

What objects tell about actions

Cees Snoek

Qualcomm Technologies
Netherlands B.V.



University of Amsterdam
The Netherlands



Goal: action recognition



Balance Beam



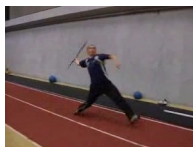
Blowing Candles



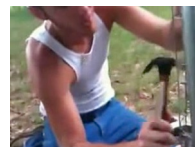
Bowling



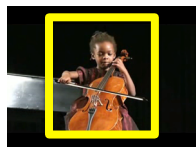
Brushing Teeth



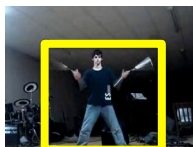
Javelin Throw



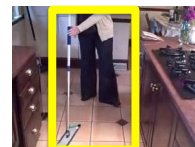
Hammering



Playing Cello



Nunchucks



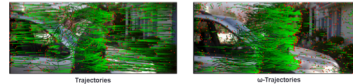
Mopping Floor

Dan Oneata, PhD Thesis, 2015

Actions: state-of-the-art

Camera motion compensated trajectories [Wang & Schmid, ICCV13]

Local descriptors: HOG, HOF, MBH



Fisher vector video encoding [Perronnin et al, CVPR10]

Power and L2 normalization on PCA reduced vectors

Stacking multiple layers [Peng et al, ECCV14]

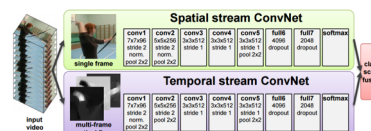
UCF101		THUMOS14 val		THUMOS14 test		Hollywood2		HMDB51	
Soomro et al. [41]	43.9%	Varol et al. [47]	62.3%	Varol et al. [47]	63.2%	Zhu et al. [60]	61.4%	Zhu et al. [60]	54.0%
Cai et al. [11]	83.5%			Oneata et al. [31]	67.2%	Vig et al. [48]	61.9%	Oneata et al. [30]	54.8%
Wu et al. [55]	84.2%					Jain et al. [14]	62.5%	Wang et al. [51]	57.2%
Wang et al. [52]	85.9%					Oneata et al. [30]	63.3%	Peng et al. [32]	59.8%
Peng et al. [32]	87.7%					Wang et al. [51]	64.3%		
								Peng et al. [33]	66.8%

Motion is the key ingredient in modern action recognition

Deep action learning

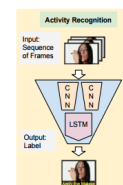
Two stream CNN

Simonyan & Zisserman, NIPS 2014



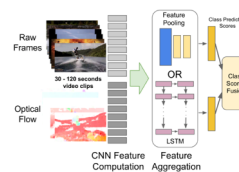
CNN outputs connected to LSTM

Donahue et al., CVPR 2015



Two streams and LSTM on snippets

Ng et al., CVPR 2015



Inspiration from language acquisition

Children first learn nouns, then verbs.

Nouns provide semantic and syntactic frames to aid in mapping the verb to its meaning.

Nouns pave the way for learning verbs?

[Gentner & Boroditsky, 2009](#)


PRELUDE: OBJECTS

www.image-net.org

Learning nouns from ImageNet

WordNet for images
14M images for 21K synsets

Yearly ImageNet competition
Automatically label 1.4M images with 1K objects
Measure top-5 classification error



Output

- Scale
- T-shirt
- Steel drum ✓
- Drumstick
- Mud turtle

Output

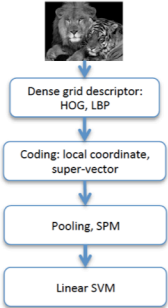
- Scale
- T-shirt
- Giant panda ✗
- Drumstick
- Mud turtle

Slide credit: Andrej Karpathy

Objects: state-of-the-art

Year 2010

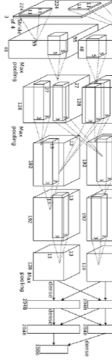
NEC-UIUC



Lin et al. CVPR11

Year 2012


SuperVision



Krizhevsky et al. NIPS12


Year 2014

GoogLeNet

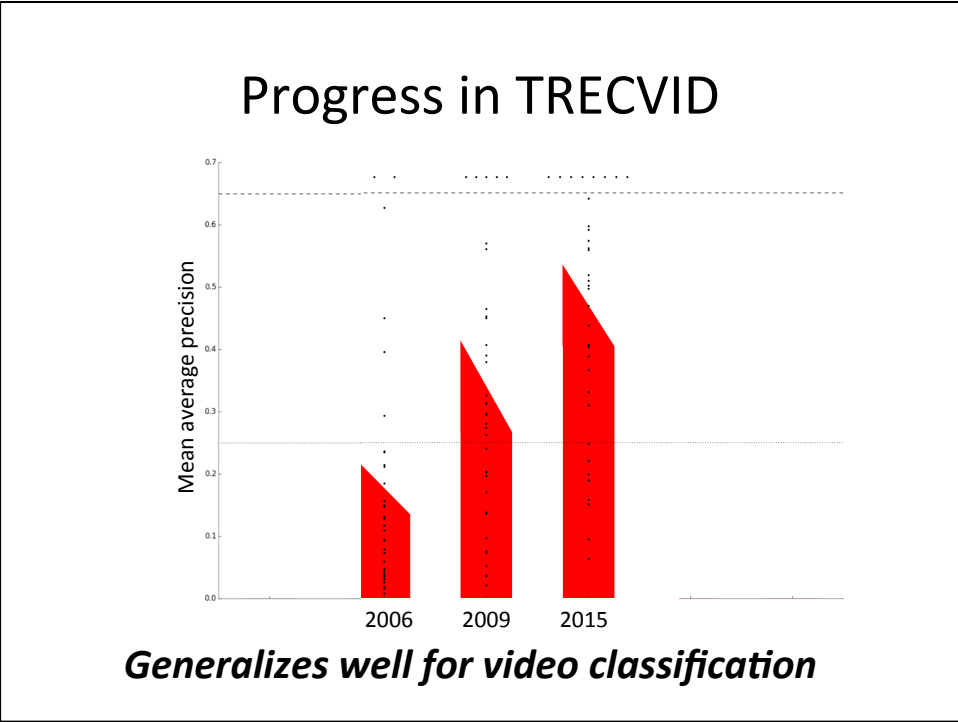
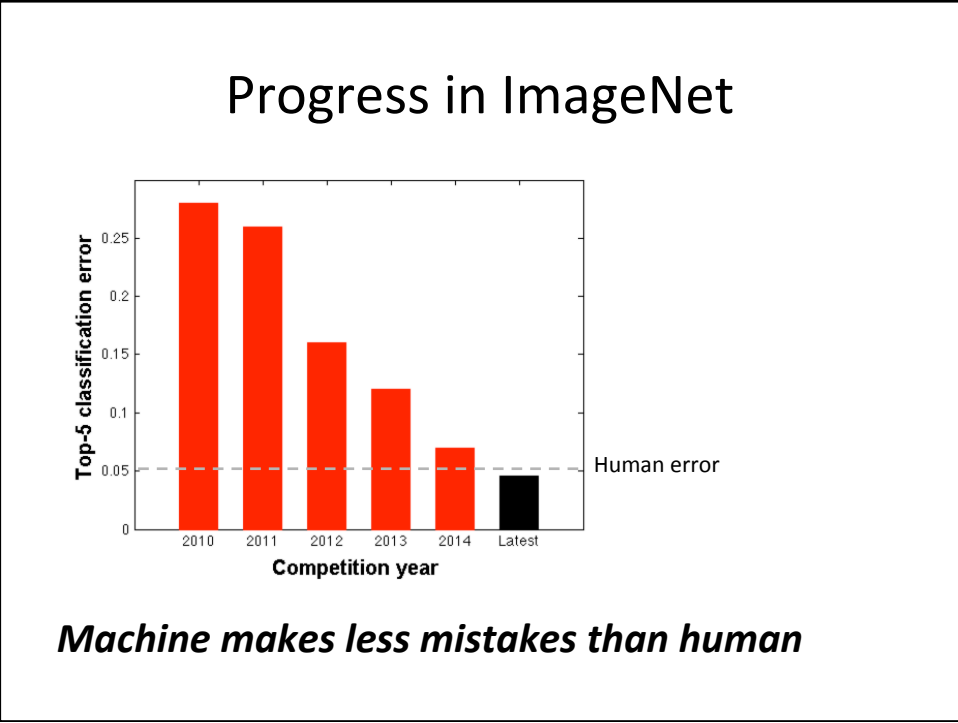


Szegedy et al. CVPR15

VGG



Simonyan et al. ICLR15



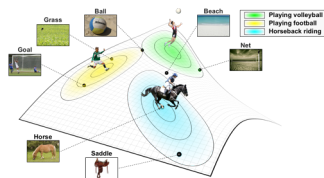
Outline

15,000 objects

Action classification for actions: a) Balance Beam, b) Bench Press and c) Floor Gymnastics

Action localization for actions: a) Lifting, b) Diving and c) Saving Bench

Supervised action recognition



Contribution

Empirical study on the benefit of having **objects** in the video representation for action recognition.




Mihir Jain





Jan van Gemert


What do 15,000 object categories tell us about classifying and localizing actions?
Mihir Jain, Jan van Gemert, and Cees Snoek. In *CVPR 2015*.


6 video datasets with 180 actions

UCF101  101 classes / 13,320 clips / web video

THUMOS14  101 classes / 15,915 clips / web video

Hollywood2  12 classes / 1,707 clips / movies

HMDB51  51 classes / 6,766 clips / diverse video

UCF Sports  10 classes / 150 clips / sports broadcasts

KTH  6 classes by 25 actors

Encoding video by 15,000 objects

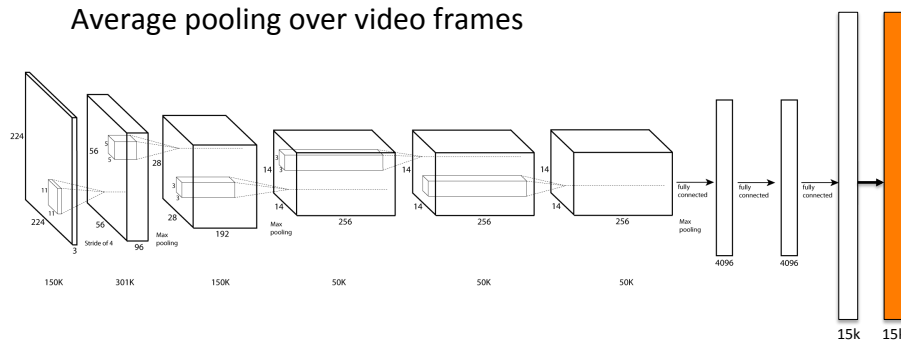
Krizhevsky-style cuda-convnet with dropout [\[NIPS12\]](#)

Convolutional neural network with 8 layers with weights

Trained using error back propagation

Learns from annotations for 15,000 ImageNet object categories

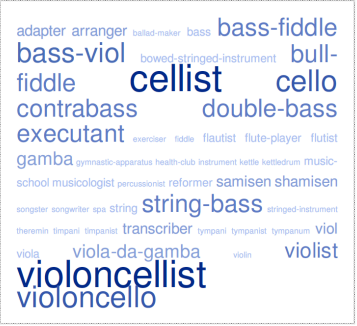
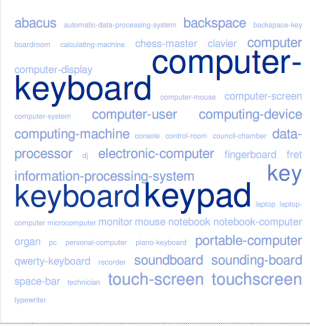


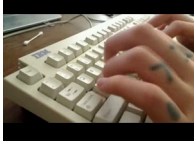

Average pooling over video frames



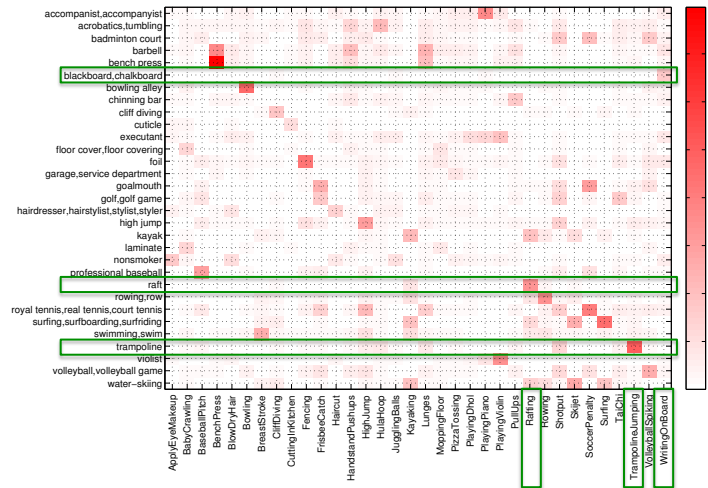
Experiment 1

OBJECTS: WHAT AND WHERE?

What objects emerge in actions?

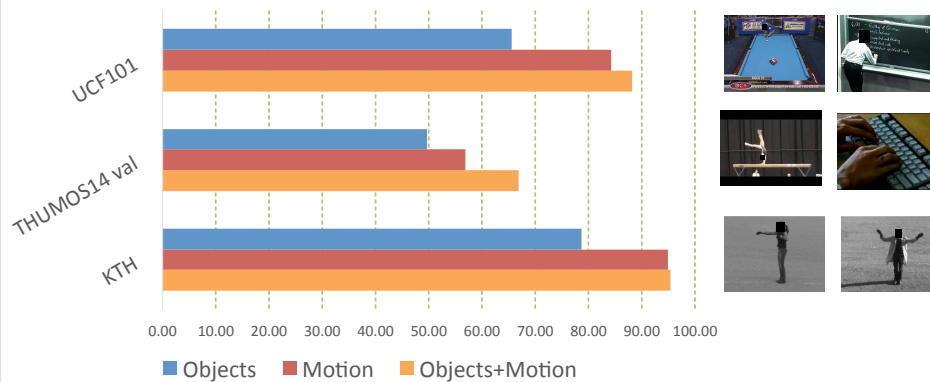
 <p>adapter arranger ballad-maker bass bass-fiddle bass-viol bowed-stringed-instrument bull-fiddle cellist cello contrabass double-bass executant exerciser fiddle flautist flute-player flutist gamba gymnastic-apparatus health-club instrument kettle drum music-school musicologist percussionist reformer samisen shamisen songster songwriter spa string-bass stringed-instrument theremin timpani timpanist transcriber tympani tympanist tympanum viol viola viola-da-gamba violin violinist violoncellist violoncello</p>	 <p>abacus automatic-data-processing-system backspace backspace-key boardroom calculating-machine chess-master clavier computer computer-display computer-keyboard computer-mouse computer-screen computer-system computer-user computing-device computing-machine console control-room council-chamber data-processor electronic-computer fingerboard fret information-processing-system key keyboard keypad laptop laptop-computer computer microcomputer monitor mouse notebook notebook-computer organ pc personal-computer piano-keyboard portable-computer qwerty-keyboard recorder soundboard sounding-board space-bar technician touch-screen touchscreen typewriter</p>	 <p>acrobatics balance-beam ballet-dancer barbell bar beam bench-press dumbbell exerciser exercising-weight figure-skating flat-bench free-weight gymnastic-apparatus gymnastic-exercise gymnastics health-club health-spa high-bar hop horizontal-bar pusher racquetball reformer rollerblading server singles skateboarding skating spa weight thruster tightrope-walking tread-wheel treadmill trapezium tumbler uneven-bars uneven-parallel-bars typewriter</p>
 <p>Playing Cello</p>	 <p>Typing</p>	 <p>Bodyweight squats</p>

Object responses per action



Object responses seem to make sense for most actions

Objects aid action classification?



Objects combined with motion always improve accuracy

Motion reliant actions



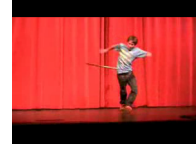
Wall Pushups



Tai Chi



Jumping Jack



Hula Hoop



Jump Rope



Trampoline Jumping



Lunges



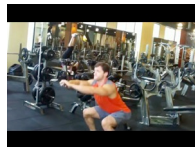
Uneven Bars



Pull Ups



Military Parade



Bodyweight Squats



Boxing Speed Bag

Object related actions



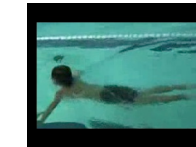
Playing Piano



Billiards



Baseball Pitch



Breast Stroke



Head Massage



Mixing



Soccer Penalty



Frisbee Catch



Rock Climbing Indoor



Archery

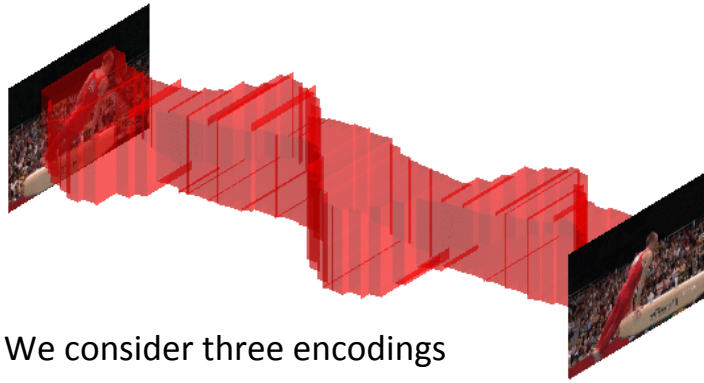


Cutting in Kitchen



Sumo Wrestling

Where do objects aid most?

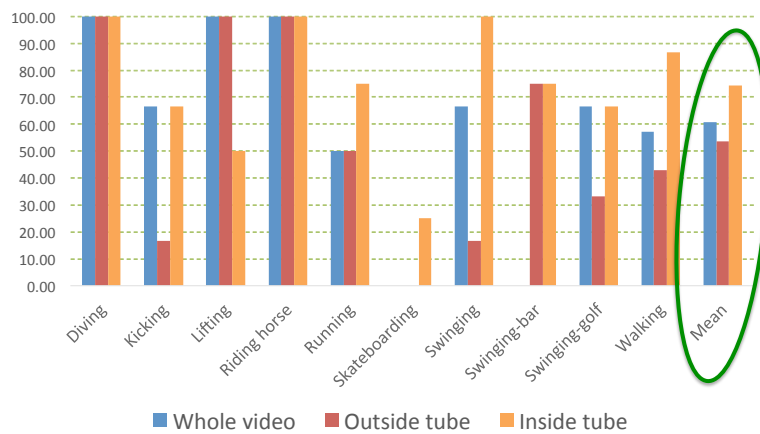


We consider three encodings

- Whole video
- Outside tube
- Inside tube

Animation credit: Jan van Gemert

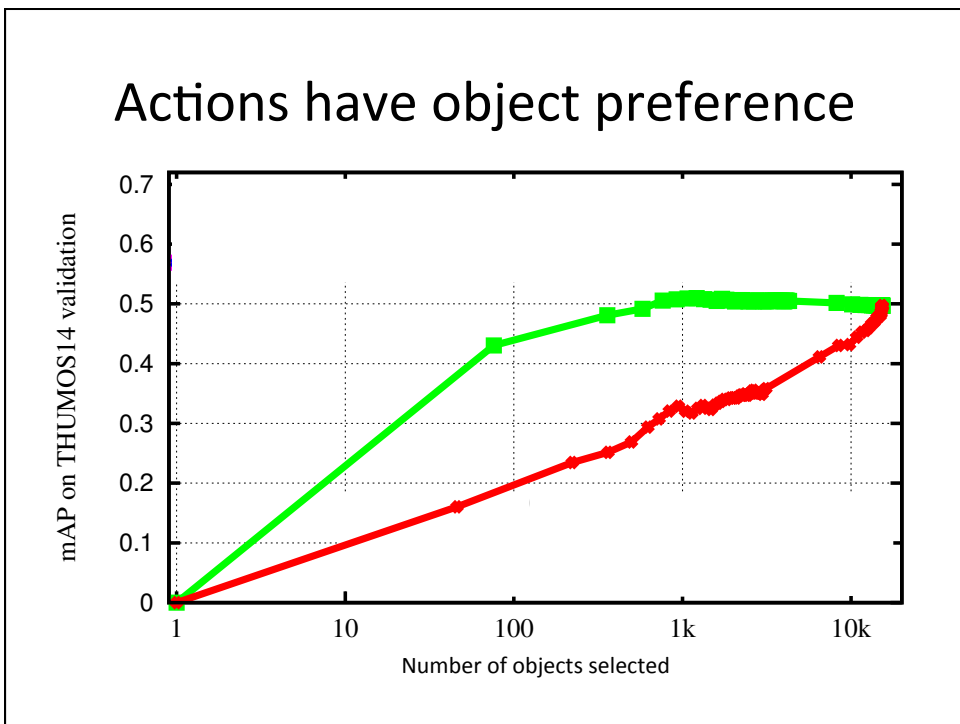
Where do objects aid most?



Objects aid most close to and involved in the action

Experiment 2

OBJECTS: SELECT AND GENERALIZE?



Learning what objects matter per action

HMDB51 and UCF101 share 12 action classes

We learn on training sets of HMDB51 and UCF101 what objects matter most per action

We test action classification on HMDB51 test set

Object-action relations are generic

	Motion	HMDB51
Brush hair	96.7	96.7
Climb	87.8	92.2
Dive	87.8	84.4
Golf	98.9	98.9
Handstand	90.0	90.0
Pullup	91.1	92.2
Punch	85.6	88.9
Pushup	72.2	88.9
Ride bike	76.7	91.1
Shoot ball	86.7	93.3
Shoot bow	92.2	94.4
Throw	37.8	36.7
Mean	83.6	87.5

Experiment 3

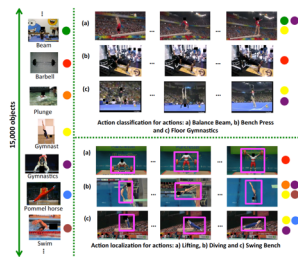
ACTIONS: STATE-OF-THE-ART

Action classification

UCF101		THUMOS14 val		THUMOS14 test		Hollywood2		HMDB51	
Soomro <i>et al.</i> [41]	43.9%	Varol <i>et al.</i> [47]	62.3%	Varol <i>et al.</i> [47]	63.2%	Zhu <i>et al.</i> [60]	61.4%	Zhu <i>et al.</i> [60]	54.0%
Cai <i>et al.</i> [1]	83.5%			Oneata <i>et al.</i> [31]	67.2%	Vig <i>et al.</i> [48]	61.9%	Oneata <i>et al.</i> [30]	54.8%
Wu <i>et al.</i> [55]	84.2%			Jain <i>et al.</i> [14]	62.5%	Wang <i>et al.</i> [51]	57.2%	Wang <i>et al.</i> [51]	57.2%
Wang <i>et al.</i> [52]	85.9%			Oneata <i>et al.</i> [30]	63.3%	Peng <i>et al.</i> [32]	59.8%	Peng <i>et al.</i> [32]	59.8%
Peng <i>et al.</i> [32]	87.7%			Wang <i>et al.</i> [51]	64.3%				
								Peng <i>et al.</i> [33]	66.8%
Objects	65.6%		49.7%		44.7%		38.4%		38.9%
Motion	84.2%		56.9%		63.1%		64.6%		57.9%
Objects + Motion	88.1%		66.8%		70.8%		66.2%		61.1%

Objects combined with motion is powerful
 Complementary to other advances [Peng *et al.*, ECCV14]
 State-of-the-art on several datasets

Outline



Supervised action recognition



Unsupervised action recognition

Contribution

Objects2action, a semantic word embedding spanned by a skip-gram model of thousands of object categories. Recognizes actions without the need for video examples.



Mihir Jain



Jan van Gemert



Thomas Mensink

Objects2action: Classifying and localizing actions without any video example.
Mihir Jain, Jan van Gemert, Thomas Mensink, and Cees Snoek. Submitted.

Lampert et al PAMI 2013,
and many others

Zero-shot recognition practice

Classify test videos by (predefined) mutual relationship using class-to-attribute mappings



Problems of attributes

Attributes are difficult to define and annotate

Demands hold-out action train classes a priori to guide the knowledge transfer

Our action recognition does not need any video data nor action annotation as prior knowledge

Objects2action

Simple convex combination of known classifiers

$$\mathcal{C}(v) = \underset{z}{\operatorname{argmax}} \sum_y p_{vy} g_{yz}$$

Test video
Object representation
Object/action affinities

$$g_{yz} = s(y)^T s(z),$$

where $s() = \text{word2vec}$

Mikolov et al NIPS 2013

Average vs Fisher Word Vectors

Objects and actions may come as multiple words

FieldHockeyPenalty \rightarrow "FieldHockeyPenalty Field Hockey Penalty"

Default is to **average** word vectors, simply ignore relations

$$s_A(c) = \frac{1}{|w|} \sum_{w \in c} s(w).$$

We introduce the **Fisher Word Vector**

model distribution over words, as a sort of topic model

$$s_F(c) = [\mathcal{G}_{\mu_1}^c, \mathcal{G}_{\sigma_1}^c, \dots, \mathcal{G}_{\mu_k}^c, \mathcal{G}_{\sigma_k}^c]^T.$$

Sparsity per action and per video

Not all objects contribute to specific actions

Cat seems unlikely to be relevant for kayaking

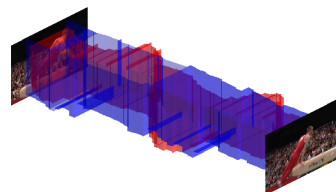
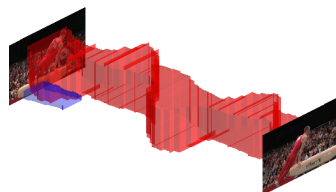
We consider two sparsity metrics

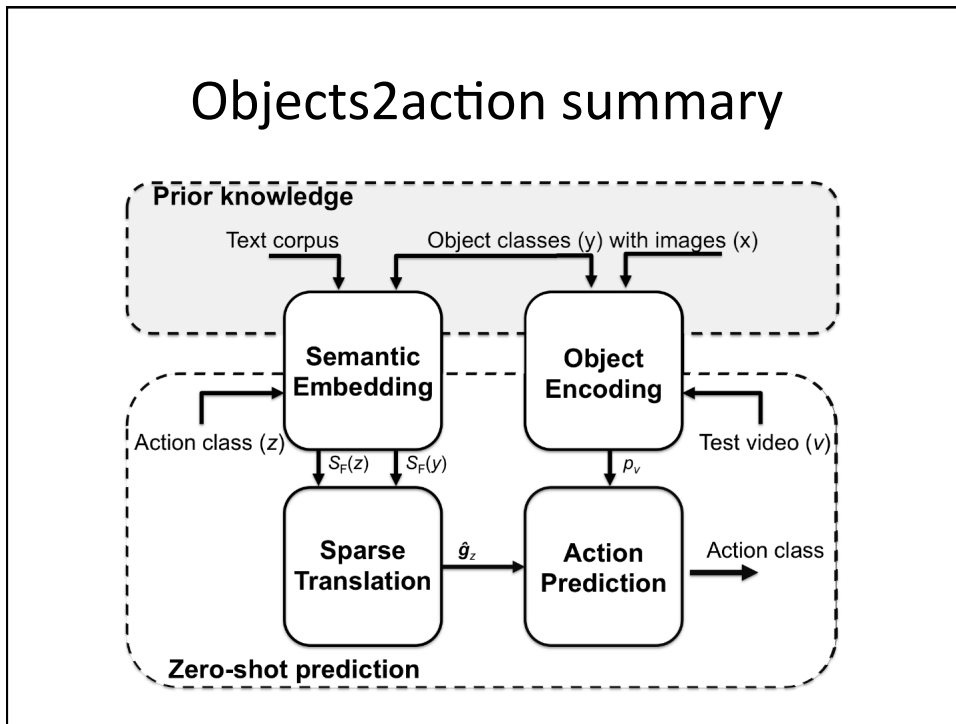
Selecting most responsive objects to a given action

Selecting most responsive objects to a given video

Zero-shot action localization

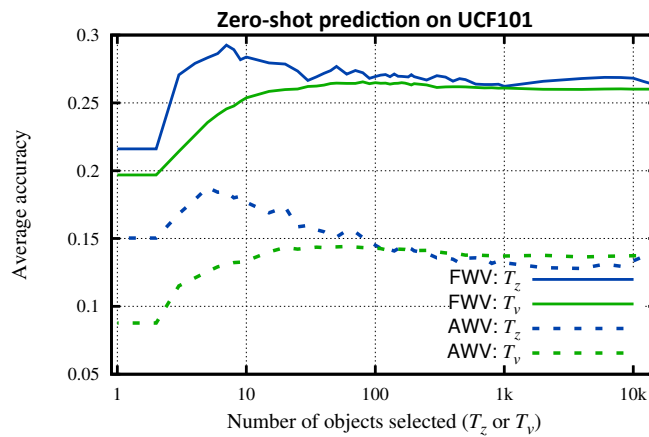
1. Generate several action tube proposals [\[Jain et al, CVPR14\]](#)
2. Encode tubes with objects
3. Zero-shot prediction for all tubes, select best one
4. Compute AUC for various overlap thresholds








EXPERIMENTS

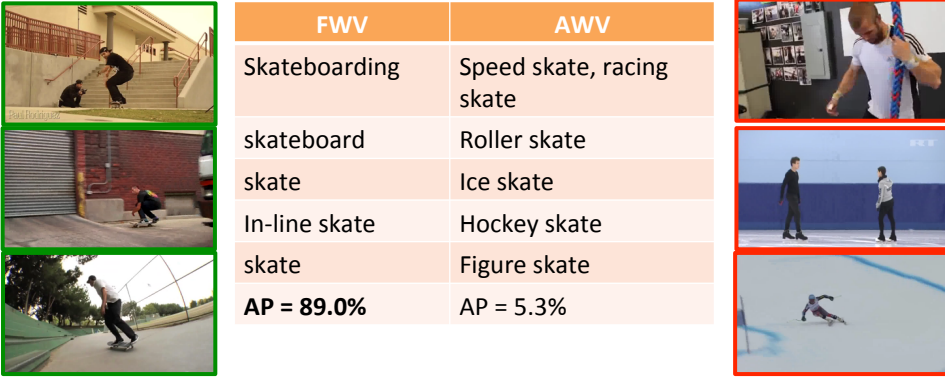
Word aggregation and sparsity







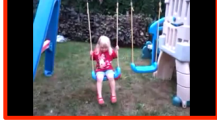
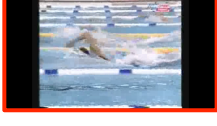

Fisher word vector much better than averaging
Selecting most prominent objects per action suffices

Results for *Skate boarding*

	FWV	AWV
	Skateboarding	Speed skate, racing skate
	skateboard	Roller skate
	skate	Ice skate
	In-line skate	Hockey skate
	skate	Figure skate
	AP = 89.0%	AP = 5.3%



Results for *Salsa spin*

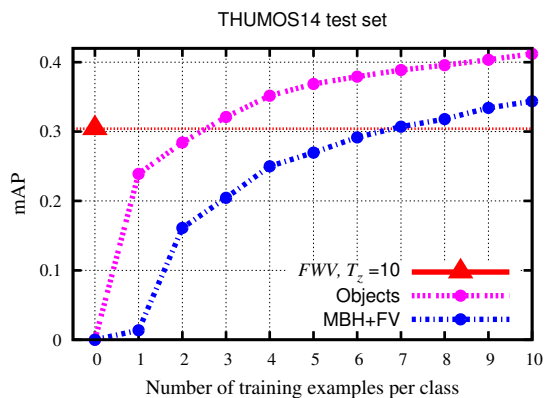
	FWV	AWV	
	Salsa	Spin dryer, spin drier	
	Spin dryer, spin drier	Spinning rod	
	Dancing-master, dance master	Chili sauce	
	guacamole	Spinning wheel	
	swing	Kick starter, kick start	
	AP = 22.0%	AP = 0.8%	

Object2action baselines

<i>Embedding</i>	<i>Sparsity</i>	<i>UCF101</i>	<i>HMDB51</i>	<i>THUMOS14</i>	<i>UCF Sports</i>
AWV	None	13.7%	8.0%	3.4%	13.9%
	Video	14.3%	7.7%	10.0%	13.9%
	Action	17.7%	9.9%	16.5%	28.1%
FWV	None	26.0%	14.2%	22.9%	23.1%
	Video	26.5%	14.5%	25.0%	23.1%
	Action	28.4%	15.5%	30.4%	28.9%
<i>Supervised</i>		63.9%	35.1%	56.3%	60.7%

Not competitive with supervised alternative, but promising

Objects2action vs few-shot learning



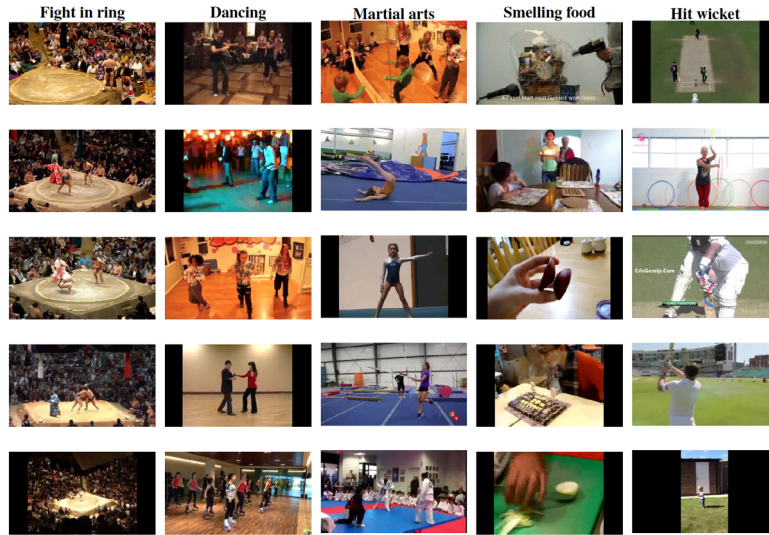
Object representation more effective for few-shot
Object2action best for less than three examples

Object transfer versus action transfer

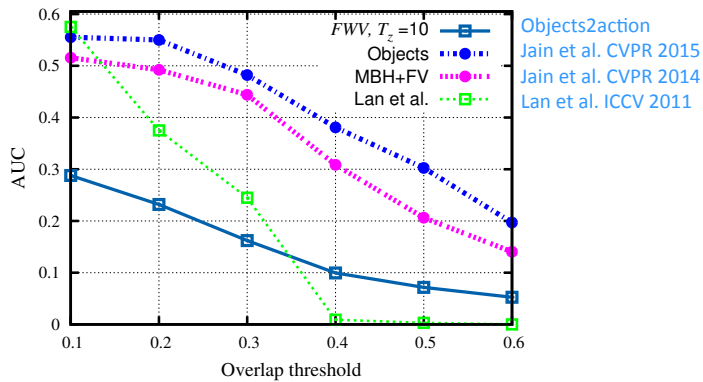
Method	Train	Test	UCF101	HMDB51
Action attributes	Even	Odd	16.2%	—
	Odd	Even	14.6%	—
Action labels	Even	Odd	15.4%	12.8%
	Odd	Even	15.9%	13.9%
Objects2action	ImageNet	Odd	35.2%	16.2%
		Even	38.7%	24.2%

Objects2action much better than alternative transfers

Never seen action in THUMOS



Zero-shot action localization



**Competitive with supervised alternative from 2011
(for high-overlap threshold)**

Conclusion

Objects matter for actions

Actions have object preference, relation is generic

Facilitates recognition without video and action examples

www.ceessnoek.info

Thank you

dr. Cees Snoek



www.ceessnoek.info



cgmsnoek@uva.nl



twitter.com/cgmsnoek