Events

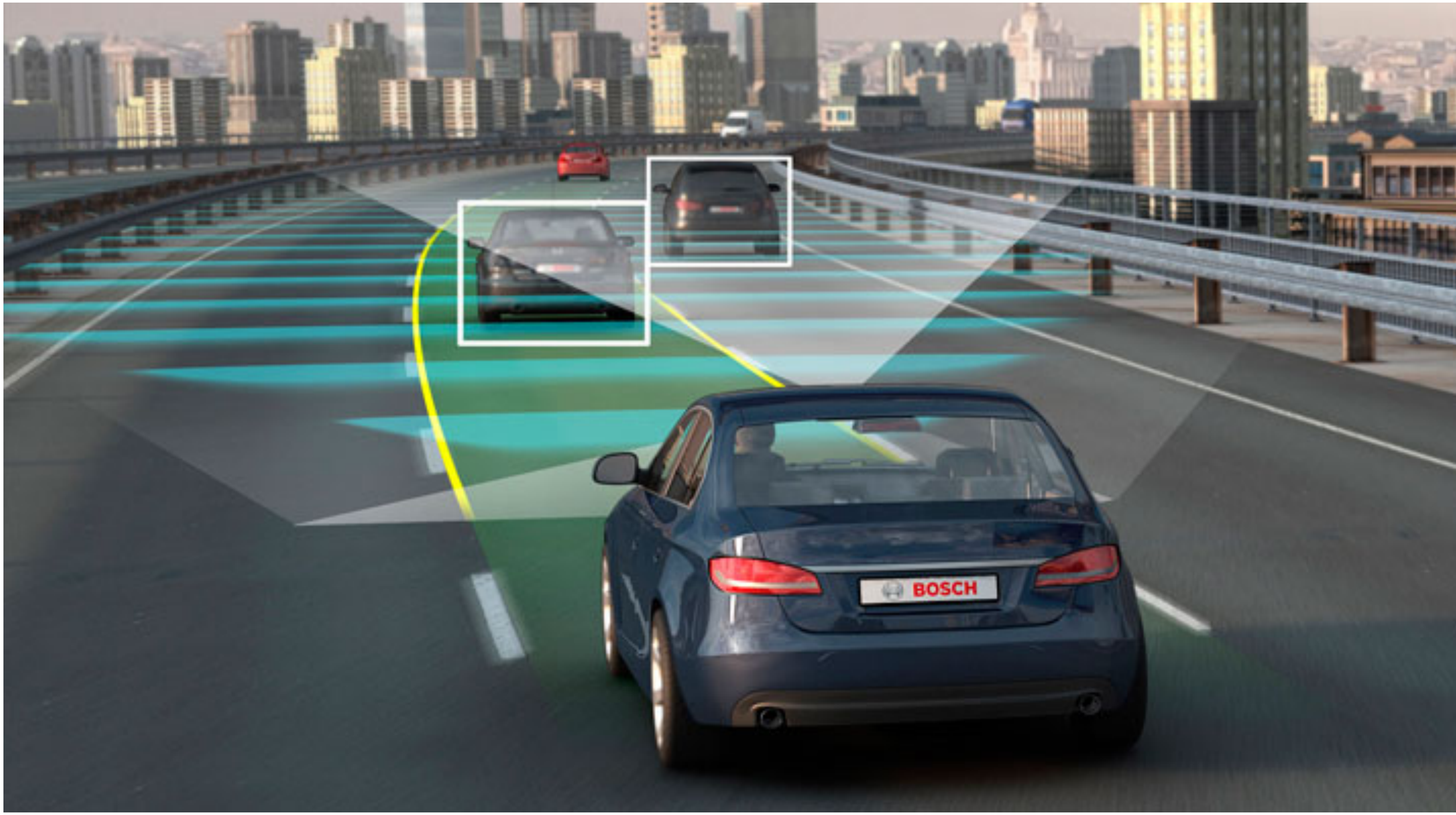
Cees Snoek

University of Amsterdam
The Netherlands

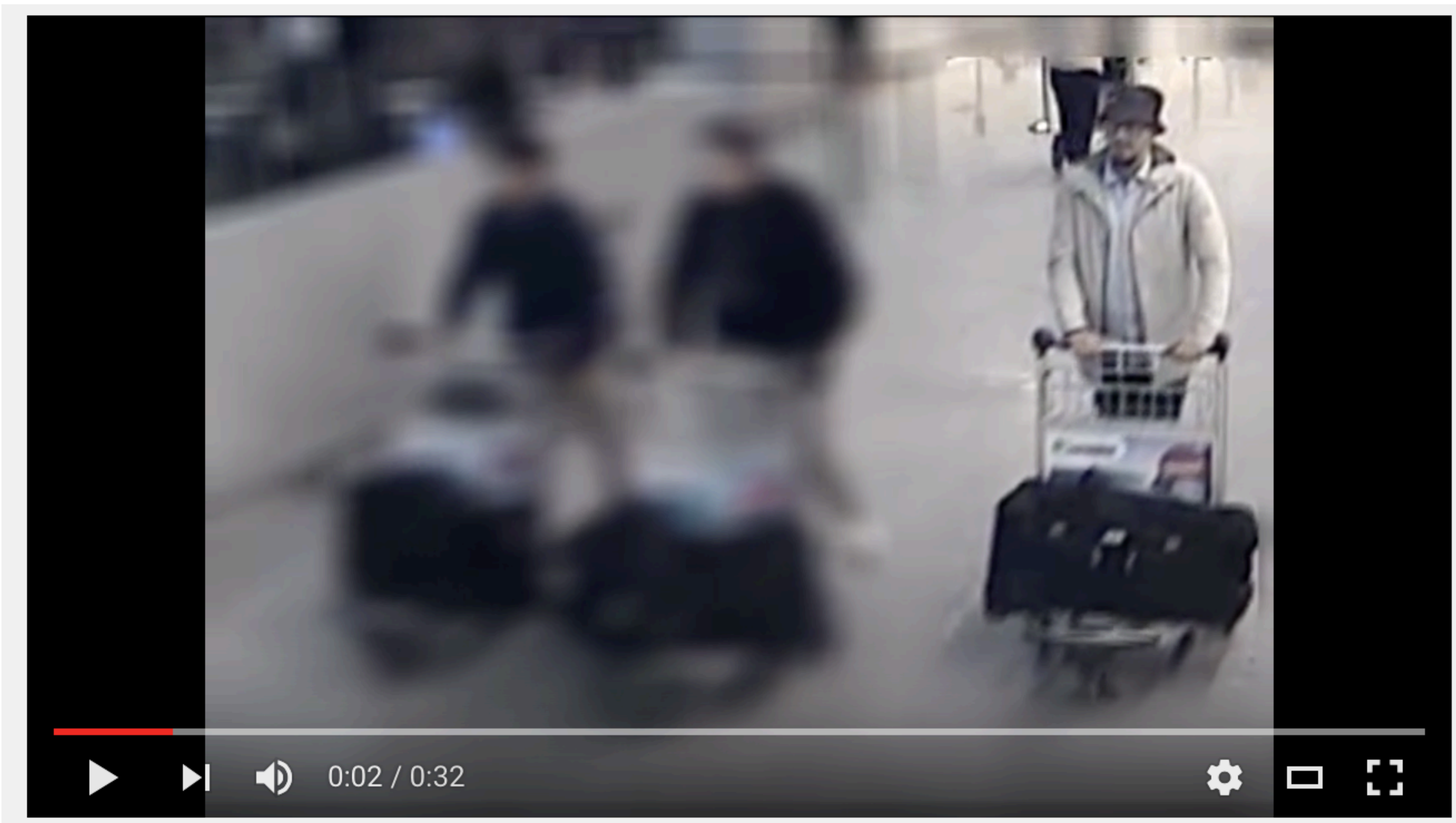
Deep Vision Lab

Motivation: Internet of things that video

Technology: self-driving cars



Forensics: Analyzing terrorist behavior

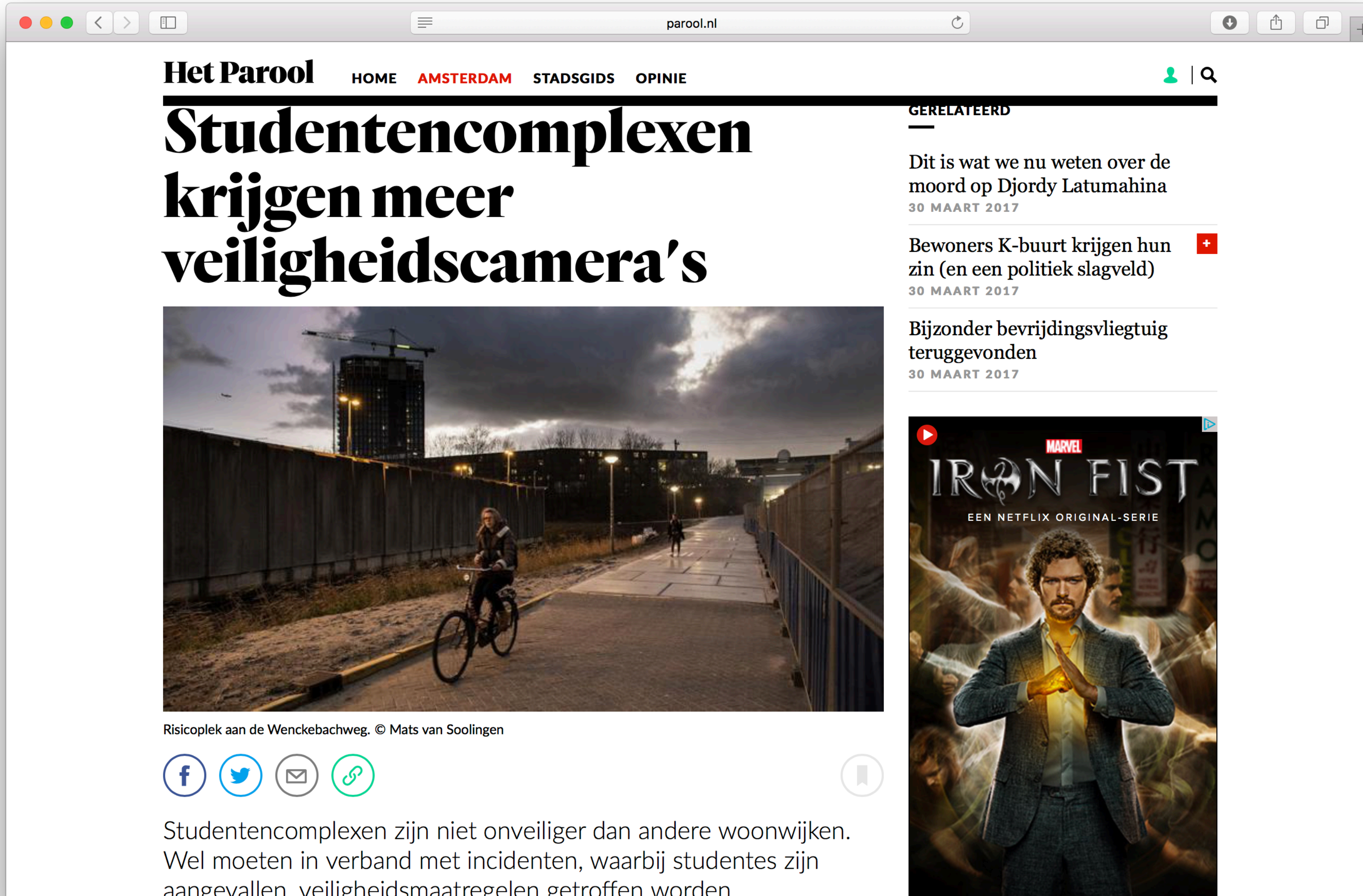


Well-being: elderly monitoring

Figure 1. Examples of interaction patterns in a nursing home

Chen et al. MM 2004

Safety: preventive monitoring



parool.nl

Het Parool HOME AMSTERDAM STADSGIDS OPINIE

Studentencomplexen krijgen meer veiligheidscamera's

Risicoplek aan de Wenckebachweg. © Mats van Soolingen

Studentencomplexen zijn niet onveiliger dan andere woonwijken. Wel moeten in verband met incidenten, waarbij studentes zijn aangevallen, veiligheidsmaatregelen getroffen worden.

GERELATEERD

Dit is wat we nu weten over de moord op Djordy Latumahina
30 MAART 2017

Bewoners K-buurt krijgen hun zin (en een politiek slagveld)
30 MAART 2017

Bijzonder bevrijdingsvliegtuig teruggevonden
30 MAART 2017

What is an event?

News events: *earthquake, abdication, product launch*

Sport events: *scoring goal, ace serve, slam dunk*

Social events: *concert, debates, exhibitions*

Every day events: interactions of people and objects

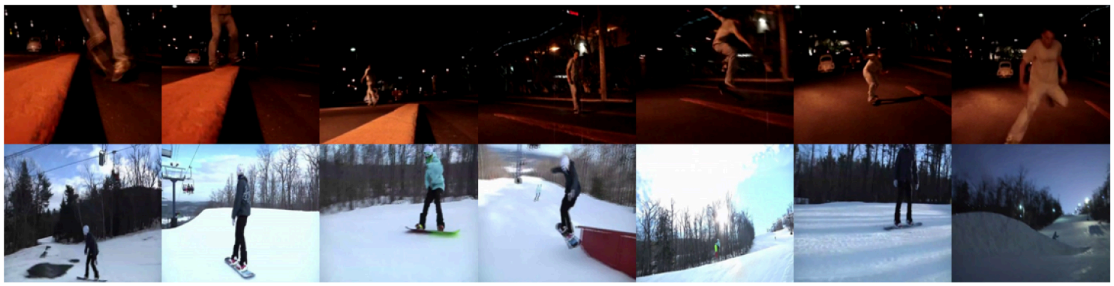
Repairing an appliance

Working on sewing project

Grooming an animal

Birthday party

Goal



Board trick

Assembling a shelter

Birthday party

Recognize and explain event as it happens in video

This lecture

We study event recognition

I. Data, data, data

II. Event classification

III. Event retrieval

Prelude

DATA, DATA, DATA

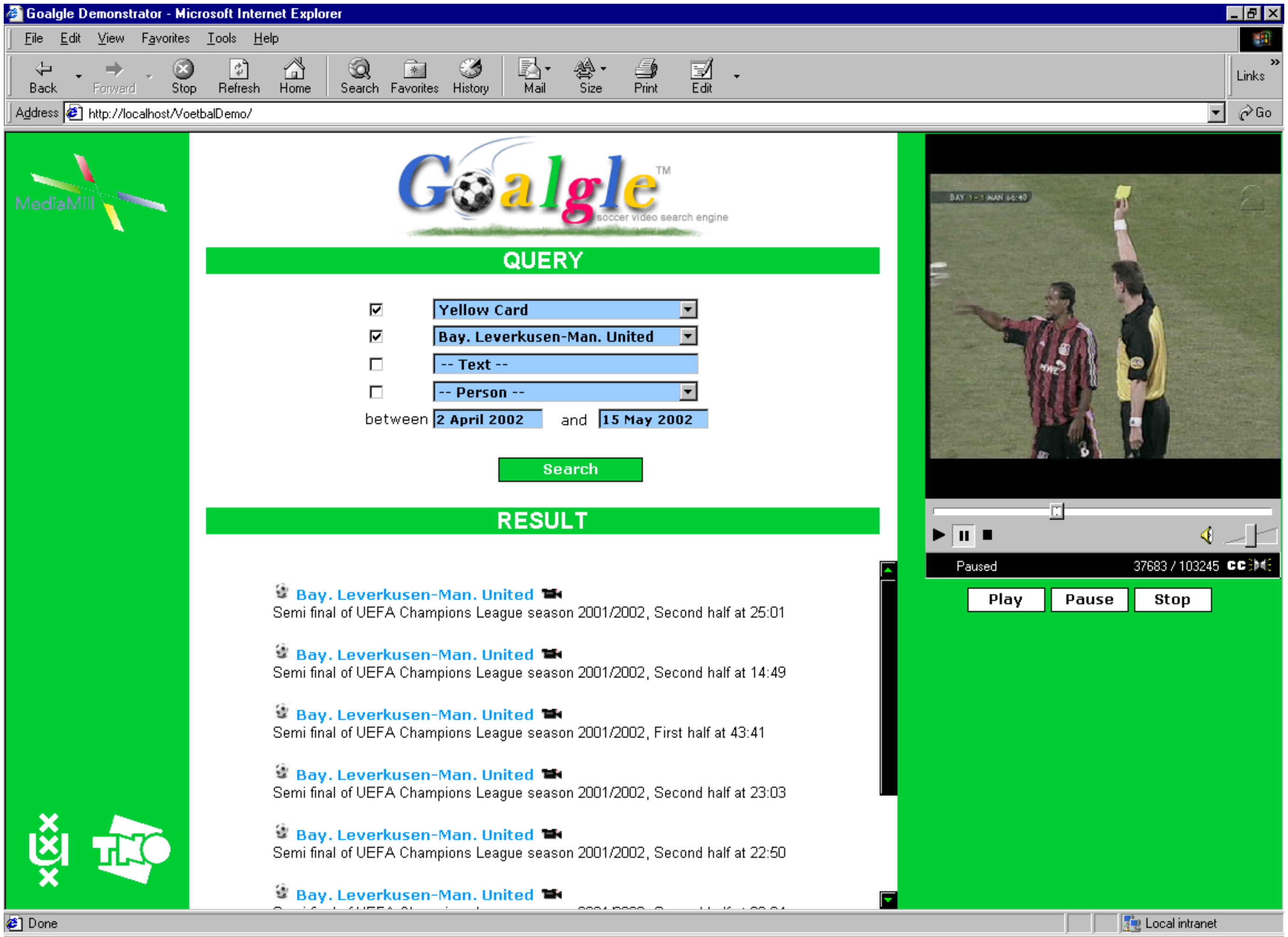
The early years 1995-2010

Progress was slow

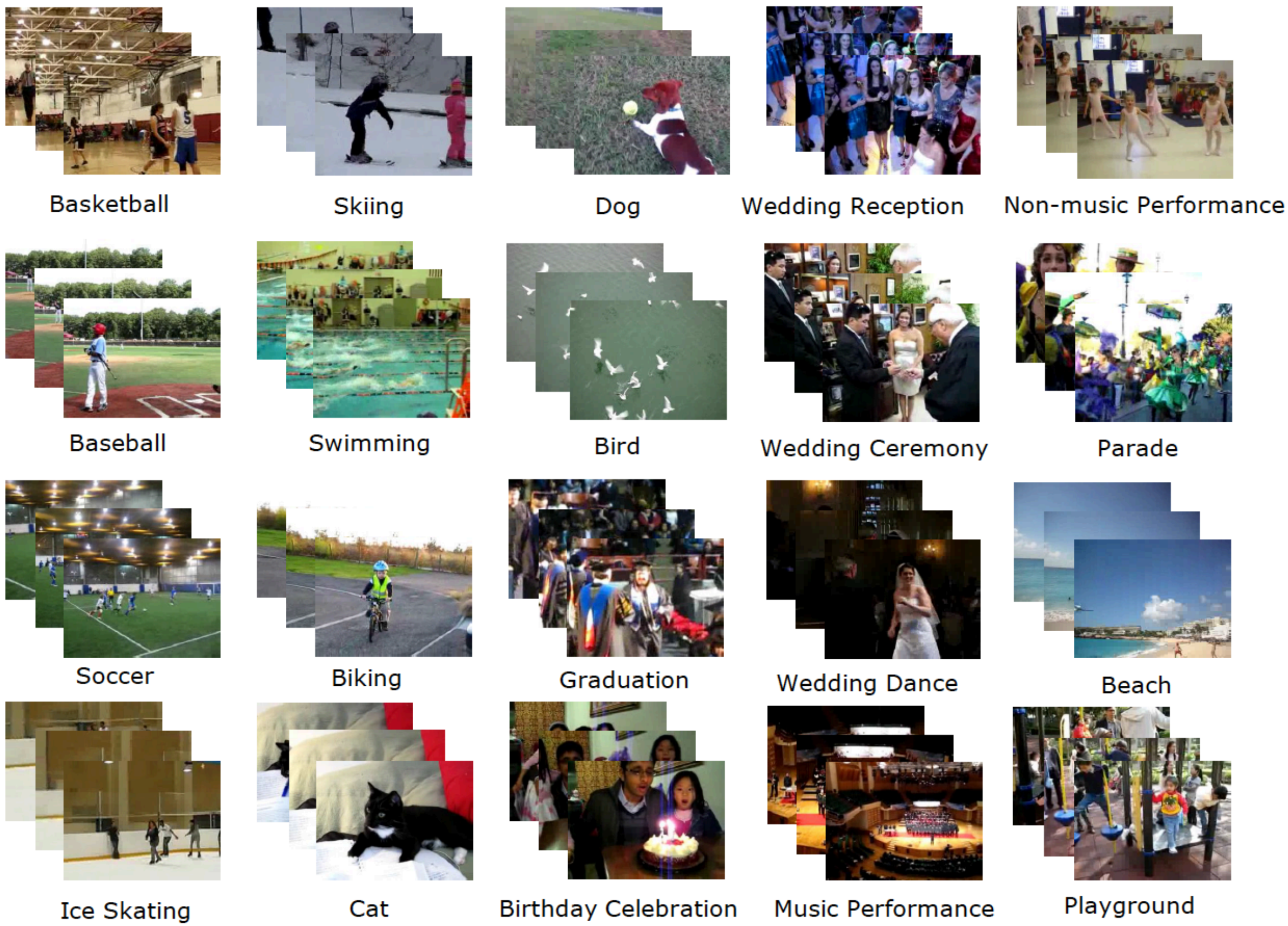
- Lack of data
- Lack of benchmarks
- Lack of community
- Lack of urgency

Snoek et al., ICME 2003

Goalgle: 9 hrs of test video…



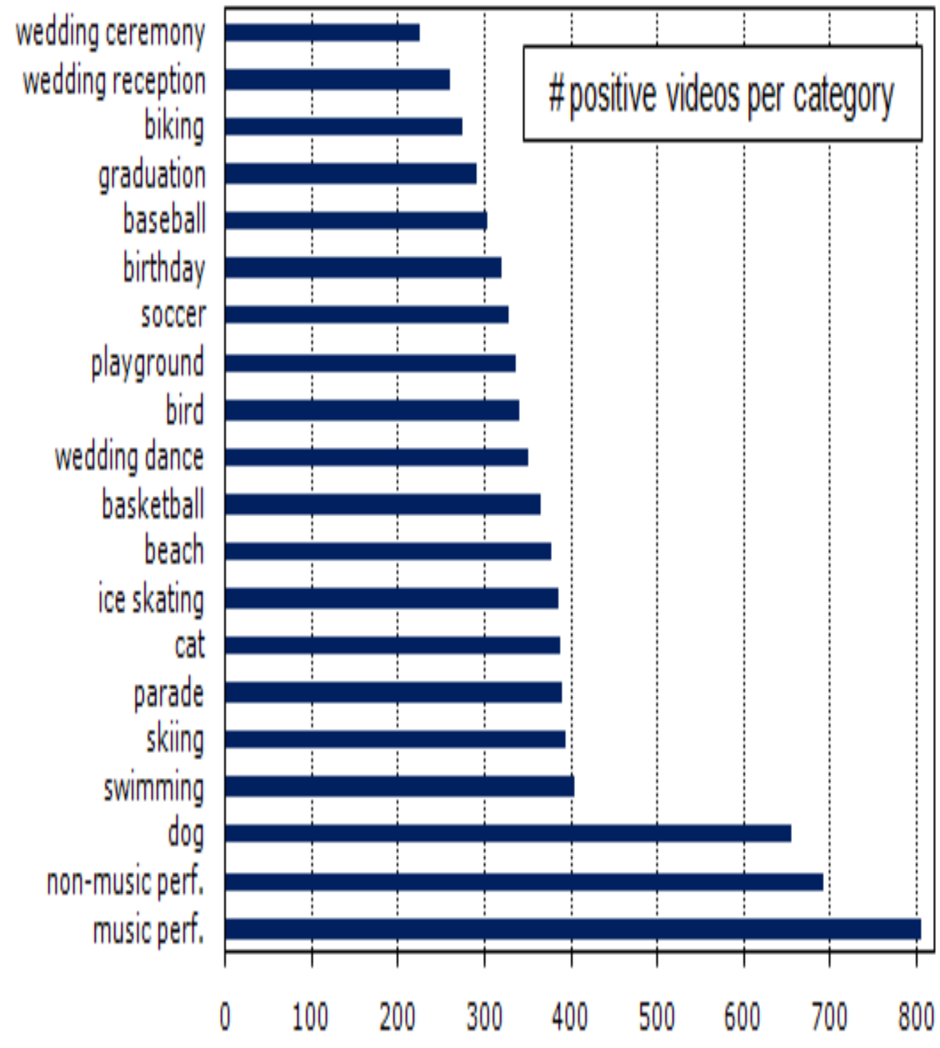
CCV: Columbia Consumer Video Database



http://www.ee.columbia.edu/ln/dvmm/CCV/

CCV snapshot

- # videos: 9,317
  - (210 hrs in total)
- video genre
  - unedited consumer videos
- video source
  - YouTube.com
- average length
  - 80 seconds
- # defined categories
  - 20
- annotation method
  - Amazon Mechanical Turk



http://trecvid.nist.gov/

TRECVID benchmark

International competition

Promote progress in video retrieval research

Open data, tasks, evaluation *and* innovation

Internet video collections

Collection Name	Designated Uses	Target sizes	Annotation
Pilot	<u>2010</u> Development collection Test collection	 1,723 clips 1,742 clips (100 hours)	Clip content annotation for both sets
Development (DEV)	<u>2011</u> Split into two subsets: (1) Transparent (DEV-T) (2) Opaque (DEV-O) <u>2012-2015</u> (1) and (2) merged to a single training collection	44K clips, (~ 1400 hours)	<u>For MED ‘11:</u> Clip content annotation for the transparent subset <u>After MED ‘11:</u> Clip content annotation for the opaque subset
Progress	<u>2012-2015</u>: test collection	120K clips, 4000 hrs	No clip content annotation
Novel 1	<u>2014</u>: test collection	120K clips, 4000 hrs.	No clip content annotation
Novel 2	<u>2015</u>: test collection	120K clips, 4000 hrs.	No clip content annotation

The TRECVID MED '11 events

Training Events

Process-Observed Events

Attempting a board trick
Feeding an animal
Landing a fish
Working on a woodworking project

Life Events

Wedding ceremony

Testing Events

Process-Observed Events

Changing a vehicle tire
Getting a vehicle unstuck
Grooming an animal
Making a sandwich
Parkour
Repairing an appliance
Working on a sewing project

Life Events

Birthday party
Flash mob gathering
Parade

Example Event Kit

Event Name:

Working on a woodworking project

Definition:

One or more people fashion an object out of wood.

Event Explication:

Woodworking is a popular hobby that involves crafting an object out of wood. Typical woodworking projects may range from creating large pieces of furniture to small decorative items or toys. The process for making objects out of wood can include cutting wood into smaller pieces .... (continues)

Evidential Description:

scene: Often indoors in a workshop, garage, artificial lighting. Occasionally outdoors

objects/people: Woodworking tools (automatic or non-automatic saws, sander, knife), paint, stains, sawhorses, toolbox, safety goggles

activities: Cutting and shaping wood, attaching pieces of wood together, smoothing/sanding wood

audio: power tool sounds; hand tool sounds (hammer, saw, etc.); narration of process

Exemplars:

HVC334271.mp4, HVC393428.mp4, HVC875424.mp4, etc.

Mnemonic

Textual Definition

Expresses event domain specific knowledge to understand the event definition

Textual listing of attributes that are often associated with the event

Specific clips from the "Event Kits" data set that are known to contain the event being defined.

Target User:

An ***Internet information analyst*** or ***experienced Internet searcher*** with event-specialized knowledge.

Part I

CLASSIFICATION

Chapter 1

FEATURE ENCODING

Several slides by: Yu-Gang Jiang

Solution 1: Feature encoding

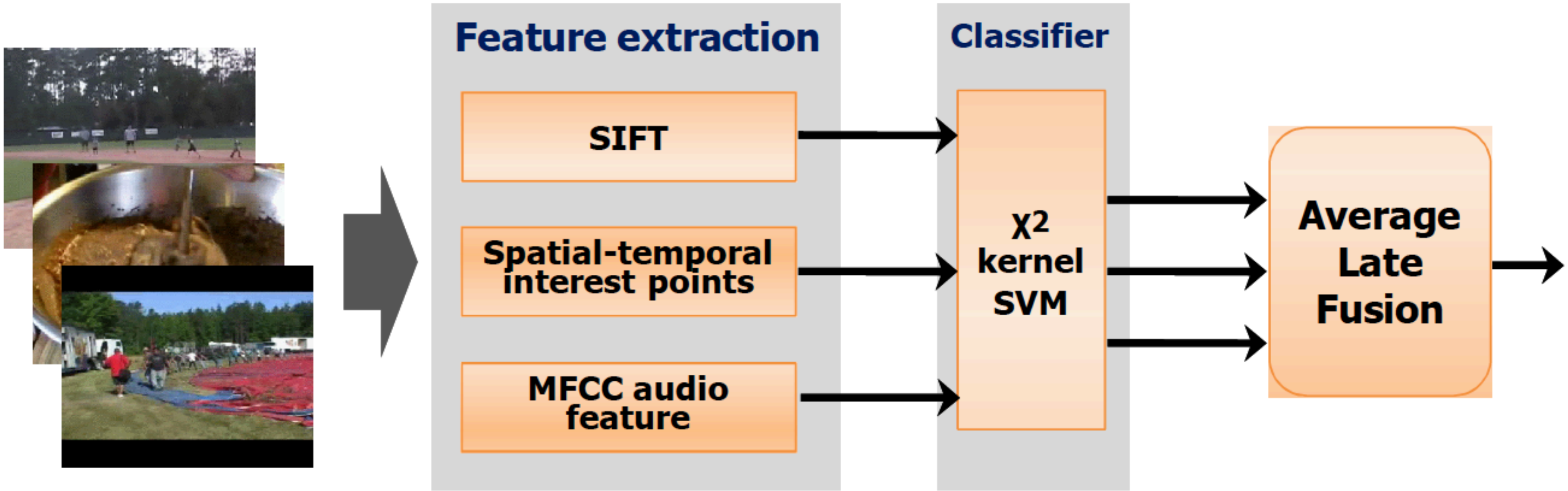
Represent video as low-level feature vector

- Image features: SIFT variations, deep learning, *etc.*
- Audio features: MFCC, AUD, *etc.*
- Text features: ASR, OCR, *etc.*
- Motion features: STIP, dense trajectories, *etc.*

Good recognition accuracy, no interpretation

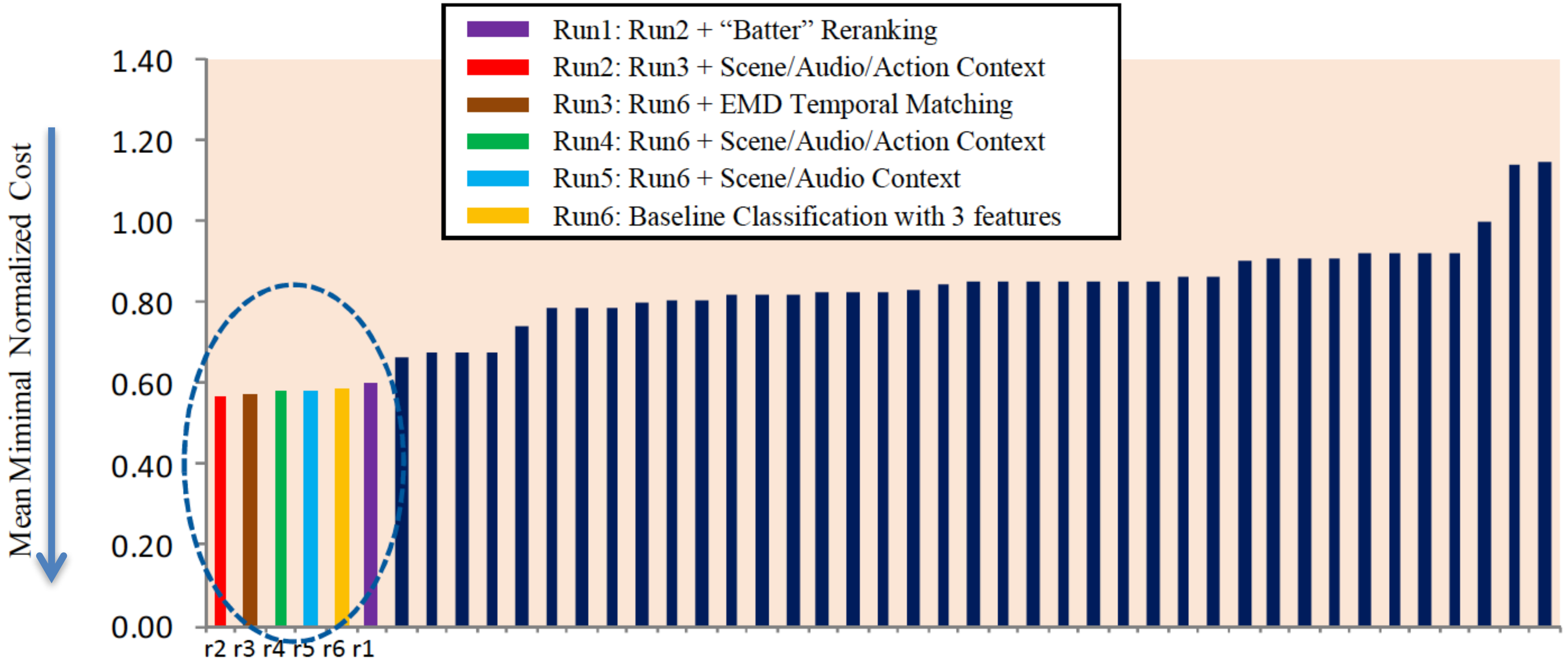
Y.G. Jiang et al. TRECVID10
P. Natarjan et al., CVPR12
Wang et al., ICCV13
and many others

Winner TRECVID 2010



Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Subh Bhattacharya, Dan Ellis, Mubarak Shah, Shih-Fu Chang, **Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching**, NIST TRECVID Workshop, 2010.

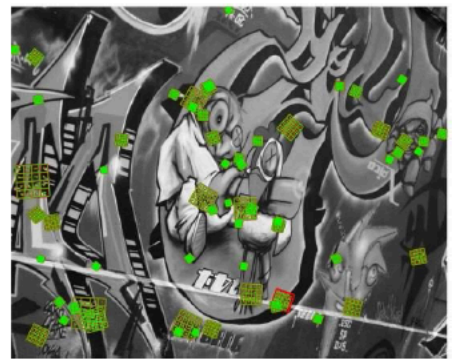
Contribution per modality



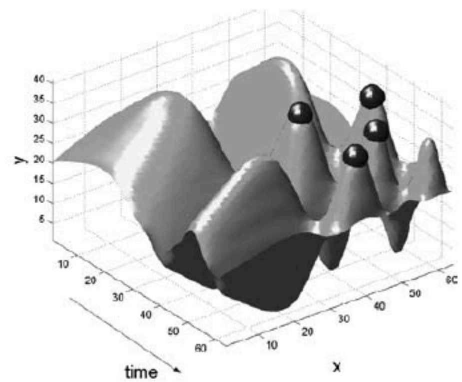
More is better, feature fusion strong fundament

Audiovisual features

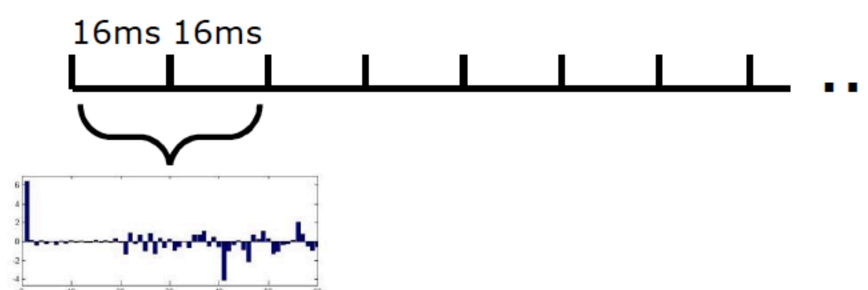
- *SIFT (visual)*
  - *D. Lowe, IJCV 04.*



- *STIP (visual)*
  - *I. Laptev, IJCV 05.*

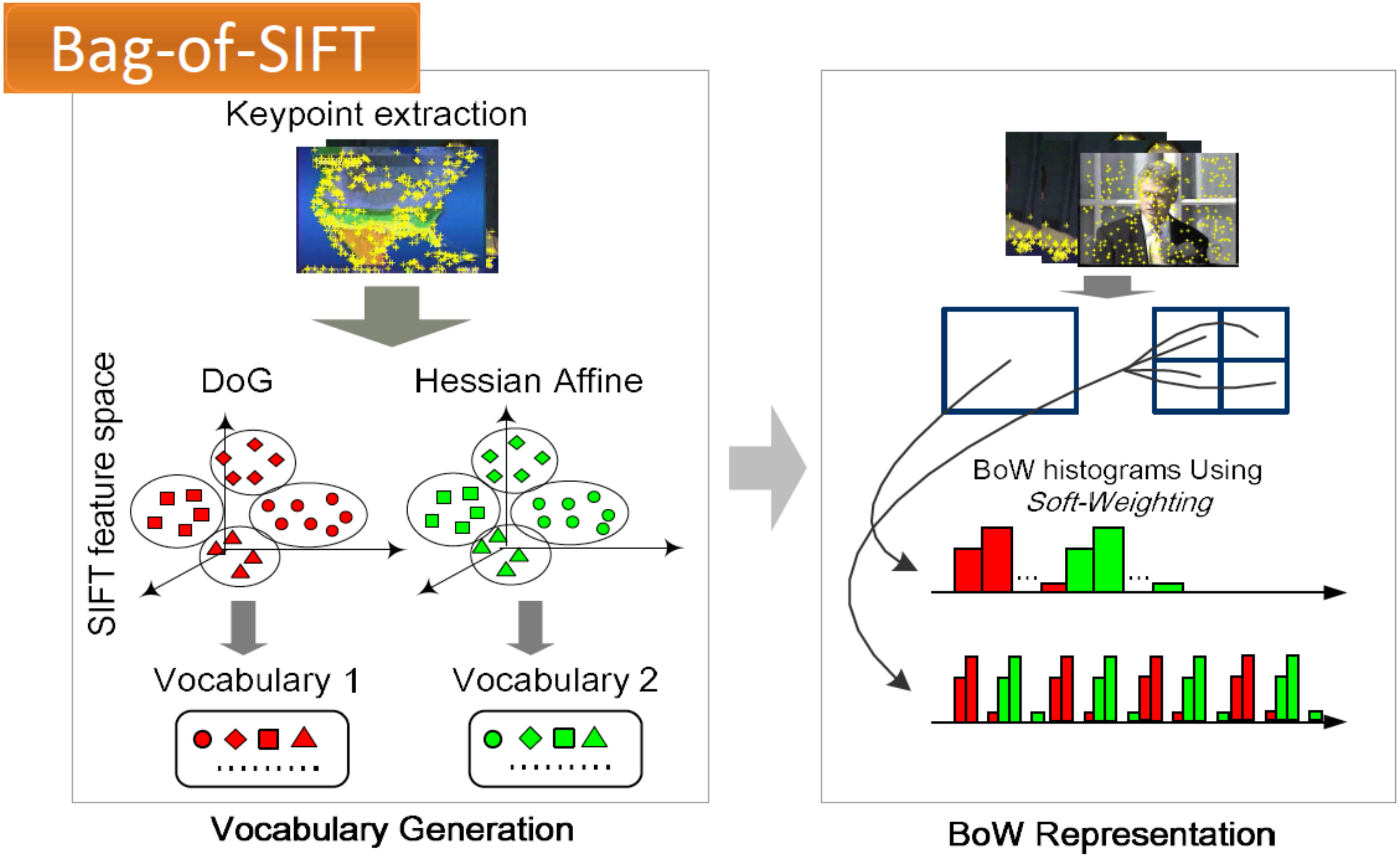


- *MFCC (audio)*



Bag-of-X representation

- X = SIFT / STIP / MFCC
- Soft weighting (Jiang, Ngo and Yang, ACM CIVR 2007)



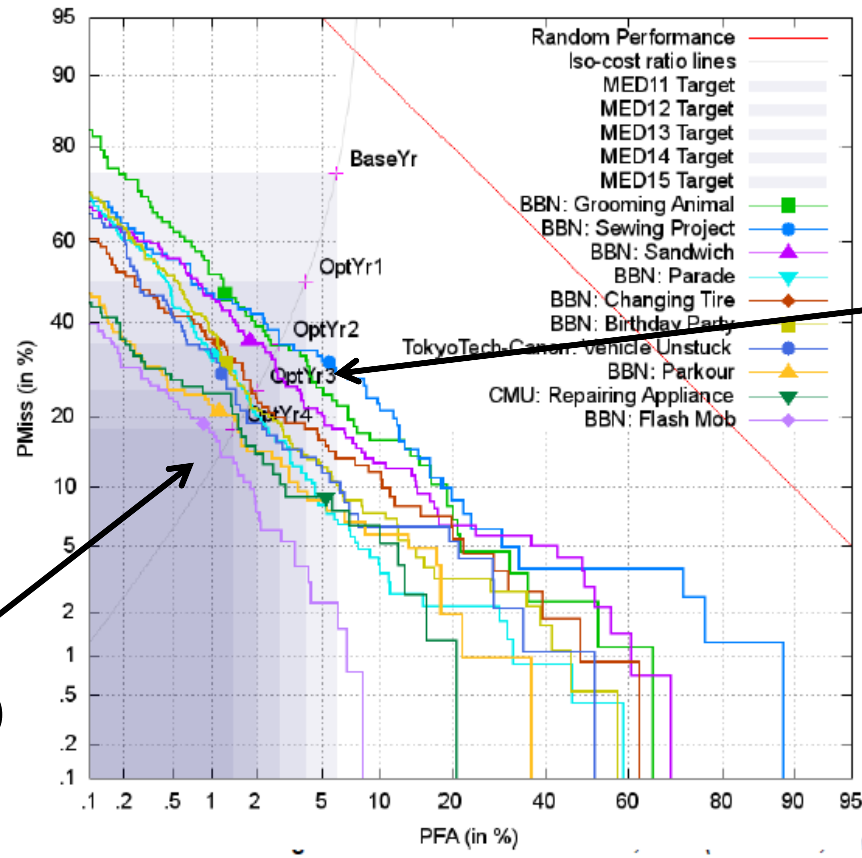
Results

- Measured by Average Precision (AP)

	Assembling a shelter	**Batting a run in**	**Making a cake**	***Mean AP***
Visual STIP	0.468	0.719	0.476	0.554
Visual SIFT	0.353	0.787	0.396	0.512
Audio MFCC	0.249	0.692	0.270	0.404
STIP+SIFT	0.508	0.796	0.476	0.593
STIP+SIFT+MFCC	**0.533**	**0.873**	**0.493**	**0.633**

- STIP works the best for event detection
- The 3 features are highly complementary!

2011 event detection results

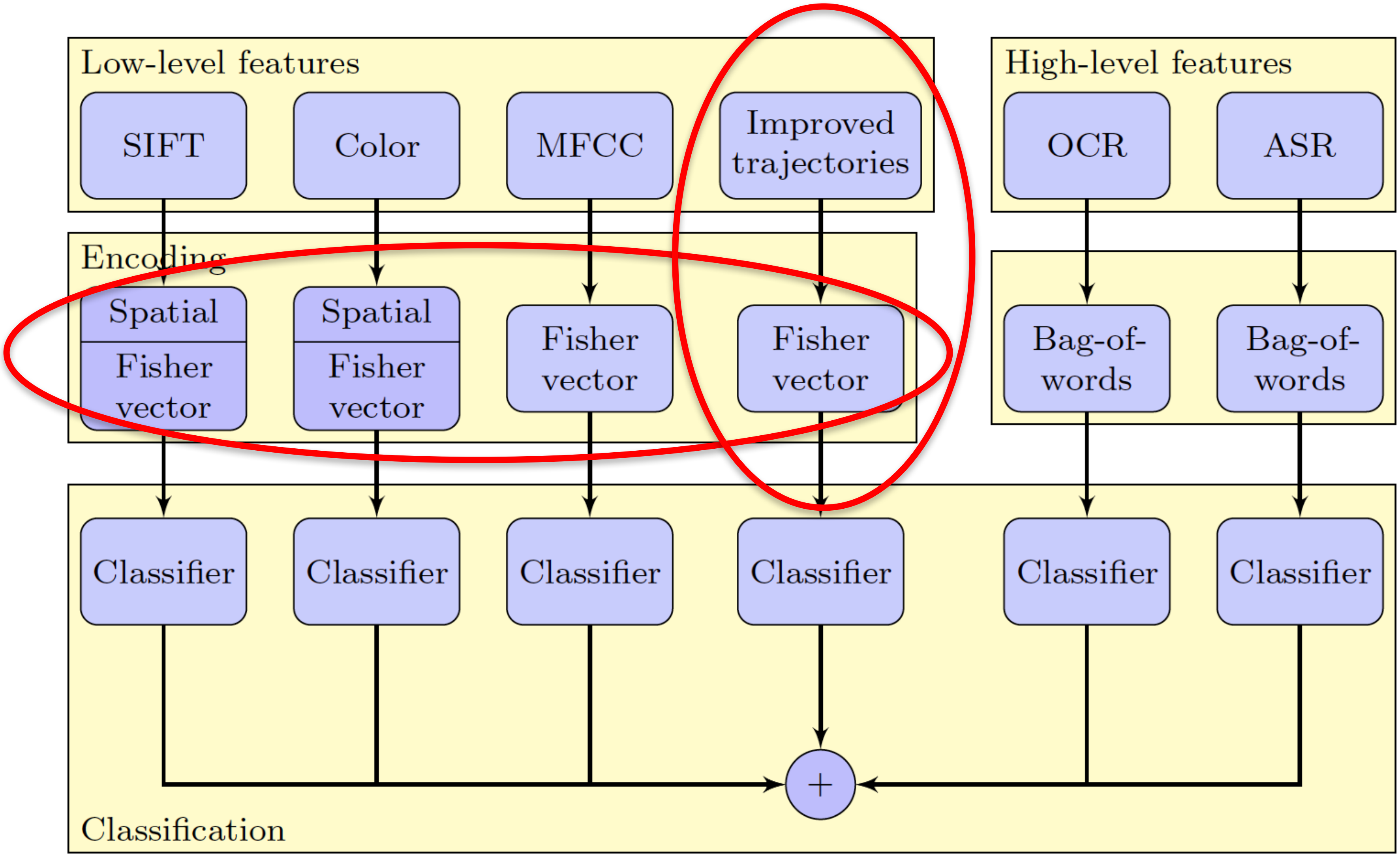


Hard: Grooming an animal

Easy: Flash mob

All systems rely predominantly on bag-of-features, no notion whether event really happened

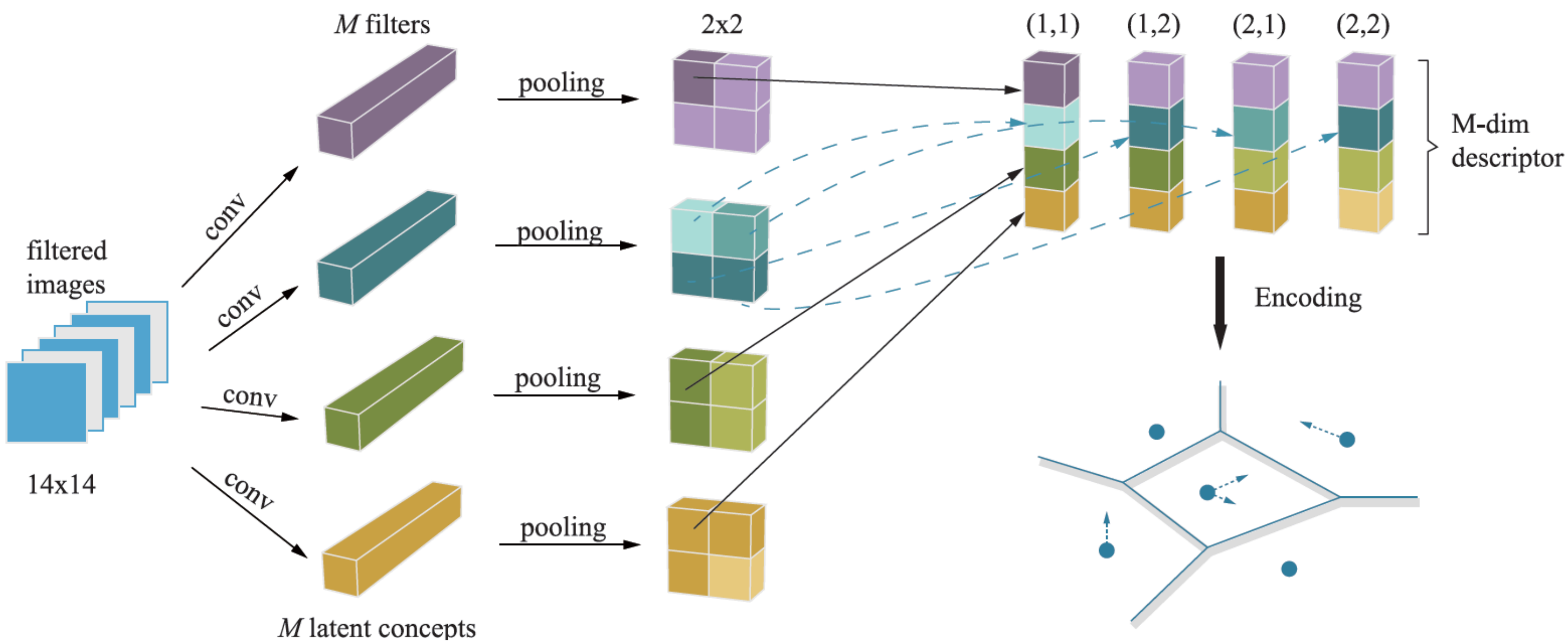
2012 & 2013 winner: Inria LEAR



Xu et al, CVPR 2015

2014 winner: CMU

Winning system combined many multimedia features, with huge computation budget, deep learning key?

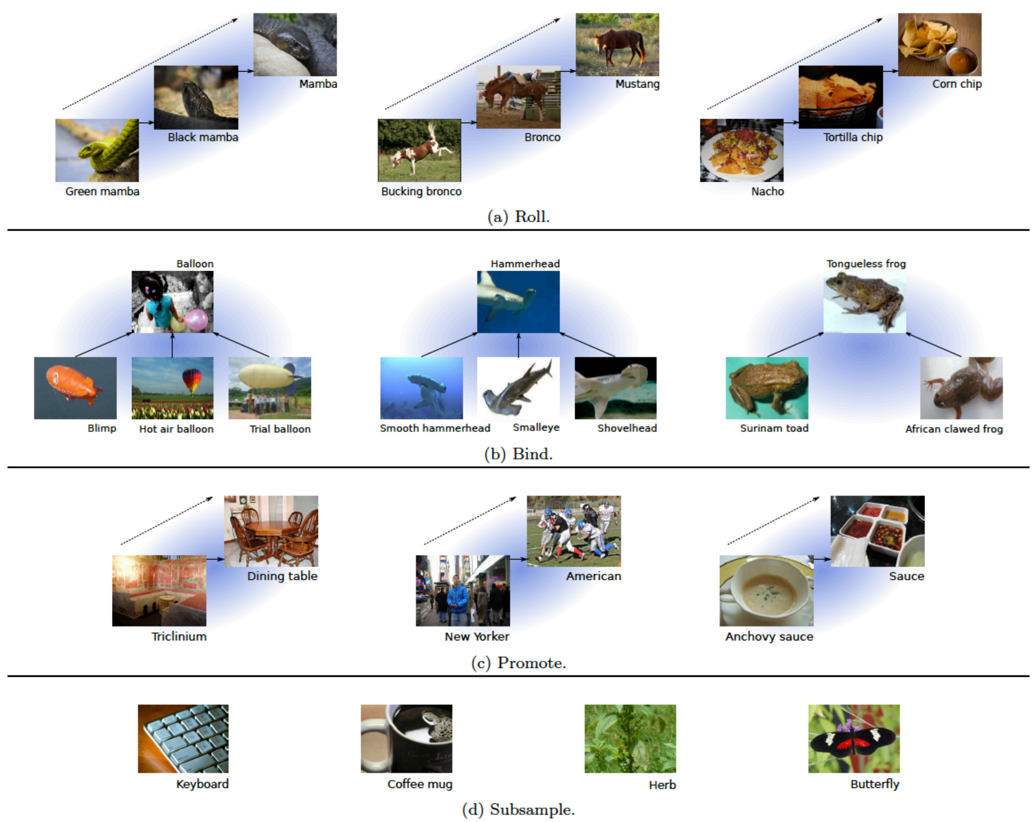


Mettes et al, ICMR 2016

2015 winner: ImageNet-Shuffle - UvA

Leverage complete, but reorganized ImageNet for pre-training

Outperform standard networks, maintain benefits of fusion



http://tinyurl.com/imagenetshuffle

Conclusion on feature encodings

- The combination of audio-visual features is key for good video event recognition
  - ~~MBH + Fisher vector best single feature~~
  - Best single feature from deep convolutional nets
- Many start to explore temporal deep learning
  - 3D convolutions
  - Recurrent neural networks
  - …

Good recognition accuracy, limited interpretation

Chapter 2

SEMANTIC ENCODING

Joint work with Amirhossein Habibian & Masoud Mazloom

Solution 2: Semantic encoding

Represent video as concept score histogram

- Detectors from deep learning, Fisher vectors, *etc.*
- Annotated examples from ImageNet, Flickr, *etc.*

Vocabulary for semantic encoding mostly driven by ad hoc concept detector availability.

Naphade et al. TMM02
Ebadollahi et al., ICME06
Snoek et al., PAMI06
Gkalelis et al., CBMI11
Merler et al., TMM12
and many others

Habibian et al., CVIU 2014
Mazloom et al. TMM 2014

Semantic encodings for video

1. How many concepts?
2. What concept types?
3. Which concepts?
4. How accurate?
5. How to select?

Experimental setup

MED: TRECVID Multimedia Event Detection 2012

13,274 videos (66% train, 34% test)

25 event categories, *marriage proposal, grooming animal, etc.*

CCV: Columbia Consumer Video

9,317 videos (50% train, 50% test)

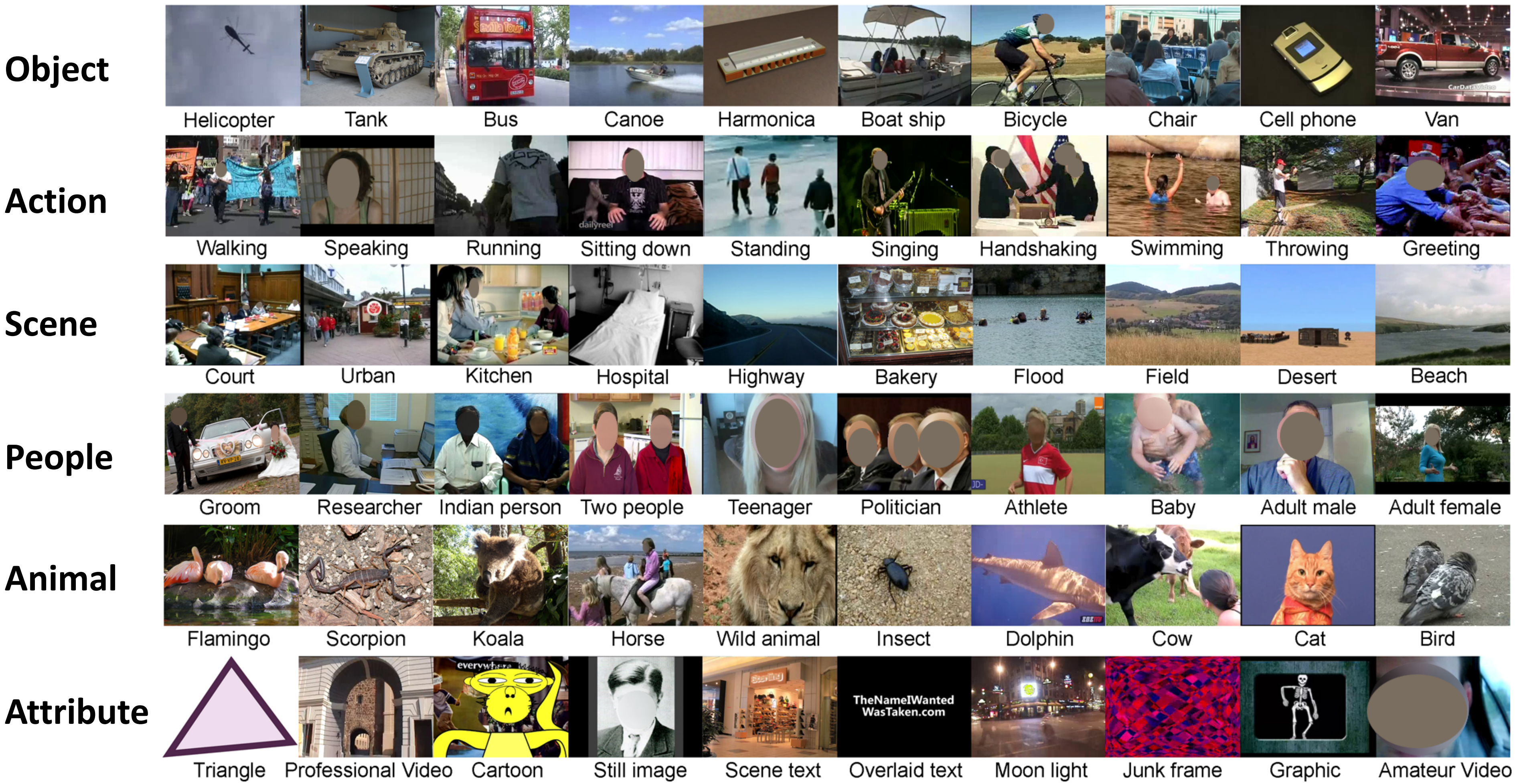
15 event categories, *music performance, graduation, etc.*

Vocabulary sampled from 1,346 concept detectors

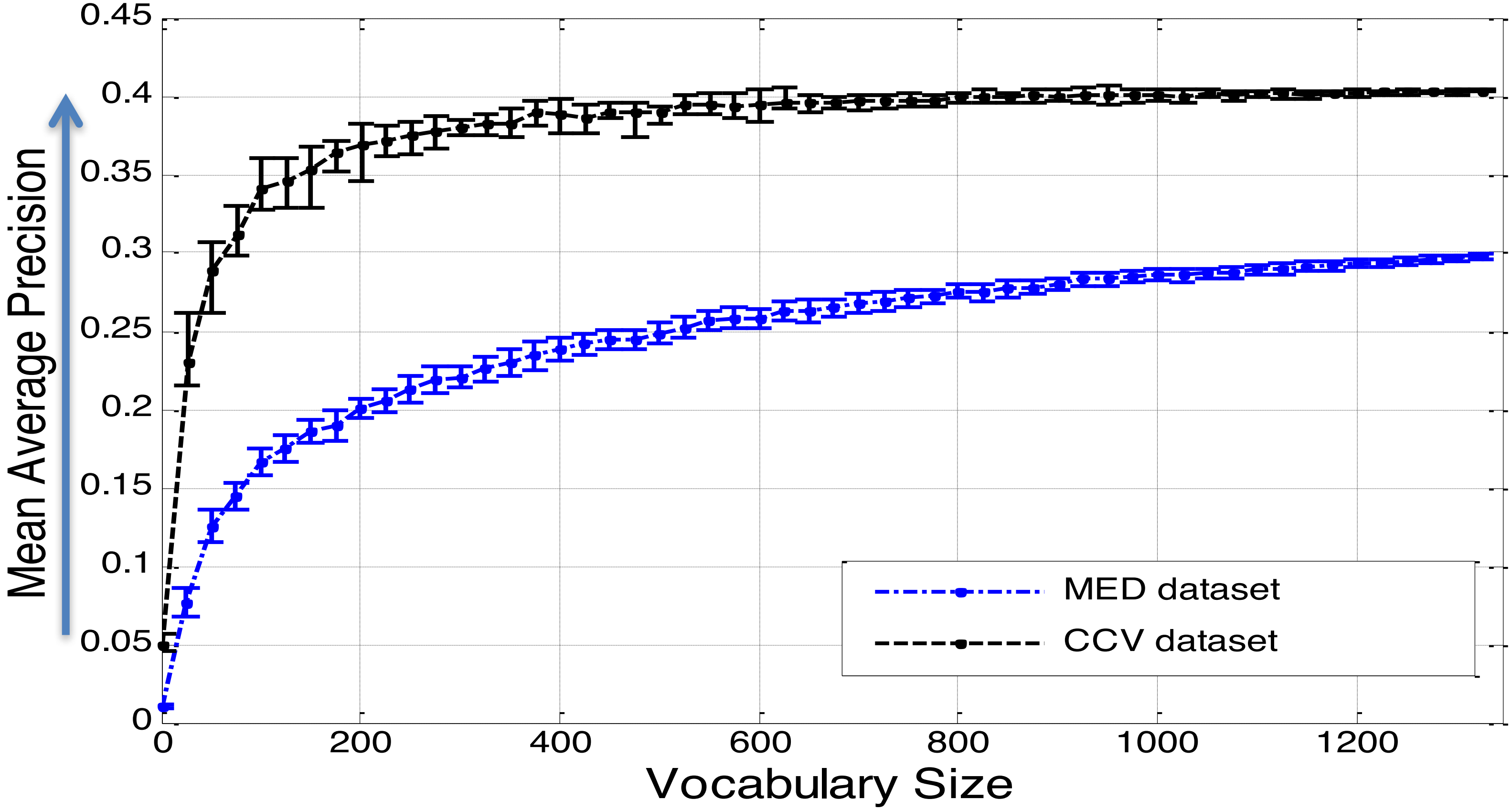
Annotations by ImageNet Challenge11 and TRECVID SIN12

Color Fisher coding with spatial pooling and linear SVM

Concepts categorized by type



1. How many concepts?



More is better, but include at least 200

2. What concept types?

Derive the vocabulary concepts

- ***Single***: Only from a particular concept type?
- ***Joint***: From various concept types?

Scene *(128)*	
`Single`	`Joint`
0.142	**0.168**

2. What concept types?

MED	**Object *(670)***		**Action *(34)***		**Scene *(128)***		**People *(78)***		**Animal *(321)***		**Attribute *(45)***	
Vocab.	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint
MAP	0.259	**0.279**	0.067	**0.076**	0.142	**0.168**	0.082	**0.123**	0.158	**0.239**	0.063	**0.082**

Small difference

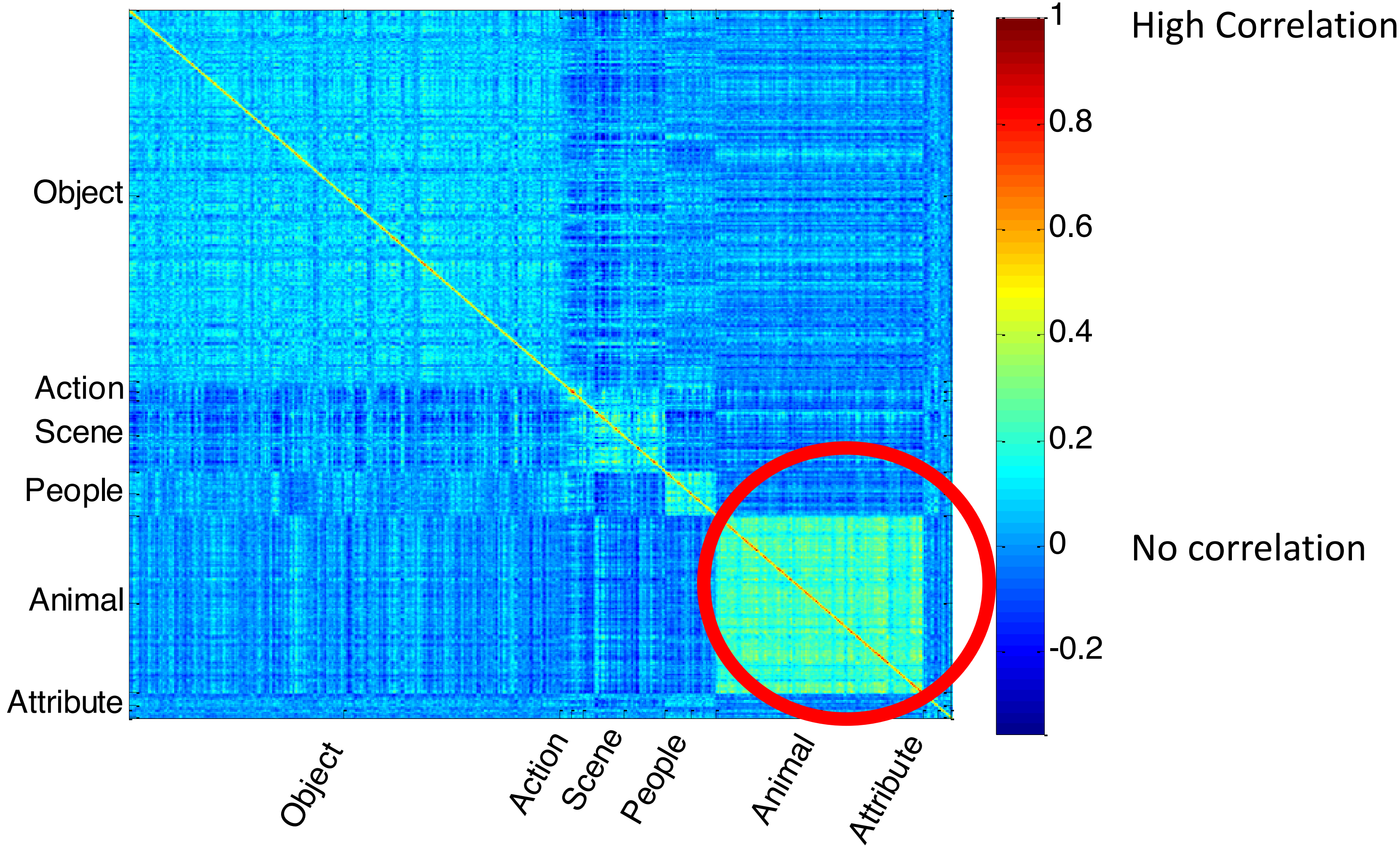
Big difference

CCV	**Object *(670)***		**Action *(34)***		**Scene *(128)***		**People *(78)***		**Animal *(321)***		**Attribute *(45)***	
Vocab.	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint
MAP	0.307	**0.335**	0.197	**0.217**	0.249	**0.285**	0.229	**0.265**	0.265	**0.310**	0.178	**0.220**

In general, a diverse vocabulary is better

Event	Animal *(321)*	
	Single	Joint
Attempting board trick	0.120	**0.271**
Feeding animal	**0.073**	0.045
Landing fish	0.323	**0.36**
Wedding ceremony	0.162	**0.388**
Working wood working project	0.116	**0.167**
Birthday party	0.139	**0.239**
Changing vehicle tire	0.054	**0.153**
Flash mob gathering	0.415	**0.475**
Getting vehicle unstuck	0.294	**0.338**
Grooming animal	**0.146**	0.127
Making sandwich	0.07	**0.176**
Parade	0.126	**0.275**
Parkour	0.089	**0.356**
Repairing appliance	0.104	**0.259**
Working sewing project	0.194	**0.238**
Attempting bike trick	0.129	**0.392**
Cleaning appliance	0.029	**0.058**
Dog show	**0.555**	0.512
Giving directions location	0.016	**0.029**
Marriage proposal	0.018	**0.05**
Renovating home	0.085	**0.192**
Rock climbing	0.309	**0.322**
Town hall meeting	0.266	**0.379**
Winning race without vehicle	0.088	**0.138**
Working metal crafts project	0.019	**0.038**

Concept correlations



Plotted for MED dataset

Semantic encodings for video

1. How many concepts?
2. What concept types?
3. **Which concepts?**
4. How accurate?
5. How to select?

3. Which concepts?

General/specific concepts are identified manually

- General: human, vegetation, outdoor etc.
- Specific: salmon, cheese, sand castle etc.

Derive the vocabulary concepts

- Only from specific concepts?
- Only from general concepts?
- Mixture of specific and general concepts?

3. Which concepts?

MED dataset

Vocabulary	Specific	General	Mixture
MAP	0.094	0.117	**0.130**

CCV dataset

Vocabulary	Specific	General	Mixture
MAP	0.208	0.232	**0.260**

Specific and general concepts should be mixed

Event	Specific	General	Mixture
Attempting board trick	0.090	0.108	**0.130**
Feeding animal	0.041	0.042	**0.045**
Landing fish	0.113	0.107	**0.139**
Wedding ceremony	0.071	0.14	**0.164**
Working wood working project	**0.083**	0.065	0.073
Birthday party	0.078	0.135	**0.138**
Changing vehicle tire	0.058	0.062	**0.071**
Flash mob gathering	0.301	0.284	**0.337**
Getting vehicle unstuck	0.195	0.246	**0.282**
Grooming animal	0.064	0.079	**0.081**
Making sandwich	0.059	0.089	**0.119**
Parade	0.073	**0.203**	0.161
Parkour	0.104	**0.226**	0.210
Repairing appliance	**0.111**	0.098	0.101
Working sewing project	0.076	0.075	**0.082**
Attempting bike trick	0.044	0.08	**0.09**
Cleaning appliance	**0.125**	0.092	0.123
Dog show	0.219	0.178	**0.23**
Giving directions location	0.028	0.019	**0.053**
Marriage proposal	0.013	0.017	**0.025**
Renovating home	0.023	0.074	**0.083**
Rock climbing	0.178	0.156	**0.194**
Town hall meeting	0.064	**0.226**	0.158
Winning race without vehicle	0.102	0.102	**0.117**
Working metal crafts project	**0.040**	0.021	**0.036**

4. How accurate?

How important is the concept detector accuracy?

Decrease concept detector accuracies to observe how event detection performance responds

Approach: Train less sophisticated detectors

Approach: Four detector settings

All examples / ColorSIFT / Spatial Pyramids

30% of examples / ColorSIFT / Spatial Pyramids

30% of examples / SIFT / Spatial Pyramids

30% of examples / SIFT

Habibian et al., submitted

Train less sophisticated detectors

MED dataset

Detectors	100% Examples ColorSIFT Spatial Pyramid	30% Examples ColorSIFT Spatial Pyramid	30% Examples SIFT Spatial Pyramid	30% Examples SIFT
MAP	**0.206**	0.189	0.182	0.185

CCV dataset

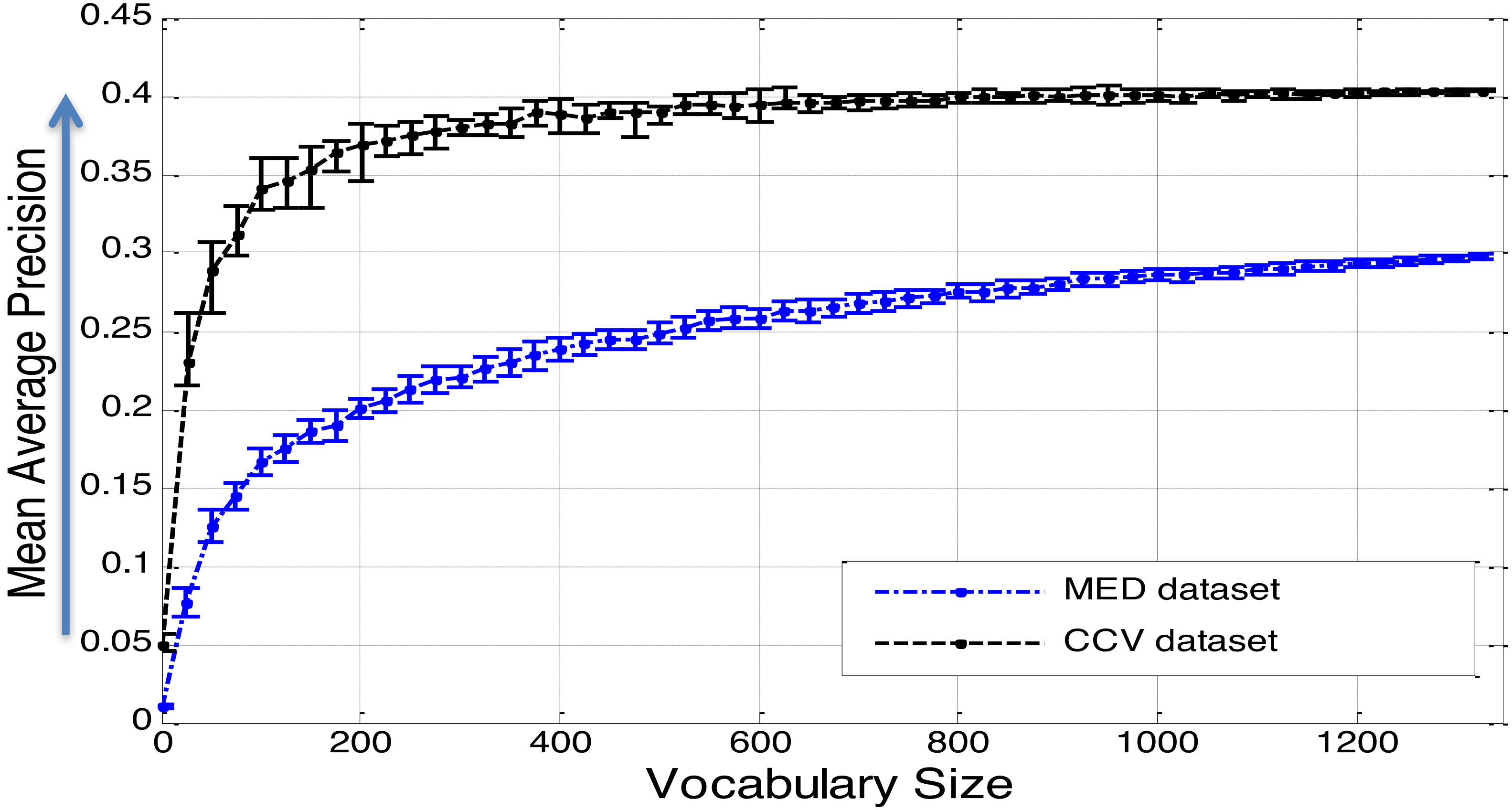
Detectors	100% Examples ColorSIFT Spatial Pyramid	30% Examples ColorSIFT Spatial Pyramid	30% Examples SIFT Spatial Pyramid	30% Examples SIFT
MAP	0.359	**0.371**	0.354	0.353

More sophisticated detectors have only minor influence on the overall event recognition accuracy.

Semantic encodings for video

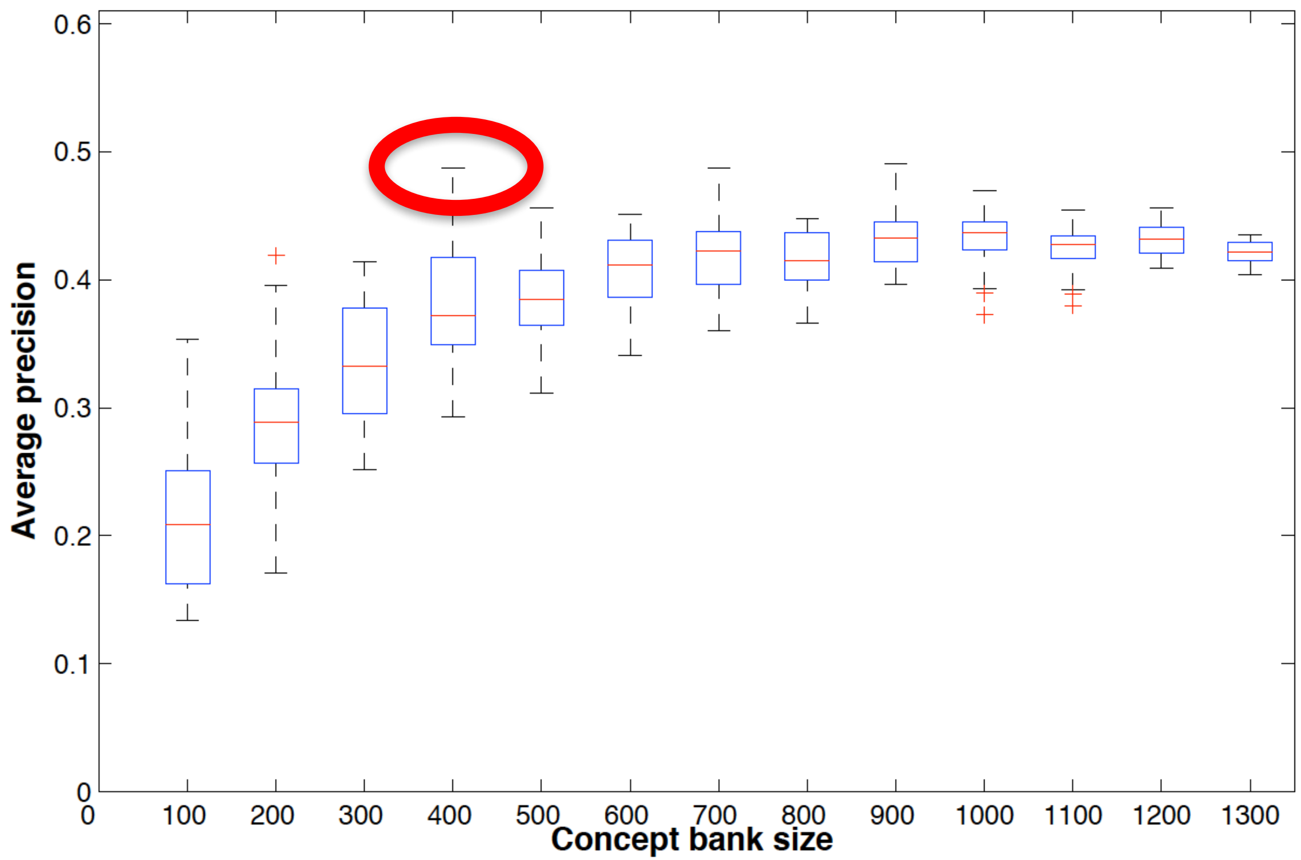
1. How many concepts?
2. What concept types?
3. Which concepts?
4. How accurate?
5. **How to select?**

5. Motivation



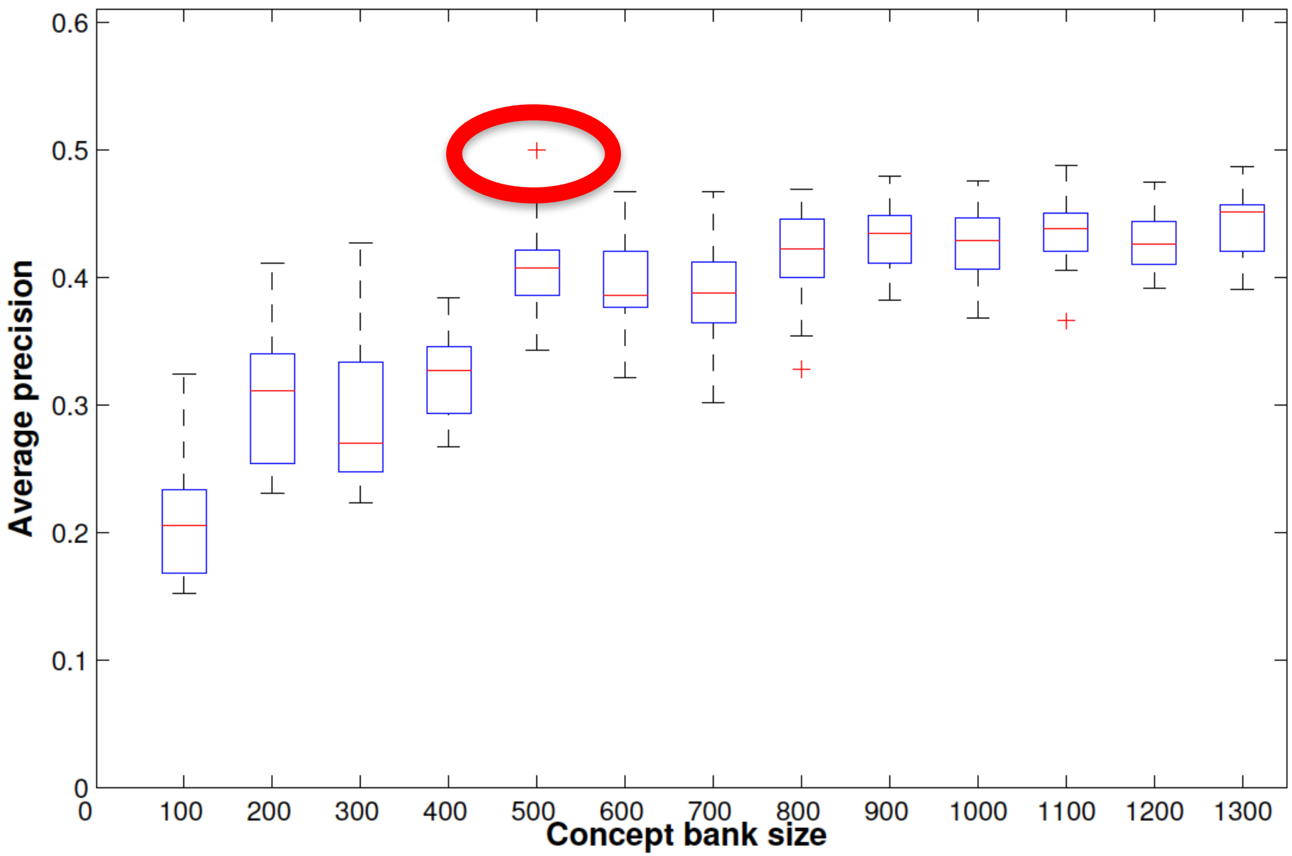
More is better, but include at least 200

Example for: *Landing a fish in*



A vocabulary of 400 concepts is more accurate than using all

Example for: *Wedding ceremony*



A vocabulary of 500 concepts is more accurate than using all

Mazloom et al., TMM 2014

Value of individual concepts

Board trick			**Wedding ceremony**			**Flash mob gathering**		
Concept	AP	Positives	Concept	AP	Positives	Concept	AP	Positives
Skating	0.194	1,300	*Church*	0.396	1,300	*Crowd*	0.280	2,341
Road	0.171	1,096	*Altar*	0.324	1,300	*3 or more people*	0.214	2,099
Snow	0.162	1,013	*Gown*	0.306	1300	*People marching*	0.205	624
Snowplow	0.123	540	*Groom*	0.288	1,280	Street battle	0.202	1,300
Ski	0.119	1,096	*Suit*	0.251	1,300	Meeting	0.186	340
Basketball			**Swimming**			**Parade**		
Concept	AP	Positives	Concept	AP	Positives	Concept	AP	Positives
Basketball	0.488	1,300	*Swimming*	0.698	1,300	*People marching*	0.318	624
Throw ball	0.485	811	*Swimming pool*	0.621	1,300	*Urban scenes*	0.155	1,403
Throwing	0.432	1,300	*Underwater*	0.432	1,300	*Police van*	0.150	1,300
Indoor sport venue	0.355	1,300	Stingray	0.227	1,300	*3 or more people*	0.138	2,099
Gym	0.337	153	*Waterscape/Waterfront*	0.211	604	*Streets*	0.135	1,300

Note the semantic correspondence between good performing concepts and events

Research question 5.

Is it possible to learn the semantic encoding of an event from examples?

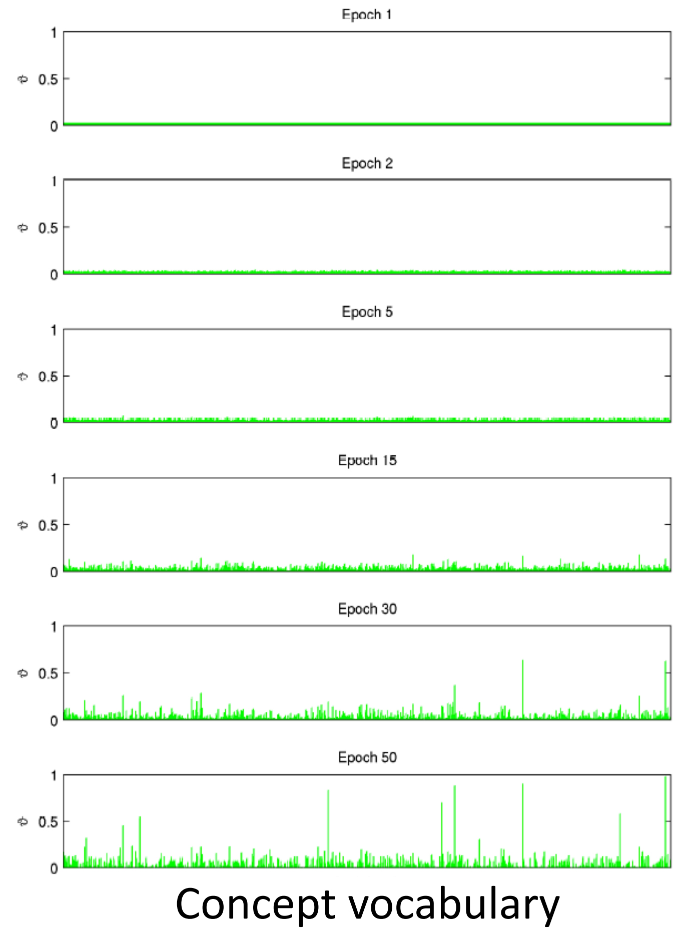
Idea

Formalize subset selection as importance sampling

Cross-entropy optimization

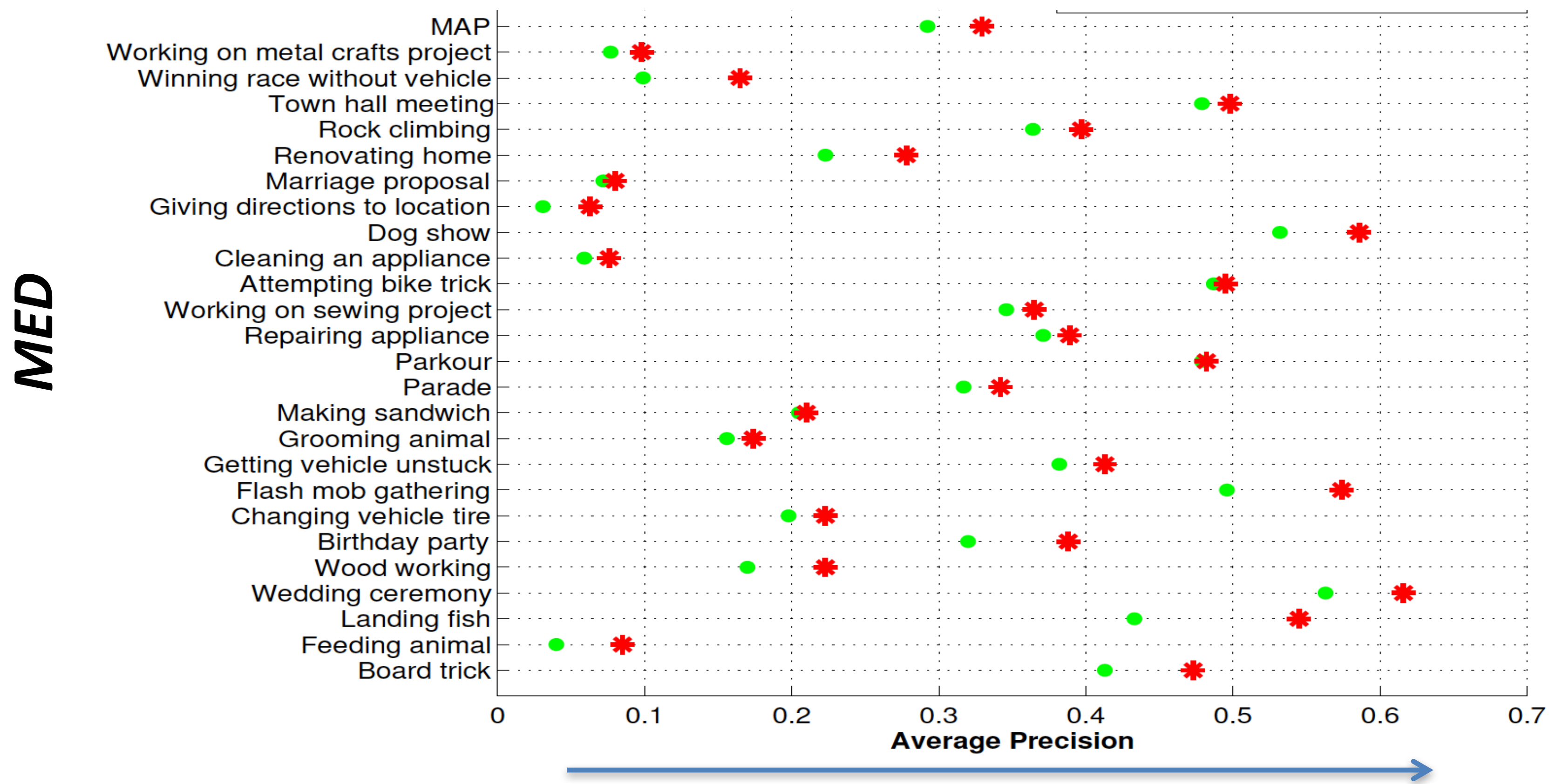
1. Sample semantic subset
2. Evaluate semantic subset
3. Update sampling parameters

Near-optimal solution



Mazloom et al., TMM 2014

All concepts (•) vs selected concepts (*)



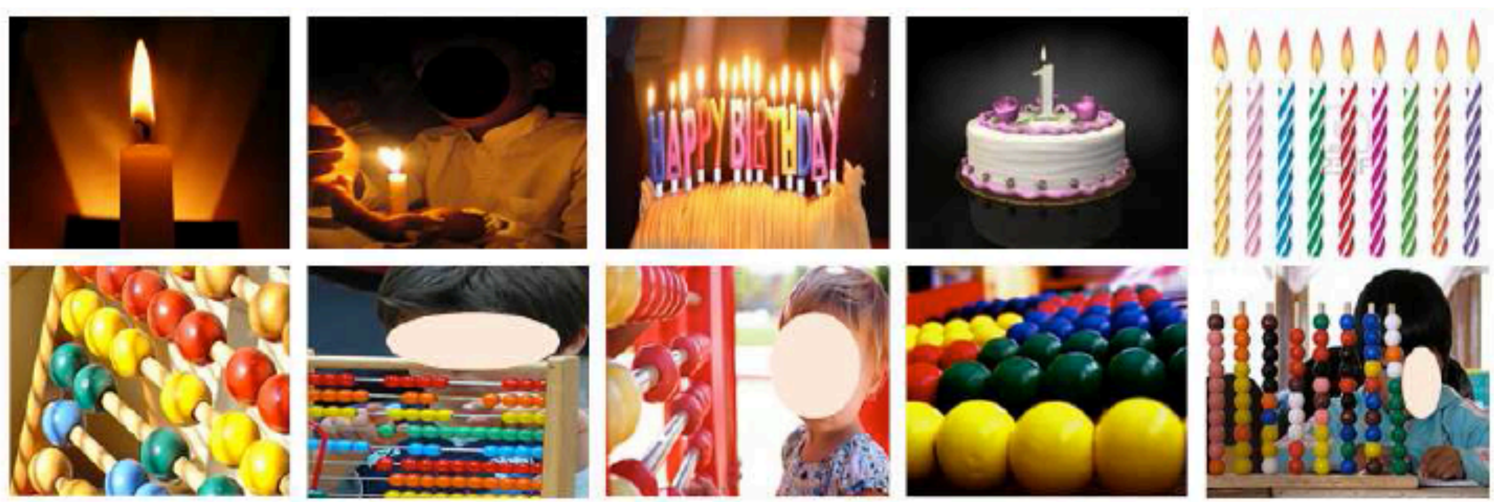
Encoding based on selected concepts always better

Concept subsets are descriptive

Font size correlates with importance

Failure case

Why is an 'Abacus' descriptive for Birthday?



Example training examples for candle and abacus

Recommendations

For event recognition using semantic encodings

1. Include at least 200 detectors
2. Diversity of concept types is important
3. Both the general and specific concepts are required
4. Concept detector accuracy is not critical
5. A descriptive concept subset can be learned from examples

Amirhossein Habibian and Cees G. M. Snoek, "**Recommendations for Recognizing Video Events by Concept Vocabularies**," *Computer Vision and Image Understanding*, vol. 124, pp. 110-122, 2014.

Part II

RETRIEVAL

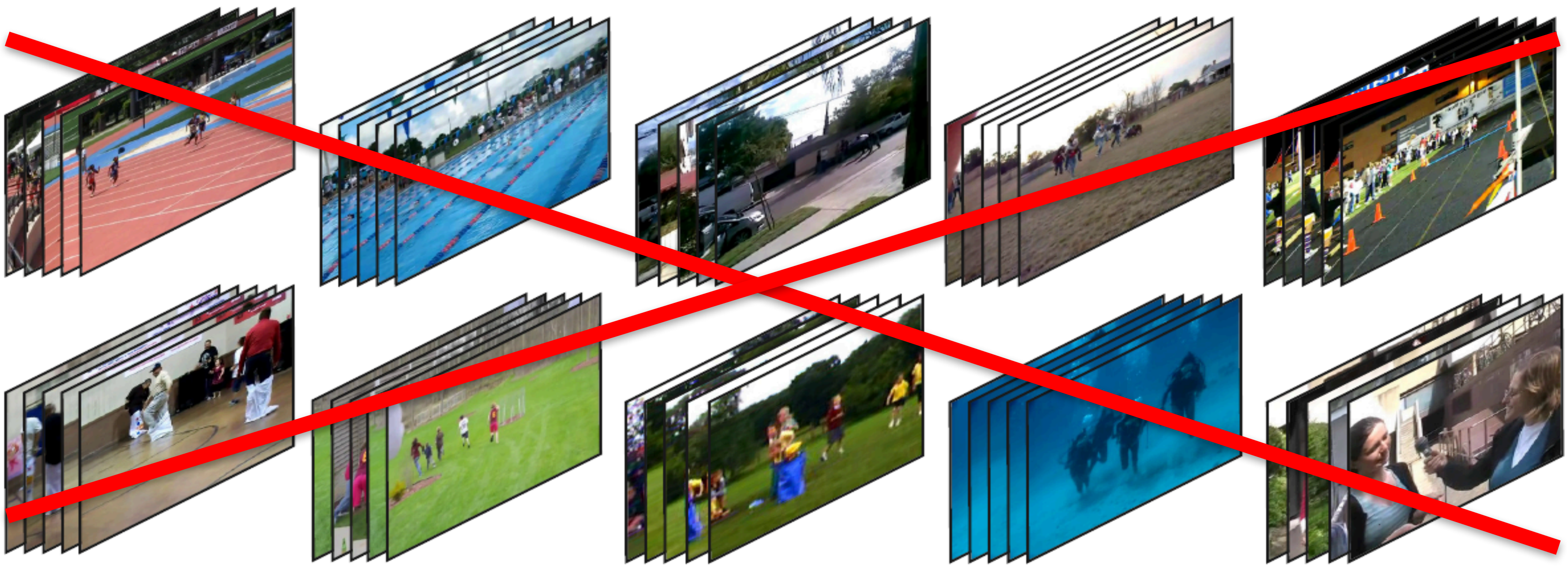
Joint work with Amirhossein Habibian & Masoud Mazloom

Hypothesis

As events become more and more specific, it is unrealistic to assume that ample examples to learn from will be commonly available.




Goal

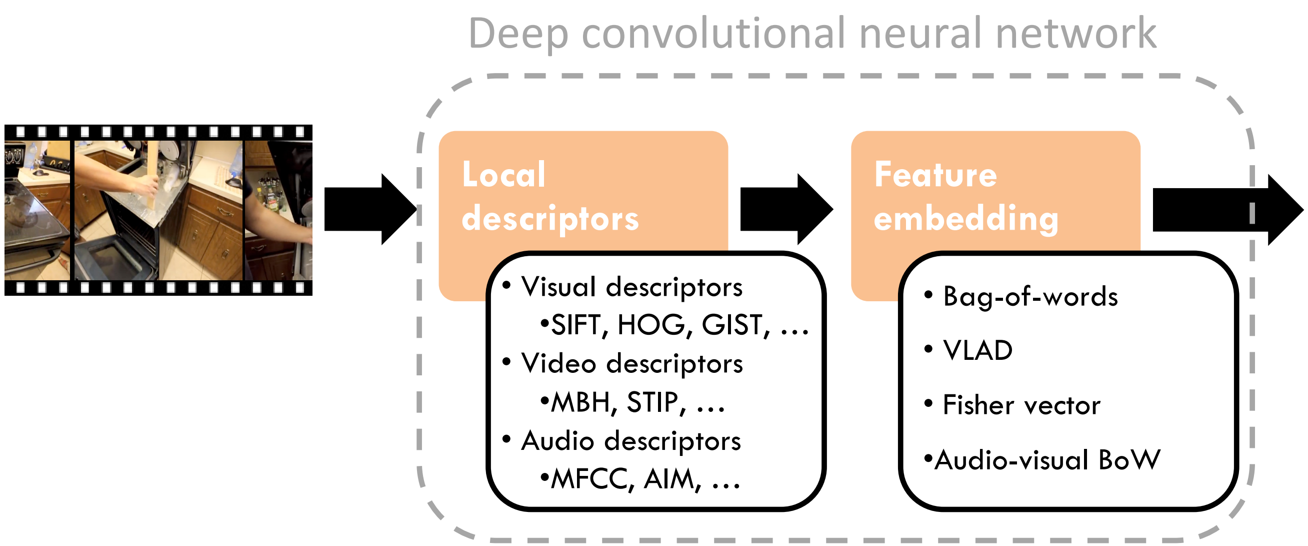


Event Name: Winning a race without a vehicle

Definition: An individual (or more) succeeds in reaching a pre-determined destination before all other individuals, without vehicle assistance or assistance of a horse or other animal. Racing generally involves accomplishing a task in less time than other competitors. The only type of racing considered relevant for the purposes of this event is the type where the task is traveling to a destination, completed by a person(s) without assistance of a vehicle or animal. Different types of races involve different types of human ...

Feature embedding fails

Representing videos as histograms of low-level features



Problem: demands examples

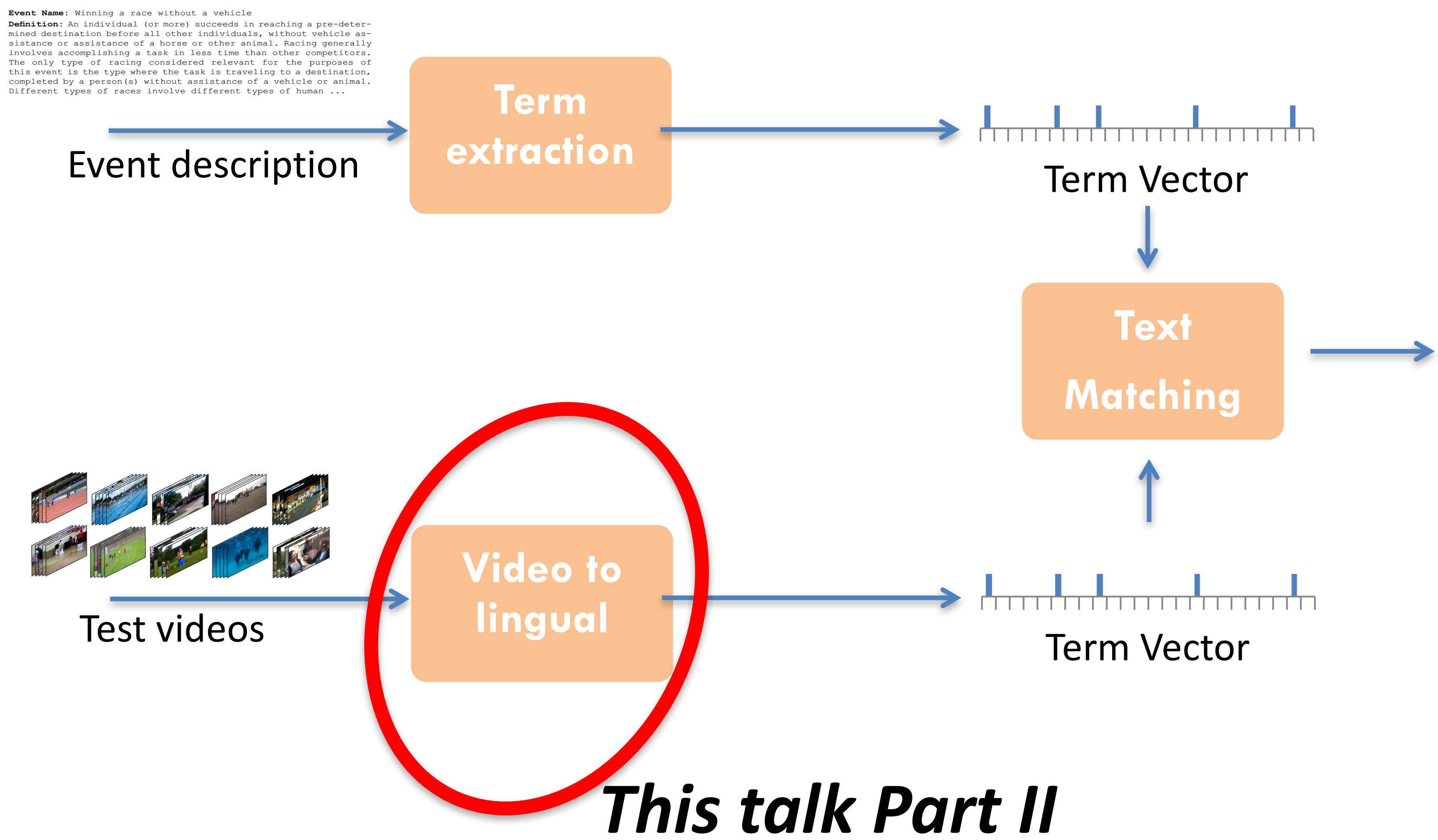
[Jiang et al., TRECVID 2010] [Natarajan et al., CVPR 2012] [Wang et al., ICCV 2013] and many others

Solution

The key to event recognition when examples are absent is to have a **lingual** video representation.

Once the video is represented in a textual form, standard retrieval metrics can be used

Event recognition, **without examples**



This part: three lingual representations

Concept embedding

Tag embedding

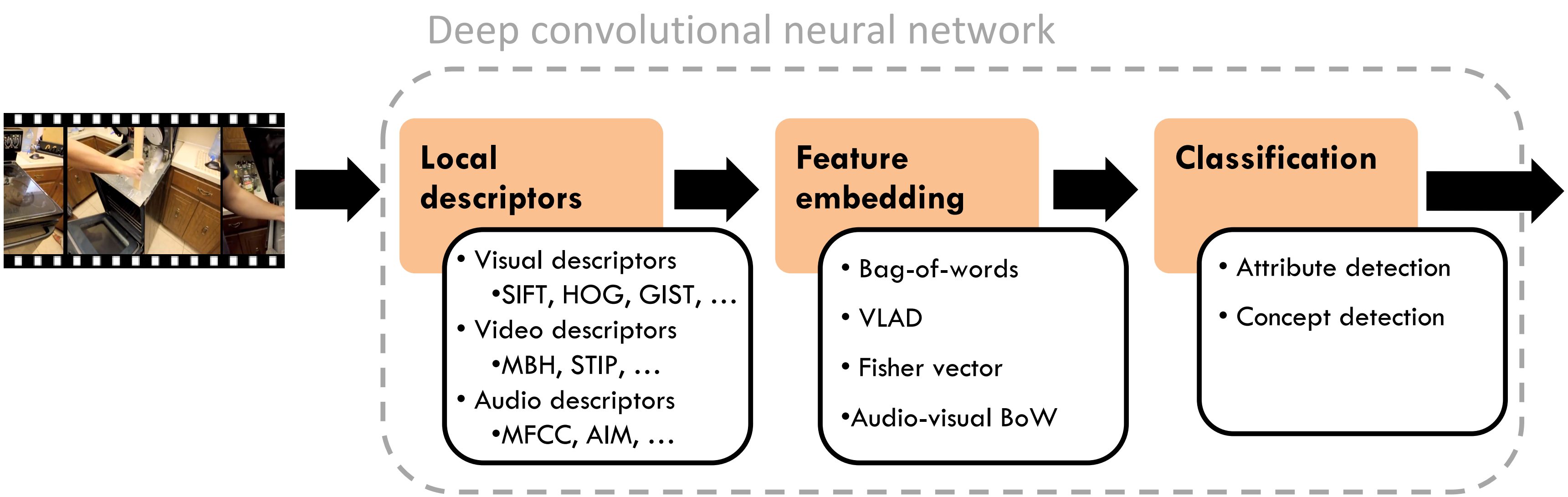
Video2vec embedding

Chapter 3

CONCEPT EMBEDDING

Concept embedding

Representing videos as histograms of concept scores



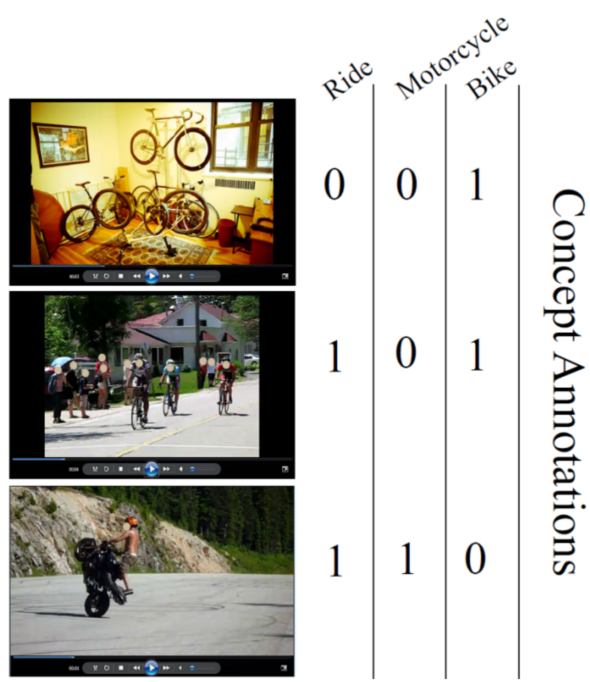
Problem: define, annotate and train concept classifiers

[Ebadollahi et al., ICME 2006] [Merler et al., TMM 2012] [Habibian et al., CVIU 2014] and many others

Label composition trick

Expanding the labels by logical operations

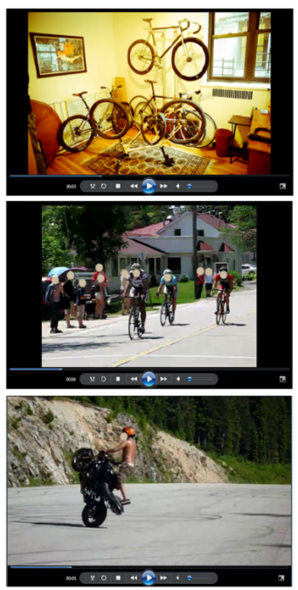
- AND, OR, …



Label composition trick

Expanding the labels by logical operations

- AND, OR, …



Ride	Motorcycle	Bike	Bike-AND-Ride	Bike-OR-Motorcycle
0	0	1	0	1
1	0	1	1	1
1	1	0	0	1

Concept Annotations

Motivation

Expanding the vocabulary for *free*

Composite concepts can be easier to detect

- boat-AND-sea
- bear-AND-cage
- man-OR-woman

Composite concepts can be more indicative of the event

- bike-AND-ride for *attempting a bike trick*

Learning composite concepts

For a vocabulary of n concepts, there are B_n disjoint compositions

- Bell number: $B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$
- Not all of them are useful

Which concepts should be composed together?

- NP-hard problem, equivalent to set-partitioning
- Approximated by a greedy search algorithm

Habibian et al. ICMR 2014

Qualitative results

Top ranked videos for *flash mob gathering*

Most dominant concepts in the video representation

Detected Videos	Composite Concepts
	Group-AND-Dance-AND-Shopping Celebrating-OR-Marching Performance-OR-Music People-OR-Girl Surprise-OR-Party
	Group-AND-Dance-AND-Shopping Band-OR-Singining Inside-OR-School Performance-OR-Music Surprise-OR-Party
	Group-AND-Dance -AND-Shopping Practice-OR-Gym Living-AND-Room Street-OR-Inside Performance-OR-Music

Composite concepts

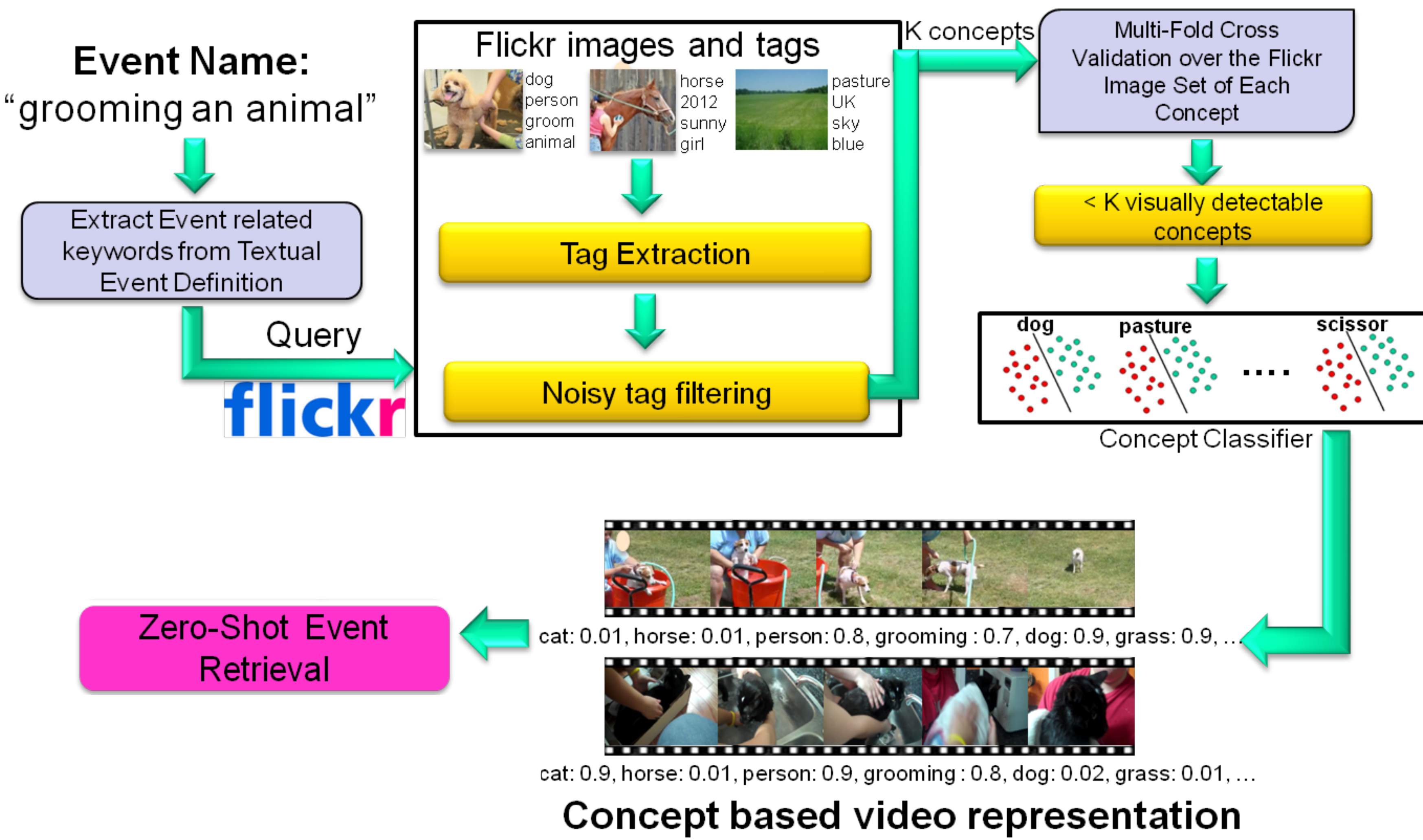
Label composition leads to a more comprehensive concept embedding

Still need to define, annotate and train concept classifiers

Greedy search algorithm slow

Chen et al. ICMR 2014
Wu et al. CVPR 2014

Discovering concepts from the web



Drawbacks of concept discovery

Big computational effort

Many concepts are rare, insufficient examples to train reliable visual classifiers

Selection is based on visual prediction accuracy only, descriptiveness is ignored

Contextual information is lost, since concepts are learned independently by binary classifiers.

Chapter 4

TAG EMBEDDING

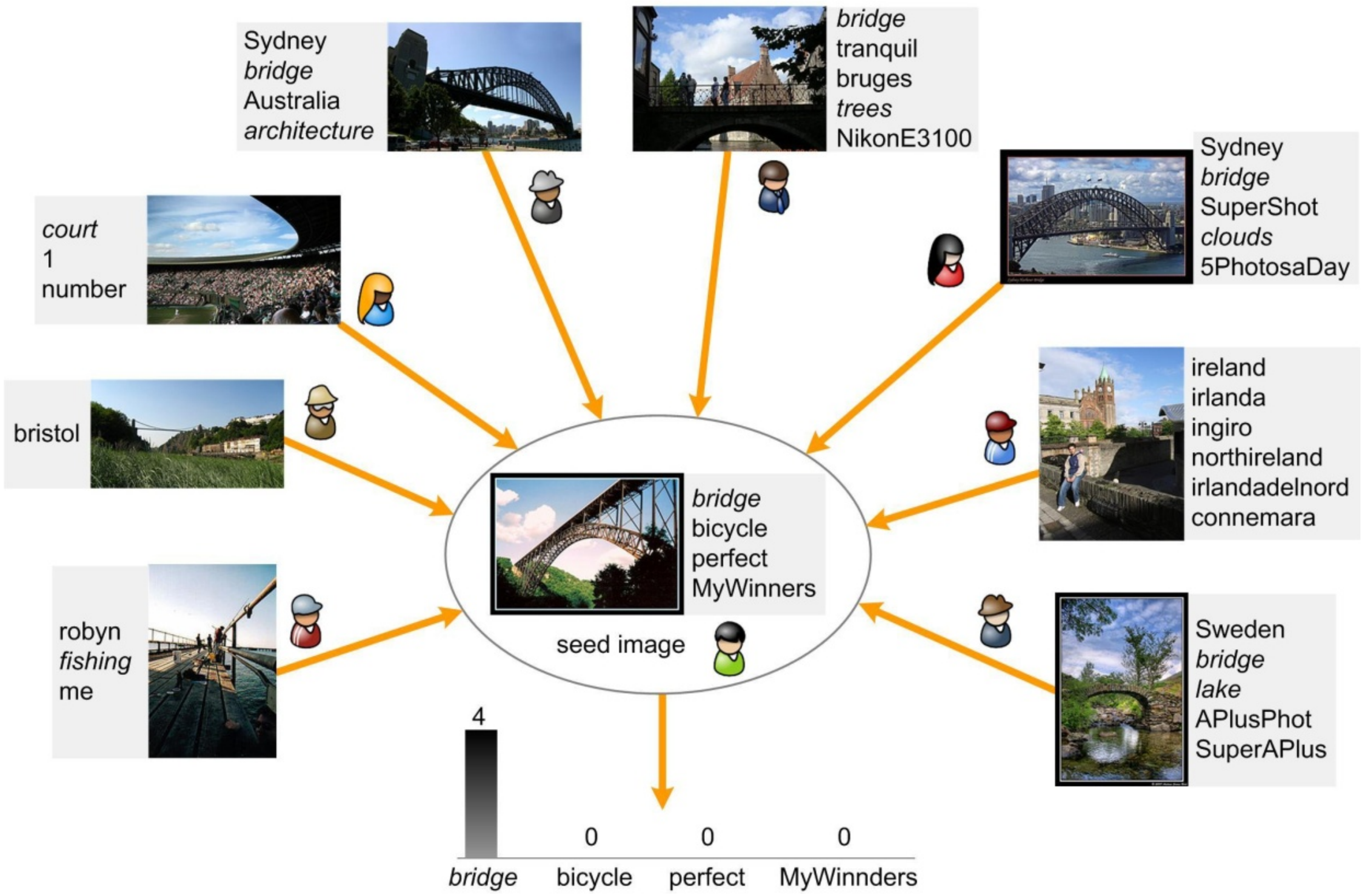
Masoud Mazloom, Xirong Li, and Cees G. M. Snoek,
TagBook: A Semantic Video Representation without Supervision for Event Detection,
IEEE Transactions on Multimedia, in press.

Idea

Embedding based on freely available social tagged videos only

Without the need for training any intermediate concept detectors

Inspiration



TagBook: embedding derived from social tags

Social-tagged web videos

Video data

Tags

woman, outdoor, metal-crafts-project, welding machine

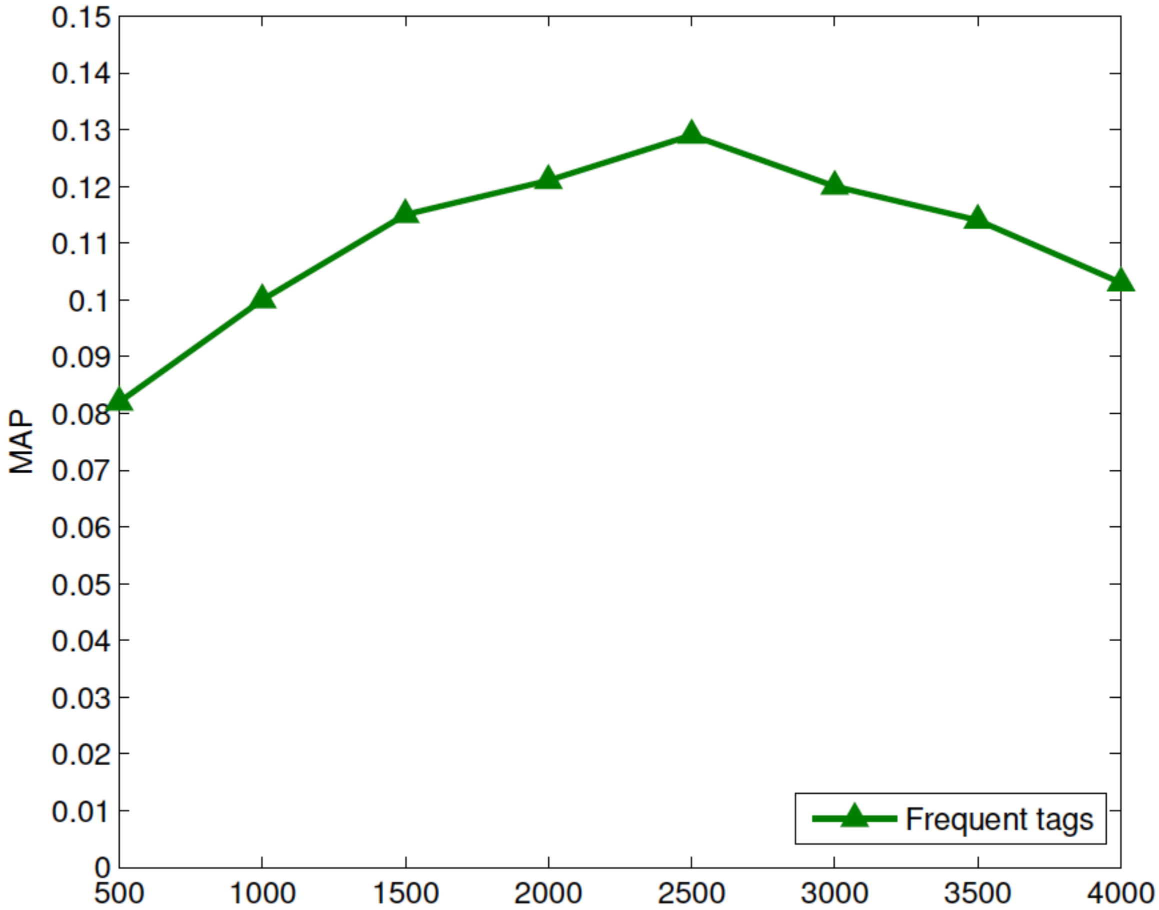
man, kitchen, metallic, cleaning, oven, spray, glasses,

man, snowboard, snow, board-trick,

man, climb-on, wall, gym, rock-climbing

TagBook = {woman, outdoor, metal-crafts-project, welding machine, man, kitchen,…, wall, gym, rock-climbing}

TagBook dimension



It is advantageous to select most frequent tags in TagBook

Chapter 5

VIDEO2VEC EMBEDDING

Amirhossein Habibian, Thomas Mensink, and Cees G. M. Snoek.
Video2vec Embeddings Recognize Events when Examples are Scarce.
IEEE Transactions on Pattern Analysis and Machine Intelligence. In press.
Previously best paper ACM Multimedia 2014.

Research question

Can we **learn the embedding** from videos and their stories?

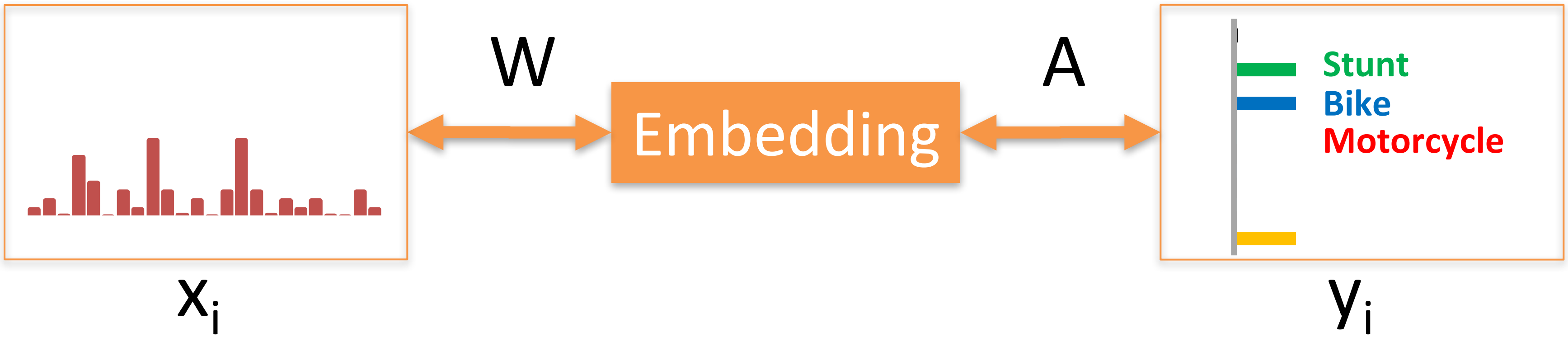
Video

Story

Story usually highlights the key concepts in video

Videos and stories are freely available, *i.e.* YouTube

Multimedia embeddings



Joint space where $x_i\,W \approx y_i\,A$

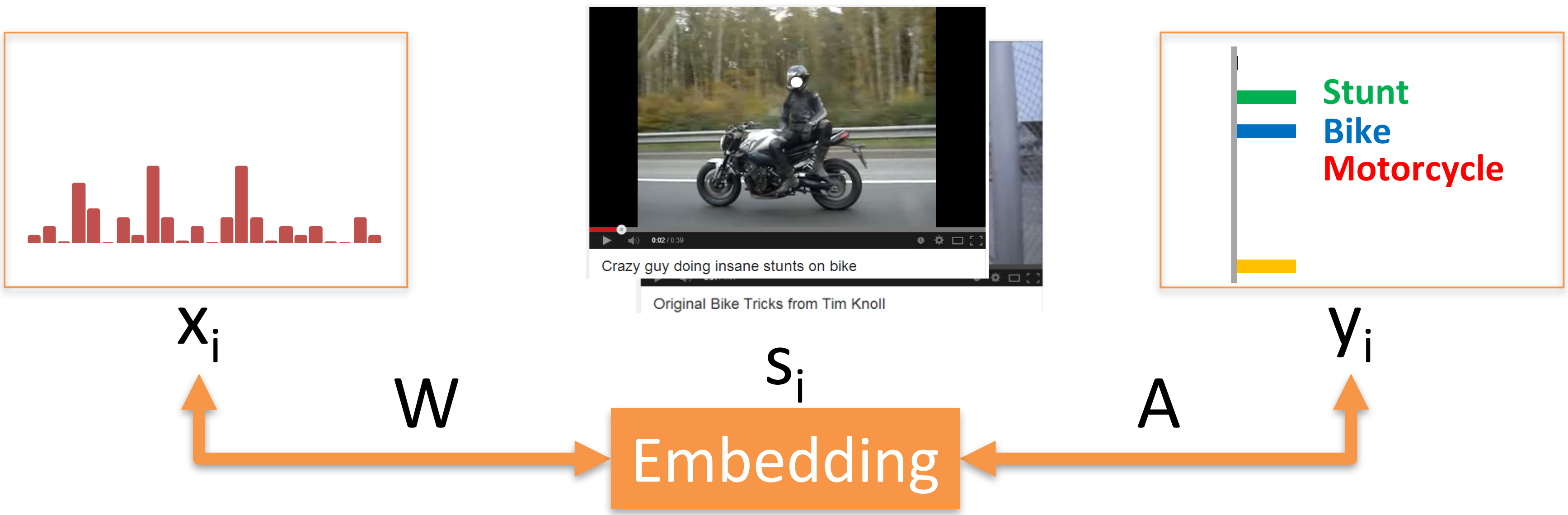
Explicitly relate training W and A from multimedia

W = Visual projection matrix   individual term classifiers

A = Textual projection matrix   select/group terms

[Rasiwasa et al., MM 2010] [Weston et al., IJCAI 2011] [Akata et al., CVPR 2013] [Das et al., WSDM 2013]

Video2vec: Embed the story of a video



Design criteria: learn W and A such that

Descriptiveness: preserve video descriptions

Predictability: recognize terms from video content

Key observation: Compelling forces

Why is this important?

Grouping terms:

Number of classes is reduced

Training classifiers per group:

More positive examples available per group

We can train from freely available web data

Key contribution: Joint optimization

Jointly optimize for descriptiveness and predictability

$$L_{\mathrm{VS}}(\boldsymbol{A}, \boldsymbol{W}) = \min_{\boldsymbol{S}} L_d(\boldsymbol{A}, \boldsymbol{S}) + L_p(\boldsymbol{S}, \boldsymbol{W})$$

Hyperparameter: size of the embedding S

L_d Loss function for descriptiveness

L_p Loss function for predictability

Video2vec connects the two loss functions

Video2vec objectives: descriptiveness

Objective 1: The Video2vec embedding should be **descriptive**

$$L_d(\boldsymbol{A}, \boldsymbol{S}) = \frac{1}{N}\sum_{i=1}^{N} \|\underbrace{\boldsymbol{y}_i}_{\text{Original transcriptions}} - \underbrace{\boldsymbol{A}\boldsymbol{s}_i}_{\text{Reconstructed terms}}\|_2^2 + \underbrace{\lambda_a \Omega(\boldsymbol{A}) + \lambda_s \Psi(\boldsymbol{S})}_{\text{Regularizers}}$$

Essentially latent semantic indexing with L2 rather than an L1 norm

Video2vec objectives: predictability

Objective 2: The Video2vec embedding should be **predictable**

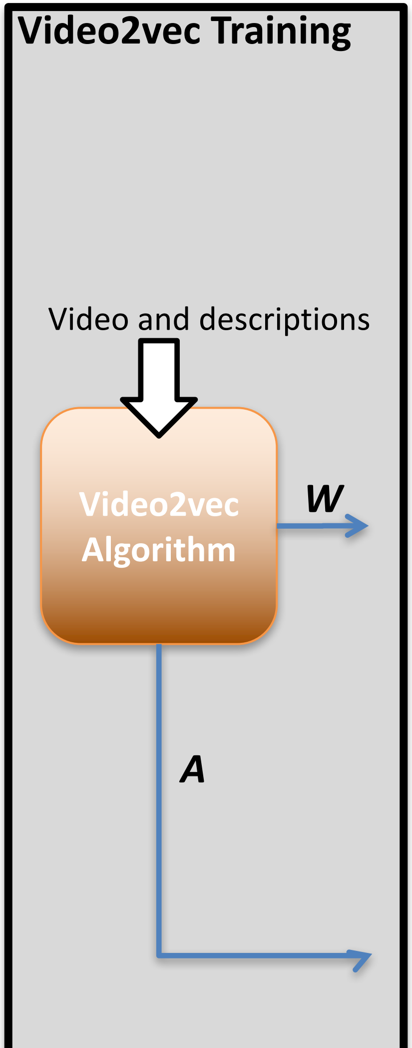
$$L_p(\boldsymbol{S}, \boldsymbol{W}) = \frac{1}{N} \sum_{i=1}^{N} \| \boldsymbol{s}_i - \boldsymbol{W}^\top \boldsymbol{x}_i \|_2^2 + \lambda_w \Theta(\boldsymbol{W})$$

Video2vec embedding

Video feature embedding

Regularizer

Video2vec: Training



Set of videos and their captions

Encode video features x_i

Any feature (combination) will do

Encode video descriptions y_i

Bag-of-words of terms

[Habibian MM 2014]

VideoStory46K dataset

Videos and title descriptions from YouTube

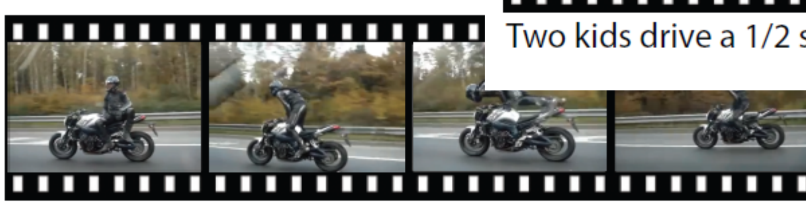
46K videos, 19K unique terms in descriptions

Seeded from video event descriptions

Filters to remove low quality videos

Cute tabby cat gives her dog a bath

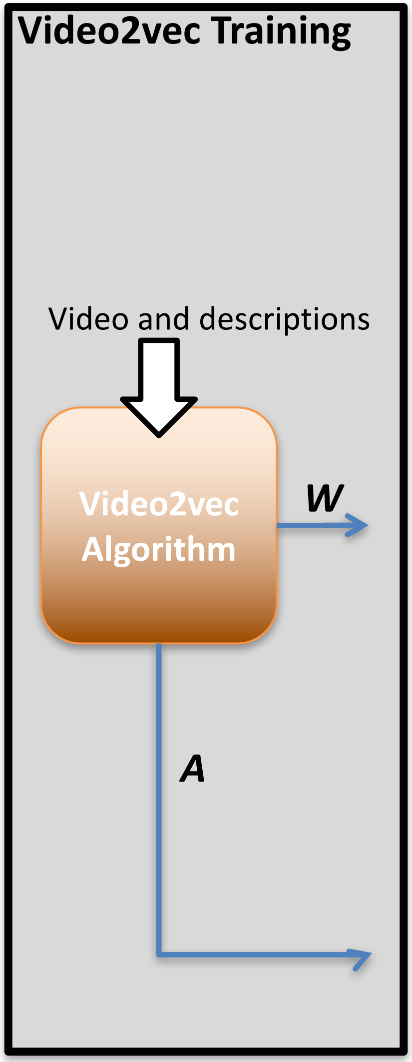
Two kids drive a 1/2 size Jeep through mud



Crazy guy doing insane stunts on bike.

Available for download: www.mediamill.nl

Video2vec: Training (2)



Using *Stochastic Gradient Descent:*

Choose random sample

Compute sample gradient wrt objective

$$\nabla_{\boldsymbol{A}} L_{\mathrm{VS}} = -2\left(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{s}_t\right)\boldsymbol{s}_t^{\top} + \lambda_a \boldsymbol{A},$$

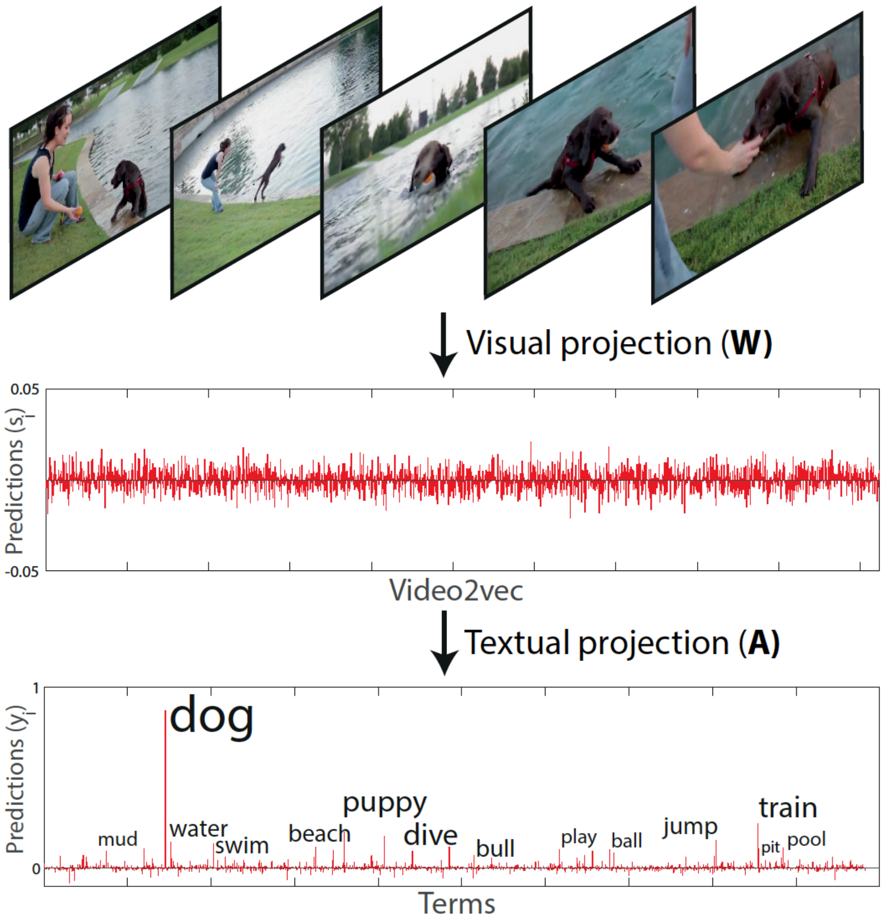
$$\nabla_{\boldsymbol{W}} L_{\mathrm{VS}} = -2\,\boldsymbol{x}_t\left(\boldsymbol{s}_t - \boldsymbol{W}^{\top}\boldsymbol{x}_t\right)^{\top} + \lambda_w \boldsymbol{W}, \text{ and}$$

$$\nabla_{\boldsymbol{s}_t} L_{\mathrm{VS}} = 2\left[\boldsymbol{s}_t - \boldsymbol{W}^{\top}\boldsymbol{x}_t - \boldsymbol{A}^{\top}\left(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{s}_t\right)\right] + \lambda_s \boldsymbol{s}_t.$$

Update parameters with step-size η

[Bottou ICCS 2010]

Video2vec at work



1. Project visual features

$$\boldsymbol{s}_i = \boldsymbol{W}^\top \boldsymbol{x}_i,$$

2. Translate to text

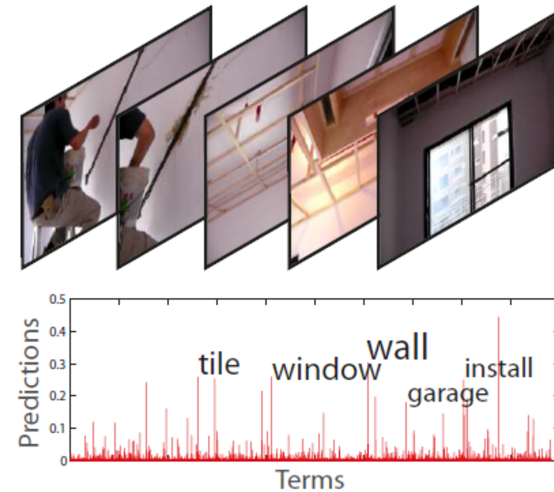
$$\hat{\boldsymbol{y}}_i = \boldsymbol{A}\boldsymbol{s}_i,$$

3. Cosine distance for matching

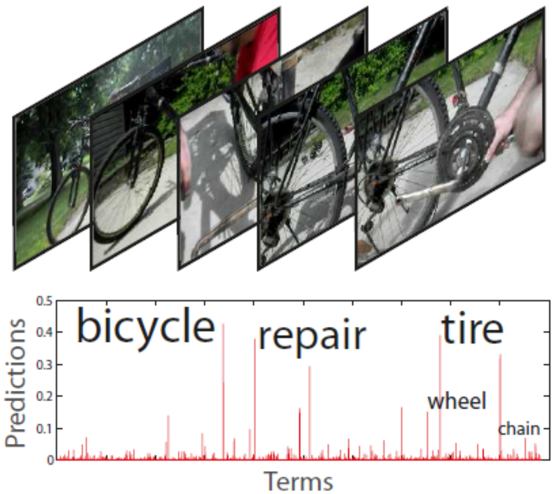
$$s_e(\boldsymbol{x}_i) = \frac{\boldsymbol{y}^{e\top}\hat{\boldsymbol{y}}_i^e}{||\boldsymbol{y}^e|| \quad ||\hat{\boldsymbol{y}}_i^e||}$$

Video2vec predicted terms

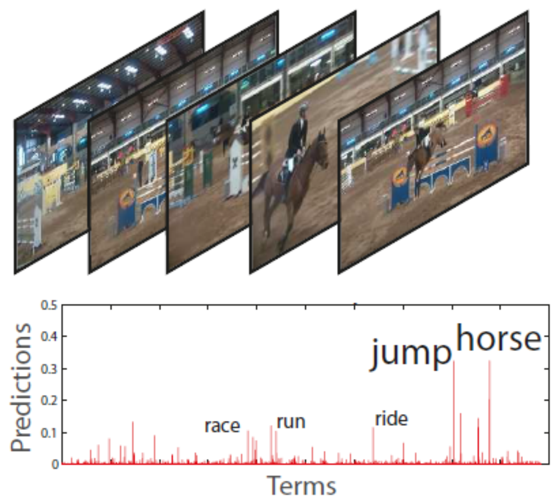
renovating a home



non-motorized vehicle repair



horse riding competition



State-of-the-art event retrieval

Authors	Published	mAP
Habibian et al.	ICMR 2014	6.4
Ye et al.	MM 2015	9.0
Chang et al.	IJCAI 2015	9.6
Mazloom et al.	ICMR 2015	11.9
Wu et al.	CVPR 2014	12.7
Jiang et al.	AAAI 2015	12.9
Mazloom et al.	TMM 2016	12.9
Liang et al.	MM 2015	18.3
Habibian et al.	TPAMI 2017	20.0

State-of-the-art event retrieval

Authors	Published	mAP
Concept embedding	ICMR 2014	6.4
Ye et al.	MM 2015	9.0
Chang et al.	IJCAI 2015	9.6
Mazloom et al.	ICMR 2015	11.9
Wu et al.	CVPR 2014	12.7
Jiang et al.	AAAI 2015	12.9
Tag embedding	TMM 2016	12.9
Liang et al.	MM 2015	18.3
Video2vec embedding	TPAMI 2017	20.0

State-of-the-art: event classification

Authors	Published	mAP
Habibian et al.	MM 2014	19.6
Nagel et al.	BMVC 2015	21.8
Li et al.	ICCV 2013	23.7
Tang et al.	CVPR 2012	26.8
Sun et al.	CVPR 2014	28.7
Chang et al.	MM 2015	30.9
ImageNet-shuffle	ICMR 2016	34.8
Video2vec embedding	TPAMI 2017	37.1

Conclusions

Event recognition without examples demands lingual representation

Concept embedding has too many limitations

Tag embedding is simple, yet surprisingly effective

Video2vec's descriptiveness & predictability is appealing