Event Recognition by Learning

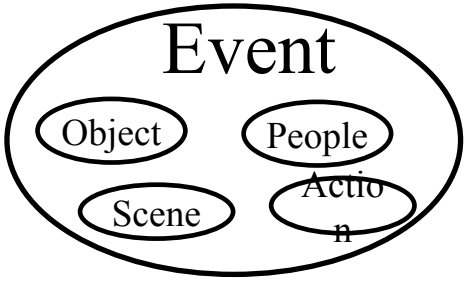
Amir Habibian

Qualcomm Research, Amsterdam

27 Feb 2017

What is an event?

Interaction of **people** and **objects** under a certain **scene**



Examples

- **Personal events:** *marriage proposal, grooming an animal*
- **Traffic events:** *accident, traffic jam*
- **Security events:** *breaking a lock, leaving a bag unattended*

Event: Winning a race without a vehicle

Why event recognition is hard?

Large variation in examples (semantic variance)

- Depending on the context, may involve various objects, actions and scenes

Event: Feeding an animal

Limited number of training examples

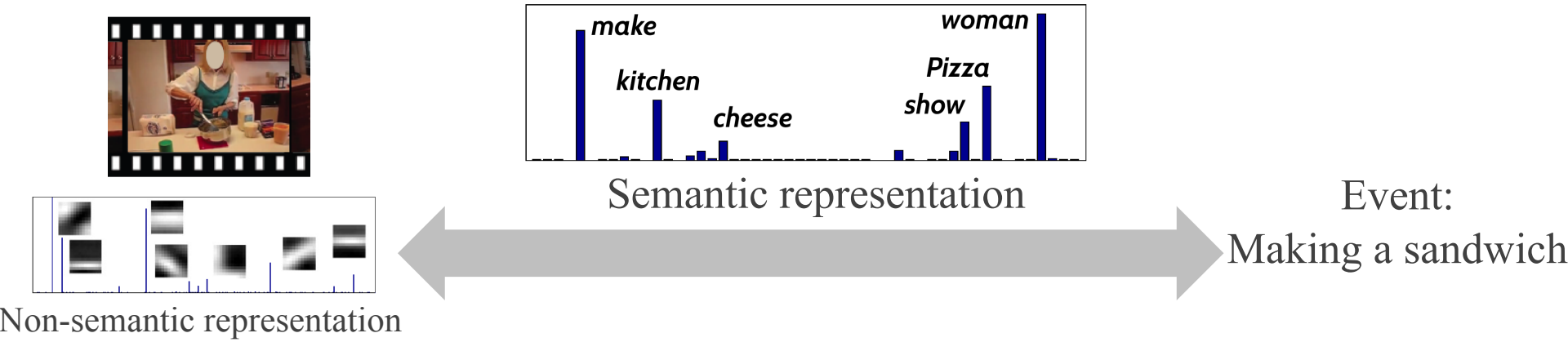
- More specific than individual object, action, and scenes

Video representations for event recognition

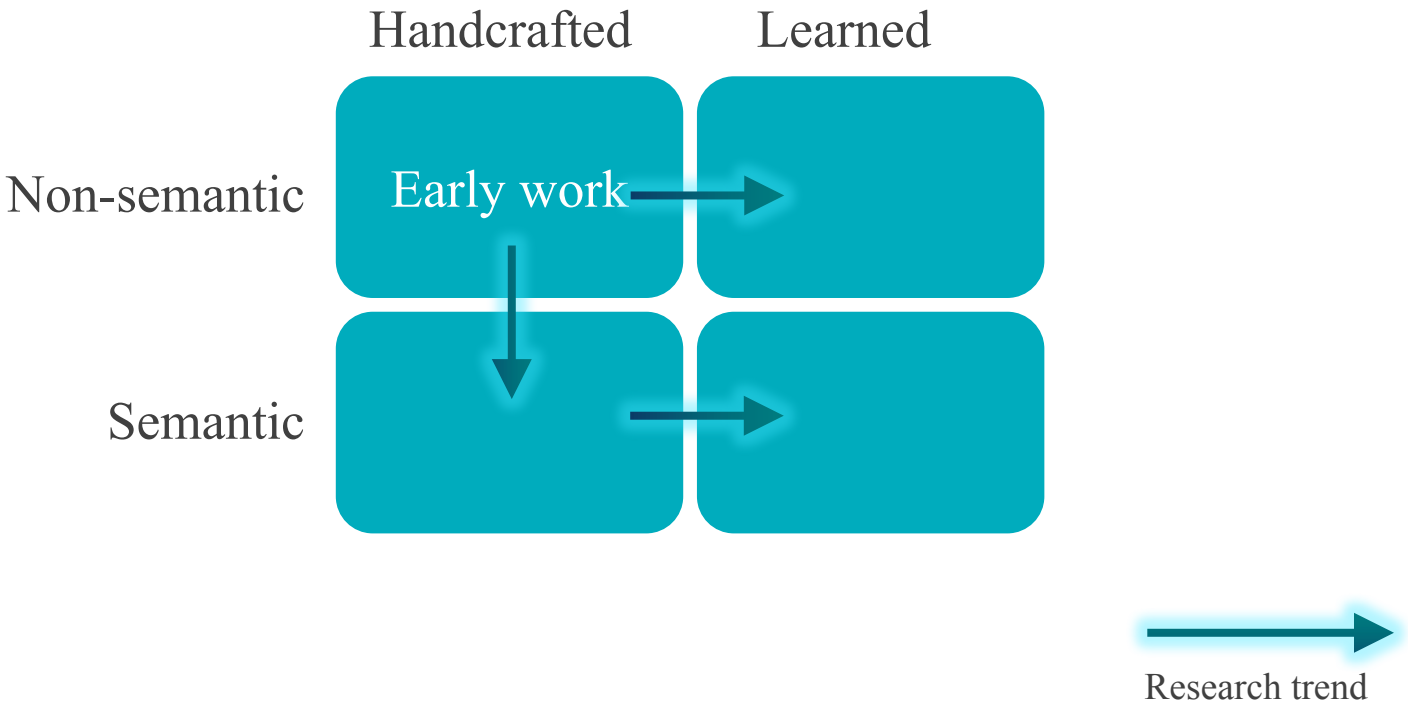
Neither shallow BoW nor deep learned representations fit well

- BoW are not discriminative enough to handle the large variations
- Not enough training examples to train a deep neural network

SOTA rely on pre-trained semantic encoders to represent videos

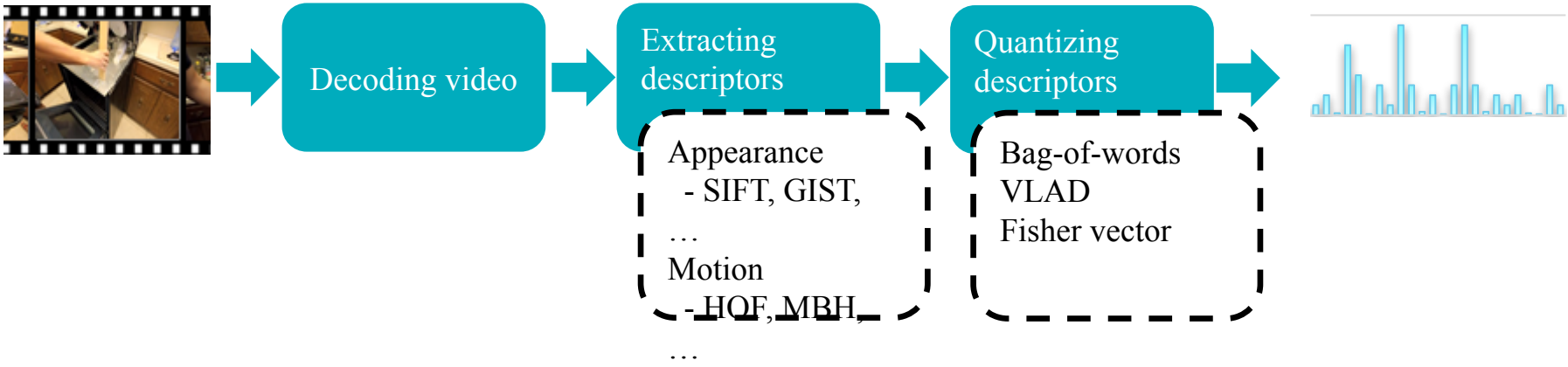


Video representations for event recognition



Non-semantic representation (handcrafted)

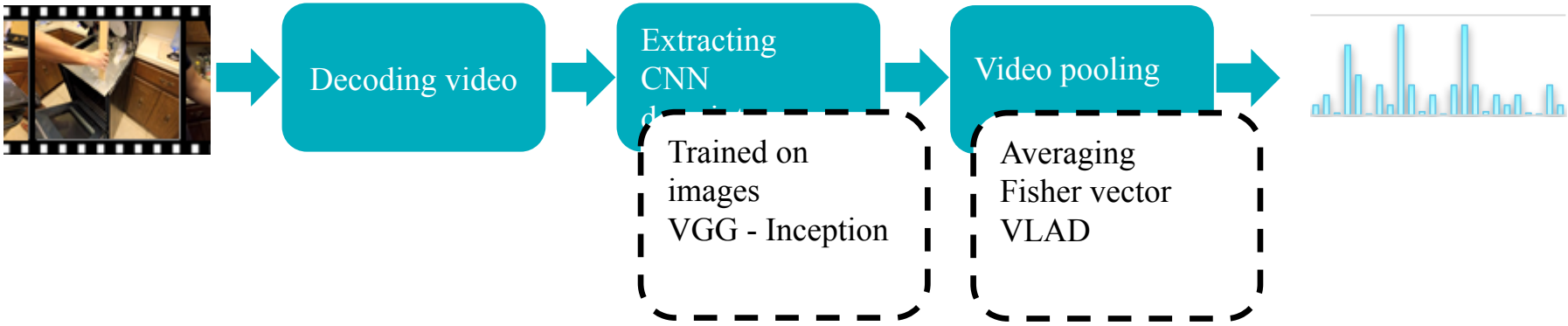
Aggregation of handcrafted descriptors over video



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

Non-semantic representation (learned)

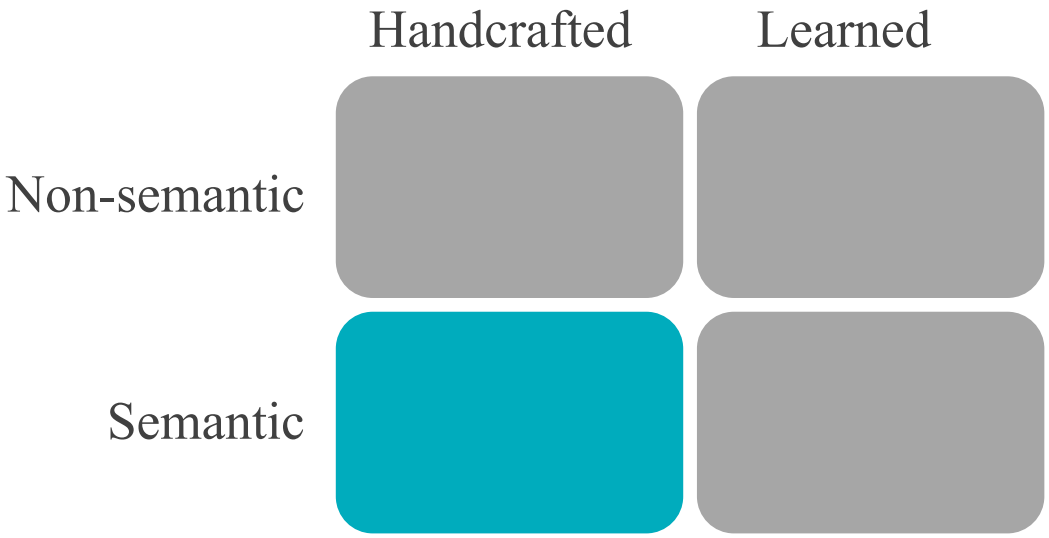
Aggregation of CNN descriptors over video



More effective and efficient compared to the handcrafted

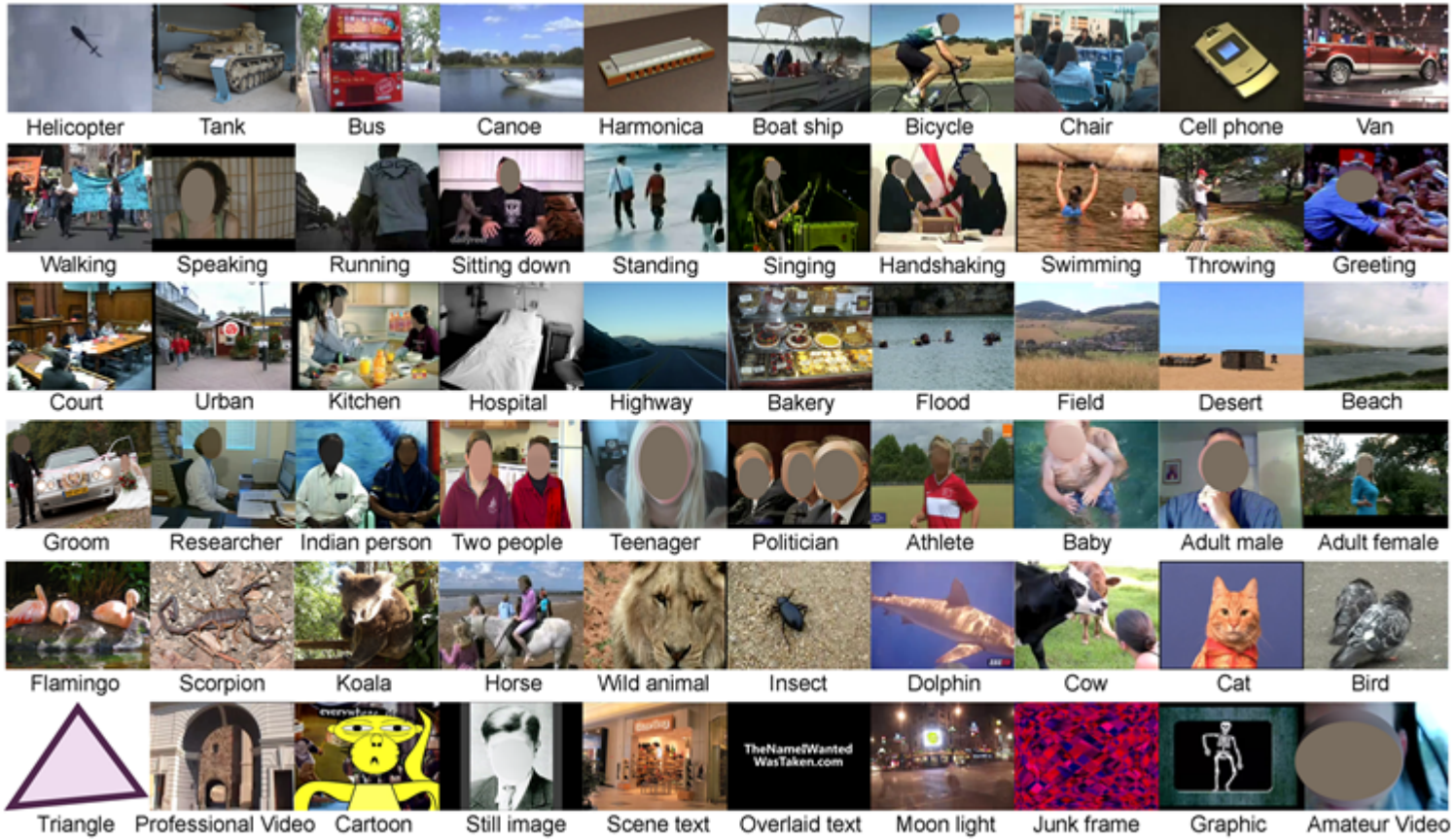
[Xu et al., CVPR 2015] [Nagel et al., BMVC 2015]

Video representations for event recognition



Semantic representation (handcrafted)

Handcraft a vocabulary of concept detectors



Handcrafting concept vocabulary

The vocabulary is created in three steps:

1. Identifying the concepts to be included in the vocabulary
2. Providing training examples per concept
3. Training concept classifiers

Involves lots of annotation effort

- To identify which concepts to include
- To provide training examples per concept

Handcrafted vocabulary

Key questions

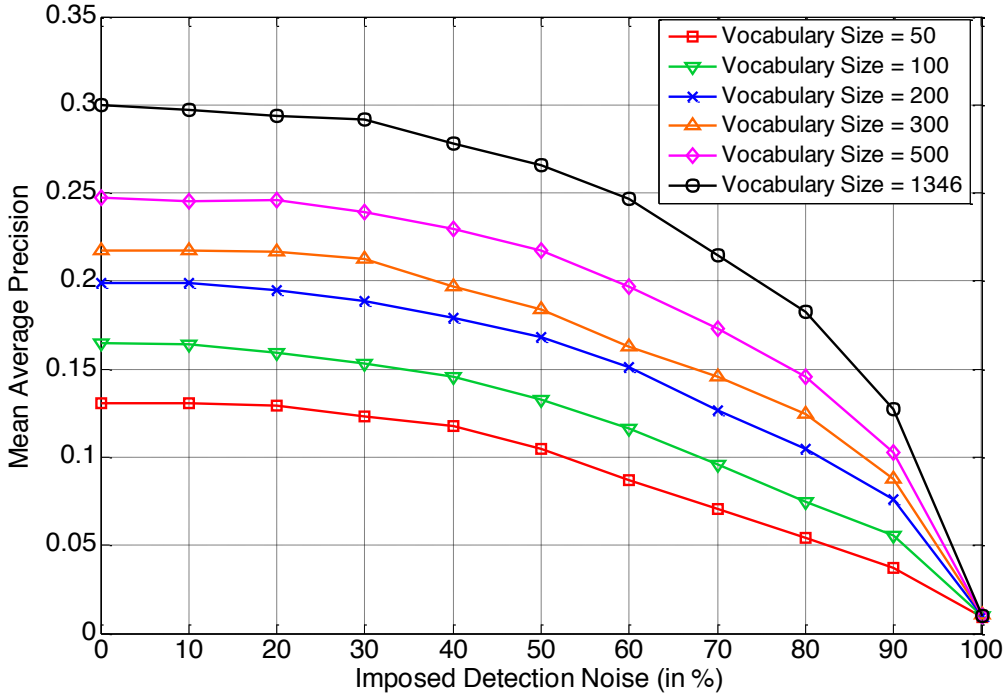
- How many concepts to include in the vocabulary?
- How accurate should the concept detectors be?
- What concept types to include in the vocabulary?
- Which concepts to include in the vocabulary?
- ...

A. Habibian, K. van de Sande, and C. Snoek, ICMR
A. Habibian and C. Snoek, CVIU'14

Quantity vs Quality

Impact of concept detector accuracies on event recognition

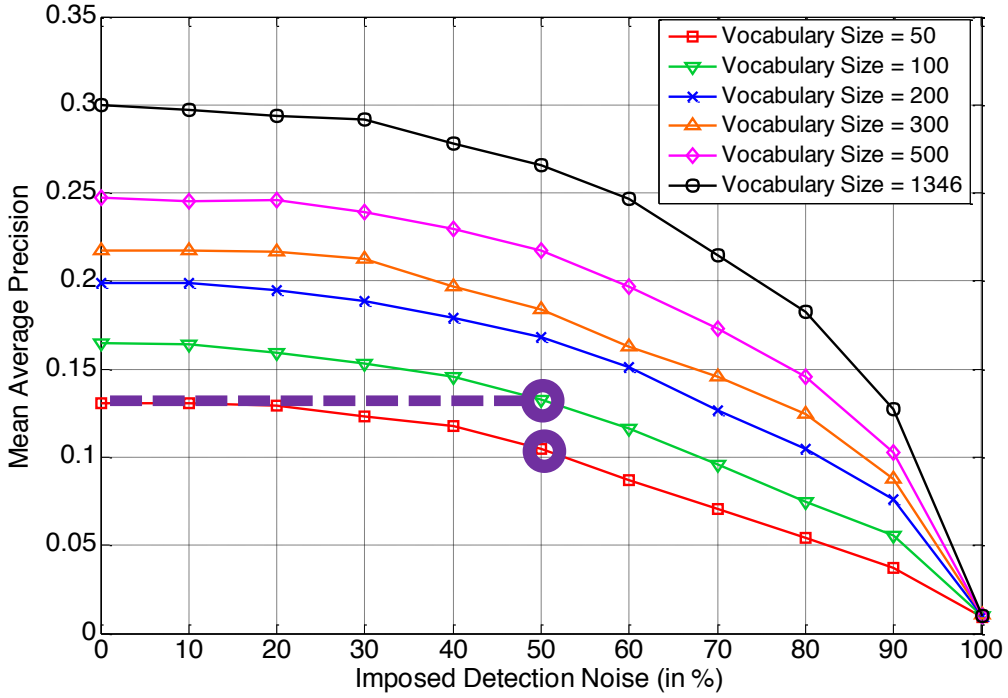
Impose noise on concept detector predictions



Quantity vs Quality

Impact of concept detector accuracies on event recognition

Impose noise on concept detector predictions



Make the vocabulary larger rather than more accurate

Conclusion

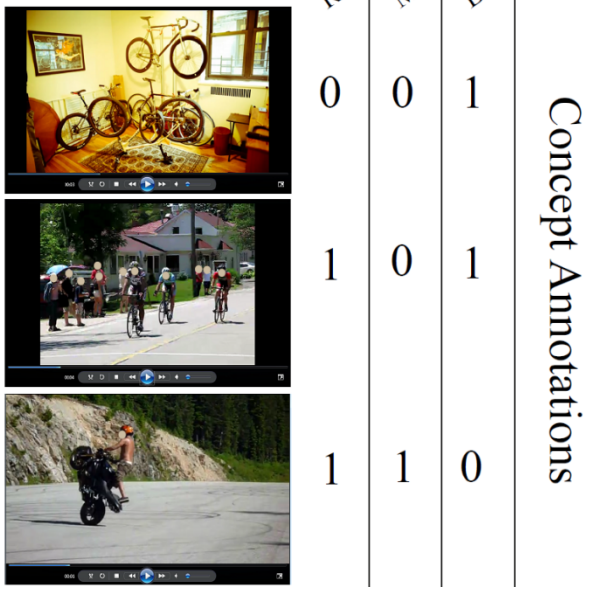
Comprehensive set of concepts from various types are needed

It requires lots of annotation effort …

Label composition trick

Expanding the labels by logical operations

- AND, OR, …



A. Habibian, T. Mensink, and C. Snoek, ICMR

Label composition trick

Expanding the labels by logical operations

- AND, OR, …

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

Qualitative results

Top ranked videos for *flash mob gathering*

Most dominant concepts in the video representation

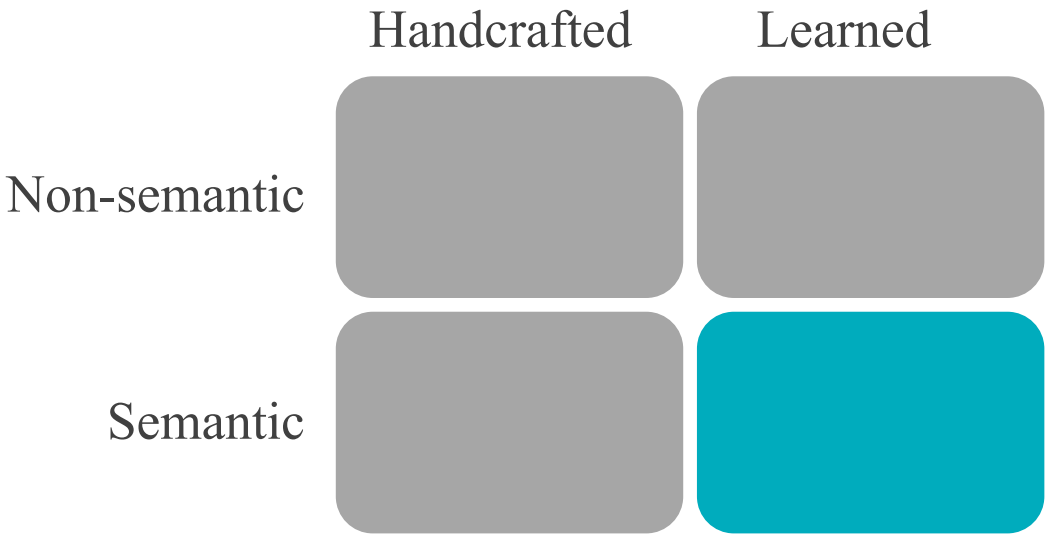
Detected Videos	Composite Concepts
	Group-AND-Dance-AND-Shopping Celebrating-OR-Marching Performance-OR-Music People-OR-Girl Surprise-OR-Party
	Group-AND-Dance-AND-Shopping Band-OR-Singining Inside-OR-School Performance-OR-Music Surprise-OR-Party
	Group-AND-Dance -AND-Shopping Practice-OR-Gym Living-AND-Room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

Conclusion

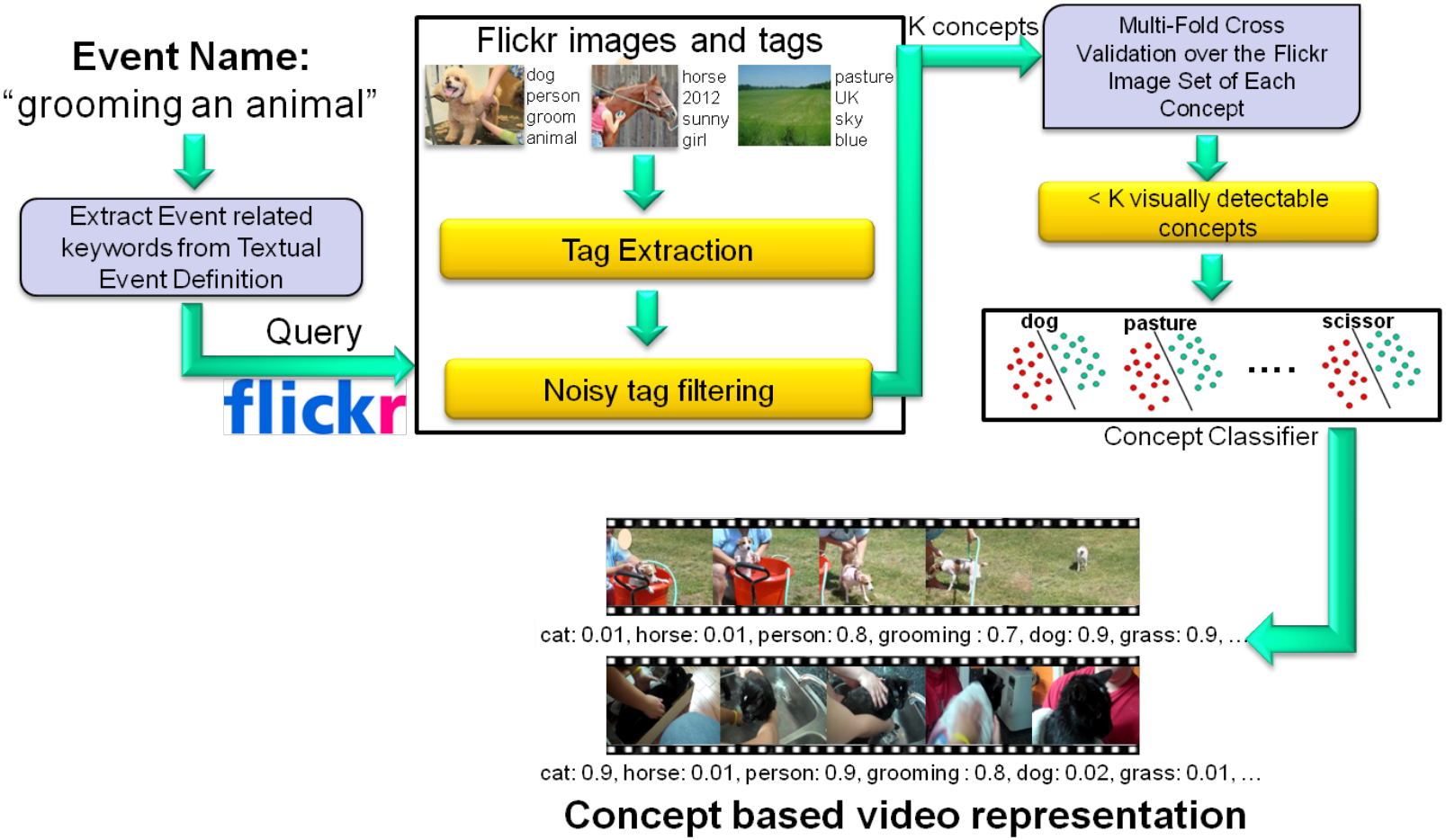
More comprehensive vocabulary by composing the concepts

Still grounded on the handcrafted concepts …

Video representations for event recognition



Discovering concepts from the web

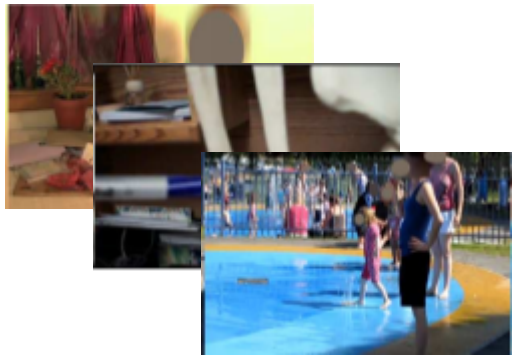


[Wu et. al. CVPR'14] [Chen et al., ICMR

Video2Vec embedding

Learn the mutual underlying subspace between videos and descriptions

Videos



...

Descriptions

A woman folds and packages a scarf she has made.

A woman points out bones on a skeleton for lab practical for an anatomy class.

A mother at a fountain tries to get her daughter to step on the water jets.

...

Semantic space

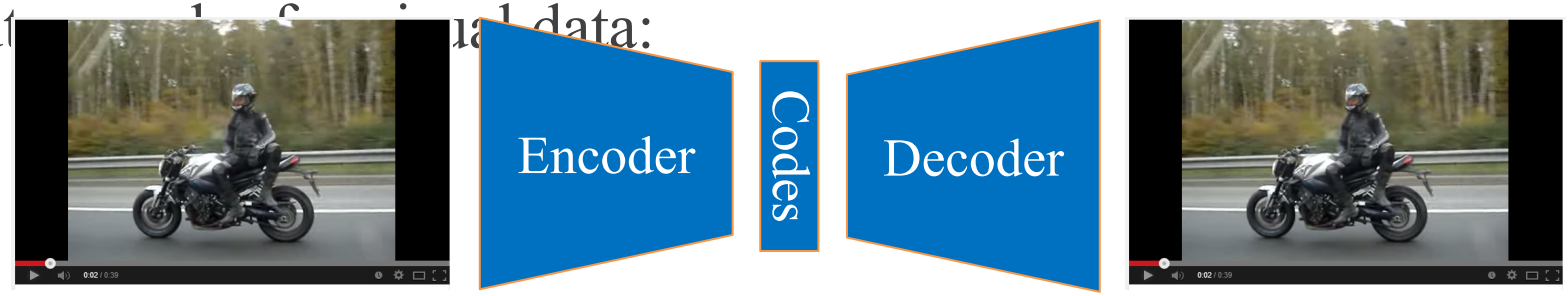
A. Habibian, T. Mensink, and C. Snoek, PAMI, In press

Autoencoder

Learn a compact representation by which the input could be reconstructed

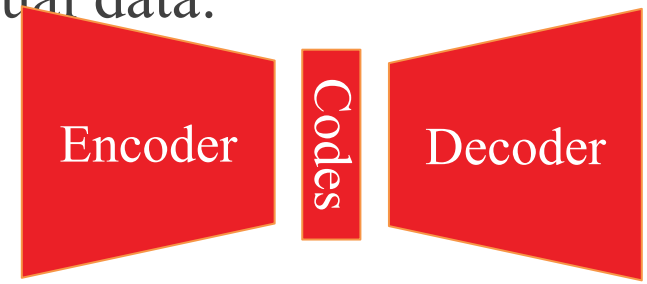
- Codes as data representation

Aut[illegible] data:



Autoencoder for textual data:

Crazy guy doing insane stunts on bike.

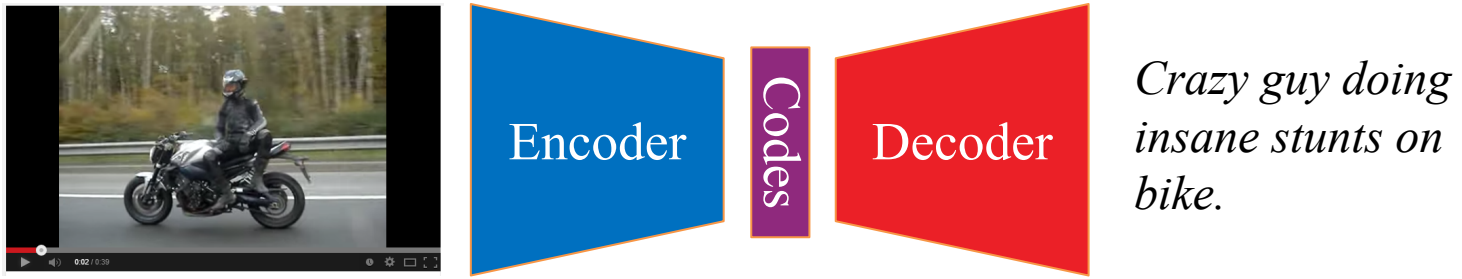


Crazy guy doing insane stunts on bike.

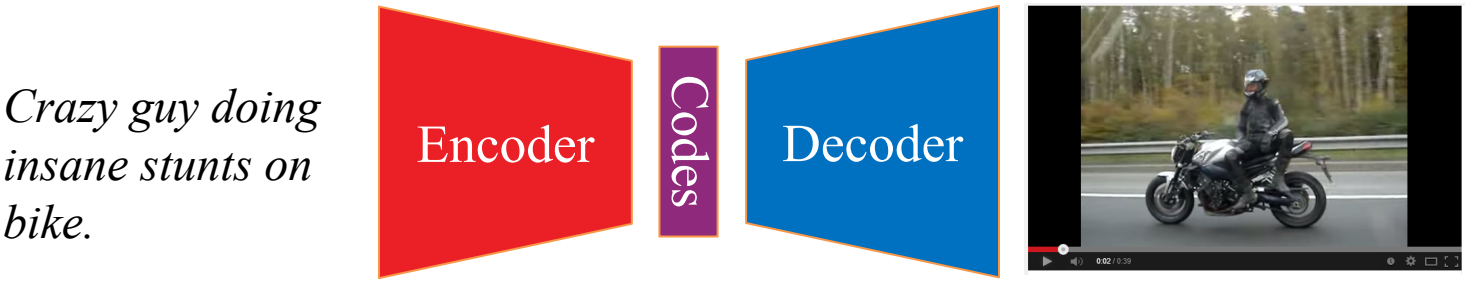
Video2Vec embedding

Reconstruct the other view of data

- Reconstruct the textual view from visual view



- Reconstruct the visual view from textual view:



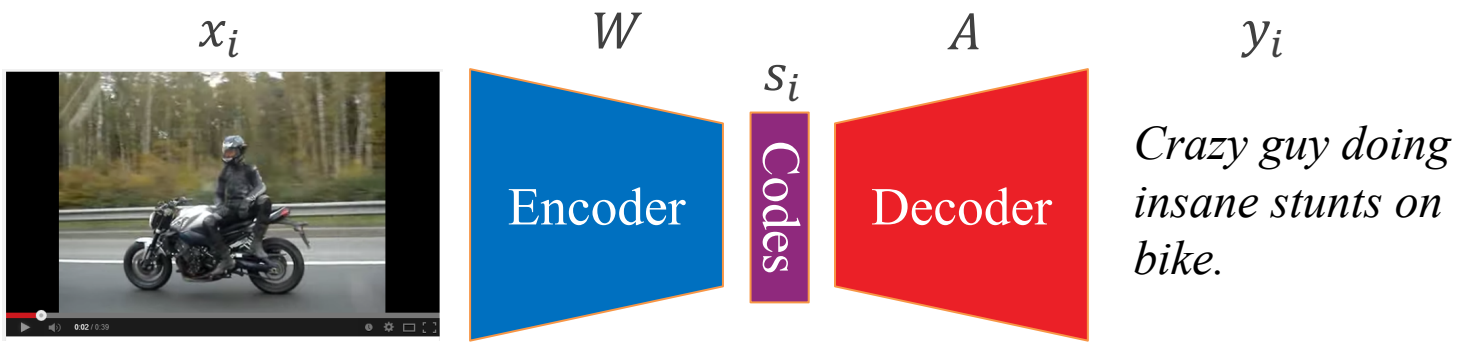
Video2Vec embedding

Reconstruct the other view of data

- Reconstruct the textual view from visual view

$$\mathcal{L}(y, \bar{y}) = \|y_i - A\,W\,x_i\|^2$$

- W: encodes visual features into codes
- A: decodes codes into textual features



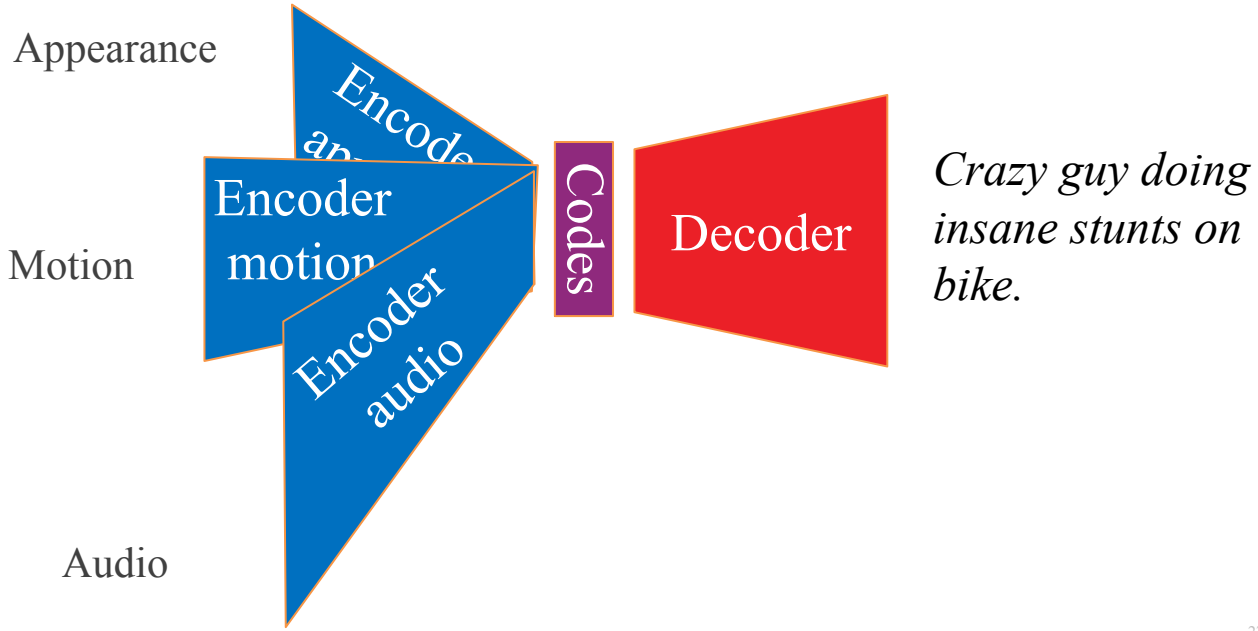
Multimodal encoding

Train a different encoder to encode every video channel

- Appearance, Motion, and audio

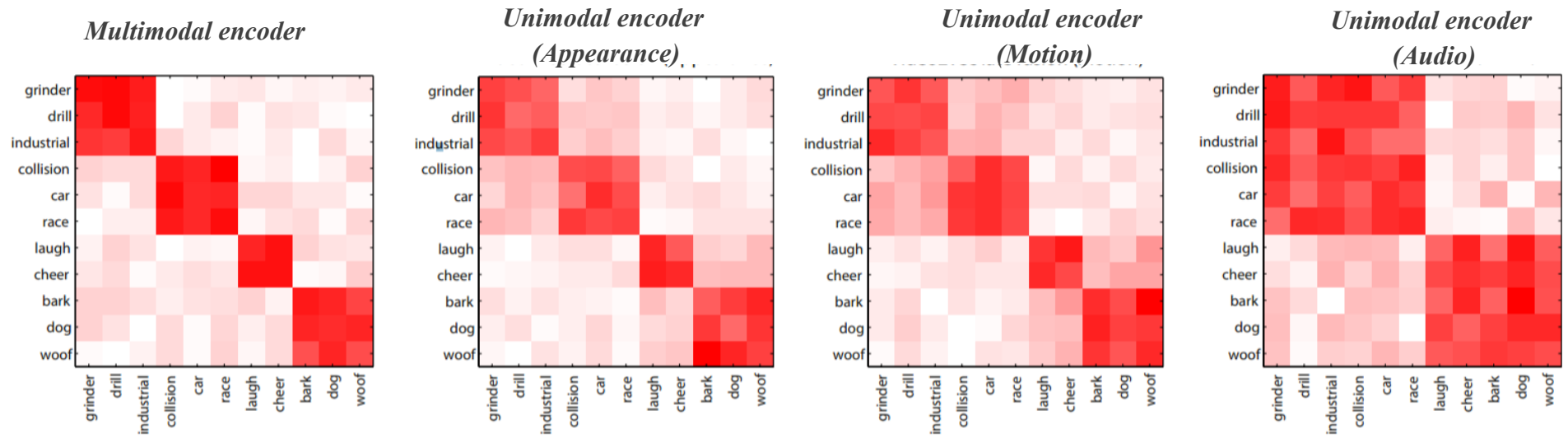
Share the codes to enforce the common structures across modalities

- Acts as a regularizer



Multimodal encoding

Visualizing the decoder (A) as A x A^T



The multimodal encoder better learns the semantic relations

Impact of multimodal encoding

Joint encoding of multiple modalities lead to a better representation

MED 2013			
Event	Appearance	+ Motion	+ Audio
Birthday party	37.1	38.8	**43.4**
Changing a vehicle tire	64.7	65.5	**67.2**
Flash mob gathering	55.3	63.8	**65.0**
Getting a vehicle unstuck	59.3	65.2	**65.3**
Grooming an animal	24.3	**29.4**	28.2
Making a sandwich	16.7	18.9	**21.3**
Parade	33.9	**44.7**	44.0
Parkour	61.3	**72.8**	72.2
Repairing an appliance	44.5	49.7	**57.4**
Working on a sewing project	47.2	**48.0**	46.0
Attempting a bike trick	8.8	**9.3**	8.9
Cleaning an appliance	10.5	13.2	**15.5**
Dog show	81.6	84.9	**85.9**
Giving directions to a location	0.6	**1.0**	0.9
Marriage proposal	0.3	0.4	**0.5**
Renovating a home	11.4	12.1	**13.8**
Rock climbing	13.7	**14.9**	14.4
Town hall meeting	40.2	38.9	**44.8**
Winning a race without a vehicle	21.9	**28.8**	27.8
Working on a metal crafts	15.2	18.0	**20.4**
mAP	32.4	35.9	**37.1**

Task specific decoding

Autoencoders rely on ℓ_2 loss to measure reconstruction error:

$$\mathcal{L}(y, \bar{y}) = \|y - \bar{y}\|^2$$

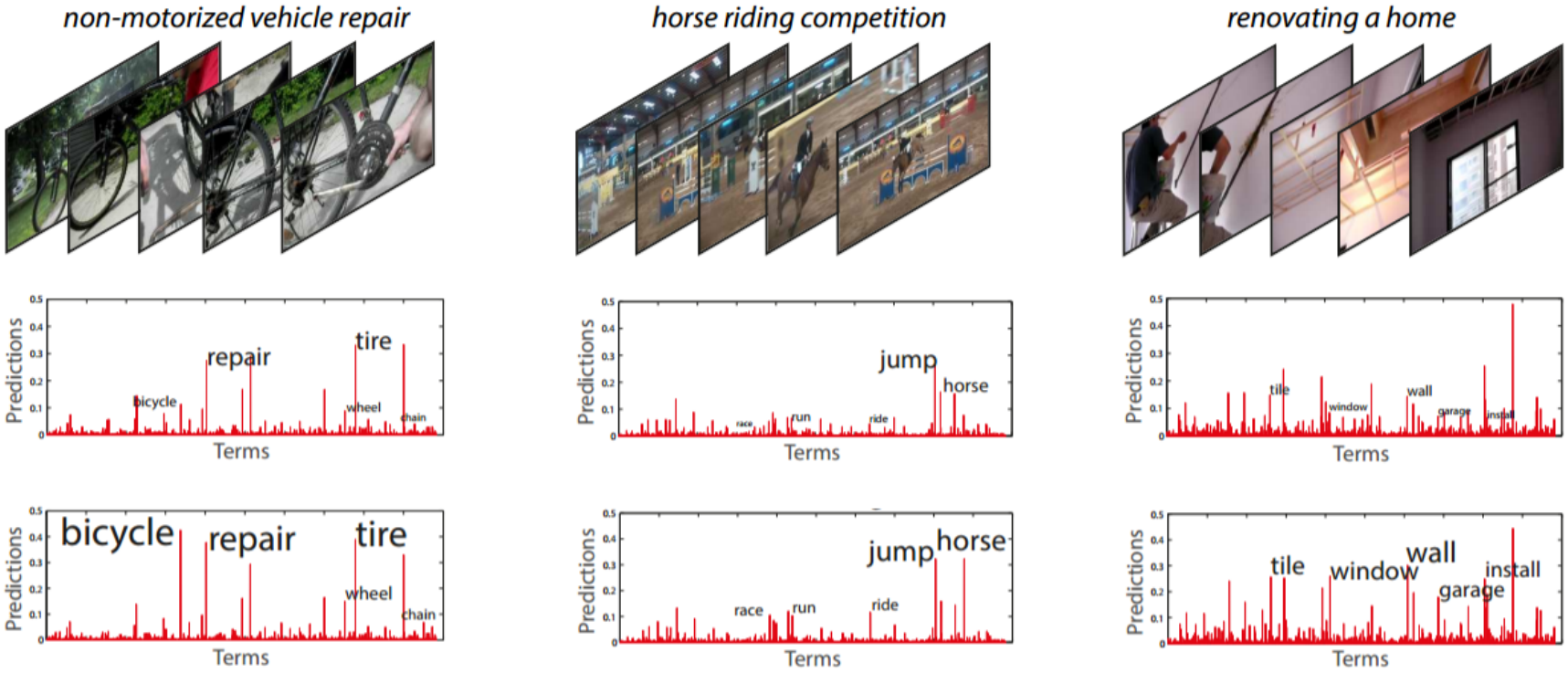
The error in reconstructing all of the words are treated equally

We replace the ℓ_2 loss with:

$$\mathcal{L}(y, \bar{y}) = \| H_t (y - \bar{y})\|^2$$

H_t is a diagonal matrix determining the importance of each word per task

Task specific decoding



Middle: standard decoder

Bottom: task specific decoder

Impact of event specific decoding

Event specific decoding lead to a better representation

- For the both unimodal and multimodal encoders

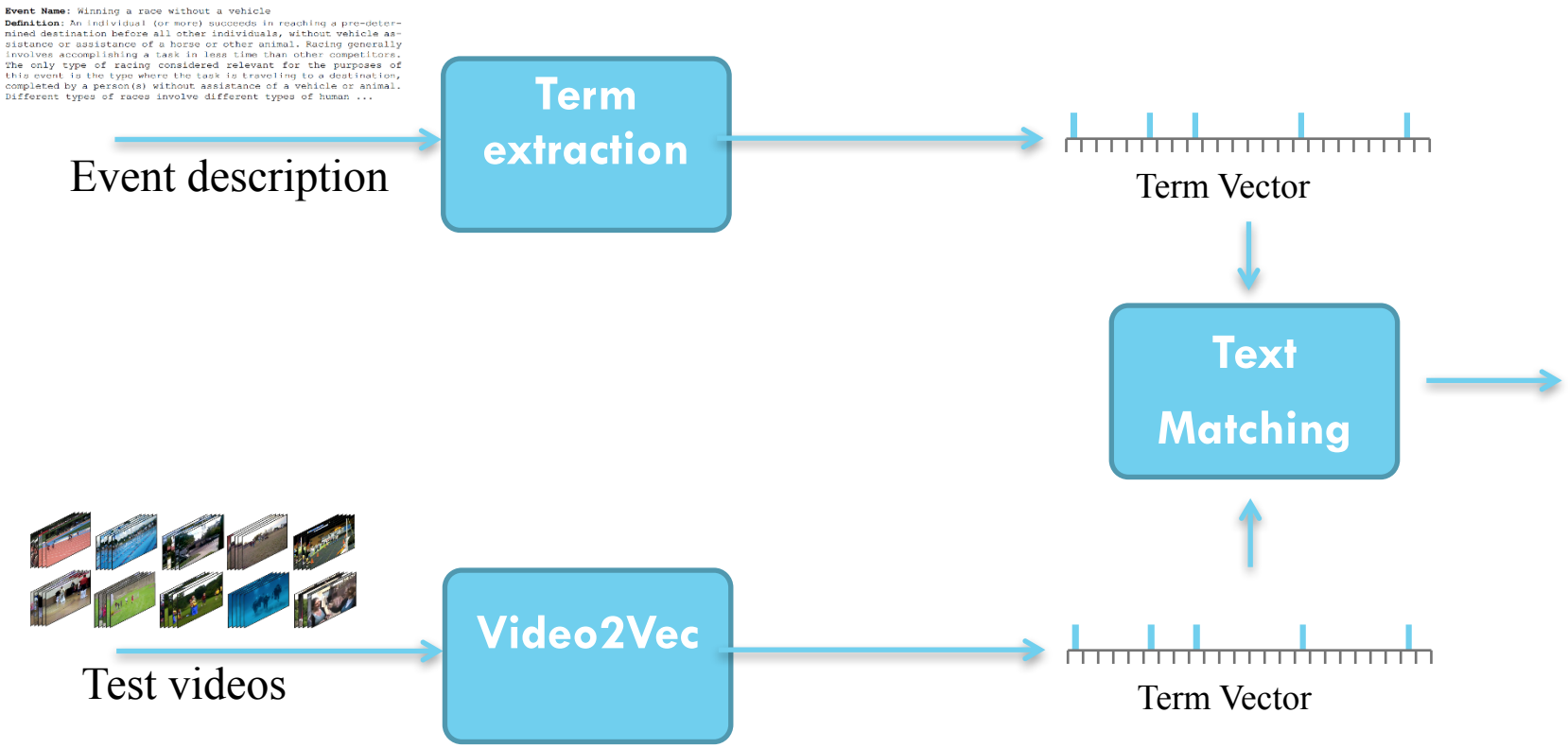
MED 2013				
Event	Video2vec	Video2vec$^{\mathcal{F}}$	Video2vec$_{\mathcal{E}}$	Video2vec$_{\mathcal{E}}^{\mathcal{F}}$
Birthday party	24.6	32.5	30.3	**37.4**
Changing a vehicle tire	**43.9**	36.0	42.1	38.2
Flash mob gathering	14.5	30.1	22.4	**33.8**
Getting a vehicle unstuck	**40.2**	29.9	36.0	32.2
Grooming an animal	18.7	21.4	**28.0**	23.4
Making a sandwich	19.4	15.5	**21.1**	17.1
Parade	17.6	32.5	26.4	**38.2**
Parkour	26.1	34.6	28.0	**40.3**
Repairing an appliance	39.8	42.4	41.4	**46.3**
Working on a sewing project	30.8	34.3	**36.4**	36.0
Attempting a bike trick	**8.8**	5.1	7.1	5.9
Cleaning an appliance	8.2	**12.6**	9.1	12.5
Dog show	4.0	5.8	5.8	**5.9**
Giving directions to a location	0.6	0.7	0.7	**0.8**
Marriage proposal	0.3	**0.8**	0.4	0.7
Renovating a home	5.2	5.3	**6.8**	6.4
Rock climbing	**1.6**	1.1	1.4	1.0
Town hall meeting	1.9	2.5	9.2	**9.6**
Winning a race without a vehicle	**9.4**	8.1	8.4	7.9
Working on a metal crafts	1.6	4.1	4.9	**7.0**
mAP	15.9	17.8	18.3	**20.0**

Zero-shot event recognition

Event recognition with video examples

1. Train the embedding on a collection of videos and their descriptions
   - Videos and their captions downloaded from YouTube
2. Use the trained embedding to encode event videos
3. Train and use the event classifier on the encoded representations
   - SVM

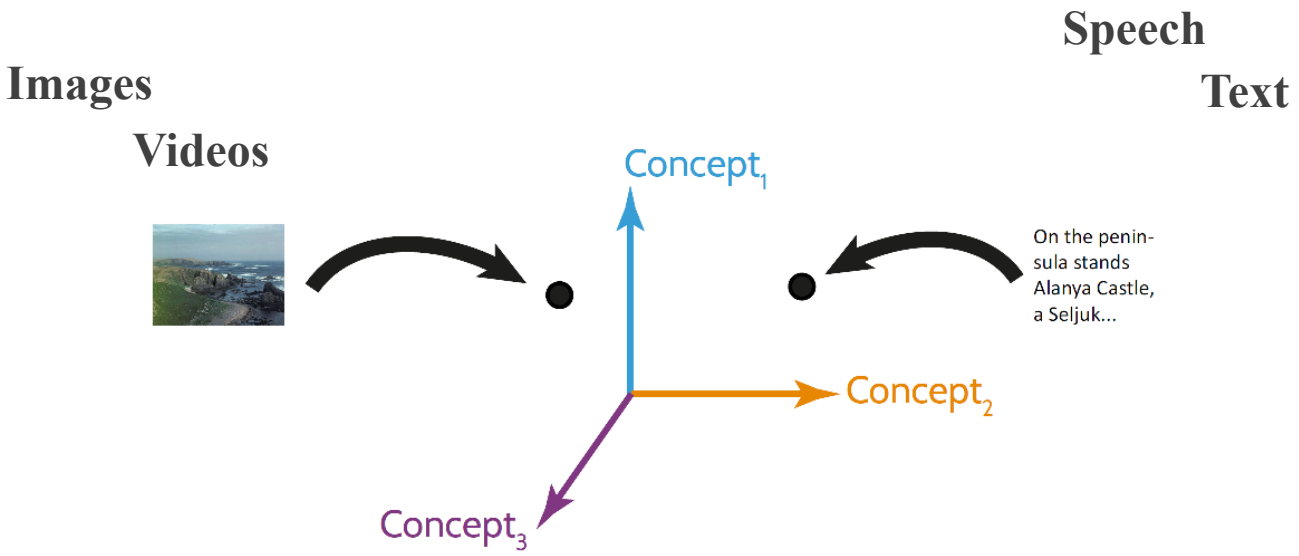
Event recognition without video examples



Applications

Application 1: Cross-modal retrieval

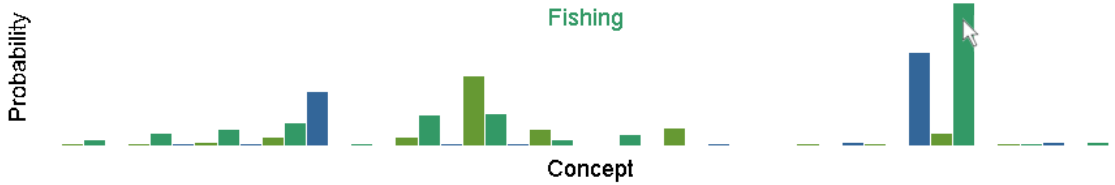
Represent the all modalities in a mutual semantic space



A. Habibian, T. Mensink, and C. Snoek, ICMR

Application 1: Cross-modal retrieval

Video Query:



Translations:

← 3 →

Two women take a boat ride to a mountain.

People on a boat go crabbing.

Two men fish on a boat.

two men are landing fish

Woman catches a salmon.

Fishing with sunlights reflecting from the water

A person captures a fish from the river.

a man is landing fish from the rocks

a girl and a boy are landing fish

diver catching fish under water

A. Habibian and C. Snoek, MM'

Application 1: Cross-modal retrieval

Text Query: Guys snowboarding, skiing and doing tricks

Translations: ← 1 →

A. Habibian and C. Snoek, MM'

Application 2: On-the-fly event search

Efficiency

- Representing videos by a compact set of concepts

Few exemplars

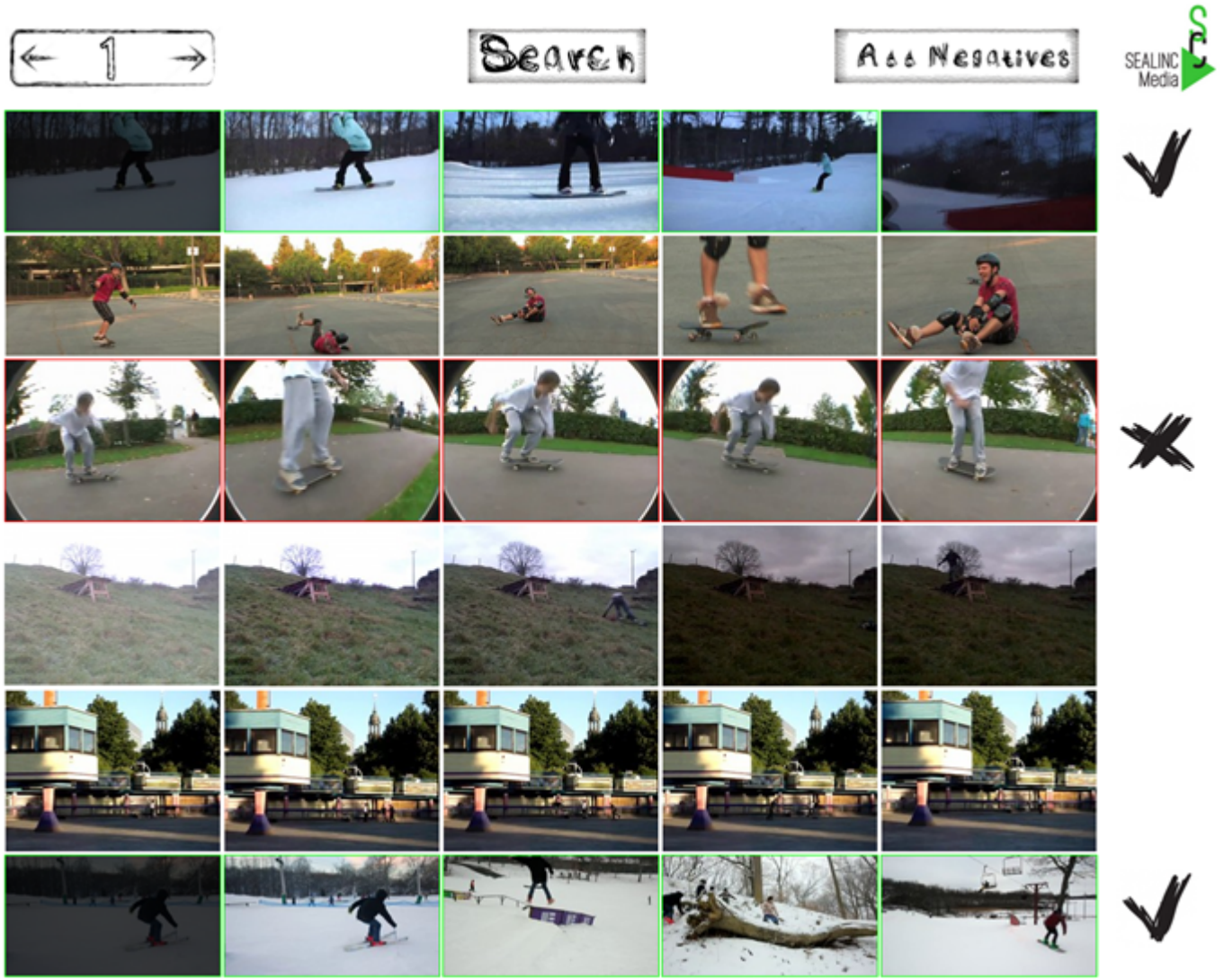
- Transfer learning from vocabulary training examples

Recounting

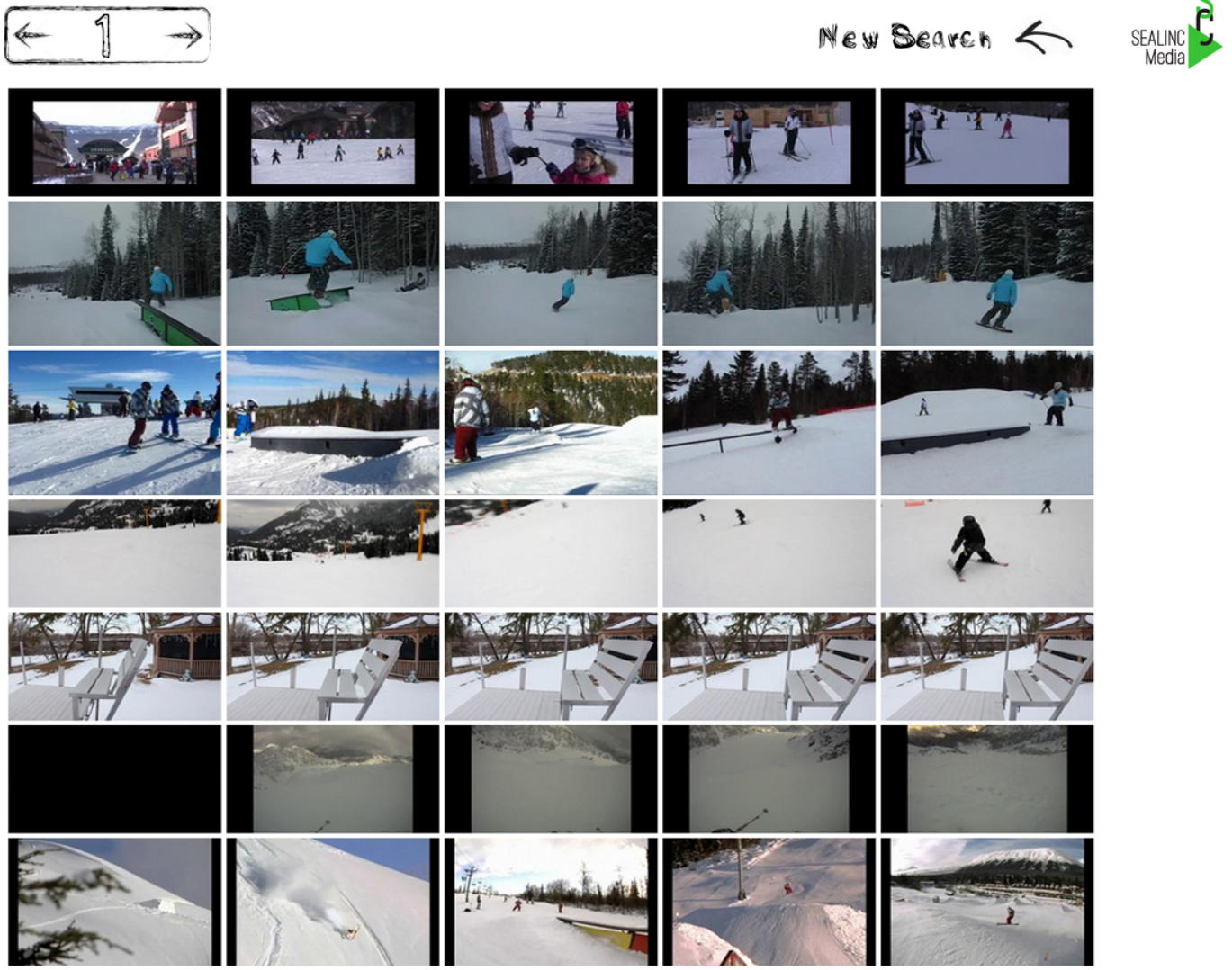
- Interpretable video representation

A. Habibian, M. Mazloom, and C. Snoek, ICMR
M. Mazloom, A. Habibian, and C.Snoek, MM'1

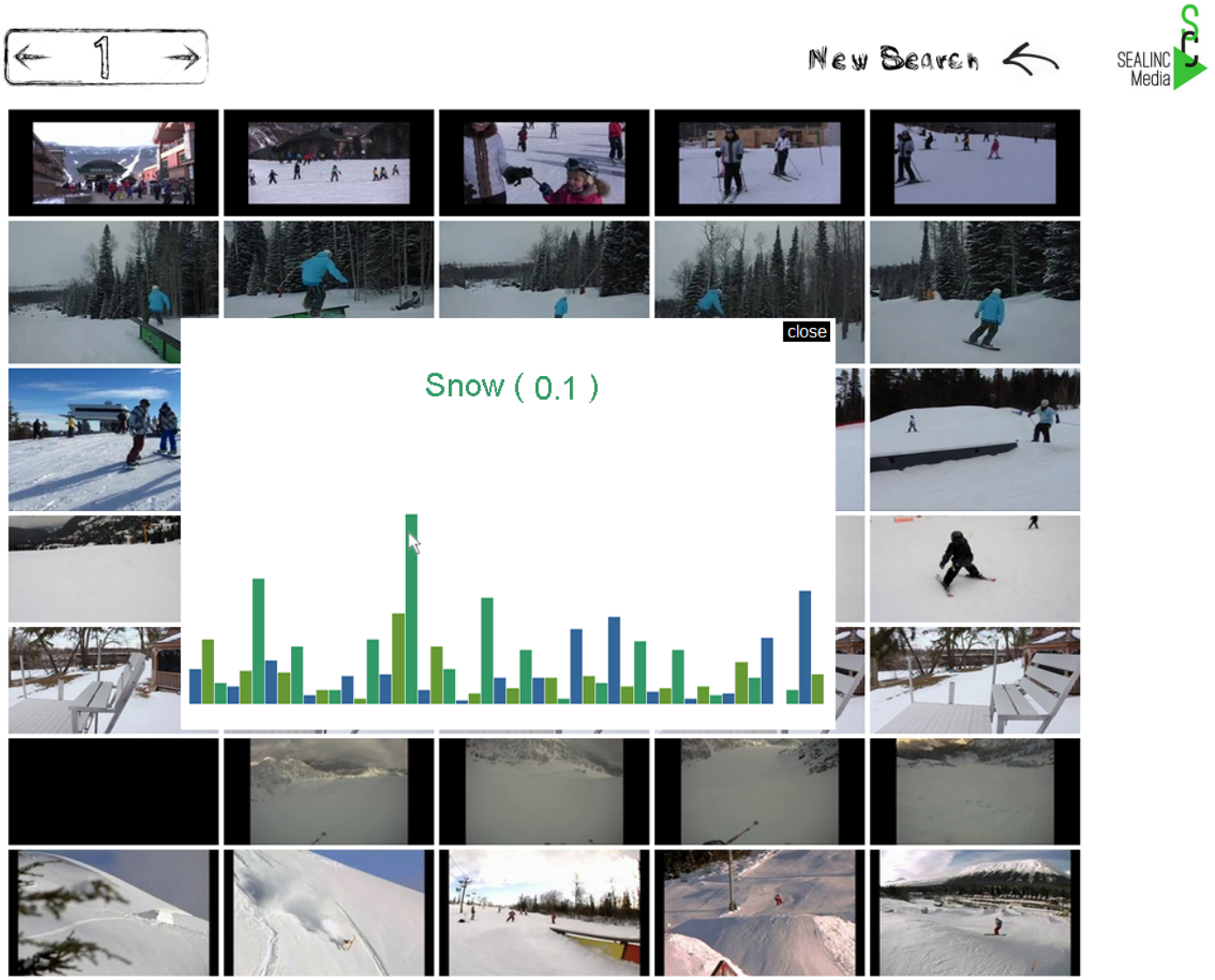
Application 2: On-the-fly event search



Application 2: On-the-fly event search

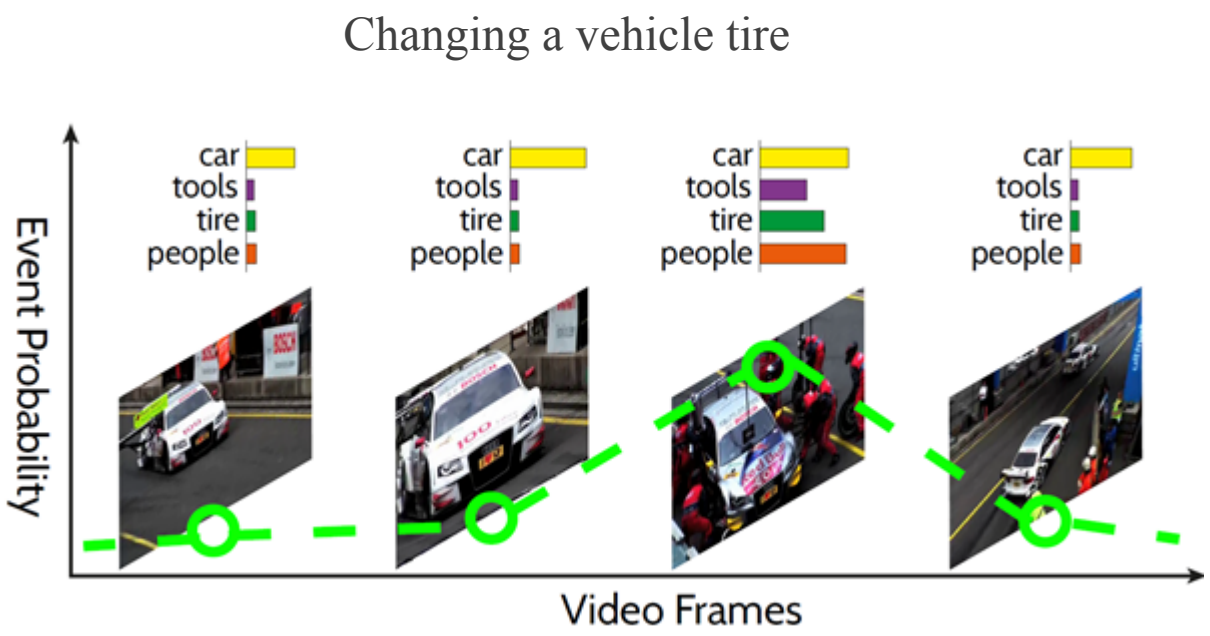


Application 2: On-the-fly event search



Application 3: Video summarization

Localizing the event over time by following its concepts
Summarizing long videos, *i.e. GoPro footages*



M.Mazloom, A. Habibian and C. Snoek, ICMR

Thanks !

habibian.a.h@gmail.com