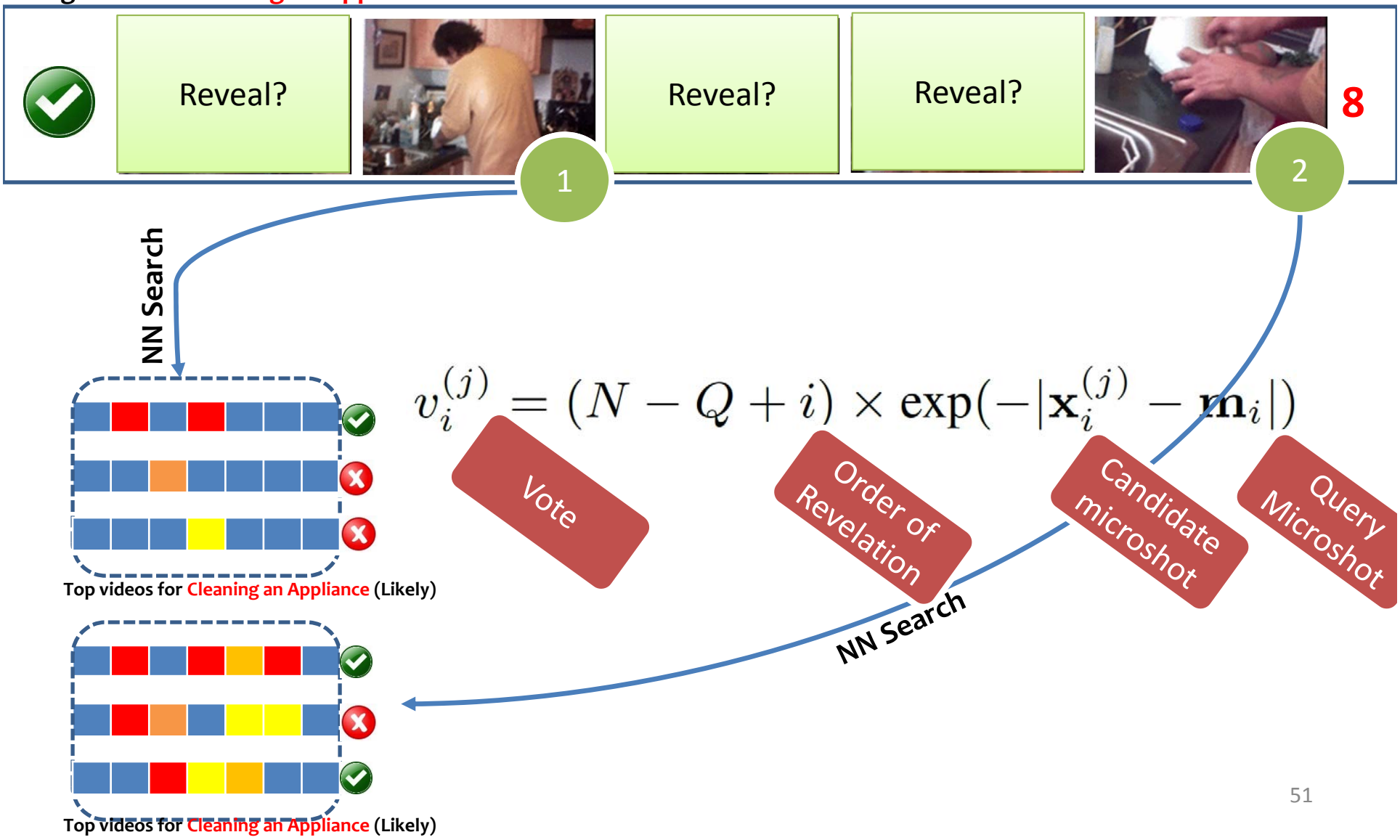
$$v_i^{(j)} = (N - Q + i) \times \exp(-|\mathbf{x}_i^{(j)} - \mathbf{m}_i|)$$

NN Search

Positives for Cleaning an Appliance (Likely)

Rewarding Decisive Microshots

Target Event: Cleaning an Appliance



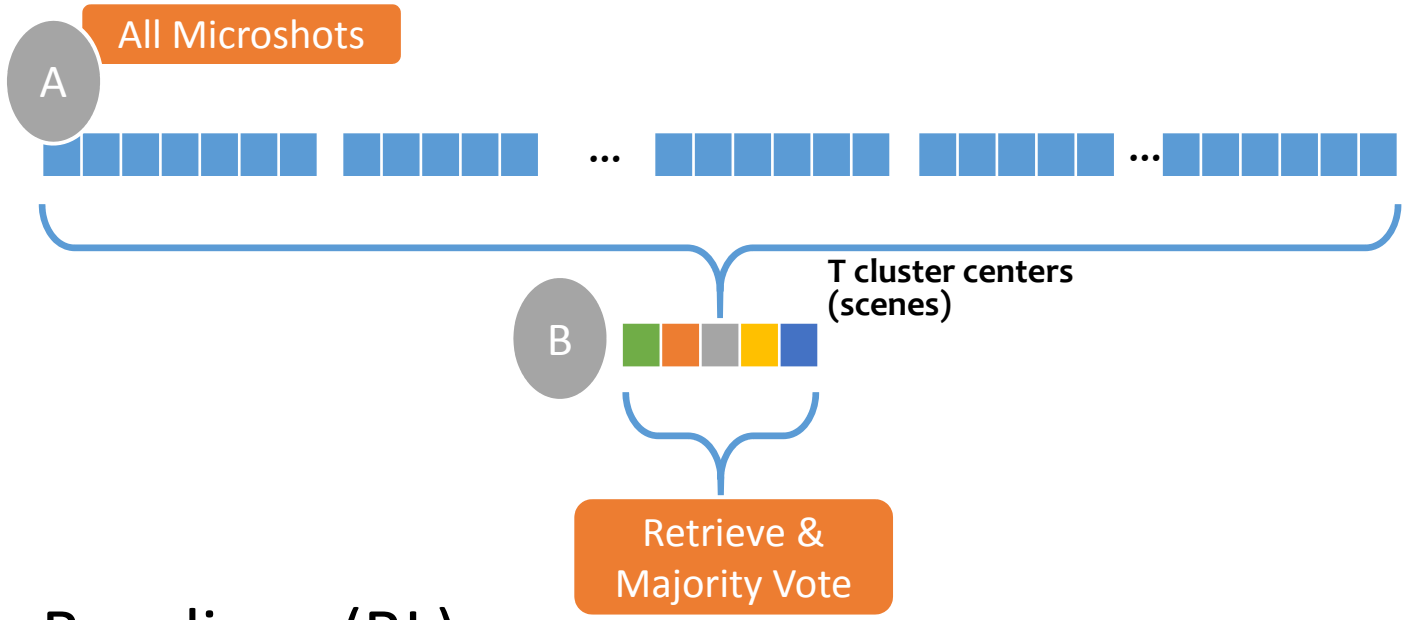
Experiments

ID	Event Name	[N]	ID	Event Name	[N]
E006	Birthday	173	E007	Changing Tire	111
E008	Flash-mob	173	E009	Vehicle Unstuck	132
E010	Grooming Animal	138	E011	Making Sandwich	126
E012	Parade	138	E013	Parkour	112
E014	Repairing Appl.	123	E015	Sewing Project	120
E021	Bike-trick	200	E022	Giving Direcitons	200
E023	Dog-show	200	E024	Wedding	200
E025	Marriage Proposal	200	E026	Renovating Home	200
E027	Rock-climbing	200	E028	Town-hall Meet	200
E029	Winning Race	200	E030	Metal crafts	200

Events
Beekeeping
Wedding shower
Non-motorized Vehicle repai
Fixing musical instrument
Horse riding competition
Felling a tree
Parking a vehicle
Playing fetch
Tailgating
Tuning musical instrument

- NIST Multimedia Event Detection TEST Dataset 2011-12 : 20 Events
- NIST Multimedia Event Detection ADHOC Dataset 2013 : 10 Events

Baselines



- Baselines (BL) :
  - A. Use all microshots
  - B. Use automatic microshot selection using scene aligned pooling [7]

[7] Liangliang Cao et. al. Scene aligned pooling for complex video recognition. In ECCV, 2012.

Retrieval Results

Events	BL-A	BL-B	MNE
Beekeeping	3.47	4.12	20.96
Wedding shower	2.87	2.05	17.23
Non-motorized Vehicle repair	2.56	3.35	16.90
Fixing musical instrument	3.52	3.09	19.26
Horse riding competition	4.60	5.21	21.46
Felling a tree	5.47	5.25	20.86
Parking a vehicle	3.09	6.11	17.04
Playing fetch	2.73	4.08	16.62
Tailgating	1.75	3.15	15.48
Tuning musical instrument	3.95	4.06	18.26
Mean Average Precision	3.41	4.07	18.47

- MED13 ADHOC data set
- Only using MNE, absolute performance gain ~ 14%

Use all microshots Scene Aligned Pooling based microshot selection

Retrieval Results

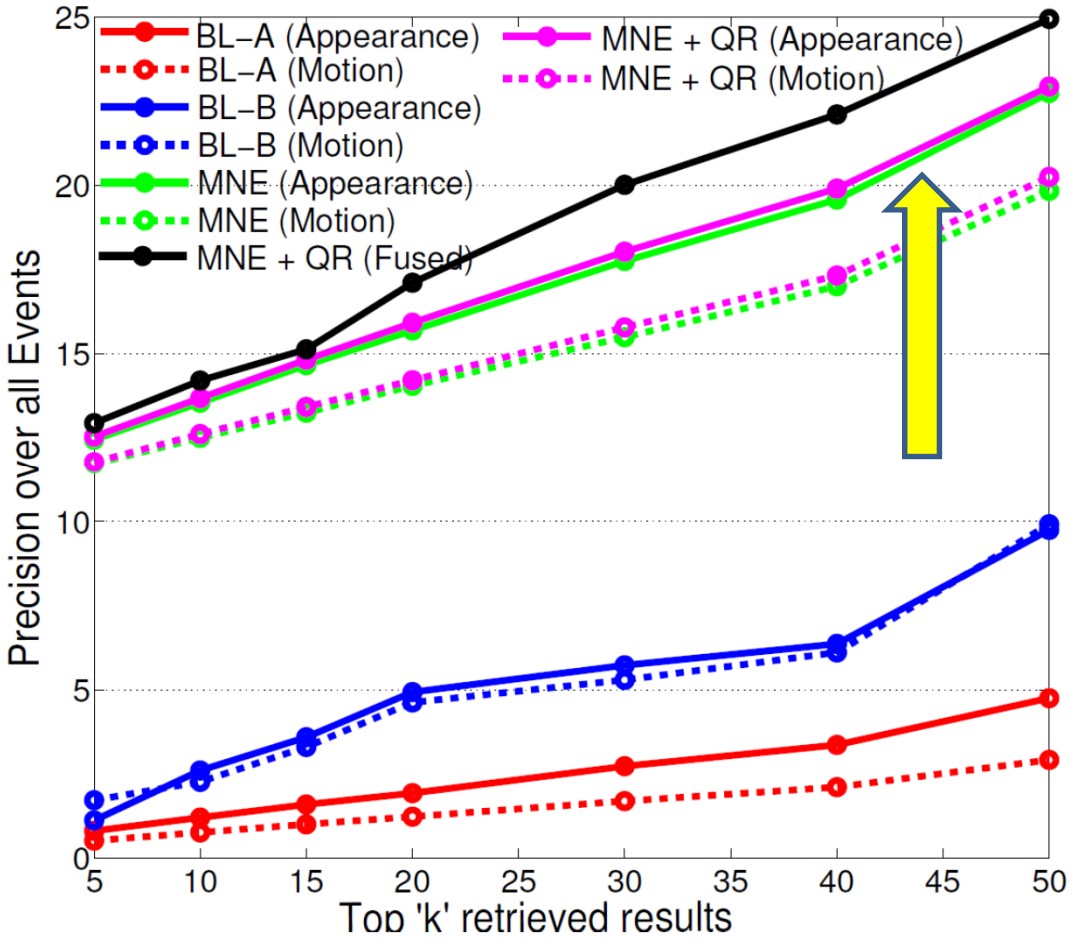
Events	BL-A	BL-B	MNE	MNE+QR
Beekeeping	3.47	4.12	20.96	20.86
Wedding shower	2.87	2.05	17.23	17.42
Non-motorized Vehicle repair	2.56	3.35	16.90	17.09
Fixing musical instrument	3.52	3.09	19.26	19.69
Horse riding competition	4.60	5.21	21.46	21.91
Felling a tree	5.47	5.25	20.86	21.27
Parking a vehicle	3.09	6.11	17.04	17.35
Playing fetch	2.73	4.08	16.62	16.74
Tailgating	1.75	3.15	15.48	14.97
Tuning musical instrument	3.95	4.06	18.26	18.56
Mean Average Precision	3.41	4.07	18.47	18.59

- Only using MNE, absolute performance gain ~ 14%
- Quick Rejection does not drastically improve the performance
- But can significantly speedup

Use all microshots | Scene Aligned Pooling based microshot selection | QR : Quick Rejection (Filtering Negatives)

Retrieval with Minimal Needed Evidences Achieves 4X Performance

- Cp. baselines: (A) all shots (B) scene clusters (Cao, et al, ECCV12)



What Concepts are Critical for Human Judgement?

- Discover needed concepts for humans
- Concepts those cannot be found from textual description of an event

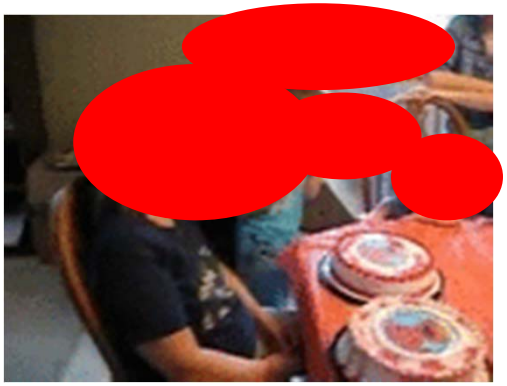
Birthday Party

Positive Concepts: "Yes, because I see …"

balloon **cake** candle chair clapping dining_table food gift
group_of_people indoor party_hat **person** singing
smiling_face table wine

Must-Have Concepts: "No, because I don't see …"

balloon **cake** candle person



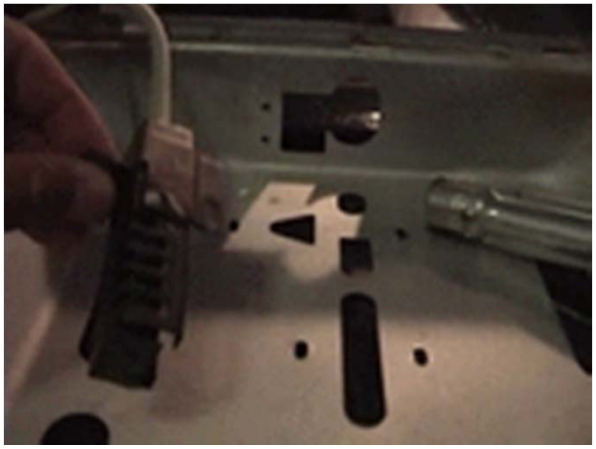
Concepts parsed from Textual Event Kit

person park streamer anniversary balloon outdoor **birthday**
blowing **cake** candle restaurant celebration children host conical
cupcake party indoor lit shiny/colorful singing food game gift guest home honor

Repairing an Appliance

Positive Concepts: "Yes, because I see …"

blender bolt closing_oven_door dewatering_machine grinder hand hand_holding_wire machine machine_parts metallic_sink nonfunctional_appliance opened_machine oven person_using_screwdriver person_using_tool person_using_tools radiator range screwdriver sink tool visible_hands wire



Must-Have Concepts: "No, because I don't see …"

indoor pointing_to_appliance pointing_to_applicances tool visible_hands

Concepts parsed from Textual Event Kit

adjusting_machine_part air_conditioners basement black_object clothes_dryers coffee_makers dishwasher drying_cabinet electric_toothbrushes food_processors freezer garage hair_dryers hand_mixers indoor induction_cookers kitchen kitchen_stove lifting_machine_parts machine_parts metallic_object microwave_ovens person person_bending_over person_holding_objects person_squatting power_tools rag refrigerator removing_machine_part replacing_machine_part screwing_parts small_machines stand_mixers toaster toaster_oven tool trash_compactor unscrewing_parts washing_machine water_heater white_object

Minimal Evidence Approach

- Discovers unique concepts needed for humans
- Must-Have concepts can be used for quick rejection
- Positive concepts can be used for efficient detection

Does Minimal Evidence help Event Recognition?

Positive Videos

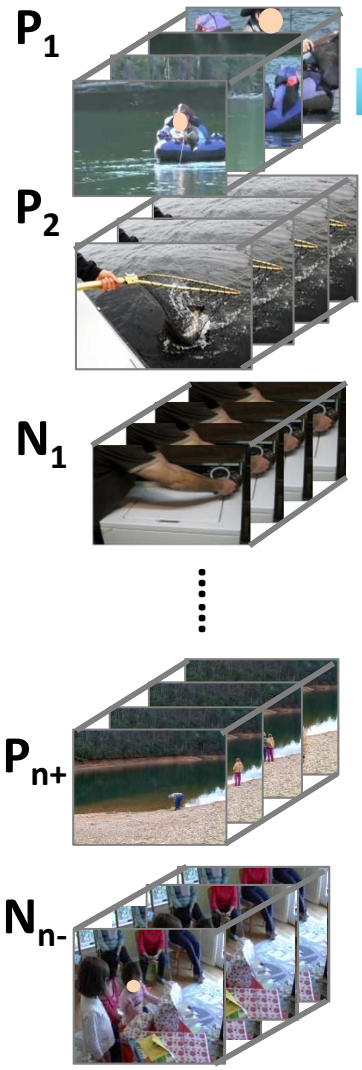
Where are evidences?

Negative Videos

K.T. Lai, F. X. Yu, M. Chen, S.-F. Chang. **Video Event Detection by Inferring Temporal Instance Labels**. In *CVPR* 2014.

Learn Key Evidences

K. Lai et al, ECCV 2014



1. Decompose videos into static & dynamic instances

static instances
(SIFT BoW frames)

Dynamic instances
(3, 5, 10, 15, 20secs MBH video clips)

2. Learn static/dynamic instance codebook by k-means

3. Create instance codewords

Static instances

5-sec dynamic instances

15 -sec dynamic instances

$$Cs = \{c_{s1}, c_{s2}, c_{s3}, ...\}$$

$$Cd = \{c_{d1}, c_{d2}, c_{d3}, ...\}$$

Embed video instances into codewords via max similarity measure:

$$s(S_i, c_l^s) = \max \text{pooling from video } S_i \text{ to codeword } c_l^s$$

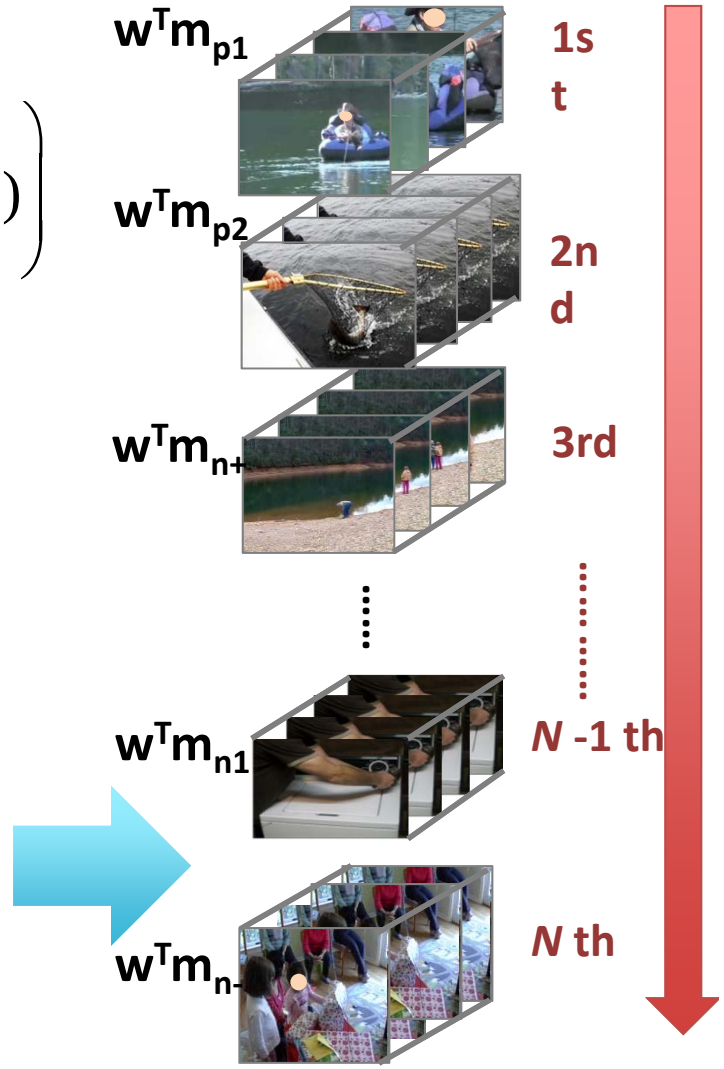
$$s(S_i, c_l^s) = \max_{1 \le j \le n_i} \left(\exp(-\frac{d(s_{ij}, c_l^s)}{\sigma}) \right)$$

$$\text{embedded representation for video S}$$

$$m_i = \left[s(S_i, c_1^s), \ldots s(S_i, c_{Gs}^s), s(D_i, c_1^d), \ldots s(D_i, c_{Gd}^d) \right]^T$$

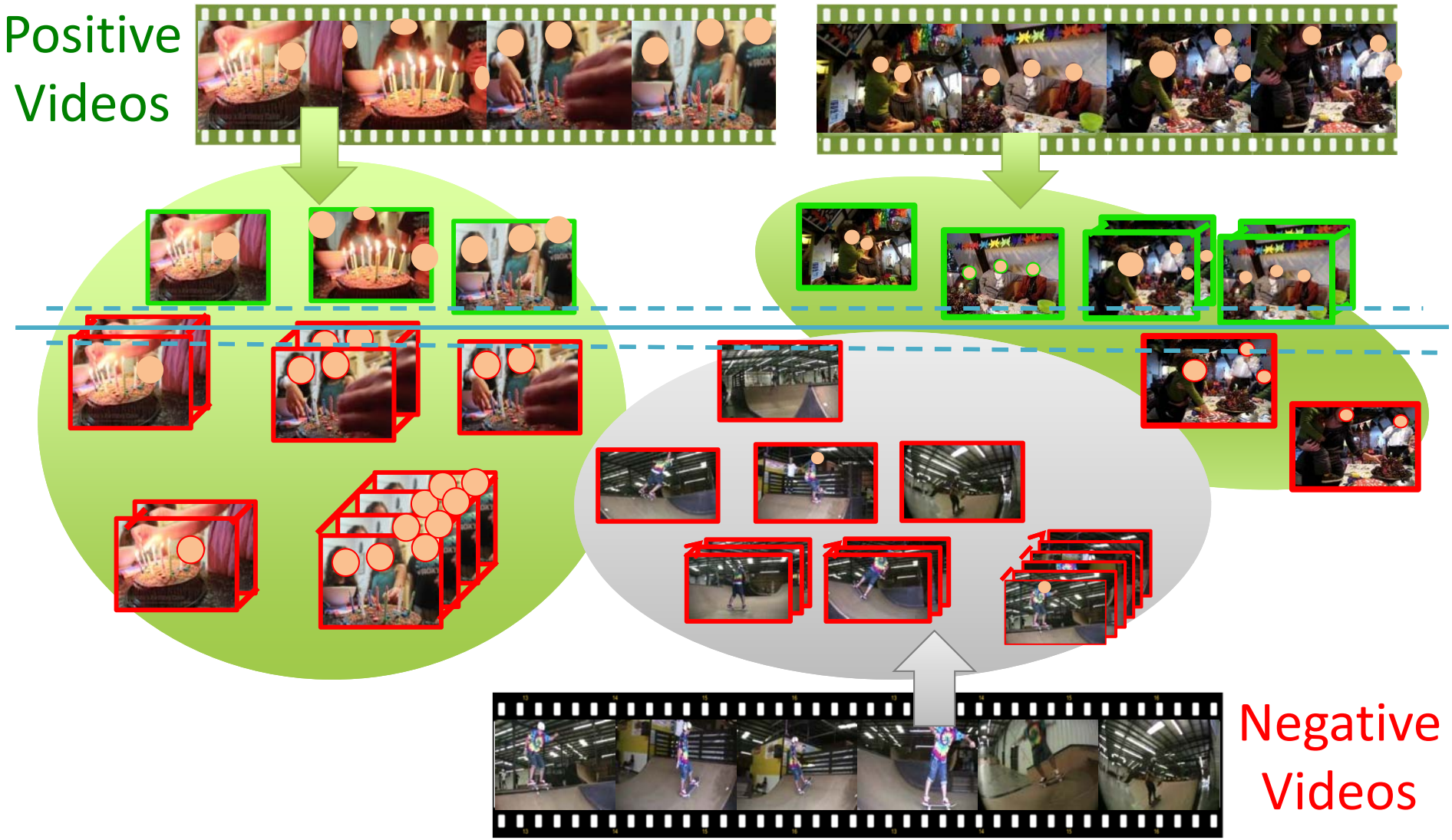
$$\min_{w} \quad \|w\|_1 + \lambda \max_{1 \le n \le N^-} \left(\frac{1}{N^+} \sum_{p=1}^{N^+} \max(1 - w^T (m_p^+ - m_n^-), 0) \right)$$

- Infinite push loss (Agarwal, 2011) used to push positive videos to the top ranks
- Other ranking orders ignored
- L_1 norm used to select key codewords and evidences
- Optimize with Alternating Direction Method of Multipliers



Event Name (006 - 015)	DMS [21]	VD-HMM [22]	SPP [10]	MKL-KLSVM[23]	MILES (SVM-$l1$)	Our Method
Birthday party	2.25%	4.38%	6.08%	6.24%	5.08%	**7.45%**
Change a vehicle tire	0.76%	0.92%	3.96%	**24.62%**	9.50%	14.44%
Flash mob gathering	8.30%	15.29%	35.28%	37.46%	33.77%	**40.87%**
Get a vehicle. unstuck	1.95%	2.04%	8.45%	**15.72%**	7.38%	7.72%
Groom an animal	0.74%	0.74%	**3.05%**	2.09%	1.76%	1.83%
Make a sandwich	1.48%	0.84%	4.95%	**7.65%**	3.13%	4.86%
Parade	2.65%	4.03%	8.95%	12.01%	14.34%	**17.69%**
Parkour	2.05%	3.04%	24.62%	10.96%	20.14%	**25.3%**
Repair an appliance	4.39%	10.88%	19.81%	**32.67%**	25.81%	31.75%
Work on sewing project	0.61%	5.48%	6.53%	7.49%	4.66%	**8.34%**
mean AP	2.52%	4.77%	12.27%	15.69%	12.56%	**16.02%**

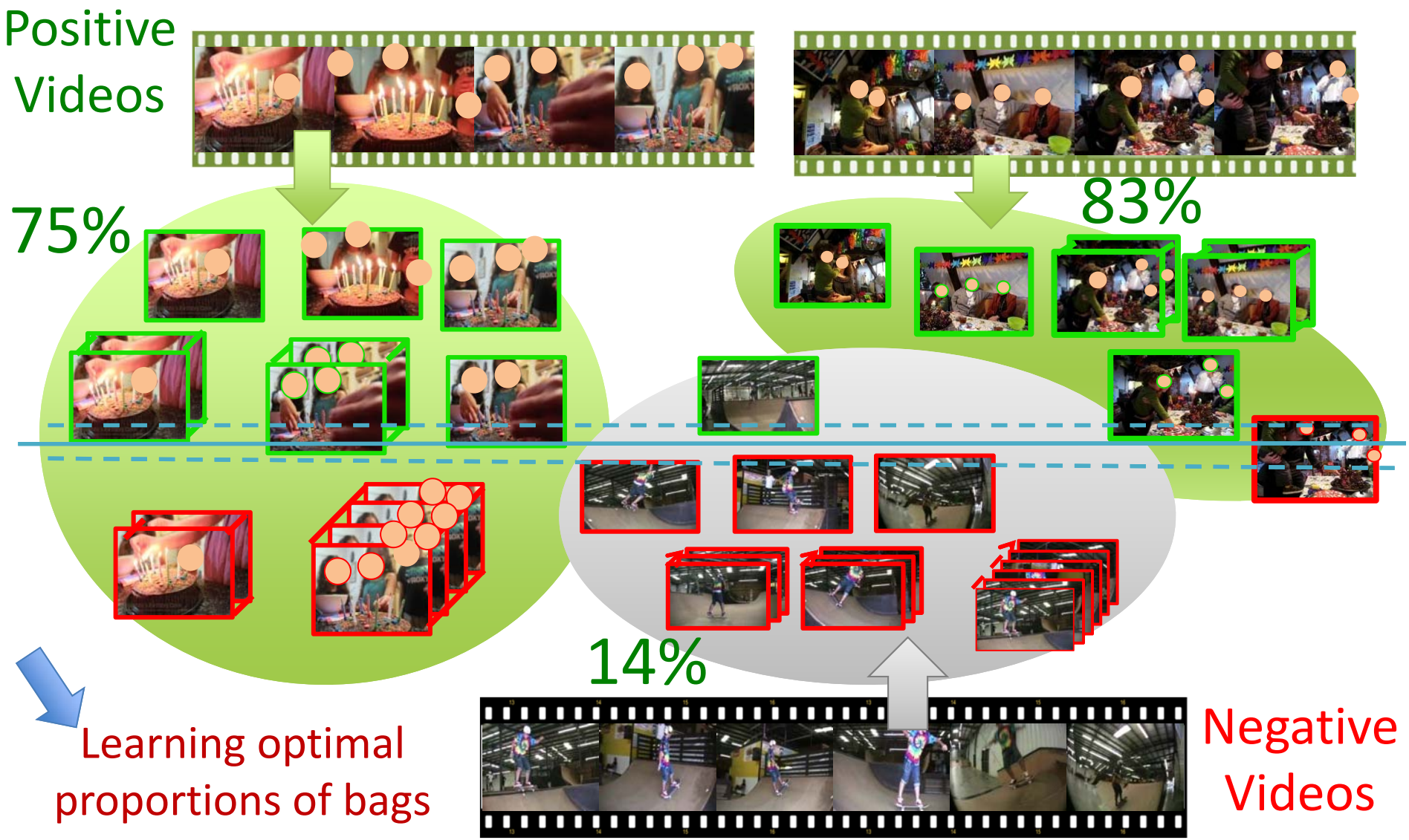
Another Formulation: Multiple Instance Learning



MIL assumptions not right

- Negative videos may still have partial evidences
- Positive videos should have more (critical) evidence proportions

Relax MIL Setting – Soft Bag Proportions



The ∝SVM Algorithm

F. Yu; D. Liu; S. Kumar; T. Jebara; S.-F. Chang. ∝SVM for learning with label proportions. **ICML13**

- Large-margin framework:

Microshot as instances

Video as Bags, Learn label proportion for each bag

$$\min_{\mathbf{y},\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} L(y_i, \mathbf{w}^T\varphi(\mathbf{x}_i) + b) + C_p \sum_{k=1}^{K} L_p\left(\tilde{p}_k(\mathbf{y}), p_k\right)$$

$$\text{s.t.} \quad \forall_{i=1}^{N}, \quad y_i \in \{-1, 1\}.$$

- Generalizes the classic SVM.
- Naturally spans supervised/semi-supervised learning and clustering.
- Learned with alternate optimization or a relaxed convex form

Dealing with Unknown Proportion

- But bag proportion is unknown a priori
- Set initial proportion P_m to 1 (0) for positive (negative) videos

$$\min_{\{\mathbf{y}^m\}_{m=1}^M, \mathbf{w}, b} \frac{1}{2}||\mathbf{w}||^2 + C \sum_{m=1}^{M} \sum_{i=1}^{N_m} L\left(y_i^m, (\mathbf{w}^\top \mathbf{x}_i^m + b)\right) + C_p \sum_{m=1}^{M} |p_m(\mathbf{y}^m) - P_m| \quad (5)$$

$$s.t. \quad P_m = \begin{cases} 1 & \text{if } Y_m = 1 \\ 0 & \text{if } Y_m = -1 \end{cases}, m = 1, \cdots, M.$$

Infer latent variables: bag proportion and instance labels

Learn SVM classifier w

Optimization Procedure

- Alternating optimization for p-SVM
  1. Fix instance labels $\boldsymbol{y}$ and learn $\boldsymbol{w}$ and b. The problem becomes a classic SVM

  2. Fix $\boldsymbol{w}$ and b *and update instance labels* $\boldsymbol{y}$*, calculate positive instance proportions* $p(\boldsymbol{y})$

  3. Optimize proportions of videos independently

     3-1. Compute all possible loss values (prediction loss + proportion loss) by flipping instance labels one by one

     3-2. Sort all loss values to find optimal instance proportion

Complexity Analysis

- Running time of one iteration

  = SVM training time + instance sorting time

  = $O(N + I_{max} \log I_{max})$, I_{max} = max number of instances in a video

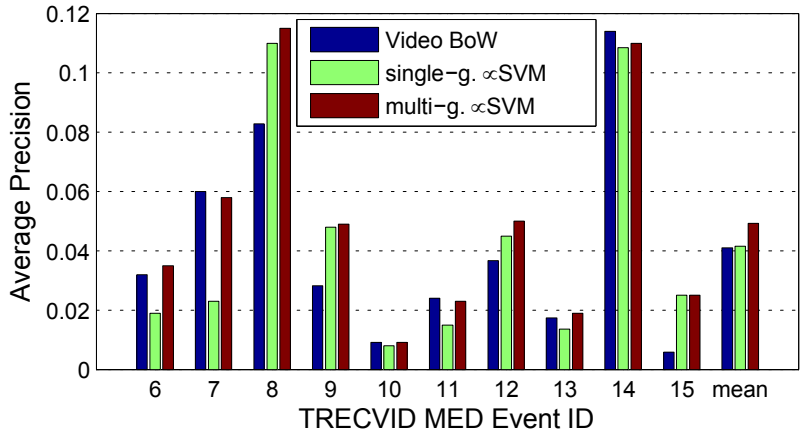
- SVM training time >> instance sorting time, so the complexity is the same as SVM

Also Learn Best Granularities for Event Evidences

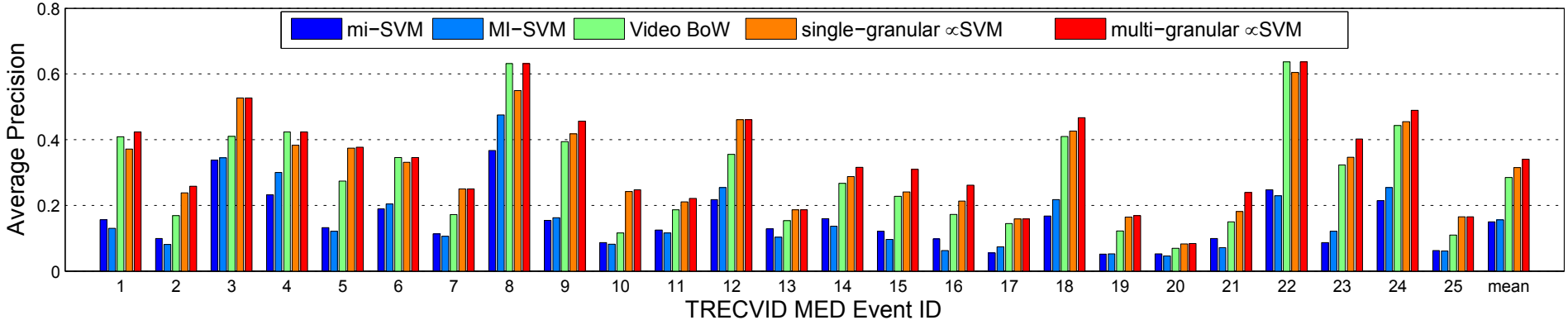
Optimal Granularity	Events
Single frame (SIFT)	5. Woodworking; 7. Changing a tire; 19. Give directions to location; 22. Rock climbing; 24. Win race without vehicle; 25. Metal craft project
3 sec (MBH)	2. Feeding animals; 3. Landing a fish; 4. Wedding ceremony; 8. Flash mob gathering; 11. Making sandwich; 12. Parade; 13. Parkour; 18. Dog show; 23. Town Hall Meeting
5 sec (MBH)	6. Birthday party; 9. Getting vehicle unstuck; 10. Grooming animal; 14. Repairing appliance 16. Attempting bike trick; 17. Cleaning appliance; 21. Renovating a home
10 sec (MBH)	none
15 sec (MBH)	1. Attempting board trick; 15. Sewing project
20 sec (MBH)	20. Marriage proposal

Experimental Results

- 20% performance gain on MED11 dataset



- 10% performance gain on MED12 dataset



Rock Climbing

Detected Evidences

Optimal granularity: 1 frame

Rock Climbing

Detected Evidences

Optimal granularity: 1 frame

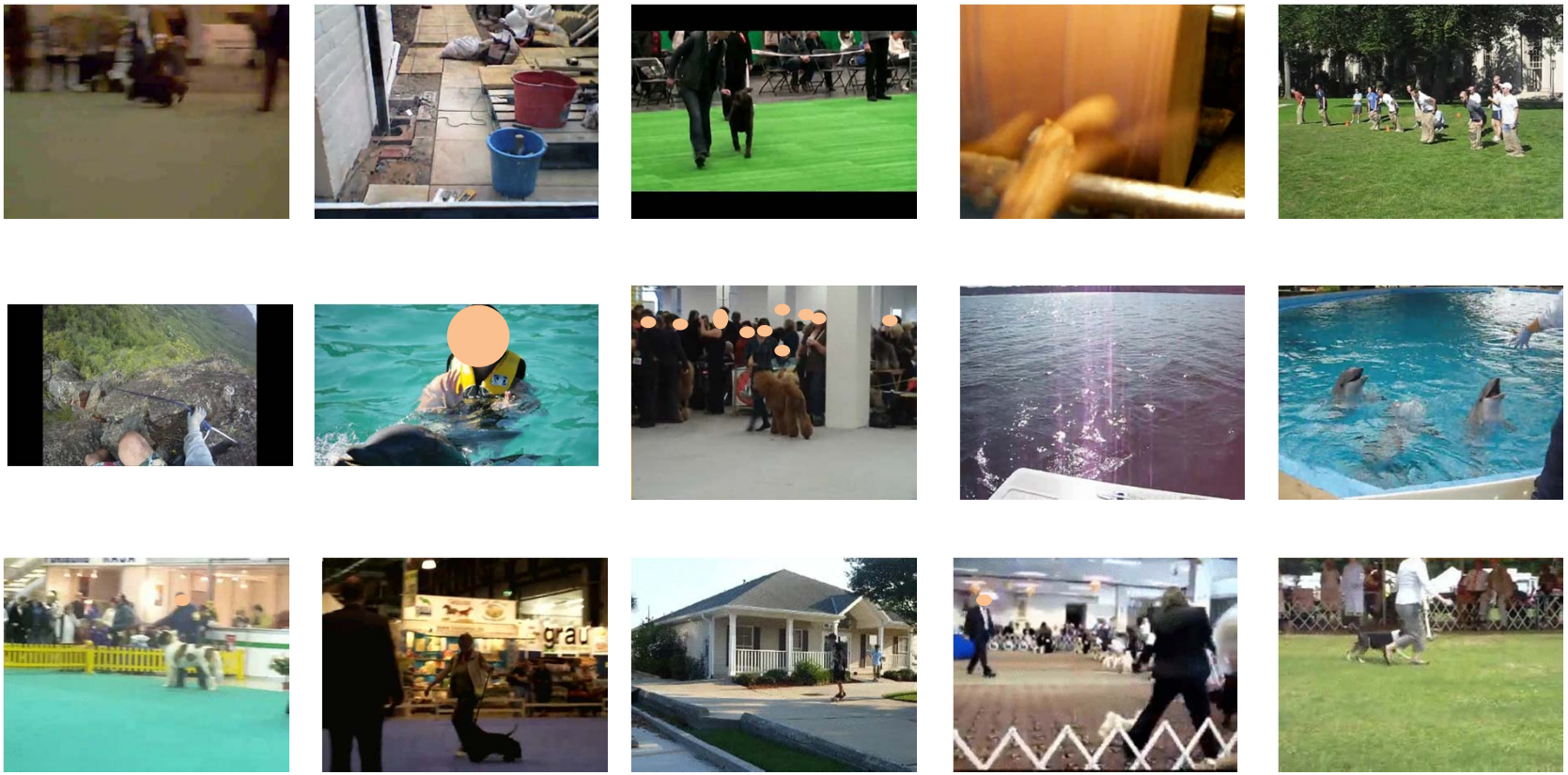
Winning Race without a Vehicle

Detected Evidences

Optimal granularity: 1 frame

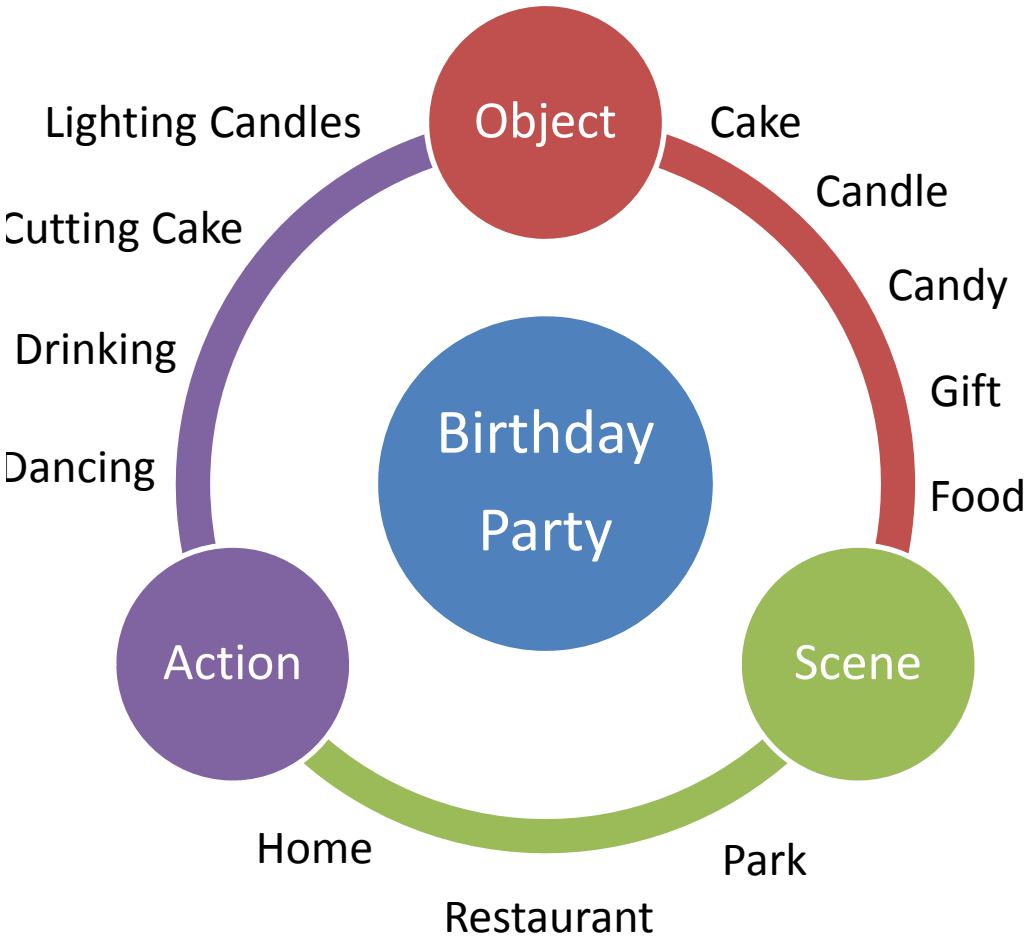
Winning Race without a Vehicle

Detected Evidences for Dog Show



Optimal granularity:
3 seconds

Conclusions



- Concepts as Event mid-level representation
- Questions:
  - How to find/combine them?
  - How to train classifiers?
  - How to locate evidences in videos?
  - How to incorporate temporal dynamics?
- Tools :
  - Preparing EventNet
  - About 1,000 events & ~12,000 Concepts

References

1. Yu-Gang Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, Mubarak Shah. **High-Level Event Recognition in Unconstrained Videos**. International Journal of Multimedia Information Retrieval, 2(2):73-101, 2013.
2. Jiawei Chen, Yin Cui, Guangnan Ye, Dong Liu, Shih-Fu Chang. **Event-Driven Semantic Concept Discovery by Exploiting Weakly Tagged Internet Images**. In *ACM International Conference on Multimedia Retrieval (ICMR), full paper (oral)*, 2014.
3. Yin Cui, Dong Liu, Jiawei Chen, and Shih-Fu Chang. "Building A Large Concept Bank for Representing Events in Video." *arXiv preprint arXiv:1403.7591* (2014).
4. J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In IEEE CVPR, 2013.
5. S. Bhattacharya, F.X. Yu, S.-F. Chang. **Minimally Needed Evidence for Complex Event Recognition in Unconstrained Videos**. In *ACM Conf. on Multimedia Retrieval (ICMR),* April 2014.
6. Kuan-Ting Lai, Dong Liu, Ming-Syan, Chen, Shih-Fu Chang. **Recognizing Complex Events in Videos by Learning Key Static-Dynamic Evidences**. In *European Conference on Computer Vision (ECCV)*, September 2014.
7. S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In NIPS, 2002
8. Kuan-Ting Lai, Felix X. Yu, Ming-Syan Chen, Shih-Fu Chang. **Video Event Detection by Inferring Temporal Instance Labels**. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), oral*, Columbus, OH, June 2014.
9. Felix Yu, Dong Liu, Sanjiv Kumar, Tony Jebara, Shih-Fu Chang. **$\propto$ SVM for learning with label proportions**. In *International Conference on Machine Learning (ICML) (full oral)*, Atlanta, GA, June 2013.