Video Analysis

Video to Language , Highlight Detection, Video Classification

Tao Mei (tmei@microsoft.com)

Joint work with Ting Yao and Yong Rui, MSR Asia

Microsoft

Microsoft Project Oxford: Adding Smart to Your Applications

A portfolio of REST APIs and SDKs which enable developers to write applications which understand the content within the rapidly growing set of multimedia data

Easy to use

Project Oxford allows you to focus on your application by easily including these services across platforms through simple REST APIs

Microsoft Project Oxford Services

PROJECT OXFORD

Vision APIs

- Analyze Image
- OCR
- Generate Thumbnail

Face APIs

- Face Detection
- Face Grouping
- Face Identification

Speech APIs

- Speech Recognition
- Text to Speech
- Speech Intent Recognition

LUIS
(Language Understanding Intelligent Service)

- Detect Intent
- Determine Entities
- Improve Models

Video APIs

- Summarization
- Stabilization
- Motion detection
- **Analysis**

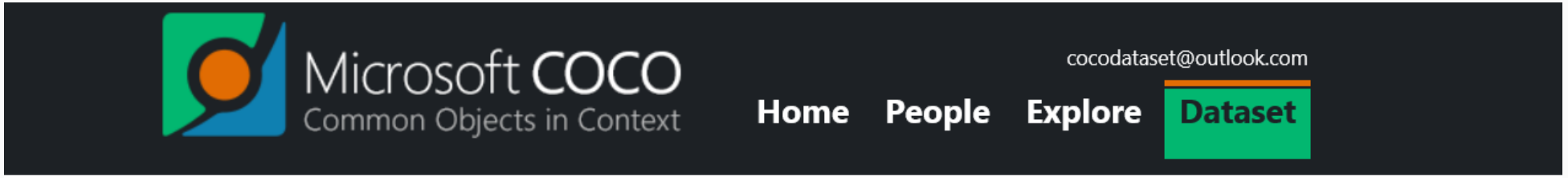
Video to Sentence

Microsoft

Video to Language

- Video description (from individual concepts to natural sentence)
  - Robotic vision
  - Movie description for blind people
  - Incident report for surveillance videos
- Video indexing
  - Learning embedding models from language-video pairs

Image captioning competition



Overview | Download | Evaluate | Leaderboard | Challenges

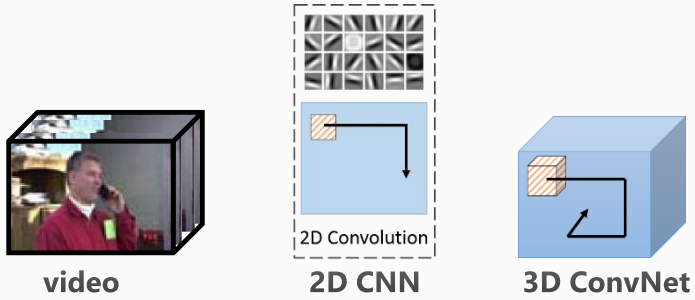
Table-C5 | Table-C40 | Table-human

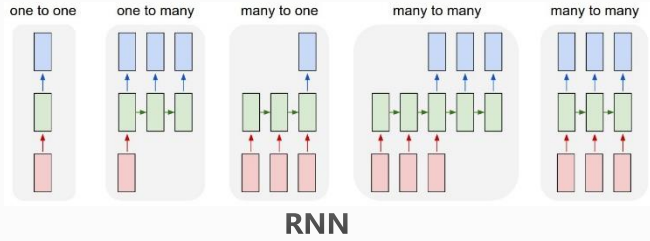
Last update: June 8, 2015. Visit CodaLab for the latest results.

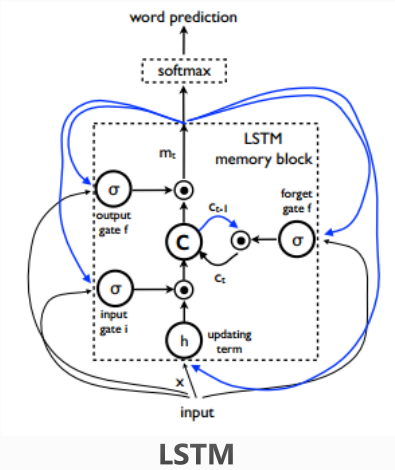
	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	
Google[4]	0.943	0.254	0.53	0.713	0.542	0.407	0.309	[CVPR 2015 oral; arxiv @ 2014-11-17]
MSR Captivator[9]	0.931	0.248	0.526	0.715	0.543	0.407	0.308	[arxiv @ 2015-05-07]
m-RNN[15]	0.917	0.242	0.521	0.716	0.545	0.404	0.299	[arxiv @ 2015-04-25]
MSR[8]	0.912	0.247	0.519	0.695	0.526	0.391	0.291	[CVPR 2015 poster; arxiv @ 2014-11-18]
Nearest Neighbor[11]	0.886	0.237	0.507	0.697	0.521	0.382	0.28	[arxiv @ 2015-05-27]
m-RNN (Baidu/ UCLA)[16]	0.886	0.238	0.524	0.72	0.553	0.41	0.302	[NIPS 2014 workshop; arxiv @ 2014-12-20]
Berkeley LRCN[2]	0.869	0.242	0.517	0.702	0.528	0.384	0.277	[CVPR 2015 oral; arxiv @ 2014-11-17]
Human[5]	0.854	0.252	0.484	0.663	0.469	0.321	0.217	
Montreal/Toronto[10]	0.85	0.243	0.513	0.689	0.515	0.372	0.268	[arxiv @ 2015-02-10]

Challenges for video-to-sentence

- Video-to-sentence is still under-explored
- Learning video representation
  - visual objects (AlexNet, GoogLeNet, VGG)
  - temporal dynamics (C3D, optical flow)
  - audio (MFCC, Spectrum-SIFT)
- Deep neural network design
  - filter: 2D CNN/3D CNN
  - multi-layer RNN (LSTM)
- Sequence learning
  - sequence vs. static frames (pooling/alignment)
  - semantic relationship between entire sentence and video content







How does video-to-sentence work?

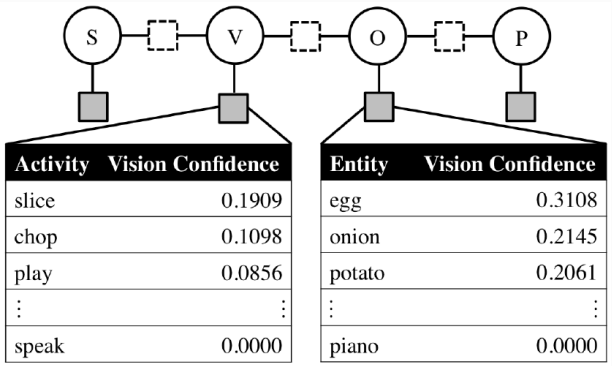
- Language template-based model [UT Austin'14, SUNY-Byffalo'15]
  - SVO detection -> template-based sentence generation

Predicting the best words for describing:
Subject (S) - Verb (V) - Object (O)

S: man, V: play, O: guitar

Generating sentence using template:
"determiner (a/the) - Subject - Verb (tense) - Preposition (optional) - determiner (a/the) - Object (optional)"

Sentence: "a man is playing a guitar"



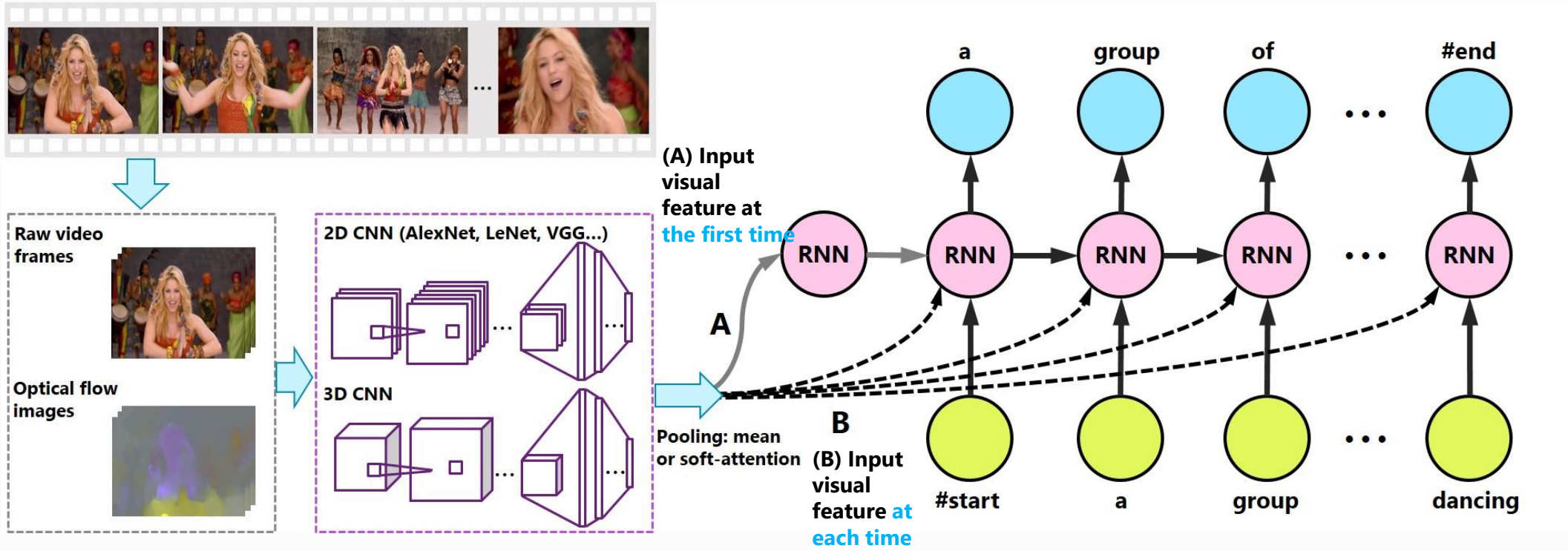
Activity	Vision Confidence
slice	0.1909
chop	0.1098
play	0.0856
⋮	⋮
speak	0.0000

Entity	Vision Confidence
egg	0.3108
onion	0.2145
potato	0.2061
⋮	⋮
piano	0.0000

factor graph model (FGM)

How does video-to-sentence work?

- RNN-based model [UC Berkeley'14'15, UdeM'15]
  - decoding (temporal) video representation into sequence of words



- **UC Berkeley'14:** AlexNet + mean pooling + B
- **UdeM'15:** (GoogLeNet + 3D CNN) + soft-attention + B
- **UC Berkeley'15:** (VGG + Optical Flow) + sequence encoding-decoding
- **MSRA:** (VGG 2D CNN + 3D CNN) + mean pooling + A + joint learning

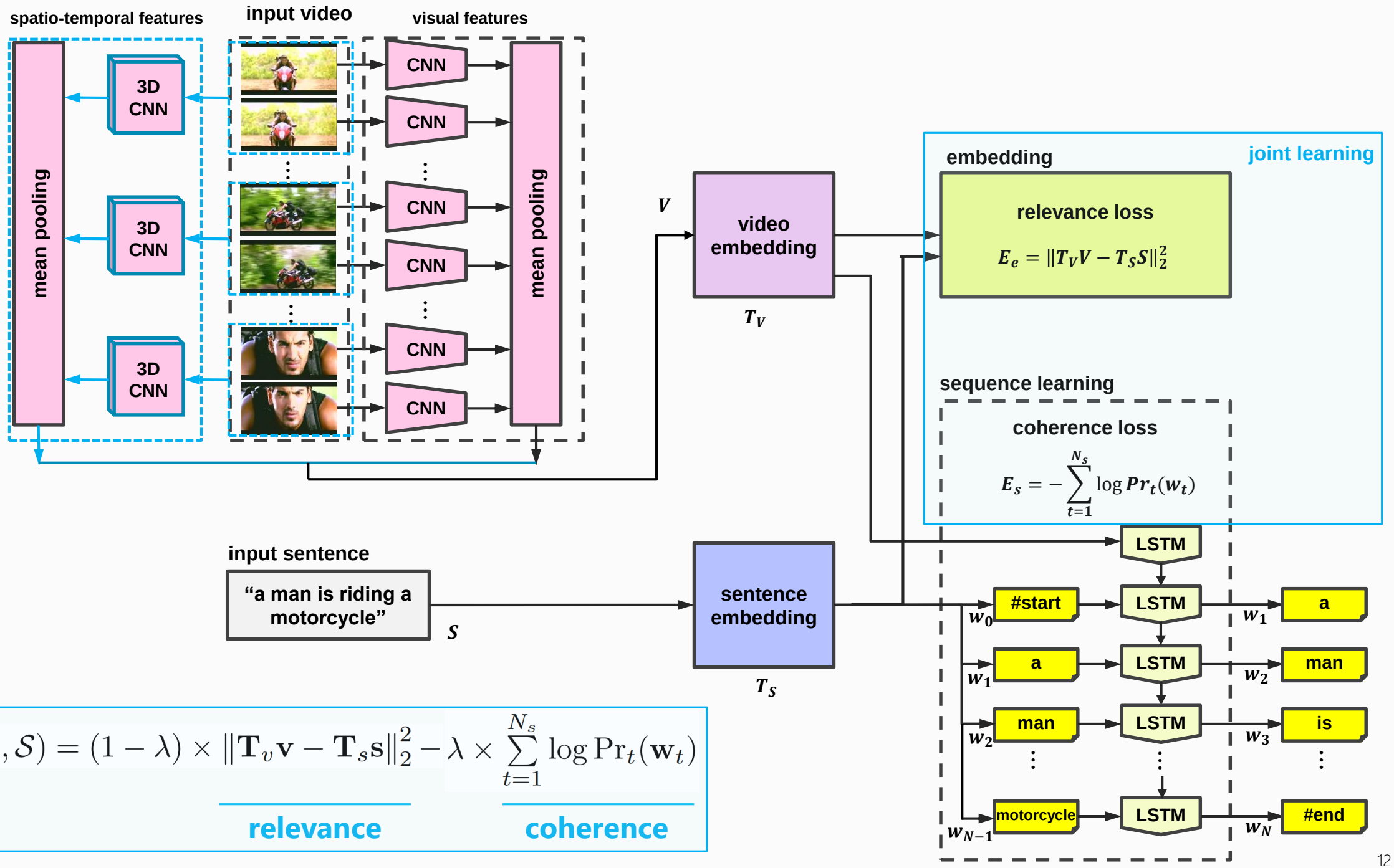
Our work: joint embedding and translating

- Key issues in sentence generation
  - *relevance*: relationship between sentence (S, V, O) semantics and video content
  - *coherence*: sentence grammar

LSTM: a man is playing a guitar
LSTM-E: a man is playing a piano

LSTM: a man is dancing
LSTM-E: a group of people are dancing

- Joint learning: relevance + coherence
  - Holistically looking at both entire sentence semantics and video content
  - Learning powerful video representation: 2D CNN (visual) + C3D (motion)



$$E(\mathcal{V},\mathcal{S}) = (1-\lambda) \times \|\mathbf{T}_v\mathbf{v} - \mathbf{T}_s\mathbf{s}\|_2^2 - \lambda \times \sum_{t=1}^{N_s} \log \Pr_t(\mathbf{w}_t)$$

relevance coherence

Evaluations

- Dataset (MSR Video Description Corpus, a.k.a. YouTube2Text)
  - 1,970 Youtube video snippets (1,200 training, 100 validation, 670 testing)
  - 10-25 sec for each clip
  - ~40 human-generated sentences for each clip (by AMT)
  - dictionary: 15,903 -> 7,000; 45 S-groups, 218 V-groups, 241 O-groups
- Training: 12 hrs in one single CPU; testing: ~5 sec per clip

1. a man is petting a dog
2. a man is petting a tied up dog
3. a man pets a dog
4. a man is showing his dog to the camera
5. a boy is trying to see something to a dog

1. a man is playing the guitar
2. a men is playing instrument
3. a man plays a guitar
4. a man is singing and playing guitar
5. the boy played his guitar

1. a kitten is playing with his toy
2. a cat is playing on the floor
3. a kitten plays with a toy
4. a cat is playing
5. a cat tries to get a ball

1. a man is singing on stage
2. a man is singing into a microphone
3. a man sings into a microphone
4. a singer sings
5. the man sang on stage into the microphone

Performance

The accuracy of S-V-O triplet prediction.

Model	Team	Subject%	Verb%	Object%
FGM	UT Austin, COLING (2014/08)	76.42	21.34	12.39
CRF	SUNY-Buffalo, AAAI (2015/01)	77.16	22.54	9.25
CCA	Stanford, CVPR (2010/06)	77.16	21.04	10.99
JEM	SUNY-Buffalo, AAAI (2015/01)	78.25	24.45	11.95
LSTM	UC Berkeley, NAACL (2014/12)	71.19	19.40	9.70
LSTM-E	**MSRA, arxiv (2015/05)**	**80.45**	**29.85**	**13.88**

The performance of sentence generation.

Model	Team	METEOR%	BLEU@4%
LSTM	UC Berkeley, NAACL (2014/12)	26.9	31.2
SA	UdeM, arxiv (2015/02)	29.6	42.2
S2VT	UC Berkeley, arxiv (2015/05)	29.8	--
LSTM-E	**MSRA, arxiv (2015/05)**	**31.0**	**45.3**

Video-to-Sentence results (within YouTube2Text)

Human: a kitten is playing with his toy
LSTM: a cat is playing with a mirror
LSTM-E: a kitten is playing with a toy

Human: a man is singing on the stage
LSTM: a man is playing a flute
LSTM-E: a man is singing

Human: a group of people are dancing
LSTM: a man is dancing
LSTM-E: a group of people are dancing

Human: a person is playing a piano keyboard
LSTM: a man is playing a guitar
LSTM-E: a man is playing a piano

Human: a man is talking on a cell phone
LSTM: a woman is talking
LSTM-E: a man is talking on a phone

Human: a man is riding his motocycle
LSTM: a man is riding a car
LSTM-E: a man is riding a motorcycle

Video-to-Sentence results (out of YouTube2Text)

A car is running

A man is cutting a piece of meat

A man is performing on a stage

A man is riding a bike

A man is singing

A panda is walking

A woman is riding a horse

A man is flying in a field

What if applying image captioning tech to video?

Video-to-sentence:

LSTM-E: a group of people are dancing

Image-to-sentence (keyframe-based): http://deeplearning.cs.toronto.edu/i2t

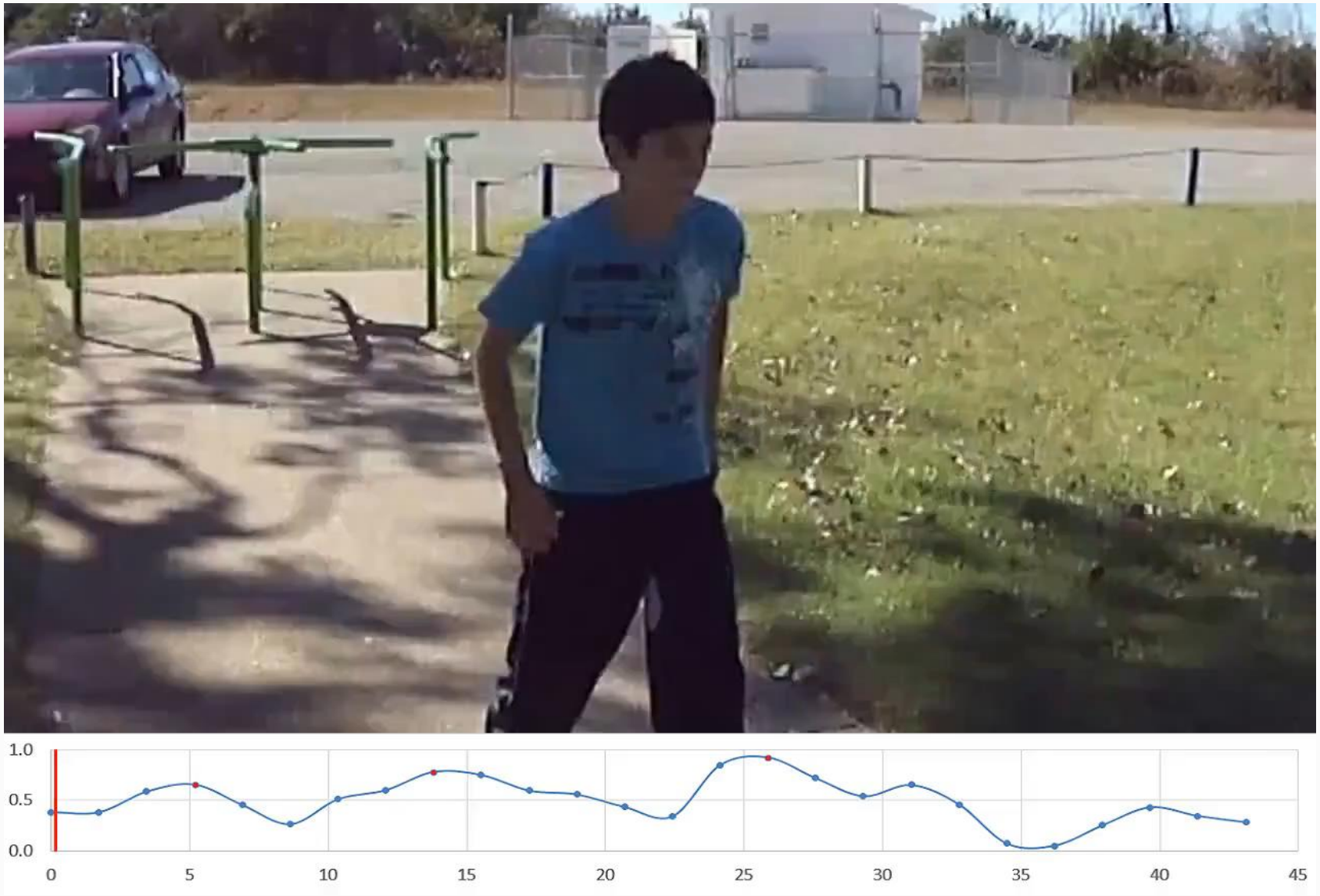
a group of people are jumping up on a stage look on a horse

the two people are standing in the front of their heads

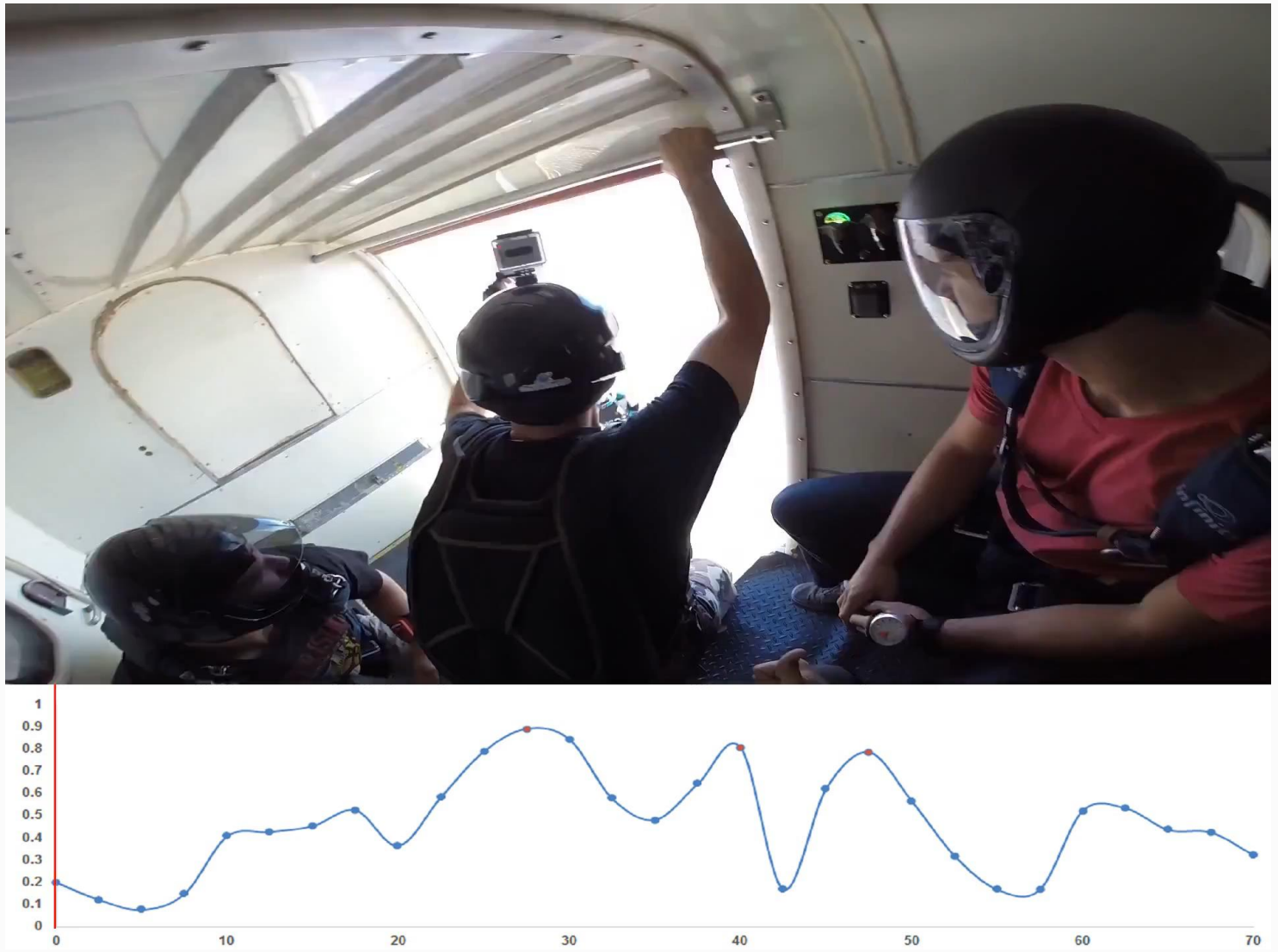
a group of people standing around next to each other

Highlight detection

Example: parkour (highlight + timelapse 4X + music)



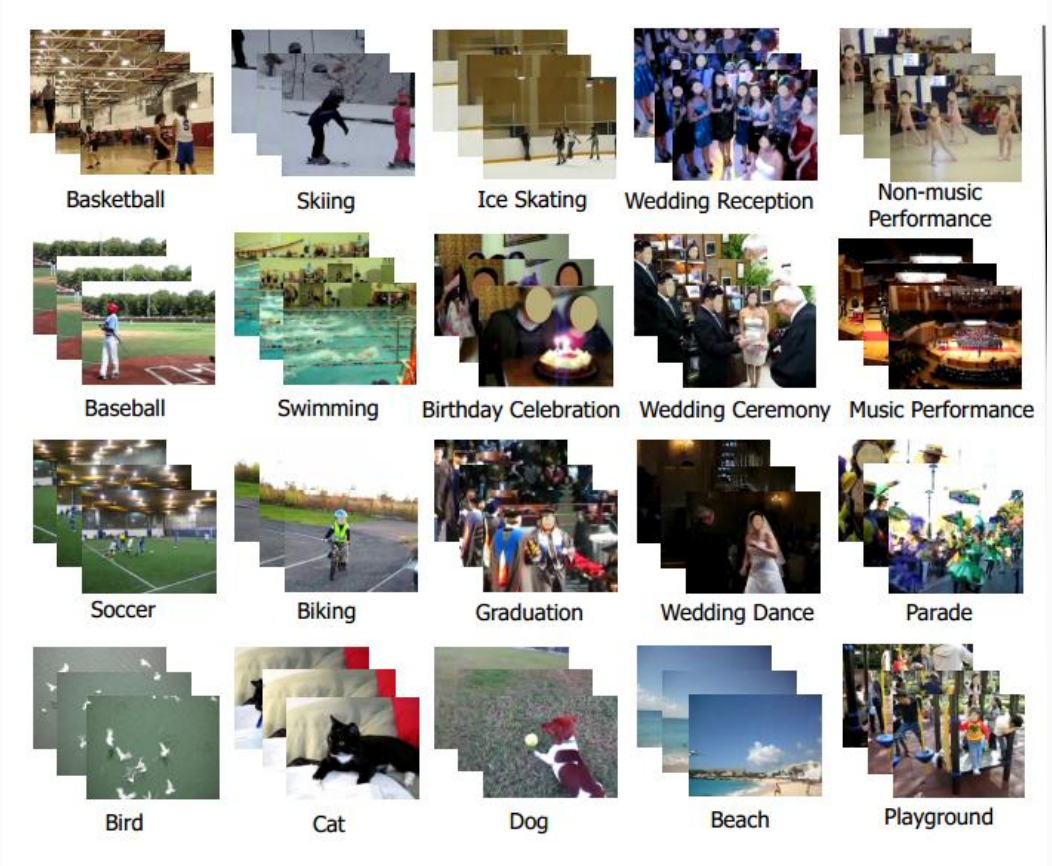
Example: GoPro video

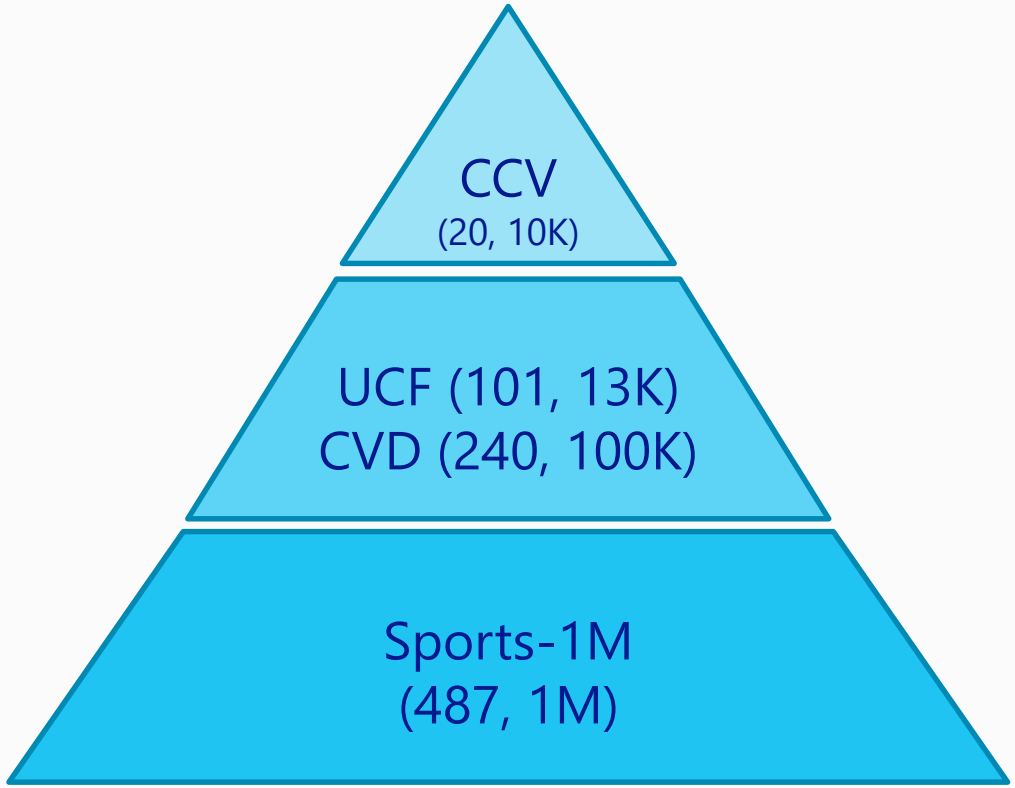


Video classification

Action recognition from video

- Examples of video categories (CCV-20)

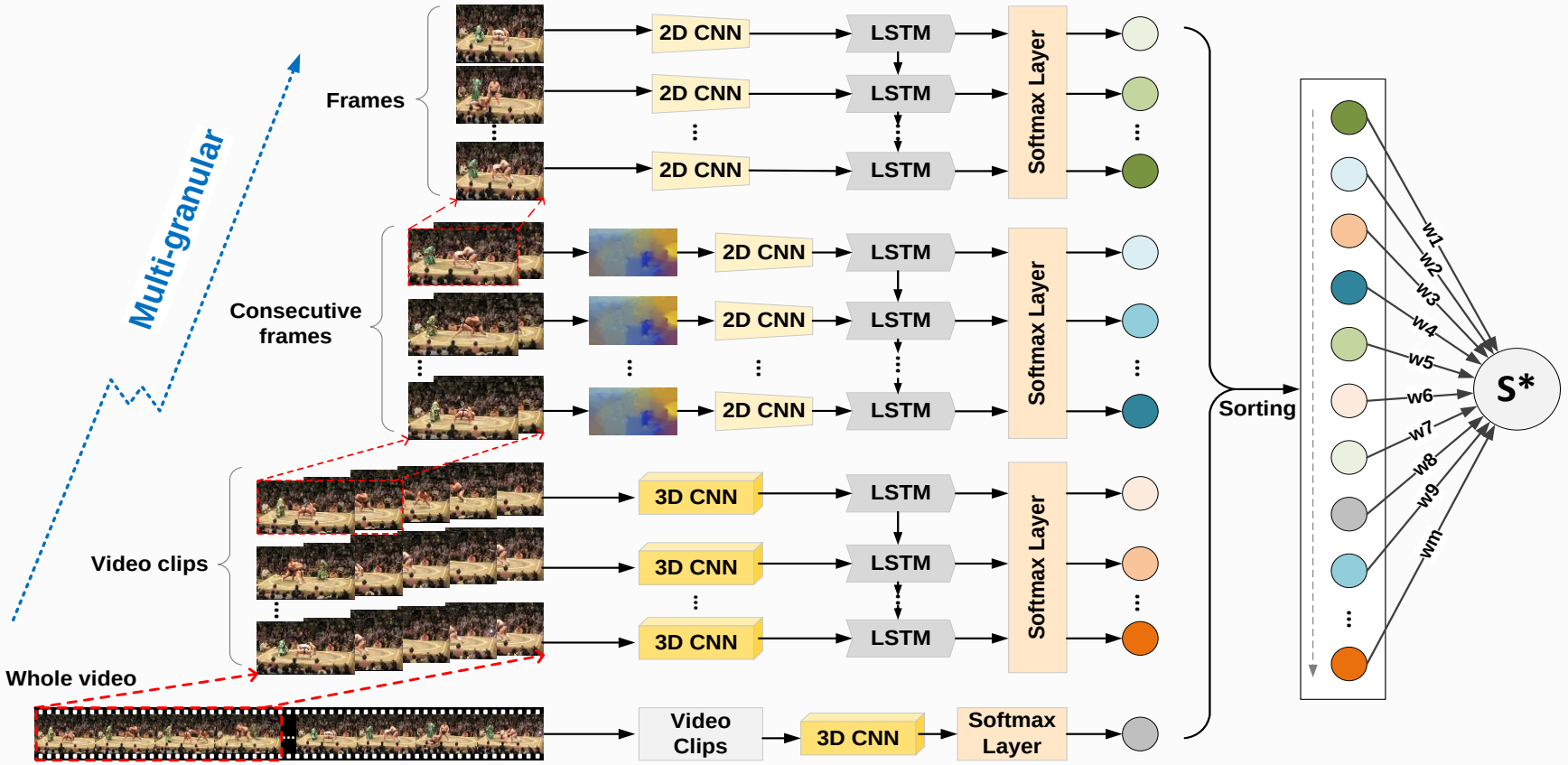




Action recognition

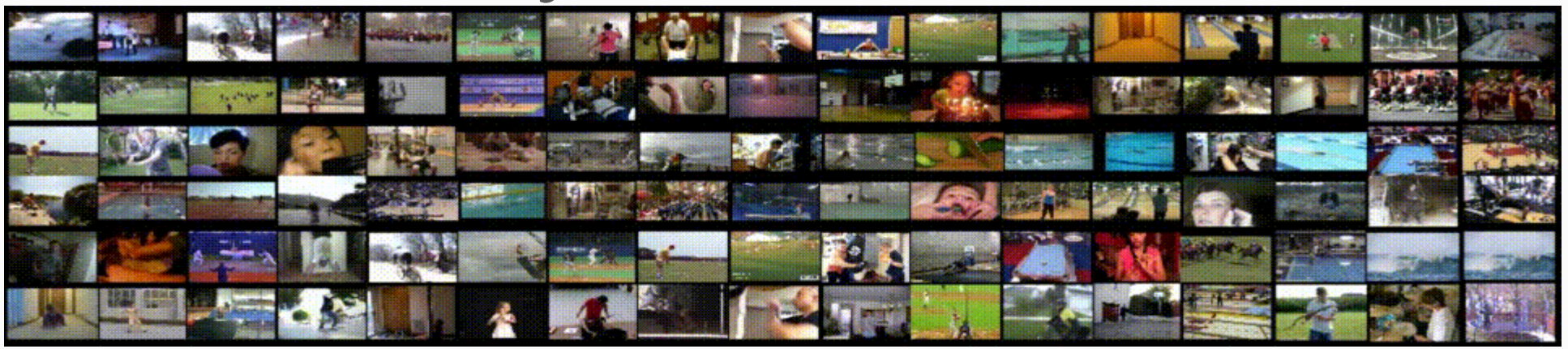
Framework

- Multi-granular spatiotemporal architecture
  - deep feature learning representation for video
  - multi-granular streams (frame + optical flow + clip + video)
  - relative importance learning for each component



THUMOS Challenge 2015

In conjunction with CVPR'15



Rank	Entry	Run1	Run2	Run3	Run4	Run5
1	U. of Tech., Sydney & CMU	0.7384	0.7157	0.7011	0.6913	0.647
2	MSR Asia (MSM)	0.6861	0.6869	0.6878	0.6886	0.6897
3	Zhejiang University	0.6876	0.6643	0.6859	0.6809	0.5625
4	INRIA_LEAR	0.6814	0.6811	0.5395	0.6739	0.6793
5	CUHK & Shenzhen Inst. Adv. Tech.	0.4894	0.5746	0.6803	0.6576	0.6604
6	University of Amsterdam	0.6798	NA	NA	NA	NA
7	Tianjin University	0.6666	0.6551	0.6324	0.5514	0.5357
8	USC & Tsinghua U.	0.6354	0.6398	0.6346	0.5639	0.6357
9	MIL + U. Tokyo	0.6159	0.6172	0.6174	0.6087	0.4986