

Minimally Needed Evidence for Complex Event Recognition

Subhabrata Bhattacharya, Felix X. Yu, Shih-Fu Chang

Acknowledgments: This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20070. 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.

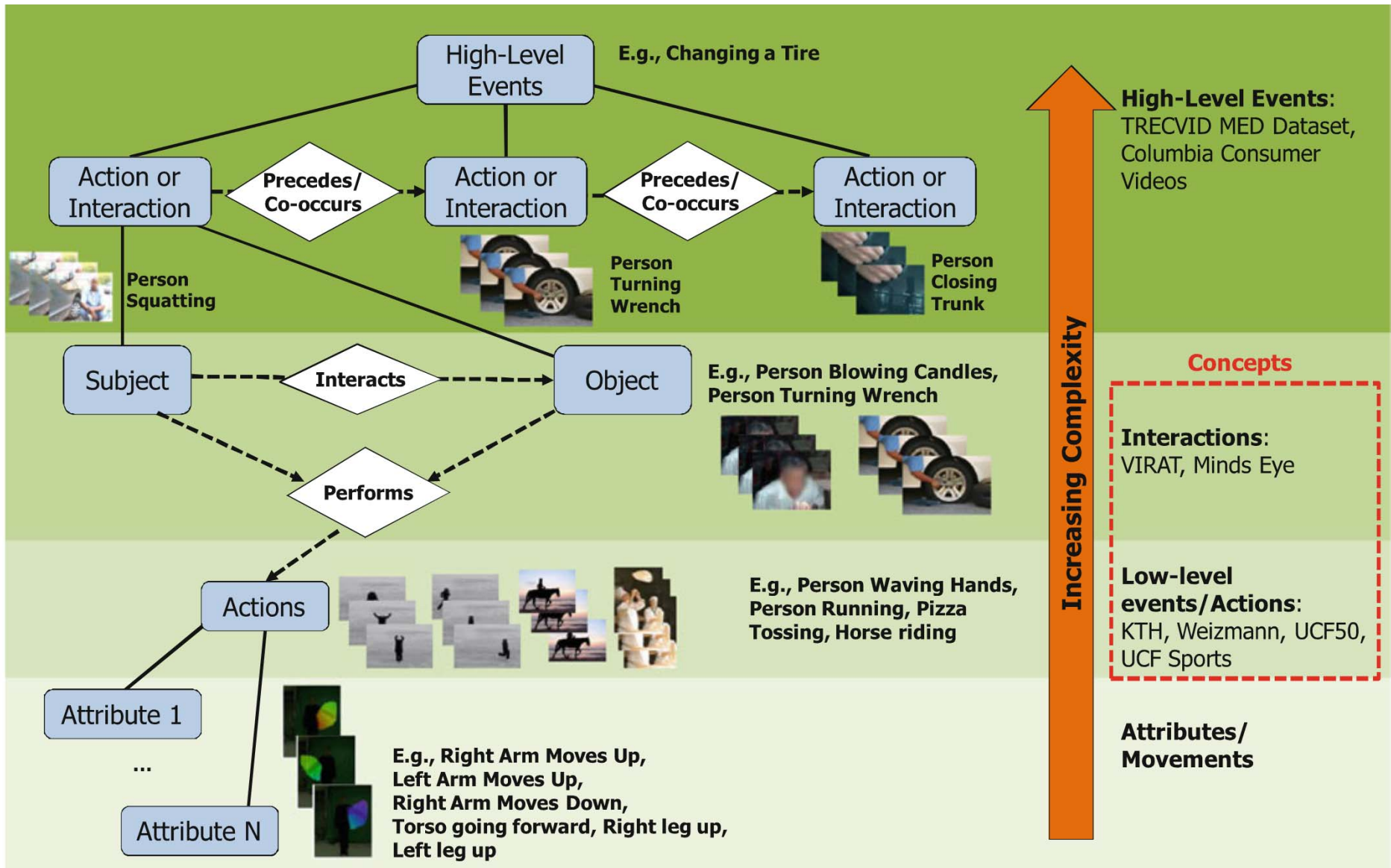


COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science



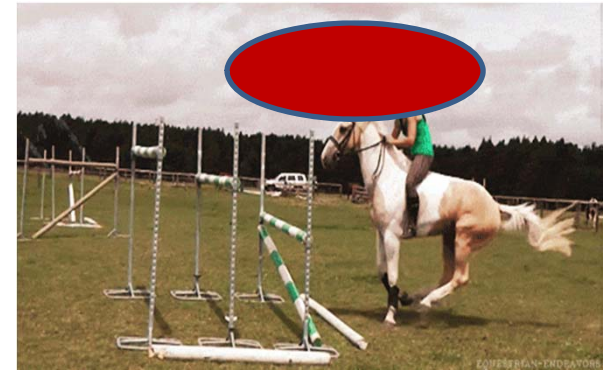
Complex (High-level) Events

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



Detecting Complex Events

What “event” is described in these videos?



Detecting Complex Events

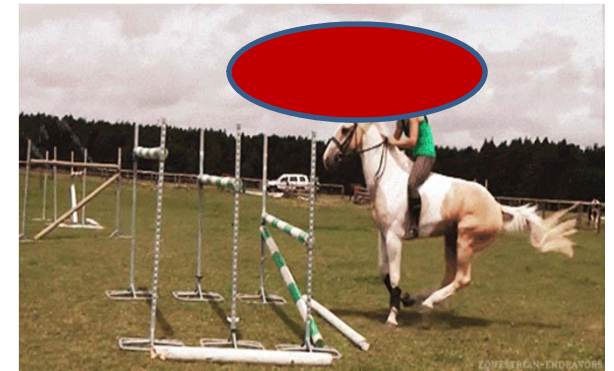
What “event” is described in these videos?



Felling a Tree



Attempting Board Trick



Horseriding Competition

Applications

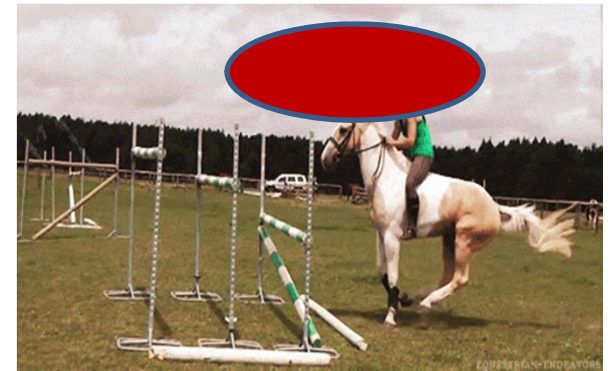
What “event” is described in these videos?



Felling a Tree



Attempting Board Trick



Horseriding Competition

Automatic Detection will help:



people marching, person walking,
person clapping, vehicle moving,
person dancing

...



people marching, person dancing,
person clapping, person walking,
vehicle moving

Summarization for Analytics

Applications

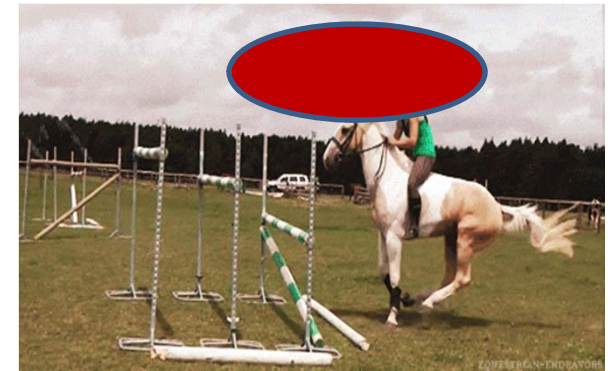
What “event” is described in these videos?



Felling a Tree



Attempting Board Trick



Horseriding Competition

Automatic Detection will help:



people marching, person walking,
person clapping, vehicle moving,
person dancing



people marching, person dancing,
person clapping, person walking,
vehicle moving

Video Summarization



It's flavor at
First sight.



Targeted Advertisement/Learning

Current Paradigm



Passive **Linear Video Playback** and Judge (Conventional)



Current Paradigm



Passive **Linear Video Playback** and Judge (Conventional)



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


Watch and Click [1]

During this study, you need to complete **3 tasks**.
In each task, a video clip will be played **2 times**.

Please watch each video carefully from the beginning to the end.

Whenever the video reaches its **highlights**, please press **SPACE key** on your keyboard.

Previous Next



[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

Related Work



Passive **Linear Video Playback** and Judge (Conventional)



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

Watch and Click [1]

During this study, you need to complete **3 tasks**.
In each task, a video clip will be played **2 times**.

Please watch each video carefully from the beginning to the end.

Whenever the video reaches its **highlights**, please press **SPACE key** on your keyboard.

Previous

Next



[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

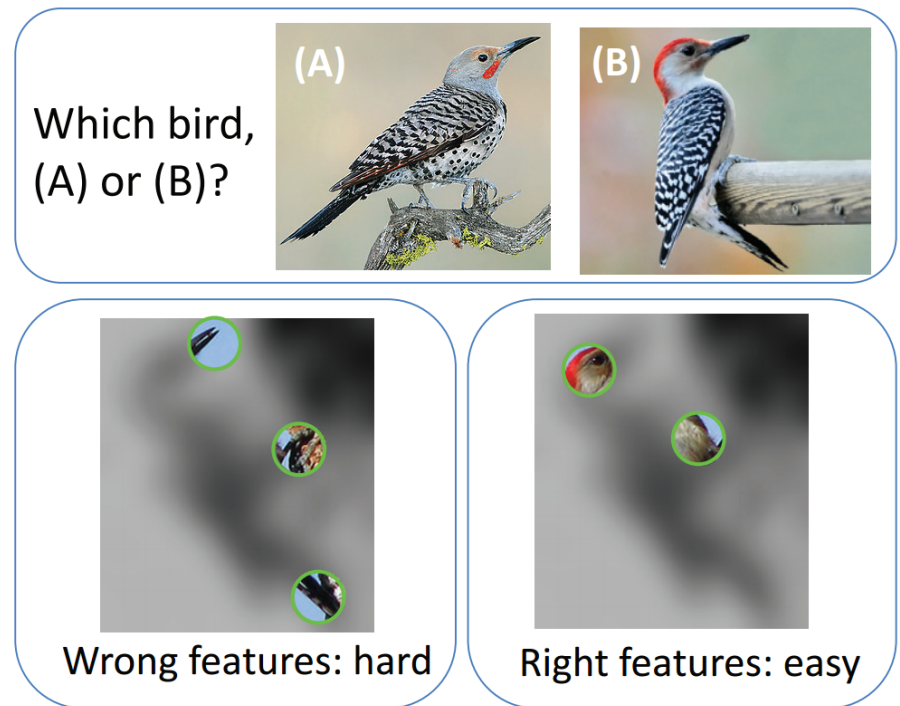
Motivation

- Do humans actually employ linear playback to judge a complex event?
- 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.

Motivation

- Do humans actually employ linear playback to judge a complex event?
- Bubble-game [5] to identify “Human” discovered discriminative patches for fine-grained recognition



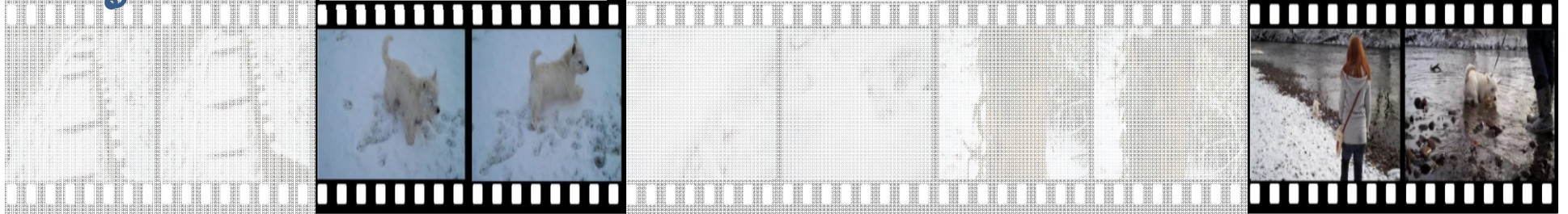
[5] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In IEEE CVPR, 2013.

Evidences needed for human judgment?

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




- Not necessarily linear
- Not necessarily sequential





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





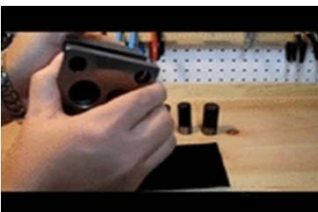

Lets take a test ...

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 ...

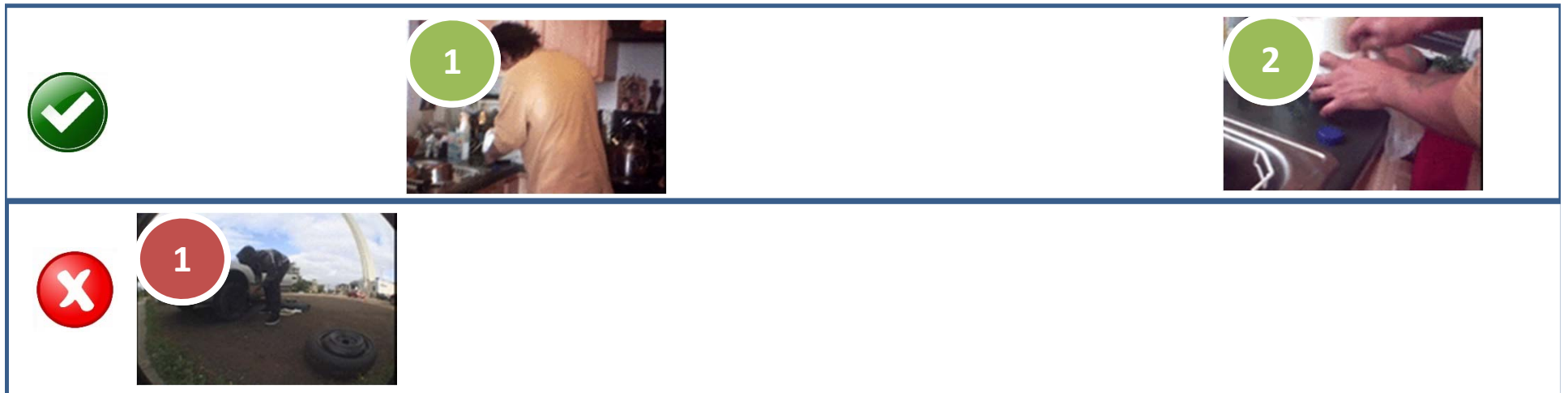
Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

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

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

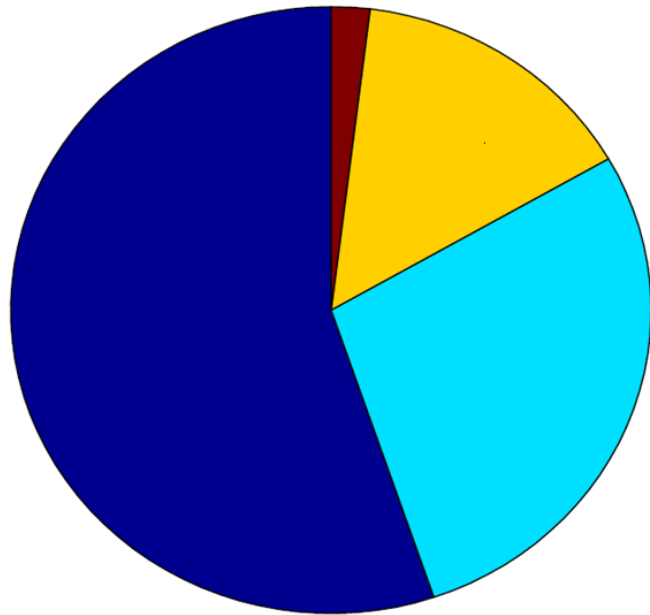
Minimally Needed Evidence

For the event Cleaning an appliance:

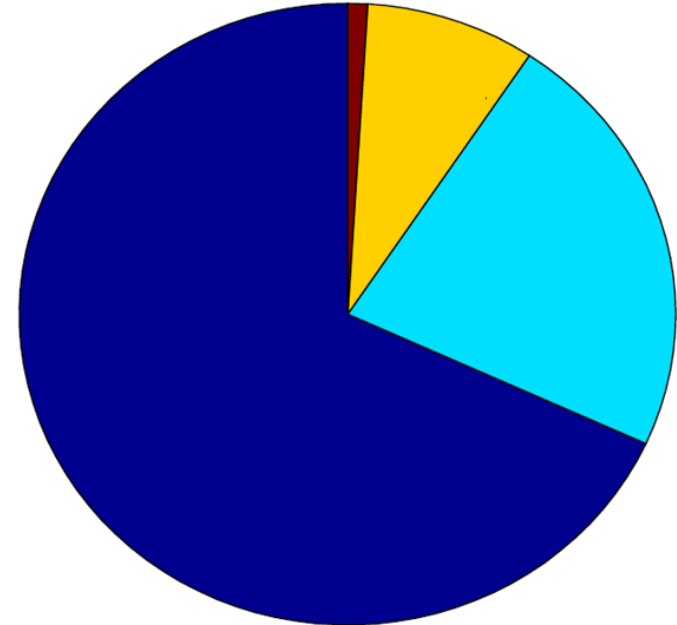


- **Practical way** of finding Minimally Needed Evidence
→ **Event Quiz Interface**
- **Clever annotation tool** → Enables **judicious use** of Human feedback
- Can **reduce computational overhead** for feature extraction

Surprisingly, Humans can



Correctly Identify (positives)



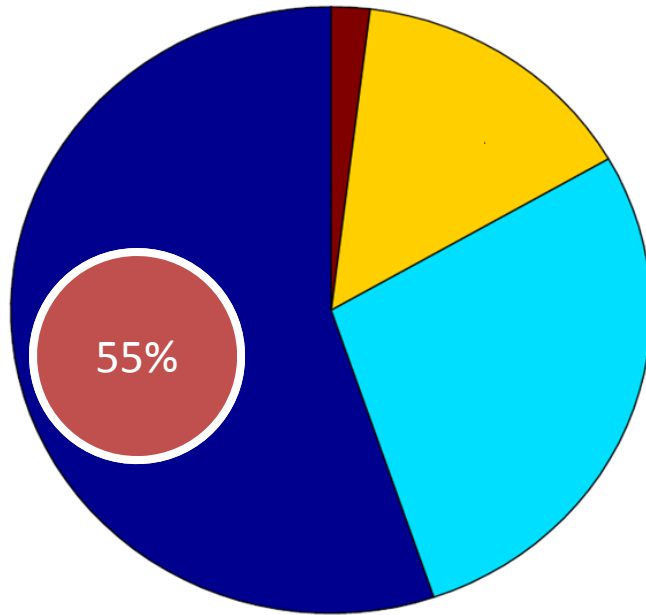
Correctly Reject (negatives)

Number of microshots Revealed

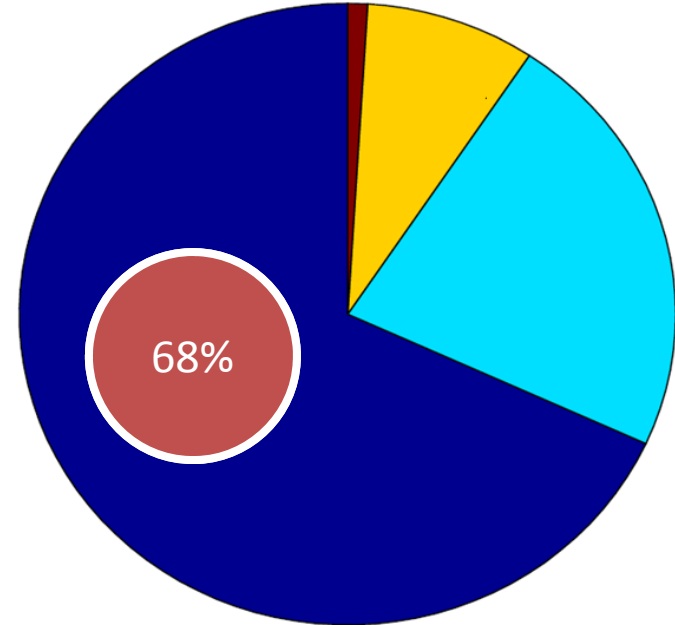


- Correctly judge an event **in ~87% cases** from just **1 microshot** (1.5s of footage)

Surprisingly, Humans can



Correctly Identify (positives)



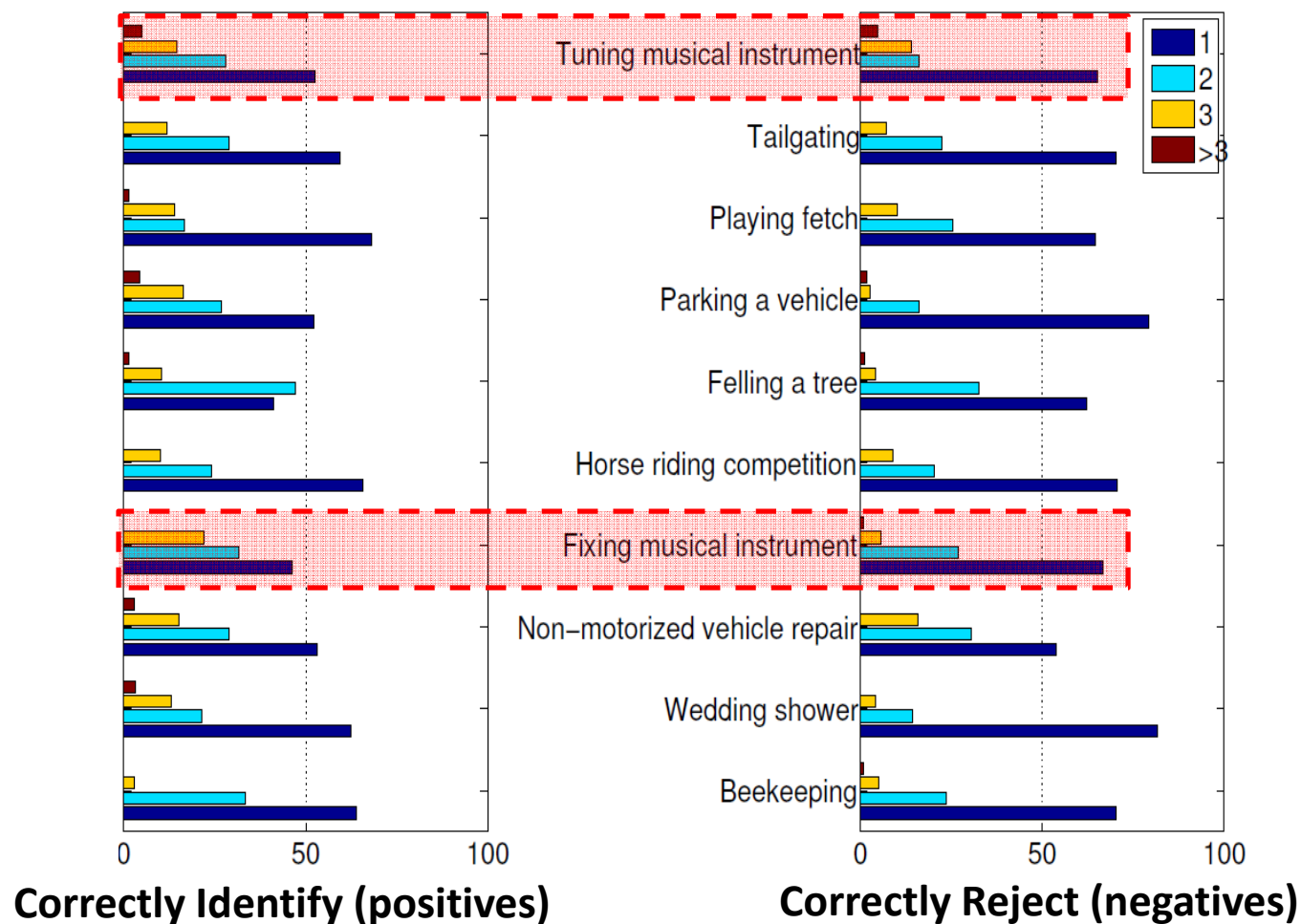
Correctly Reject (negatives)

Number of microshots Revealed



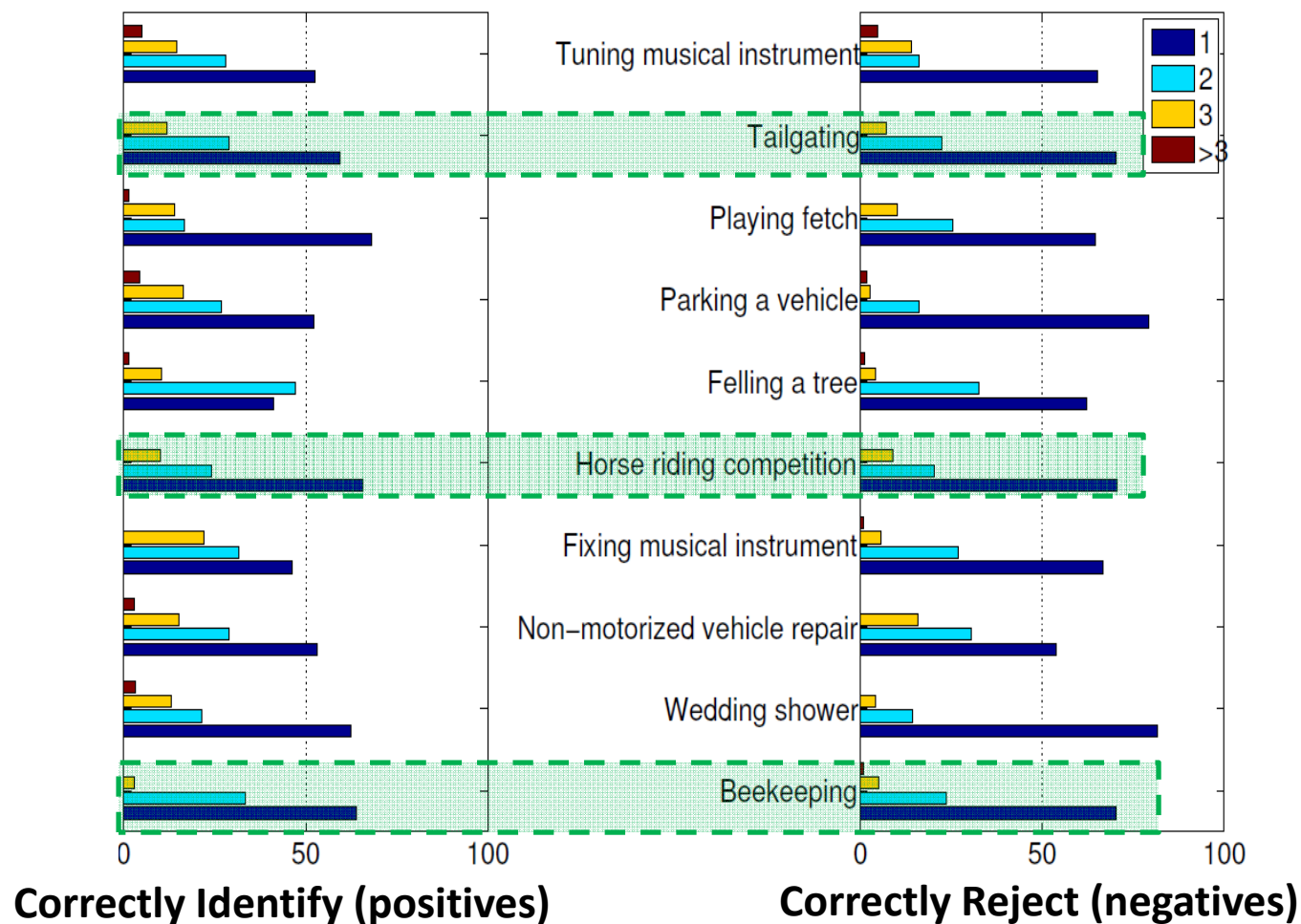
- Correctly identify a video containing an event in videos in 55% cases
- Correctly reject a video for not containing an event in 68% cases

Additionally- Event Complexity varies



- “Tuning a musical instrument” **more visually challenging** than “Fixing musical instrument” (take 4% more microshot revelations than other events)

Additionally- Event Complexity varies



- “Tailgating”, “Horseriding Competition” and “Beekeeping” **require less evidence microshots**

Microshot Selection

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

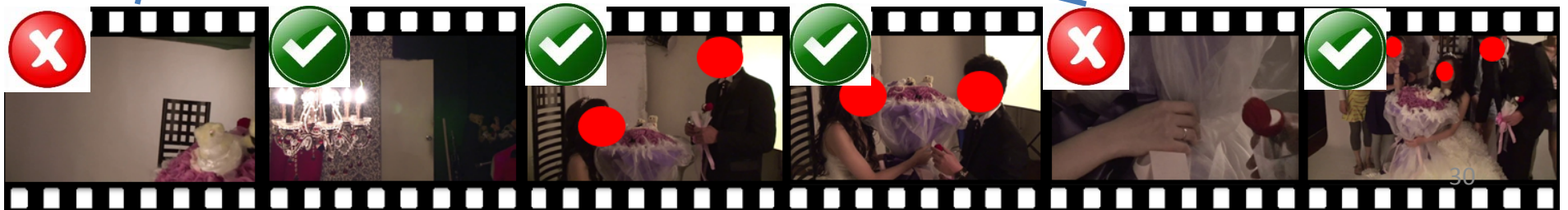


Microshot Selection

- 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)



Microshot Selection

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

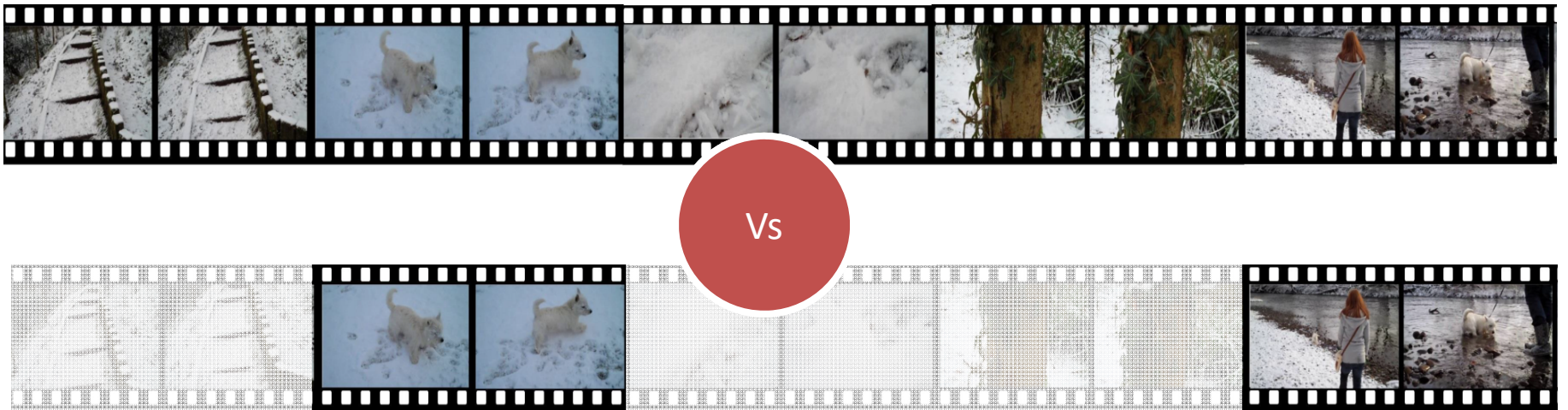


- Divide video into non-overlapping 1.5s blocks; Filter **out non-interesting** microshots (low **appearance** + **motion** entropy)

$$I = -\alpha \underbrace{\sum_{i=1}^M P_a(i) \log[P_a(i)]}_{\text{Appearance Component}} - \beta \underbrace{\sum_{i=1}^M P_m(i) \log[P_m(i)]}_{\text{Motion Component}}$$

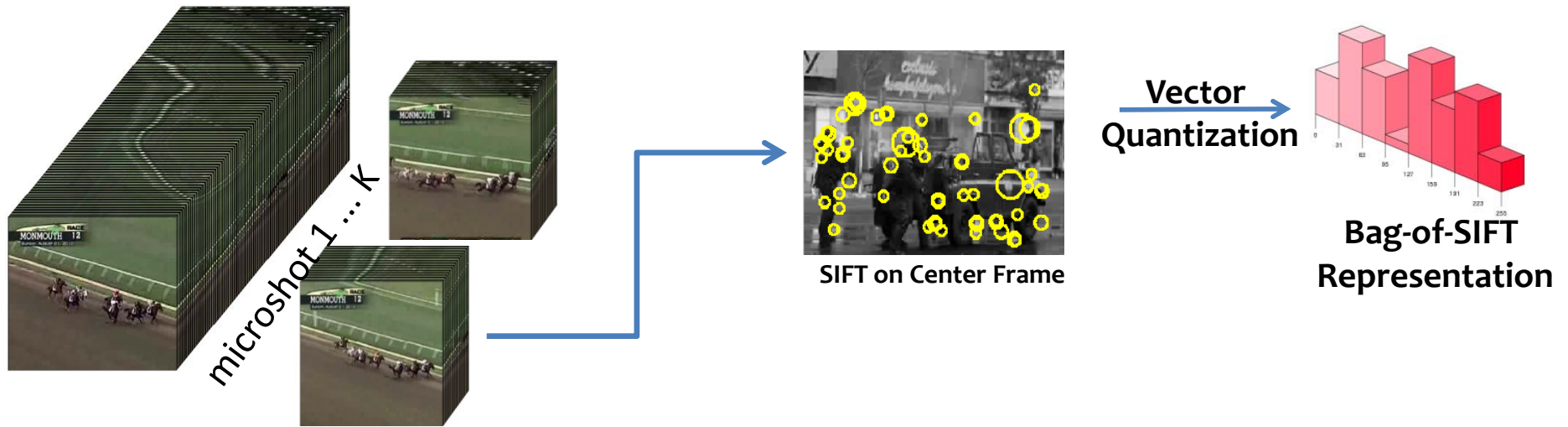
Hypothesis

- Human Discovered MNE provide better event representation for recognition



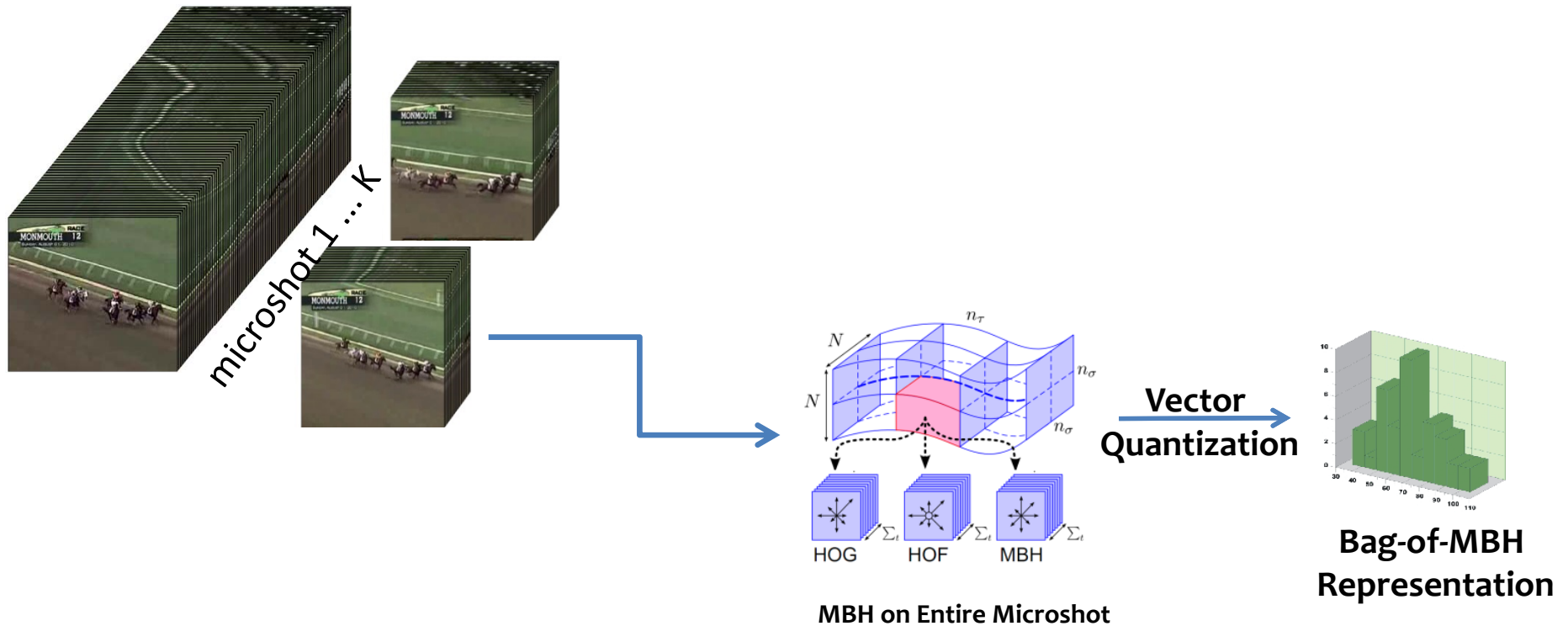
- To validate: Basic Retrieval experiment

Representation for Retrieval



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

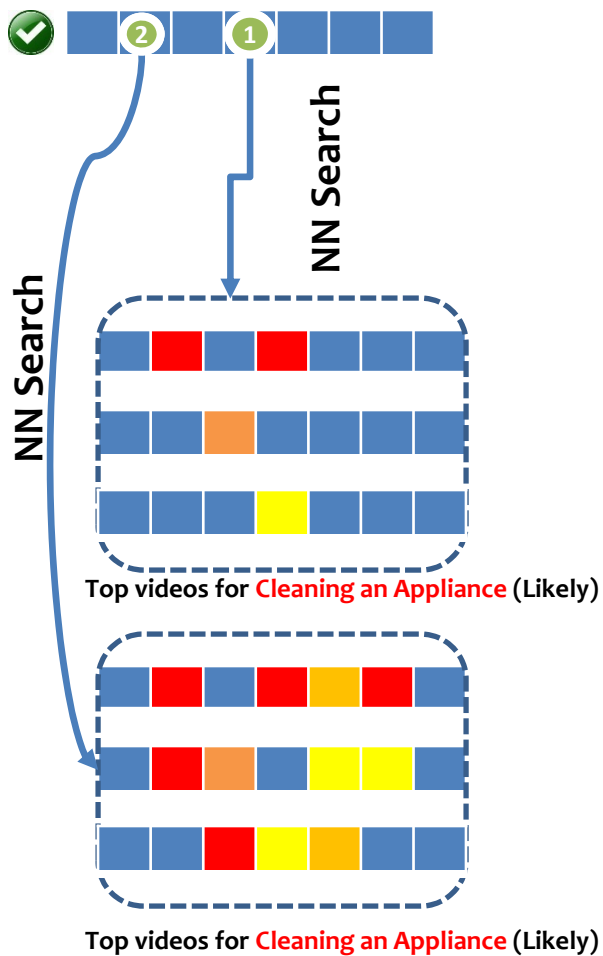
Representation for Retrieval



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

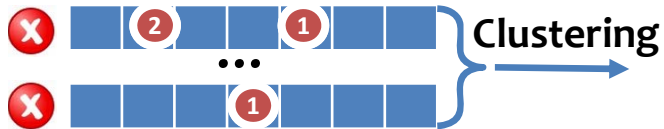
Retrieval Methodology

Target Event: **Cleaning an Appliance**



Use Negative Cues for Quick Reject

Target Event: **Cleaning an Appliance**



Top-3 Negative Clusters for
“Cleaning an Appliance”

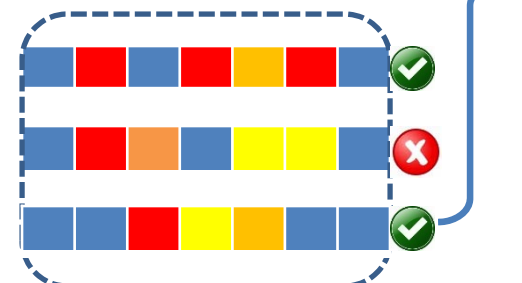
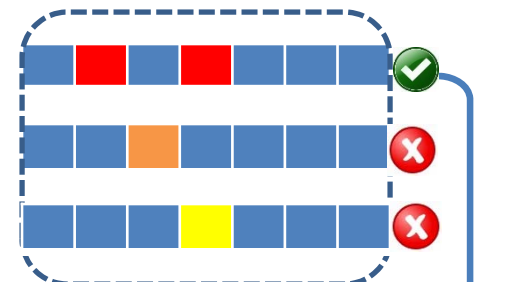
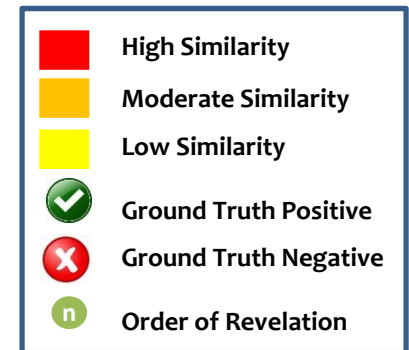
Train 1-class
SVM



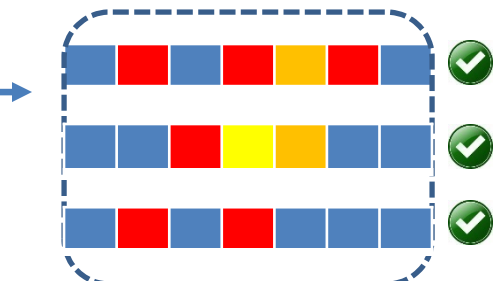
1-class model to
filter non (**Cleaning
an Appliance**)

Retrieval Methodology

Target Event: **Cleaning an Appliance**



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



Top videos for **Changing an Appliance** (Likely)

Rewarding Decisive Microshots

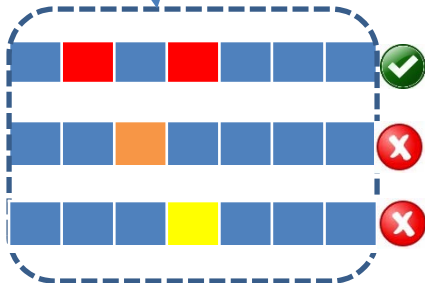
Target Event: **Cleaning an Appliance**

Timeline of a video clip:

- Green checkmark icon
- Green box: Reveal?
- Image: Person in a kitchen
- Green box: Reveal?
- Green box: Reveal?
- Image: Hands cleaning a sink
- Red number: 8

A green circle with the number 1 is positioned below the first image, with an arrow pointing to the NN Search diagram below.

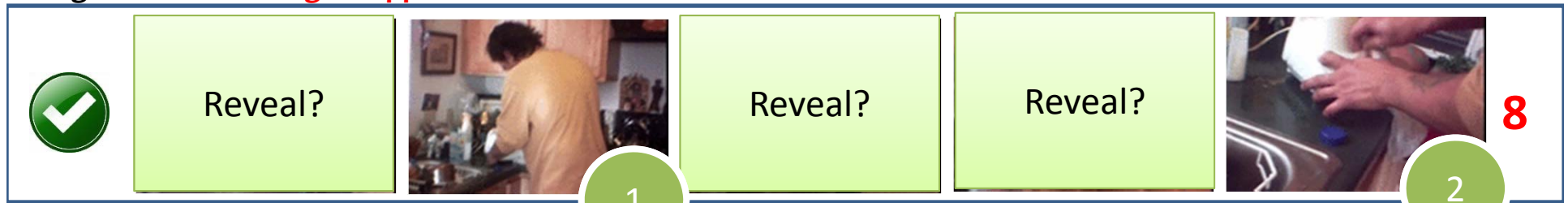
NN Search



Positives for **Cleaning an Appliance** (Likely)

Rewarding Decisive Microshots

Target Event: **Cleaning an Appliance**

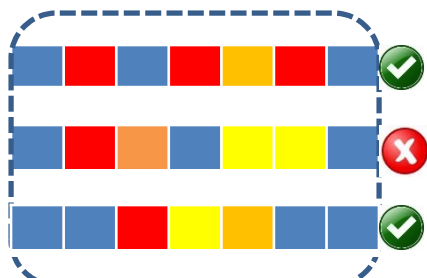


NN Search



Positives for **Cleaning an Appliance** (Likely)

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



Positives for **Cleaning an Appliance** (Likely)

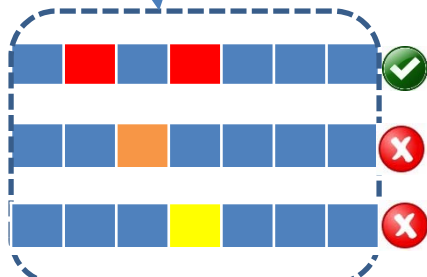
NN Search

Rewarding Decisive Microshots

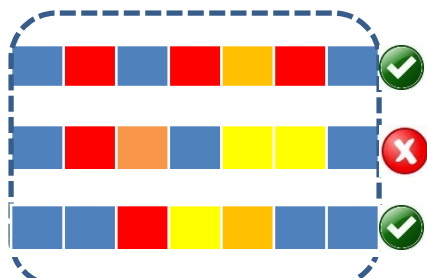
Target Event: **Cleaning an Appliance**



NN Search



Top videos for **Cleaning an Appliance** (Likely)



Top videos for **Cleaning an Appliance** (Likely)

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

Vote

Order of
Revelation

Candidate
microshot

Query
Microshot

NN Search

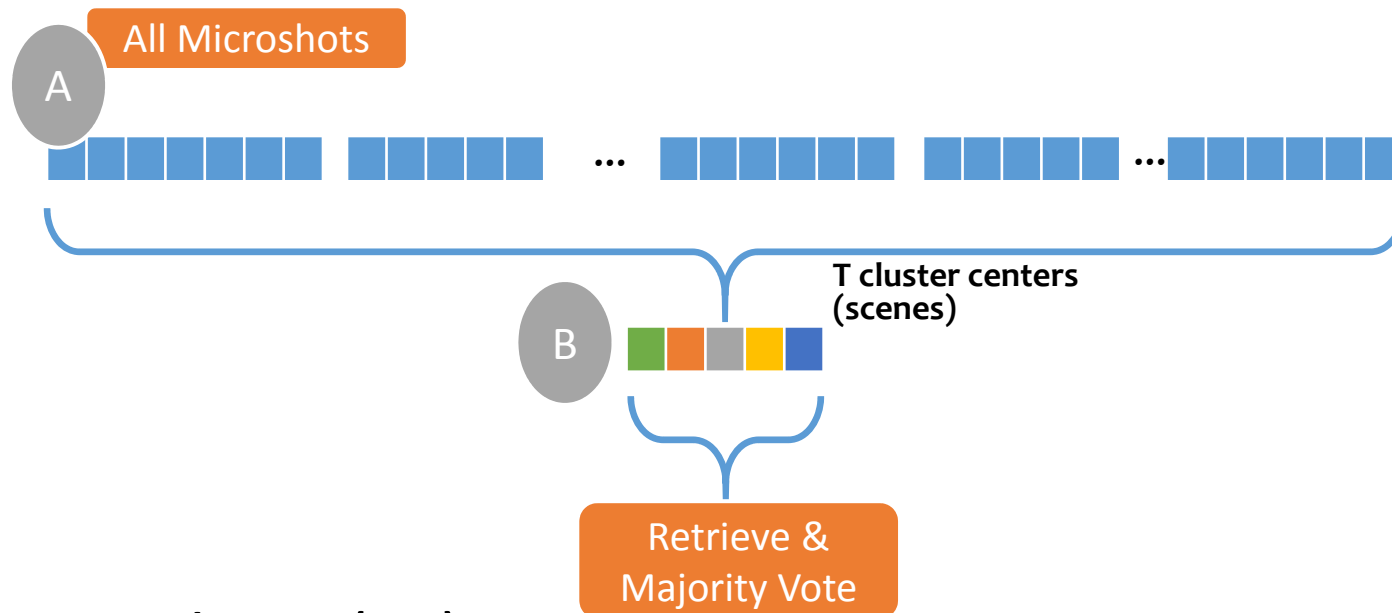
Experiments

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 Directions	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 repair
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

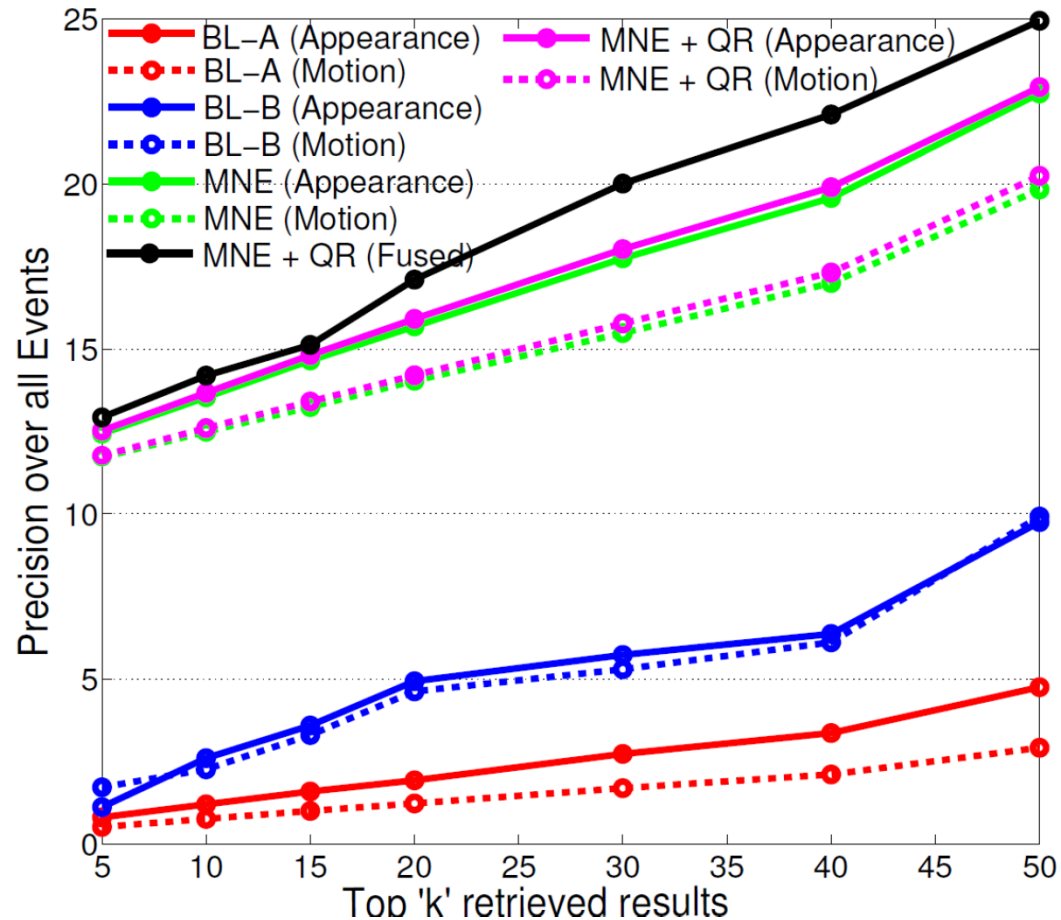
- MED13 ADHOC data set
- Only using MNE, absolute performance gain ~ 14%

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

Additional Results



- Early fusion of motion/appearance with MNE+QR yields best performance

Comparison with Concept Retrieval

Events	[17]	MNE+QR (Fused)
Birthday party	15.9	35.6
Changing vehicle tire	11.8	23.4
Flash mob	30.5	34.3
Vehicle unstuck	15.9	28.2
Grooming animal	21.4	28.5
Making sandwich	13.8	31.6
Parade	22.1	34.8
Parkour	21.9	37.1
Repairing appliance	22.1	32.1
Sewing project	10.1	24.4
Bike trick	11.0	21.3
Cleaning appliance	7.2	19.1
Dog show	13.0	24.5
Giving directions	11.4	22.9
Marriage proposal	2.5	16.3
Renovating home	20.7	29.1
Rock climbing	6.5	18.5
Town hall meeting	5.5	12.3
Winning race w/o vehicle	8.5	18.6
Metal crafts project	3.0	16.2
Mean average precision	13.1	25.4

- ~12% absolute gain over a state-of-the-art algorithm

Qualitative MNEs

The 'Repairing an Appliance' MNE is presented as a vertical sequence of three film strips. The top film strip shows a hand applying a substance to a sink drain, followed by a close-up of the drain, and a red product box labeled 'Kwik Seal'. The middle film strip shows a person working on a green printed circuit board with various electronic components. The bottom film strip shows a person's hands working on the interior of an appliance. Each film strip is connected to a red rectangular box on a blue horizontal bar below it. The caption 'Repairing an Appliance' is centered at the bottom.

Repairing an Appliance

The 'Birthday Party' MNE is presented as a vertical sequence of three film strips. The top film strip shows a group of people gathered around a table with a large green birthday cake. The middle film strip shows a group of children sitting at a table with colorful party supplies. The bottom film strip shows a person sitting at a table with balloons and a large, decorated birthday cake. Each film strip is connected to a red rectangular box on a blue horizontal bar below it. The caption 'Birthday Party' is centered at the bottom.

Birthday Party

Concepts behind the Evidence

- Discovering new concepts – that cannot be mined from textual description of an event

Does the event **Birthday party** exist in these videos?

Game Play Rules and Objectives Possible Evidences for Birthday party

1. Answer (in Yes/No) if you see **Birthday party** in minimal

2. Please list (in comma-separated) the evidences that helped you

When done, click "Start a New Session". Hit F5 to discard

3. Possible Evidences for Birthday party

Cake, Balloons, Table, Chairs, Blowing candles, People Clapping, Kids running, Smiling Faces, Indoor living room, Hugging, Cheering, Shaking hands, Birthday caps, Face painting, Drinking, Eating food, Opening gift box.

1. Describe the evidences in your own words or phrases so that

(a) Another person looking ONLY at these phrases can identify the event without seeing the video.

(b) Your description SHOULD NOT have the same name as the event itself.

2. Feel free to suggest your own evidences that may not be listed above.


YES! Because:
Candles, Cake, Smiling face, People

YES! Because:
Candles, Cupcakes, Party hat, Kids, Table

NO! Because:
Outdoors, Swimming Competition, Lap pool

Video #0 Yes No

Show Microshot



Show Microshot

Show Microshot


Show Microshot

10

I think the video describes "Birthday party" because I SEE: Candles, Cake, Smiling Face, People

Video #1 Yes No

Show Microshot



Show Microshot


Show Microshot

Show Microshot

10

I think the video describes "Birthday party" because I SEE: cupcakes, Candle, Party hat, Kids, Table

Microshot Timestamp



Show Microshot

Show Microshot

10

I think the video DOES NOT describe "Birthday party" because I DONT SEE: Cakes, Candles

I think the video DOES NOT describe "Birthday party" because I SEE: Outdoors, Swimming competition, Lap pool

 Start a New Session

 Submit Labels and Search

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

Desired Concepts: "No, because I don't see ..."

balloon **cake** candle **person**

Concepts parsed from Textual Event Kit

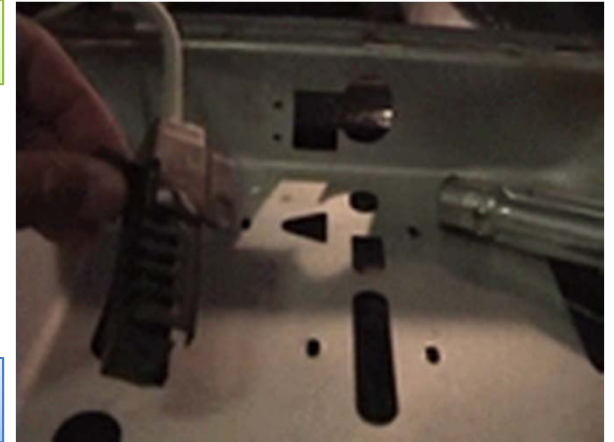
person park streamer anniversary balloon outdoor **birthday**
blowing **cake** candle restaurant celebration children host conical
cupcake party indoor lit shiny/colorful singing food game gift guest home honor



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



Desired Concepts: "No, because I don't see ..."

indoor pointing_to_appliance pointing_to_appliances **tool**
visible_hands

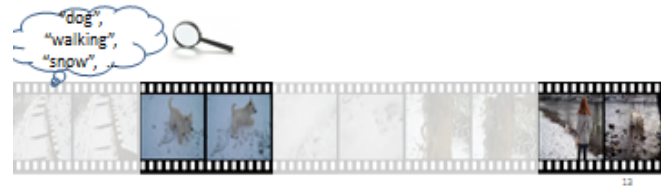


Concepts parsed from Textual Event Kit

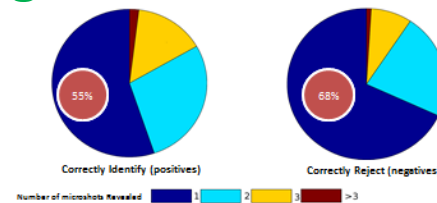
adjusting_machine_part air_conditioners basement black_object clothes_dryers coffee_makers
dishwasher drying_cabinet electric_toothbrushes food_processors freezer
garage hair_dryers hand_mixers indoor induction_cookers kitchen kitchen_stove lifting_machine_parts
machine_parts metallic_object microwave_ovens **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

Summary

- **Leverage human cognitive capability** in recognizing complex events in videos



- **Introduces a novel quiz UI** to find Minimally Needed Evidence
- **Surprising findings of human strategies in event labeling**



- **Discover novel concepts** that cannot be mined directly from textual descriptions of an event



spreading butter,
moving hands