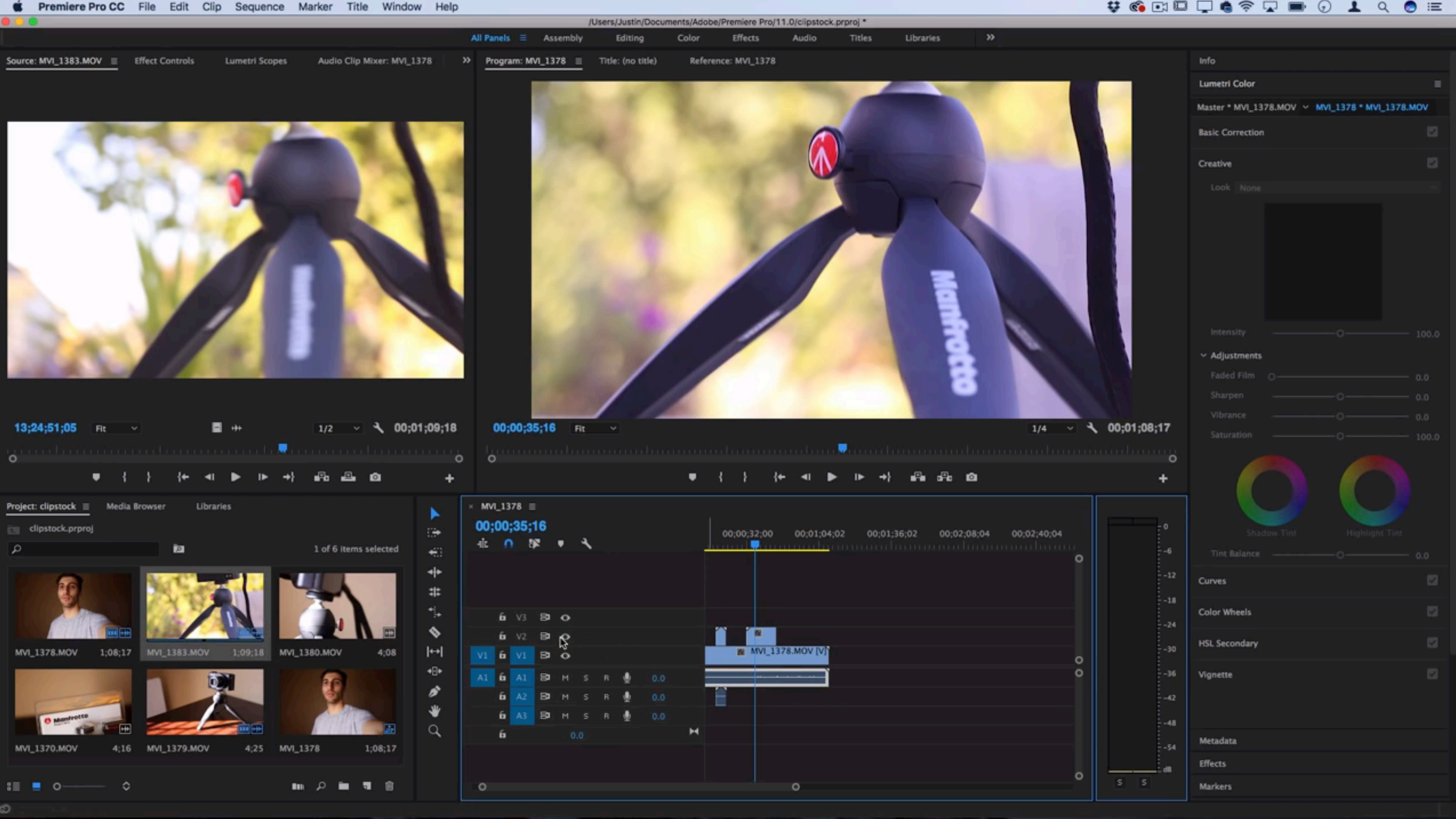
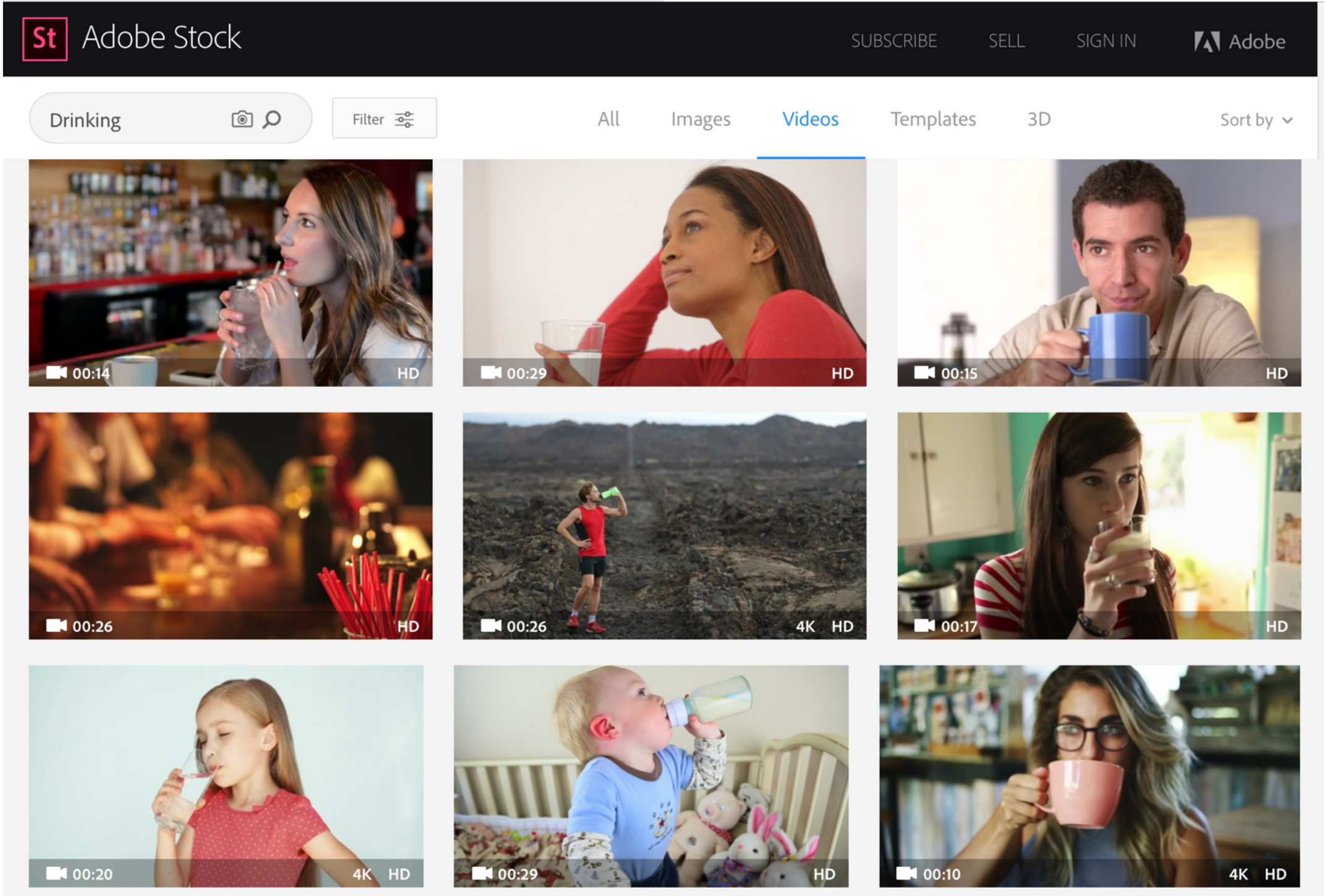
Learning from Video: Recognizing Actions and Localizing Moments with Natural Language

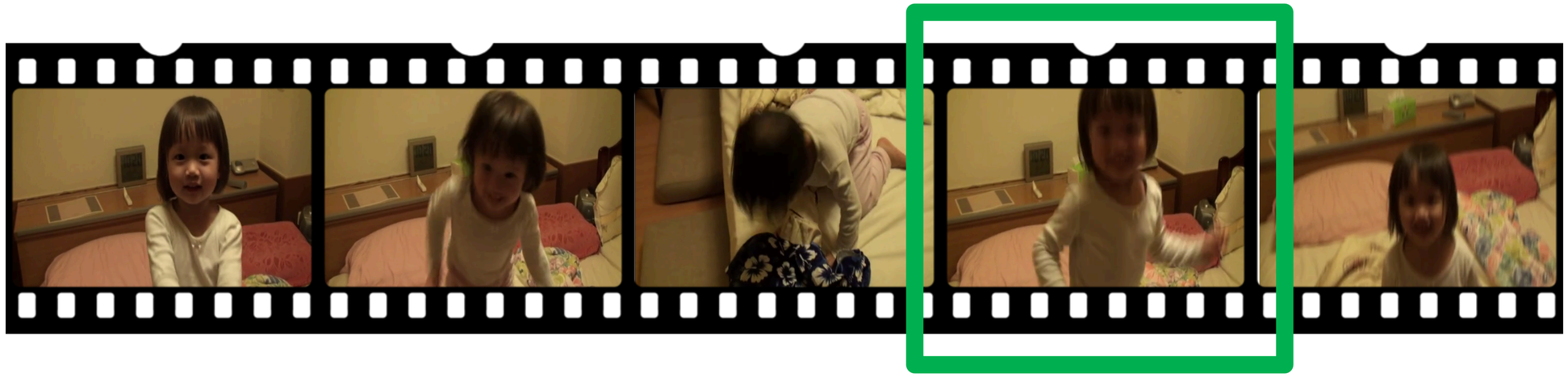
Bryan Russell

Adobe Research

Sound check

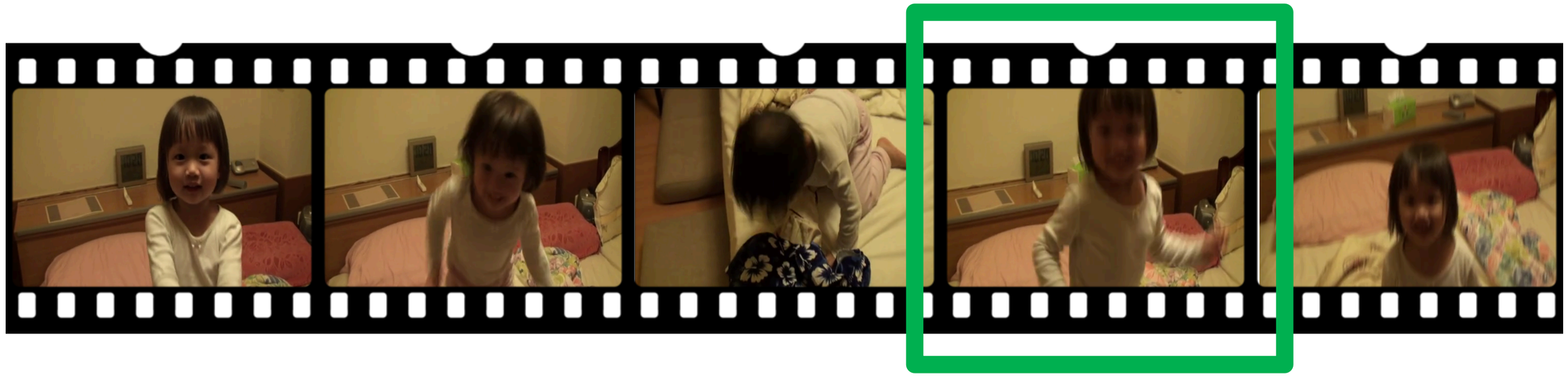

Premiere Pro CC File Edit Clip Sequence Marker Title Window Help
/Users/Justin/Documents/Adobe/Premiere Pro/11.0/clipstock.prproj *
All Panels Assembly Editing Color Effects Audio Titles Libraries
Source: MVI_1383.MOV Effect Controls Lumetri Scopes Audio Clip Mixer: MVI_1378
Program: MVI_1378 Title: (no title) Reference: MVI_1378
13;24;51;05 Fit 1/2 00;01;09;18
00;00;35;16 Fit 1/4 00;01;08;17
Info
Lumetri Color
Master * MVI_1378.MOV MVI_1378 * MVI_1378.MOV
Basic Correction
Creative
Look None
Intensity 100.0
Adjustments
Faded Film 0.0
Sharpen 0.0
Vibrance 0.0
Saturation 100.0
Shadow Tint
Highlight Tint
Tint Balance 0.0
Curves
Color Wheels
HSL Secondary
Vignette
Metadata
Effects
Markers
Project: clipstock Media Browser Libraries
clipstock.prproj
1 of 6 items selected
MVI_1378.MOV 1;08;17
MVI_1383.MOV 1;09;18
MVI_1380.MOV 4;08
MVI_1370.MOV 4;16
MVI_1379.MOV 4;25
MVI_1378 1;08;17
MVI_1378
00;00;35;16
00;00;32;00 00;01;04;02 00;01;36;02 00;02;08;04 00;02;40;04
V3 V2 V1 A1 A2 A3
MVI_1378.MOV [V]


Adobe Stock
SUBSCRIBE
SELL
SIGN IN
Adobe
Drinking
Filter
All
Images
Videos
Templates
3D
Sort by
00:14 HD
00:29 HD
00:15 HD
00:26 HD
00:26 4K HD
00:17 HD
00:20 4K HD
00:29 HD
00:10 4K HD



Query by object: girl

Query by action: jumping



Query by natural language:
“The little girl jumps after falling”

ActionVLAD: Learning spatio-temporal aggregation for action classification

Rohit Girdhar (CMU)
Abhinav Gupta (CMU)
Deva Ramanan (CMU)
Josef Sivic (INRIA Paris, Adobe)
Bryan Russell (Adobe)

CVPR 2017

Input: video

Casual team of financial traders take a break from work …

File: #80552741 | Author: hotelfoxtrot69

Download comp image

Search similar contents

Output: tags

Objects/People:
- Computer screen
- Women

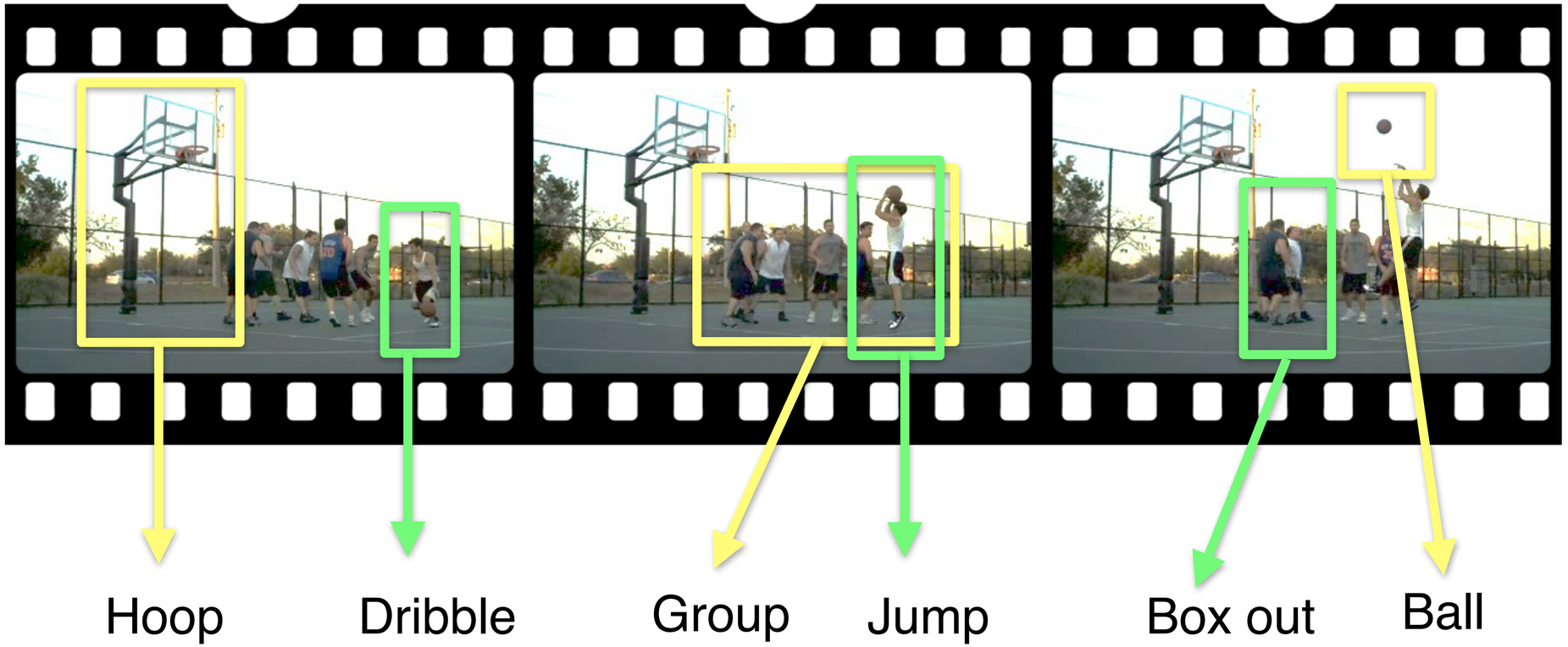
Scene:
- Workplace
- Office

Actions:
- Laughing
- High-five

Attributes:
- Filmed (not cartoon)

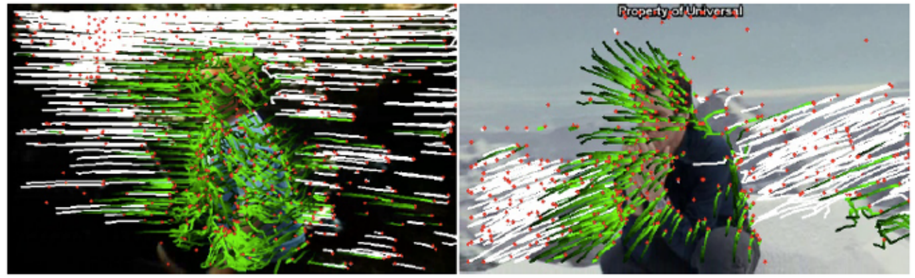
Challenge: What is a good spatiotemporal representation?

Example: “Basketball shoot”



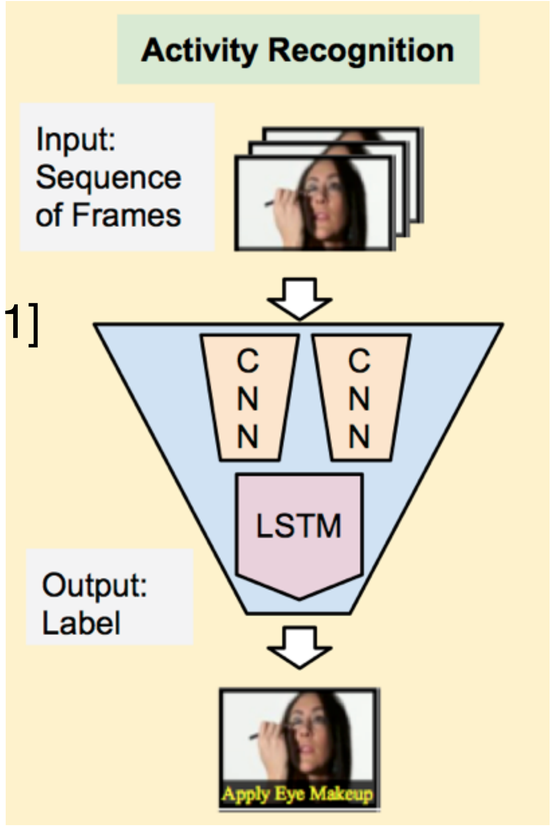
Prior work

iDT: H. Wang and C. Schmid, 2013

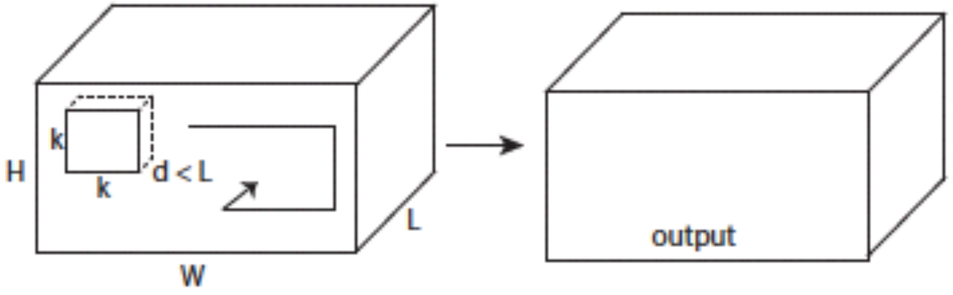


Also: [H. Wang, et al., 2009], [H. Wang, et al. 2011]

LRCN: J. Donohue, et al., 2015

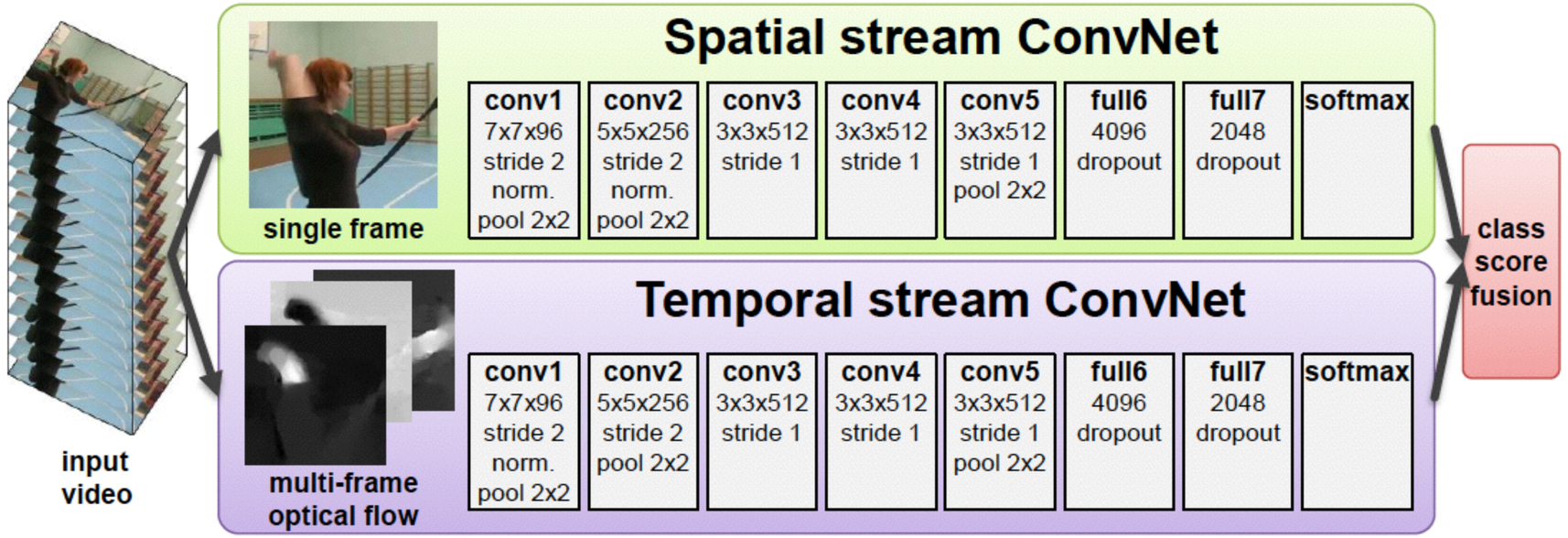


C3D: D. Tran, et al., 2015

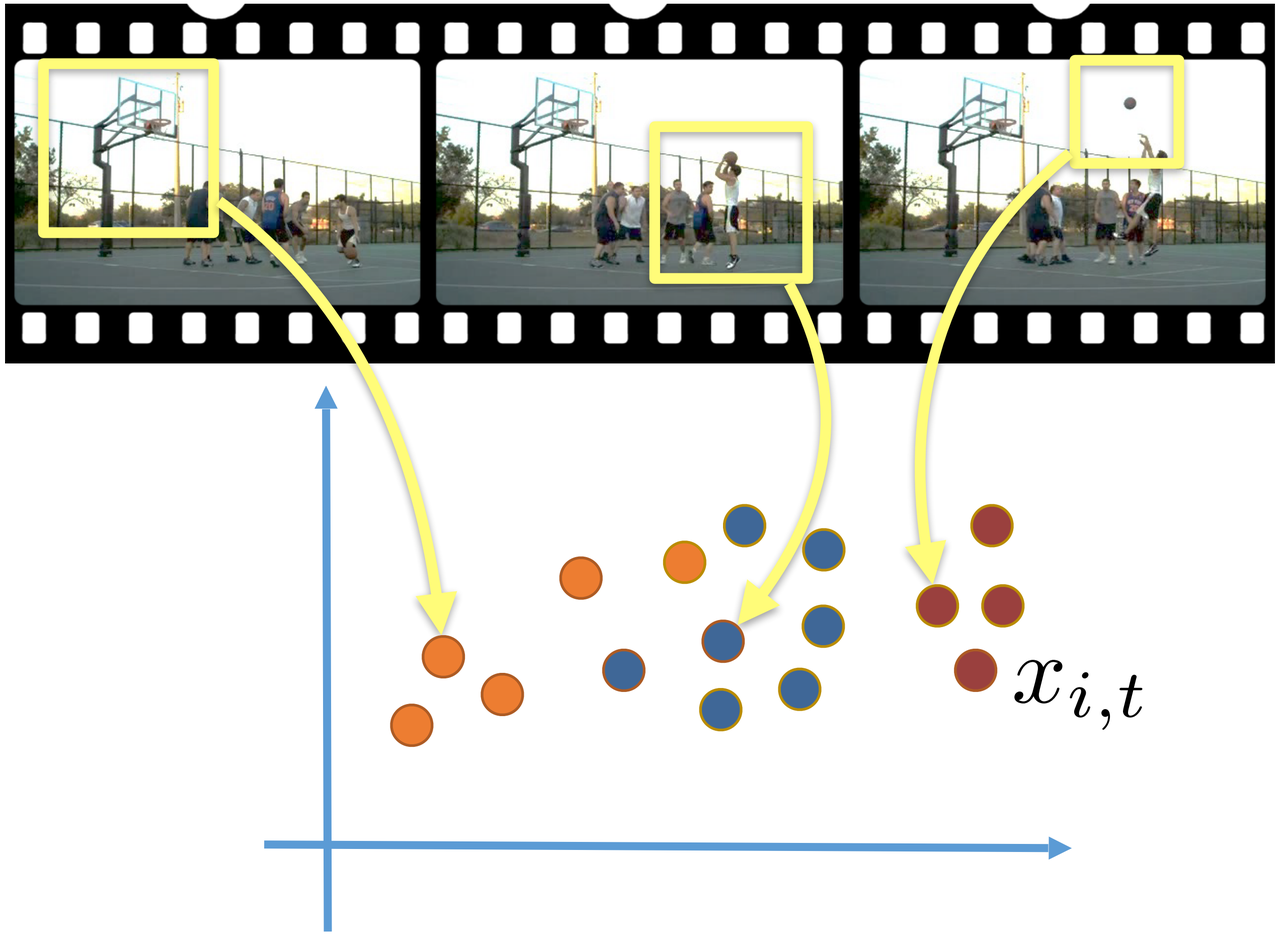


Two-stream networks

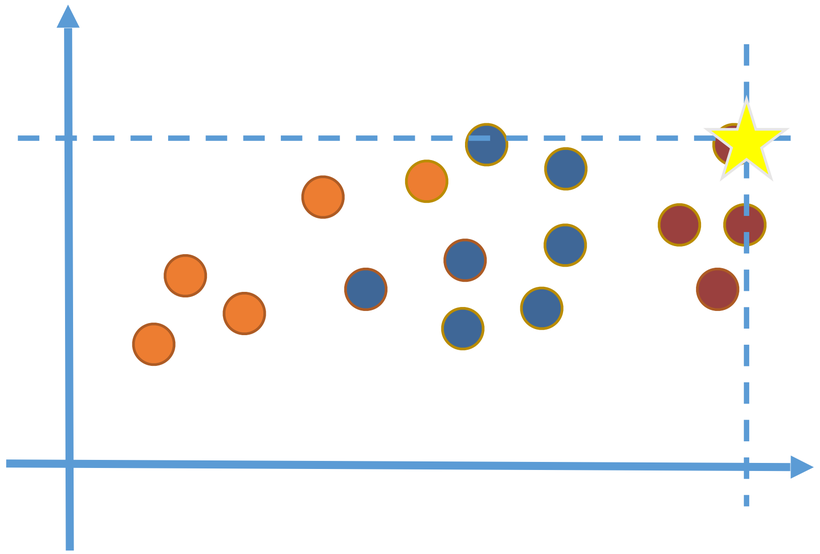
K. Simonyan and A. Zisserman, 2014



See also: [L. Wang, et al., 2015], [L. Wang, et al., 2016], [G. Varol, et al., 2016], [W. Zhu, et al., 2016], [X. Wang, et al., 2016], [C. Feichtenhofer, et al., CVPR 2016], [C. Feichtenhofer, et al., NIPS 2016]

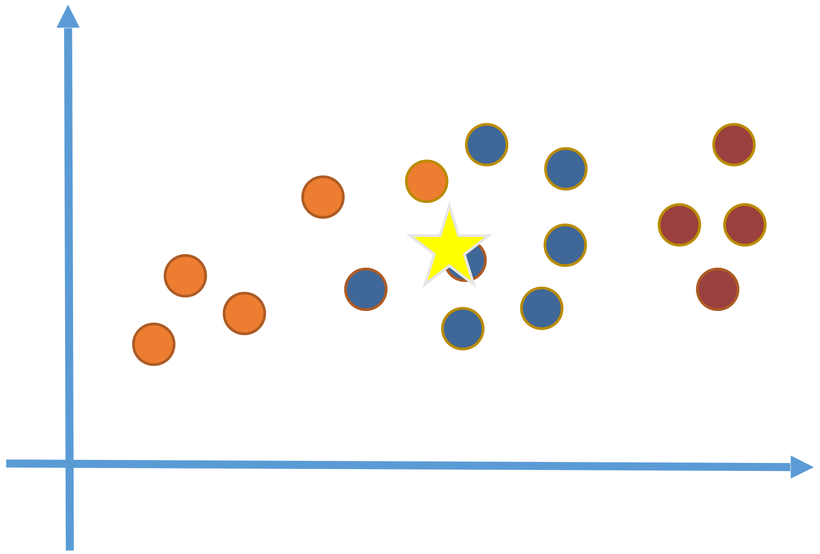

$x_{i,t}$

Max pooling



$$f[j] \leftarrow \max_{\substack{t \in \{1,\ldots,T\} \\ i \in \{1,\ldots,N\}}} x_{i,t}[j]$$

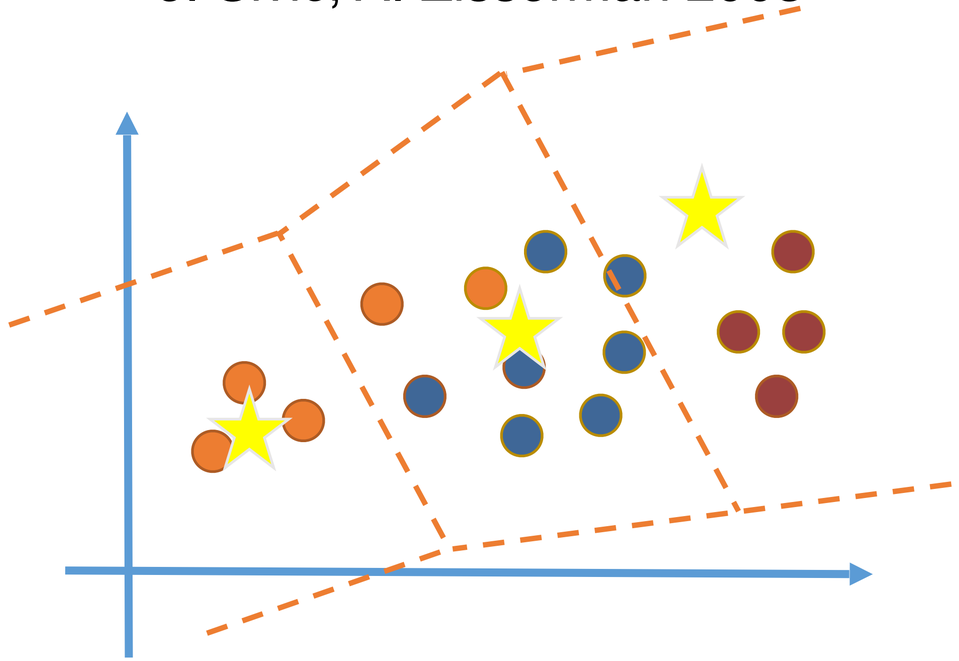
Average pooling



$$f[j] \leftarrow \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}[j]$$

Bag of “action words”

J. Sivic, A. Zisserman 2003

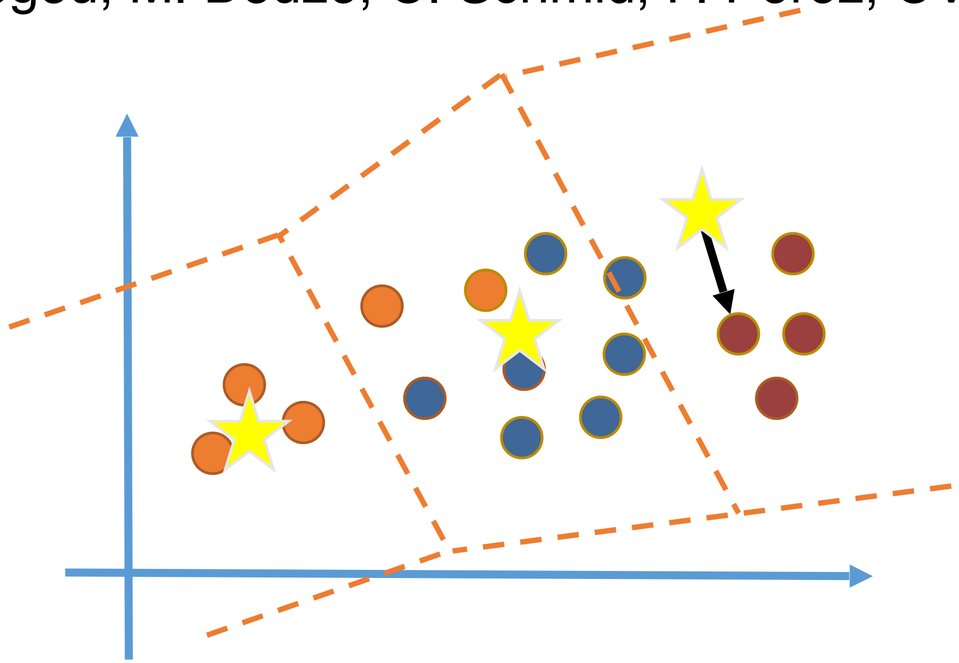


$$f[k] \leftarrow \sum_{t=1}^{T} \sum_{i=1}^{N} a_k(x_{i,t})$$

Hard Assignment

VLAD pooling

H. Jégou, M. Douze, C. Schmid, P. Pérez, CVPR 2010

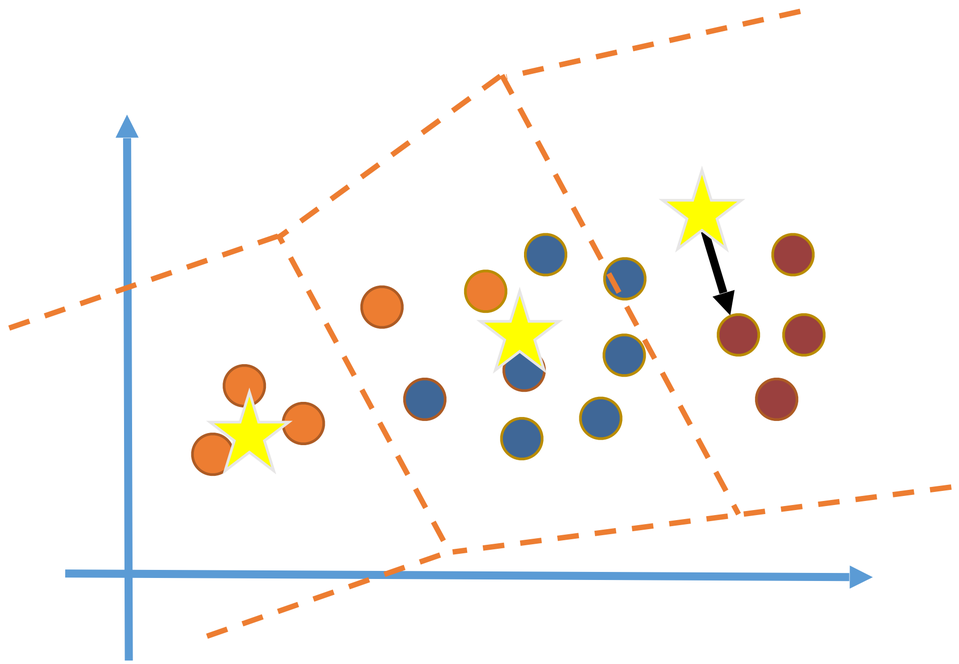


$$f[j,k] \leftarrow \sum_{t=1}^{T} \sum_{i=1}^{N} a_k(x_{i,t}) \left(x_{i,t}[j] - c_k[j]\right)$$

Hard assignment

Residual

Ours: ActionVLAD spatio-temporal pooling



$$f[j,k] \leftarrow \sum_{t=1}^{T}\sum_{i=1}^{N} \frac{\exp^{-\alpha||x_{i,t}-c_k||^2}}{\sum_{k'}\exp^{-\alpha||x_{i,t}-c_{k'}||^2}} \left(x_{i,t}[j]-c_k[j]\right)$$

Soft assignment

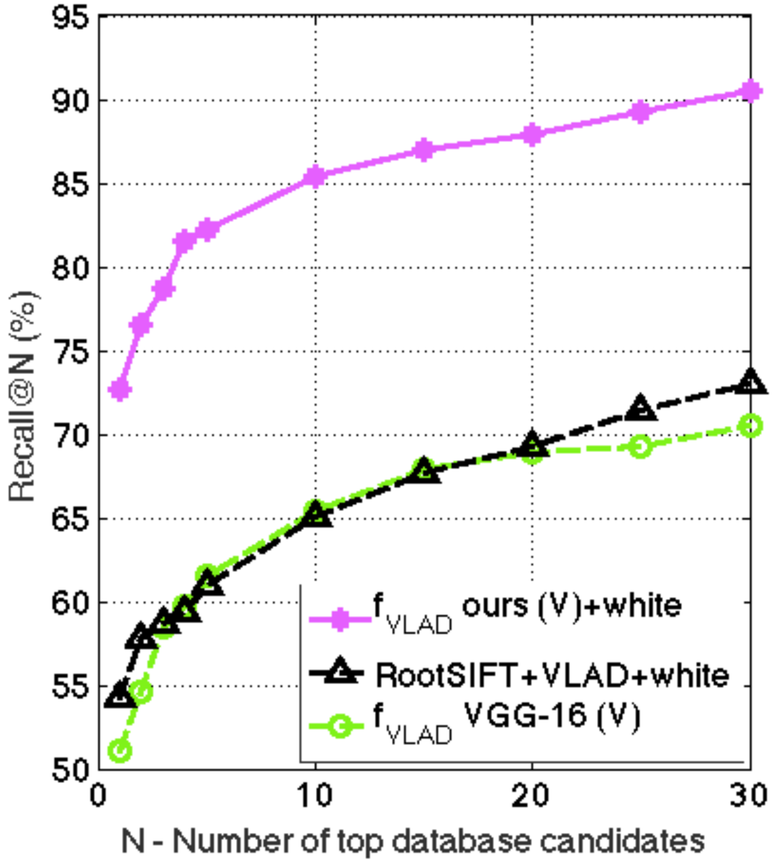
Residual

NetVLAD – Weakly-supervised place recognition

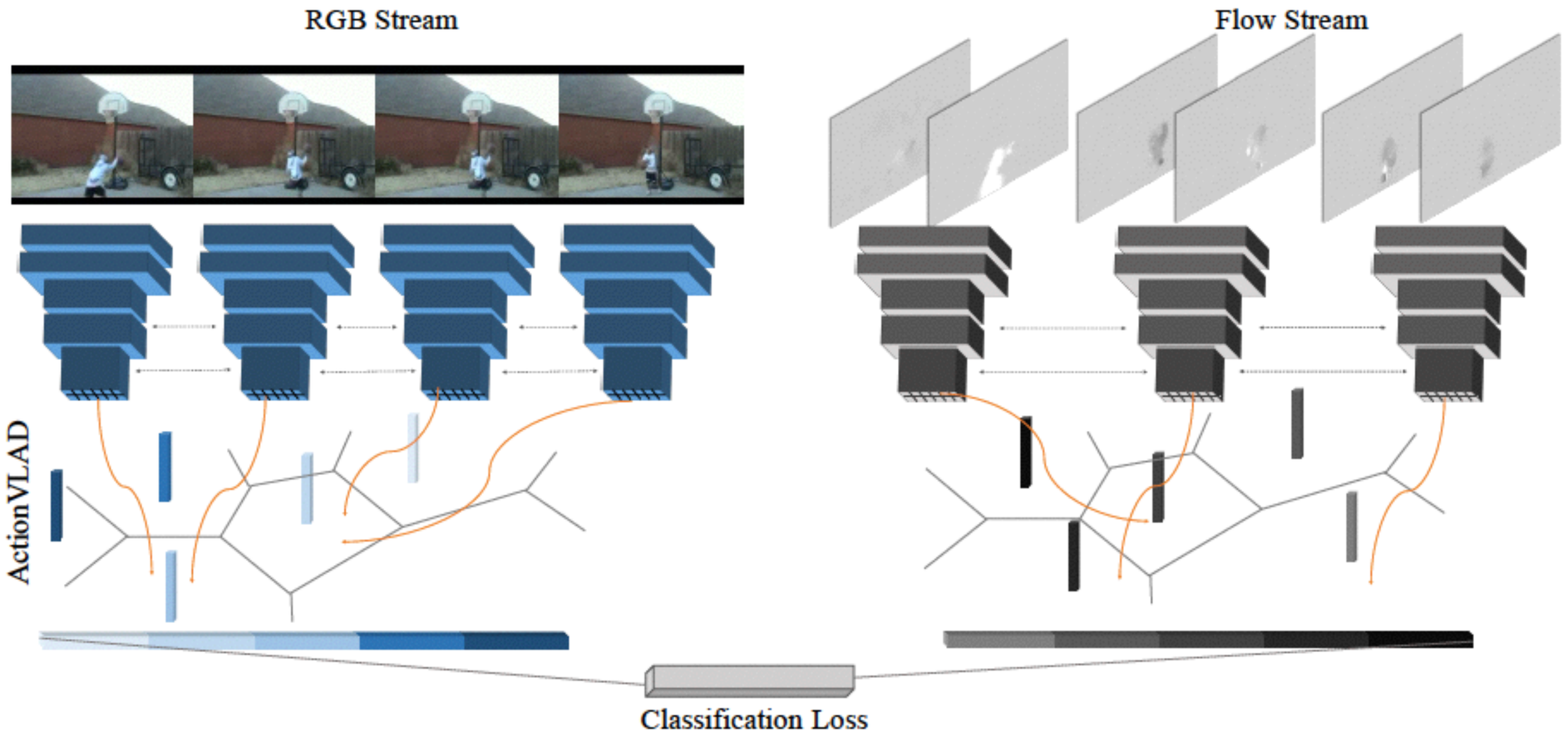
Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, Josef Sivic (CVPR'16)

State-of-art performance on all place recognition benchmarks

Tokyo 24/7



Our architecture



- Base CNN architecture is VGG or BN-inception
- ActionVLAD after conv5 layer
- Late fuse RGB and flow streams

	UCF101	HMDB51
Spatio-Temporal ConvNet [26]	65.4	-
LRCN [13]	82.9	-
C3D [49]	85.2	-
Factorized ConvNet [45]	88.1	59.1
VideoDarwin [16]	-	63.7
Two-Stream + LSTM [31] (GoogLeNet)	88.6	-
Two-Stream ConvNet [42] (VGG-M)	88.0	59.4
Two-Stream ConvNet [57, 59] (VGG-16)	91.4	58.5
Two-Stream Fusion [15] (VGG-16)	92.5	65.4
TDD+FV [55]	90.3	63.2
RNN+FV [29]	88.0	54.3
Transformations [59]	92.4	62.0
LTC [51]	91.7	64.8
KVMF [63]	**93.1**	63.
ActionVLAD (LateFuse, VGG-16)	92.7	**66**
DT+MVSV [9]		
iDT+FV [53]	8	
iDT+HSV [33]	87.9	
MoFAP [56]	88.3	
C3D+iDT [49]	90.4	
Two-Stream Fusion+iDT [15]	93.5	
LTC+iDT [51]	92.7	
ActionVLAD (VGG-16) + iDT	**93.6**	**9.8**
TSN (BN-Inception, 3-modality) [58]	94.2	69.4
DT+Hybrid architectures [12]	92.5	70.4
ST-ResNet+iDT [14]	94.6	70.3

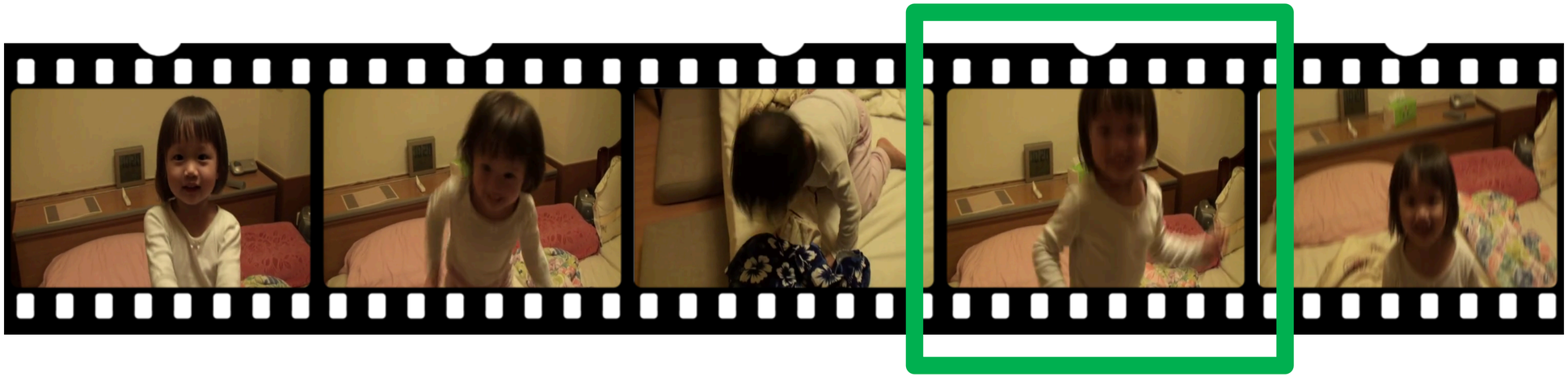
A. Miech, I. Laptev, J. Sivic
winner Youtube-8M
challenge

Localizing Moments in Video with Natural Language

Lisa Anne Hendricks (UC Berkeley)
Trevor Darrell (UC Berkeley)
Eli Shechtman (Adobe)
Josef Sivic (INRIA Paris, Adobe)
Oliver Wang (Adobe)
Bryan Russell (Adobe)

To appear at ICCV 2017

Text query: “The little girl jumps after falling”



Related: Object Retrieval by Text Query

Query: “Chair on the left”

Hu, et al. “Natural Language Object Retrieval”. CVPR 2016
Mao, et al. “Generation and Comprehension of Unambiguous Object Descriptions.”. CVPR 2016

Challenge: temporal language

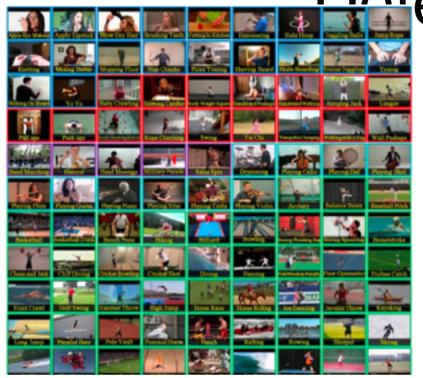
The moment before the children appear

The lady begins to entertain the baby

Challenge: Where to find a dataset?

- Temporally grounded sentence descriptions
- Open-world natural language
- Referring expressions
- Unedited personal videos

Action Recognition/ Detection



UCF101
HMDB-51
Kinetics

Sports-1M

ActivityNet, AVA

Video Description

MSR-VTT

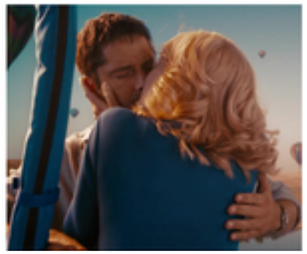
A black and white horse runs around.

A woman gives a speech on a news channel.

TACoS

898: The man takes out a cutting board.
1300: He washes the carrot.
1500: He takes out a knife.

Movie Description



Abby clasps her hands around his face and kisses him

MovieQA
VideoMCC
Dense captioning events

Referring Expressions in Images

Query: Man in the middle with blue shirt and blue shorts.

Kazemzadeh, et al. "ReferItGame: Referring to Objects in Photographs of Natural Scenes." *EMNLP*. 2014.

Query: Woman in a red shirt.

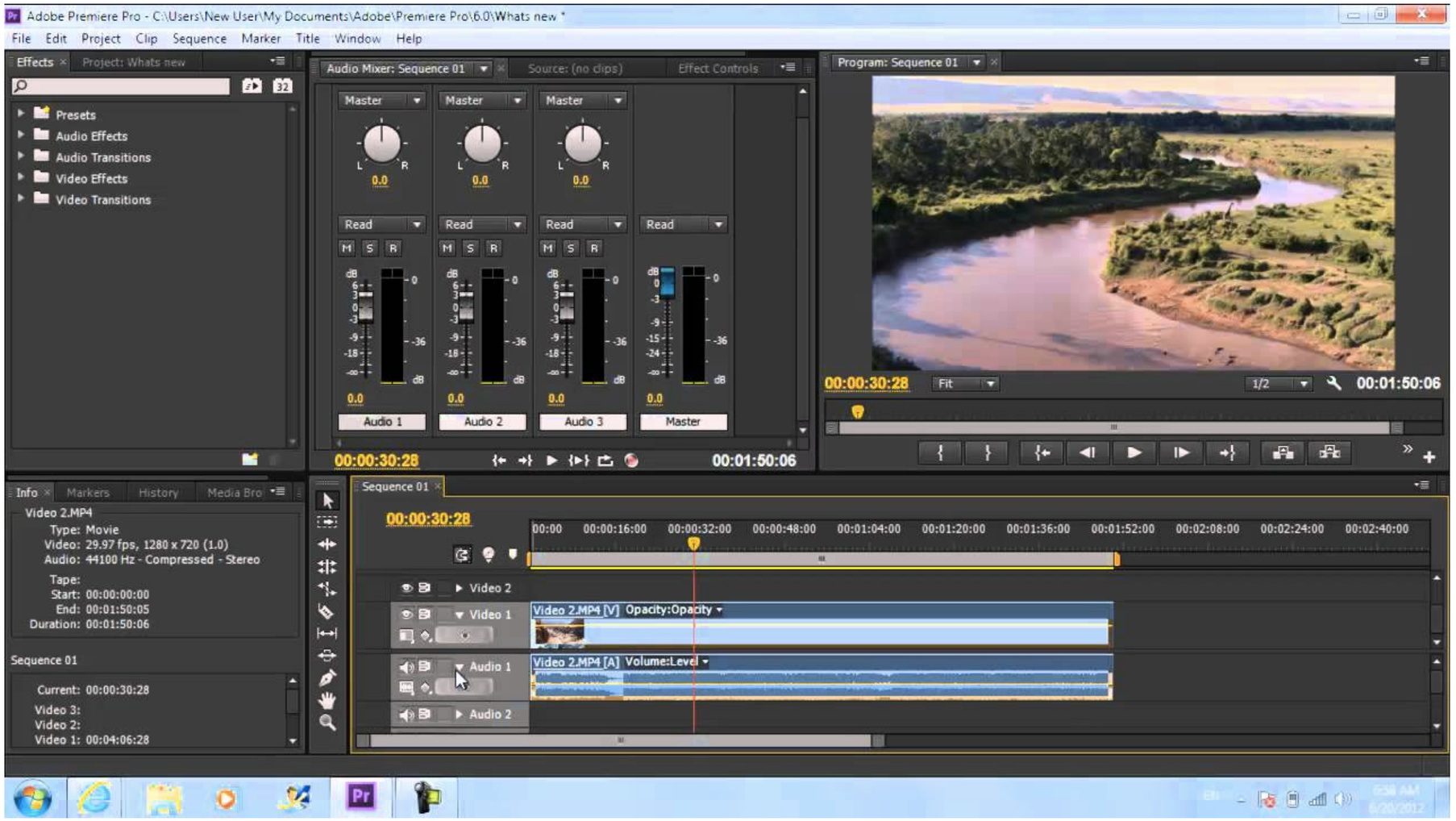
Mao, et al. "Generation and comprehension of unambiguous object descriptions." *CVPR*. 2016

Why Is Collecting Data Hard?

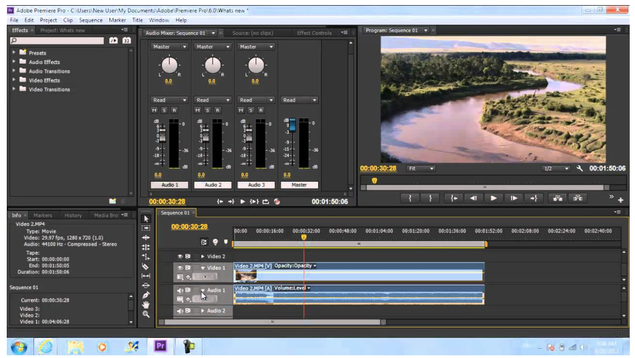
YouTube

flickr

Why Is Collecting Data Hard?



Why Is Collecting Data Hard?



Solution: View videos as a timeline.

Annotating a video with natural language

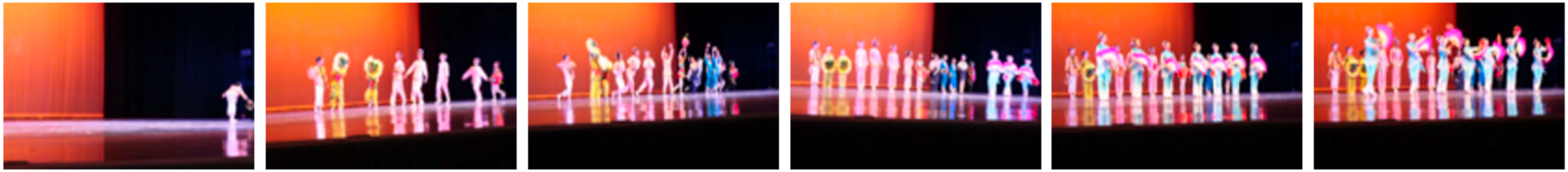
Example 1

The car drives away

A person gets out of the vehicle and walks away.

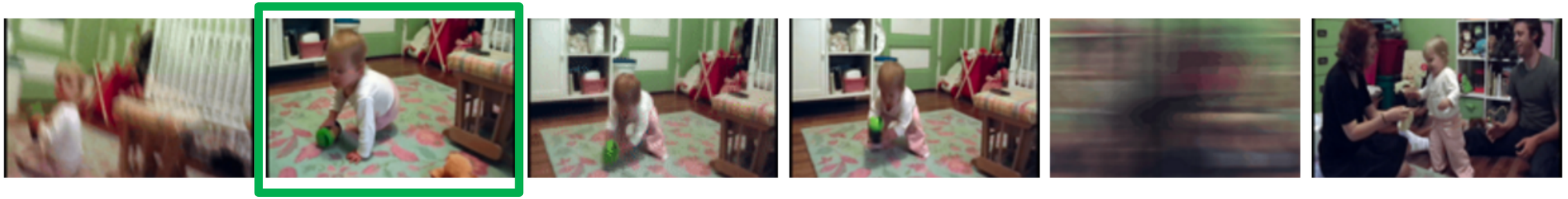
An old car rolls up to the front of a tent.

Example 2

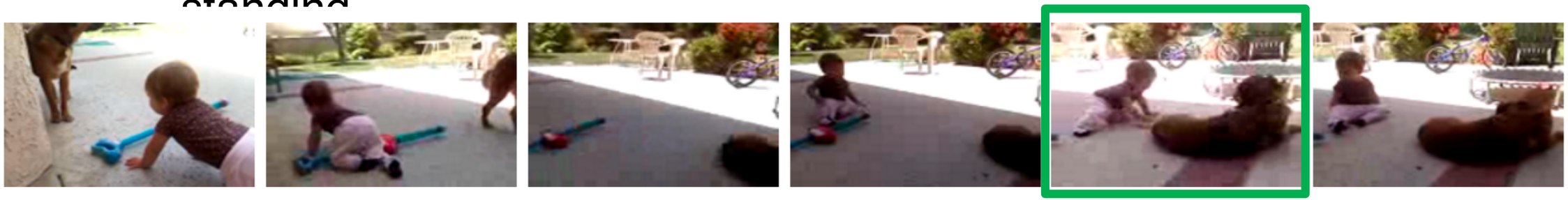


Second set of dancers come on.

What do babies do?



The **baby** goes from crouching to



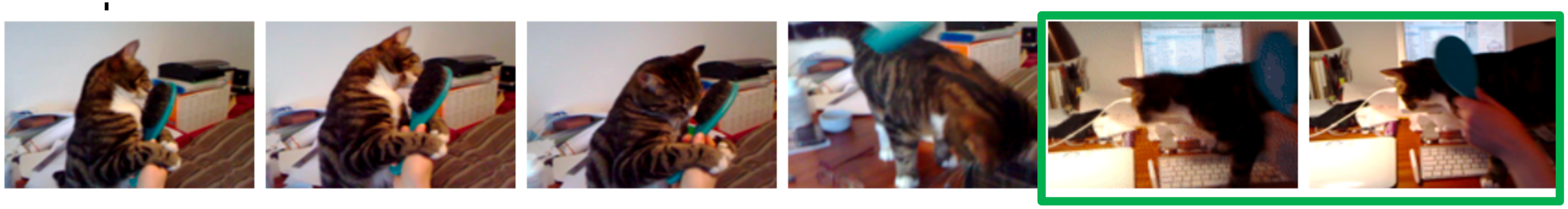
A **baby** pulls a dogs tail.

The **baby** takes something out of his

What walks?

Two ladies **walk** past the

A white dog wearing a pink cast on it’s leg **walks** over to stand by a gray

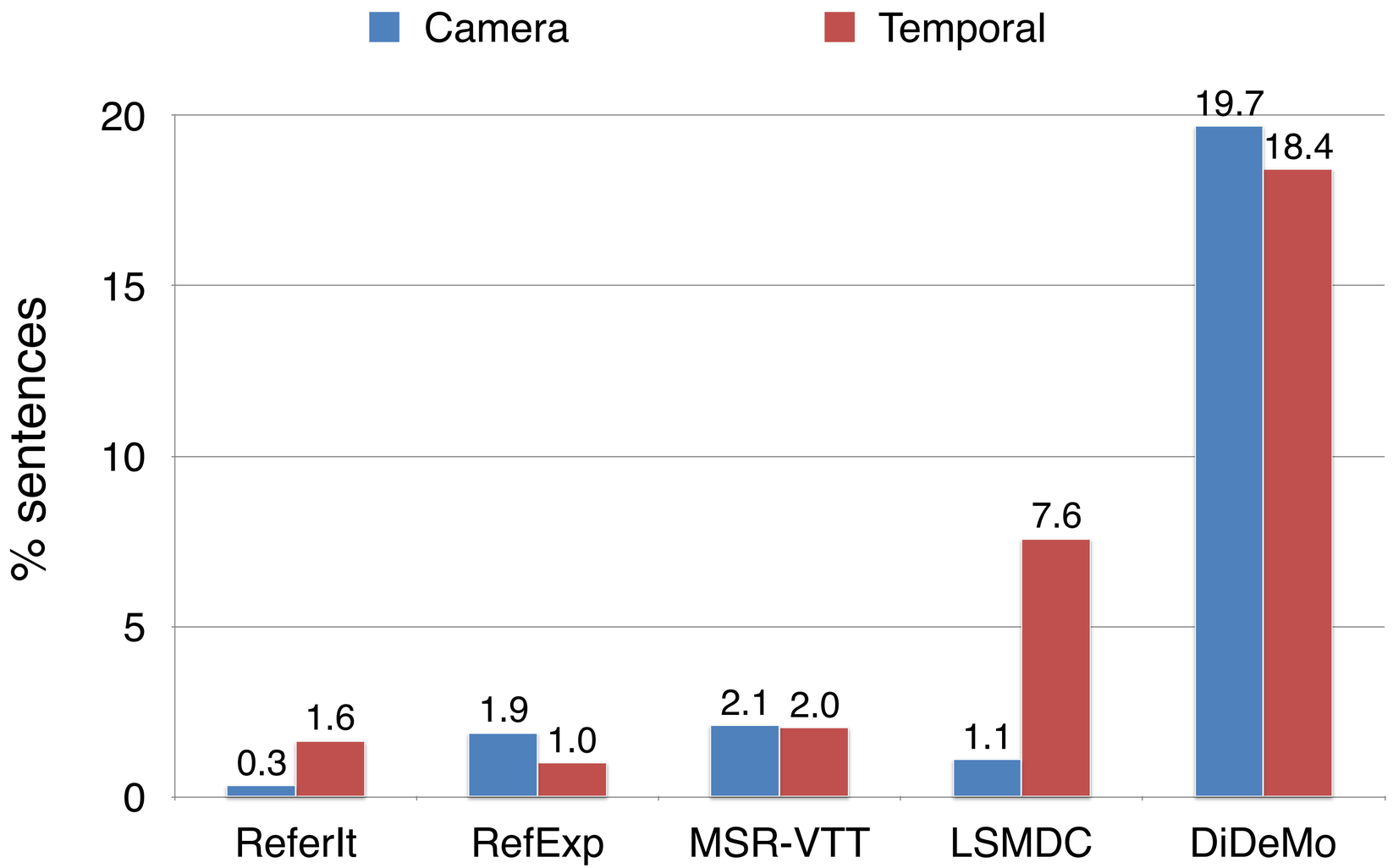


The cat **walks** across the

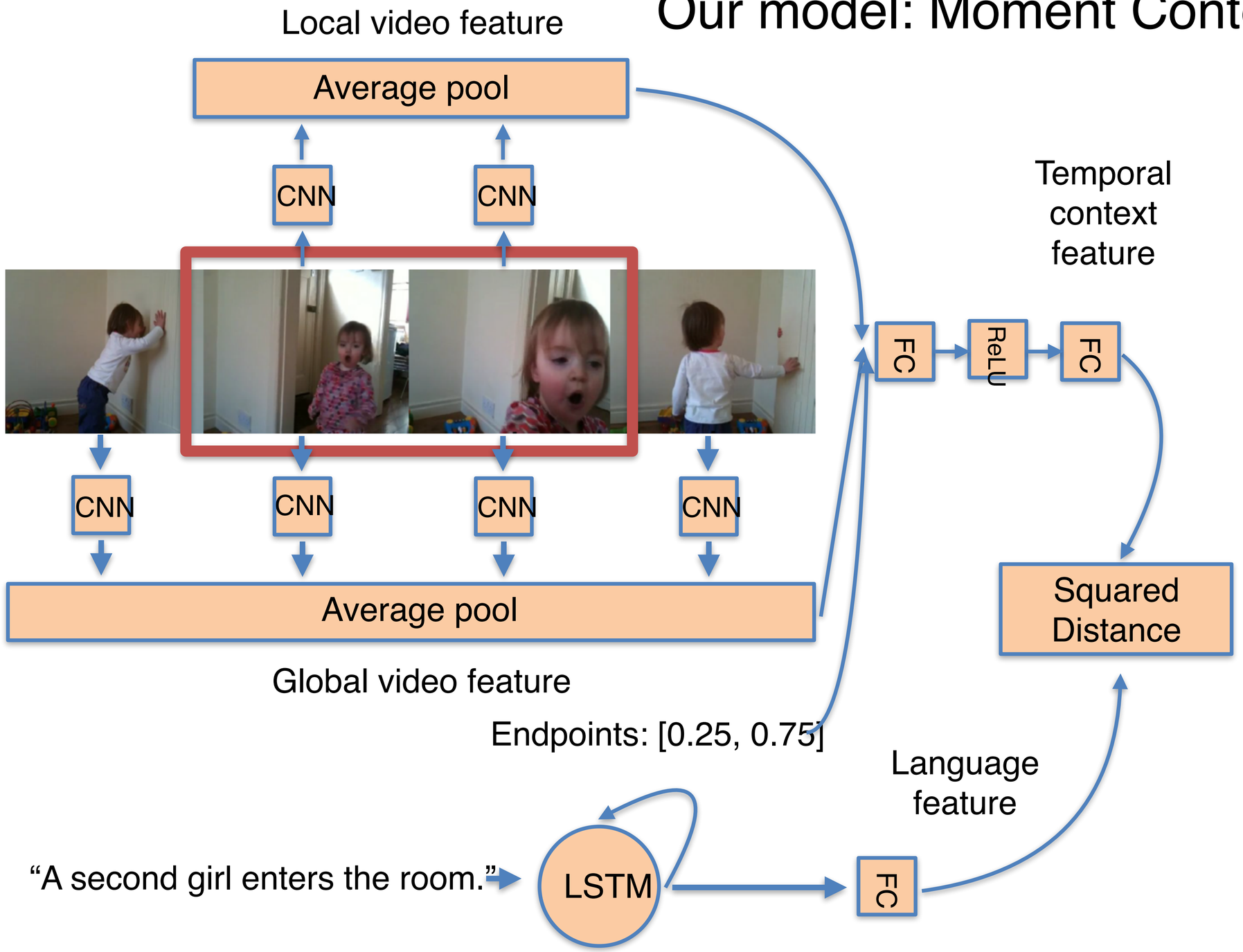
Our dataset: Distinct Describable Moments (DiDeMo)

- Dataset statistics:
  - # videos: 10,464
  - # descriptions: 40,543

DiDeMo sentences



Our model: Moment Context Networ




cat jumps up and spins around frantically